Scheduling in Wireless and Healthcare Networks

by

Yiqiu Liu

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2020 by the
Graduate Supervisory Committee:

Lei Ying, Chair
Pengyi Shi
Weina Wang
Junshan Zhang
Yanchao Zhang

ARIZONA STATE UNIVERSITY

August 2020

ABSTRACT

This dissertation studies the scheduling in two stochastic networks, a co-located wireless network and an outpatient healthcare network, both of which have a cyclic planning horizon and a deadline-related performance metric.

For the co-located wireless network, a time-slotted system is considered. A cycle of planning horizon is called a frame, which consists of a fixed number of time slots. The size of the frame is determined by the upper-layer applications. Packets with deadlines arrive at the beginning of each frame and will be discarded if missing their deadlines, which are in the same frame. Each link of the network is associated with a quality of service constraint and an average transmit power constraint. For this system, a MaxWeight-type problem for which the solutions achieve the throughput optimality is formulated. Since the computational complexity of solving the MaxWeight-type problem with exhaustive search is exponential even for a single-link system, a greedy algorithm with complexity $O(n \log(n))$ is proposed, which is also throughput optimal.

The outpatient healthcare network is modeled as a discrete-time queueing network, in which patients receive diagnosis and treatment planning that involves collaboration between multiple service stations. For each patient, only the root (first) appointment can be scheduled as the following appointments evolve stochastically. The cyclic planing horizon is a week. The root appointment is optimized to maximize the proportion of patients that can complete their care by a class-dependent deadline. In the optimization algorithm, the sojourn time of patients in the healthcare network is approximated with a doubly-stochastic phase-type distribution. To address the computational intractability, a mean-field model with convergence guarantees is proposed. A linear programming-based policy improvement framework is developed, which can approximately solve the original large-scale stochastic optimization in queueing networks of realistic sizes.

DEDICATION

To my mother, father, Yifan and Yunzhou

ACKNOWLEDGMENTS

No two PhD journeys are alike. Mine is tough but rewarding. I owe a great gratitude to all the people who have helped me along the way.

First of all, I would like to thank my adviser Prof. Lei Ying for his guidance and encouragement. In the past five years, he has provided numerous advice to my research. His knowledge has constantly enlightened me and his attention to details has deeply inspired me.

I would like to thank Prof. Pengyi Shi for giving me the opportunity to contribute to the literature of healthcare systems, which has always been my wish. Prof. Shi has dedicated a tremendous amount of effort to help me finish the second part of this dissertation.

I would like to thank Prof. Weina Wang, Prof. Junshan Zhang and Prof. Yanchao Zhang for serving on my dissertation committee and providing valuable feedback.

I also want to thank my friends and academic siblings. Especially thanks to Xin, Dheeraj and Hairi for spending an enormous amount of time discussing mathematical problems with me.

Most importantly, I would like to express my deepest gratitude to my wife Yifan Zhao and my son Yunzhou Liu. Having both of you in my life has been a miracle to me. I am forever grateful to your love, support and company.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Scheduling algorithms are key components in the operation of many stochastic systems, such as the data centers, production lines and computer operating systems. Among the various stochastic systems, a large portion of them have cyclic planning horizons and deadline-related performance metrics (see Pinedo (2012)). For example, a manufacturing center needs to make plans for production each year based on the prediction of market trends, and each order needs to be fulfilled within a certain deadline in order to acquire the revenue. For these stochastic systems, well designed scheduling algorithms can boost the performance (e.g. revenue for companies and average query delay for data centers) while reduce the operation cost.

1.1 Background

Decision-making in the form of scheduling algorithms plays an important role in the operation of many stochastic systems. A one-size-fits-all solution is to first model these stochastic systems with the Markov decision process (MDP) and then apply the dynamic programming technique (see Bertsekas (1995)). However, this method often suffers from the 'curse of dimensionality' where the state space is usually huge and the computation complexity is intractable. Recently, as deep learning (see Goodfellow *et al.* (2016)) advances, traditional dynamic programming has evolved into a large family of deep reinforcement learning algorithms (see Sutton and Barto (2018) and Arulkumaran *et al.* (2017)). Nonetheless, most of the deep reinforcement learning are data-driven and has low training efficiency, resulting in demand for huge training samples and time. To circumvent these limitations, application-specific scheduling algorithms are getting constant attentions. As they are customized for the specific stochastic system, they usually feature low computation complexities and guaranteed performance.

In the era of Internet-of-Things, there is a booming demand for wireless networks that support reliable real-time communications. A good solution is the frame-based wireless network (see Hou *et al.* (2009)) that features a cyclic planning horizon, where a fixed number of time slots are grouped into a frame. Meanwhile, with the prevalence of personal wearable devices such as smart watches and smart glasses, demand for wireless networks with lower transmit powers is also growing. In our work, we develop a scheduling algorithm under the frame-based framework to support reliable (high packet delivery ratio) and low transmit power communications.

As health care moves toward more consolidation, the resulting outpatient healthcare networks will serve an increasingly diverse patient population in terms of conditions treated but also in terms of geography and medical urgency. A key performance metric in assessing the service quality is the fraction of patients that are able to complete their itineraries by a target deadline, referred as the itinerary completion rate. In our work, we leverage the capacity allocation via patient admission scheduling to improve the itinerary completion rates, which is essentially designing the root appointment template.

1.2 Literature Review

1.2.1 Linerature Review for Scheduling in Wireless Network

Due to these emerging real-time applications of wireless networks, there has been a great interest in the development of scheduling algorithms in wireless networks to support packets with hard deadlines (see Tarello *et al.* (2008); Hou *et al.* (2009);

Jaramillo *et al.* (2011); Kang *et al.* (2013); Yang *et al.* (2015); Kang *et al.* (2015); Singh and Kumar (2016); Aditya and Rahul (2016); Liu and Ying (2016); Ewaisha and Tepedelenlioğlu (2017); Deng and Hou (2017); Zuo *et al.* (2017)). In the seminal work of Hou *et al.* (2009), Hou, Borkar and Kumar proposed a frame-based framework to tackle the problem of scheduling packets with hard deadlines and proposed a deficit counter for each data flow to measure whether the fraction of packets dropped exceeds the maximum packet dropping rate. Assuming frame-based traffic flows such that packets arrive at the beginning of each frame and the deadlines are at the end of the same frame, they proved that a low complexity scheduling algorithm, called Largest-Debt-First (LDF), is throughput optimal in co-located networks. The frame-based framework has later been generalized in Jaramillo *et al.* (2011), where packets may arrive in the middle of a frame and the deadlines may be earlier than the end of the frame. An algorithm inspired by the MaxWeight scheduling algorithm Tassiulas and Ephremides (1993) has been proposed in Jaramillo *et al.* (2011) and proved to be throughput optimal. Besides the frame-based traffic models, a geometric approach has been introduced in Kang *et al.* (2013) for general packet arrivals and deadline distributions without the frame structure, and has been used in Kang *et al.* (2013, 2015); Du and de Veciana (2016) to quantify the efficiency ratio of LDF. Providing end-to-end hard deadlines in multihop wireless networks has also been studied recently in Liu and Ying (2016); Singh and Kumar (2016); Deng and Hou (2017), where decentralized routing and scheduling solutions have been proposed to support end-to-end hard deadlines.

Despite these significant advances on wireless scheduling with hard deadlines, only a few work simultaneously addresses both hard deadlines and average power constraints. For example, Tarello *et al.* (2008) considers a finite time horizon problem of finding minimum energy to transmit all packets of each link and proposes the optimal policy through dynamic programming; Aditya and Rahul (2016) considers the problem of transmitting deadline-constrained packets over a single wireless link, and proposes an online algorithm that minimizes the transmit power; Ewaisha and Tepedelenlioğlu (2017) considers hard deadlines and average power constraints in a wireless network with a Bernoulli packet arrival and derives the optimal policy with the Lyapunov optimization techniques; Zuo *et al.* (2017) proposes a near optimal scheduling and power control algorithm assuming that each link has exactly one packet to transmit in each frame, and the packet arrives at the beginning of the frame and should be delivered before the end of the frame.

1.2.2 Linerature Review for Scheduling in Healthcare Network

The analytical framework developed in scheduling in healthcare network incorporates a number of features, including: (1) capacity optimization in an outpatient network; (2) a discrete-time queueing network model with multiple classes of patients, class-specific deadlines, blocking, and fork-join structure; (3) a phase-type representation of the sojourn time; and (4) a mean-field approximation of the stochastic blocking process. While each of these features has been studied in the literature, which we review here, to the best of our knowledge there is no comprehensive framework that integrates all of them.

Outpatient network. Outpatient scheduling has been well studied in single-station systems (see the survey paper Cayirli and Veral (2003)). In recent years, a growing amount of appointment scheduling literature has focused on multi-station systems (e.g., Wang *et al.* (2018), Wang *et al.* (2019), and Diamant *et al.* (2018)). At a high level, our context differs from the appointment scheduling research that assumes control over the timing of each patient's appointment. In contrast, our setting requires that we focus on a priori appointment *allocation*, since (1) each followup appointment is generated stochastically after the previous appointment is complete, and (2) appointments must be scheduled as soon as possible in keeping with the destination clinic business model. Hence, scheduling of subsequent appointments of an itinerary cannot reasonably be optimized, but instead follow an earliest available appointment protocol.

Wang *et al.* (2018) study patient scheduling in a two-station system, where patient care is coordinated between anesthesia and internal medicine. In this paper, similar to Huang *et al.* (2015), patient overflows are directly penalized in the objective, whereas our model endogenizes this effect through network blocking. Kazemian *et al.* (2017) uses a simulation approach to perform heuristic real-time advance daily scheduling of both clinic and surgery visits to minimize provider overtime in a system with multiple patient classes and service completion deadlines. Deglise-Hawkinson *et al.* (2018) study capacity planning in an integrated care environment, but focuses on patient scheduling to minimize time to obtain the initial appointment, whereas we model time to complete an itinerary. In addition, a number of dynamic scheduling papers model customers with different priorities that may consume single or multiple resources with holding costs for delays or overtime costs (Patrick *et al.*, 2008; Feldman *et al.*, 2014; Gocgun and Ghate, 2012). However, dynamic scheduling is not feasible in our context and hence our modeling framework differs significantly.

Hospital scheduling. The literature on hospital patient flow scheduling also considers networks of services like the one in our setting, where patients are routed to different hospital units, e.g., Bretthauer *et al.* (2011); Dai and Shi (2019); Chow *et al.* (2011); Helm and Van Oyen (2014), just to name a few. These works focus on capacity planning or scheduling to reduce workload variability, patient delay, and/or probability of exceeding capacity, where the time to complete treatment is assumed to be exogenous to the scheduling policy; we endogenize this completion time to be dependent on the network blocking in our model.

Huang *et al.* (2015) study the scheduling of patients to doctors in an ED with multiple classes of patients with class-dependent deadlines, modeling the system as a single station (doctor) with a feedback loop for in-process patients and applying a holding cost to penalize long sojourn times. The authors leverage heavy-traffic analysis and establish asymptotic optimality for their proposed policies. He *et al.* (2019) employ a hybrid robust-stochastic approach to a similar problem setting. In contrast, we consider a general, large-scale network and, rather than a cost-based approach, we optimize a completion rate (probability), which requires detailed distributional modeling. Similar to our work, Baron *et al.* (2017) develop a queueing network model with fork-join structure and study the impact of strategic idleness in reducing the chance of excessively long waits (deadline of 15 - 20 minutes) at each station. We model deadline violation for the entire itinerary and strategic idleness is not an option. Bretthauer *et al.* (2011) explicitly model blocking in hospitals and develop a queueing-based heuristic for tandem systems. They minimize the total blocking probabilities when designing capacity strategies, which is different from our objective.

Queueing network and phase-type service time. Queueing networks with phase-type distributions have been extensively studied in the literature, e.g., Gómez-Corral (2004); Liu and Whitt (2012). Phase-type models have also been widely used in sojourn time approximation (Ozawa, 2006; Haviv and van der Wal, 2008). One of the most relevant papers is Gue and Kim (2015), which develops a phase-type approximation of the sojourn time distribution for a network of multi-server queues. The main idea is to approximate service time and waiting time distributions by phase-type distributions, based on previous work by Asmussen and Møller (2001). Different from the queueing models studied in these papers, we consider a discrete-time queue on a daily time scale. Given our context of root appointment template design, we assume that arrivals come in a batched manner and each appointment takes one fixed slot. Further, we incorporate salient features including fork-join structure, time-varying arrivals and capacities. Due to these differences, the queueing dynamics are significantly different from those studied in the literature, requiring the development of new methods for characterizing the blocking probabilities and sojourn time distributions. Our numerical study (see Appendix E) shows that traditional methods can result in a significant bias in calculating the sojourn time distribution. More importantly, our contribution lies in connecting blocking probabilities with a phase-type model of sojourn time that is not studied in traditional settings. To facilitate evaluating the sojourn time distribution, we replace the blocking probability distribution with a point mass and prove asymptotic convergence.

Mean-field approximation. The final stream of literature related to this paper is in mean-field approximation. To establish the asymptotic convergence of the blocking probability distribution to the equilibrium solution of the mean field model, our proof leverages the Stein's method framework developed in Braverman and Dai (2017), who study steady-state diffusion approximations in $M/Ph/N+M$ queue; also see the tutorial (Braverman *et al.*, 2017) on applying this framework in queueing models and the references there for this line of work. Gurvich (2014) independently develops a method to prove a steady-state convergence that is similar to Theorem 1 in Braverman and Dai (2017). Ying (2018) applies this framework to characterize steady-state convergence rates to the equilibrium solution of the mean-field model. Gast *et al.* (2018) extend Ying (2018) to discrete-time Markov chains. Our discrete-time model differs from them in two main aspects. First, traditional mean-field analysis assumes the population size to be constant, whereas the total patient count in our setting is random and has an unbounded support in the steady state. This also brings the challenge when dealing with patient counts outside a bounded set, whereas most mean-field analysis papers work with bounded sets. Second, the generator of our mean-field model does not have a continuous second derivative but its first derivative can be proved to be Lipschitz. Dai and Shi (2017); Feng and Shi (2018) also apply Stein's framework for steady-state approximation in discrete-time queues, but they focus on diffusion approximation; we need the point mass to facilitate calculation, and hence, focus on the mean-field approximation.

1.3 Summary of Contributions

We first summarize the contributions to the scheduling in wireless networks in the following.

- We consider a frame-based time-slotted L-link system, in which a frame consists of T consecutive time slots. Packets arrive at the beginning of each frame, and need to be delivered before their deadlines. We assume the packets that arrive at the same frame and for the same link have the same deadline, but the deadline can vary from frame to frame. We further assume the channel conditions are static within a frame and vary from frame to frame. Given such network and traffic models, we formulated an optimization problem similar to that in Jaramillo *et al.* (2011), which is a variation of the classical MaxWeight problem. Following the standard Lyapunov analysis, it is shown that any scheduling algorithm that solves the optimization problem is throughput optimal.

- Using exhaustive search to solve the MaxWeight optimization problem is computationally expensive. Even for a single link system, the computational complexity is proved to be exponential in the summation of the deadline and the number of packets. Therefore, we propose a greedy algorithm, named PDMax. PDMax schedules packet deadlines and incremental weight gains calculated by solving an optimal power control problem defined for a single link. We prove PDMax is throughput optimal, and has computational complexity $O(LT\log(LT))$. We remark that in contrast to Hou *et al.* (2009); Jaramillo *et al.* (2011), the objective function of our MaxWeight problem is not linear in the number of scheduled packets because the transmit power is a nonlinear function of the number of packets transmitted. Because of that, packet-by-packet greedy algorithms (such as those in Hou *et al.* (2009); Jaramillo *et al.* (2011)) are no longer the right approach. The key innovation of PDMax is to map packet scheduling to time-slot scheduling where time slots are allocated to links in a greedy fashion based on the incremental gains. The incremental gains of a link are the increases of the objective function when more time slots are allocated to the link. They are calculated by solving an optimal power control problem whose objective function again is not linear (but is convex).

- Our simulation results confirm that PDMax outperforms the greedy-MaxWeight algorithm and LDF algorithm by achieving higher throughput and lower average transmit power with significant margins.

Next we summarize both the technical and piratical contributions to the scheduling in healthcare networks.

- *Analytical framework.* We introduce a discrete-time queueing network for modeling patient flow in Section 3.1.1 and formulate the template allocation problem to maximize itinerary completion as a stochastic optimization program in Section 3.1.2, with the objective function depending on the entire distribution of the itinerary completion time. The objective is non-linear in the template allocation decision variables since a patient's itinerary completion time not only depends on those who started their itineraries earlier but also on patients who enter the network later. These complex dynamics play out in the relationship between the template allocations and the blocking in the network of services, which is itself a random process.

As a building block, we first show that, in complex networks with all key features being incorporated (time-varying arrivals and capacities, stochastic itineraries, parallel appointments), we are able to characterize the itinerary completion time via a doubly-stochastic phase-type distribution, which is driven by the stochastic blocking process in the network. Different from previous works that directly assume phase-type service times or approximate sojourn time with phase-type approximations, our characterization is *exact* in such complex networks.

- *Solution algorithm.* Although the itinerary time characterization is exact, there is no simple closed-form analytical expression in the appointment allocation decisions. Combined with the large state space of the queueing network, it makes traditional optimization methods intractable. To overcome this intractability, we use a policy iteration framework to approximately solve the large-scale stochastic optimization. In the policy evaluation step, instead of evaluating the doubly-stochastic distribution that involves a large matrix power calculation and numerical integration over the distribution of the blocking process, we leverage a mean-field model to replace the blocking distribution with a point mass. This significantly reduces the computational burden. We provide a rigorous justification for this replacement by characterizing the convergence rate of the blocking process in Chapter 4. Our convergence proof builds upon the Stein's method framework (Braverman and Dai, 2017; Ying, 2018; Gast *et al.*, 2018), but requires non-trivial adaption to account for the fact that the total patient count in our model is random and can go unbounded; the conventional mean-field model assumes a fixed population size.

In the policy improvement step, we optimize capacity allocation in each iteration with the itinerary time distribution calculated using blocking probabilities from the previous iteration. This allows us to formulate the non-linear stochastic optimization as a linear program (LP) for each iteration. We add constraints to ensure that the blocking probabilities in each iteration do not deviate from the previous step too much, with provable bounds. We integrate the policy evaluation and policy improvement steps in a framework that iteratively updates the template to maximize the itinerary completion rates. In addition to the mathematical justifications, we show via a comprehensive numerical study that (1) the policy evaluation is remarkably accurate, and (2) the iterative optimization generates significantly improved templates in a range of different settings. We also show the importance of incorporating the features that make our problem difficult (e.g. parallel appointments, evaluating the full *distribution* of itinerary time) by showing that our comprehensive framework significantly outperforms simpler optimization approaches that ignore these features.

- *Value of an integrated approach.* Through our case study at the Mayo Clinic, we show that our optimization approach can significantly improve the itinerary completion rates and is computationally efficient for realistic network sizes (e.g., 26 stations). Furthermore, we identify the drivers of the improvement from our optimal template compared to the current practice, showing the value of our integrated approach. We show that template design is a multifaceted problem, and that ignoring any of the complex drivers of itinerary completion failure can lead to poor performance. The necessity of simultaneously accounting for all these complexities makes simpler optimizations and manual template design fall well short of the optimal, which highlights the practical importance of our comprehensive optimization algorithm.

- *Broader applications.* While we provide an example of how to apply our modeling framework to a destination care center, the concept of managing differentiated deadlines in a stochastic queueing network applies more broadly to a number of other business settings. Examples include new product introduction (NPI) and job shop prototyping with multiple customers. In NPI, a new product is often started with the required steps not entirely know in advance. Different new products will also have different deadlines. Deciding when to start NPI's and how much capacity to allocate to them falls squarely within our framework. In job shop prototyping, different jobs may follow different paths and require revisits due for rework. Further, jobs from different customers may require different deadlines. Hence, a similar timing and capacity allocation scheme is needed.

Chapter 2

WIRELESS SCHEDULING UNDER DEADLINE AND POWER CONSTRAINTS

2.1 Co-Located Wireless Network

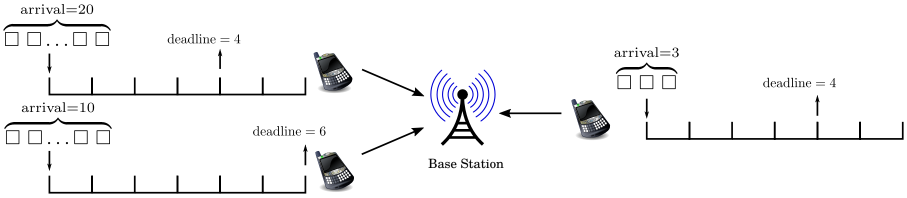


Figure 2.1: A Co-Located Network with Three Links and Frame Size 6.

We consider a co-located wireless network, e.g. an uplink/downlink cellular network as described in the following. The network consists of L links (also called users) which share a single frequency band. Assume the network is time-slotted. In each time slot, at most one link is allowed to transmit due to interference. We further assume time slots are grouped into frames such that each frame consists of T consecutive time slots. Throughout this paper, we define $\mathcal{L} = \{1, 2, \ldots, L\}$ and $\mathcal{T} = \{1, 2, \ldots, T\}$.

We assume packets arrive at the beginning of each frame. Let $\boldsymbol{A}_l(j)$ denote the number of packets arrive at link l at the beginning of frame j. We assume $\boldsymbol{A}_l(j)$ are independently and identically distributed over frames and independent across links. Furthermore, we assume $\boldsymbol{A}_l(j) \leq A_{\max}$ for all j and l, and the $\boldsymbol{A}_l(j)$ packets have the same deadline $1 \leq \boldsymbol{D}_l(j) \leq T$ which however can vary from frame to frame and from link to link. Each link l is associated with two constraints determined by the upper layer applications and devices: (1) the quality of service (QoS) constraint such that the packet dropping probability should not exceed p_l, i.e. the long-term average packet delivery ratio should be at least $1 - p_l$; and (2) the average power constraint such that the long-term average transmit power of link l should not exceed $\frac{1}{T}\beta_l$ (Watt). [1]

We further assume that the link bandwidth is B_l, and in the j-th frame, the channel gain of the link is $\boldsymbol{G}_l(j)$ and the noise level at the receiver is $\boldsymbol{N}_l(j)$ throughout the frame. Without loss of generality, we assume $\boldsymbol{N}_l(j) = 1$ and use $\boldsymbol{G}_l(j)$ to represent the channel condition so as to simplify the notations. Under the additive white Gaussian noise channel, in the t-th time slot, the relationship between the bandwidth B_l, channel condition $\boldsymbol{G}_l(j) = g_l$, transmit power w_{lt}, and number of packets transmitted s_{lt} is

$$B_l \log_2 (1 + w_{lt} g_l) = \frac{s_{lt} Z_l}{\Delta t},$$

where Z_l is the packet size and Δt is the time slot duration. Equivalently, we have

$$w_{lt} = \frac{1}{g_l}\left(2^{s_{lt}\frac{Z_l}{B_l \Delta t}} - 1\right).$$

We assume $\frac{Z_l}{B_l \Delta t} = 1$ and define function $f : \mathbb{R}_+ \times \mathbb{Z}_+ \to \mathbb{R}_+$ as:

$$f(x, y) = \frac{1}{x}\left(2^y - 1\right).$$

It follows that $w_{lt} = f(g_l, s_{lt})$.

Our model can be applied to the video streaming and conferencing scenarios, where the data arrives frame by frame. The video processing system can tolerate certain packet loss ratio. For users with mobile devices such as cellphones and laptops, maintaining a low average transmit power usage is critical to extending the battery life.

[1] Denote the power level of link l in each time slot as $\boldsymbol{f}_t$ (Watt). This constraint is equivalent to $\mathbb{E}\left[\sum_{t \in \mathcal{T}} \boldsymbol{f}_t\right] \leq \beta_l$.

2.2 Problem Formulation

In this section, we introduce the mathematical formulation of the problem, which is similar to Jaramillo *et al.* (2011).

We first define the rate-power region $\mathcal{C}(a, d, g)$ under given arrival $\boldsymbol{A} = a$, deadline $\boldsymbol{D} = d$, channel conditions $\boldsymbol{G} = g$, which is analogous to the capacity region. The elements in the rate-power region are the tuples $(\mu, \phi) = (\{\mu_l\}_{l\in\mathcal{L}}, \{\phi_l\}_{l\in\mathcal{L}})$, where μ_l and ϕ_l are long term average transmission rate and power level that can be achieved through time sharing among elements in the set of schedule s that satisfies the following constraints:

$$\sum_{1\le t\le d_l} s_{lt} \le a_l, \forall l \in \mathcal{L} \tag{2.1}$$

$$\sum_{d_l<t\le T} s_{lt} = 0, \forall l \in \mathcal{L} \tag{2.2}$$

$$\sum_{l\in\mathcal{L}} \mathbb{1}_{\{s_{lt}>0\}} \le 1, \forall t \in \mathcal{T}, \tag{2.3}$$

which is defined as $\mathcal{S}(a, d)$. Note that inequality (2.1) states that the number of packets link l transmits cannot exceed the total number of available packets; inequality (2.2) states that after deadline d_l, link l has no packets to transmit; and inequality (2.3) states that at most one link can be scheduled to transmit at each time slot. In other words, $\mathcal{C}(a, d, g)$ is the convex hull of $\mathcal{S}(a, d)$, which can be written as

$$\mathcal{C}(a, d, g) = \Bigg\{ (\bar{\mu}, \bar{\phi}) | \text{ there exists } p(s) \text{ such that}$$

$$\sum_{\bar{s}\in\mathcal{S}(a,d)} p(\bar{s}) \sum_{t\in\mathcal{T}} \bar{s}_{lt} \ge \bar{\mu}_l, \ \forall l$$

$$\sum_{\bar{s}\in\mathcal{S}(a,d)} p(\bar{s}) \sum_{t\in\mathcal{T}} f(g_l, \bar{s}_{lt}) \le \bar{\phi}_l, \ \forall l \Bigg\},$$

where $p(s) = \Pr(\boldsymbol{S} = s)$ is a probability distribution of the feasible schedule $\boldsymbol{S}$. The rate-power region under given joint distribution

$$p(a, d, g) = \Pr(\boldsymbol{A} = a, \boldsymbol{D} = d, \boldsymbol{G} = g\}$$

is

$$\begin{aligned} \mathcal{C} = \Bigg\{ (\mu, \phi) | \mu_l = & \sum_{(a,d,g)} \bar{\mu}_{(a,d,g)} p(a, d, g), \\ \phi_l = & \sum_{(a,d,g)} \bar{\phi}_{(a,d,g)} p(a, d, g), \\ & \text{for some } \left(\bar{\mu}_{(a,d,g)}, \bar{\phi}_{(a,d,g)}\right) \in \mathcal{C}(a, d, g) \Bigg\}. \end{aligned}$$

After defining the rate-power region, our task can be formulated as finding a tuple (μ, ϕ) in the rate-power region satisfying the QoS and average power constraints, which can be written as

$$\begin{aligned} & \max_{(\mu,\phi) \in C} \ 1 \\ & \text{s.t. } \mu_l \geq \lambda_l (1 - p_l), \forall l \in \mathcal{L} \\ & \phi_l \leq \beta_l, \forall l \in \mathcal{L}, \end{aligned} \tag{2.4}$$

where λ_l is the mean of the packet arrival process $\boldsymbol{A}_l(j)$.

Problem (2.4) can be solved by the virtual queue techniques. Define $\boldsymbol{S}(j)$ to be the schedule adopted in frame j where $\boldsymbol{S}_{lt}(j)$ is the number of packets link l transmits in time slot t of frame j. Each link maintains two counters (virtual queues): a deficit queue $\boldsymbol{\delta}_l(j)$ and a power queue $\boldsymbol{\theta}_l(j)$. The two virtual queues keep track of the progress of fulfilling the QoS constraints and the average transmit power constraints. They are updated at the end of each frame as follows:

$$\boldsymbol{\delta}_l(j + 1) = \left[\boldsymbol{\delta}_l(j) + \boldsymbol{A}_l(j)(1 - p_l) - \sum_{t \in \mathcal{T}} \boldsymbol{S}_{lt}(j) \right]^+ \tag{2.5}$$

$$\boldsymbol{\theta}_l(j + 1) = \left[\boldsymbol{\theta}_l(j) - \beta_l + \sum_{t \in \mathcal{T}} f\left(\mathbf{G}_l(j), \boldsymbol{S}_{lt}(j)\right) \right]^+ . \tag{2.6}$$

Note that due to the omission of time slot duration in the definition of rate-power region, for consistency, we do not need to scale $\boldsymbol{\theta}_l$ with time slot duration.

By Theorem 1 in Jaramillo *et al.* (2011), when the expected values of the virtual queue sizes are finite, both QoS and average power constraints are satisfied, which is stated in the following lemma.

Lemma 2.2.1. *If a joint power-control and scheduling algorithm selects schedule s^* at frame j such that*

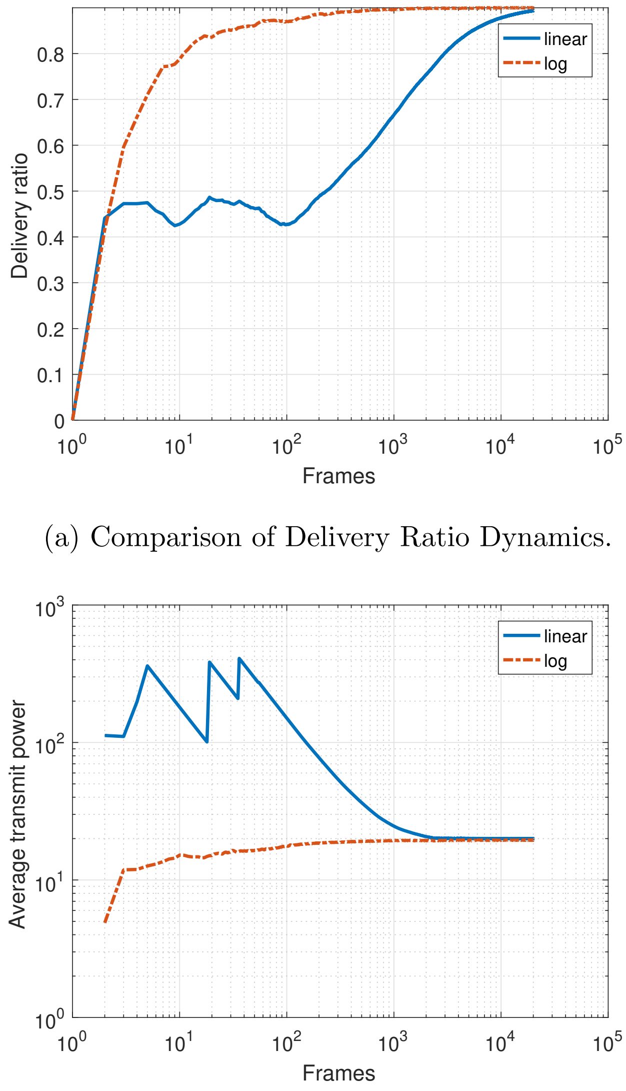
$$s^* \in \arg\max_{s \in \mathcal{S}(a,d)} \sum_{l \in \mathcal{L}} \sum_{t \in \mathcal{T}} \boldsymbol{\delta}_l(j) s_{lt} - \log(\boldsymbol{\theta}_l(j) + 1) f(\boldsymbol{G}_l(j), s_{lt}) \tag{2.7}$$

where $(\boldsymbol{A}(j), \boldsymbol{D}(j)) = (a, d)$ and update $\boldsymbol{\delta}(j)$ and $\boldsymbol{\theta}(j)$ with (2.5) and (2.6), then the algorithm is throughput optimal. □

The proof follows the standard Lyapunov drift analysis (a comprehensive introduction of the Lyapunov drift method can be found in Srikant and Ying (2014)), and is presented in appendix 2.6.1 for the completeness of the paper.

Remark: An algorithm that solves (2.7) with $\boldsymbol{\theta}_l(j)$ as the weight in the second term is also throughput optimal, which can be proved by considering a quadratic Lyapunov function. We use $\log(\boldsymbol{\theta}_l(j) + 1)$ in the objective function instead of $\boldsymbol{\theta}$ because $f(\cdot, \cdot)$ is an exponential function in s_{lt}, which brings the issue that the weight $\boldsymbol{\theta}$ for the virtual power queue grows too fast. With large $\boldsymbol{\theta}$, the algorithm will choose to transmit small amount of packets, which results in slow convergence rates. This is shown in Figure 2.2, where the parameters are chosen as described in Section 2.4 with a delivery ratio (i.e. $1 - p_l$) target of 0.9 and an average transmit power target of 2 Watt under arrival rate 17 packets/frame for each user.

Remark: Solving (2.7) with exhaustive search, even for a single link case, has a computational complexity of order $O(e^{a+T})$, with the assumption that there are a packets to transmit by the deadline of T. The proof is presented in appendix A.



(a) Comparison of Delivery Ratio Dynamics.

(b) Comparison of Average Transmit Power Dynamics

Figure 2.2: Comparison of the Convergence Rates with θ and $\log(\theta + 1)$ in the Objective Functions.

Remark: If we replace the objective function (2.7) by

$$\max \sum_{t \in \mathcal{T}} \boldsymbol{\delta}_l(j) s_{lt}$$

(i.e. ignore the power queue), and assume at most one packet can be transmitted in each time slot, then the problem becomes the classical job scheduling problem (e.g. see chapter 6.6.2 of Brassard and Bratley (1996)), which is to schedule a set of jobs such that each job is associated with a profit and a deadline. A well-known greedy algorithm that maximizes the total profit is to iteratively schedule the job with the maximum profit among all remaining jobs in the idle time slot that is closest to its deadline. For example, consider three jobs

$$\{(10, 2), (1, 1), (2, 2)\},$$

where the first number represents the profit and the second number represents the deadline. The greedy job scheduling algorithm first schedules job 1 at time slot 2, and then job 3 at time slot 1. In our setting, the profit is the virtual queue length of the link the packet is associated with. However, due to power control, a link can transmit multiple packets in one time slot and the "profits" of transmitting multiple packets are not additive. To overcome this issue, we transfer packet scheduling to time-slot scheduling and assume that link l has d_l virtual jobs with the same deadline d_l. The profit of the kth virtual job of link l is set to be $\Delta W_{l,k}$ such that

$$\sum_{t=1}^{k} \Delta W_{l,t} = \max_{s_l} \sum_{t=1}^{k} \boldsymbol{\delta}_l(j) s_{lt} - \log(\boldsymbol{\theta}_l(j) + 1) f(\boldsymbol{G}_l(j), s_{lt})$$

subject to constraints (2.1) and (2.2) for link l. Therefore, $\Delta W_{l,k}$ is the incremental gain when the number of time slots assigned to link l increases from $k - 1$ to k, and can be calculated by solving a single link power control problem (see details in the algorithm description). The virtual jobs are then scheduled using the greedy algorithm mentioned earlier. However, one *unique* constraint we have is that to get "profit" $\Delta W_{l,k}$ from virtual job k, link l should have also scheduled the virtual jobs from 1 to $k - 1$ because $\Delta W_{l,k}$ is calculated based on the assumption that link l is

Table 2.1: Summary of Notations

Notation	Description
$s = [s_{lt}]_{l \in \mathcal{L}, t \in \mathcal{T}}$	Schedule for number of packets link l transmits at time slot t
$\delta = [\delta_l]_{l \in \mathcal{L}}$	Deficit queue length for link l
$\theta = [\theta_l]_{l \in \mathcal{L}}$	Power queue length for link l
$\gamma = [\gamma_l]_{l \in \mathcal{L}}$	Log-scaled power queue length for link l
$W = [W_{l,t}]_{L \times T}$	Weight matrix
$\Delta W = [\Delta W_{l,t}]_{L \times T}$	Incremental weight gain matrix
$\lfloor x \rfloor$	$\max\{r : r \in \mathbb{Z}, r \leq x\}$
$\lceil x \rceil$	$\lfloor x \rfloor + 1$
$[x]^+$	$\max(0, x)$

given k time slots to transmit. We will show that for each l, $\Delta W_{l,k}$ is decreasing in k. Therefore, under PDMax, virtual job k of link l will be scheduled only if virtual jobs 1 to $k-1$ have also been selected.

2.3 PDMax: Power-Deadline Constrained MaxWeight

In this section, we introduce our joint power-control and scheduling algorithm, PDMax, to support both the QoS constraint and the average transmit power constraint. PDMax is throughput optimal for the traffic models defined in Section 2.1. For the convenience of readers, we summarize the notations in Table 2.1. Note that the $\lceil \cdot \rceil$ opereator is not the traditional ”ceiling” operator. In our definition, if x is an integer, $\lceil x \rceil = x + 1$.

We present PDMax together with a simple example at each step to help readers better understand the proposed algorithm. The example is a co-located network with 3 links as shown in Figure 2.1. Each frame consists of $T = 6$ time slots with the following virtual queue states and parameters at the beginning of frame j:

- The deficit queue lengths $\delta = [45,\ 10,\ 1]$,
- The power queue lengths $\theta = [205,\ 1,\ 100]$,
- The channel conditions $g = [501,\ 1566,\ 1099]$,
- Number of packet arrivals $a = [20,\ 11,\ 3]$,
- Deadlines for each link $d = [4,\ 6,\ 4]$,
- Packet dropping rate requirements $p = [0.1,\ 0.2,\ 0.3]$,
- Average power constraints $\beta = [20,\ 30,\ 40]$.

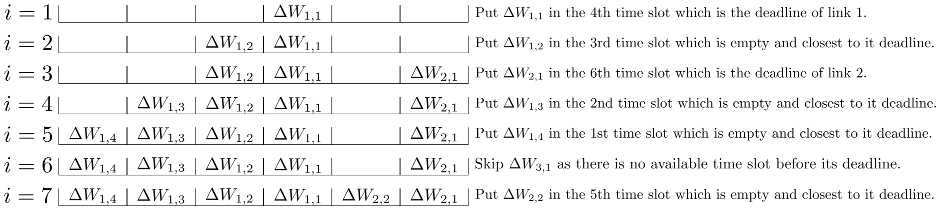
PDMax solves the optimization problem (2.7) for each frame. We now drop the frame index j to simplify the notations.

Step 1: We first calculate x_l^* for all links $l \in \mathcal{L}$,

$$x_l^* = \left[\min \left\{ \arg\max_{x \in \mathcal{X}_l} \delta_l x - \gamma_l \left(2^x - 1\right) \right\} \right]^+ \tag{2.8}$$

where $\gamma_l = \frac{1}{g_l} \log(\theta_l + 1)$ and

$$\mathcal{X}_l = \left\{ 0, a_l, \left\lfloor \log_2 \left(\frac{\delta_l}{\gamma_l \ln 2} \right) \right\rfloor, \left\lceil \log_2 \left(\frac{\delta_l}{\gamma_l \ln 2} \right) \right\rceil \right\}. \tag{2.9}$$



$i = 1$				$\Delta W_{1,1}$			Put $\Delta W_{1,1}$ in the 4th time slot which is the deadline of link 1.
$i = 2$			$\Delta W_{1,2}$	$\Delta W_{1,1}$			Put $\Delta W_{1,2}$ in the 3rd time slot which is empty and closest to it deadline.
$i = 3$			$\Delta W_{1,2}$	$\Delta W_{1,1}$		$\Delta W_{2,1}$	Put $\Delta W_{2,1}$ in the 6th time slot which is the deadline of link 2.
$i = 4$		$\Delta W_{1,3}$	$\Delta W_{1,2}$	$\Delta W_{1,1}$		$\Delta W_{2,1}$	Put $\Delta W_{1,3}$ in the 2nd time slot which is empty and closest to it deadline.
$i = 5$	$\Delta W_{1,4}$	$\Delta W_{1,3}$	$\Delta W_{1,2}$	$\Delta W_{1,1}$		$\Delta W_{2,1}$	Put $\Delta W_{1,4}$ in the 1st time slot which is empty and closest to it deadline.
$i = 6$	$\Delta W_{1,4}$	$\Delta W_{1,3}$	$\Delta W_{1,2}$	$\Delta W_{1,1}$		$\Delta W_{2,1}$	Skip $\Delta W_{3,1}$ as there is no available time slot before its deadline.
$i = 7$	$\Delta W_{1,4}$	$\Delta W_{1,3}$	$\Delta W_{1,2}$	$\Delta W_{1,1}$	$\Delta W_{2,2}$	$\Delta W_{2,1}$	Put $\Delta W_{2,2}$ in the 5th time slot which is empty and closest to it deadline.

Figure 2.3: The Process of Generating List V^*.

We then generate weight matrix $W = \{W_{l,k}\}_{l\in\mathcal{L},k\in\mathcal{K}}$ and incremental weight gain matrix $\Delta W = \{\Delta W_{l,k}\}_{l\in\mathcal{L},k\in\mathcal{K}}$ such that

$$W_{l,k} = \begin{cases} 0 & \text{if } x_l^* = 0 \\ k\left(\delta_l x_l^* - \gamma_l(2^{x_l^*} - 1)\right) & \text{if } x_l^* \geq 1, k \leq \frac{a_l}{x_l^*} \\ \delta_l a_l + \gamma_l k \left(1 - 2^{\lfloor \frac{a_l}{k} \rfloor}\left(1 + \frac{a_l}{k} - \lfloor \frac{a_l}{k} \rfloor\right)\right) & \text{if } x_l^* \geq 1, k > \frac{a_l}{x_l^*} \end{cases} \tag{2.10}$$

and

$$\Delta W_{l,k} = \begin{cases} W_{l,k} - W_{l,k-1} & \text{if } k > 1 \\ W_{l,k} & \text{if } k = 1. \end{cases} \tag{2.11}$$

Remark: Recall that the objective of PDMax is to maximize the total weight defined by deficit queues and power queues in (2.7). In the calculation above, $W_{l,k}$ represents the maximum weight gain of allocating k time slots to link l, and $\Delta W_{l,k}$ represents the incremental weight gain when the number of time slots allocated to link l increases from $k-1$ to k.

Example for Step 1: As for the simple example and link 1, we have

$$\gamma_1 = \frac{1}{G_1}\log(\theta_1 + 1) = 0.0153$$
$$\mathcal{X}_1 = \left\{0, a_1, \left\lfloor \log_2\left(\frac{\delta_1}{\gamma_1 \ln 2}\right)\right\rfloor, \left\lceil \log_2\left(\frac{\delta_1}{\gamma_1 \ln 2}\right)\right\rceil\right\} = \{0, 20, 12, 13\}$$
$$x_1^* = \left[\min\left(\arg\max_{x\in\mathcal{X}_1} \delta_1 x - \gamma_1\left(2^x - 1\right)\right)\right]^+ = 12$$

Thus for link 1 and $k = 1$, we will have

$$W_{1,1} = k\left(\delta_1 x_1^* - \gamma_1(2^{x_1^*} - 1)\right) = 477.17$$
$$\Delta W_{1,1} = W_{1,1} = 477.17$$

For link 1 and $k = 2$, we will have

$$W_{1,2} = \delta_1 a_1 + \gamma_1 k \left(1 - 2^{\lfloor \frac{a_1}{k} \rfloor}\left(1 + \frac{a_1}{k} - \left\lfloor \frac{a_1}{k} \right\rfloor\right)\right) = 868.61$$
$$\Delta W_{1,2} = W_{1,2} - W_{1,1} = 391.44$$

Step 2: We sort the entries of the ΔW matrix, which contains the incremental weight gains $\Delta W_{l,k}$ for all links $l \in \mathcal{L}$ and number of time slots $k \in \mathcal{T}$. Each entry $\Delta W_{l,k}$ also represents a virtual job. PDMax assigns a subset of all virtual jobs to the T time slots in the frame with at most one virtual job per time slot. Note that each packet is associated with a deadline and will be dropped if missing the deadline. Therefore $\Delta W_{l,k}$ can only be scheduled at a time slot no later than the d_l-th time slot. We consider $\Delta W_{l,k}$ in a descending order, where ties are broken according to index k. For given $\Delta W_{l,k}$, we check the time slots no later than d_l. If there are multiple empty time slots, we assign $\Delta W_{l,k}$ to the latest one (i.e., the one closest to d_l); otherwise, we skip $\Delta W_{l,k}$. Let

$$V^* = [(l_1, k_1), \ldots, (l_T, k_T)]$$

denote the ordered list of subscriptions associated with the incremental gains assigned to the T time slots. The number of time slots allocated to each link can be calculated from V^*.

Remark: As discussed earlier, this step is motivated by the well-known greedy scheduling algorithm for jobs with profits and deadlines Brassard and Bratley (1996), where the idea is to schedule the most profitable job among the remaining jobs to the latest possible remaining time slots. PDMax views ΔW_l, a T-dimensional vector, as the profits of the T virtual jobs belonging to link l.

Example for Step 2: Continue the example in Step 1. The sorted incremental gains are shown in Table 2.2, where we only include the first seven values. The step-by-step generation of the ordered list is illustrated in Figure 2.3, where the process takes 7 iterations indexed by i. Notice that $\Delta W_{3,1}$ was skipped because there is no idle time slot before its deadline when it was considered.

Step 3: In this step, the algorithm decides the link and transmit power for each time slot in a frame. For link l, the number of time slots k_l assigned to link l is first

Table 2.2: First Seven Largest Elements in ΔW

$\Delta W_{1,1}$	$\Delta W_{1,2}$	$\Delta W_{2,1}$	$\Delta W_{1,3}$	$\Delta W_{1,4}$	$\Delta W_{3,1}$	$\Delta W_{2,2}$
477.17	391.44	108.69	26.53	2.961	2.958	1.25

calculated from the ordered list V^* as the following, which is to count the number of times link l appears in V^*:

$$k_l = \sum_{(l',k')\in V^*} \mathbb{1}_{\{l'=l\}}.$$

Then link l transmits packets as follows:

1. If $0 < k_l \le \frac{a_l}{x_l^*}$ and $x_l^* \neq 0$, then link l transmits x_l^* packets with power $\frac{1}{g_l}\left(2^{x_l^*} - 1\right)$ in each assigned time slots;

2. If $k_l > \frac{a_l}{x_l^*}$ and $x_l^* \neq 0$, then in each of the first $k_l c$ assigned time slots, link l transmits $\left\lfloor \frac{a_l}{k_l} \right\rfloor$ packets with power $\frac{1}{g_l}\left(2^{\left\lfloor \frac{a_l}{k_l} \right\rfloor} - 1\right)$, and in each of the remaining time slots, link l transmits $\left\lceil \frac{a_l}{k_l} \right\rceil$ packets with power $\frac{1}{g_l}\left(2^{\left\lceil \frac{a_l}{k_l} \right\rceil} - 1\right)$, where $c = 1 - \left(\frac{a_l}{k_l} - \left\lfloor \frac{a_l}{k_l} \right\rfloor\right)$.

Remark: The intuition is that that x_l^* is the optimal number of packets to be transmitted if only one time slot is assigned to link l. Therefore, if link l has enough packets to transmit (case (1)), then it always sends x_l^* packets in each time slot. If link l does not have enough number of packets to transmit (case (2)), then a water-filling strategy that balances the number of packets transmitted in each time slot is optimal.

Example for Step 3: In our example, we have that $k = [4, 2, 0]$ and $x^* = [12, 14, 8]$. Therefore,

1. For link 1, $k_1 = 4 > \frac{a_1}{x_1^*}$ and $c = 1$. Thus link 1 should transmit $\left\lfloor \frac{a_1}{k_1} \right\rfloor = 5$ packets in each of the 4 time slots assigned.

2. For link 2, $k_2 = 2 > \frac{a_2}{x_2^*}$ and $c = 1 - \frac{a_2}{k_2} + \left\lfloor \frac{a_2}{k_2} \right\rfloor = 0.5$. Thus link 2 should transmit $\left\lfloor \frac{a_2}{k_2} \right\rfloor = 5$ packets in the first assigned time slot and $\left\lceil \frac{a_2}{k_2} \right\rceil = 6$ packets in the second assigned time slot.

3. For link 3, since $k_3 = 0$, it is not scheduled in the current frame.

Finally we generate the schedule for current frame as

$$s^* = \begin{pmatrix} 5 & 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 6 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Step 4: At the end of the frame, we update the deficit queue and power queue for each link according to (2.5) and (2.6).

Example for Step 4: In the example, for link 1, we have

$$\delta_1(j+1) = [45 + 20 \times 0.9 - 20]^+ = 43$$
$$\theta_1(j+1) = [1 - 20 + 0.2475]^+ = 0.$$

Theorem 2.3.1. *The schedule s^* generated by PDMax is a solution to optimization problem (2.7). Therefore, PDMax is throughput optimal. The computational complexity of PDMax is $O(LT \log(LT))$.* □

The proof of throughput optimality is presented in the appendix and consists of two steps. In the first step, we compute the closed-form solution for a single link with k time slots to transmit. In the second step, we prove that the greedy approach solves the network-wide optimization problem based on the closed form solution from the first step. We then analyze the computational complexity of PDMax to conclude the proof of the theorem.

2.4 Numerical Results

We simulated a co-located wireless network with $L = 6$ users and frame size of $T = 10$ time slots. We assumed bandwidth $B = 20$MHz, packet size $Z = 1.6$Mbits, and time slot length $\Delta t = 80$ms. Each channel was assumed to be a Gaussian channel with mean of 20 dB/Watt and variance of 10 dB/Watt^2. We compared PDMax with the greedy-MaxWeight algorithm and the Largest-Deficit-First (LDF) algorithm.

The greedy-MaxWeight algorithm prioritizes links according to the value of

$$h(\delta_l, \gamma_l, x_l^*) = \delta_l x_l^* - \gamma_l \left(2^{x_l^*} - 1\right)$$

in order to achieve a higher total weight. When a link is selected, it iteratively allocates its packets to the available time slots before its deadline. At the rth iteration, it allocates $\min\{x_l^*, \mathbf{A}_l(j) - (r-1)x_l^*\}$ packets to the last available time slot before its deadline, where x_l^* is defined in (2.8) and is the optimal number of packets to transmit over a single time slot for link l.

Consider the example in Section 2.3. Since $h(\delta_1, \gamma_1, x_1^*) = 263.03$, $h(\delta_2, \gamma_2, x_2^*) = 129.54$, and $h(\delta_3, \gamma_3, x_3^*) = 2.64$, link 1 has the highest priority. The greedy-MaxWeight algorithm first figures out the available time slots for link 1 are slot $1, 2, 3, 4$. Then starting from the last available one which is the 4th time slot, link 1 allocates $x = \min(x_1^*, a_1) = \min(12, 20) = 12$ packets. In the second last available time slot, which is the 3rd time slot, it allocates $x = \min(x_1^*, a_1 - x_1^*) = \min(12, 20 - 12) = 8$ packets. After that, the greedy-MaxWeight algorithm moves to link 2 which is second on the priority list. The available time slots for link 2 are $1, 4, 5, 6$. Link 2 allocates $\min(14, 11) = 11$ packets to the 6th time slot. For link 3, it schedules 3 packets in the 2nd time slot.

In the LDF algorithm, links are prioritized according to deficit queue lengths. We further require that the average transmit power of link l during each frame is no more than $\frac{1}{T}\beta_l$ Watt, i.e. the summation of the power level at each time slot (”energy”) is at most β_l. The number of packets to be transmitted at each time slot then depends on the number of remaining packets and the amount of remaining energy. For link l, starting from the last available time slot, in each time slot available to link l, it schedules $x = \min(\bar{a}_l, \log_2\lfloor 1 + \bar{\beta}_l G_l \rfloor)$ packets, where $\bar{a}_l$ is the number of packets that have not been scheduled and $\bar{\beta}_l$ is the remaining energy that has not been reserved. The algorithm updates the two values to be $\bar{a}_l - x$ and $\bar{\beta}_l - \frac{(2^x-1)}{G_l}$ when moving to the second last available time slot. The algorithm moves to the next link on the priority list when one of the three events occurs: 1) the unreserved energy is zero, 2) all packets have been scheduled, and 3) current time slot is later than the deadline d_l.

Consider the example in Section 2.3. Link 1 has the highest priority. LDF first figures out the available time slots for link 1, which are time slots $1, 2, 3, 4$. LDF considers the 4th time slot, and schedules $x = \min(\bar{a}_1, \log_2\lfloor 1 + \bar{\beta}_1 G_1 \rfloor) = \min(20, \log_2\lfloor 1 + 20 \times 501 \rfloor) = 13$ packets. The algorithm then updates $\bar{a}_1$ to be $\bar{a}_1 - x = 7$ and $\bar{\beta}_1$ to be $\bar{\beta}_1 - \frac{2^x-1}{G_l} = 3.65$. Similarly, LDF schedules

$$x = \min(\bar{a}_1, \log_2\lfloor 1 + \bar{\beta}_1 G_1 \rfloor) = \min(7, \log_2\lfloor 1 + 3.65 \times 501 \rfloor) = 7$$

packets in the 3rd time slot which is the second last available one. After that, all packets belonging to link 1 have been scheduled, so LDF moves to link 2 which is the second on the priority list. The available time slots for link 2 are 1,2,5,6. LDF schedules $\min(\bar{a}_2, \log_2\lfloor 1 + \bar{\beta}_1 G_1 \rfloor) = \min(11, \log_2\lfloor 1 + 30 * 1566 \rfloor) = 11$ packets for time slot 6. For link 3, LDF schedules 3 packets in the 2nd time slot.

We compared the performances of the PDMax, greedy-MaxWeight and LDF on three performance metrics: arrival rate, delivery ratio and average transmit power level. We conducted three sets of simulations. In each set of the simulations, we fixed one of the performance metrics and then plotted the tradeoff curve of the remaining two quantities for each of the three algorithms. For each set of parameters, we simulated $100,000$ frames and set deadlines for each link in each frame as a discrete uniform random variable drawn from $\left[\frac{T}{2}, T\right]$. The number of packets arrived at the beginning of each frame and channel condition are independently and identically distributed for all links.

1. In the first scenario, we fixed the arrival rate and plotted the delivery ratio versus the average transmit power. In particular, we set the number of arrivals at the beginning of each frame to be a binomial random variable with mean 20 for all the links. We varied the minimum delivery ratio, i.e. $1 - p_l$, from 0.65 to 0.975 with step size of 0.025 for all links. For a given delivery ratio, we varied the average power constraint in our simulations to identify the minimum average power level required to satisfy the minimum delivery ratio. The results are shown in Figure 2.4, where the power level for a given delivery ratio is the minimum average power level required by the corresponding algorithm. From the figure, we can see that to guarantee a delivery ratio of 0.9, PDMax requires average transmit power to be at least 5.5 Watt, while the greedy-MaxWeight algorithm requires at least 80 Watt and the LDF algorithm requires more than 100 Watt. Therefore, PDMax only requires 5% of the average transmit power required by the other two algorithms.

2. In the second scenario, we plotted the arrival rate versus the average transmit power with a fixed delivery ratio. In particular, we set the average power constraint to be 2 Watt for all links and then varied the minimum delivery ratios from 0.65 to 0.975 with a step size of 0.025 for all links. For each delivery ratio, we varied the arrival rate to identify the maximum arrival rate that each

algorithm can support for given delivery ratio and average power constraint. The results are shown in Figure 2.5. Again we can observe that PDMax can support significantly higher arrival rates under all delivery ratios; and in most cases, it doubles the maximum arrival rate compared to LDF.

3. In the third scenario, we plotted the arrival rate versus the average transmit power with a fixed delivery ratio. We set the delivery ratio to be 0.9 and varied the arrival rate from 10 to 20 with a step size of 1 for all links. For each arrival rate, we varied the average power constraint for all links to identify the minimum average power required to support the arrival rate with given delivery ratio. The result is shown in Figure 2.6. We again observe that the average transmit power required by PDMax is much smaller than the other two algorithms.

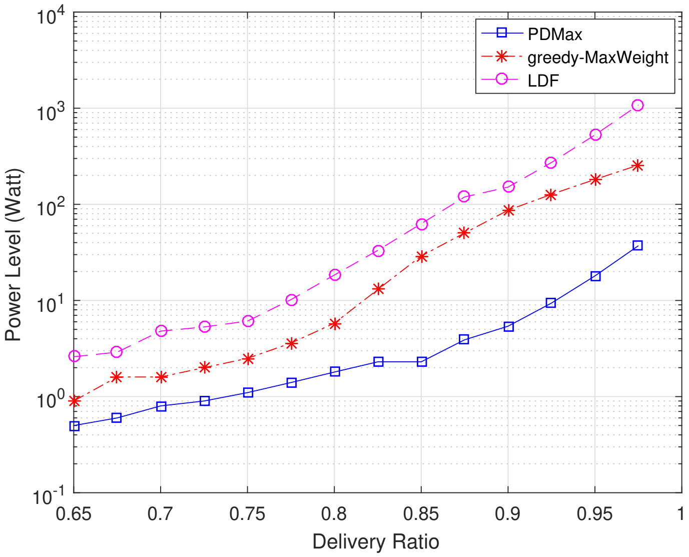


Figure 2.4: Comparison of the Minimum Average Transmit Power Required for Supporting a Fixed Arrival Rate.

In a summary, all three simulation results show that PDMax significantly outperforms the other two algorithms.

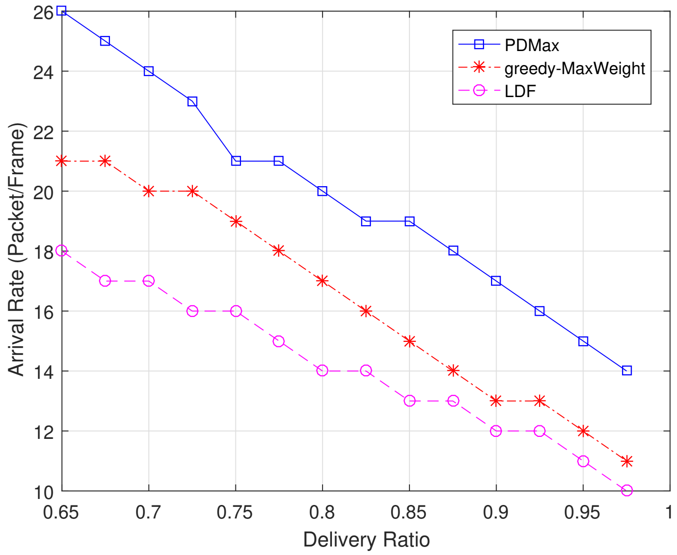


Figure 2.5: Comparison of the Maximum Arrival Rates Each Algorithms Can Support for a Fixed Average Transmit Power.

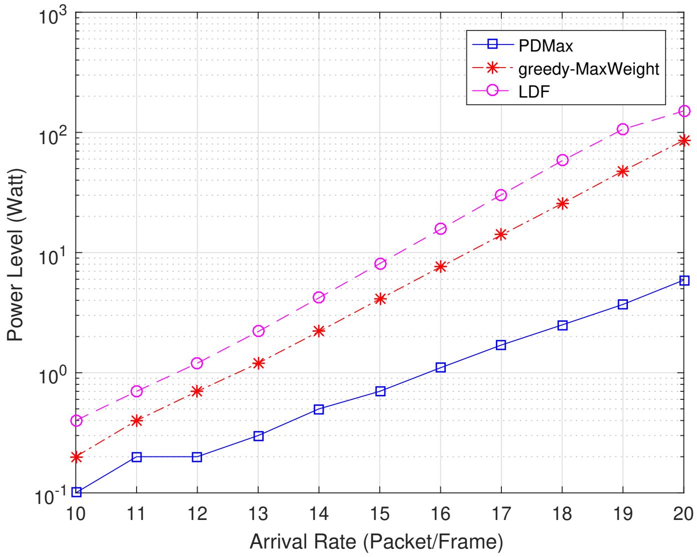


Figure 2.6: Comparison of the Minimum Average Transmit Power Required for Supporting a Fixed Delivery Ratio.

2.5 Generalization of PDMax Algorithm

Our algorithm can be further generalized in several ways. We address two of them.

- For a general value of $\frac{Z_l}{B_l \Delta t}$, let $\alpha_l = 2^{\frac{Z_l}{B_l \Delta t}}$, then

  $$f_l(x, y) = \frac{1}{x} \left(\alpha_l^y - 1\right).$$

  We can substitute 2^x, $\log_2(\cdot)$ and $\ln 2$ with α_l^x, $\log_{\alpha_l}(\cdot)$ and $\ln \alpha_l$ for each link l in the algorithm and in the proof, respectively.

- In practice, the maximum transmit power is finite, which is equivalent to that the maximum number of packets that can be transmitted in each time slot is finite, i.e. $s_{lt} \leq s_l^{\max}$. We can add $s_l^{\max}$ to the set $\mathcal{X}_l$ in step 1 of the algorithm.

2.6 Proofs

2.6.1 Proof of Lemma 2.2.1

We prove that if an algorithm solves the optimization problem defined in (2.7), then for any μ and ϕ such that

$$(\mu + \epsilon, \phi - \epsilon) \in \mathcal{C} \tag{2.12}$$

for some $\epsilon > 0$, the virtual queues $(\boldsymbol{\delta}_l(j), \boldsymbol{\theta}_l(j))$ defined in (2.5) and (2.6) are positive recurrent.

We consider the following Lyapunov function $V : \mathbb{R}_+^L \times \mathbb{R}_+^L \rightarrow \mathbb{R}_+$

$$V(\delta, \theta) = \sum_{l \in \mathcal{L}} \left(\frac{1}{2} \delta_l^2 + \int_0^{\theta_l} \log(x + 1) dx \right).$$

Define $\tilde{\theta}_l = \theta_l + \sum_{t\in\mathcal{T}} f\left(g_l, s^*_{lt}(j)\right) - \beta_l$ and $\tilde{\boldsymbol{A}}_l(j) = \boldsymbol{A}_l(j)(1-p_l)$, then we have the following:

$$\begin{aligned}
&\mathbb{E}\left(V(\boldsymbol{\delta}(j+1), \boldsymbol{\theta}(j+1)) - V(\boldsymbol{\delta}(j), \boldsymbol{\theta}(j)) \middle| \begin{pmatrix} \boldsymbol{\delta}(j) = \delta \\ \boldsymbol{\theta}(j) = \theta \end{pmatrix} \right) \\
\leq &\sum_{a,d,g} \sum_{l\in\mathcal{L}} \left(\frac{1}{2}\left(\delta_l + \tilde{a}_l - \sum_{t\in\mathcal{T}} s^*_{lt}(j)\right)^2 - \frac{1}{2}\delta_l^2 \right. \\
&\left. + \int_{\theta_l}^{\left[\tilde{\theta}_l\right]^+} \log(x+1)dx \right) p(a,d,g) \\
\leq &\sum_{a,d,g} \sum_{l\in\mathcal{L}} \left(\frac{1}{2}\tilde{a}_l^2 + \frac{1}{2}\left(\sum_{t\in\mathcal{T}} s^*_{lt}(j)\right)^2 + \delta_l \left(\tilde{a}_l - \sum_{t\in\mathcal{T}} s^*_{lt}(j)\right) \right. \\
&\left. + \int_{\theta_l}^{\left[\tilde{\theta}_l\right]^+} \log(x+1)dx \right) p(a,d,g).
\end{aligned}$$

For $\theta_l > \beta_l$, we have that $\left[\tilde{\theta}_l\right]^+ = \tilde{\theta}_l$. Assuming $\theta_l > \beta_l$, by the extreme value theorem, we have

$$\begin{aligned}
&\int_{\theta_l}^{\left[\tilde{\theta}_l\right]^+} \log(x+1)dx \\
\leq &\left(\tilde{\theta}_l - \theta_l\right) \log\left(\tilde{\theta}_l + 1\right) \\
= &\left(\tilde{\theta}_l - \theta_l\right) \log\left(\theta_l + 1\right) + \left(\tilde{\theta}_l - \theta_l\right) \log\left(\frac{\tilde{\theta}_l + 1}{\theta_l + 1}\right),
\end{aligned}$$

where

$$\begin{aligned}
&\left(\tilde{\theta}_l - \theta_l\right) \log\left(\frac{\tilde{\theta}_l + 1}{\theta_l + 1}\right) \\
\leq &\left(\sum_{t\in\mathcal{T}} f(g_l, a_l) - \beta_l\right) \log\left(1 + \frac{\sum_{t\in\mathcal{T}} f(g_l, a_l) - \beta_l}{\beta_l + 1}\right) \\
:= &C_1(a,d,g)
\end{aligned}$$

and $C_1(a,d,g)$ is independent of the lengths of the virtual queues. can be bounded by a constant independent of δ_l and θ_l because $\theta_l > \beta_l$ and $\tilde{\theta}_l - \theta_l$ is bounded.

For the case $\theta_l \leq \beta_l$, by adding and subtracting the term $\left(\tilde{\theta}_l - \theta_l\right) \log\left(\theta_l + 1\right)$, we can have

$$\begin{aligned}\int_{\theta_l}^{\left[\tilde{\theta}_l\right]^+} \log(x+1)dx = &\left(\tilde{\theta}_l - \theta_l\right) \log\left(\theta_l + 1\right) \\ &+ \int_{\theta_l}^{\left[\tilde{\theta}_l\right]^+} \log(x+1)dx - \left(\tilde{\theta}_l - \theta_l\right) \log\left(\theta_l + 1\right),\end{aligned}$$

where

$$\begin{aligned}&\int_{\theta_l}^{\left[\tilde{\theta}_l\right]^+} \log(x+1)dx - \left(\tilde{\theta}_l - \theta_l\right) \log\left(\theta_l + 1\right) \\ \leq &\left(\sum_{t \in \mathcal{T}} f(g_l, a_l)\right) \log\left(1 + \sum_{t \in \mathcal{T}} f(g_l, a_l)\right) \\ &+ \left|\sum_{t \in \mathcal{T}} f(g_l, a_l) - \beta_l\right| \log(\beta_l + 1) := C_2(a, d, g)\end{aligned}$$

and $C_2(a, d, g)$ is independent of the lengths of the virtual queues. Therefore, we have

$$\begin{aligned}&\mathbb{E}\left(V(\boldsymbol{\delta}(j+1), \boldsymbol{\theta}(j+1)) - V(\boldsymbol{\delta}(j), \boldsymbol{\theta}(j)) \middle| \begin{pmatrix} \boldsymbol{\delta}(j) = \delta \\ \boldsymbol{\theta}(j) = \theta \end{pmatrix} \right) \\ \leq &\sum_{a,d,g} \sum_{l \in \mathcal{L}} p(a, d, g) \left(\delta_l \left(\tilde{a}_l - \sum_{t \in \mathcal{T}} s^*_{lt} \right) \right. \\ &+ \log(\theta_l + 1) \left(\sum_{t \in \mathcal{T}} f\left(g_l, s^*_{lt}(j)\right) - \beta_l \right) \\ &+ \left. \left(\frac{1}{2}\tilde{a}_l^2 + \frac{1}{2} \left(\sum_{t \in \mathcal{T}} s^*_{lt}(j) \right)^2 + C_1(a, d, g) + C_2(a, d, g) \right) \right).\end{aligned}$$

Following the definition of the rate-power region $\mathcal{C}$ and condition (2.12), we get

$$\begin{aligned}&\mathbb{E}\left(V(\boldsymbol{\delta}(j+1), \boldsymbol{\theta}(j+1)) - V(\boldsymbol{\delta}(j), \boldsymbol{\theta}(j)) \middle| \begin{pmatrix} \boldsymbol{\delta}(j) = \delta \\ \boldsymbol{\theta}(j) = \theta \end{pmatrix} \right) \\ \leq &- \epsilon \sum_{l \in \mathcal{L}} (\delta_l + \log(\theta_l + 1)) + C',\end{aligned}$$

where C' is a constant independent of δ and θ. From Theorem 3.3.7 in Srikant and Ying (2014), we can conclude that the process is positive recurrent.

2.6.2 Proof of Theorem 2.3.1

To prove the main theorem, we first prove following lemmas concerning the power control of a single link with k time slots for transmitting packets.

Lemma 2.6.1. *Consider a link with $kx > 0$ packets to transmit over k time slots, and define set*

$$\mathcal{Y}_{(kx)}^{(k)} = \left\{ y : y = [y_1, \ldots, y_k], \sum_{t=1}^{k} y_t = kx, y_t \in \mathbb{Z}_+ \right\},$$

and

$$y_{(kx)}^{(k)*} = [x, \ldots, x].$$

Then

$$y_{(kx)}^{(k)*} \in \arg\max_{y \in \mathcal{Y}_{(kx)}^{(k)}} H(y),$$

where

$$H(y) = \sum_{t=1}^{k} \delta y_t - \gamma(2^{y_t} - 1).$$

Proof. Let us consider a general solution

$$y = [x + \epsilon_1, \ldots, x + \epsilon_k].$$

Since

$$\sum_{t=1}^{k} y_t = kx,$$

it follows that

$$\sum_{t=1}^{k} \epsilon_t = 0.$$

Then maximizing $H(y)$ is equivalent to solving the following optimization problem, which is convex:

$$\begin{aligned} \max_{\epsilon=[\epsilon_1,\ldots,\epsilon_k]} \quad & \sum_{t=1}^{k} \left(\delta(x+\epsilon_t) - \gamma(2^{x+\epsilon_t}-1)\right) \\ \text{subject to} \quad & \sum_{t=1}^{k} \epsilon_t = 0. \end{aligned}$$

The Lagrangian of this optimization problem is

$$L(\epsilon,\xi) = \delta kx + \gamma k - \gamma 2^x \sum_{t=1}^{k} 2^{\epsilon_t} + \xi \sum_{t=1}^{k} \epsilon_t.$$

From the KKT conditions Srikant and Ying (2014), we have

$$\begin{aligned} \frac{\partial L}{\partial \epsilon_t} &= \xi - \gamma 2^{x+\epsilon_t} \ln 2 = 0, \\ \xi \sum_{t=1}^{k} \epsilon_t &= 0. \end{aligned} \tag{2.13}$$

It is easy to see that the only solution to (2.13) is

$$\epsilon = [0,\ldots,0].$$

Due to the convexity of the objective function, we conclude that $\epsilon^* = [0,\ldots,0]$ is the global maximum, and the lemma holds. □

Lemma 2.6.2. *Consider a single link with $kx+m$ packets to be transmitted over k time slots, where $0 \le m < k$ and $x > 0$. Define*

$$\mathcal{Y}^{(k)}_{(kx+m)} = \left\{ y : y = [y_1,\ldots,y_k], \sum_{t=1}^{k} y_t = kx+m, y_t \in \mathbb{Z}_+ \right\},$$

and

$$y^{(k)*}_{(kx+m)} = \left[\underbrace{x+1,\ldots,x+1}_{m}, \underbrace{x,\ldots,x}_{k-m} \right].$$

We have

$$y^{(k)*}_{(kx+m)} \in \operatorname*{arg\,max}_{y \in \mathcal{Y}^{(k)}_{(kx+m)}} H(y).$$

Proof. When $m = 0$, Lemma 2.6.2 is equivalent to Lemma 2.6.1. Let's consider the case that $m \geq 1$ and use induction. Define

$$h(x) = \delta x - \gamma(2^x - 1).$$

For $m = 1$, let $y = [y_1, \ldots, y_k]$ be an arbitrary element in $\mathcal{Y}^{(k)}_{(kx+1)}$ and we construct $\hat{y}$ by ordering elements in y from the maximum to the minimum such that $\hat{y}_1 \geq \cdots \geq \hat{y}_k$. It is easy to verify that $\hat{y} \in \mathcal{Y}^{(k)}_{(kx+1)}$ and $H(y) = H(\hat{y})$.

For $\hat{y}_1$, the first element of $\hat{y}$, it must satisfy that $\hat{y}_1 \geq x + 1$ other wise $\sum_{i=1}^{k} \hat{y}_i \leq kx$. If $\hat{y}_1 = x + 1$, then

$$H(\hat{y}) = h(x+1) + H\left(\hat{y}_{-1}\right),$$

where $\hat{y}_{-1} = [\hat{y}_2, \ldots, \hat{y}_k]$ and $\hat{y}_{-1} \in \mathcal{Y}^{(k-1)}_{((k-1)x)}$. From Lemma 2.6.1 we have $H\left(\hat{y}_{\{1\}}\right) \leq H\left(y^{(k-1)*}_{((k-1)x)}\right)$ and $y^{(k-1)*}_{((k-1)x)} = [\underbrace{x, \ldots, x}_{k-1}]$. It follows that

$$H(y) = H(\hat{y}) \leq h(x+1) + H\left(y^{(k-1)*}_{((k-1)x)}\right) = H\left(y^{(k)*}_{(kx+1)}\right).$$

Now if $\hat{y}_1 > x + 1$, then $\exists i \in \{2, \ldots, k\}$ such that $\hat{y}_i < x + 1$, because otherwise,

$$\sum_{t=1}^{k} \hat{y}_t = \sum_{t=1}^{k} y_t > kx + 1$$

which means $y \notin \mathcal{Y}^{(k)}_{(kx+1)}$.

Let $\Delta_u = x+1-\hat{y}_1$ and $\Delta_d = \hat{y}_i - x - 1$, where $\Delta_u \geq 1$ and $\Delta_d \geq 1$. It follows that

$$\begin{aligned}
&h(\hat{y}_1) + h(\hat{y}_i) - h(x+1) - h(x+1+\Delta_u-\Delta_d) \\
=&\delta\hat{y}_1 - \gamma(2^{\hat{y}_1}-1) + \delta\hat{y}_i - \gamma(2^{\hat{y}_i}-1) \\
&-\delta(x+1) + \gamma(2^{x+1}-1) \\
&-\delta(x+1+\Delta_u-\Delta_d) + \gamma(2^{x+1+\Delta_u-\Delta_d}-1) \qquad (2.14) \\
=&\gamma\left(2^{x+1} + 2^{x+1+\Delta_u-\Delta_d} - 2^{x+1+\Delta_u} - 2^{x+1-\Delta_d}\right) \\
=&\gamma 2^{x+1}(2^{\Delta_u}-1)(2^{-\Delta_d}-1) \\
<&0.
\end{aligned}$$

Let $y_{\{1,i\}} = [y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_k]$, and we have

$$\begin{aligned}
H(y) =&H(\hat{y}) \\
=&h(\hat{y}_1) + h(\hat{y}_i) + H(\hat{y}_{\{1,i\}}) \\
<&h(x+1) + h(x+1+\Delta_u-\Delta_d) + H(\hat{y}_{\{1,i\}}) \\
\leq&h(x+1) + H\left(y^{(k-1)*}_{((k-1)x)}\right) \\
=&H\left(y^{(k)*}_{(kx+1)}\right)
\end{aligned}$$

Therefore we conclude that for $m=1$

$$y^{(k)*}_{(kx+1)} = [x+1, \underbrace{x, \ldots, x}_{k-1}] \in \arg\max_{y \in \mathcal{Y}^{(k)}_{kx+1}} H(y).$$

To use induction, suppose that for $1 \leq n < k-1$, we have

$$y^{(k)*}_{(kx+n)} = [\underbrace{x+1, \ldots, x+1}_{n}, \underbrace{x, \ldots, x}_{k-n}] \in \arg\max_{y \in \mathcal{Y}^{(k)}_{(kx+n)}} H(y).$$

Now consider $m = n+1$, and let $y = [y_1, \ldots, y_k]$ be an arbitrary element in $\mathcal{Y}^{(k)}_{(kx+n+1)}$. We construct $\hat{y}$ by sorting elements in y from the maximum to the minimum, i.e. $\hat{y}_1 \geq \cdots \geq \hat{y}_k$. Then we have $H(\hat{y}) = H(y)$.

If $\hat{y}_1 = x + 1$, then $H(\hat{y}) = h(x+1) + H\left(y_{\{1\}}\right)$, where $\hat{y}_{\{1\}} = [\hat{y}_2, \ldots, \hat{y}_k]$ and $\hat{y}_{\{1\}} \in \mathcal{Y}^{(k-1)}_{((k-1)x+n)}$. By the assumption of the induction, we have

$$\begin{aligned} y^{(k-1)*}_{((k-1)x+n)} =&[\underbrace{x+1, \ldots, x+1}_{n}, \underbrace{x, \ldots, x}_{k-1-n}] \\ &\in \underset{y \in \mathcal{Y}^{(k-1)}_{((k-1)x+n)}}{\arg\max} \; H(y). \end{aligned}$$

It follows that

$$\begin{aligned} H(y) =&H(\hat{y}) \\ =&h(x+1) + H\left(\hat{y}_{\{1\}}\right) \\ \leq&h(x+1) + H\left(y^{(k-1)*}_{((k-1)x+n)}\right) \\ =&H\left(y^{(k)*}_{(kx+n+1)}\right). \end{aligned}$$

If $\hat{y}_1 > x + 1$, then $\exists i \in \{2, \ldots, k\}$ such that $\hat{y}_i < x + 1$, because otherwise

$$\sum_{t=1}^{k} \hat{y}_t = \sum_{t=1}^{k} y_t > kx + k,$$

which means $y \notin \mathcal{Y}^{(k)}_{(kx+n+1)}$. Let $\hat{y}_1 = x + 1 + \Delta_u$, $\hat{y}_i = x + 1 - \Delta_d$, where $\Delta_u \geq 1$ and $\Delta_d \geq 1$. According to inequality (2.14), we have

$$\begin{aligned} H(y) =&H(\hat{y}) \\ =&h(\hat{y}_1) + h(\hat{y}_i) + H(\hat{y}_{\{1,i\}}) \\ <&h(x+1) + h(x+1+\Delta_u - \Delta_d) + H(\hat{y}_{\{1,i\}}) \\ \leq&h(x+1) + H\left(y^{(k-1)*}_{((k-1)x+n)}\right) \\ =&H\left(y^{(k)*}_{(kx+n+1)}\right). \end{aligned}$$

Therefore, when $m = n + 1$, we conclude

$$y^{(k)*}_{(kx+n+1)} = [\underbrace{x+1, \ldots, x+1}_{n+1}, \underbrace{x, \ldots, x}_{k-n-1}] \in \underset{y \in \mathcal{Y}^{(k)}_{(kx+n+1)}}{\arg\max} \; H(y).$$

In a summary, by induction, we conclude that for $1 \leq m < k$, the following relation holds:

$$y^{(k)*}_{(kx+m)} = [\underbrace{x+1, \ldots, x+1}_{m}, \underbrace{x, \ldots, x}_{k-m}] \in \arg\max_{y \in \mathcal{Y}^{(k)}_{(kx+m)}} H(y).$$

□

Lemma 2.6.3. *Consider a single-link network with $a \geq 1$ packets to transmit over k time slots such that $0 < k \leq d$. The optimal schedule, denoted by s^*, that solves (2.7) for the single-link network satisfies*

1. *If $x^* = 0$, then $s^* = [0, \ldots, 0]$.*
2. *If $x^* \geq 1$ and $0 < k \leq \frac{a}{x^*}$, then*

$$s^* = [x^*, \ldots, x^*].$$

3. *If $x^* \geq 1$ and and $k > \frac{a}{x^*}$, then*

$$s^* = \left[\underbrace{\left\lfloor \frac{a}{k} \right\rfloor, \ldots, \left\lfloor \frac{a}{k} \right\rfloor}_{kc}, \underbrace{\left\lceil \frac{a}{k} \right\rceil, \ldots, \left\lceil \frac{a}{k} \right\rceil}_{k(1-c)} \right],$$

where $c = 1 - \left(\frac{a}{k} - \left\lfloor \frac{a}{k} \right\rfloor\right)$.

Proof. It is trivial to prove the result when $x^* = 0$ so we focus on the proof when $x^* \geq 1$.

We first prove that for a single-link network to transmit a certain number of packets in a given number of time slots, the strategy defined by Lemma 2.6.2 solves (2.7). Then we prove under such strategy, the more packets transmitted, the higher the total weight will be.

Consider $0 < k \leq \frac{a}{x^*}$, which implies $kx^* \leq a$. Let s be an arbitrary element in $\mathcal{S}(a, k)$. We first construct $\hat{s}$ from s by ordering elements in s_i from the maximum to the minimum, i.e. $\hat{s}_1 \geq \cdots \geq \hat{s}_k$. It follows that $H(s) = H(\hat{s})$. Then we construct

$$\hat{s}^{(i)} = [\underbrace{x^*, \ldots, x^*}_{i}, \hat{s}_{i+1}, \ldots, \hat{s}_k].$$

It follows that

$$H(\hat{s}) = H(\hat{s}^{(0)}) \leq \cdots \leq H(\hat{s}^{(k)}).$$

Thus we have

$$H(\hat{s}^{(k)}) \geq H(\hat{s}) = H(s), \forall s \in \mathcal{S}(a, k),$$

which implies that

$$s^* = \hat{s}^{(k)} = [x^*, \ldots, x^*] \in \arg\max_{s \in \mathcal{S}(a,k)} H(s)$$

when $1 \leq k \leq \frac{a}{x^*}$.

For $k > \frac{a}{x^*}$ that implies $kx^* > a$ and $\frac{a}{k} < x^*$, we have $\lfloor \frac{a}{k} \rfloor < \lceil \frac{a}{k} \rceil \leq x^*$. Here we first prove that if all packets are required to be transmitted, then the strategy defined in Lemma 2.6.2 is optimal. Let

$$\mathcal{S}_{(a)}^{(k)} = \left\{ s : s = [s_1, \ldots, s_k], \sum_{t=1}^{k} s_t = a, s_t \in \mathbb{Z}_+ \right\}.$$

We consider two cases of c in the following.

1. $c = 1$ implies $\frac{a}{k} = \lfloor \frac{a}{k} \rfloor$. From Lemma 2.6.1, we have

$$s_{(a)}^{(k)*} = \left[\left\lfloor \frac{a}{k} \right\rfloor, \ldots, \left\lfloor \frac{a}{k} \right\rfloor \right] \in \arg\max_{s \in \mathcal{S}_{(a)}^{(k)}} H(s).$$

2. If $0 < c < 1$, then we have $a = k\lfloor \frac{a}{k} \rfloor + k(1-c)$ and $0 < k(1-c) < k$. By Lemma 2.6.2, we have

$$s^{(k)*}_{(a)} = s^{(k)*}_{(k\lfloor \frac{a}{k} \rfloor + k(1-c))} = \left[\underbrace{\left\lfloor \frac{a}{k} \right\rfloor + 1, \dots, \left\lfloor \frac{a}{k} \right\rfloor + 1}_{k(1-c)}, \underbrace{\left\lfloor \frac{a}{k} \right\rfloor, \dots, \left\lfloor \frac{a}{k} \right\rfloor}_{kc} \right]$$

$$= \left[\underbrace{\left\lfloor \frac{a}{k} \right\rfloor, \dots, \left\lfloor \frac{a}{k} \right\rfloor}_{kc}, \underbrace{\left\lceil \frac{a}{k} \right\rceil, \dots, \left\lceil \frac{a}{k} \right\rceil}_{k(1-c)} \right] \in \arg\max_{s \in \mathcal{S}^{(k)}_{(a)}} H(s)$$

Therefore, we have

$$s^{(k)*}_{(a)} = \left[\underbrace{\left\lfloor \frac{a}{k} \right\rfloor, \dots, \left\lfloor \frac{a}{k} \right\rfloor}_{kc}, \underbrace{\left\lceil \frac{a}{k} \right\rceil, \dots, \left\lceil \frac{a}{k} \right\rceil}_{k(1-c)} \right] \in \arg\max_{s \in \mathcal{S}^{(k)}_{(a)}} H(s).$$

Now we prove the total weight increases as the number of packets transmitted using the policy in Lemma 2.6.2 increases. Let $m \in \mathbb{Z}_+$ and $m \leq a-1$. We first note

$$s^{(k)*}_{(a-m)} = \left[\underbrace{\left\lfloor \frac{a-m}{k} \right\rfloor, \dots, \left\lfloor \frac{a-m}{k} \right\rfloor}_{kc_m}, \underbrace{\left\lceil \frac{a-m}{k} \right\rceil, \dots, \left\lceil \frac{a-m}{k} \right\rceil}_{k(1-c_m)} \right] \in \arg\max_{s \in \mathcal{S}^{(k)}_{(a-m)}} H(s),$$

where $c_m = 1 - \left(\frac{a-m}{k} - \left\lfloor \frac{a-m}{k} \right\rfloor \right)$; and

$$s^{(k)*}_{(a-m-1)} = \left[\underbrace{\left\lfloor \frac{a-m-1}{k} \right\rfloor, \dots, \left\lfloor \frac{a-m-1}{k} \right\rfloor}_{kc_{m-1}}, \underbrace{\left\lceil \frac{a-m-1}{k} \right\rceil, \dots, \left\lceil \frac{a-m-1}{k} \right\rceil}_{k(1-c_{m-1})} \right]$$

$$\in \arg\max_{s \in \mathcal{S}^{(k)}_{(a-m-1)}} H(s),$$

where $c_{m-1} = 1 - \left(\frac{a-m-1}{k} - \left\lfloor \frac{a-m-1}{k} \right\rfloor \right)$. We also consider two cases of c_m.

1. If $c_m = 1$, then $\left\lfloor \frac{a-m}{k} \right\rfloor = \frac{a-m}{k}$. Since $0 < \frac{1}{k} < 1$, we have $\left\lceil \frac{a-m-1}{k} \right\rceil = \frac{a-m}{k}$ and $\left\lfloor \frac{a-m-1}{k} \right\rfloor = \frac{a-m}{k} - 1$. Thus $1 - c_{m-1} = 1 - \frac{1}{k}$ and $c_{m-1} = \frac{1}{k}$. It follows that

$$s^{(k)*}_{(a-m)} = \left[\underbrace{\frac{a-m}{k}, \dots, \frac{a-m}{k}}_{k} \right]$$

$$s^{(k)*}_{(a-m-1)} = \left[\frac{a-m}{k} - 1, \underbrace{\frac{a-m}{k}, \dots, \frac{a-m}{k}}_{k-1} \right]$$

2. If $0 < c_m < 1$, then we have $\frac{a-m}{k} > \left\lfloor \frac{a-m}{k} \right\rfloor$. In addition, $\frac{a-m}{k} \geq \left\lfloor \frac{a-m}{k} \right\rfloor + \frac{1}{k}$, because otherwise we would have

$$k\left\lfloor \frac{a-m}{k} \right\rfloor < a - m < k\left\lfloor \frac{a-m}{k} \right\rfloor + 1$$

which means $a - m \notin \mathbb{Z}_+$. Thus

$$\frac{a-m-1}{k} = \frac{a-m}{k} - \frac{1}{k} \geq \left\lfloor \frac{a-m}{k} \right\rfloor$$

and

$$\frac{a-m-1}{k} < \frac{a-m}{k} \leq \left\lceil \frac{a-m}{k} \right\rceil,$$

which implies $\left\lfloor \frac{a-m-1}{k} \right\rfloor = \left\lfloor \frac{a-m}{k} \right\rfloor$ and $\left\lceil \frac{a-m-1}{k} \right\rceil = \left\lceil \frac{a-m}{k} \right\rceil$. It follows that

$$\begin{aligned} 1 - c_{m-1} =& \frac{a-m-1}{k} - \left\lfloor \frac{a-m-1}{k} \right\rfloor \\ =& \frac{a-m-1}{k} - \left\lfloor \frac{a-m}{k} \right\rfloor \\ =& 1 - c_m - \frac{1}{k} \end{aligned}$$

and

$$\begin{aligned} c_{m-1} =& \left\lceil \frac{a-m-1}{k} \right\rceil - \frac{a-m-1}{k} \\ =& \left\lceil \frac{a-m}{k} \right\rceil - \frac{a-m-1}{k} \\ =& c_m + \frac{1}{k}. \end{aligned}$$

Therefore we have that $s^{(k)*}_{(a-m)}$ and $s^{(k)*}_{(a-m-1)}$, can be specified by the following form

$$s^{(k)*}_{(a-m)} = \left[\underbrace{\left\lfloor \frac{a-m}{k} \right\rfloor, \ldots, \left\lfloor \frac{a-m}{k} \right\rfloor}_{kc_m}, \underbrace{\left\lceil \frac{a-m}{k} \right\rceil, \ldots, \left\lceil \frac{a-m}{k} \right\rceil}_{k(1-c_m)}\right]$$

$$s^{(k)*}_{(a-m-1)} = \left[\underbrace{\left\lfloor \frac{a-m}{k} \right\rfloor, \ldots, \left\lfloor \frac{a-m}{k} \right\rfloor}_{kc_m+1}, \underbrace{\left\lceil \frac{a-m}{k} \right\rceil, \ldots, \left\lceil \frac{a-m}{k} \right\rceil}_{k(1-c_m)-1}\right].$$

Since

$$\left\lfloor \frac{a-m}{k} \right\rfloor < \left\lceil \frac{a-m}{k} \right\rceil < x^*,$$

we have

$$h\left(\left\lfloor \frac{a-m}{k} \right\rfloor\right) < h\left(\left\lceil \frac{a-m}{k} \right\rceil\right),$$

which implies that

$$H\left(s^{(k)*}_{(a-m)}\right) - H\left(s^{(k)*}_{(a-m-1)}\right) = h\left(\left\lceil \frac{a-m}{k} \right\rceil\right) - h\left(\left\lfloor \frac{a-m}{k} \right\rfloor\right) > 0.$$

Thus

$$H\left(s^{(k)*}_{(a)}\right) > \cdots > H\left(s^{(k)*}_{(1)}\right).$$

Therefore we can conclude that

$$\begin{aligned} s^{(k)*}_{(a)} &= \left[\underbrace{\left\lfloor \frac{a}{k} \right\rfloor, \ldots, \left\lfloor \frac{a}{k} \right\rfloor}_{kc}, \underbrace{\left\lceil \frac{a}{k} \right\rceil, \ldots, \left\lceil \frac{a}{k} \right\rceil}_{k(1-c)}\right] \\ &\in \arg\max_{s \in \mathcal{S}^{(k)}_{(a)}} H(s) \\ &\in \arg\max_{s \in \mathcal{S}(a,k)} H(s), \end{aligned}$$

when $k > \frac{a}{x^*}$, which finishes the proof. □

Lemma 2.6.4. *The elements in the incremental weight gain matrix generated in PDMax satisfies*

$$\Delta W_{l,1} \geq \cdots \geq \Delta W_{l,T} \geq 0.$$

Proof. When $x_l^* = 0$, $\Delta W_{l,1} = \cdots = \Delta W_{l,T} = 0$. Now assume $x_l^* \geq 1$. Define $h(x) = \delta_l x - \gamma_l(2^x - 1)$ and $H(s) = \sum_{t=1}^{T} h(s_t)$. We consider the following four cases.

1. If $k + 1 \leq \frac{a}{x^*}$ and $k \leq \frac{a}{x^*}$, we have

   $$W_{l,k} - W_{l,k-1} = h(x^*).$$

   Thus

   $$\Delta W_{l,1} = \Delta W_{l,2} = \cdots = \Delta W_{l,\left\lfloor \frac{a}{x^*} \right\rfloor}.$$

2. If $k + 1 > \frac{a}{x^*}$ and $k \leq \frac{a}{x^*}$, we have

   $$\Delta W_{l,k} = W_{l,k} - W_{l,k-1} = h(x^*)$$

   and

   $$\Delta W_{l,k+1} = W_{l,k+1} - W_{l,k} \leq (k+1)h(x^*) - kh(x^*) = h(x^*).$$

   Thus in this case, $\Delta W_{l,k+1} \leq \Delta W_{l,k}$.

3. If $k + 1 > \frac{a}{x^*}$, $k > \frac{a}{x^*}$ and $k - 1 \leq \frac{a}{x^*}$, we have

   $$\Delta W_{l,k+1} - \Delta W_{l,k} = W_{l,k+1} + W_{l,k-1} - 2W_{l,k}.$$

   Since $2W_{l,k} = H\left(s(2a, 2k, \delta, \gamma)^*\right)$ and $\exists s \in \mathcal{S}_{(2a-\epsilon)}^{(2k)}$ for some $\epsilon \geq 0$ such that $W_{l,k+1} + W_{l,k-1} = H(s)$, we have

   $$H\left(s(2a, 2k, \delta, \gamma)^*\right) \geq H\left(s(2a - \epsilon, 2k, \delta, \gamma)^*\right) \geq H(s),$$

   which is equivalent to $\Delta W_{l,k+1} \leq \Delta W_{l,k}$.

4. For the last case, if we have the conditions $k+1 > \frac{a}{x^*}$, $k > \frac{a}{x^*}$ and $k-1 > \frac{a}{x^*}$, we have

$$\Delta W_{l,k+1} - \Delta W_{l,k} = W_{l,k+1} + W_{l,k-1} - 2W_{l,k}.$$

Since

$$2W_{l,k} = H\left(s(2a, 2k, \delta, \theta)^*\right)$$

and $\exists s \in \mathcal{S}_{(2a)}^{(2k)}$, such that $W_{l,k+1} + W_{l,k-1} = H(s)$, we have

$$H\left(s(2a, 2k, \delta, \gamma)^*\right) \geq H(s),$$

which is equivalent to $\Delta W_{l,k+1} \leq \Delta W_{l,k}$.

In a summary, we have $\Delta W_{l,1} \geq \cdots \geq \Delta W_{l,T}$. □

Define $\mathcal{V} = \{V : \max\{V_t(1) = l\} \leq d_l, \forall l \in \mathcal{L}\}$ to be the set of all feasible lists that satisfies the deadline constraints, where $V_t(1)$ is the first element of V_t. By Brassard and Bratley (1996), we have that

$$V^* \in \underset{V \in \mathcal{V}}{\arg\max} \sum_{t \in \mathcal{T}} \Delta W_{V_t(1), V_t(2)}, \tag{2.15}$$

where V^* is defined in step 2 of PDMax.

By Lemma 2.6.4, if $(l,k) \in V^*$ and $(l,k-1) \notin V^*$ for some $k \geq 2$, we can replace (l,k) by $(l,k-1)$ in V^* without decreasing the value of $\sum_{t \in \mathcal{T}} \Delta W_{V_t^*(1), V_t^*(2)}$. We repeat this for every such $(l,k) \in V^*$ until for every $(l,k) \in V^*$ where $k \geq 2$, we have $(l,k-1) \in V^*$. By doing so, we get a new list $\bar{V}^*$ such that

$$\sum_{t \in \mathcal{T}} \Delta W_{V_t^*(1), V_t^*(2)} \leq \sum_{t \in \mathcal{T}} \Delta W_{\bar{V}_t^*(1), \bar{V}_t^*(2)}$$

and if $(l,k) \in \bar{V}^*$ then $(l,k-1) \in \bar{V}^*, \ldots, (l,1) \in \bar{V}^*$.

It follows that

$$\sum_{t \in \mathcal{T}} \Delta W_{V_t^*(1), V_t^*(2)} \leq \sum_{t \in \mathcal{T}} \Delta W_{\bar{V}_t^*(1), \bar{V}_t^*(2)} = \sum_{l \in \mathcal{L}} \sum_{t=1}^{k_l} \Delta W_{l,t} = \sum_{l \in \mathcal{L}} W_{l,k_l},$$

where $k_l = \sum_{t \in \mathcal{T}} \mathbb{1}_{\{V_t^*(1)=l\}}(t)$ denotes number of time slots each link l is allocated in schedule s^*.

For s^* generated by PDMax, we have that

$$\begin{aligned} \tilde{H}(s^*) :=& \sum_{l \in \mathcal{L}} \sum_{t \in \mathcal{T}} h_l(s^*_{lt}) = \sum_{l \in \mathcal{L}} \sum_{t=1}^{k_l} h_l(s(k_l)^*_{lt}) \\ =& \sum_{l \in \mathcal{L}} W_{l,k_l} = \sum_{t \in \mathcal{T}} \Delta W_{V_t^*(1), V_t^*(2)}, \end{aligned}$$

where

$$h_l(x) = \delta_l x - \gamma_l (2^x - 1).$$

Given an arbitrary schedule $s^{'} \in \mathcal{S}(a, d)$, we let

$$k_l^{'} = \sum_{t \in \mathcal{T}} \mathbb{1}_{\{s^{'}_{l,t} > 0\}}(t)$$

denote number of time slots allocated to link l in schedule $s^{'}$. Then we have

$$\sum_{t \in \mathcal{T}} h_l(s^{'}_{lt}) \leq W_{l,k_l^{'}}, \ \forall l \in \mathcal{L}.$$

Following (2.15), we have

$$\begin{aligned} \tilde{H}(s^{'}) = \sum_{l \in \mathcal{L}} \sum_{t \in \mathcal{T}} h_l(s^{'}_{lt}) \leq& \sum_{l \in \mathcal{L}} W_{l,k_l^{'}} = \sum_{l \in \mathcal{L}} \sum_{t=1}^{k_l^{'}} \Delta W_{l,t} = \sum_{t \in \mathcal{T}} \Delta W_{V_t^{'}(1), V_t^{'}(2)} \\ \leq& \sum_{t \in \mathcal{T}} \Delta W_{V_t^*(1), V_t^*(2)} = \tilde{H}(s^*), \end{aligned}$$

where $V^{'}$ is the list consists of $(l, 1), \ldots, (l, k_l^{'})$ for each l in which

$$\mathbb{1}_{\{s^{'}_{lt} > 0\}}(t) = \mathbb{1}_{\{V_t^{'}(1)=l\}}(t), \forall t \in \mathcal{T}.$$

The first part of Theorem 2.3.1 is therefore proved.

Chapter 3

CAPACITY ALLOCATION TO IMPROVE ITINERARY COMPLETION IN HEALTHCARE NETWORK

3.1 Itinerary Completion Optimization Model and Solution Algorithm

Consider a discrete time queueing network of U stations, with $\mathcal{U} = \{0, 1, \ldots, U-1\}$ denoting the set of stations. Each station $u \in \mathcal{U}$ corresponds to a service in the care network that patients may need to use, such as the diagnostic clinic or general surgery. Each time unit represents a single day. The decision is the number of appointment slots in the "root" service to allocate to each type of target patient (breast cancer patients in our case study) on each day. Patients are then scheduled into these root appointment slots after the template is designed. After the initial root appointment of a patient's itinerary, the patient requests subsequent appointments randomly at different services. This sequence of services visits represents the patient's care path. Figure 3.1 illustrates a simplified example of a patient's path in breast cancer treatment planning.

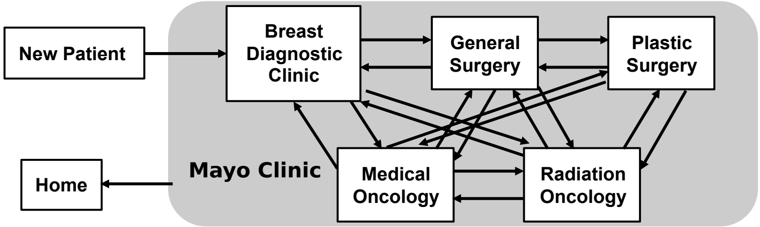


Figure 3.1: Simplified Example of the Flow Model for Breast Cancer Patients.

We model K types of target patients, with $\mathfrak{K} = \{0, 1, \ldots, K-1\}$ denoting the set of patient types. A patient type could be a combination of patient diagnosis and home location. For example, for the same diagnosis such as breast cancer, national or international patients who travel to Mayo Clinic often require different resources than local or regional patients. Thus, each type $k \in \mathfrak{K}$ patient has its own care pathway distributions. Each patient is assumed to have at most one appointment at each station on a given day, where the capacity of each station is represented by the total number of appointment slots available on a given day. The objective is to maximize the proportion of patients that complete their itinerary by their patient-type specific target deadlines.

In Section 3.1.1, we specify the queueing network for modeling the patient flow. In Section 3.1.2, we characterize the itinerary completion time (ICT) and how it depends on the network blocking profile that is generated by the initial appointment allocation decision.

3.1.1 A Queueing Network for Patient Flow

In this section, we specify the patient flow model describing the patient's movement in the healthcare network that serves as the building block for the stochastic optimization to maximize itinerary completion rates. We first specify the model inputs: initial root appointments, patient care paths, and capacities. Then we characterize the underlying stochastic process that captures the system status as a function of these inputs. While the root appointments are the decision variables, we consider them as given in this section and leave the decision model formulation to the next section.

Patient arrivals: root appointments. We denote the number of root appointment slots to reserve for type k patients on day d as

$$\Theta_{k,d}, \quad \forall k \in \mathfrak{K},\ d \in \{1, \ldots, D\},$$

where D is the length of the appointment cycle. We use $\Theta = \{\Theta_{k,d}\}$ to denote the vector of all the decision variables. For the ease of exposition, we set $D = 5$ in the rest of this paper unless otherwise specified, since Mayo Clinic, and many healthcare service providers, adopt a static template for each day of the week (Mon-Fri) that repeats every week. We refer to $d \in \{1, \ldots, 5\}$ as weekday d, reserving t for days in the absolute sense. We define $d(t)$ as the weekday associated with day t. Due to long waiting lists, we assume that these root appointment slots are always filled, though our modeling framework can accommodate probabilistic arrivals.

Itinerary. An itinerary involves a series of treatment and diagnosis stages, where a patient must complete a set of appointments in one stage before moving to the next. The itinerary for patient type k is specified by the set of tuples

$$\mathcal{C}_k = \{(u, s, p^k_{u,s}) : u \in \mathcal{U},\ s = 0, 1, 2, \ldots, S,\ p^k_{u,s} \in (0, 1]\},$$

where $(u, s, p^k_{u,s})$ indicates that an appointment at station u is required in the s^{th} stage of treatment with probability $p^k_{u,s}$ for type k patients. $s = 0$ denotes the start of a patient's itinerary, i.e., the root appointment, and S is the maximum number of stages. We account for two important features of the care path:

1. Parallel service: appointments at multiple stations could be required in a single stage, i.e., same value of s for multiple u's;

2. Stochastic itinerary: visits over the care path are probabilistic, i.e., a random subset of resources is required for a given realization of a care path.

For exposition, we assume all itineraries start with a root appointment in station 0 (the root service) for any patient type, because we are modeling a particular illness, which in many cases starts with a consult in a specific specialty; e.g. breast diagnostic clinic (BDC) for breast cancer patients. However, the analytical framework and solution we develop can be easily adapted to different and potentially random starting locations of the root appointments.

Table 3.1 illustrates a possible itinerary from our partner's data. The root appointment of each itinerary starts in station 0, the Breast Diagnostic Clinic (BDC); each entry in the table denotes the probability that an appointment that is needed from the service (station) in each stage. Patients following this itinerary start in the BDC. In addition to this root appointment, they need a follow-up consult at BDC in stage 1 with 100% chance. The patients also need consults in medical oncology with 2% and 9% chances in stage 1 and stage 2, respectively. Other entries in the table can be interpreted similarly.

	BDC	Med Onco	Rad Onco	Gen Surg	Plastic Surg
u_i	0	1	2	3	4
	1 (root)				
stage 1	1	0.02	0.01	0.02	0.00
stage 2	0.08	0.09	0.04	1	0.28
stage 3	0.05	0.01	0.01	0.01	0.03
stage 4	0.02	0.02	0.01	0.05	0.04

Table 3.1: Sample of a Care Path for National Breast Cancer Patients.

Exogenous patients. Each service in the network serves both the target patients and other patients, which we call exogenous patients. For example, breast cancer patients typically require a general surgery consult as part of their treatment planning, though many other types of patients also have general surgery consults. We do not model the detailed itineraries of exogenous patients. Instead, we assume that the number of requests for an appointment from exogenous patients at station u on day d is given by a random variable denoted as $\Lambda_u^e(d)$.

Capacity. Let $C_{u,d}$ denote capacity, i.e. the maximum number of appointment slots available in service u on day d. Let $C = \{C_{u,d}\}$ represent the capacity vector. If a patient needs an appointment from a service that is full, blocking occurs.

Blocking. The capacity of each station is used to serve exogenous patients as well as target patients that are arriving for root appointments or returning for subsequent visits later in their care path. Target patients arriving for their root appointment in station 0 are guaranteed an appointment slot because these appointments are booked in advance of the patient arriving to the clinic. Target patients arriving to subsequent appointments in their itinerary join a random ordering of all patients requesting appointment at a given service. We model the flow this way because we are modeling a capacity allocation decision and not a scheduling decision for individual patients. Hence, we do not model the exact timing of each appointment within a day and instead use this random ordering to capture the scheduling complexities at a high level.

Let $N_u(t)$ denote the total number of patients requesting an appointment (excluding root appointments) in station u on day t, from either the target or the exogenous patient groups. Under random ordering, each patient requesting an appointment from u is blocked with probability:

$$B_u(t) = \frac{\left(N_u(t) - C_{u,d(t)}\right)^+}{N_u(t)}, \tag{3.1}$$

where $d(t)$ is the weekday for day t, and $x^+ = \max(x, 0)$ for any real number x. We abuse the notation slightly for ease of exposition, since the capacity for station 0 is in fact $\tilde{C}_{0,d(t)} = C_{0,d(t)} - \sum_{k \in \mathfrak{K}} \Theta_{k,d(t)}$ after excluding the reserved capacity for root appointments; the same convention is used in the rest of the paper unless specified otherwise. From (3.1), the blocking probability, $B_u(t)$, is itself a stochastic process since it depends on $N_u(t)$.

If a target patient is blocked, they try again the next day to obtain an appointment at the same service. If the patient is not blocked, they move to the next stage of their itinerary on the following day. For analytical tractability, we assume that the blocked exogenous patients will exit the system. However, in one of our case studies we relax this assumption and allow exogenous patients to retry.

Patient flow model. We specify the stochastic process that models the dynamics of the queueing network with inputs Θ, $\{\mathcal{C}_k : k \in \mathfrak{K}\}$, and $C_{u,d}$. For ease of exposition, we present the patient flow model where at most one (but potentially different) resource is required in each stage; the full model with multiple parallel appointments is specified in Appendix B.2.

For a given station u, let $M^{B,k}_{u,s}(t)$ and $M^{NB,k}_{u,s}(t)$ denote the number of blocked and non-blocked type k patients in station u that are in stage s at the end of day t, with $M^B_u(t) = \sum_{k,s} M^{B,k}_{u,s}(t)$ being the total number of blocked patients at the end of day t. Let $N^k_{u,s}(t+1)$ denote the total number of type k patients in stage s requesting an appointment in station u on day $t+1$, which can be calculated as

$$N^k_{u,s}(t+1) = M^{B,k}_{u,s}(t) + \tilde{M}^k_{u,s}(t) + \Lambda_u(t+1). \tag{3.2}$$

The $M^{B,k}_{u,s}(t)$ blocked patients will retry to obtain an appointment in station u on day $t+1$. The random variable (r.v.) $\Lambda_u(t+1) \sim Bin\left(\Theta_{k,d(t)},\ p^k_{u,1}\right)$ represents the number of new patients who need to visit station u after their root appointments in stage 0. The r.v. $\tilde{M}^k_{u,s}(t)$ counts the patients who have finished their appointments in stage $s-1$ at the end of day t, requesting an appointment at station u in stage s on day $t+1$. It follows that $\tilde{M}^k_{u,s}(t) = \sum_{\tilde{u}} \mathbf{e}_u \cdot Mult(M^{NB,k}_{\tilde{u},s-1}(t),\ [p^k_{u_0,s}, p^k_{u_1,s}, \ldots, p^k_{u_{U-1},s}])$, where each term in the summation denotes the amount of patients, out of those

$M^{NB,k}_{\tilde{u},s-1}(t)$, that will come to station u ($\mathbf{e}_u$ denotes the variable for station u in the multinomial outcomes). The total number of patients requesting appointments at station u on day $t+1$ is thus given by

$$N_u(t+1) = \sum_k \sum_s N^k_{u,s}(t+1) + \Lambda^e_u(d(t+1)),$$

and the blocking probability $B_u(t+1)$ can then be calculated using (3.1). Based on $B_u(t+1)$, we can calculate the transitions from $N_u(t+1)$ to $\{M^{B,k}_{u,s}(t+1), \forall k, s\}$ based on hyper-geometric and multinomial distributions; see details in Appendix B. The patient-count process $\left\{N^k_{u,s}(t), t = 0, 1, \ldots, \forall u, s, k\right\}$ then forms a discrete-time Markov chain (DTMC).

Lemma 3.1.1. *Let q_u be the smallest probability, among all stages, that a patient will not return to station u after completing an appointment in u at the current stage. Let $p_u = \max_{s,k} p^k_{u,s}$ be the largest probability, among all stages and patient types, that a patient will visit station u. Under the sufficient condition*

$$q_u C_{u,d} > \Theta_{k,d} p^k_{u,1} + \sum_{v \neq u} C_{v,d-1} p_u, \quad \forall u, d, \tag{3.3}$$

1. *the DTMC $\{N^k_{u,s}(t)\}$ is positive recurrent with a unique stationary distribution;*
2. *the steady-state distribution of $\{N^k_{u,s}(t)\}$ is periodic.*

Proof. It is sufficient to prove for each station u, that the total patient-count N_u will not explode. Since the exogenous patients do not retry, they do not affect the stability of the system and we exclude them from the rest of the analysis. In other words, we consider that $N_u(t)$ only includes requests from the target patients on each day t. We use the Lyapunov function $V(x) = x$. Denote the conditional expectation $\mathbb{E}\Big[\ \cdot \ \mid N_u(t) = n\Big]$ as $\mathbb{E}_n$. Condition on n, note that all blocked target patients

$M_u^B(t) = \sum_{k,s} M_{u,s}^{B;k}(t) = (n - C_{u,d(t)})^+$. Then, the Lyapunov drift, conditional on $N_u(t) = n$, equals

$$\begin{aligned}
& \mathbb{E}\left[V(N_u(t+1)) - V(N_u(t)) \,\middle|\, N_u(t) = n\right] \\
\leq\ & \mathbb{E}_n\left[M_u^B(t) + (1-q_u)\big(n - M_u^B(t)\big) + \sum_{v \neq u} C_{v,d(t)} p_u + \Theta_{k,d(t+1)} p_{u,1}^k\right] - n \\
=\ & \mathbb{E}_n\left[q_u M_u^B(t) + (1-q_u)n + \Theta_{k,d(t+1)} p_{u,1}^k + \sum_{v \neq u} C_{v,d(t)} p_u\right] - n \\
=\ & (1-q_u)n + q_u \cdot (n - C_{u,d(t)})^+ - n + \Theta_{k,d(t+1)} p_{u,1}^k + \sum_{v \neq u} C_{v,d(t)} p_u.
\end{aligned}$$

For the inequality in the second row, we use two facts (i) for all the $n - \sum_{k,s} m_{u,k,s}$ patients who completed appointments at station u on day t, at most $(1-q_u)$ of them will request an appointment from station u in the next stage; (ii) there are at most $C_{v,d(t)}$ patients could have completed appointments at another station $v \neq u$ on day t, and at most p_u of them in the previous stage and need to visit u for the next stage (p_u is an upper bound since if the patient needs to visit multiple stations in the last stage, she may not be able to move to the next stage due to blocking).

When $n > C_{u,d(t+1)}$ is large, this drift equals

$$\begin{aligned}
& (1-q_u)n + q_u(n - C_{u,d(t+1)}) - n + \Theta_{k,d(t)} p_{u,1}^k + \sum_{v \neq u} C_{v,d(t)} p_u \\
=\ & -q_u C_{u,d(t+1)} + \Theta_{k,d(t)} p_{u,1}^k + \sum_{v \neq u} C_{v,d(t)} p_u,
\end{aligned}$$

which is negative from Condition (3.3). As a result, by the Foster-Lyapunov theorem, the DTMC is positive recurrent and has a unique stationary distribution.

The periodicity of the Markov chain comes from the fact that the arrivals from the target and exogenous patients, as well as the capacities are periodic with the same period T (while the care pathway is time-stationary). Thus, the transition matrix for the patient-count DTMC is also periodic with the same period T. Hence, it is straightforward to prove that, when the system is in the steady state, $\big(N(t), \ldots, N(t+D-1)\big)$ has the same distribution as $\big(N(t+D), \ldots, N(t+2D-1))\big)$. □

From this DTMC, we can also characterize the dynamics for the blocking probability process $\{B_u(t), t = 0, 1, \ldots, \forall u\}$, based on which we will calculate the itinerary completion time as described in the following section.

3.1.2 Itinerary Completion Optimization Model

The goal of our healthcare partner, and the impetus for this project, is to redesign the root appointment allocation to improve the *itinerary completion rate*, where the time needed to finish the entire itinerary does not exceed a type-dependent deadline. In this section, we first characterize the itinerary completion time (ICT) distribution and then formulate the optimization model.

Itinerary completion. Let $L_{i,k,t}$ denote the ICT for patient i of type k, who starts her itinerary at time t. This ICT is equivalent to the sojourn time in the queueing network of our patient flow model. The distribution of $L_{i,k,t}$ depends on the blocking probabilities $\{B_u(s),\ s \in [t, t+L_{i,k,t}]\}$ along the patient's itinerary, which is correlated among time periods, all patients in the system, and the decision variables, Θ.

Our main objective is to minimize the total penalty costs associated with patients who fail to complete their itineraries by a preset deadline, subject to capacity and throughput constraints. Let $\tau_{k,d}$ denote the itinerary completion deadline for type k patients starting on weekday d. The long-run average cost associated with a given Θ is given by

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left\{\sum_{t=1}^{T}\sum_{k} w_k \sum_{i=1}^{\Theta_{k,d(t)}} \mathbb{1}\left(L_{i,k,t} \geq \tau_{k,d(t)}\right)\right\}. \tag{3.4}$$

Here, w_k represents the relative weight for different types of patients. For example, our healthcare partner suggests a higher w_k for national and international patients, since they have to travel long distances to receive care. By Lemma 3.1.1, this long-run average cost can be reformulated as

$$\sum_{k\in\mathfrak{K}}\sum_{d=1}^{D} w_k \cdot \Theta_{k,d} \cdot \mathbb{P}_\infty(L_{k,d} \geq \tau_{k,d} \mid \Theta), \tag{3.5}$$

where $\mathbb{P}_\infty$ is the steady-state distribution of the itinerary completion time for type k patients who start their itineraries on weekday d, and $L_{k,d}$ is the steady-state version for $L_{i,k,t}$. Note, $\mathbb{P}_\infty$ depends on the DTMC $\{N^k_{u,s}(t)\}$, which further depends on the template Θ, the capacities C, and itineraries $\mathcal{C}$. Let $\mathrm{B} = \{B_{u,d}\}$ denote the corresponding set of random variables characterizing the steady-state blocking probabilities in station u on day d, where $B_{u,d} = \lim_{t\to\infty} B_u(t)$ is the limit taken on the lattice where $d(t) = d$. The following theorem characterizes the distribution of $L_{k,d}$.

Proposition 3.1.1. *The steady-state itinerary completion time for a type k patient starting itinerary on day d, denoted by $L_{k,d}$, follows a doubly-stochastic distribution given by*

$$\mathbb{P}_\infty(L_{k,d} \leq x|\Theta) = \int_{b\in[0,1]^{u\cdot d}} \big(1 - \mathbf{e}_d(\mathbf{T}_\mathrm{b})^x \cdot \mathbf{1}\big) dF_B(b|\Theta), \tag{3.6}$$

where $\mathbf{T}_b$ denotes the generator matrix for a phase-type distribution, $b = \{b_{u,d}\}$ is the vector of (realized) blocking probabilities for each u and d from the joint distribution with CDF $F_\mathrm{B}(\cdot|\Theta)$, $\mathbf{e}_d$ is the unit vector with 1 in the d^{th} column and zero elsewhere, and $\mathbf{1}$ is the unit vector of ones.

Proof. First, $\mathrm{B} = \{B_{u,d}\}$ is well defined by Lemma 3.1.1. Next, under the random ordering assumption, for a given set of realized blocking probabilities $b = \{b_{u,d}\}$, in each stage of her itinerary, a patient is either able to obtain an appointment from station u on weekday d with probability $b_{u,d}$ and moves to the next stage, or is blocked and remains in station u to retry on day $d + 1$, with probability $1 - b_{u,d}$. Thus, the sojourn time through the network is the same as the absorption time of a DTMC governed by the generator matrix $\mathbf{T}_\mathrm{b}$ that depends on b, and this sojourn time follows a phase-type distribution. The structure of $\mathbf{T}_\mathrm{b}$ is fully specified in Sections 3.3.1 and Appendix B.3. The CDF of the phase-type distribution for a patient starting on day d is given by

$$H_{k,d}(x) = 1 - \mathbf{e}_d(\mathbf{T}_b)^x \cdot \mathbf{1}. \tag{3.7}$$

The distribution of $L_{k,d}$ in (3.6) is obtained by unconditioning on the blocking probabilities. □

Stochastic optimization model. Let $W_{u,d}$ denote the workload, defined as the (random) number of appointments requested at service u on day d combining both target and exogenous patients. Let θ_k be the weekly throughput requirement for type k patients, i.e. the volume of patients that should be seen each week in accordance with management goals. We assume that $\{\theta_k\}$'s are chosen such that the system is stable. The optimization is given by:

$$\min_{\Theta} \sum_{k \in \mathfrak{K}} \sum_{d=1}^{D} w_k \cdot \Theta_{k,d} \cdot \mathbb{P}_\infty(L_{k,d} \geq \tau_{k,d} \mid \Theta) \tag{3.8}$$

$$s.t. \quad \sum_{d=1}^{D} \Theta_{k,d} \geq \theta_k, \quad \forall k \in \mathfrak{K}, \tag{3.9}$$

$$\sum_{k \in \mathfrak{K}} \Theta_{k,d} \leq C_{0,d}, \quad \forall d = 1, \ldots, D, \tag{3.10}$$

$$\mathbb{P}_\infty(W_{u,d} \geq C_{u,d} \mid \Theta) \leq \gamma_{u,d}, \quad \forall u \in \mathcal{U}, d = 1, \ldots, D, \tag{3.11}$$

$$\Theta_{k,d} \in \mathbb{R}^+. \tag{3.12}$$

Constraints (3.9) and (3.10) are throughput and capacity constraints. Constraints (3.11) are service level-type constraint that ensures that the chance that the workload at each station exceeds capacity is smaller than $\gamma_{u,d}$ to avoid excessive blocking of exogenous patients. Healthcare management can choose their desired target service levels by setting $1 - \gamma_{u,d}$ properly.

It is worth to emphasize that, although we focus on the service metric given in (3.8) for this paper, the modeling framework and the analytical methods we developed can be adapted to other metrics that depend on the distribution of the sojourn time, $L_{k,d}$. In addition, we can further incorporate additional constraints on the ICT distributions for each type of patient, as we will demonstrate in the case study in Section 3.4. By incorporating multiple ICT constraints, e.g., for regional patients 20% must complete within two days and 70% within four days, we can actually control the distribution of ICT, not just the mean. This flexibility gives our template optimization framework a precision that would not be possible if optimizing to a mean or variance-based objective, which in turn would not be possible without the proceeding detailed modeling of ICT distributions.

3.2 Challenges and Scalable Iterative Algorithm

In this section, we discuss the challenges associated with solving optimization problem (3.8) -(3.12) and then present a solution approach to overcome these challenges.

3.2.1 Analytical and Computational Challenges.

The correlations between the template, Θ, and the ICT, $L_{k,d}$ make the stochastic optimization problem both mathematically complex and computationally intractable. Several major barriers include: (i) $L_{k,d}$ has a doubly stochastic distribution depending on the blocking probabilities, which have a complex dependence on Θ. The joint distribution, $F_{\mathrm{B}}(\cdot|\Theta)$ requires solving the steady-state distribution of high-dimensional DTMC, which is of size $\bar{n}^{K \cdot U \cdot S}$, where $\bar{n}$ is an upper bound on $N_{u,s}^{k}(t)$. For example, the state space size is 15^{200} in our case study if we cap the patient count at $\bar{n} = 15$ (a conservative estimate). (ii) Numerically evaluating the integration (3.6) requires evaluating the phase-type distribution at sufficiently many realization $b = \{b_{u,d}\}$ from $F_{\mathrm{B}}(\cdot|\Theta)$, which is difficult when there is a large number of combinations of u, d. (iii) Since there is no analytical form to evaluate (3.6) for a given Θ, the only method for directly solving the optimization is to perform an exhaustive search. This is computationally intractable *even if* we could evaluate $L_{k,d}$ efficiently, say, using simulation, since the decision space is of size $\prod_{k,d}(\theta_k)^d$, which, in our case study would be approximately 25^{20}.

At a high level, the challenges above stem from the (i) the strong correlation between all the ICT random variables, $\{L_{k,d}\ \forall k, d\}$, and (ii) the dependence of $L_{k,d}$ on the entire matrix of decision variables $\Theta = \{\Theta_{k,d}\ \forall k, d\}$. This is because follow-up appointments generated from all root appointments occupy the same set of resources in the queueing network over intersecting time periods. Next, we provide a high level overview of our novel approach to overcome these challenges and develop a scalable algorithm to optimize itinerary completion for large queueing networks.

3.2.2 Scalable Iterative Algorithm.

To solve the optimization problem (3.8) – (3.12), we develop an iterative algorithm that decouples the correlation between decisions, Θ, and the deadline-violation probability $\mathbb{P}_\infty(L_{k,d} \geq \tau_{k,d})$. This algorithm iterates between two steps: (1) performance evaluation of the itinerary completion distribution that addresses the first two challenges, and (2) policy improvement on the decision variables that addresses the third challenge.

1. **Performance evaluation.** In each iteration $i + 1$, we evaluate $\mathbb{P}_\infty(L_{k,d} \leq x)$ using the template generated in the previous iteration, $\Theta^{(i)}$. To evaluate

$\mathbb{P}_\infty(L_{k,d} \leq x)$, we develop a mean-field model in Section 3.3.2 that allows us to replace the integration in (3.6) over $F_{\mathrm{B}}(\cdot|\Theta^{(i)})$, with the *point mass* for the blocking probabilities in iteration i, $\beta^{(i)} = \{\beta^{(i)}_{u,d}\}$, where $\beta^{(i)}$ is the equilibrium solution from the mean-field model under $\Theta^{(i)}$. That is,

$$\mathbb{P}_\infty(L_{k,d} \leq x) = \int_{b\in[0,1]^{u\cdot d}} \big(1 - \mathbf{e}_d(\mathbf{T}_{\mathrm{b}})^x \cdot \mathbf{1}\big) dF_B(b) \approx \big(1 - \mathbf{e}_d(\mathbf{T}_{\beta^{(\mathrm{i})}})^x \cdot \mathbf{1}\big). \quad (3.13)$$

Since the mean-field model is a deterministic system, its equilibrium is easy to solve versus solving the high-dimensional DTMC. Additionally, we only need to evaluate the matrix power calculation once on this point mass versus numerically evaluating an integral. To justify the replacement in (3.13), we rigorously show in Chapter 4 the asymptotic convergence of the steady-state blocking distribution, F_{B}, to the point mass β from the mean-field model, via the Stein's method framework.

2. **Policy improvement.** We use a policy improvement approach to decouple the dependence between calculating $\mathbb{P}_\infty(L_{k,d} \leq x)$ and optimizing the decision variable Θ. To get the updated template, $\Theta^{(i+1)}$, we first replace $\mathbb{P}_\infty(L_{k,d} \leq x)$ with (3.13), where $\beta^{(i)}$ is calculated from the mean-field model using the previous template $\Theta^{(i)}$. This decouples $L_{k,d}$ and Θ to be optimized in the current iteration as follows.

$$\min_{\Theta} \quad \sum_{k\in\mathfrak{K}}\sum_{d=1}^{D} w_k \cdot \Theta_{k,d} \cdot \big(\mathbf{e}_d(\mathbf{T}_{\beta^{(\mathrm{i})}})^{\tau_{k,d}} \cdot \mathbf{1}\big) \tag{3.14}$$

$$s.t. \quad \sum_{d=1}^{D} \Theta_{k,d} \geq \theta_k, \quad \forall k \in \mathfrak{K}, \tag{3.15}$$

$$\sum_{k\in\mathfrak{K}} \Theta_{k,d} \leq C_{0,d}, \quad \forall d = 1,\dots,D, \tag{3.16}$$

$$\left|\beta_{u,d}(\Theta) - \beta_{u,d}^{(i)}\right| \leq \epsilon, \quad \forall u \in \mathcal{U}, d = 1,\dots,D, \tag{3.17}$$

$$\mathbb{E}_{mf}[N_{u,d}|\ \Theta](1-\gamma_{u,d}) \leq C_{u,d}, \quad \forall u \in \mathcal{U}, d = 1,\dots,D, \tag{3.18}$$

$$\Theta_{k,d} \in \mathbb{R}^{+}. \tag{3.19}$$

Here, $\beta_{u,d}(\Theta)$ in constraint (3.17) and $\mathbb{E}_{mf}[N_{u,d}|\ \Theta]$ in constraint (3.18) are calculated using the deterministic approximation in the mean-field model, with $\mathbb{E}_{mf}[N_{u,d}|\ \Theta]$ being the mean-field version for $W_{u,d}$ in constraint (3.11), and $\beta_{u,d}(\Theta) = (\mathbb{E}_{mf}[N_{u,d}|\ \Theta] - C_{u,d})^{+}/\mathbb{E}_{mf}[N_{u,d}|\ \Theta]$. Constraint (3.18) is the mean-field version for constraints (3.11). Constraint (3.17) ensures the network blocking profile is sufficiently close to the blocking profile from the previous iteration. Thus, $\mathbb{P}_{\infty}(L_{k,d} \leq x)$ evaluated using (3.13), is sufficiently close between iteration i and $i+1$ so that the itinerary completion probabilities calculated using $\Theta^{(i)}$ will be a good approximation within the feasible set of decision variables that solve for template $\Theta^{(i+1)}$. In Section 3.3.3, we provide a performance bound on this approximation by bounding the gap between the ICT distribution obtained from the template in the previous iteration and the ICT distributions in the feasible set of the current iteration.

The key here is that there are sufficiently many templates Θ in the feasible set that can generate similar blocking profiles, but they can lead to very different ICT distributions for different types of patients, since ICT also depends on the timing of the itinerary start. In addition, the expected workload, $\mathbb{E}_{mf}[N_{u,d}]$, is calculated by taking the expectation of the arrivals and departures, adjusted by the expected number of blocked patients each day using $\beta^{(i)}$, so that constraints (3.17) and (3.18) are *linear* in the decision variable Θ. Given the linear objective function and the linear constraints in (3.14) to (3.19), the optimization program here is a linear program (LP) that can be solved efficiently for large (realistic-sized) networks with a commercial optimization software such as `CPLEX`. We conclude with two remarks for the iterative algorithm.

Remark 3.2.1 (Initialization)**.** *To obtain an initial set of blocking probabilities for the policy improvement, one can use the historical template to compute the corresponding blocking probabilities. An alternative approach is to develop a workload smoothing optimization as a pre-processing stage to minimize the blocking probabilities across the system, since high blocking along the itinerary can extend the ICT. The full details of the workload smoothing are given in Appendix D.*

Remark 3.2.2 (Refinement for low-blocking settings)**.** *The mean-field model captures the blocking on the "fluid scale," which is best when blocking is significant. In settings where the workload is often below the capacity such that the fluid-scale blocking probability is 0 (but blocking still occurs due to stochastic fluctuations), we refine the approximation to the blocking probabilities using an offered load approach. At a high-level, we approximate the aggregate workload distribution of each station u on each day d for a given template Θ with a normal r.v. having mean $\mu_{u,d}(\Theta)$ and standard deviation $\sigma_{u,d}(\Theta)$, where the mean and standard deviation are calculated by assuming there is no capacity constraint (Massey and Whitt, 1993). See Appendix D.1 for a detailed calculation.*

3.3 Itinerary Completion Time Approximation: Exact Analysis and Approximations

In this section, we take the template Θ as given and focus on the ICT evaluation. We first characterize the doubly-stochastic distribution (3.6) by specifying the generator matrix T_b in Section 3.3.1. To address the computational difficulty in the exact analysis of this doubly-stochastic distribution, we then introduce the mean-field model in Section 3.3.2, which provides an equilibrium solution for β that is the input to (3.13). The mean-field approximation is justified in (3.13) by proving the asymptotic convergence from the distribution of $B = \{B_{u,d}\}$ to the point mass β and characterizing the convergence rate in Chapter 4. We also translate the error in the blocking probabilities to a performance bound in the ICT distribution (our main metric of interests) in Section 3.3.3.

We focus on the setting where a patient visits one resource in each stage in this section; in Appendix B.3, we incorporate the feature that patients may need to visit multiple stations in parallel in each stage of their itineraries, examples of which are prevalent in our data.

3.3.1 Phase-Type Representation in the Base Setting

We first review the idea of the phase-type representation of ICT in (3.6). Conditioning on a given realization of the blocking probabilities $b = \{b_{u,d}\} \sim F_{\mathrm{B}}(\cdot|\Theta)$, the time to complete an appointment station on a patient's itinerary is a Bernoulli trial with a failure represented by the patient being blocked. This follows from the random ordering assumption, which is appropriate for our capacity planning level of analysis as explained in Section 3.1.1. The time to complete all appointments on a patient's itinerary is thus driven by an underlying Markov chain with transition matrix based on (i) the probability of completing the current appointment and (ii) if successful, the transition to the next appointment. Thus, the ICT is the sum of geometric distributions with time-varying success probabilities, which forms a discrete phase-type distribution (Casale, 2010).

Single stage. For illustration, we start by considering the simplest setting with a single stage (requiring service at station u) for a given patient type k. For any realization of the blocking probabilities $b = \{b_{u,d}\}$, the patient is either able to get the appointment at service u or is blocked (with probability $1 - b_{u,d}$ or $b_{u,d}$, respectively). If the patient is blocked, she will retry to get the appointment the next day until being successful; otherwise, her itinerary finishes. For the given b, the event of getting the appointment or being blocked becomes independent Bernoulli trials among different days. Then, the ICT of type k patient starting on weekday d, $L_{k,d}$ resembles a geometric distribution except the success probability is time-varying. The transitions of these success probabilities are driven by a DTMC that can be characterized by one common *generator matrix*:

$$\left[\begin{array}{c|c} \mathbf{T}_b & \mathbf{T}_b^0 \\ \hdashline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{ccccc|c} 0 & b_{u,1} & 0 & 0 & 0 & 1 - b_{u,1} \\ 0 & 0 & b_{u,2} & 0 & 0 & 1 - b_{u,2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ b_{u,5} & 0 & 0 & 0 & 0 & 1 - b_{u,5} \\ \hdashline 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right], \tag{3.20}$$

where $\mathbf{T}_b \cdot \mathbf{1} + \mathbf{T}_b^0 = \mathbf{1}$ and $\mathbf{1}$ is the vector of 1's. Each state in $\mathbf{T}_b$ represents the blocking status on day d, while the column of $\mathbf{T}_b^0$ represents the absorbing state – itinerary completion. For example, the first row in the generator matrix indicates that the initial attempt is made on day 1 (Monday). With probability $b_{u,1}$ the patient is blocked and moves to Tuesday (column 2) to try to obtain the appointment; with probability $1 - b_{u,1}$ the patient obtains the appointment and moves to the absorbing state. Other rows can be interpreted similarly except when the initial attempt is made on day 5, the next retrial has to be made on day 1 of the next week, as the system is periodic with $D = 5$ as the period. The last row indicates that the success state is an absorbing state. The CDF of $L_{k,d}$ is given in (3.7), while its probability mass function (pmf) is

$$h_{k,d}(x) = \mathbf{e}_d(\mathbf{T}_b)^{x-1} \cdot \mathbf{T}_b^0, \tag{3.21}$$

To explain (3.21), note that, for a patient whose initial trial is on day d, the probability that her first success is on day x equals the probability of no success in the past $x - 1$ days, $(\mathbf{T}_b)^{x-1}$, and that a success occurs on day x, $\mathbf{T}_b^0$. Here, $\mathbf{e}_d$ adjusts the initial starting day to be day d. The CDF in (3.7) has a similar explanation; also see Latouche and Ramaswami (1999).

Probabilistic requirements in multiple stages. We now extend the above phase-type representation to the setting where an itinerary involves multiple stages and probabilistic resource requirements. Here, we still model each stage with only one station as a building block. For notational simplicity, we drop the type k in the description below when denoting the itinerary for a given type patient $\mathcal{C} = \{(u_0, 0, 1), (u_1, 1, p_{u_1,1}), \ldots, (u_n, n, p_{u_n,n})\}$, where $p_{u_s,s}$ denotes the probability that this particular type patient requires service from station u_s in stage s and $p_{u_0,0} = 1$ for the root appointment.

When a patient completes service at station u_s in stage s, she requests an appointment at the next station u_{s+1} if this service is required with probability $p_{u_{s+1},s+1}$. If stage $s + 1$ is skipped, the patient transitions to request an appointment at station u_{s+2} with probability $p_{u_{s+2},s+2}$, or not requiring stage $s + 2$ either, in which case the patient transitions to request an appointment at station u_{s+3} and so forth.

The generator matrix for the transitions between phases in $L_{k,d}$ follows

$$\left[\begin{array}{c|c} \mathbf{T}_b & \mathbf{T}_b^0 \\ \hline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{cccccc|c} \mathbf{T}^{\mathbf{1}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{2}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{3}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{4}}_{\mathbf{u_1}} & \dots & \mathbf{T}^{\mathbf{n}}_{\mathbf{u_1}} & \mathbf{T}^{\mathbf{0}}_{\mathbf{u_1}} \\ \mathbf{0} & \mathbf{T}^{\mathbf{1}}_{\mathbf{u_2}} & \mathbf{T}^{\mathbf{2}}_{\mathbf{u_2}} & \mathbf{T}^{\mathbf{3}}_{\mathbf{u_2}} & \dots & \mathbf{T}^{\mathbf{n-1}}_{\mathbf{u_2}} & \mathbf{T}^{\mathbf{0}}_{\mathbf{u_2}} \\ & \ddots & \ddots & & & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{T}^{\mathbf{1}}_{\mathbf{u_n}} & \mathbf{T}^{\mathbf{0}}_{\mathbf{u_n}} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & 1 \end{array}\right]. \tag{3.22}$$

Each state in the transition matrix represents the blocking status in (s, d), a combination of which stage s the patient is in and on which workday d. The matrix block $\mathbf{T}^{\mathbf{1}}_{\mathbf{u_s}}$ is defined similarly as $\mathbf{T_u}$ in (3.20) with $b_{u_s,d}$ replacing $b_{u,d}$. $\mathbf{T}^{\mathbf{1}}_{\mathbf{u_s}}$ characterizes the transitions within stage s; i.e. an appointment has not yet been obtained at station u_s. The other matrices on each row characterize the transitions out of stage s to stage $s + j - 1$ or to the absorbing state as follows:

$$\mathbf{T}^{\mathbf{j}}_{\mathbf{u_s}} = p_{u_{s+j-1},s+j-1} \prod_{m=s+1}^{s+j-2} (1 - p_{u_m,m}) \begin{bmatrix} 0 & 1 - b_{u_s,1} & 0 & 0 & 0 \\ 0 & 0 & 1 - b_{u_s,2} & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 - b_{u_s,5} & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.23}$$

for $j = 2, \dots, n - s + 1$. $\mathbf{T}^{\mathbf{j}}_{\mathbf{u_s}}$ represents the transition from stage s to stage $s + j - 1$ given it is not skipped. The transition directly to the absorbing stage, given that all remaining stages are skipped, is given by

$$\mathbf{T}^{\mathbf{0}}_{\mathbf{u_i}} = \prod_{m=s+1}^{n} p_{u_m,m} \cdot \begin{bmatrix} 1 - b_{u_s,1} & 1 - b_{u_s,2} & \dots & 1 - b_{u_s,5} \end{bmatrix}'. \tag{3.24}$$

We use the convention that an empty product is equal to 1 for (3.23) and (3.24). The CDF and pmf of $L_{k,d}$ have the same form as in (3.7) and (3.21), where we replace $\mathbf{T}_b$ and $\mathbf{T}_b^0$ with the ones from (3.22).

3.3.2 Mean-Field Model and Numerical Illustration

As highlighted in Section 3.2, evaluating $\mathbb{P}_\infty(L_{k,d} \geq T_{k,d}|\Theta)$ for a given schedule Θ relies on the steady-state blocking probability distribution $F_{\mathrm{B}}(\cdot|\Theta)$, which needs to be solved from a high-dimensional DTMC. Even if we are able to efficiently compute the distribution $F_{\mathrm{B}}(\cdot|\Theta)$, calculating $L_{k,d}$ for each realized $b \sim F_{\mathrm{B}}(\cdot|\Theta)$ still requires evaluating integration that involves matrix power, as demonstrated in Section 3.3.1. To address this challenge, we leverage the mean-field model for the DTMC and replace the steady-state distribution of $F_{\mathrm{B}}(\cdot|\Theta)$ with a point mass $\beta = \{\beta_{u,d}\}$ that is the equilibrium solution of the mean-field model. We specify the mean-field model next, and will rigorously prove the asymptotic convergence of the blocking probability distribution to the point mass β in Chapter 4. Before the technical details, we first numerically demonstrate that the distribution of $b_{u,d}$ converges to the point mass $\beta_{u,d}$ when the system size is large.

Numerical validation. As highlighted in Section 3.2, evaluating $\mathbb{P}_\infty(L_{k,d} \geq T_{k,d}|\Theta)$ for a given template Θ relies on the steady-state blocking probability distribution $F_{\mathrm{B}}(\cdot|\Theta)$, which needs to be solved from a high-dimensional DTMC. Even if we are able to efficiently compute the distribution $F_{\mathrm{B}}(\cdot|\Theta)$, calculating $L_{k,d}$ for each realized $b \sim F_{\mathrm{B}}(\cdot|\Theta)$ still requires evaluating an integral over matrix powers. To address this challenge, we leverage a mean-field model for the DTMC and replace the steady-state distribution of $F_{\mathrm{B}}(\cdot|\Theta)$ with a point mass $\beta_\infty = \{\beta_{u,d}\}$, which comes from the equilibrium solution of the mean-field model. Before presenting the technical details, we first illustrate numerically how the distribution of $F_{\mathrm{B}}(\cdot|\Theta)$ converges to the point mass β_∞ as the system size grows.

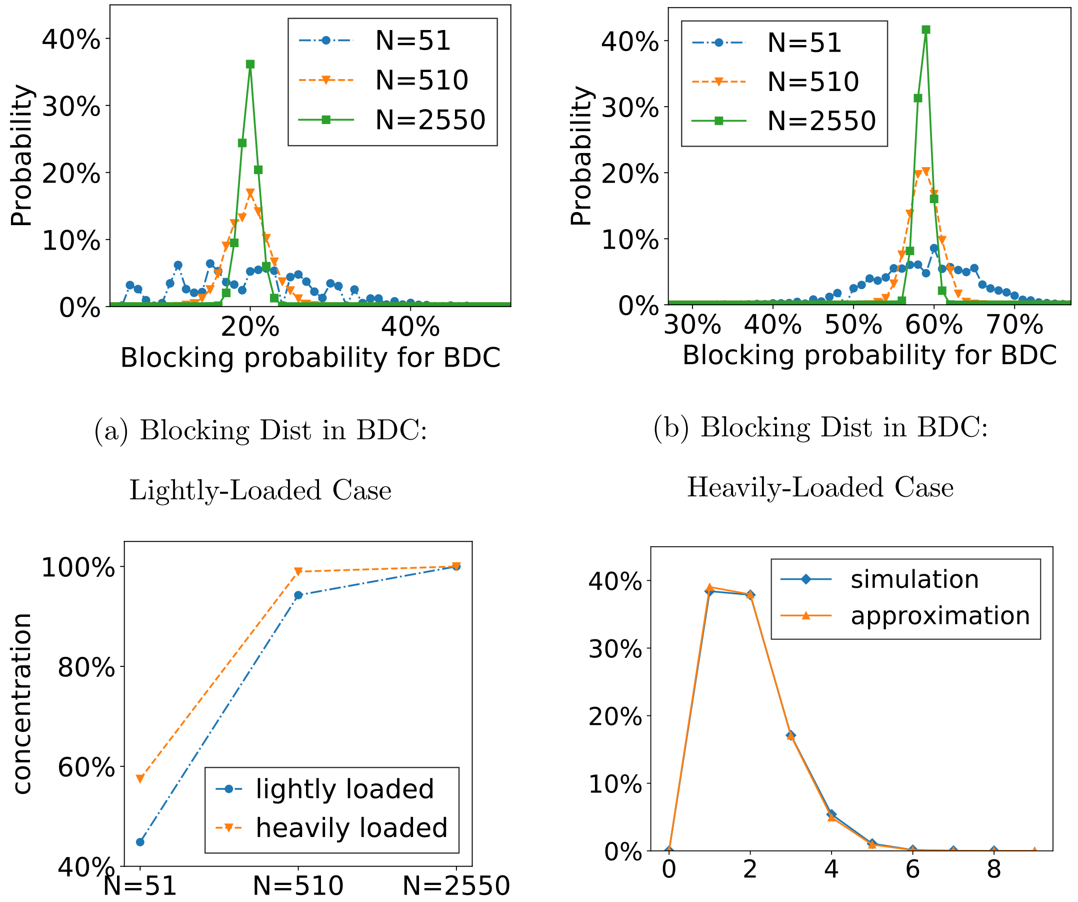
Figures 3.2(a) shows the blocking probability distribution for BDC on Monday. This distribution comes from simulating a five-station network used as the baseline in our case study in Section 3.4. The network is parameterized via our partner's data using the historical appointment template. We model four types of patients with *time-varying* arrivals and capacities, with possible parallel appointments. The average capacity for BDC is $N = 51$, and we scale this baseline system (along with capacities of the other four stations as well as arrivals) by 10 and 50 times proportionally. As the system size increases, we observe that the distributions of the blocking probabilities quickly become concentrated near a point mass – which is what we will prove in Chapter 4. The capacity for BDC is $N = 51$. We consider four types of patients and the arrivals and capacities are time-varying as estimated from the data. We scale this baseline system by 10 and 50 times proportionally.

Figure 3.2 (b) plots the blocking probability distribution for BDC on Monday under a scenario with a higher average system load.

The level of concentration, defined as the fraction of blocking probabilities that fall within the range of the point mass ± 0.05, is illustrated in Figure 3.2(c) for the original scenario as well as a scenario with a higher average system load.

Figure 3.2(d) demonstrates that, even if the blocking probabilities are moderately concentrated, as in the current system ($N = 51$), the approximation for the ICT distribution is already very close to that from the simulation. Applying the Kolmogorov-Smirnov (KS) statistic, we find that the median and maximum distances between the two distributions are less than 2 % and 7% respectively across all patient types and starting days for all the experimental settings in the case study. See Appendix E for more numerical results.

Dynamics of mean-field model. The point mass β_∞ comes from the equilibrium solution of the mean-field model, which serves as a deterministic approximation for the proportion of blocked patients in the original stochastic system, $U(t) = \{U_u(t),\ u = 0, \dots, U-1\}$. This is also known as the occupancy measures (Ying, 2018). For each



(a) Blocking Dist in BDC: Lightly-Loaded Case

(b) Blocking Dist in BDC: Heavily-Loaded Case

(c) Level of Concentration

(d) ICT Distribution for National Patients Starting on Monday under $N = 51$

Figure 3.2: Illustration of Blocking Probability Approximation

station u, $U_u = \sum_{k,s} M_{u,s}^{B,k}(t) / \sum_{k,s} N_{u,s}^k(t)$, where $N_{u,s}^k(t)$ and $M_{u,s}^{B,k}(t)$ are respectively the number of target patients requesting an appointment and number of blocked at station u on day t. Let $m_{u,s}^{B,k}(t)$ and $m_{u,s}^{NB,k}(t)$ be the deterministic counterparts of $M_{u,s}^{B,k}(t)$ and $M_{u,s}^{NB,k}(t) = N_{u,s}^k(t) - M_{u,s}^{B,k}(t)$, and $n_{u,s}^k(t+1)$ be the counterpart for $N_{u,s}^k(t+1)$. By taking the expectation of the random quantities in (3.2), we get

$$n_{u,s}^k(t+1) \quad = \quad m_{u,s}^{B,k}(t) + \sum_{\tilde{u}} p_{u,s}^k \cdot m_{\tilde{u},s-1}^{NB,k}(t) + p_{u,1}^k \cdot \Theta_{k,d(t)}.$$

Consequently, for the total number of patients in station u on day $t+1$, we have the following:

$$n_u(t+1) \;\; = \;\; \sum_k \sum_s n^k_{u,s}(t+1) + \lambda^e_u(d(t+1)) \tag{3.25}$$

where $\lambda^e_u(d(t+1))$ is the expectation of the exogenous arrivals $\Lambda^e_u(d(t+1))$. Then, with $\beta_u(t+1) = \left[n_u(t+1) - C_{u,d(t+1)}\right]^+ / n_u(t+1)$ being the blocking probability in the deterministic system, we have

$$m^{B,k}_{u,s}(t+1) = \beta_u(t+1) \cdot n^k_{u,s}(t+1), \quad m^{NB,k}_{u,s}(t+1) = n^k_{u,s}(t+1) - m^{B,k}_{u,s}(t+1). \tag{3.26}$$

Finally, let $\mu_u(\cdot)$ is the deterministic counterpart for $U_u(\cdot)$. We have

$$\mu_u(t+1) = \frac{\sum_{k,s} m^{B,k}_{u,s}(t+1)}{\sum_{k,s} n^k_{u,s}(t+1)} = \beta_u(t+1).$$

Under the same stability condition given in Lemma 3.1.1, we can show the deterministic system has a unique equilibrium solution β_∞ by verifying the Lynapunov condition.

3.3.3 Bounding the Gap in ICT Distributions

Consider two sets of blocking probabilities such that $|\beta_{u,d} - \beta'_{u,d}| \leq \epsilon$ for all u, d. In this section, we translate this ϵ gap between the blocking probabilities into the gap between the corresponding ICT distributions and establish an upper bound for the latter gap.

Lemma 3.3.1. *Let $p = \max_d \beta_d$ be the largest blocking probability for a type k patient across all days d. Let Θ and Θ' be two feasible schedules in the program* (3.14) *to* (3.19)*, with the induced blocking probabilities satisfying $|\beta_{u,d} - \beta'_{u,d}| \leq \epsilon$, $\forall u, d$. Let S be the maximum number of stages in a care pathway. Then*

$$\Big|\mathbb{P}_\infty(L_{k,d} \leq x|\Theta) - \mathbb{P}_\infty(L_{k,d} \leq x|\Theta')\Big| \leq O(x^S p^{x-1} \cdot \epsilon) + o(\epsilon).$$

Proof. For notional convenience, we omit the station u index and re-index β_d with $d = d \bmod 5 \in \{0, \dots, 4\}$. In other words, we omit the mod operator in the index of day for the blocking probabilities. Under this labeling scheme we can write

$$\mathbb{P}_\infty(L_{k,d} \leq x|\Theta) = 1 - \mathbf{e}_d \cdot (\mathbf{T}_\beta)^x \cdot \mathbf{1} = 1 - \prod_{\tilde{d}=d}^{d+x-1} \beta_{\tilde{d}}, \tag{3.27}$$

$$\Big|\mathbb{P}_\infty(L_{k,d} \leq x|\Theta) - \mathbb{P}_\infty(L_{k,d} \leq x|\Theta')\Big| \leq \sum_{d_0=d}^{d+x-1} \prod_{\tilde{d}=d}^{d+x-1} \left(\left(\beta_{\tilde{d}}\right)^{\mathbb{1}\{\tilde{d} \neq d_0\}} \cdot \epsilon \right) + o(\epsilon). \tag{3.28}$$

The second line follows from the assumption that $|\beta_d - \beta'_d| \leq \epsilon$. Replacing $\beta_{\tilde{d}}$ in (3.27) with $\beta_{\tilde{d}} \pm \epsilon$ and subtracting the two distributions yields (3.28). Since $p = \max_d \beta_d$, it is easy to show by replacing β_d with p for all d that (3.28) can be bounded by

$$\sum_{d_0=d}^{d+x-1} \prod_{\tilde{d}=d}^{d+x-1} \left(\left(\beta_{\tilde{d}}\right)^{\mathbb{1}\{\tilde{d} \neq d_0\}} \cdot \epsilon \right) + o(\epsilon) \leq x \cdot p^{x-1} \cdot \epsilon + o(\epsilon). \tag{3.29}$$

To extend to multiple stages, we leverage the property that the generator matrix $\mathbf{T}_\beta$ is upper triangular and use an induction to prove the performance bound. We first illustrate with the two-stage case and then extend the general S-stage case.

$$\mathbf{T}_{\mathcal{C}} = \begin{bmatrix} \mathbf{T}^1_{u_1} & \mathbf{T}^2_{u_1} \\ \mathbf{0} & \mathbf{T}^1_{u_2} \end{bmatrix}, \tag{3.30}$$

$$\mathbf{T}_{\mathcal{C}}(\epsilon) = \begin{bmatrix} \mathbf{T}^1_{u_1}(\epsilon) & \mathbf{T}^2_{u_1}(\epsilon) \\ \mathbf{0} & \mathbf{T}^1_{u_2}(\epsilon) \end{bmatrix}, \tag{3.31}$$

$$\mathbf{T}^1_{u}(\epsilon) = \begin{bmatrix} 0 & b_{u,1}+\epsilon & 0 & 0 & 0 \\ 0 & 0 & b_{u,2}+\epsilon & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{u,5}+\epsilon & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.32}$$

$$\mathbf{T}^2_{u}(\epsilon) = \begin{bmatrix} 0 & 1-b_{u,1}-\epsilon & 0 & 0 & 0 \\ 0 & 0 & 1-b_{u,2}-\epsilon & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1-b_{u,5}-\epsilon & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.33}$$

Let $\mathbf{T}^{2,(x-1)}_{u_1}$ be the upper right block of $(\mathbf{T}_{\mathcal{C}})^{x-1}$. It is easy to show that:

$$(\mathbf{T}_{\mathcal{C}})^x = \begin{bmatrix} \left(\mathbf{T}^1_{u_1}\right)^x & \mathbf{T}^1_{u_1}\cdot\mathbf{T}^{2,(x-1)}_{u_1} + \mathbf{T}^2_{u_1}\cdot\left(\mathbf{T}^1_{u_2}\right)^{x-1} \\ \mathbf{0} & \left(\mathbf{T}^1_{u_2}\right)^x \end{bmatrix}, \tag{3.34}$$

Now consider

$$\left|\mathbb{P}_\infty(L_{k,d} \le x|\Theta) - \mathbb{P}_\infty(L_{k,d} \le x|\Theta^{(i)})\right| \le \left|\mathbf{e}_d \cdot (\mathbf{T}_{\mathcal{C}}(\epsilon))^x \cdot \mathbf{1} - \mathbf{e}_d \cdot (\mathbf{T}_{\mathcal{C}})^x \cdot \mathbf{1}\right|. \tag{3.35}$$

Note that the terms $\mathbf{e}_d \cdot (\mathbf{T}_{\mathcal{C}}(\epsilon))^x \cdot \mathbf{1}$ that do not contain ϵ cancel with the terms of $\mathbf{e}_d \cdot (\mathbf{T}_{\mathcal{C}})^x \cdot \mathbf{1}$. For example, selecting all non-ϵ terms from $\mathbf{e_1}(\mathbf{T}^1_{u_1}(\epsilon))^3 \cdot \mathbf{1} = (\beta_{u_1,1} + \epsilon)(\beta_{u_1,2}+\epsilon)(\beta_{u_1,3}+\epsilon)$ yields $\beta_{u_1,1}\cdot\beta_{u_1,2}\cdot\beta_{u_1,3} = \mathbf{e}_1(\mathbf{T}^1_{u_1})^x \cdot \mathbf{1}$. Thus, we can cancel all the terms from the upper left block of $\mathbb{P}_\infty(L_{k,d} \le x|\Theta)$. To calculate the remaining error terms after canceling all terms without ϵ, consider terms that are linear in ϵ. For example, the ϵ-linear terms from $\mathbf{e_1}(\mathbf{T}^1_{u_1}(\epsilon))^3 \cdot \mathbf{1} = (\beta_{u_1,1}+\epsilon)(\beta_{u_1,2}+\epsilon)(\beta_{u_1,3}+\epsilon)$

are $\beta_{u_1,1} \cdot \beta_{u_1,2} \cdot \epsilon + \beta_{u_1,1} \cdot \beta_{u_1,3} \cdot \epsilon + \beta_{u_1,2} \cdot \beta_{u_1,3} \cdot \epsilon$. Note, there are $\binom{x}{x-1}$ such terms in $(\mathbf{T}_{\mathcal{C}}(\epsilon))^x \cdot \mathbf{1}$. The remainder of the terms are $O(\epsilon)$. Hence, the error contributed by the upper left block of $\mathbf{e}_d\,(\mathbf{T}_{\mathcal{C}}(\epsilon))^x \cdot \mathbf{1}$ can be written as

$$\epsilon \cdot \sum_{d_0=d}^{d+x-1} \prod_{d_1=d}^{d+x-1} \left(\left(b_{u_1,d}\right)^{\mathbb{1}\{d_1 \neq d_0\}} \right) + o(\epsilon)$$

To get a uniform bound, let $\bar{b} = \max_{i,d}\{b_{u_i,d}\}$. The bound can then be written as

$$x \cdot (\bar{b})^{x-1}\epsilon + o(\epsilon) \tag{3.36}$$

Consider the upper right block of the matrix.

$$\mathbf{T}^{2,(1)}_{u_1} = \mathbf{T}^{2}_{u_1} \tag{3.37}$$

$$\mathbf{T}^{2,(2)}_{u_1} = \mathbf{T}^{1}_{u_1} \cdot \mathbf{T}^{2,(1)}_{u_1} + \mathbf{T}^{2}_{u_1} \cdot (\mathbf{T}^{1}_{u_2}) = \mathbf{T}^{1}_{u_1} \cdot \mathbf{T}^{2}_{u_1} + \mathbf{T}^{2}_{u_1} \cdot (\mathbf{T}^{1}_{u_2}) \tag{3.38}$$

$$\mathbf{T}^{2,(3)}_{u_1} = \mathbf{T}^{1}_{u_1} \cdot \mathbf{T}^{2,(2)}_{u_1} + \mathbf{T}^{2}_{u_1} \cdot (\mathbf{T}^{1}_{u_2})^2 = (\mathbf{T}^{1}_{u_1})^2 \cdot \mathbf{T}^{2}_{u_1} + \mathbf{T}^{1}_{u_1} \cdot \mathbf{T}^{2}_{u_1} \cdot \mathbf{T}^{1}_{u_2} + \mathbf{T}^{2}_{u_1} \cdot (\mathbf{T}^{1}_{u_2})^2 \tag{3.39}$$

$$\mathbf{T}^{2,(x)}_{u_1} = \sum_{j=0}^{x-1} (\mathbf{T}^{1}_{u_1})^j \cdot \mathbf{T}^{2}_{u_1} \cdot (\mathbf{T}^{1}_{u_2})^{(x-j-1)} \tag{3.40}$$

From (3.40), each term in the sum captures the probability of spending j days in stage 1 (station u_1) and at least $n-j-1$ days in stage 2. Summing them up, we see that $\mathbf{T}^{2,(n)}_{u_1}$ is in fact just the probability that stage 1 has been completed but stage 2 has not yet been completed by time n. Multiplying $\mathbf{T}^{2,(x-1)}_{u_1}(\epsilon)$ by $\mathbf{e}_d \cdot \mathbf{T}^{1}_{u_1}$ on the left and $\mathbf{1}$, intuitively each term from (3.40) becomes of the form

$$(b_{u_1,d} + \epsilon) \left(\prod_{d_0=d+1}^{d+1+j-1} (b_{u_1,d_0} + \epsilon) \right) \cdot (1 - b_{u_1,d+1+j-1}) - \epsilon) \prod_{d_0=d+1+j}^{d+1+n-2} b_{u_2,d_0} + O(\epsilon)$$

Canceling the ϵ-constant terms and isolating the terms that are linear in ϵ, we find the error for each term is

$$\begin{aligned}
&\epsilon \cdot (1 - b_{u_1,(d+1+j-1)})) \sum_{d_1=d+1}^{d+x-1} \left(\prod_{d_0=d}^{d+1+j-1} (b_{u_1,d_0})^{\mathbb{1}\{d_0 \neq d_1\}} \right) \prod_{d_0=d+1+j}^{d+x-1} (b_{u_2,d_0})^{\mathbb{1}\{d_0 \neq d_1\}} \\
&- \epsilon \left(\prod_{d_0=d}^{d+1+j-1} (b_{u_1,d_0}) \right) \prod_{d_0=d+1+j}^{d+x-1} (b_{u_2,d_0}) + O(\epsilon)
\end{aligned} \tag{3.41}$$

Note, each term j has $x-1$ terms, each of which are a multiplication of $x-1$ terms. Further, we consider $j = 0, \ldots, x-1$. Hence we have x^2 terms. To get a uniform bound, let $\bar{b} = \max_{i,d}\{b_{u_i,d}\}$. We have a bound on (3.41) of

$$x(x-1) \cdot (\bar{b})^{x-1}\epsilon + o(\epsilon). \tag{3.42}$$

This bound dominates the $x \cdot (\bar{b})^{x-1}\epsilon$. Thus, the total error bound is given by

$$\left| \mathbb{P}_\infty(L_{k,d} \leq x|\Theta) - \mathbb{P}_\infty(L_{k,d} \leq x|\Theta^{(i)}) \right| \leq O(\epsilon \cdot x^2(\bar{b})^{x-1}) + o(\epsilon). \tag{3.43}$$

For S stages, the error terms can be determined through a recursive way. In stage s (i.e. the s^{th} block of the top row of blocks (which is all we need) taken to the n^{th} power gives that

$$\mathbf{T}^{S,(n)}_{u_1 \to u_s} = \sum_{j=0}^{n-1} (\mathbf{T}^1_{u_1})^j \cdot \left(\sum_{i=2}^{s} \mathbf{T}^i_{u_1} \cdot \mathbf{T}^{S-i+1,(n-j-1)}_{u_i \to u_s} \right). \tag{3.44}$$

That is, each term $\mathbf{T}^{i,(n-j-1)}_{u_i}$ is the block in stage i for the $n-j-1$ power of the phase-type matrix that excludes the first $i-1$ stations. Intuitively, this means we stay in station 1 for j units of time (being blocked) and then jump to stage i and stay in stage i for at least $n-j-1$ units of time. Plugging in the error bounds derived for smaller number of stages eventually gives us the final bound of

$$\left| \mathbb{P}_\infty(L_{k,d} \leq x|\Theta) - \mathbb{P}_\infty(L_{k,d} \leq x|\Theta^{(i)}) \right| \leq O(\epsilon \cdot x^S b^{x-1}) + o(\epsilon)$$

□

Remark 3.3.1. *In the single-stage scenario, we can further obtain an uniform bound for the gap in the ICT distributions. That is, the maximum distance is given by*

$$\max_{x\in\mathbb{N}} \left|\mathbb{P}_\infty(L_{k,d} \leq x|\Theta) - \mathbb{P}_\infty(L_{k,d} \leq x|\Theta')\right| \leq x^* p^{(x^*-1)} \cdot \epsilon + o(\epsilon), \tag{3.45}$$

where $x^* = \left\lceil \frac{p}{1-p} \right\rceil$*. This can be shown by noting that the sequence* $s_x = x \cdot p^{x-1}$ *on* $x \in \mathbb{N}$ *is either strictly decreasing in* x*, or is unimodal in* x*, first increasing and then decreasing. Since* $\lim_{x\to\infty} \frac{x}{x+1} = 1$*, there exists* x^* *such that the sequence is increasing prior to* x^* *and decreasing afterward, which implies that the maximum distance between the two ICT distributions can be found at* $x^* = \min\{x : \frac{x}{x+1} \geq p\}$*, which gives us the results in* (3.45)*.*

3.4 Case Studies of Itinerary Completion Improvement

In this section, we present a comprehensive case study applying our optimization to the problem of improving itinerary completion for breast cancer (BC) patients in our partner healthcare network. While our analytical framework and algorithm is easily scalable to optimizing itinerary completion rates for an ensemble of services offered across the entire healthcare network, we focus on the BC patients as the *target patients* in the case study as a proof of concept. The BC service line volunteered to provide contextualization and data as an initial pilot, and this service provides a sufficiently rich network and complexity of itineraries to demonstrate the full capabilities of our method. In Section 3.4.1, we introduce the dataset, the network setting of our healthcare partner, and the model parameterization for our numerical experiments. In Section 3.4.2, we compare our optimal template with the historical template for the current BC network, demonstrating that (1) our optimization algorithm can efficiently solve a large-scale, 26-station network, and (2) can significantly improve itinerary completion rates. In Section 3.4.3, we demonstrate the importance of an integrated optimization approach by highlighting the pitfalls and challenges of manual template design. In Section 3.4.4 we perform sensitivity analyses where we relax several analytical assumptions and generate generalizable insights to settings beyond our case study.

3.4.1 Dataset and Model Parametrization

Datasets. In this analysis, we leverage two separate datasets from Mayo clinic that span from 2006 to 2011: (1) patient appointment data and (2) staffing plan data. The patient appointment data contains the itinerary for each patient, the type of patient, their geo-code (e.g. local, national), appointment area (e.g. general surgery), appointment type (e.g. physician consult), and appointment day and time. The staffing data contains how many and which type of staff were scheduled to work and how many appointments were seen in each service for each day.

Network. Among all itineraries, 26 services were utilized by more than 0.5% of BC patients. Figure 3.1 shows a simplified diagram of the patient flow for five key breast cancer services (stations): breast diagnostic clinic (BDC), medical oncology (Med Onco), radiation oncology (Rad Onco), general surgery (Gen Surg), and plastic surgery (Plas Surg).

Patient types. Since we focus on one type of diagnosis, we differentiate the patient type by geo-code (international, national, regional, and local) in the case study. National/international patients are the *priority* patients and the others are non-priority, because the former are time-sensitive due to their travel constraints. Figure 3.3a displays the historical average number of root appointments for priority and non-priority patients; Table 3.2 provides details.

Care path. Care path probabilities, $p_{u,k,d}$, are estimated from the fraction of each patient type, k, requiring an appointment in service u on day d of their itinerary. In

	International	National	Regional	Local	Total
Mon	2%	36%	27%	35%	10.91
Tue	2%	45%	22%	32%	10.02
Wed	1%	43%	26%	30%	13.29
Thu	2%	38%	29%	32%	13.87
Fri	3%	36%	24%	36%	8.23
Total	2%	40%	26%	33%	56.33

Table 3.2: The Average Number of Root Appointments Allocated by Day of Week from Historical Data, and the Proportion from Each Patient Type

estimating care paths, we reduce the bias caused by blocking by using patient flow data from periods of low congestion during 2006 – 2011, where there was little to no blocking. Table 3.1 in Section 3.1 shows these probabilities for national patients in the first four stages.

Exogenous workload. We model non-BC patients as an exogenous, random arrival stream, and do not control their arrival allocations in this case study. But note that our model is fully capable of optimizing the appointment allocations beyond BC patients in an integrated system-wide implementation. We approximate the workload distribution from these exogenous arrivals as a normal random variable truncated at zero, with mean and variance parameters estimated from historical data. Table 3.3 summarizes the average total workload for each station on each day of the week, along with the percentage of workload contributed by BC patients.

Capacity. To estimate the capacity, we assume each patient takes one fixed slot, which is reasonable since appointment lengths are generally standard within each service. However, different stations may have different service times and/or hours of

Total workload	BDC	Med Onco	Rad Onco	Gen Surg	Plas Surg
Mon	48.5	123.6	91.9	56.0	54.0
Tue	43.6	145.1	95.5	64.5	59.8
Wed	53.2	141.4	102.7	68.3	54.5
Thu	55.7	124.1	108.2	65.9	53.5
Fri	33.9	102.2	56.6	38.3	38.9
% of BC	65%	2%	2%	26%	8%
Capacity	51	152	118	72	60

Table 3.3: Estimated Total Average Workload, Capacity, and Proportion of Workload Contributed by BC Patients for Each Station in the 5-Station Network

operation. To estimate capacities, $C_{u,d}$, we first estimate the appointment capacity per FTE (full-time equivalent) using the 95% percentile of the historical workload per FTE. We chose 95% since our healthcare partner indicated that patients are usually being "squeezed" into overtime in the top 5% of days worked. Leveraging the staffing data from a separate data set, we multiply capacity per FTE by the average number of staff on duty to obtain the total capacity; see Table 3.3.

Optimization objective and algorithm. In the baseline, we maximize the proportion of priority patients that complete their treatment by Friday. The target deadline is $T_d = 6 - d$ for patients starting their itineraries on workday $d = 1, \ldots, 5$. We set $w_k = 1$ in (3.14) for priority patients and 0 otherwise. In Section 3.4.3, we consider a setting where target completion rates for regional patients are incorporated as constraints.

For the iterative policy improvement algorithm, we obtain the initial blocking probabilities from the pre-processing workload smoothing optimization (see details in

Appendix D.1). We set the target service level, γ, to be close to the historical service level for each station to ensure that the optimal template does not significantly impact access for exogenous patients. We add a small tolerance term $\epsilon = 0.03$ in (3.17) to solve the optimization efficiently.

Simulation platform. We develop a discrete-event simulation using Python to evaluate the blocking probabilities and itinerary completion in the queueing network from various template designs; it also serves as a benchmark to evaluate the accuracy of the ICT approximations developed in Section 3.3.1 and Appendix B.3. We calculate performance metrics using batch means to obtain means and confidence intervals. We simulate the system for 20,000 weeks and divide it into 10 batches, where the first batch is excluded as a warm-up period. All simulation experiments are run on a desktop computer with an Intel i7-8700 CPU and 64GB of RAM.

We describe the flow of events in our discrete-even simulation, which is calibrated with the parameters introduced above. At the beginning of each day, we first determine the number of new patients. Patients getting root appointments arrive according to the template $\Theta = \{\Theta_{k,d}\}$ and are guaranteed to get their root appointments. To account for non-integer Θ, we generate arrivals according to a random variable $\lfloor \Theta_{k,d} \rfloor + \text{Bern}(\Theta_{k,d} - \lfloor \Theta_{k,d} \rfloor)$, where Bern$(p)$ is the Bernoulli random variable with parameter p, s.t. $(1 + p)\lfloor \Theta_{k,d} \rfloor = \Theta_{k,d}$. For each new patient of type k starting on day d, we generate a priori a realization of her care path by considering $S \cdot U$ independent Bernoulli trials, one for each station in each stage, where each trial is given by Bern$(p^k_{u,s})$. We generate exogenous arrivals from the fitted Gaussian distribution (truncated at zero). Each station maintains a pool of appointments, which records the unique ID (generated at admission time) of each patient that currently has an appointment at station u to be finished. To determine which appointments are fulfilled on a given day d in station u, we randomly shuffle the list of patients requesting an appointment and admit patients in their randomized order up to the capacity $C_{u,d}$, with the remaining (target) patients being blocked, staying in the pool of appointments for the next day; exogenous patients are lost in the no-retrial setting. If a patient is not blocked, they join the pool of appointments for the next station(s) of their pre-generated care path on the following day. On each day of the simulation, we record the total number of patients blocked, the blocking probability (fraction of patients blocked), and the number of exogenous patients blocked. For each patient, we also record her ICT and whether or not they complete by the target deadlines.

3.4.2 Itinerary Completion Results for Mayo Clinic Breast Cancer Patients

To demonstrate the computational efficiency of our algorithm, we solve the optimization for for the 26-station network for BC patients as described above. Table 3.4 reports the itinerary completion rates under the historical and optimal templates for the 26-station network. The results are obtained by simulating each template on the 26-station network. The optimal template, which is solved from our iterative optimization algorithm, can significantly improve itinerary completion for national and international patients, with only a small reduction for regional patients. While the optimal template benefits priority patients, seemingly at the expense of local patients, local patients live within 50 miles of the clinic and completing by Friday is not a significant concern. Further, the average itinerary completion *time* under the optimal template is actually shorter for regional and local patients due to lower blocking rates; 3.11 days under the historical template versus 3.03 days under the optimal template. For patients with no travel restrictions, the completion time is likely more important than completing by Friday.

We also compare the 26-resource network solution with the smaller 5-resource network solution (the five key resources in Figure 3.1); the results are nearly the same. As a result, we analyze the 5-resource network in the remainder of this case study to generate clearer insights.

Computation time. The 26-station network solves in about 30 minutes per iteration, compared with 5 minutes per iteration for the 5-station network. This suggests that the computational time increases linearly in the network size, demonstrating the scalability of our algorithm.

	International	National	Regional	Local
Historical	$38.0 \pm 2.3\%$	$58.8 \pm 0.2\%$	$62.3 \pm 0.5\%$	$62.9 \pm 0.4\%$
Optimal	$94.1 \pm 1.3\%$	$93.1 \pm 0.3\%$	$56.7 \pm 0.6\%$	$22.5 \pm 0.4\%$
5-resource template	$94.5 \pm 0.9\%$	$92.9 \pm 0.2\%$	$55.7 \pm 0.5\%$	$21.4 \pm 0.3\%$

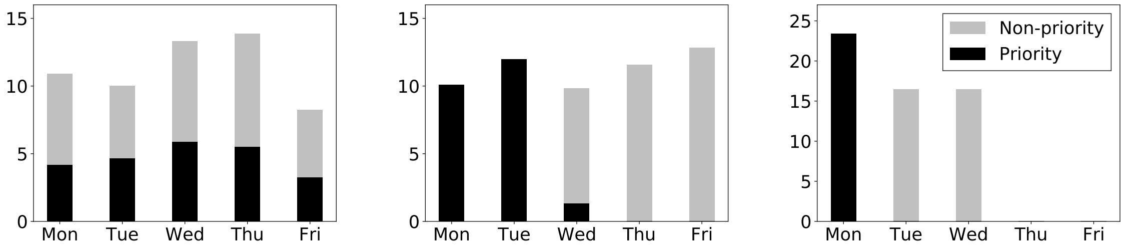
Table 3.4: Itinerary Completion Rates with 95% Confidence Intervals for the Historical Template, Optimal Template Solved from the 26-Station Full Network, and Optimal Template Solved from a Smaller, 5-Station Critical Resource Network

3.4.3 Pitfalls of Manual Template Design: Value of an Integrated Approach

In this section, we analyze the performance of our network optimization and highlight the key drivers behind poor itinerary completion rates for BC patients. First, the historical template does poorly by allocating too many priority patient slots near the end of the week, where they have little chance to complete their itineraries by Friday. The solution to this seems obvious: move priority appointments to the beginning of the week. However, we demonstrate the myopic nature of this approach and illustrate several pitfalls in template design that lead to (i) direct blocking, (ii) overflow blocking from non-priority patients, and (iii) network blocking. These occur because of a failure to consider (i) subsequent appointments generated from the root appointment, (ii) itineraries of non-priority patients, and (iii) workloads at other services in the network, respectively.

Direct blocking: front-loaded template.

Comparing the historical template in Figure 3.3(a) with the optimal template in Figure 3.3(b) we see a migration of priority patients from mid-late week to Monday and Tuesday primarily. To demonstrate that moving patients earlier in the week is only part of the benefit of the optimization, we also design a *front-loaded* template that moves priority patients to Monday to give them the greatest buffer between their root appointment and their deadline; see Figure 3.3(c). However, as we detail below, the benefits of this additional buffer for priority patients in the front-loaded template are dampened by the increase in direct blocking caused by overloading BDC on earlier days of the week. We call this direct blocking because most patients require a follow-up in BDC after their root appointment, and overloading the BDC service lengthens the entire itinerary.



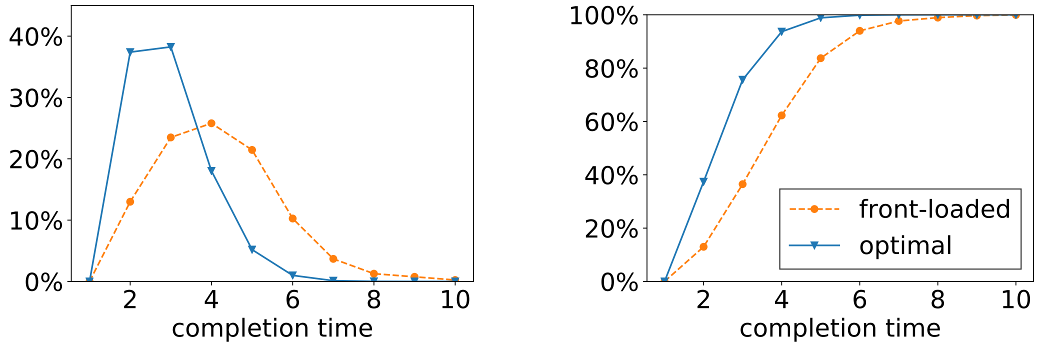
(a) Historical (b) Optimal (c) Front-loaded

Figure 3.3: Historical Template, Optimal Template, and Front-Loaded Template

Table 3.5 shows that the front-loaded template falls well short of the optimal performance. This can be explained by the ICT distributions shown in Figure 3.4. Under the optimal template, nearly all patients complete their itinerary within five

	International	National	Regional	Local
Historical	40.4% ± 1.6%	62.8% ± 0.3%	65.7% ± 0.4%	65.4% ± 0.5%
Front-loaded	75.0% ± 1.5%	84.1% ± 0.3%	73.8% ± 0.3%	73.6% ± 0.4%
Optimal	96.6% ± 1.0%	95.4% ± 0.2%	62.2% ± 0.4%	22.9% ± 0.2%

Table 3.5: Itinerary Completion Rates with 95% Confidence Intervals for the Historical Template, a Front-loaded Template, and the Optimal Template.



(a) Probability Distribution Function (PDF)

(b) Cumulative Distribution Function (CDF)

Figure 3.4: PDF and CDF of Itinerary Completion Time for National and International Patients Admitted on Monday

days, whereas 17% of patients in the front-loaded template have ICTs of at least five days, despite the fact that most patients only require three to four stages. This protracted ICT time is caused by direct blocking being as high as 45%-50% on Monday through Wednesday at BDC, where much of the initial diagnosis and treatment planning occurs. Blocking is less than 2% across all weekdays under the optimal template.

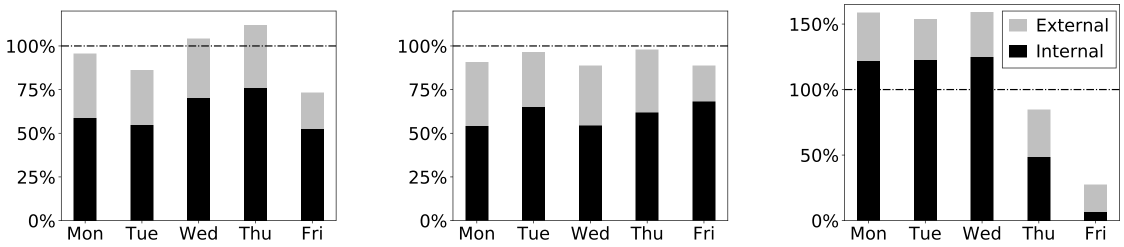
To further illustrate the impact of this direct blocking, Table 3.6 shows the overall average ICT and the average time to complete each stage for national patients admitted on Monday. ICTs are over a day longer under the front-loaded template, with much of the delays occurring in the first stage of the care path. This can be particularly frustrating for patients, as they travel to the clinic with the expectation of a quick turnaround only to find that they must wait days to get their second appointment.

	Average ICT	stage 1	stage 2	stage 3	stage 4
			Optimal		
require 3 stages	3.05 ± 0.01	1.00 ± 0.00	1.03 ± 0.01	1.02 ± 0.00	
require 4 stages	4.06 ± 0.01	1.00 ± 0.00	1.03 ± 0.01	1.02 ± 0.01	1.01 ± 0.01
			Front-loaded		
require 3 stages	4.28 ± 0.01	2.14 ± 0.01	1.06 ± 0.00	1.08 ± 0.01	
require 4 stages	5.35 ± 0.04	2.14 ± 0.03	1.06 ± 0.01	1.10 ± 0.01	1.05 ± 0.01

Table 3.6: Average ICT by Care Path Stage for National Patients Admitted on Monday

Note that stages two to four are also completed more quickly in the optimal template; this is due to reduced blocking in General Surgery and Plastic Surgery, which are the primary services required later in the itinerary. This failure to consider the other services on the care path is the third pitfall (network blocking), which we discuss in Section 3.4.3.

The analysis here highlights the importance of smoothing the workload to reduce blocking across the network. Figure 3.5 plots the workloads resulting from the three templates. Compared with the other templates, the optimal template creates lower internal (controllable by the template) workload on Wednesdays and Thursdays in response to the higher external (non-controllable) workloads. This generates a much smoother utilization across different days of the week. The front-loaded template highlights the importance of workload smoothing, since simply pushing priority patients as early in the week as possible is not sufficient. In contrast, our optimization performs a careful allocation of appointments across the week to balance the timing of priority patient root appointments with the more subtle cause for non-completion: blocking on the care path.



(a) Historical template

(b) Optimal template

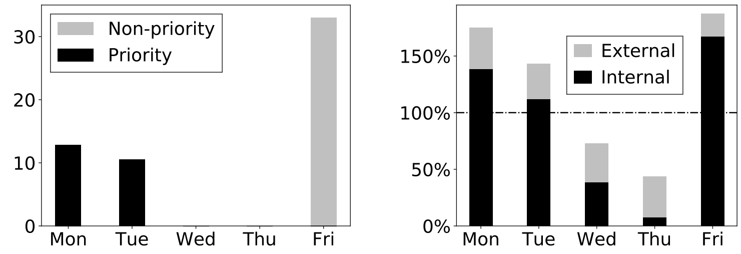
(c) Front-loaded template

Figure 3.5: Utilization of the BDC and Decomposition by Internal v.s. External

Overflow blocking and non-priority patients.

In the front-loaded template, direct blocking at BDC caused itinerary completion failures. To alleviate direct blocking, one might spread out the priority patients over the earlier days in the week, while allocating non-priority patients later in the week to clear the middle of the week for priority patients to complete their itineraries; see Figure 3.6(a) for this *historical-revised* template. Perhaps surprisingly, this template only performs marginally better than the front-loaded template in the completion rates (86% national and 74% international, versus 84% national and 75% international). Despite reducing direct blocking, blocking remains high in BDC on Monday

(50%) and Tuesday (35%). Here, however, the blocking is caused by non-priority patients that overflow to Monday of the next week from both blocking on Friday and regular subsequent appointments in stage two; note the high internal workloads on Monday and Tuesday shown in Figure 3.6(b). Focusing solely on priority patients does little to improve their completion rates and can significantly hurt regional patients. Though regional patients are not as sensitive to the Friday deadline, ignoring their completion rates entirely can be myopic. A more strategic approach would optimize priority patient completion with a guarantee on the fraction of regional patients that are also able to meet their completion target. Figure 3.7 plots the trade-off between completion rate of priority patients vs regional patients by adding a constraint for regional patient completion, which is varied on the x-axis. While the shape of this curve is intuitive, generating this output to support strategic management decision making would not be possible without our integrated framework.



(a) Historical revised template

(b) Utilization of BDC

Figure 3.6: Performance under the Historical Revised Template.

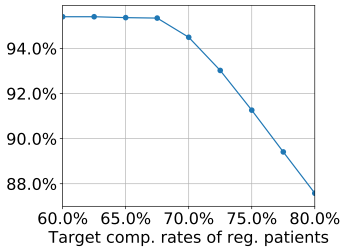


Figure 3.7: Completion Rates of Priority Patients v.s. Regional Patients.

Network Blocking.

The front-loaded and historical-revised templates highlight the importance of reducing blocking in BDC, where root appointments occur. However, in a coordinated care system, blocking in other services also affects itinerary completion rates. Here, we return to the last scenario considered in the previous subsection, in which we optimize priority patient itineraries while ensuring with at least a 75% completion rate for regional patients. We then compare the network-optimal template (solved from our optimization approach) with a *network-agnostic* template from an optimization that considers only BDC; see Table 3.7.

	International	National	Regional	Local
Network-optimal	$96.3 \pm 0.7\%$	$91.0 \pm 0.2\%$	$74.8 \pm 0.7\%$	$6.8 \pm 0.2\%$
Network-agnostic	$96.4 \pm 0.5\%$	$95.3 \pm 0.1\%$	$63.1 \pm 0.4\%$	$14.4 \pm 0.2\%$

Table 3.7: Impact of Network Dynamics: Comparing a Network-Agnostic Optimization with the Network Optimization

While the network-agnostic template assumes it achieves the target 75% completion rate for regional patients when optimizing with only BDC, the actual completion rate, as reported in Table 3.7, only reaches 63% when simulated in the five-resource network. This shows that a simpler model focusing on one main resource can introduce significant bias. Indeed, if the volume of exogenous patients in general surgery were to increase by 20%, which is common in practice due to shifting service-line strategies, regional completion drops even lower to 58%, while the network-optimal performance is barely affected. This ability to adjust to dynamic changes in auxiliary service-lines can be valuable for proactive decision making, and was a main impetus for this research.

Comparing the network-optimal and network-agnostic templates in Figure E.4 in Appendix E.3, it is easy to see where the network-agnostic one goes wrong. It allocates too many regional patient apopintment slots on Thursday (30% more than the network optimal), assuming that they will be guaranteed to complete their appointments at other services (e.g. General Surgery) without any delay. There are two key problems with this template. First, any delay in the broader network for a patient starting late in the week is likely to cause a completion failure. Second, BC patients often require a general surgery consult on the first or second stage after the root appointment at BDC. Thus, allocating too many patients on Thursday can exacerbate network blocking by overloading general surgery. This analysis highlights not only the importance of network-driven design, but also that it is critical to account for parallel appointments (as in Appendix B.3), though this feature has been largely overlooked in the literature.

Value of an integrated approach.

We conclude this section by highlighting the following key insights into the pitfalls of template design:

1. The historical template allocates too many appointments for the priority patients late in the week, leading to low itinerary completion rates. However, simply moving these priority patients to the beginning of the week is not necessarily beneficial due to direct blocking.

2. Although we optimize itinerary completion for priority patients, ignoring the impact of non-priority patients can backfire due to overflow blocking.

3. It is crucial to coordinate the network of care services; ignoring network dynamics can lead to significantly reduced performance relative to expectations.

3.4.4 Impact of Exogenous Retrials

For analytical tractability, we made the assumption that exogenous patients exit the system if they are blocked. To test the robustness of our model to this assumption, we conduct simulation experiments where exogenous patients are allowed to retry until they are able to obtain an appointment. This analysis further highlights the importance of our optimization approach, which is robust to exogenous retrials, while other templates are not.

Table 3.8 reports the completion rates of the front-loaded, historical revised and optimal templates. Compared to the scenario without exogenous patient retrials (see Table 3.5), the itinerary completion rate for priority patients drops by 7% in the front-loaded and historical revised templates while the rate for regional patients drops by 12%. Exogenous retrials create a cascading effect, propagating blocking across the days of the week. For example, in the front-loaded template the blocking on Monday only increases by 5%, whereas the blocking on Tuesday jumps by 14% due to the carry-over from exogenous patients blocked on Monday. This cascading effect pushes itineraries successively later in the week. The compounding effect of cascading blocking caused by the retrials that occur in the actual system makes it even more difficult to heuristically design a template that can work well in practice. On the other hand, the optimal solution has nearly identical performance to the scenario without retrials. The primary reason for the robustness of the optimal solution to the retrial assumption is that it accounts for the complex blocking dynamics across the network, which reduces the number of retrials and hence the impact of retrials on itinerary completion. This highlights the importance of accounting for blocking, which would be nearly impossible without a queueing network optimization framework like the one we develop in this paper.

	International	National	Regional	Local
Historical	$40.0 \pm 2.3\%$	$62.1 \pm 0.2\%$	$64.9 \pm 0.5\%$	$64.8 \pm 0.5\%$
Front-loaded	$68.3 \pm 1.8\%$	$76.8 \pm 0.4\%$	$60.7 \pm 0.7\%$	$60.5 \pm 0.5\%$
Historical Revised	$66.4 \pm 1.1\%$	$79.1 \pm 0.4\%$	$0.0 \pm 0.0\%$	$0.0 \pm 0.0\%$
Optimal	$96.5 \pm 0.4\%$	$95.0 \pm 0.2\%$	$62.0 \pm 1.0\%$	$22.7 \pm 0.1\%$

Table 3.8: Itinerary Completion Rates under Historical, Front-Loaded and Historical Revised Templates and the Optimal Template from Our Algorithm When Exogenous Patient Retrials are Allowed

Chapter 4

MEAN-FIELD ANALYSIS FOR APPROXIMATION OF BLOCKING PROBABILITIES

In this Chapter, we state the main results showing that the distributions of the blocking probabilities converge to point masses, in both the transient state and steady state. Specifically, in a network with n stations, for the transient analysis we show that the distribution of $U(t) = \{U_u(t),\ \forall u = 1, \ldots, n\}$ converges to the point mass $\beta(t) = \{\beta_u(t),\ \forall u = 1, \ldots, n\}$ as the size of the system $N \to \infty$; for their steady-state counterparts, we show that $U_\infty = \{U_{u,\infty}\}$ converge to the point mass $\beta_\infty = \{\beta_{u,\infty}\}$ as $N \to \infty$.

4.1 Transient Analysis

We begin by stating a general version of the results. That is, for a class of functions h satisfying certain conditions, the difference between $\mathbb{E}[h(U(t))]$ and $h(\beta(t))$ for each $t \geq 0$ is of order $O(1/N)$, where N is a scaling factor for the system size (e.g., capacity). Then, by choosing h to be a quadratic function in Corollary 4.1.1, we show the convergence of the stochastic system in mean square, which implies the convergence in probability.

In the standard mean-field framework, the total population size is fixed at N, and it is sufficient to describe the system dynamics using the occupancy measures $\{U_u(t)\}$ and $\{\mu_u(t)\}$. However, in our setting the total number of patients, $N_u(t)$, is changing in every period, preventing us from applying the existing results to establish the desired convergence results. To overcome this, we introduce an auxiliary variable, $V_u(t) = N_u(t)/N^q \in \mathcal{V}$ for some $q \geq 0$. For given N and q, it is equivalent to record

$V_u(t)$ as to record $N_u(t)$, which maintains the Markovian properties when tracking the dynamics from time 0 to t. Setting q properly allows us to establish the desired convergence results when $N \to \infty$. It is worth emphasizing that the occupancy measures, $\{U_u(t)\}$, are the actual variable of interest, not $\{V_u(t)\}$. Thus, the output of the testing functions h in Theorem 4.1.1 below depends only on $U_u(t)$'s. Examples of such functions include $h(u_1, v_1, \ldots, u_n, v_n) = u_1$ and $h(u_1, v_1, \ldots, u_n, v_n) = (u_1 - \bar{u})^2$ with $\bar{u}$ being some constant.

Theorem 4.1.1. *Consider a function $h : [0,1]^n \times \mathcal{V}^n \to \mathbb{R}$ that is continuous and twice differentiable, where the first derivative of h is $(1/\gamma)$-Lipschitz, i.e., $|h'(a) - h'(b)| \leq \frac{1}{\gamma}||a - b||$. Assume the following initial condition: $N_u(0) = n_u(0) = c_{u,0}N$ for each station u, where $c_{u,0}$ does not depend on N; and $U_u(0) = \mu_u(0)$ for each u. Then, for any fixed $t \geq 0$, if $q \geq 3/2$, we have that*

$$\Big|\mathbb{E}\left[h\left(U_1(t), V_1(t), \ldots, U_n(t), V_n(t)\right)\right] - h(\mu_1(t), v_1(t), \ldots, \mu_n(t), v_n(t))\Big| \leq \frac{c_t}{N}, \quad (4.1)$$

where $c_t > 0$ is a constant that is independent of N, q.

The main proof uses an induction argument, where a key lemma for the induction establishes a similar bound as in (4.1) for each t, conditioning on the state in $t-1$. This key lemma is proved by performing a Taylor expansion around $h(\mu_1(t-1), v_1(t-1), \ldots, \mu_n(t-1), v_n(t-1))$, where we show that the stochastic system and mean-field model agree in expectation for the one-step transition and hence the first-order term of the Taylor expansion can be cancelled out. Then, it is sufficient to simply bound the remainder term by $O(1/N)$. The complete proof is detailed in Appendix F.1.

Considering a particular testing function $h\big(u_1, v_1, \ldots, u_n, v_n\big) = \big(u_u - \beta_u(t)\big)^2$ for a given station $u \in \{1, \ldots, n\}$ gives us the following corollary.

Corollary 4.1.1. *Under the same conditions in Theorem 4.1.1, $U_u(t) \to \beta_u(t)$ in mean square as $N \to \infty$ for any given $t \geq 0$ and station u.*

Proof. Recall that $\mu_u(t) = \beta_u(t)$ for a given $u \in \{1, \ldots, n\}$. It is easy to verify that $h\big(u_1, v_1, \ldots, u_n, v_n\big) = \big(u_u - \beta_u(t)\big)^2$ satisfies the conditions for Theorem 4.1.1. Applying the theorem gives us $\mathbb{E}\left[(U_u(t) - \beta_u(t))^2\right] = O(1/N)$, i.e., convergence in mean square as $N \to \infty$. Note, mean-square convergence implies convergence in probability. □

4.2 Steady-State Analysis

For the steady-state analysis, we focus on the time-stationary setting where $\lambda_u(t) = \lambda_u$, $\lambda_{e,u}(t) = \lambda_{e,u}$, and $C_u(t) = C_u$ for each u. We denote $(U_{u,\infty}, V_{u,\infty})$ and $(\mu_{u,\infty}, v_{u,\infty})$ as the steady-state version of $(U_u(t), V_u(t))$ and $(\mu_u(t), v_u(t))$. In addition, $\mu_{u,\infty} = \beta_{u,\infty}$, the steady-state blocking probability, for each station u. We denote (U_∞, V_∞) and (μ_∞, v_∞) as the vector of these steady-state variables from all stations in the stochastic and deterministic systems.

Theorem 4.2.1. *Under the stability condition* (3.3)*,*

$$\mathbb{E}\left[||(U_\infty, V_\infty) - (\mu_\infty, v_\infty)||_2^2\right] = O\left(\frac{1}{N}\right). \tag{4.2}$$

The proof follows the Stein's method framework developed in Braverman and Dai (2017), Ying (2018), and Gast *et al.* (2018). The main difficulties in our setting include (i) we have a discrete-time system, not a continuous-time system as studied in most previous papers; (ii) the varying population size requires us to introduce the auxillary variable V_∞. Further, this variable can be unbounded, requiring us to conduct analysis separately on a bounded set and outside the bounded set; previous papers such as Gast *et al.* (2018) mostly work with bounded sets. We give a sketch of the proof below, and relegate the complete proof to Appendix F.2.

Proof. Sketch of proof. The key to the Stein's method framework is that, instead of directly bounding the difference between (U_∞, V_∞) and (μ_∞, v_∞), we bound the difference between the value functions from the Poisson equation with respect to the deterministic system. Consider a given state $(u, v) = (u_1, \ldots, u_n, v_1, \ldots, v_n)$. Let $G_t(u, v)$ and $\Psi_t(u, v)$ denote the generators for t-step transitions in the stochastic system and the mean-field model, respectively. The Poisson equation with respect to Ψ_1 can be written as:

$$f_g(u, v) = g(u, v) - g\big(\mu_\infty, v_\infty\big) + f_g\big(\Psi_1(u, v)\big),$$

or equivalently

$$g(u, v) - g\big(\mu_\infty, v_\infty\big) = f_g(u, v) - f_g\big(\Psi_1(u, v)\big). \tag{4.3}$$

Here, $g(u, v) = \sum_i (u_i - \mu_{i,\infty})^2 + (v_i - v_{i,\infty})^2$ and f_g is the (relative) value function, given as

$$f_g(u, v) = \sum_{t=0}^{\infty} \left[g\big(\Psi_t(u, v)\big) - g\big(\mu_\infty, v_\infty\big) \right].$$

f_g is well-defined since the deterministic system has a unique equilibrium solution under the stability condition (3.3). Next, taking expectation of (4.3) with respect to $(u, v) \sim (U_\infty, V_\infty)$,

$$\mathbb{E}\left[g\big(U_\infty, V_\infty\big) - g\big(\mu_\infty, v_\infty\big) \right] = \mathbb{E}\left[f_g(U_\infty, V_\infty) - f_g\big(\Psi_1(U_\infty, V_\infty)\big) \right].$$

Then, using the basic adjoint relationship $\mathbb{E}\left[f_g\big(G_1(U_\infty, V_\infty)\big) - f_g(U_\infty, V_\infty) \right] = 0$ for the stochastic system and adding this 0 term to the above equation, we get

$$\begin{aligned} \mathbb{E}\left[g\big(U_\infty, V_\infty\big) - g\big(\mu_\infty, v_\infty\big) \right] &= \mathbb{E}\left[f_g(U_\infty, V_\infty)) - f_g\big(\Psi_1(U_\infty, V_\infty)\big) \right] \\ &\quad + \mathbb{E}\left[f_g\big(G_1(U_\infty, V_\infty)\big) - f_g(U_\infty, V_\infty) \right] \\ &= \mathbb{E}\left[f_g\big(G_1(U_\infty, V_\infty)\big) - f_g\big(\Psi_1(U_\infty, V_\infty)\big) \right]. \end{aligned}$$

Now we have achieved *generator coupling* on the right-hand side of the above equation (Braverman and Dai, 2017). To prove (4.2), we just need to bound

$$\begin{aligned} & \mathbb{E}\left[f_g\big(G_1(U_\infty, V_\infty)\big) - f_g\big(\Psi_1(U_\infty, V_\infty)\big)\right] \\ = \; & \mathbb{E}\Big[\sum_{t=0}^{\infty}\left[g\big(\Psi_t(G_1(U_\infty, V_\infty))\big) - g\big(\Psi_t(\Psi_1(U_\infty, V_\infty))\big)\right]\Big]. \end{aligned}$$

This coupling allows us to work directly with the one-step generator G_1 and the deterministic transition generators Ψ_t. The rest of the proof involves:

1. We first consider when (u, v) are in a bounded set and show that there exists a T^* for any (u, v) in this set, such that the system enters a "contraction" region after T^*. After entering the contraction region, there exists a constant $(1-\epsilon) < 1$ such that, for any pair of states (u, v) and (u', v') that are sufficiently close to the equilibrium (μ_∞, v_∞), we have $|\Psi_1(u, v) - \Psi_1(u', v')| \leq (1-\epsilon)||(u, v) - (u', v')||$. Thus, we can bound $\mathbb{E}\left[\sum_{t=T^*+1}^{\infty}\left[g\big(\Psi_t(G_1(u, v))\big) - g\big(\Psi_t(\Psi_1(u, v))\big)\right]\right]$ by $O(1/N)$, leveraging the transient analysis combined with this contraction mapping. For each $t \leq T^*$, we also apply the results from the transient analysis by verifying $h = g \circ \Psi_t$ satisfies the conditions required for testing function in Theorem 4.1.1.

2. Next we analyze states outside the bounded set. Following the framework in Ying (2017), we show that, for any $v > r_b$ in the unbounded set, the deterministic process will be absorbed back into the bounded set in finite steps T_v. We then calculate the expected error incurred while in the unbounded set. The key insight is that we can bound this error by $\mathbb{E}\left[V_\infty - r_b\right] N^{-1/2}$; then replacing r_b with v_∞ and squaring the difference inside the expectation allows us to move this error term to the LHS of the bounding equation and combine it with $\mathbb{E}\left[||(U_\infty, V_\infty) - (\mu_\infty, v_\infty)||_2^2\right]$ – which is what we are trying to bound. Then,

combining the results on the bounded set, we show that the remaining terms on the RHS of the bounding equation is $O(1/N)$, which establishes the final result.

□

Finally, note that $B_u(t) = \mathbb{E}[U_u(t)|N(t)]$ and $B_u = \mathbb{E}[U_{u,\infty}|N_\infty]$, while the distributions of $U_u(t)$ and $U_{u,\infty}$ converge to the point masses $\beta_u(t)$ and $\beta_{u,\infty}$, which do not depends on $N_u(t)$ or N_∞. Thus, the distributions of $B_u(t)$ and B_u also converge to the same point mass. In other words, when the system size N is large, we can approximately replace the blocking probability of each patient by the deterministic number β, justifying approximation (3.13). We conclude our analysis by providing a bound on the phase-type distribution approximation in the policy evaluation step of our optimization algorithm.

Chapter 5

CONCLUSION

In this dissertation, we study scheduling in a co-located wireless network under both deadline and average power level constraints. We first formulate an optimization problem such that any power-control and scheduling algorithm that solves the optimization problem is throughput optimal. Then we propose a low complexity algorithm, named PDMax, and proved its throughput optimality. We compared the performance of PDMax with greedy-MaxWeight and LDF through simulations and showed that our algorithm outperforms the other two algorithms by achieving higher throughput and using lower average transmit power.

We also develop an optimization approach to a queueing network model for priority appointment allocation in a network of healthcare services with patient classes that have different time-sensitivities. We apply our new approach to meet itinerary completion deadlines for national/international patients at the Mayo Clinic. To capture the sojourn time in the queueing network, we design a phase-type approximation that we rigorously justify using mean field theory, with provable error bounds. We leverage this phase-type model in an iterative decomposition approach and provide bounds on the error introduced by the decomposition. This approach transforms the non-linear stochastic optimization into a sequence of tractable LP models that can efficiently optimize sojourn times relative to class-specific deadlines in large scale networks. Finally we present a case study of improving itinerary completion for breast cancer patients at the Mayo Clinic. In this study, we demonstrate that template design is a complex and multifaceted problem in which multiple aspects of the system's dynamics must be considered, including network blocking, timing of root appointments for priority patients, and the impact of non-priority and exogenous patients on itinerary completion. We illustrate some of the pitfalls of templates that fail to consider all the factors that impact itinerary completion, which makes manual template design extremely challenging. Simultaneously addressing all of these factors requires an analytical model, which we show can significantly improve itinerary completion for breast cancer patients traveling long distances to receive care.

REFERENCES

Aditya, D. and V. Rahul, "Online energy efficient packet scheduling with a common deadline", in "Proc. Int. Symp. Modelling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)", pp. 1–8 (Tempe, AZ, 2016).

Arulkumaran, K., M. P. Deisenroth, M. Brundage and A. A. Bharath, "Deep reinforcement learning: A brief survey", IEEE Signal Processing Magazine **34**, 6, 26–38 (2017).

Asmussen, S. and J. R. Møller, "Calculation of the steady state waiting time distribution in GI/PH/c and MAP/PH/c queues", Queueing systems **37**, 1-3, 9–29 (2001).

Baron, O., O. Berman, D. Krass and J. Wang, "Strategic idleness and dynamic scheduling in an open-shop service network: Case study and analysis", Manufacturing & Service Operations Management **19**, 1, 52–71 (2017).

Bertsekas, D. P., *Dynamic programming and optimal control*, vol. 1 (Athena scientific Belmont, MA, 1995).

Brassard, G. and P. Bratley, *Fundamentals of algorithmics*, vol. 33 (Prentice Hall Englewood Cliffs, 1996).

Braverman, A. and J. Dai, "Stein's method for steady-state diffusion approximations of $M/Ph/n + M$ systems", The Annals of Applied Probability **27**, 1, 550–581 (2017).

Braverman, A., J. Dai and J. Feng, "Stein's method for steady-state diffusion approximations: an introduction through the erlang-a and erlang-c models", Stochastic Systems **6**, 2, 301–366 (2017).

Bretthauer, K. M., H. S. Heese, H. Pun and E. Coe, "Blocking in healthcare operations: A new heuristic and an application", Production and Operations Management **20**, 3, 375–391 (2011).

Casale, G., "Approximating passage time distributions in queueing models by bayesian expansion", Performance Evaluation **67**, 11, 1076–1091 (2010).

Cayirli, T. and E. Veral, "Outpatient scheduling in health care: A review of literature", Production and Operations Management **12**, 4, 519–549 (2003).

Chow, V. S., M. L. Puterman, N. Salehirad, W. Huang and D. Atkins, "Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation", Production and Operations Management **20**, 3, 418–430 (2011).

Dai, J. and P. Shi, "A two-time-scale approach to time-varying queues in hospital inpatient flow management", Operations Research **65**, 2, 514–536 (2017).

Dai, J. and P. Shi, "Inpatient overflow: An approximate dynamic programming approach", Manufacturing & Service Operations Management **21**, 4, 894–911 (2019).

Davio, M., "Kronecker products and shuffle algebra", IEEE Transactions on Computers **30**, 2, 116–125 (1981).

Deglise-Hawkinson, J., J. E. Helm, T. Huschka, D. L. Kaufman and M. P. Van Oyen, "A capacity allocation planning model for integrated care and access management", Production and operations management **27**, 12, 2270–2290 (2018).

Deng, H. and I.-H. Hou, "On the capacity requirement for arbitrary end-to-end deadline and reliability guarantees in multi-hop networks", arXiv preprint arXiv:1704.04857 (2017).

Diamant, A., J. Milner and F. Quereshy, "Dynamic patient scheduling for multi-appointment health care programs", Production and Operations Management **27**, 1, 58–79 (2018).

Du, Y. and G. de Veciana, "Efficiency and optimality of largest deficit first prioritization: Resource allocation for real-time applications", in "Proc. IEEE Int. Conf. Computer Communications (INFOCOM)", (San Francisco, CA, 2016).

Ewaisha, A. E. and C. Tepedelenlioğlu, "Optimal power control and scheduling for real-time and non-real-time data", IEEE Transactions on Vehicular Technology **67**, 3, 2727–2740 (2017).

Feldman, J., N. Liu, H. Topaloglu and S. Ziya, "Appointment scheduling under patient preference and no-show behavior", Operations Research **62**, 4, 794–811 (2014).

Feng, J. and P. Shi, "Steady-state diffusion approximations for discrete-time queue in hospital inpatient flow management", Naval Research Logistics (NRL) **65**, 1, 26–65 (2018).

Gast, N., D. Latella and M. Massink, "A refined mean field approximation of synchronous discrete-time population models", Performance Evaluation **126**, 1–21 (2018).

Gocgun, Y. and A. Ghate, "Lagrangian relaxation and constraint generation for allocation and advanced scheduling", Computers & Operations Research **39**, 10, 2323–2336 (2012).

Gómez-Corral, A., "Sojourn times in a two-stage queueing network with blocking", Naval Research Logistics (NRL) **51**, 8, 1068–1089 (2004).

Goodfellow, I., Y. Bengio and A. Courville, *Deep learning* (MIT press, 2016).

Gue, K. R. and H. H. Kim, "An approximation model for sojourn time distributions in acyclic multi-server queueing networks", Computers & Operations Research **63**, 46–55 (2015).

Gurvich, I., “Diffusion models and steady-state approximations for exponentially ergodic Markovian queues”, The Annals of Applied Probability **24**, 6, 2527–2559 (2014).

Haviv, M. and J. van der Wal, “Mean sojourn times for phase-type discriminatory processor sharing systems”, European Journal of Operational Research **189**, 2, 375–386 (2008).

He, S., M. Sim and M. Zhang, “Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach”, Management Science **65**, 9, 4123–4140 (2019).

Helm, J. E. and M. P. Van Oyen, “Design and optimization methods for elective hospital admissions”, Operations Research **62**, 6, 1265–1282 (2014).

Hou, I.-H., V. Borkar and P. R. Kumar, “A theory of QoS for wireless”, in “Proc. IEEE Int. Conf. Computer Communications (INFOCOM)”, pp. 486–494 (Rio de Janeiro, Brazil, 2009).

Huang, J., B. Carmeli and A. Mandelbaum, “Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback”, Operations Research **63**, 4, 892–908 (2015).

Jaramillo, J. J., R. Srikant and L. Ying, “Scheduling for optimal rate allocation in ad hoc networks with heterogeneous delay constraints”, IEEE J. Sel. Areas Commun. **29**, 979–987 (2011).

Kang, X., I.-H. Hou and L. Ying, “On the capacity requirement of largest-deficit-first for scheduling real-time traffic in wireless networks”, in “Proc. ACM Int. Symp. Mobile Ad Hoc Networking and Computing (MobiHoc)”, (Hangzhou, China, 2015).

Kang, X., W. Wang, J. J. Jaramillo and L. Ying, “On the performance of largest-deficit-first for scheduling real-time traffic in wireless networks”, in “Proc. ACM Int. Symp. Mobile Ad Hoc Networking and Computing (MobiHoc)”, pp. 99–108 (Bangalore, India, 2013).

Kazemian, P., M. Y. Sir, M. P. Van Oyen, J. K. Lovely, D. W. Larson and K. S. Pasupathy, “Coordinating clinic and surgery appointments to meet access service levels for elective surgery”, Journal of biomedical informatics **66**, 105–115 (2017).

Latouche, G. and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modelling, 1st edition. Chapter 2: PH Distributions* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999).

Liu, X. and L. Ying, “Spatial-temporal routing for supporting end-to-end hard deadlines in multi-hop networks”, in “Proc. IEEE Conf. Information Sciences and Systems (CISS)”, pp. 262–267 (Princeton, NJ, 2016).

Liu, Y. and W. Whitt, “A fluid model for many-server queues with time-varying arrivals and phase-type service distribution”, ACM SIGMETRICS Performance Evaluation Review **39**, 4, 43–43 (2012).

Massey, W. and W. Whitt, “Networks of infinite-server queues with nonstationary Poisson input”, Queueing Syst **13**, 1, 183–250 (1993).

Ozawa, T., “Sojourn time distributions in the queue defined by a general QBD process”, Queueing Systems **53**, 4, 203–211 (2006).

Patrick, J., M. L. Puterman and M. Queyranne, “Dynamic multipriority patient scheduling for a diagnostic resource”, Operations Research **56**, 6, 1507–1525 (2008).

Pinedo, M., *Scheduling: theory, algorithms, and systems*, vol. 5 (Springer, 2012).

Robbins, H., “A remark on stirling’s formula”, The American mathematical monthly **62**, 1, 26–29 (1955).

Singh, R. and P. R. Kumar, “Throughput optimal decentralized scheduling of multi-hop networks with end-to-end deadline constraints: Unreliable links”, arXiv preprint arXiv:1606.01608 (2016).

Srikant, R. and L. Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective* (Cambridge University Press, 2014).

Sutton, R. S. and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).

Tarello, A., J. Sun, M. Zafer and E. Modiano, “Minimum energy transmission scheduling subject to deadline constraints”, Wireless Networks **14**, 5, 633–645 (2008).

Tassiulas, L. and A. Ephremides, “Dynamic server allocation to parallel queues with randomly varying connectivity”, IEEE Trans. Inf. Theory **39**, 466–478 (1993).

Wang, D., D. J. Morrice, K. Muthuraman, J. F. Bard, L. K. Leykum and S. H. Noorily, “Coordinated scheduling for a multi-server network in outpatient pre-operative care”, Production and Operations Management **27**, 3, 458–479 (2018).

Wang, D., K. Muthuraman and D. Morrice, “Coordinated patient appointment scheduling for a multistation healthcare network”, Operations Research **67**, 3, 599–618 (2019).

Yang, L., Y. E. Sagduyu, J. Zhang and J. H. Li, “Deadline-aware scheduling with adaptive network coding for real-time traffic”, IEEE/ACM Trans. Netw. **23**, 5, 1430–1443 (2015).

Ying, L., “Stein’s method for mean field approximations in light and heavy traffic regimes”, Proceedings of the ACM on Measurement and Analysis of Computing Systems **1**, 1, 1–27 (2017).

Ying, L., “On the approximation error of mean-field models”, Stochastic Systems **8**, 2, 126–142 (2018).

Zuo, S., H. Deng and I.-H. Hou, “Energy efficient algorithms for real-time traffic over fading wireless channels”, IEEE Trans. Wireless Commun. **16**, 3, 1881–1892 (2017).

APPENDIX A

COMPUTATIONAL COMPLEXITY FOR PDMAX ALGORITHM

For a link to transmit a number of packets in T consecutive time slots, we will first prove that the number of possible schedules is $\frac{\prod_{i=1}^{T}(a+i)}{T!}$ in Theorem A.0.1 and then utilize the Stirling's approximation formula to show it is at the order of $O(e^{a+T})$ in Corollary A.0.1. Lemma A.0.1 is used for proving an equation that is later used in Theorem A.0.1. Lemma A.0.2 is used in the proof of Corollary A.0.1.

Lemma A.0.1. *The following equality holds for any* $a = 0, 1, 2, \ldots$ *and* $t = 1, 2, 3, \ldots$

$$\begin{aligned}1 + \sum_{i=1}^{a} \prod_{j=0}^{i-1} \frac{a-j}{a+t-j} =& 1 + \frac{a}{a+t} + \frac{a(a-1)}{(a+t)(a+t-1)} + \ldots \\ &+ \frac{a(a-1)\cdots 1}{(a+t)(a+t-1)\cdots(t+1)} = \frac{a+t+1}{t+1}\end{aligned}$$

Proof. We use the induction for the proof. For $a = 0, 1, 2$, we have

$$\begin{aligned}1 + \sum_{i=1}^{0} \prod_{j=0}^{i-1} \frac{0-j}{0+t-j} =& 1 = \frac{0+t+1}{t+1}, \text{ for all } t = 1, 2, \ldots \\ 1 + \sum_{i=1}^{1} \prod_{j=0}^{i-1} \frac{1-j}{1+t-j} =& 1 + \frac{1}{t+1} = \frac{1+t+1}{t+1}, \text{ for all } t = 1, 2, \ldots \\ 1 + \sum_{i=1}^{2} \prod_{j=0}^{i-1} \frac{2-j}{2+t-j} =& 1 + \frac{2}{t+2} + \frac{2 \times 1}{(t+2)(t+1)} = \frac{2+t+1}{t+1}, \text{ for all } t = 1, 2, \ldots\end{aligned}$$

For $a = n$, we assume the following equality holds for all $t = 1, 2, \ldots$

$$\begin{aligned}1 + \sum_{i=1}^{n} \prod_{j=0}^{i-1} \frac{n-j}{n+t-j} =& 1 + \frac{n}{n+t} + \frac{n(n-1)}{(n+t)(n+t-1)} + \ldots \\ &+ \frac{n(n-1)\cdots 1}{(n+t)(n+t-1)\cdots(t+1)} = \frac{n+t+1}{t+1}.\end{aligned}$$

Then for the case of $a = n + 1$, by applying the induction hypothesis, we have the following:

$$\begin{aligned}
1 + \sum_{i=1}^{n+1} \prod_{j=0}^{i-1} \frac{n+1-j}{n+1+t-j} =& 1 + \frac{n+1}{n+1+t} + \frac{(n+1)n}{(n+1+t)(n+t)} + \dots \\
&+ \frac{(n+1)n(n-1)\cdots 1}{(n+1+t)(n+t)(n+t-1)\cdots(t+1)} \\
=& 1 + \frac{n+1}{n+1+t}\left(1 + \frac{n}{n+t} + \frac{n(n-1)}{(n+t)(n+t-1)} + \dots \right. \\
&\left. + \frac{n(n-1)\cdots 1}{(n+t)(n+t-1)\cdots(t+1)}\right) \\
=& 1 + \frac{n+1}{n+1+t}\frac{n+t+1}{t+1} \text{ (by assumption)} \\
=& 1 + \frac{n+1}{t+1} = \frac{(n+1)+t+1}{t+1},
\end{aligned}$$

for all $t = 1, 2, \dots$ □

Theorem A.0.1. *For a single link system with T time slots in a frame and a packets to transmit, denote $N(a, T)$ as the number of possible schedules, we have*

$$N(a, T) = \frac{\prod_{i=1}^{T}(a+i)}{T!}$$

Proof. We enumerate all the choices for each time slot:

- For time slot 1, possible numbers of packets to transmit are 0, $1, \dots, a$. Assume we choose s_1 packets to transmit.
- For time slot 2, possible numbers of packets to transmit are 0, $1, \dots, a - s_1$. Assume we choose s_2 packets to transmit.
  $\cdots$
- For time slot T, possible numbers of packets to transmit are 0, $1, \dots, a - (s_1 + s_2 + \dots + s_{T-1})$. Assume we choose s_T packets to transmit.

Thus the number of possible schedules are

$$N(a, T) = \sum_{s_1=0}^{a} \sum_{s_2=0}^{a-s_1} \cdots \sum_{s_T=0}^{a-(s_1+\dots+s_{T-1})} 1 = \sum_{s_T=0}^{a} \sum_{s_{T-1}=0}^{a-s_T} \cdots \sum_{s_1=0}^{a-(s_2+\dots+s_T)} 1. \tag{A.1}$$

The second equality comes from looking at the procedure of deciding the schedule backwards, i.e. from the last time slot to the first time slot. We will then use induction for the proof.

We begin with the case of $T = 1$ and $T = 2$. By applying (A.1), we have that the number of possible schedules for these two cases are

$$
\begin{aligned}
N(a,1) &= \sum_{s_1=0}^{a} 1 = (a+1) \text{ for all } a = 0,1,2,\ldots \\
N(a,2) &= \sum_{s_1=0}^{a} \sum_{s_2=0}^{a-s_1} 1 = \sum_{s_1=0}^{a} (a - s_1 + 1) = \frac{(a+2)(a+1)}{2 \times 1} \\
&\text{for all } a = 0,1,2,\ldots
\end{aligned}
$$

Assume that for $T = t$, we have

$$N(a,t) = \frac{\prod_{i=1}^{t}(a+i)}{t!} \text{ for all } a = 0,1,2,\ldots,$$

Then for $T = t+1$, we have

$$
\begin{aligned}
N(a,t+1) &= \sum_{s_{t+1}=0}^{a} \left(\sum_{s_t=0}^{a-s_{t+1}} \cdots \sum_{s_1=0}^{a-(s_2+\cdots+s_{t+1})} 1 \right) \\
&= \sum_{s_{t+1}=0}^{a} N(a - s_{t+1}, t) \\
&= \sum_{s_{t+1}=0}^{a} \frac{\prod_{i=1}^{t}(a - s_{t+1} + i)}{t!} \text{ (by assumption)} \\
&=\frac{\prod_{i=1}^{t}(a+i)}{t!} + \frac{\prod_{i=1}^{t}(a-1+i)}{t!} + \cdots + \frac{\prod_{i=1}^{t} i}{t!} \\
&=\frac{\prod_{i=1}^{t}(a+i)}{t!} \left(1 + \frac{a}{a+t} + \frac{a(a-1)}{(a+t)(a+t-1)} + \ldots \right. \\
&\quad \left. + \frac{a(a-1)\cdots 1}{(a+t)(a+t-1)\cdots(t+1)} \right) \\
&=\frac{\prod_{i=1}^{t}(a+i)}{t!} \frac{a+t+1}{t+1} \text{ (by Lemma A.0.1)} \\
&=\frac{\prod_{i=1}^{t+1}(a+i)}{(t+1)!},
\end{aligned}
$$

for all $a = 0,1,2,\ldots$ □

Lemma A.0.2. *For any $n = 7,8,9,\ldots,$ we have the inequalities*

$$\left(\frac{n}{3}\right)^n < n! < \left(\frac{n}{2}\right)^n.$$

Proof. By Robbins (1955) we have that for any $n = 1, 2, \ldots$ the following inequalities hold:

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}}.$$

To prove the first inequality, since $e < 3$, for any $n = 1, 2, \ldots$ we have that

$$\begin{aligned}\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n+1}} >& \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \\ =& \sqrt{2\pi n} \left(\frac{n}{e}\right)^n > \sqrt{2\pi n} \left(\frac{n}{3}\right)^n > \left(\frac{n}{3}\right)^n\end{aligned}$$

To prove the second inequality, since $e > \sqrt{2\pi} e^{\frac{1}{12n}}$ for any $n = 1, 2, \ldots$, it follows that

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}} < e n^{n+\frac{1}{2}} e^{-n}.$$

To compare the RHS of the above to $(n/2)^n$, for integer $n \geq 7$, we have that

$$\frac{e n^{n+\frac{1}{2}} e^{-n}}{\left(\frac{n}{2}\right)^n} = e\sqrt{n} \left(\frac{2}{e}\right)^n < 1.$$

Thus for any $n = 7, 8, 9, \ldots$, the following inequality holds

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}} < (n/2)^n .$$

□

Corollary A.0.1. *For a single link system with T time slots in a frame and a packets to transmit, the number of possible schedules is at least at the order of $O(e^{a+T})$.*

Proof. By Theorem A.0.1, we have

$$N(a, T) = \frac{\prod_{i=1}^{T}(a+i)}{T!} = \frac{(a+T)!}{a!\, T!}.$$

By Lemma A.0.2, we have that for sufficiently large n,

$$\left(\frac{n}{3}\right)^n < n! < \left(\frac{n}{2}\right)^n.$$

Thus, a lower bound for $N(a, T)$ under sufficiently large a and T is

$$\begin{aligned}N(a, T) >& \frac{\left(\frac{a+T}{3}\right)^{a+T}}{(T/2)^T (a/2)^a} = \frac{\left(\frac{a+T}{3}\right)^T \left(\frac{a+T}{3}\right)^a}{(T/2)^T (a/2)^a} \\ =& \left(\frac{2}{3}\right)^T \left(1 + \frac{a}{T}\right)^T \left(\frac{2}{3}\right)^a \left(1 + \frac{T}{a}\right)^a \\ =& \left(\frac{2}{3}\right)^{a+T} \left(1 + \frac{a}{T}\right)^T \left(1 + \frac{T}{a}\right)^a,\end{aligned}$$

Since $\lim_{T\to\infty} \left(1 + \frac{a}{T}\right)^T = e^a$ and $\lim_{a\to\infty} \left(1 + \frac{T}{a}\right)^a = e^T$, we have that for sufficiently large $a + T$,

$$N(a, T) > \left(\frac{2}{3}\right)^{a+T} e^{a+T}.$$

□

APPENDIX B

ADDITIONAL DETAILS ON THE PATIENT FLOW MODEL, MEAN FILED MODEL AND ITINERARY COMPLETION TIME FOR SETTINGS WITH PARALLEL APPOINTMENTS

B.1 Settings Without Parallel Appointment

Denote the total number of target patients blocked at the end of day $t+1$ as $M_u^B(t+1)$. Recall that $N_u(t+1)$ is the total number of requests and $B_u(t+1)$ is the blocking probability. Conditioning on $N_u(t+1) = n$ and $\Lambda_u^e(d(t+1)) = \lambda_e$, and denoting $\big(N_u(t+1) - C_{u,d(t+1)}\big)^+ = b$, we have that $M_u^B(t+1)$ follows the hypergeometric distribution

$$\mathbb{P}\big(M_u^B(t+1) = k | b, \lambda_e\big) = \frac{c(n-\lambda_e, k)\cdot c(\lambda_e, b-k)}{c(n,b)}, \tag{B.1}$$

where $c(a,b)$ is the binomial coefficient. Note that M_u^B does not follow a binomial distribution because, in total, there are exactly b patients blocked. Thus, we are sampling b patients *without* replacement from the joint pool of target patients and exogenous patients, whereas the binomial distribution assumes each patient has an independent Bernoulli trial, i.e., sampling *with* replacement. In other words, the blocking events are correlated among patients.

Once we have $M_u^B(t+1)$, under the random ordering assumption, the joint distribution of $\{M_{u,s}^{B,k}(t+1), \forall k,s\}$ follows the multinomial distribution with parameters $M_u^B(t+1)$ and probabilities $\Big\{\frac{N_{u,s}^k(t+1)}{\sum_{k,s} N_{u,s}^k(t+1)}, \forall k,s\Big\}$, and $M_{u,s}^{NB,k}(t+1) = N_{u,s}^k(t+1) - M_{u,s}^{B,k}(t+1)$.

B.2 Patient Flow and Meal-Field Models for Settings with Parallel Appointments

To specify the patient flow model with potentially parallel appointments in each stage, we denote $\mathcal{R}_{k,s} = \{u_1, u_2, \dots, u_n\}$ as the set of resources that are required to complete stage s for a type k patient. We further define the following vector tracking the appointment completion status for resource group $\mathcal{R}_{k,s}$ as

$$\mathcal{B}_{\mathcal{R}_{k,s}} = (a_{u_1}, a_{u_2}, \dots, a_{u_n}),$$

where for each $u_j \in \mathcal{R}_{k,s}$, $a_{u_j} = 0$ indicates that the appointment at resource u_j has not yet been completed and $a_{u_j} = 1$ indicates that it has been completed. Given $N_{k,s}$ total possible resources that can be used by a type k patient at stage s, each resource group is one possible combination of the $N_{k,s}$ resources. That is, $\mathcal{R}_{k,s} \in \mathcal{P}(\{u_1, \dots, u_{N_{k,s}}\})$, where $\mathcal{P}$ is the power set.

For illustration purposes, we explain the patient flow model by considering the case where each stage contains only *two* stations for the patient to visit; the model framework can be generalized easily. In this scenario, the blocking status vector $\mathcal{B}_{\mathcal{R}_{k,s}}$ can take four possible values: $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$, where $(1,1)$ represents that the patient finished all appointments required in the current stage and is ready to move to the next stage on her care path.

Now, we define the following patient count that differentiates not only by k,s but also by the blocking status. That is, we denote $M_{\mathcal{R},s}^{k,\mathcal{B}}(t)$ that counts the total number of type k patients whose blocking status is $\mathcal{B} = (a_1, a_2)$ in stage s, at the end of day t. We drop the index k,s from $\mathcal{B}$ and $\mathcal{R}$ for notational simplicity. The total number of target patients requesting an appointment from station u on day $t+1$ is a random variable that follows:

$$N_{u,s}^{k}(t+1) = \sum_{\mathcal{R}:u\in\mathcal{R}} \sum_{\mathcal{B}:a_u=0} M_{\mathcal{R},s}^{k,\mathcal{B}}(t) + \tilde{M}_{\mathcal{R},s}^{k}(t) + \Lambda_u(t+1).$$

Here, the first double-summation represents all the patients who are in stage s and need to visit station u, yet have not finished their appointments at u. The second term represents the patients who have finished all appointments in stage $s-1$ by the end of day t and now need to visit station u in stage s. That is,

$$\tilde{M}_{\mathcal{R},s}^{k}(t) = \sum_{\mathcal{R}_{s-1}} Mult\big(M_{\mathcal{R},s-1}^{k,(1,1)}(t), p_{\mathcal{R}_{s-1},\mathcal{R}_s}\big),$$

where each term in the summation denotes the number of patients, out of those $M_{\mathcal{R},s-1}^{k,(1,1)}(t)$ who completed all appointments in stage $s-1$, that request appointments from resource group $\mathcal{R}_s$ in stage s with probability $p_{\mathcal{R}_{s-1},\mathcal{R}_s}$, where this patient count follows a multinomial distribution.

Transitions in the Stochastic System

Once we get $N_{u,s}^{k}(t+1)$, we can then define $N_u(t+1)$ similarly as in Section 3.1.1 and calculate $B_u(t+1)$ in (3.1). To obtain $M_{\mathcal{R},s}^{k,\mathcal{B}}(t+1)$, we first calculate the following intermediate variables: $N_{\mathcal{R},s}^{b,k,\mathcal{B}}(t+1;u)$, which counts the number of patients, out of all patients in the same category determined by $(k,s,\mathcal{B},\mathcal{R})$, that are blocked at station u on day $t+1$. As in Section 3.1.1, we get $M_u^B(t+1)$ from the hypergeometric distribution that samples blocked patients *without* replacement from the joint pool of the target patients and exogenous patients. Then, the joint distribution of $N_{\mathcal{R},s}^{b,k,\mathcal{B}}(t+1;u)$'s follow a multinomial distribution with parameters $M_u^B(t+1)$ and the proportions of patients from the corresponding category of $(k,s,\mathcal{B},\mathcal{R})$.

With these intermediate variables, we can characterize the transitions in the patient counts to the new blocking status. Here, we give a sketch of this characterization using an example from the two-resource setting. Assume that for a given $k,s,\mathcal{R}$, there are three patients in status $(0,0)$ who need appointments from both stations u_1 and u_2. Conditioning on $N_{\mathcal{R},s}^{b,k,\mathcal{B}}(t+1;u_1)=2$ and $N_{\mathcal{R},s}^{b,k,\mathcal{B}}(t+1;u_2)=1$, we then need to enumerate over all possible sequences of the three patients. Index the three patients with $1,2,3$. Then we have the following possibilities for the blocking status in u_1 and u_2:

$$\{(1,2),3;(1,3),2;(2,3),1\} \text{ for } u_1, \quad \{(1),2,3;(2),1,3;(3),1,2\} \text{ for } u_2,$$

where the patients in the parenthesis are blocked. We then enumerate over all the possible combinations of the blocking status between the two stations and then get the new $M_{\mathcal{R},s}^{k,\mathcal{B}}(t+1)$. For example, if we have $(1,2),3$ in u_1 and $(1),2,3$ in u_2, then we know that, out of the three patients, one of them stays in status $(0,0)$, one transitions to status $(0,1)$, and one transitions to $(1,1)$. Other cases can be derived similarly.

Mean-Field Model

Once we have the stochastic patient flow model, the corresponding mean-field model can be written by taking the expectation of the random quantities, similar to what we show in Section 3.3.2. In particular, in the mean-field model, the transitions from $m^{k,\mathcal{B}}_{\mathcal{R},s}(t)$'s to $m^{k,\mathcal{B}}_{\mathcal{R},s}(t+1)$'s (the deterministic counterparts for $M^{k,\mathcal{B}}_{\mathcal{R},s}(\cdot)$) are much simplified. For example, for a patient in a starting status $(0,0)$, the transition probabilities into status $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$ are simply $\beta_{u_1}(t+1)\beta_{u_2}(t+1)$, $\big(1-\beta_{u_1}(t+1)\big)\beta_{u_2}(t+1)$, $\beta_{u_1}(t+1)\big(1-\beta_{u_2}(t+1)\big)$, and $\big(1-\beta_{u_1}(t+1)\big)\big(1-\beta_{u_2}(t+1)\big)$, respectively, where $\beta_u(t+1)$ is the blocking probability on day $t+1$ in the mean-field model. The stability condition (3.3) ensures that we have an equilibrium solution $\beta = \{\beta_{u,d}\}$, which can be solved numerically from the mean-field model.

B.3 Itinerary Completion Time for Settings With Parallel Appointments

In this section, we generalize the characterization of the ICT distribution to the setting where patients may require parallel appointments from multiple stations in the same stage, in addition to probabilistic resource requirements and multiple stages of treatment presented in Section 3.3. These features are critical to itinerary completion as seen in our data, but have not been considered in prior works such as Casale (2010). Modeling r parallel appointments requires calculating the generator matrix for the maximum of r phase-type distributions. In the interest of space, we focus on considering parallel appointments in a single stage here and relegate the extension to the general multi-stage model to Appendix C.

Consider r parallel appointments at stations $u_1, u_2, \ldots, u_r$. The time to obtain an appointment from each station is characterized by the phase-type random variables $X_1, X_2, \ldots, X_r$. The time to complete all appointments in this stage is thus given by $X = \max\{X_1, X_2, \ldots, X_r\}$. If we directly apply results on the maximum of multiple phase-type distributions, e.g., see Davio (1981), the generator matrix for X involves a recursive calculation. Given r phase-type RVs, each with a generator matrix $\mathbf{T}_i$ $(i = 1, \ldots, r)$ of size $N \times N$, the generator matrix $\mathbf{T}_X$ is of size $\big((N+1)^r-1\big) \times \big((N+1)^r-1\big)$. As r increases, the size of the generator matrix will experience exponential growth. In our case study, for a five-station network, $\mathbf{T}_X$ (in one stage) has a size of $7,775 \times 7,775$, with $60,450,625$ entries. This makes solving any realistically sized networks computationally intractable. To overcome this computational challenge, we develop a new state transformation, which leads to an exact, yet compact representation of the generator matrix.

A Compact Representation for the Generator Matrix.

We develop an alternate, equivalent representation for the generator matrix, $\mathbf{T}_X$, by leveraging special structures of the network sojourn problem. The key is to introduce a transformed state by noting that (i) transitions can only occur between the current day and the next day, i.e., from day 1 to day 2, or day 2 to day 3; and (ii) the time for attempting to obtain parallel appointments is synchronized, e.g., it is impossible to have a state where an appointment for station 1 is attempting to get scheduled on day 1 while another appointment for station 2 is attempting to get scheduled on day 3. To specify the state transformation, we first define the Kronecker product.

Definition B.3.1. *The Kronecker Product of matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$, $A \otimes B$, *is an operation on two matrices, such that if* $\mathbf{A}$ *is an* $m \times n$ *matrix and* $\mathbf{B}$ *is a* $p \times q$ *then* $A \otimes B$ *is a* $mp \times nq$ *matrix that can be written as follows:*

$$A \otimes B = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}. \tag{B.2}$$

The next proposition shows the generator matrix of X using the state transformation.

Proposition B.3.1. *Let* $X_1, \dots, X_r$ *be* r *phase-type r.v.. Then* $X = \max_{j=1}^{r} X_j$ *follows a phase-type distribution with the following generator matrix:*

$$\mathbf{V} = \begin{bmatrix} & \textit{Day 1} & \textit{Day 2} & \textit{Day 3} & \textit{Day 4} & \textit{Day 5} & \textit{Absorb} \\ \textit{Day 1} & \boldsymbol{0} & \mathbf{V}_{1,2} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \mathbf{V}_1^0 \\ \textit{Day 2} & \boldsymbol{0} & \boldsymbol{0} & \mathbf{V}_{2,3} & \boldsymbol{0} & \boldsymbol{0} & \mathbf{V}_2^0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \textit{Day 5} & \mathbf{V}_{5,1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \mathbf{V}_5^0 \\ \textit{Absorb} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & 1 \end{bmatrix}. \tag{B.3}$$

Here, $\mathbf{V_{i,i+1}}$ *and* $\mathbf{V_i^0}$ *are specified through the Kronecker product as follows*

$$\left[\begin{array}{c|c} \mathbf{V_{i,i+1}} & \mathbf{V_i^0} \\ \hline \boldsymbol{0} & 1 \end{array}\right] = \bigotimes_{j=1}^{n} A_{u_j,i}, \tag{B.4}$$

with $A_{u_j,i}$ *being the following* 2×2 *matrix that depends on the blocking probabilities* $\beta_{u_j,i}$

$$A_{j,i} = \begin{bmatrix} \beta_{u_j,i} & 1 - \beta_{u_j,i} \\ 0 & 1 \end{bmatrix}. \tag{B.5}$$

Proof. We give a sketch of proof by construction. Each matrix block, $\mathbf{V_{i,i+1}}$ or $\mathbf{V_i^0}$, represents the transitions from day i to day $i+1$. In this case, either at least one of the r appointments has not yet been finished and must retry on day $i+1$, given by $\mathbf{V_{i,i+1}}$; or all r appointments are completed to reach the absorbing state, given by $\mathbf{V_i^0}$. Consider the following state representation for the underlying DTMC that governs the transition in $\mathbf{V_{i,i+1}}$ or $\mathbf{V_i^0}$: $(a_1, a_2, \dots, a_L)$, where $a_j = 1$ means that the patient has completed her appointment at station j and $a_j = 0$ means that the patient has not yet completed her appointment at station j. Under this state representation, the transition probabilities in $\mathbf{V_{i,i+1}}$ or $\mathbf{V_i^0}$ are characterized by all possible outcomes of Bernoulli retrial for all stations with $a_j = 0$ on day i (i.e., the stations that have not yet been completed), which can be seen to be equivalent to the Kronecker product in (B.4). □

In the case study of the five-station network in Section 3.4, this compact state representation only requires multiplication of matrices of size 31×31, having 961 entries, compared to the original matrix size with $60,450,625$ entries. Appendix C specifies the generator matrix in the most general multi-stage setting with both parallel appointments and probabilistic resource requirements. Once we characterize the generator matrix, to calculate the ICT distribution, we then replace the stochastic blocking probabilities by their point mass from the mean-field model. The mean-field model for the general setting with parallel appointments is detailed in Appendix B.2.

APPENDIX C

A COMPREHENSIVE FRAMEWORK FOR PHASE-TYPE APPROXIMATION

We consider the transition from one stage (with multiple appointments) to next stage (with multiple appointments). We first define a set, $\mathcal{R} = \{u_1, u_2, \dots, u_n\}$, that contains the group of resources that are required to complete a particular stage of treatment. We define the state space of the phase-type distribution for completing all the appointments in group $\mathcal{R}$ as

$$\mathcal{S}_{\mathcal{R}} = \{(a_{u_1}, a_{u_2}, \dots, a_{u_n}, d)\},$$

where for each $u_j \in \mathcal{R}$, $a_{u_j} = 0$ indicates that the appointment at resource u_j has not yet been completed on day d and $a_{u_j} = 1$ indicates that it has been completed. For a given group of resources, $\mathcal{R}$, the size of the state space is $5 \cdot (2^{|\mathcal{R}|} - 1)$, which is the same as the size of the matrix block that defines the phase-type distribution for completing all the appointments in resource group $\mathcal{R}$. We adjust by -1 because if not all a_{u_j}'s can be 1; otherwise, it should not be in this stage anymore. Given N total possible resources that can be used by the patient, each group, $\mathcal{R}$, is one possible combination of the N resources. That is, $\mathcal{R} \in \mathcal{P}(\{u_1, \dots, u_N\})$, where $\mathcal{P}$ is the power set, or the set of all possible subsets of $\{u_1, \dots, u_N\}$.

C.1 Deterministic Resource Requirements

We first consider the case where each group of appointments must be finished before moving to the next stage, i.e., no probabilistic resource requirements. For the ease of exposition, we begin by illustrating the phase-type generator for a simpler setting with m stages, where each stage contains only *two* stations for the patient to visit. That is, let $s = 1, 2, \dots, m$ denote one of the m stages, and $u_{s,j} \in \{u_1, \dots, u_N\}$ $(j = 1, 2,\ s = 1, \dots, m)$ denote the j^{th} station the patient needs to visit in stage s. The state for stage s is given by

$$\mathcal{S}_{\mathcal{R}_s} =(a_{u_{s,1}}, a_{u_{s,2}}, d)$$

where $d = 1, \dots, 5$ represents the day of week. We specify the transition matrix for the general setting at the end of this subsection.

The transition matrix has a similar block structure as the one shown in (3.22), where we replace the blocking $b_{u_j,d}$ probabilities (or non-blocking probabilities) at each station in blocks $T^1_{u_j}$ (or $T^2_{u_j}$) by a matrix that captures the blocking (or non-blocking) probabilities for finishing all of the appointments in the current *resource group*, instead of in a single station as in (3.22). We describe this new structure by comparing with each block entry in (3.22).

Block Corresponding to Transitions Within One Stage

We first characterize $V^1_{\mathcal{R}_s}$, which corresponds to $T^1_{u_s}$ in (3.22). As mentioned, for exposition, here we consider both resources are needed in stage s, i.e. $\mathcal{R}_s = \{u_{s,1}, u_{s,2}\}$ for $s = 1, 2, \dots, m$, Note that there can be at most three combinations for $(a_{u_{s,1}}, a_{u_{s,2}})$: $(0, 0)$, $(0, 1)$, and $(1, 0)$, denoting that neither appointment was able to be scheduled, the appointment in $u_{s,2}$ was able to be scheduled but not $u_{s,1}$, and the appointment in $u_{s,1}$ was able to be scheduled but not $u_{s,2}$, respectively.

The transition from day d to day $d+1$ corresponding to sample paths where the patient has not completed both appointments in stage s is thus given by:

$$\mathbf{V}^1_{\mathcal{R}_s}(d,d+1) = \left[\begin{array}{llll} & (0,0,d+1) & (0,1,d+1) & (1,0,d+1) \\ (0,0,d) & \beta_{u_{s,1},d}\cdot\beta_{u_{s,2},d} & \beta_{u_{s,1},d}(1-\beta_{u_{s,2},d}) & (1-\beta_{u_{s,1},d})\beta_{u_{s,2},d} \\ (0,1,d) & 0 & \beta_{u_{s,1},d} & 0 \\ (1,0,d) & 0 & 0 & \beta_{u_{s,2},d} \end{array}\right], \tag{C.1}$$

where $\mathbf{V}^1_{\mathcal{R}_s}(d,d+1)$ is the multi-appointment analogue of the $\beta_{u_s,d}$ in (3.22), that is, the matrix that represents the time to complete both jobs $u_{s,1}$ and $u_{s,2}$ in stage s – also the maximum of two phase-type distributions.

As a result, block $V^1_{\mathcal{R}_s}$ can be written similarly to $T^1_{u_s}$ as

$$\mathbf{V}^1_{\mathcal{R}_s} = \left[\begin{array}{llllll} & \mathit{Day}\ 1 & \mathit{Day}\ 2 & \mathit{Day}\ 3 & \mathit{Day}\ 4 & \mathit{Day}\ 5 \\ \mathit{Day}\ 1 & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_s}(1,2) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathit{Day}\ 2 & \mathbf{0} & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_s}(2,3) & \mathbf{0} & \mathbf{0} \\ \mathit{Day}\ 3 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_s}(3,4) & \mathbf{0} \\ \mathit{Day}\ 4 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_s}(4,5) \\ \mathit{Day}\ 5 & \mathbf{V}^1_{\mathcal{R}_s}(5,1) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array}\right], \tag{C.2}$$

Note that the process will stay in block $\mathbf{V}^1_{\mathcal{R}_s}$ until all the appointments in group $\mathcal{R}_s$ have been able to be successfully scheduled. The transitions to the next stage, $s+1$, where all appointments in stage s have been completed are given below.

Block Corresponding to Transitions from One Stage to Another Stage

Let $\mathbf{V}^2_{\mathcal{R}_s\to\mathcal{R}_{s+1}}(d,d+1)$ represent the transition from the group of resources, $\mathcal{R}_s$ on day d to the group of resources $\mathcal{R}_{s+1}$ on day $d+1$. This block is the multi-appointment analogue of to $T^2_{u_s}$ in (3.22).

Recall, for illustration, we let $\mathcal{R}_{s+1} = \{u_{s+1,1}, u_{s+1,2}\}$, which indicates that two resources are required in stage $s+1$. Hence specifying the transition from states $(a_{u_{s,1}}, a_{u_{s,2}}, d)$ to states $(a_{u_{s+1,1}}, a_{u_{s+1,2}}, d+1)$ gives us

$$\mathbf{V}^2_{\mathcal{R}_s,\mathcal{R}_{s+1}}(d,d+1) = \left[\begin{array}{llll} & (0,0,d+1) & (0,1,d+1) & (1,0,d+1) \\ (0,0,d) & (1-\beta_{u_{s,1},d})(1-\beta_{u_{s,2},d}) & 0 & 0 \\ (0,1,d) & (1-\beta_{u_{s,1},d}) & 0 & 0 \\ (1,0,d) & (1-\beta_{u_{s,2},d}) & 0 & 0 \end{array}\right], \tag{C.3}$$

Note, when the process initially enters $s+1$ on day $d+1$, none of the new apointments generated for this stage have been completed yet. Hence it is only possible to transition to state $(0,0,d+1)$ (indicating that none of the new appointments for stage $s+1$ have yet been completed) among all the states for $(a_{u_{s+1,1}}, a_{u_{s+1,2}}, d+1)$.

As a result, block $\mathbf{V}^2_{\mathcal{R}_s \to \mathcal{R}_{s+1}}$ can be written as follows by stacking all the terms of $\mathbf{V}^2_{\mathcal{R}_s,\mathcal{R}_{s+1}}(d, d+1)$:

$$\mathbf{V}^2_{\mathcal{R}_s \to \mathcal{R}_{s+1}} = \begin{bmatrix} \mathbf{0} & \mathbf{V}^2_{\mathcal{R}_s,\mathcal{R}_{s+1}}(1,2) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}^2_{\mathcal{R}_s,\mathcal{R}_{s+1}}(2,3) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}^2_{\mathcal{R}_s,\mathcal{R}_{s+1}}(3,4) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}^2_{\mathcal{R}_s,\mathcal{R}_{s+1}}(4,5) \\ \mathbf{V}^2_{\mathcal{R}_s,\mathcal{R}_{s+1}}(5,1) & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{C.4}$$

Full Transition Matrix

Now we are ready to specify the full transition matrix.

$$\left[\begin{array}{c|c} \mathbf{T}_{\mathcal{C}} & \mathbf{T}^{\mathbf{0}}_{\mathcal{C}} \\ \hdashline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{cccccc|c} \mathbf{V}^{\mathbf{1}}_{\mathcal{R}_\mathbf{1}} & \mathbf{V}^{\mathbf{2}}_{\mathcal{R}_\mathbf{1} \to \mathcal{R}_\mathbf{2}} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^{\mathbf{1}}_{\mathcal{R}_\mathbf{2}} & \mathbf{V}^{\mathbf{2}}_{\mathcal{R}_\mathbf{2} \to \mathcal{R}_\mathbf{3}} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ & \ddots & \ddots & & & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{V}^{\mathbf{1}}_{\mathcal{R}_\mathbf{m}} & \mathbf{V}^{\mathbf{0}}_{\mathcal{R}_\mathbf{m}} \\ \hdashline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & 1 \end{array}\right] \tag{C.5}$$

Here, $\mathbf{V}^{\mathbf{0}}_{\mathcal{R}_\mathbf{m}}$ is defined similarly to $\mathbf{V}^{\mathbf{2}}_{\mathcal{R}_\mathbf{s} \to \mathcal{R}_\mathbf{s+1}}$ $(s < m)$, except for each block $\mathbf{V}^0_{\mathcal{R}_m}(d, d+1)$, all transitions are to the absorbing state.

Extension to General Numbers of Stations in Each Stage

Using the compact form (B.4), we can easily extend the above descriptions to settings where there are more than two stations to be visit in each stage. Consider the process is in set of resources $\mathcal{R}_s$ in stage s and transitions to resources $\mathcal{R}_{s+1}$ in stage $s+1$. Then we can define the transitions as follows:

$$\left[\begin{array}{c|c} \mathbf{V}^1_{\mathcal{R}_s}(d, d+1) & \mathbf{V}^2_{\mathcal{R}_s}(d, d+1) \\ \hdashline \mathbf{0} & 1 \end{array}\right] = \bigotimes_{u_j \in \mathcal{R}_s} A_{u_j,d}, \tag{C.6}$$

which gives us $\mathbf{V}^1_{\mathcal{R}_s}(d, d+1)$ to specify the block $\mathbf{V}^1_{\mathcal{R}_s}$ and also

$$\mathbf{V}^2_{\mathcal{R}_r,\mathcal{R}_{s+1}}(d, d+1) = \begin{bmatrix} \mathbf{V}^2_{\mathcal{R}_r}(d, d+1) & \mathbf{0}_{|\mathcal{S}(\mathcal{R}_s)| \times |\mathcal{S}(\mathcal{R}_{s+1})|-1} \end{bmatrix}. \tag{C.7}$$

Here, $\mathbf{0}_{|\mathcal{S}(\mathcal{R}_s)| \times |\mathcal{S}(\mathcal{R}_{s+1}|}$ in (C.7) is a $|\mathcal{S}(\mathcal{R}_s)| \times |\mathcal{S}(\mathcal{R}_{s+1}|$ matrix of zeros, and $A_{u_j,i}$ is the following 2×2 matrix:

$$A_{u_j,i} = \begin{bmatrix} \beta_{u_j,i} & 1 - \beta_{u_j,i} \\ 0 & 1 \end{bmatrix}. \tag{C.8}$$

C.2 Probabilisitic Resource Requirements

In this section, we consider the case where not each resource is required to visit in each stage. The basic setting for this probabilistic resource requirement case is similar as described above. Consider N total possible resources, the possible resource groups form the powerset $\mathcal{P}(\{u_1, \ldots, u_N\})$. Again, for the purposes of exposition, we first present a simplified setting, where in each stage $s = 1, \ldots, m$, there are two possible resources that *may* require appointments, $u_{s,1}, u_{s,2}$. We then specify the general form at the end of this subsection.

In the simplified setting, the possible groups of resources for stage s are given by $\mathcal{R}_s \in \mathcal{P}(\{u_{s,1}, u_{s,2}\} = \{\{u_{s,1}, u_{s,2}\}, \{u_{s,1}\}, \{u_{s,2}\}, \emptyset\}$. In other words, for stage s, there are four possible outcomes: visiting both stations, visiting $u_{s,1}$, visiting $u_{s,2}$, or skip this stage, with probabilities $\mathbb{P}(\{u_{s,1}, u_{s,2}\}) = \nu_{u_{s,1}}\nu_{u_{s,2}}$, $\mathbb{P}(\{u_{s,1},\}) = \nu_{u_{s,1}}(1 - \nu_{u_{s,2}})$, and $\mathbb{P}(\{u_{s,2},\}) = (1 - \nu_{u_{s,1}})\nu_{u_{s,2}}$, and $\mathbb{P}(\emptyset) = (1 - \nu_{u_{s,1}})(1 - \nu_{u_{s,2}})$, respectively. In our phase-type generator matrix specified below, we omit the last one, the null set $\emptyset$, since it contains no resources. Let $\mathcal{R}_{s,k}$, $k = 1, \ldots, 3$ be the set of appointments that need to be completed prior to exiting stage s, representing each of the three non-empty outcomes.

To capture all possible resource groups, we first need to enlarge the state space to be

$$\big(S_{\mathcal{R}_{s,1}}, S_{\mathcal{R}_{s,2}}, S_{\mathcal{R}_{s,3}}, d\big),$$

where $S_{\mathcal{R}_{s,k}}$ the state space defined above for a given resource group $\mathcal{R}_{s,k}$, i.e., the tuple of a_u's for all $u \in \mathcal{R}_{s,k}$. For example, $S_{\mathcal{R}_{s,1}} = \big\{(a_{u_{s,1}}, a_{u_{s,2}})\big\}$ since appointments are required in both resources for group $\mathcal{R}_{s,1}$. Similarly, $S_{\mathcal{R}_{s,2}} = \big\{a_{u_{s,1}}\big\} = \{0\}$ since only $u_{s,1}$ is required in resource group $\mathcal{R}_{s,2}$ (and we only need $a_{u_{s,1}} = 0$ to track whether stage s is finished or not when there is a single resource).

Block Corresponding to Transitions Within One Stage

The transitions within a single stage are defined by the phase-type block matrix U_r^1, which is the probabilistic analogue to V_r^1 defined in the previous subsection. In the simplifed setting, we have

$$\mathbf{U}_r^1 = \begin{bmatrix} \mathbf{V}^1_{\mathcal{R}_{s,1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_{s,2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}^1_{\mathcal{R}_{s,3}} \end{bmatrix}, \tag{C.9}$$

Here, each block represents the phase-type transitions for attempting to complete all the appointments in the corresponding resource group. That is, block $V^1_{\mathcal{R}_{s,k}}$ captures the transitions prior to completing all the appointments in the resource group $\mathcal{R}_{s,k}$ and can be specified as the general form of (C.2) with $\mathbf{V}^1_{\mathcal{R}_{s,k}}(d, d+1)$ given by (C.6). The reason why $\mathbf{U}_r^1$ has a diagonal structure is that, once a process enters a given block corresponding to group $\mathcal{R}_{s,k}$, it will stay in that block until all the appointments for resources in the group have been scheduled. Then the process will leave the block according and transitions to stage $r + 1$ with probability defined by the blocks that follow. Note that $\mathcal{R}_{s,k}$'s are not necessarily all the same size, as the size of each block is determined by the number of resources in that block.

Block Corresponding to Transitions from Stage s to Stage $s+1$.

In the setting with probabilistic resource requirements, transition from stage s to $s+1$ indicates that the patient has completed all appointments required in stage s. We specify the block for transitions out of stage s (i.e. completion of all required appointments), U_r^2, which corresponds to $T_{u_s}^2$ in (3.22). Again, for illustration, we focus on the simplified setting where there are only two stations that may need to be visited in stage s and stage $s+1$, i.e., each stage has three possible resource groups (excluding the empty set) denoted as $\mathcal{R}_{s,k}$ and $\mathcal{R}_{s+1,k}$ ($k = 1, 2, 3$), respectively. In this simplified setting, the block is given by

$$\mathbf{U}_r^2 = \left[\begin{array}{c|ccc} & \mathcal{R}_{s+1,1} & \mathcal{R}_{s+1,2} & \mathcal{R}_{s+1,3} \\ \hline \mathcal{R}_{s,1} & p_{s+1,1}\cdot\mathbf{V}^2_{\mathcal{R}_{s,1}\to\mathcal{R}_{s+1,1}} & p_{s+1,2}\cdot\mathbf{V}^2_{\mathcal{R}_{s,1}\to\mathcal{R}_{s+1,2}} & p_{s+1,3}\cdot\mathbf{V}^2_{\mathcal{R}_{s,1}\to\mathcal{R}_{s+1,3}} \\ \mathcal{R}_{s,2} & p_{s+1,1}\cdot\mathbf{V}^2_{\mathcal{R}_{s,2}\to\mathcal{R}_{s+1,1}} & p_{s+1,2}\cdot\mathbf{V}^2_{\mathcal{R}_{s,2}\to\mathcal{R}_{s+1,2}} & p_{s+1,3}\cdot\mathbf{V}^2_{\mathcal{R}_{s,2}\to\mathcal{R}_{s+1,3}} \\ \mathcal{R}_{s,3} & p_{s+1,1}\cdot\mathbf{V}^2_{\mathcal{R}_{s,3}\to\mathcal{R}_{s+1,1}} & p_{s+1,2}\cdot\mathbf{V}^2_{\mathcal{R}_{s,3}\to\mathcal{R}_{s+1,2}} & p_{s+1,3}\cdot\mathbf{V}^2_{\mathcal{R}_{s,3}\to\mathcal{R}_{s+1,3}} \end{array}\right], \tag{C.10}$$

While (C.10) may seem complex, it is easily interpreted. First, we have added labels for the rows and columns. The row labels, $\mathcal{R}_{s,k}$, indicate which resource set the patient has just completed in stage s. The column labels, $\mathcal{R}_{s+1,\ell}$, represent which resource group the patient will require in stage $s+1$ of treatment. Note these labels are for exposition and do not represent the matrix states, nor do they directly indicate the size of the blocks, which is fully defined by the $\mathbf{V}$'s. Inside the matrix, $p_{s+1,\ell}$ is the probability of requiring resource group $\mathcal{R}_{s+1,\ell}$ in stage $s+1$. Thus, each block entry (row $\mathcal{R}_{s,k}$ to column $\mathcal{R}_{s+1,\ell}$) represents the probability of finishing the remaining appointments of resource group $\mathcal{R}_{s,k}$ and transitioning to resource group $\mathcal{R}_{s+1,\ell}$, denoted by $\mathbf{V}^2_{\mathcal{R}_{s,k}\to\mathcal{R}_{s+1,\ell}}$; this event occurs with probability $p_{s+1,\ell}$. Recall from before that, for a given k, the matrix blocks $\mathbf{V}^2_{\mathcal{R}_{s,k}\to\mathcal{R}_{s+1,\ell}}$ only differ from each other by the number of zero blocks required to expand the state space to the appropriate size for resource group $\mathcal{R}_{s+1,\ell}$, which may be different for each ℓ.

Block Corresponding to Transitions from Stage s to Stage $s+d$.

The transitions from stage s to $s+d$ are defined by matrix U_s^{d+1}, which has the exact same form as (C.10) except that we replace resource group $\mathcal{R}_{s+1,\ell}$ with resource group $\mathcal{R}_{s+d,\ell}$ and resource group probabilities $p_{s+1,\ell}$ with $p_{s+d,\ell}$, where

$$p_{s+d,1} = \nu_{u_{s+d,1}}\nu_{u_{s+d,2}}\Pi_{q=s+1}^{s+d-1}(1-\nu_{u_{q,1}})(1-\nu_{u_{q,2}})$$

and $p_{s+d,2}$, $p_{s+d,3}$ can be defined similarly by replacing the first two terms with $\nu_{u_{s+d,1}}(1-\nu_{u_{s+d,2}})$ and $(1-\nu_{u_{s+d,1}})\nu_{u_{s+d,2}}$, respectively.

Full Transition Matrix

Now we are ready to specify the full transition matrix with at total of m possible stages.

$$\left[\begin{array}{c|c} \mathbf{T}_{\mathcal{C}} & \mathbf{T}^{\mathbf{0}}_{\mathcal{C}} \\ \hline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{cccccc|c} \mathbf{U^1_1} & \mathbf{U^2_1} & \mathbf{U^3_1} & \mathbf{U^4_1} & \dots & \mathbf{U^m_1} & \mathbf{U^0_1} \\ \mathbf{0} & \mathbf{U^1_2} & \mathbf{U^2_2} & \mathbf{U^3_2} & \dots & \mathbf{U^{m-1}_2} & \mathbf{U^0_2} \\ & \ddots & \ddots & & & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{U^1_m} & \mathbf{U^0_m} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & 1 \end{array}\right], \tag{C.11}$$

$$\mathbf{U}^0_s = \Pi^{m-s}_{q=s+1}(1-\nu_{u_{q,1}})(1-\nu_{u_{q,2}}) \cdot \begin{bmatrix} \mathbf{V}^0_{R_{s,1}} \\ \mathbf{V}^0_{R_{s,2}} \\ \mathbf{V}^0_{R_{s,3}} \end{bmatrix} \tag{C.12}$$

Note that the last column, $\mathbf{U}^0_s$, captures transitions to the absorbing state, where each block $\mathbf{V}^0_{R_{s,k}}$ is defined similarly as $\mathbf{V}^0_{R_m}$ in (C.5).

APPENDIX D

REFINED BLOCKING APPROXIMATION AND WORKLOAD SMOOTHING

D.1 Refined Blocking Approximation: Offered-Load Approximation for Workload Distribution

To obtain analytical forms for the blocking probability, we leverage the normal approximation for the distribution of the steady-state workload, $D^{\infty}_{u,d}$, and characterize its mean and standard deviation. We approximate $D^{\infty}_{u,d}$ with a normal r.v. with mean $\mu_{u,d}(\Theta)$ and standard deviation $\sigma_{u,d}(\Theta)$. We denote the pdf of this normal r.v. as $\phi_{u,d}(\cdot)$. Then, we approximate the probability that there are x patients in need of service from the station as $\pi(x) = \int_{x-0.5}^{x+0.5} \phi_{u,d}(s)ds$, and thus, the blocking probability as:

$$\beta_{u,d}(\Theta) \approx \sum_{x \geq C_{u,d}} \pi(x) \cdot \frac{x - C_{u,d}}{x}, \tag{D.1}$$

where $C_{u,d}$ is the capacity constraint for station u on weekday day d. Because the queueing system we study has batch arrivals instead of continuous Poisson arrivals, the traditional normal excess probability for approximating blocking does not apply here. For a given static template $\Theta = \{\Theta_{k,d}\}$, we first consider the deterministic case where $\Theta_{k,d}$ type k patients start their itinerary on each workday $d = 1, \ldots, 5$ in each week.

Next, we specify how to calculate $\mu_{u,d}(\Theta)$ and $\sigma_{u,d}(\Theta)$. Since a patient may take more than one week to finish her itinerary, the total workload on a given day could come from patients scheduled in this week and all previous weeks. Let $\bar{n}$ be the maximum number of weeks that a patient is allowed to spend in the system during one itinerary. Suppose we are in week 0 and consider weekday d. Then, to account for workloads from earlier weeks that accumulate to d, we need to trace back all days that start from the current day up to day d in $\bar{n}$ weeks earlier. Let $p_{u,d,k}(t)$ be the probability that a type k patient starts her itinerary $t = \tilde{d} + 5n$ days ago and requires an appointments from station u on the current day d. For notational convenience, we further define $\hat{p}^{n,\tilde{d}}_{u,d,k} = p_{u,d,k}(\tilde{d} + 5n)$. Then, the mean and variance of the workload in station u on day d follow:

$$\mu_{u,d}(\Theta) = \sum_{k \in \mathfrak{K}} \sum_{\tilde{d}=1}^{5} \sum_{n=0}^{\bar{n}-1} \Theta_{k,\tilde{d}} \cdot \hat{p}^{n,\tilde{d}}_{u,d,k}. \tag{D.2}$$

$$\sigma^2_{u,d}(\Theta) = \sum_{k \in \mathfrak{K}} \sum_{\tilde{d}=1}^{5} \sum_{n=0}^{\bar{n}-1} \Theta_{k,\tilde{d}} \cdot \hat{p}^{n,\tilde{d}}_{u,d,k}(1 - \hat{p}^{n,\tilde{d}}_{u,d,k}). \tag{D.3}$$

The probability distribution $p_{d,k,u}(\cdot)$ can be estimated from the data to capture different care pathways and the possible correlations among the visits on the pathway.

D.2 Workload Smoothing

In this section, we present the pre-processing stage of performing workload smoothing to obtain an initial schedule for our iterative policy optimization framework. The main idea is to minimize blocking (i.e., the event where a patient can't get an appointment on the day they request) over all services across the week. To explain the rationale, recall that our eventual objective is to minimize the deadline-violation probability. The more likely a patient is going to be blocked, the longer the sojourn time of each patient is, and thus, the more likely the deadline is violated.

The objective of this workload smoothing stage is to find an initial schedule Θ such that the overall probabilities that the workload at each station exceeds the capacity are minimized. However, the blocking probabilities are not linear in the decision variable Θ. To linearize the objective, we approximate $\phi_{u,d}(\cdot)$ with a piecewise linear function that anchors on the set of discrete points $m(i)$. Then, for example, we can calculate the integral over $\phi_{u,d}(\cdot)$ by a linear summation of the values in each interval $[m(i), m(i+1))$ of the piecewise linear function.

To formulate the workload smoothing optimization, we also need to design a set of constraints that sets the workload realization in each interval $[m(i), m(i+1))$ to be consistent with the blocking probability. For each interval which is characterized by the starting grid point $m(i)$, the workload realization in this interval equals $\mu_{u,d}(\Theta) + m(i)\sigma_{u,d}(\Theta)$. However, $\sigma_{u,d}(\Theta)$ is still non-convex in the decision variable Θ because it is the square root of the variance, even though the variance $\sigma^2_{u,d}(\Theta)$ is linear in Θ from (D.3). To remove this non-convexity, we propose to approximate the square root of $\sigma^2_{u,d}(\Theta)$ with the following approach based on Newton's method. That is, let $\hat{\sigma}_{u,d}$ be an initial guess for the standard deviation, and the one-step Newton's method gives us

$$\sigma_{u,d}(\Theta) \approx \frac{1}{2}\left(\frac{\sigma^2_{u,d}(\Theta)}{\hat{\sigma}_{u,d}} + \hat{\sigma}_{u,d}\right),$$

which is then linear in terms of Θ. In our application, a high level of accuracy can be achieved if $\hat{\sigma}$ is set to the standard deviation of the historical workload of the current system.

We now formally state the LP that minimizes the blocking probabilities across all stations in the healthcare network. Again, without loss of generality, we consider a planning horizon of $1, \ldots, 5$ corresponding to each workday in a week. Let $\hat{\theta}_{k,d}$ be the maximum number of type k patients allowed to be scheduled on day d, and Q be some large constant value. We have that

$$\min_{\Theta,\delta} \sum_{u\in\mathcal{U}} \sum_{d=1}^{5} \sum_{i\in\mathcal{M}'} \Big(\Phi(m(i+1)) - \Phi(m(i))\Big)\delta_{u,d,i} \tag{D.4}$$

$s.t.$

$$\mu_{u,d}(\Theta) + m(i)\cdot\frac{1}{2}\left(\frac{\sigma^2_{u,d}(\Theta)}{\hat{\sigma}_{u,d}} + \hat{\sigma}_{u,d}\right) - C_{u,d} \le Q\cdot\delta_{u,d,i} \quad \forall u\in\mathcal{U}, i\in\mathcal{M}', d=1,\ldots,5 \tag{D.5}$$

$$\delta_{u,d,i+1} \ge \delta_{u,d,i} \quad \forall i\in\mathcal{M}' \tag{D.6}$$

$$\sum_{d=1}^{5}\Theta_{k,d} \ge \theta_k \quad \forall k\in\mathfrak{K} \tag{D.7}$$

$$\Theta_{k,d}\in\mathbb{R}^+, \delta_{u,d,i}\in\{0,1\} \quad \forall k\in\mathfrak{K}, d=1,\ldots,5. \tag{D.8}$$

We explain this formulation as follows:

1. The objective (D.4) will drive the system to minimize the sum of the approximated blocking probabilities across the week over all stations in the healthcare network. (One can also modify the objective to incorporate different weights to reflect that blockage in some services may be more critical than in others.)

2. The set of auxiliary variables $\delta = \{\delta_{u,d,i}\}$, with $\delta_{u,d,i}$ being a surrogate to help maintain the consistency between the workload realization and the blocking instance. To see this and explain constraint (D.5), note that the objective drives the program to *minimize* $\delta_{u,d,i}$'s since $\Big(\Phi(m(i+1)) - \Phi(m(i))\Big)$ has the same (positive) value under our grid partition. Thus, if the realized workload on the interval $[m(i), m(i+1))$, $\mu_{u,d}(\Theta) + m(i) \cdot \frac{1}{2}\left(\frac{\sigma^2_{u,d}(\Theta)}{\hat{\sigma}_{u,d}} + \hat{\sigma}_{u,d}\right)$ is small than the capacity $C_{u,d}$, i.e., no blocking, then this minimization objective will force $\delta_{u,d,i}$ to take the value 0, the smallest value as the constraint allows. Meanwhile, if the realized workload is larger than the capacity $C_{u,d}$, $\delta_{u,d,i}$ will be set to 1; the constraint will still be satisfied since the right-hand side of (D.5) is large after multiplying by the large constant Q. The standard deviation $\sigma^2_{u,d}(\Theta)$ here is approximated with the Newton's method to linearize this constraint.

3. Constraint (D.6) is added to speed up solving the MIP.

4. Finally, constraint (D.7) says that the weekly volume meets the minimum throughput requirement.

APPENDIX E

ADDITIONAL NUMERICAL RESULTS

E.1 Numerical Validation for the Phase-Type Approximation

In this section, we numerically show that the ICT approximation is still remarkably accurate in the more general setting with parallel appointments. The experiments are performed on the five-station network as illustrated in Figure 3.1, which is parameterized using data from our healthcare partner; see Section 3.4 for details.

For a given arrival template Θ, we compare the ICT distribution obtained from the phase-type approximation (3.13) with the empirical distribution obtained from simulating the system. To obtain the blocking probabilities for the phase-type approximations, we simulate the system under the template Θ. The details of the dataset, parameterization, and simulation setup are introduced in Section 3.4 of the main paper. Figure E.1 compares the ICT distributions from simulation and from phase-type approximation for national patients who start their itineraries on Monday. Figure E.1a demonstrates that the phase-type distribution is remarkably close to the simulated distribution. Applying the Kolmogorov-Smirnov (KS) statistic, we find that the maximum distance between the ICT distributions from the simulation and the approximation is less than 6% across all patient types and starting days for all the experimental settings we have tested.

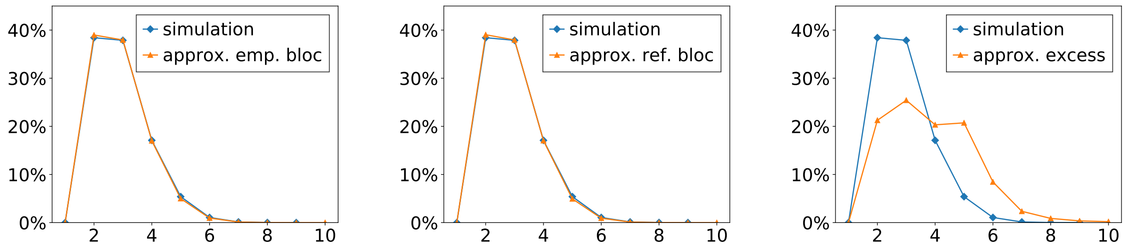
We also evaluate the phase-type approximation for ICT when using the *approximated* blocking probabilities from (D.1) as the input. Figure E.1b compares the ICT distributions from simulation and from the phase-type approximation using the approximate blocking probabilities. This figure demonstrates that the phase-type approximation quality is not significantly impacted by the blocking approximation that we use for computational efficiency. To further demonstrate the importance of using (D.1) to approximate the blocking probabilities, Figure E.1c shows the ICT distribution using the conventional method of calculating blocking using the normal *excess probability*,

$$\beta_{u,t} \approx 1 - \Phi\left(\frac{C_{u,t} - \mu_{u,t}}{\sigma_{u,t}}\right).$$

Clearly, using (D.1) to approximate blocking significantly improves the approximation accuracy. The Kolmogorov-Smirnov (KS) statistic comparing the phase-type distribution using the approximated blocking probabilities with the simulated distribution has a maximum distance of less than 7% across all patient types and stations, compared with 24% when employing the conventional excess probability; the median distance is 2% when using (D.1) versus 8% when using excess probability.

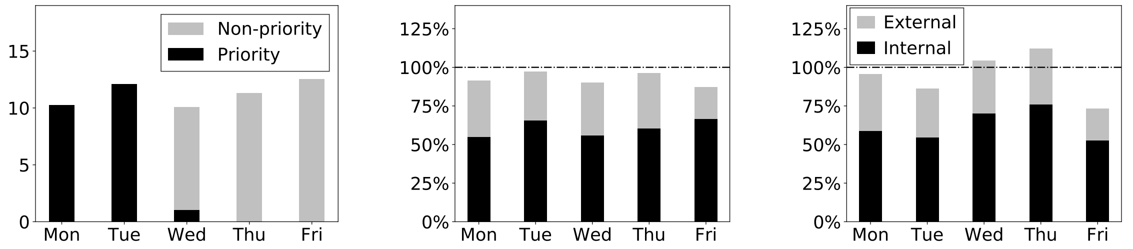
E.2 Additional Results from the Full 26-Station Network

Figure E.2a plots the optimal template for the 26-resource setting, which is very similar to optimal template for the 5-resource setting. Figure E.2b and E.2c plot the utilization of BDC for the 26-resource setting under the optimal and the historical templates respectively. The optimal template achieves a blocking probability less than 2% for all weekdays, while the blocking probability for BDC reaches 14% under the historical template.



(a) Using simulated $\beta_{u,t}$ (b) Using (D.1) for $\beta_{u,t}$ (c) Using excess probability

Figure E.1: Comparison of Simulated and Approximated ICT Distributions.



(a) Optimal template for the 26-station setting.

(b) Utilization of BDC under the optimal template

(c) Utilization of BDC under the historical template

Figure E.2: Optimal Templates and the BDC Utilization under the Optimal Template and the Historical Template for the 26-Station Setting.

Average ICT

Table E.1 shows the overall ICT and the average time to complete each stage for national patients admitted on Monday. Similar to the 5-station setting, ICTs are over a day longer under the front-loaded template, with much of the delays occurring in the first stage.

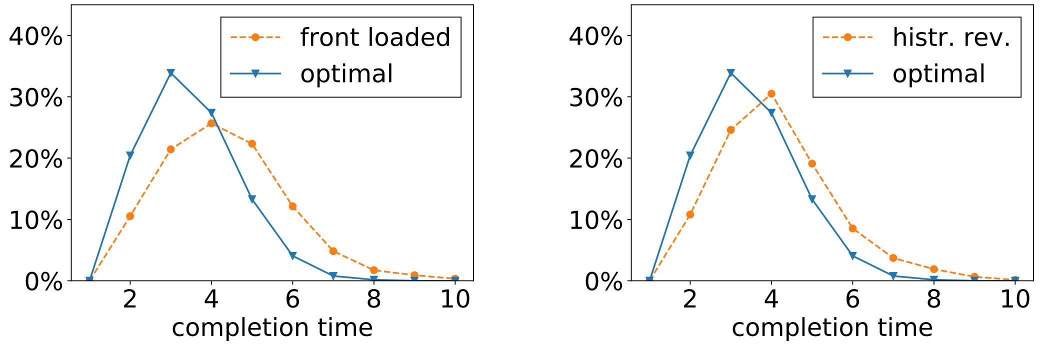
	Average ICT	stage 1	stage 2	stage 3	stage 4
	Optimal				
require 3 stages	3.06	1.02	1.03	1.02	
require 4 stages	4.13	1.04	1.04	1.03	1.02
	Front-loaded				
require 3 stages	4.28	2.14	1.06	1.07	
require 4 stages	5.35	2.15	1.06	1.08	1.07

Table E.1: Average ICT by Care Path Stage for National Patients Admitted on Monday for the 26-Station Setting

Comparison with the Front-Loaded and the Historical-Revised Templates

Under the front-loaded template, the blocking probability for BDC is above 45% from Monday to Wednesday. The historical-revised template has a comparatively lower blocking probability for BDC on Tuesday (35%). However, it reaches 70% on Friday and the patients who do not get appointments on Friday have to re-try on the next Monday, which in turn causes high blocking probability for BDC on Monday (50%).

Figure E.3 plots the ICT distributions for international and national patients admitted on Monday under the optimal, front-loaded and historical-revised templates. The curves are similar to the ones in the 5-resource setting we present in Section 3.4.3 of the main paper.



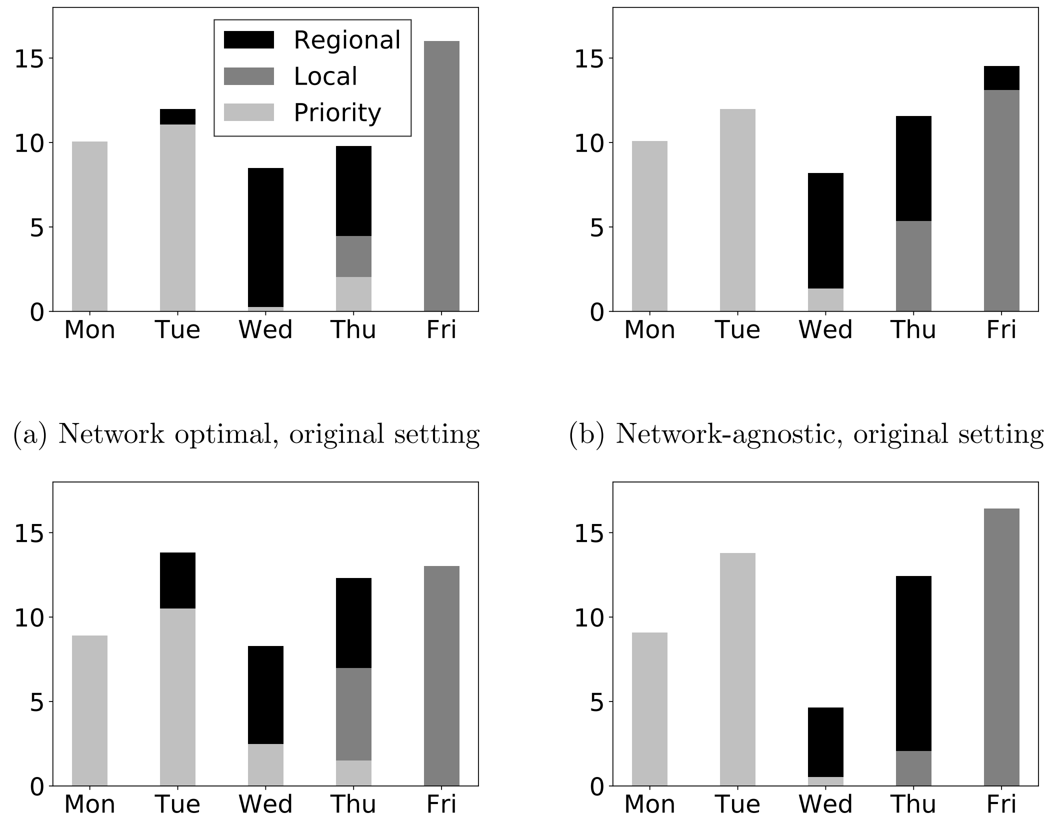
(a) Optimal v.s. the front-loaded templates

(b) Optimal v.s. the historical revised templates

Figure E.3: ICT Distribution Comparison for 26-Station Setting

E.3 Network Optimal and Network-Agnostic Templates

Figure E.4 shows the templates solved from the network-agnostic and full-network optimization problems in Section 3.4.3. Note the network agnostic template is different between the baseline scenario and the scenario with increased workload. This is because some of the constraints needed to be changed for feasibility in the increased workload scenario.



(a) Network optimal, original setting

(b) Network-agnostic, original setting

(c) Network optimal, expanded general surgery volume

(d) Network-agnostic, expanded general surgery volume

Figure E.4: Network Optimal and Network-Agnostic Templates with an Additional Constraint of 75% Completion Rate for Regional Patients.

APPENDIX F

DETAILED PROOFS FOR MEAN-FILED ANALYSIS

F.1 Proof of Theorem 4.1.1

For illustration purpose, we first detail the proof for the single-station in Section F.1.1, which serves as a building block. We then detail the proof for the general multi-station network in Sections F.1.2 and F.1.3.

F.1.1 Proof for the Single-Station Case

Stochastic System and Its Mean-Field Approximation

We start from the proof for the single-station setting with one type of patient. Let $M^B(t)$ and $M^{NB}(t)$ denote the number of blocked and non-blocked patients at the end of day t. Here, $M^{NB}(t)$ includes all patients who have finished their appointment in their current stage; each unblocked patient will stay in the system with a probability p and leave the system with probability $(1-p)$; this probability does not depend on which stage this patient is currently in.

As a result, the total number of requests from the target patients that will show up at the beginning of day $t+1$, $N(t+1)$, follows:

$$N(t+1) = M^B(t) + Bin(M^{NB}(t), p) + \lambda(t+1),$$

where $Bin(M^{NB}(t), p)$ denotes a binomial r.v. with parameters $M_{NB}(t)$ and p, corresponding to the patients who will stay, and $\lambda(t+1)$ is the number of new target patients that will show up on day $t+1$. The total number of patients, including the exogenous patients, equals $N(t+1)+\lambda^e(t+1)$. For analytical tractability, we assume that both $\lambda(\cdot)$ and $\lambda^e(\cdot)$ are deterministic. Note that, different from the notations used in the main paper, we *exclude* the exogenous patients from $N(\cdot)$, which only includes the requests from target patients.

Then, we can calculate the number of blocked and non-blocked patients at the end of day $t+1$ as:

$$M^B(t+1) = hGeo\Big(N(t+1) + \lambda^e(t+1), N(t+1), b\Big); \quad \text{(F.1)}$$

$$M^{NB}(t+1) = N(t+1) - M^B(t+1). \quad \text{(F.2)}$$

Here, $hGeo$ denotes the hyper-geometric distribution and

$$b = \big(\lambda^e(t+1) + n - C(t+1)\big)^+,$$

where $x^+ = \max(x, 0)$ for any real number x. That is, among the total number of requests $N(t+1) = n$ plus the external arrivals $\lambda^e(t+1)$, there will be b number of patients who cannot be accommodated since the capacity is $C(t+1)$. To see why it follows the hyper-geometric distribution, choosing the b patients who cannot be accommodated is equivalent to choosing b patients *without replacement* from two pools: the pool of $N(t+1) = n$ target patients, and the pool of $\lambda^e(t+1)$ external patients. As a result, the conditional probability for $M^B(t+1)$ follows

$$\mathbb{P}(M^B(t+1) = k | N(t+1) = n, b) = \frac{c(n,k)\cdot c(\lambda^e(t+1), b-k)}{c(n+\lambda^e(t+1), b)}$$

which is the hyper-geometric distribution, where $c(a, b)$ denotes the binomial coefficient. Note that, when $b = 0$, this distribution is still well-defined with a point mass on

$$\mathbb{P}(M^B(t+1) = 0 | N(t+1) = n, b = 0) = 1.$$

We define

$$U(t) = M^B(t)/N(t) \in [0, 1]$$

to be the proportion of blocked patients at the end of day t. Note that it is sufficient to track $\big(U(t), N(t)\big)$ as the state for the stochastic system, since we can recover $M^B(t)$ and $M^{NB}(t)$ from $U(t)$ and $N(t)$. Furthermore, for a given scaling constant $N \geq 1$ and a constant $q \geq 0$, it is also equivalent to track $\big(U(t), \frac{N(t)}{N^q}\big)$ as N^q remains as a constant that is independent of t; we will see why we introduce this scaling constant later in the proof. We denote

$$V(t) = \frac{N(t)}{N^q} \in \mathcal{V},$$

where $\mathcal{V}$ is the range for $V(t)$ and it depends on N, q. We assume the following the scaling for the arrival rates and the capacities.

Assumption F.1.1. *For any given scaling factor $N \geq 1$, we assume that for each $t \geq 1$, $\lambda(t) = r(t)N$, $\lambda^e(t) = r_e(t)N$, $C(t) = r_c(t)N$, where $r(t), r_e(t), r_c(t) > 0$ are constants that do not depend on N for each $t \geq 1$. Further, the initial state $N(0) = c_0 N$ where c_0 does not depend on N.*

We end this subsection with two remarks. First, $V(t)$ is an *auxiliary variable* to help track the state of the system, in particular, make sure that the stochastic system remains a Markov chain. However, $U(t)$ is the variable that we eventually care about, not $V(t)$. This is why the values of the testing functions h we use in the main theorems and the following lemmas only depend on $U(t)$, not on $V(t)$. Examples of such functions include $h(u, v) = u$ and $h(u, v) = (u - \bar{u})^2$ with $\bar{u}$ being some constant. We still keep the two arguments for $h(\cdot, \cdot)$ for the purpose of the proof.

Second, for the domain of $U(t)$ and $V(t)$, later we will show that $N(t)$ is in the order of N, the same order as the arrival rates $\lambda(\cdot), \lambda^e(\cdot)$ and the capacities $C(\cdot)$. Thus, $(M^B(t), M^{NB}(t))$ are in the same order as $N(t)$, so $U(t)$ is well-defined on $[0, 1]$. When $q > 1$, however, $V(t)$ will shrink to 0 as $N \to \infty$, so does $\mathcal{V}$. This shrinking domain does not affect our results, though, since the value of $h(\cdot, \cdot)$ does not depend on $V(t)$.

Deterministic System

A deterministic approximation for the system dynamics are as follows:

$$\begin{aligned} n(t+1) &= m^B(t) + p \cdot m^{NB}(t) + \lambda(t+1); && \text{(F.3)} \\ m^B(t+1) &= n(t+1) \cdot \beta_{t+1}; && \text{(F.4)} \\ m^{NB}(t+1) &= n(t+1) - m^B(t+1). && \text{(F.5)} \end{aligned}$$

Here, $n(\cdot)$, $m^B(\cdot)$, and $m^{NB}(\cdot)$ are the deterministic counterparts for $N(\cdot)$, $M^B(\cdot)$, and $M^{NB}(\cdot)$, respectively, and

$$\beta_t = \frac{\big(n(t) + \lambda_e(t) - C(t)\big)^+}{n(t) + \lambda_e(t)} = \max\left(1 - \frac{C(t)}{n(t) + \lambda^e(t)}, 0\right).$$

is the blocking probability on day t. We further define

$$\mu(t) = m^B(t)/n(t) = \beta_t$$

be the proportion of blocking patients. Similarly to the stochastic system, it is sufficient to track $(\mu(t), n(t))$, or equivalently, $(\mu(t), v(t))$, where $v(t) = \frac{n(t)}{N^q}$ is the counterpart for $V(t)$ defined above, with $N \geq 1$ and $q \geq 0$ being the same scaling constants we introduced previously. We define

$$\big(\mu(t+1), v(t+1)\big) = \Psi_1\left(\mu(t), v(t)\right)$$

to be the one-step transition from $\big(\mu(t), v(t)\big)$ to $\big(\mu(t+1), v(t+1)\big)$.

As with $V(t)$, $v(t)$ is the *auxiliary variable* to help track the state of the system, but does not affect the value of the testing functions, $h(\cdot,\cdot)$. We specify $\Psi_1 = (\Psi_1^u, \Psi_1^v)$ as the following:

$$\begin{aligned} v(t+1) &= \Psi_1^v(u,v) = v - (1-p)(1-u)v + \frac{\lambda(t)}{N^q}, \\ u(t+1) &= \Psi_1^u(u,v) = \max\left(1 - \frac{C(t+1)}{v(t+1)N^q + \lambda^e(t+1)}, 0\right) \\ &= \max\left(1 - \frac{C/N^q}{v - (1-p)(1-u)v + \frac{\lambda(t)+\lambda^e}{N^q}}, 0\right). \end{aligned}$$

Note that $\Psi_1^v(u,v)$ is a smooth function with continuous first and second derivatives, while $\Psi_1^u(u,v)$ is a piecewise smooth function with the non-smooth point at $v(t+1)N^q = n(t+1) = C(t+1) - \lambda^e(t+1)$; however, on intervals that do not contain this non-smooth point, both the first and second derivatives are continuous.

Transient Analysis

To prove the main theorem, we need the following lemmas first.

Lemma F.1.1. *For any given $(U(0), V(0)) = (u, v)$ with the given constants N and q, we have that*

$$\mathbb{E}_{u,v}\left[\big(U(1), V(1)\big) - \Psi_1(u,v)\right] = 0,$$

where $\mathbb{E}_{u,v}$ denotes the conditional expectation conditioning on $(U(0), V(0)) = (u, v)$.

Proof. We first show that $\mathbb{E}_{u,v}\left[V(1) - \Psi_1^v(u,v)\right] = 0$. Note that

$$N(1) = N(0)U(t) + Bin\big(N(0)(1 - U(0)), p\big) + \lambda(1),$$

and hence the conditional expected number of patients for $t = 1$ can be calculated as the following,

$$\mathbb{E}_{u,v}[N(1)] = vN^q \cdot u + vN^q \cdot (1-u)p + \lambda(1) = vN^q - vN^q \cdot (1-u)(1-p) + \lambda(1).$$

Then, we have

$$\mathbb{E}_{u,v}[V(1)] = v - v \cdot (1-u)(1-p) + \lambda(1)/N^q = \Psi_1^v(u,v).$$

Next, we show that $\mathbb{E}_{u,v}\left[U(1) - \Psi_1^u(u,v)\right] = 0$. Conditioning on $N(1) = n$ and using (F.1) and (F.2) for the relationship between $M^B(1)$, $M^{MB}(1)$ and $N(1)$, we have that

$$\begin{aligned} \mathbb{E}_{u,v}\left[U(1) - \Psi_1^u(u,v)\right] &= \mathbb{E}_{u,v}\left[\mathbb{E}\Big[M^B(1)/n - \Psi_1^u(u,v)\Big| N(1) = n\Big]\right] \\ &= 0. \end{aligned}$$

To get the second equality, we use the fact that the mean of the hyper-geometric distribution, conditioning on $N(1) = n$, equals $n\beta_1$. □

In the following lemmas as well as in the main theorem, we will consider testing functions that satisfy the following properties:

Assumption F.1.2. *The testing function $h : [0,1] \times \mathcal{V} \to \mathbb{R}$ be any continuous and twice differentiable function, where the first derivative of h is $(1/\gamma)$-Lipschitz, i.e.,*

$$|h'(a) - h'(b)| \le \frac{1}{\gamma}||a - b||.$$

Further, the testing function only depends on $U(\cdot)$, not the auxiliary variable $V(\cdot)$.

The next lemma considers a one-step transition from the initial state.

Lemma F.1.2. *Consider a function $h : [0,1] \times \mathcal{V} \to \mathbb{R}$ that satisfies Assumption F.1.2. Under the scaling given in Assumption F.1.1 and given $(U(0), V(0)) = (u,v)$ where $v = c_0 N^{1-q}$ and $q \ge 3/2$, we have that*

$$\Big|\mathbb{E}_{u,v}\left[h\left(U(1), V(1)\right)\right] - h\big(\Psi_1(u,v)\big)\Big| \le \frac{c_1}{N},$$

where $c_1 = c_1(r(1), r_e(1), r_c(1), c_0) > 0$ is a constant that depends on $r(1), r_e(1), r_c(1), c_0$ but is independent of u, N, q.

Proof. We perform Taylor expansion of h for $(U(1), V(1))$ in the neighborhood of $\Psi_1(u,v)$:

$$h\left(U(1), V(1)\right) - h\big(\Psi_1(u,v)\big) = h'\big(\Psi_1(u,v)\big) \cdot \mathcal{E} + o\left(\mathcal{E}\right),$$

where $\mathcal{E} = \big(U(1) - \Psi_1^u(u,v), V(1) - \Psi_1^v(u,v)\big)$. We get

$$\mathbb{E}_{u,v}\left[h\left(U(1), V(1)\right)\right] - h\big(\Psi_1(u,v)\big) \;=\; h'\big(\Psi_1(u,v)\big) \cdot \mathbb{E}_{u,v}[\mathcal{E}] + \mathbb{E}_{u,v}[o\left(\mathcal{E}\right)].$$

From Lemma F.1.1, we have $\mathbb{E}_{u,v}[\mathcal{E}] = 0$. Thus, it is sufficient to finish the proof by bounding $\mathbb{E}_{u,v}[o\left(\|\mathcal{E}\|\right)]$. For notational simplicity, we remove the time-index t and just denote $\lambda(1) = \lambda$, $\lambda_e(1) = \lambda_e$, and $C(1) = C$. We have

$$\begin{aligned}
\mathbb{E}_{u,v}\left[o(\|\mathcal{E}\|)\right] &\leq \frac{1}{2\gamma}\mathbb{E}_{u,v}\left[\|\mathcal{E}\|_2^2\right] \\
&= \frac{1}{2\gamma}\mathbb{E}_{u,v}\left[\big(U(1) - \Psi_1^u(u,v)\big)^2 + \big(V(1) - \Psi_1^v(u,v)\big)^2\right].
\end{aligned}$$

Here, the first inequality follows from the Taylor's theorem and Shalve-Shwartz and Zhang (2013). Then,

$$\begin{aligned}
& \mathbb{E}_{u,v}\left[\big(U(1) - \Psi_1^u(u,v)\big)^2\right] \\
= \ & \mathbb{E}_{u,v}\left[\mathbb{E}\big[\,\big(M^B(1)/n - \beta_1\big)^2\,|N(1) = n\big]\right] \\
= \ & \mathbb{E}_{u,v}\Big[\mathbb{E}\big[\mathrm{Var}\big(M^B(1)/n\big)|N(1) = n\big]\Big] \\
= \ & \mathbb{E}_{u,v}\left[\mathbb{E}\Big[\frac{1}{n^2}(\lambda_e + n - C)\frac{n}{\lambda_e + n}\frac{\lambda_e}{\lambda_e + n}\frac{C}{\lambda_e + n - 1}|N(1) = n, n > C - \lambda_e\Big]\right] \\
\leq \ & \mathbb{E}_{u,v}\left[\mathbb{E}\big[C/n^2\ |N(1) = n, n > C - \lambda_e\big]\right] \\
\leq \ & \frac{C}{\lambda^2},
\end{aligned}$$

where the third line uses the fact that the expectation of $M^B(1)/N(1)$ is simply β_1 from Lemma F.1.1, the fifth line uses the fact that $n, \lambda_e \leq (n + \lambda_e)$ and $\lambda_e + n - C \leq \lambda_e + n - 1$, and the last line uses the fact that $N(1) \geq \lambda$ for all $N(1) = n$. We have that $\frac{C}{\lambda^2} = \frac{r_c(1)}{r^2(1)N}$ is in the order of $1/N$. Thus, there exists a constant $\tilde{c}_1 > 0$ independent of u, v, N, q such that $\mathbb{E}_{u,v}\left[\big(U(1) - \Psi_1^u(u,v)\big)^2\right] \leq \tilde{c}_1/N$.

Next,

$$\begin{aligned}
& \mathbb{E}_{u,v}\left[\big(V(1) - \Psi_1^v(u,v)\big)^2\right] \\
= \ & \mathbb{E}_{u,v}\Big[\mathrm{Var}\big(V(1)\big)\Big] \leq \frac{1}{4}v^2 = \frac{1}{4}c_0^2 \cdot N^{2-2q}.
\end{aligned}$$

Here, to get the last inequality, we use the fact that

$$\lambda \leq N(1) \leq N(0) + \lambda, \text{ or } \lambda/N^q \leq V(1) \leq v + \lambda/N^q,$$

and the Popoviciu's inequality on variance for bounded random variables, i.e.,

$$\mathrm{Var}\big(V(1)\big) \leq \frac{1}{4}(v + \lambda/N^q - \lambda/N^q)^2.$$

Clearly, when $q \geq 3/2$, $\mathbb{E}_{u,v}\left[\big(V(1) - \Psi_1^v(u,v)\big)^2\right] \leq \frac{1}{4}\cdot\frac{c_0^2}{N^{2q-2}} \leq \frac{1}{4}\cdot\frac{c_0^2}{N}$. Thus, setting $c_1 = \frac{1}{2\gamma}(\tilde{c}_1 + \frac{1}{4}\cdot\frac{c_0^2}{N})$, we complete the proof. □

The following lemma is an extension of Lemma F.1.2 for the case of a general value of t.

Lemma F.1.3. *Consider a function $h : [0,1] \times \mathcal{V} \to \mathbb{R}$ that satisfies Assumption F.1.2. Under the scaling given in Assumption F.1.1 and given $(U(t-1), V(t-1)) = (u,v)$ and $q \geq 3/2$, we have that*

$$\Big|\mathbb{E}_{u,v}\left[h\left(U(t), V(t)\right)\right] - h\big(\Psi_1(u,v)\big)\Big| \leq \frac{\tilde{c}_t}{N},$$

where $\tilde{c}_t = \tilde{c}_t(r(1), \ldots, r(t), r_e(t), r_c(t), c_0) > 0$ is a constant that depends on $r_e(1)$, $r_c(1)$, c_0 and all $r(t)$'s from period 1 up to the current period t, but $\tilde{c}_t$ is independent of u, v, N, q.

The proof for Lemma F.1.3 is similar to that for Lemma F.1.2, except that when we bound $\mathbb{E}_{u,v}\left[\big(V(t) - \Psi_1^v(u,v)\big)^2\right]$, we use the fact that

$$\lambda(t) \leq N(t) \leq N(0) + \sum_{s=1}^{t} \lambda(s), \text{ or } r(t)N^{1-q} \leq V(t) \leq c_0 N^{1-q} + \sum_{s=1}^{t} r(s) N^{1-q},$$

and thus,

$$\begin{aligned} & \mathbb{E}_{u,v}\left[\big(V(t) - \Psi_1^v(u,v)\big)^2\right] \\ = \; & \mathbb{E}_{u,v}\Big[\mathrm{Var}\big(V(t)\big)\Big] \leq \frac{1}{4}\left(c_0 + \sum_{s=1}^{t-1} r(s)\right)^2 \cdot N^{2-2q}, \end{aligned}$$

which is of order $1/N$ when $q \geq 3/2$.

Next, we are ready to prove the first main theorem. We restate the theorem first and then show the proof.

Theorem 1. *Consider a function $h : [0,1] \times \mathcal{V} \to \mathbb{R}$ that satisfies Assumption F.1.2. Under the scaling given in Assumption F.1.1 and assume the following initial condition: $N(0) = n(0) = c_0 N$ where c_0 does not depend on N; and $U(0) = \mu(0)$. Then, for any fixed $t \geq 0$, if $q \geq 3/2$, we have that*

$$\Big|\mathbb{E}\left[h\left(U(t), V(t)\right)\right] - h(\mu(t), v(t))\Big| \leq \frac{c_t}{N}, \tag{F.6}$$

where $c_t = c_t(r(1), \ldots, r(t), r_e(t), r_c(t), c_0) > 0$ and is a constant that depends on $r_e(1), r_c(1), c_0$ and all $r(t)$'s from period 1 up to the current period t, but c_t is independent of N, q.

Proof. We prove by induction. The theorem holds for the case $t = 0$ by assumption. Assume that the theorem holds for some $t \geq 0$, we have for $t+1$ that

$$\begin{aligned} & \Big|\mathbb{E}[h\left(U(t+1), V(t+1)\right)] - h\left(\mu(t+1), v(t+1)\right)\Big| \\ \leq \; & \Big|\mathbb{E}\left[h\left(U(t+1), V(t+1)\right) - h\left(\Psi_1\big(U(t), V(t)\big)\right)\right]\Big| \qquad \text{(F.7)} \\ + \; & \Big|\mathbb{E}\Big[h\left(\Psi_1\big(U(t), V(t)\big)\right) - h\left(\mu(t+1), v(t+1)\right)\Big]\Big|. \qquad \text{(F.8)} \end{aligned}$$

For the term (F.7), by utilizing the definition of total expectation, we can establish the following bound:

$$\begin{aligned}
& \Big| \mathbb{E}\left[h\left(U(t+1), V(t+1)\right) - h\left(\Psi_1\big(U(t), V(t)\big)\right)\right] \Big| \\
\leq\ & \mathbb{E}\left[\Big| \mathbb{E}_{u,v}[h\left(U(t+1), V(t+1)\right)] - h\left(\Psi_1(u,v)\right) \Big| \,\Big|\, U(t) = u, V(t) = v \right] \\
\leq\ & \tilde{c}_t / N,
\end{aligned}$$

using Lemma F.1.3 and the fact that $\tilde{c}_t$ does not depend on $U(t), V(t), N, q$.

For (F.8), recall that $(\mu(t+1), v(t+1)) = \Psi_1\big(u(t), v(t)\big)$ in the deterministic system. We show in Lemma F.1.4 below that $h \circ \Psi_1$ is twice-differentiable and the first derivative of $h \circ \Psi_1$ is $\frac{1}{\gamma}$-Lipschitz. Thus, we can apply the induction hypothesis to the function $h \circ \Psi_1$ and get

$$\begin{aligned}
& \mathbb{E}\left[h\left(\Psi_1\big(U(t), V(t)\big)\right) - h\left(\mu(t+1), v(t+1)\right)\right] \\
=\ & \mathbb{E}\big[h \circ \Psi_1\big(U(t), V(t)\big) - h \circ \Psi_1\big(u(t), v(t)\big)\big] \\
\leq\ & c_t / N,
\end{aligned}$$

where c_t here depends on $c_{h \circ \Psi_1}$ and on $r_e(1), r_c(1), c_0$ and all $r(t)$'s from period 1 up to the current period t, but it is independent of N, q. Setting $c_{t+1} = \tilde{c}_t + c_t$, we complete the proof. □

Lemma F.1.4. *Let $g : [0,1] \times \mathcal{V} \to \mathbb{R}$ be any continuous and twice differentiable function, where the second derivative of g is bounded by a constant $c_h > 0$. Given a scaling factor $N \geq 1$ and $q \geq 3/2$, $g \circ \Psi_1$ is twice-differentiable and the first derivative of $g \circ \Psi_1$ is $\frac{1}{\gamma}$-Lipschitz, where this Lipschitz constant is independent of N, q.*

Proof. Given $(\mu(t), v(t)) = (u, v)$, we first specify Ψ when $vN^q + \lambda^e(t+1) > C(t+1)$. We can write $\Psi_1 = (\Psi_1^u, \Psi_1^v)$ as the following:

$$\begin{aligned}
v(t+1) &= \Psi_1^v(u, v) = v - (1-p)(1-u)v + \frac{\lambda(t+1)}{N^q} && \text{(F.9)} \\
u(t+1) &= \Psi_1^u(u, v) = \beta_{t+1} = 1 - \frac{C(t+1)}{v(t+1)N^q + \lambda^e(t+1)} && \text{(F.10)} \\
&= 1 - \frac{C/N^q}{v - (1-p)(1-u)v + \frac{\lambda(t+1) + \lambda^e(t+1)}{N^q}}. && \text{(F.11)}
\end{aligned}$$

We show at the end of this proof that their first and second derivatives are continuous.

When $vN^q + \lambda^e(t+1) \leq C(t+1)$. We can write $\Psi_1 = (\Psi_1^u, \Psi_1^v)$ as the following:

$$\begin{aligned}
v(t+1) &= \Psi_1^v(u, v) = v - (1-p)(1-u)v + \frac{\lambda(t+1)}{N^q} && \text{(F.12)} \\
u(t+1) &= \Psi_1^u(u, v) = 0. && \text{(F.13)}
\end{aligned}$$

Clearly, their first and second derivatives are also continuous.

Since the first derivative of Ψ_1 is piecewise continuous, it is sufficient to show that (u, v) come from a bounded set (which also implies that the second derivatives are bounded). Clearly, $U(t) \in [0, 1]$ which is a bounded set. For $V(t)$ and a fixed $t \geq 1$, we know that

$$r(t)N^{1-q} \leq V(t) \leq c_0 N^{1-q} + \sum_{s=1}^{t} r(s)N^{1-q}.$$

When $q \geq 3/2$, $\lim_{N\to\infty} N^{1-q} = 0$. Thus, there exists a constant c_2 such that $N^{1-q} \leq c_2$, and c_2 is independent of N. As a result, $V(t)$ is also on a bounded set, i.e., $\mathcal{V}$ is bounded.

To see why bounded set is sufficient, let x_0 denote the non-smooth point for a function $f(x)$, whose first derivative $|f'(x)| \leq L$. Then,

$$\begin{aligned} |f(a) - f(b)| &\leq |f(a) - f(x_0)| + |f(x_0) - f(b)| \\ &= \left| \int_b^{x_0} f'(x)dx \right| + \left| \int_{x_0}^{a} f'(x)dx \right| \\ &\leq \int_b^{x_0} |f'(x)|dx + \int_{x_0}^{a} |f'(x)|dx \\ &\leq L(a - b). \end{aligned}$$

Finally, we verify the first and second derivatives of $\Psi_1^u(u, v)$ and $\Psi_1^v(u, v)$ are continuous when $vN^q + \lambda^e(t + 1) > C(t + 1)$.

$$\begin{aligned} \frac{\partial \Psi_1^v}{\partial u} &= (1 - p)v, \\ \frac{\partial \Psi_1^v}{\partial v} &= p - pu + u, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \Psi_1^u}{\partial u} &= \frac{v(1-p)C/N^q}{\big(vu + vp(1-u) + (\lambda + \lambda_e)/N^q\big)^2} \\ &\leq \frac{v(1-p)C}{(vu + vp(1-u))^2 N^q}, \text{ which is bounded because } C = O(N), q \geq 3/2; \\ \frac{\partial \Psi_1^u}{\partial v} &= \frac{(u(1-p)+p)C/N^q}{\big(vu + vp(1-u) + (\lambda + \lambda_e)/N^q\big)^2} \\ &\leq \frac{(u(1-p)+p)C}{(vu + vp(1-u))^2 N^q}, \text{ which is bounded because } C = O(N), q \geq 3/2. \end{aligned}$$

The second derivatives of $\Psi_1^u(u, v)$ and $\Psi_1^v(u, v)$ can be verified similarly. This concludes the proof.

□

F.1.2 Transient Proofs for the Network Setting

Stochastic and Deterministic Systems

For the network setting, we consider a set of stations: $u \in \mathcal{U} = \{1, \ldots, n\}$. When we refer to a station u, we use u_i and station i interchangeably. For ease of exposition, we start by consider a single type of patients ($K = 1$) and the routing probabilities is the same for each stage. We denote $p_{i,j}$ as the routing probability from station i to station j after completing the current appointment at station i. We allow a non-zero probability of $(1 - \sum_{u \in \mathcal{U}} p_{i,u})$ to directly leave the system. We discuss how to extend the proof to multiple types of patients and to stage-dependent probabilities later in Section F.1.3.

Same as in the single-station setting, we denote $M_u^B(t)$ and $M_u^{NB}(t)$ as the number of blocked and non-blocked patients at the end of day t for station u. We denote the total number of requests from the target patients that will show up at the beginning of day $t+1$ for station u as $M_u(t+1)$. We have

$$M_u(t+1) = M_u^B(t) + \sum_{\tilde{u}} Mult(M_{\tilde{u}}^{NB}(t), p_{\tilde{u},u}) + \lambda_u(t+1),$$

where $Mult(M_{\tilde{u}}^{NB}(t), p_{\tilde{u},u})$ denotes a multinomial r.v., corresponding to non-blocked patients who will need to visit station u on day $t+1$, and $\lambda_u(t+1)$ is the number of target patients that will show up on day $t+1$. (According to the setting in our main paper, $\lambda_u(t+1) = 0$ for non-BDC station, but we keep this in the proof for completeness.) The total number of patients, including the exogenous patients, equals $M_u(t+1) + \lambda_u^e(t+1)$. Again, for analytical tractability, we assume that both $\lambda_u(\cdot)$'s and $\lambda_u^e(\cdot)$'s are deterministic. Here, to not confuse with $N_u(\cdot)$ used in the main paper, we use $M_u(\cdot)$ to denote the total target patients' request and *exclude* the exogenous patients from $M_u(\cdot)$; in the main paper, $N_u(\cdot)$ also includes the requests from exogenous patients.

The dynamics from $M_u(t+1)$ to $M_u^B(t+1)$ and $M_u^{NB}(t+1)$ are the same as in the single-station setting. We define

$$U_u(t) = M_u^B(t)/M_u(t) \in [0,1], \quad u \in \mathcal{U}$$

be the proportion of blocked patients at the end of day t for station u. Note that it is sufficient to track $\big(U_u(t), M_u(t), \forall u \in \mathcal{U}\big)$ as the state for the stochastic system, since we can recover $M_u^B(t)$ and $M_u^{NB}(t)$ from $U_u(t)$ and $M_u(t)$ for each station u. Furthermore, for a given scaling constant $N \geq 1$ and a constant $q \geq 0$, we introduce the auxiliary variables

$$V_u(t) = \frac{M_u(t)}{N^q} \in \mathcal{V}, \quad u \in \mathcal{U},$$

where $\mathcal{V}$ is the range for $V(t)$ and it depends on N, q. We assume the following the scaling for the arrival rates and the capacities.

Assumption F.1.3. *For any given scaling factor $N \geq 1$, we assume that for each $t \geq 1$, $\lambda_u(t) = r_u(t)N$, $\lambda_u^e(t) = r_{u,e}(t)N$, $C_u(t) = r_{u,c}(t)N$, where $r_u(t), r_{u,e}(t), r_{u,c}(t) > 0$ are constants that do not depend on N for each $t \geq 1$ and for each $u \in \mathcal{U}$. Further, for each u, the initial state $M_u(0) = c_{u,0}N$, where $c_{u,0}$does not depend on N.*

A deterministic approximation for the system dynamics can be constructed as follows:

$$\begin{aligned} m_u(t+1) &= m_u^B(t) + \sum_{\tilde{u}} p_{\tilde{u},u} \cdot m_u^{NB}(t) + \lambda_u(t+1); && \text{(F.14)} \\ m_u^B(t+1) &= m_u(t+1) \cdot \beta_{u,t+1}; && \text{(F.15)} \\ m_u^{NB}(t+1) &= m_u(t+1) - m_u^B(t+1). && \text{(F.16)} \end{aligned}$$

Here, $m_u(\cdot)$, $m_u^B(\cdot)$, and $m_u^{NB}(\cdot)$ are the deterministic counterparts for $M_u(\cdot)$, $M_u^B(\cdot)$, and $M_u^{NB}(\cdot)$, respectively, and

$$\beta_{u,t} = \frac{\big(m_u(t) + \lambda_u^e(t) - C_u(t)\big)^+}{m_u(t) + \lambda_u^e(t)} = \max\left(1 - \frac{C_u(t)}{m_u(t) + \lambda_u^e(t)}, 0\right).$$

is the blocking probability on day t. We further define

$$\mu_u(t) = m_u^B(t)/m_u(t) = \beta_{u,t}$$

be the proportion of blocking patients for station u. We track $(\mu_u(t), v_u(t))$, where $v_u(t) = \frac{m_u(t)}{N^q}$ is the counterpart for $V_u(t)$ defined above, with $N \geq 1$ and $q \geq 0$ being the same scaling constants we introduced. We define

$$\big(\mu_1(t+1), v_1(t+1), \ldots, \mu_n(t+1), v_n(t+1)\big) = \Psi_1\left(\mu_1(t), v_1(t), \ldots, \mu_n(t), v_n(t)\right)$$

be the one-step transition. We also define

$$\mu_1(t+1) = \Psi_1^{u,1}\left(\mu_1(t), v_1(t), \ldots, \mu_n(t), v_n(t)\right)$$

and

$$v_1(t+1) = \Psi_1^{v,1}\left(\mu_1(t), v_1(t), \ldots, \mu_n(t), v_n(t)\right)$$

for station $u = 1$; other stations can be define similarly.

Transient Analysis

To prove the main theorem in the multi-station setting, the key is to note that all stations function independently on how many are blocked or not today, conditioning on the total requests given at the beginning of the day. Thus, it is easy to extend the lemmas proved in the single-station setting to the multi-station setting here.

Lemma F.1.5. *For any given $(U_1(0), V_1(0), \ldots, U_n(0), V_n(0)) = (u_1, v_1, \ldots, u_n, v_n)$ with the given constants N and q, we have that*

$$\mathbb{E}_{u,v}\left[\big(U_1(1), V_1(1), \ldots, U_n(1), V_n(1)\big) - \Psi_1(u_1, v_1, \ldots, u_n, v_n)\right] = 0,$$

where $\mathbb{E}_{u,v}$ denotes the conditional expectation that is conditioning on the initial state $(u_1, v_1, \ldots, u_n, v_n)$.

Proof. We first show that $\mathbb{E}_{u,v}\left[V_1(1) - \Psi_1^{v,1}(u_1, v_1, \ldots, u_n, v_n)\right] = 0$ for station 1; the proof is the same for all other stations. Note that

$$M_1(1) = M_1(0)U_1(t) + \sum_{u \in \mathcal{U}} Mult\big(M_u(0)(1 - U_u(0)), p_{u,1}\big) + \lambda_1(1),$$

and hence,

$$\begin{aligned} \mathbb{E}_{\mu,v}[M_1(1)] &= v_1 N^q \cdot u_1 + \sum_{j=1}^{n} v_j N^q \cdot (1 - \mu_j) p_{j,1} + \lambda(1) \\ &= \Psi_1^{v,1}(u_1, v_1, \ldots, u_n, v_n). \end{aligned}$$

The proof for showing $\mathbb{E}_{u,v}\left[U_1(1) - \Psi_1^{u,1}(u_1, v_1, \ldots, u_n, v_n)\right] = 0$ is the same as that in Lemma F.1.1 for the single-station, by noting the mean of the hyper-geometric distribution, conditioning on $M_1(1) = m$, equals $m\beta_{1,1}$. □

In the following lemmas as well as in the main theorem, we will consider testing functions that satisfy the following properties:

Assumption F.1.4. *The testing function $h : [0,1]^n \times \mathcal{V}^n \to \mathbb{R}$ be any continuous and twice differentiable function, where the first derivative of h is $(1/\gamma)$-Lipschitz, i.e.,*

$$|h'(a) - h'(b)| \leq \frac{1}{\gamma} ||a - b||.$$

Further, the testing function only depends on $U_u(\cdot)$ for some station u, not on any of the auxiliary variable $V_u(\cdot)$'s.

The next lemma is the multi-station version of Lemma F.1.2.

Lemma F.1.6. *Consider a function $h : [0,1]^n \times \mathcal{V}^n \to \mathbb{R}$ that satisfies Assumption F.1.4. Under the scaling given in Assumption F.1.3 and given*

$$(U_1(0), V_1(0), \ldots, U_n(0), V_n(0)) = (u_1, v_1, \ldots, u_n, v_n)$$

where $v_u = c_{u,0} N^{1-q}$ for each $u \in \mathcal{U}$ and $q \geq 3/2$, we have that

$$\Big| \mathbb{E}_{u,v}\left[h\left(U_1(1), V_1(1), \ldots, U_n(1), V_n(1)\right)\right] - h\big(\Psi_1(u_1, v_1, \ldots, u_n, v_n)\big) \Big| \leq \frac{c_1}{N},$$

where $c_1 = c_1(r_u(1), r_{u,e}(1), r_{u,c}(1), c_{u,0}, \forall u) > 0$ is a constant that depends on all the $r_u(1), r_{u,e}(1), r_{u,c}(1), c_{u,0}$'s, but is independent of N, q and $U_1(0), \ldots, U_n(0)$.

Proof. Denote $(u, v) = (u_1, v_1, \ldots, u_n, v_n)$. We perform Taylor expansion of h in the neighborhood of $\Psi_1(u, v)$:

$$h\left(U_1(1), V_1(1), \ldots, U_n(1), V_n(1)\right) - h\big(\Psi_1(u, v)\big) = h'\big(\Psi_1(u, v)\big) \cdot \mathcal{E} + o\left(\mathcal{E}\right),$$

where $\mathcal{E} = \big(U_1(1) - \Psi_1^{u,1}(u, v), V_1(1) - \Psi_1^{v,1}(u, v), \ldots, U_n(1) - \Psi_1^{u,n}(u, v), V_n(1) - \Psi_1^{v,n}(u, v)\big)$.

Similar as in the single-station setting, using Lemma F.1.5, we have $\mathbb{E}_{u,v}[\mathcal{E}] = 0$. Thus, it is sufficient to finish the proof by bounding $\mathbb{E}_{u,v}[o(\|\mathcal{E}\|)]$. To do so, it is sufficient to bound $\mathbb{E}_{u,v}\left[\left(U_j(1) - \Psi_1^{u,j}(u,v)\right)^2\right]$ and to bound $\mathbb{E}_{u,v}\left[\left(V_j(1) - \Psi_1^{v,j}(u,v)\right)^2\right]$ for each station $j = 1, \ldots, n$. The latter can be bounded in the same way as in Lemma F.1.2 by noting the fact that

$$\lambda_j(1) \leq M_j(1) \leq \lambda_j(1) + \sum_{u \in \mathcal{U}} N_u(0), \text{ or } \lambda_j(1)/N^q \leq V_j(1) \leq \lambda/N^q + \sum_{u \in \mathcal{U}} v_u,$$

where $v_u = c_{u,0}N^{1-q}$ for each $u \in \mathcal{U}$.

The other term can be bounded as

$$\begin{aligned} & \mathbb{E}_{u,v}\left[\left(U_j(1) - \Psi_1^{u,j}(u,v)\right)^2\right] \\ = \ & \mathbb{E}_{u,v}\left[\mathbb{E}\left[\left(M_j^B(1)/n - \beta_{j,1}\right)^2 | M_j(1) = n\right]\right] \\ = \ & \mathbb{E}_{u,v}\left[\mathbb{E}\left[\mathrm{Var}\left(M_j^B(1)/n\right) | M_j(1) = n\right]\right] \\ \leq \ & \mathbb{E}_{u,v}\left[\mathbb{E}\left[C_j(1)/n^2 \ | M_j(1) = n, n > C_j(1) - \lambda_{j,e}(1)\right]\right] \\ \leq \ & \frac{C_j(1)}{\lambda_j^2(1)}, \end{aligned}$$

where $\frac{C_j(1)}{\lambda_j^2(1)} = \frac{r_{j,c}(1)}{r_j^2(1)N}$ is in the order of $1/N$. The rest of the proof can be proceeded in the same way as in Lemma F.1.2. □

The following lemma is the multi-station version of Lemma F.1.3.

Lemma F.1.7. *Consider a function $h : [0,1]^n \times \mathcal{V}^n \to \mathbb{R}$ that satisfies Assumption F.1.4. Under the scaling given in Assumption F.1.3 and given $(U_1(t-1), V_1(t-1), \ldots, U_n(t-1), V_n(t-1)) = (u_1, v_1, \ldots, u_n, v_n)$ and $q \geq 3/2$, we have that*

$$\left|\mathbb{E}_{u,v}\left[h\left(U_1(t), V_1(t), \ldots, U_n(t), V_n(t)\right)\right] - h\big(\Psi_1(u_1, v_1, \ldots, u_n, v_n)\big)\right| \leq \frac{\tilde{c}_t}{N},$$

where $\tilde{c}_t = \tilde{c}_t(r_u(1), \ldots, r_u(t), r_{u,e}(t), r_{u,c}(t), c_{u,0}, \forall u) > 0$ is a constant that depends on all the $r_{u,e}(t), r_{u,c}(t), c_{u,0}$'s and all $r_u(t)$'s from period 1 up to the current period t, but $\tilde{c}_t$ is independent of u, v, N, q.

Theorem 2. *Consider a function $h : [0,1]^n \times \mathcal{V}^n \to \mathbb{R}$ that satisfies Assumption F.1.4. Under the scaling given in Assumption F.1.3 and assume the following initial condition: $N_u(0) = n_u(0) = c_{u,0}N$ for each station u, where $c_{u,0}$ does not depend on N; and $U_u(0) = \mu_u(0)$ for each station u. Then, for any fixed $t \geq 0$, if $q \geq 3/2$, we have that*

$$\left|\mathbb{E}\left[h\left(U_1(t), V_1(t), \ldots, U_n(t), V_n(t)\right)\right] - h(\mu_1(t), v_1(t), \ldots, \mu_n(t), v_n(t))\right| \leq \frac{c_t}{N}, \quad \text{(F.17)}$$

where $c_t = c_t(r_u(1), \ldots, r_u(t), r_{u,e}(t), r_{u,c}(t), c_{u,0}, \forall u) > 0$ is a constant that depends on all the $r_{u,e}(t), r_{u,c}(t), c_{u,0}$'s and all $r_u(t)$'s from period 1 up to the current period t, but c_t is independent of N, q.

Proof. We prove by induction. The theorem holds for the case $t = 0$ by assumption. Let $(U(t+1), V(t+1)) = (U_1(t+1), V_1(t+1), \ldots, U_n(t+1), V_n(t+1))$ $(\mu(t), v(t)) = (\mu_1(t), v_1(t), \ldots, \mu_n(t), v_n(t))$. Same as in the single-station setting, assume that the theorem holds for some $t \geq 0$, we have for $t+1$ that

$$\begin{aligned} & \Big|\mathbb{E}[h\left(U(t+1), V(t+1)\right)] - h\left(\mu(t+1), v(t+1)\right)\Big| \\ \leq \; & \Big|\mathbb{E}\left[h\left(U(t+1), V(t+1)\right) - h\left(\Psi_1\big(U(t), V(t)\big)\right)\right]\Big| \\ + \; & \Big|\mathbb{E}\big[h\left(\Psi_1\big(U(t), V(t)\big)\right) - h\left(\mu(t+1), v(t+1)\right)\big]\Big|. \end{aligned}$$

For the first term on the RHS, we have

$$\begin{aligned} & \Big|\mathbb{E}\left[h\left(U(t+1), V(t+1)\right) - h\left(\Psi_1\big(U(t), V(t)\big)\right)\right]\Big| \\ \leq \; & \mathbb{E}\left[\Big|\mathbb{E}_{u,v}[h\left(U(t+1), V(t+1)\right)] - h\left(\Psi_1(u,v)\right)\Big| \,\Big|\, U(t) = u, V(t) = v\right] \\ \leq \; & \tilde{c}_t/N, \end{aligned}$$

using the multi-station version in Lemma F.1.7 and the fact that $\tilde{c}_t$ does not depend on $U(t), V(t), N, q$. To bound the second term on the RHS, we apply the induction hypothesis to $h \circ \Psi_1$ as we did in the single-station version.

□

F.1.3 Extensions to Multiple Patient Classes and Stage-Dependent Routing Probabilities

To extend to multiple classes and/or stage-dependent routing probabilities, we just need to incorporate more states for tracking the system dynamics. We illustrate how to do it for multiple classes; the extension to stage-dependent routing or the combination or both is similar.

Now consider the case we have K types of patients, and denote $p^k_{i,j}$ as the routing probability from station i to station j for type k patients, after completing the current appointment at station i. We allow a non-zero probability of $(1-\sum_{u\in\mathcal{U}} p^k_{i,u})$ to directly leave the system.

We denote $M^{B,k}_u(t)$ and $M^{NB,k}_u(t)$ as the number of blocked and non-blocked patients for each class k. For system dynamics, we need to track the total number of patients from each class that request an appointment from station u on day $t+1$, defined as

$$M^k_u(t+1) = M^{B,k}_u(t) + \sum_{\tilde{u}} Mult(M^{NB,k}_{\tilde{u}}(t), p^k_{\tilde{u},u}) + \lambda^k_u(t+1).$$

The total number of target patients' requests (across classes) that will show up at the beginning of day $t+1$ for station u, $M_u(t+1) = \sum_{k=1}^K M^k_u(t+1)$. Next, from $M_u(t+1)$ and $\lambda^e_u(t+1)$, we are able to calculate the blocking probabilities and the total number of patients blocked as in the single-class setting. Then, we get $M^{B,k}_u(t+1)$

and $M_u^{NB,k}(t+1)$ for day $t+1$ using the multinomial distribution with proportions $\{\frac{M_u^k(t+1)}{\sum_k M_u^k(t+1)}\}$.

For the convergence results, we define

$$U_u(t) = \frac{\sum_k M_u^{B,k}(t)}{\sum_k M_u^k(t)} \in [0,1], \quad u \in \mathcal{U}$$

as the proportion of blocked patients across K classes for station on day t. For the auxiliary variables, we now need to track for each class. We define

$$V_u^k(t) = \frac{M_u^k(t)}{N^q} \in \mathcal{V}, \quad u \in \mathcal{U},$$

where $\mathcal{V}$ is the range for $V_u^k(t)$ and it depends on N, q. The testing functions h will take arguments as

$$h\big(U_1(t), V_1^1(t), \ldots, V_1^K(t), \ldots, U_n(t), V_n^1(t), \ldots, V_n^K(t)\big).$$

For the deterministic system, we have

$$\begin{aligned} m_u^k(t+1) &= m_u^{B,k}(t) + \sum_{\tilde{u}} p_{\tilde{u},u}^k \cdot m_u^{NB,k}(t) + \lambda_u^k(t+1), \\ m_u^{B,k}(t+1) &= m_u^k(t+1) \cdot \beta_{u,t+1}, \quad m_u^{NB,k}(t+1) = m_u^k(t+1) - m_u^{B,k}(t+1), \end{aligned}$$

where

$$\beta_{u,t} = \frac{\big(m_u(t) + \lambda_u^e(t) - C_u(t)\big)^+}{m_u(t) + \lambda_u^e(t)}, \quad m_u(t) = \sum_{k=1}^{K} m_u^k(t).$$

We further define

$$\mu_u(t) = \frac{\sum_k m_u^{B,k}(t)}{\sum_k m_u^k(t)} = \beta_{u,t}, \quad v_u^k(t) = \frac{m_u^k(t)}{N^q}.$$

Then, the one-step transition generator is defined via

$$\begin{aligned} & \big(\mu_1(t+1), v_1^1(t+1), \ldots, v_1^K(t+1), \ldots, \mu_n(t+1), v_n^1(t+1), \ldots, v_n^K(t+1)\big) \\ = \; & \Psi_1\left(\mu_1(t), v_1^1(t), \ldots, v_1^K(t), \ldots, \mu_n(t), v_n^1(t), \ldots, v_n^K(t)\right). \end{aligned}$$

We also define $\Psi_1^{u,i}$ and $\Psi_1^{v,k,i}$ as the one-step transition to $\mu_i(t+1)$ and $v_i^k(t+1)$ for class k and station i.

Once we have the modified definitions on the states and transitions, it is straightforward to extend the proof for Lemmas F.1.5 through F.1.7 and the main theorem to the multi-class version. For example, consider the extension for Lemma F.1.5. Given $(\mu, v) = (\mu_1(0), v_1^1(0), \ldots, v_1^K(0), \ldots, \mu_n(0), v_n^1(0), \ldots, v_n^K(0))$, when showing

$$\mathbb{E}_{\mu,v}\left[V_1^k(1) - \Psi_1^{v,k,1}(\mu, v)\right] = 0$$

for station 1 (other station uses the same argument), we note that $\{M_u^{B,k}(0)\}$ follows a multinomial distribution with parameters $U_u(0) \cdot \sum_k M_u^k(0)$ and $\{\frac{M_u^k(0)}{\sum_k M_u^k(0)}\}$. This gives us $\{M_u^{NB,k}(0)\}$ correspondingly. Then,

$$M_1^k(1) = M_1^{B,k}(0) + \sum_{u \in \mathcal{U}} Multi\big(M_u^{NB,k}(0), p_{u,1}^k\big) + \lambda_1^k(1),$$

and hence,

$$\begin{aligned} \mathbb{E}_{\mu,v}[M_1^k(1)] &= v_1^k N^q \cdot \mu_1 + \sum_{j=1}^{n} v_j^k N^q \cdot (1-\mu_j) p_{j,1}^k + \lambda_1^k(1) \\ &= N^q \cdot \Psi_1^{v,k,1}(\mu, v). \end{aligned}$$

Other lemmas and the main theorem can be proceed in a similar way.

F.2 Proof of Theorem 4.2.1

In the steady-state proof, we focus on proving the *single-station* setting. The extension to the multi-station setting with potentially class- and stage-dependent routing probabilities is the same as we show in the transient analysis, i.e., incorporating proper states in tracking the system dynamics. Furthermore, we focus on the *time-stationary* setting. In the rest of this section, we first specify the necessary generators for the proof and give the roadmap for the proof under the Stein's method framework in Section F.2.1. The proof involves dealing with states in the bounded set and the unbounded set separately. We prove results for states belonging to the bounded set in Section F.2.2. Then, we prove the main theorem in Section F.2.3. The proofs for several lemmas that are used in the main proof are detailed in Section F.2.4.

F.2.1 Generators and Stein's Framework

We first define the following generators. In the deterministic system, let $\Psi_t = (\Psi_t^u, \Psi_t^v)$ $(t \geq 0)$ denote the t-step transition generator from the current state $(\mu(0), v(0))$ to $(\mu(t), v(t))$, with Ψ_0 being the self-mapping from $(\mu(0), v(0))$ to $(\mu(0), v(0)$ in the same period. In the original (stochastic) system, we use G_t to denote the t-step generator from $(U(0), V(0))$ to $(U(t), V(t))$, with G_0 being the self-mapping in the same period and G_1 being the one-step transition to $(U(1), V(1))$.

In the steady-state proof, we focus on the time-stationary setting. We also fix $q = 3/2$ for ease of exposition.

Assumption F.2.1. *For any given scaling factor $N \geq 1$, we assume that for each $t \geq 1$, $\lambda(t) = \lambda = rN$, $\lambda^e(t) = \lambda_e = r_e N$, $C(t) = C = r_c N$, where $r, r_e, r_c > 0$ are three constants that do not depend on N. Further, we fix $q = 3/2$.*

We denote (U_∞, V_∞) as the random variable following the unique stationary distribution of the stochastic system, and (μ_∞, ν_∞) as the equilibrium point of the deterministic system; these steady-state variables exist under the stability conditions we proved. Different from the transient analysis, the proof will involve bounding both the distance between U_∞ and μ_∞ and the distance between V_∞ and ν_∞. Specifically, we will consider one particular testing function

$$g(u,v) = (u-\mu_\infty)^2 + (v-\nu_\infty)^2.$$

We also denote $g_1(u,v) = (u-\mu_\infty)^2$ and $g_2(u,v) = (v-\nu_\infty)^2$. It is easy to verify that g_1 and g_2 satisfy Assumption F.1.2.

Stein's Method Framework

We first overview the Stein's method framework, which serves as the backbone of our proof. For a given state $(u,v) \in [0,1] \times \mathcal{V}$, the Poisson equation with respect to Ψ can be written as:

$$f_g(u,v) = g(u,v) - g\big(\mu_\infty, v_\infty\big) + f_g\big(\Psi_1(u,v)\big),$$

or equivalently

$$g(u,v) - g\big(\mu_\infty, v_\infty\big) = f_g(u,v) - f_g\big(\Psi_1(u,v)\big). \tag{F.18}$$

Here, f_g is the (relative) value function, given as

$$f_g(u,v) = \sum_{t=0}^{\infty} \left[g\big(\Psi_t(u,v)\big) - g\big(\mu_\infty, v_\infty\big)\right],$$

which is well-defined since the deterministic system has a unique equilibrium point.

Now, taking expectation of the Poisson equation (F.18) with respect to $(u,v) \sim (U_\infty, V_\infty)$, we get

$$\mathbb{E}\left[g\big(U_\infty, V_\infty\big) - g\big(\mu_\infty, v_\infty\big)\right] = \mathbb{E}\left[f_g(U_\infty, V_\infty) - f_g\big(\Psi_1(U_\infty, V_\infty)\big)\right].$$

Then, using the basic adjoint relationship $\mathbb{E}\left[f_g\big(G_1(U_\infty, V_\infty)\big) - f_g(U_\infty, V_\infty)\right] = 0$ for the stochastic system and adding this 0 term to the above equation, we then get

$$\begin{aligned}
&\mathbb{E}\left[g\big(U_\infty, V_\infty\big) - g\big(\mu_\infty, v_\infty\big)\right] \\
= \;& \mathbb{E}\left[f_g(U_\infty, V_\infty)) - f_g\big(\Psi_1(U_\infty, V_\infty)\big)\right] + \mathbb{E}\left[f_g\big(G_1(U_\infty, V_\infty)\big) - f_g(U_\infty, V_\infty)\right] \\
= \;& \mathbb{E}\left[f_g\big(G_1(U_\infty, V_\infty)\big) - f_g\big(\Psi_1(U_\infty, V_\infty)\big)\right].
\end{aligned}$$

Now we have achieved *generator coupling* on the right-hand side of the above equation. To bound the value $\left|\mathbb{E}\left[g\big(U_\infty, V_\infty\big) - g\big(\mu_\infty, v_\infty\big)\right]\right|$, we just need to bound

$$\begin{aligned}
&\left|\mathbb{E}\left[f_g\big(G_1(U_\infty, V_\infty)\big) - f_g\big(\Psi_1(U_\infty, V_\infty)\big)\right]\right| && \text{(F.19)}\\
= \;& \left|\mathbb{E}\Big[\sum_{t=0}^{\infty} \left[g\big(\Psi_t(G_1(U_\infty, V_\infty))\big) - g\big(\Psi_t(\Psi_1(U_\infty, V_\infty))\big)\right]\Big]\right|. && \text{(F.20)}
\end{aligned}$$

Roadmap

To apply the Stein's method framework to prove our main theorem, note that $U_\infty \in [0,1]$, while V_∞ is the steady-state counterpart of $V(t) = N(t)/N^q$ for the scaling factor N and $q \geq 3/2$. We discuss separately when V_∞ is in a bounded set and when it is not. We first state the following lemma, which shows the probability that V_∞ is not in the bounded set converges to 0 as $N \to \infty$.

Lemma F.2.1. *Under the stability condition for the DTMC $\{N(t)\}$, we have*

$$\lim_{N\to\infty} \mathbb{P}_\infty(N(t) > r_b N^q) = 0 \quad \forall q > 1,$$

where $r_b \geq 1$ is some constant that does not depend on N, q, and $\mathbb{P}_\infty$ denotes the steady-state probability of the DTMC.

The proof of this lemma is detailed in Appendix F.2.4. We denote the bounded set as $\mathcal{V}^b = [0, r_b]$ and $\tilde{\mathcal{V}}^b = \mathcal{V}/\mathcal{V}^b$ as the complement set of $\mathcal{V}^b$. Next, we first prove results for $(u,v) \in [0,1] \times \mathcal{V}^b$. Then, we prove the main theorem that considers the entire domain $[0,1] \times \mathcal{V}'$.

F.2.2 Results on States in the Bounded Set

We first consider testing function $h(u,v)$ where the outcome of h only depends on u and satisfies Assumption F.1.2. For example, $h(u,v) = g_1(u,v) = (u - \mu_\infty)^2$.

Proposition F.2.1. *Let $h : [0,1]\times\mathcal{V}^b \to \mathbb{R}$ be a function satisfying Assumption F.1.2. Conditioning on $(U_\infty, V_\infty) = (u,v) \in [0,1] \times \mathcal{V}^b$, we have that*

$$\left|\mathbb{E}\Big[\sum_{t=0}^{\infty} \big[h\big(\Psi_t(G_1(u,v))\big) - h\big(\Psi_t(\Psi_1(u,v))\big)\big]\Big]\right| \leq c_1^*/N,$$

where c_1^* is a constant that is independent of N or (u,v).

Proof. First, for a given $(u,v) \in [0,1] \times \mathcal{V}^b$, since the outcome of $h(u,v)$ only depends on u, it is equivalent to consider $h\big(\Psi_t^u(G_1(u,v))$ for $h\big(\Psi_t(G_1(u,v))$ and $h\big(\Psi_t^u(\Psi_1(u,v))$ for $h\big(\Psi_t(\Psi_1(u,v))$, where $\Psi_t^u(a,b)$ is the transition from $(\mu(0), v(0)) = (a,b)$ to $\mu(t)$ in t-steps. Note that Ψ_t^u is well-defined because the transition dynamics are deterministic as long as we know the starting state. Then, to prove the result, we use:

$$\begin{aligned}
&\left|\mathbb{E}\Big[\sum_{t=0}^{\infty} \big[h\big(\Psi_t^u(G_1(u,v))\big) - h\big(\Psi_t^u(\Psi_1(u,v))\big)\big]\Big]\right| \\
\leq\ &\left|\mathbb{E}\left[\sum_{t=0}^{T^*-1} \big[h\big(\Psi_t^u(G_1(u,v))\big) - h\big(\Psi_t^u(\Psi_1(u,v))\big)\big]\right]\right| &\text{(F.21)}\\
&+\left|\mathbb{E}\left[\sum_{t=T^*}^{\infty} \big[h\big(\Psi_t^u(G_1(u,v))\big) - h\big(\Psi_t^u(\Psi_1(u,v))\big)\big]\right]\right|. &\text{(F.22)}
\end{aligned}$$

Here, T^* is the first time that, from any $(u, v) \in [0,1] \times \mathcal{V}^b$, the maximum time *in the deterministic system* to reach a "contraction" region that the (1) blocking probability ($u(t) = \beta_t$) will not cross the zero point anymore and (2) if $\mu_\infty > 0$, the mapping from $u(t)$ to $u(t+1)$ is Lipschitz with constant $1-\epsilon$, where ϵ does not depend on (u, v) or N, i.e., it is a contraction mapping. Formally, it is defined as follows:

- If $C > \lambda_e + \frac{\lambda}{1-p}$, $T^* = \sup_{(u,v)} \{t \geq 0 : \ n(t) \leq C - \lambda_e, (u(0), v(0)) = (u, v)\}$;
- If $\frac{\lambda}{1-p} < C \leq \frac{\lambda}{1-p} + \lambda_e$, $T^* = \sup_{(u,v)} \left\{t \geq 0 : n(t) \in \left[\frac{C}{1-\mu^*+\delta} - \lambda_e, \frac{C}{1-\mu^*-\delta} - \lambda_e\right], (u(0), v(0)) = (u, v)\right\}$,

where $\delta > 0$ satisfies the conditions defined in Lemma F.2.2. Because $(u, v) \in [0,1] \times \mathcal{V}^b$ is on the bounded set, T^* is well defined. The rest of the proof leverages T^* to establish bounds for (F.21) and (F.22).

For (F.21), we establish Lemma F.2.3, which is a modified version of Lemma F.1.2 (focusing on Ψ_t^u). Based on this lemma, it is straightforward to show that

$$\left|\mathbb{E}\Big[\sum_{t=0}^{T^*-1} \big[h\big(\Psi_t^u(G_1(u,v))\big) - h\big(\Psi_t^u(\Psi_1(u,v))\big)\big]\Big]\right| \leq c_2^*/N,$$

where c_2^* is some constant that is independent of N or (u, v).

For (F.22), if $C > \lambda_e + \frac{\lambda}{1-p}$, since $u(t) = 0$ for all $t \geq T^*$, we have

$$\left|\mathbb{E}\left[\sum_{t=T^*}^{\infty} \big[h\big(\Psi_t^u(G_1(u,v))\big) - h\big(\Psi_t^u(\Psi_1(u,v))\big)\big]\right]\right| = 0.$$

If $\frac{\lambda}{1-p} < C \leq \frac{\lambda}{1-p} + \lambda_e$, by Lemma F.2.2, after entering the "contraction" region, $\Psi_1^u(u, v) \in \text{Lip}(1-\epsilon)$ with respect to u, i.e., $|\Psi_1^u(u_1, v_1) - \Psi_1^u(u_2, v_2)| \leq (1-\epsilon)|u_1 - u_2|$, where the Lipschitz constant $1-\epsilon < 1$ does not depend (u, v) or N. Given this result, we can show that

$$\begin{aligned}
& \left|h\big(\Psi_{T^*+k}^u(G_1(u,v))\big) - h\big(\Psi_{T^*+k}^u(\Psi_1(u,v))\big)\right| \\
\leq \; & \frac{1}{\gamma}\left|\Psi_{T^*+k}^u\big(G_1(u,v)\big) - \Psi_{T^*+k}^u\big(\Psi_1(u,v)\big)\right| \\
\leq \; & \frac{1-\epsilon}{\gamma}\left|\Psi_{T^*+k-1}^u\big(G_1(u,v)\big) - \Psi_{T^*+k-1}^u\big(\Psi_1(u,v)\big)\right| \ldots \\
\leq \; & \frac{(1-\epsilon)^k}{\gamma}\left|\Psi_{T^*}^u\big(G_1(u,v)\big) - \Psi_{T^*}^u\big(\Psi_1(u,v)\big)\right|,
\end{aligned}$$

and thus (F.22) can be bounded by the summation of the above geometric series as follows,

$$\begin{aligned} & \left|\mathbb{E}\left[\sum_{t=T^*}^{\infty}\left[h\big(\Psi_t^u(G_1(u,v))\big) - h\big(\Psi_t^u(\Psi_1(u,v))\big)\right]\right]\right| \\ \leq \quad & \big|\Psi_{T^*}^u\big(G_1(u,v)\big) - \Psi_{T^*}^u\big(\Psi_1(u,v)\big)\big| \cdot \sum_{k=0}^{\infty}(1-\epsilon)^k \\ \leq \quad & \frac{c_3^*}{\epsilon N}, \end{aligned}$$

where c_3^* comes from Lemma F.2.3, and $\frac{1}{\epsilon}$ comes from the sum of geometric series.

□

Applying $h(u,v) = g_1(u,v) = (u-\mu_\infty)^2$ to this proposition, we get the following corollary.

Corollary F.2.1. *Conditioning on $(U_\infty, V_\infty) = (u,v) \in [0,1] \times \mathcal{V}^b$, we have that*

$$\left|\mathbb{E}\Big[\sum_{t=0}^{\infty}\Big[\big(\Psi_t^u(G_1(u,v)) - \mu_\infty\big)^2 - \big(\Psi_t^u(\Psi_1(u,v)) - \mu_\infty\big)^2\Big]\Big]\right| \leq c_1^*/N,$$

where c_1^* is a constant that is independent of N or (u,v).

Consider testing functions $\tilde{h}(u,v)$ that only depend on v (e.g., $h(u,v) = g_2(u,v) = (v-v_\infty)^2$), we get a similar result as Proposition F.2.1.

Proposition F.2.2. *Let $\tilde{h} : [0,1] \times \mathcal{V}^b \to \mathbb{R}$ be a function satisfying Assumption F.1.2 except that the output of $\tilde{h}(u,v)$ only depends on v. Conditioning on $(U_\infty, V_\infty) = (u,v) \in [0,1] \times \mathcal{V}^b$, we have that*

$$\left|\mathbb{E}\Big[\sum_{t=0}^{\infty}\Big[\tilde{h}\big(\Psi_t(G_1(u,v))\big) - \tilde{h}\big(\Psi_t(\Psi_1(u,v))\big)\Big]\Big]\right| \leq \tilde{c}_1^*/N,$$

where $\tilde{c}_1^*$ is a constant that is independent of N or (u,v).

The proof for this proposition is almost the same as that for Proposition F.2.1, except since $\tilde{h}(u,v)$ only depends on v, it is equivalent to consider $\tilde{h}\big(\Psi_t^v(G_1(u,v))$ for $\tilde{h}\big(\Psi_t(G_1(u,v))$ and $\tilde{h}\big(\Psi_t^v(\Psi_1(u,v))$ for $\tilde{h}\big(\Psi_t(\Psi_1(u,v))$, where $\Psi_t^v(a,b)$ is the transition from $(\mu(0), v(0)) = (a,b)$ to $v(t)$ in t-steps. The proof also involves modifying Lemma F.1.2 to focusing on Ψ_t^v. Similarly, we get the following corollary by applying the testing function $\tilde{h}(u,v) = g_2(u,v) = (v-v_\infty)^2$ to Proposition F.2.2.

Corollary F.2.2. *Conditioning on $(U_\infty, V_\infty) = (u,v) \in [0,1] \times \mathcal{V}^b$, we have that*

$$\left|\mathbb{E}\Big[\sum_{t=0}^{\infty}\Big[\big(\Psi_t^v(G_1(u,v)) - v_\infty\big)^2 - \big(\Psi_t^v(\Psi_1(u,v)) - v_\infty\big)^2\Big]\Big]\right| \leq \tilde{c}_1^*/N,$$

where $\tilde{c}_1^*$ is a constant that is independent of N or (u,v).

F.2.3 Proof for the Main Theorem

For the main theorem, we focus on the testing function which is in the following form

$$g(u,v) = g_1(u,v) + g_2(u,v) = (u-\mu_\infty)^2 + (v-v_\infty)^2.$$

Proof. To prove

$$\mathbb{E}\left[||(U_\infty, V_\infty) - (\mu_\infty, v_\infty)||_2^2\right] = O\left(\frac{1}{N}\right), \tag{F.23}$$

we utilize (F.20) that considers the definition of the Poisson equation and get the following:

$$\begin{aligned} & \mathbb{E}\left[||(U_\infty, V_\infty) - (\mu_\infty, v_\infty)||_2^2\right] = \mathbb{E}\left[g\big(U_\infty, V_\infty\big)\right] \\ \leq & \left|\mathbb{E}\sum_{t=0}^{\infty}\left[g_1\big(\Psi_t^u(G_1(U_\infty,V_\infty))\big) - g_1\big(\Psi_t^u(\Psi_1(U_\infty,V_\infty))\big)\right]\right| \\ & + \left|\mathbb{E}\sum_{t=0}^{\infty}\left[g_2\big(\Psi_t^v(G_1(U_\infty,V_\infty))\big) - g_2\big(\Psi_t^v(\Psi_1(U_\infty,V_\infty))\big)\right]\right|. \end{aligned}$$

Applying the two corollaries proved in the Section F.2.2 (and recall that c_1^* and $\tilde{c}_1^*$ do not depend on N or $(u,v) \in [0,1] \times \mathcal{V}^b$), we get

$$RHS \tag{F.24}$$

$$\leq \; c_1^*/N + \tilde{c}_1^*/N$$

$$+ \; \left|\mathbb{E}\sum_{t=0}^{\infty}\left[g_1\big(\Psi_t^u(G_1(U_\infty,V_\infty))\big) - g_1\big(\Psi_t^u(\Psi_1(U_\infty,V_\infty))\big)\right]\mathbb{1}_{(U_\infty,V_\infty)\in[0,1]\times\tilde{\mathcal{V}}^b}\right| \tag{F.25}$$

$$+ \; \left|\mathbb{E}\sum_{t=0}^{\infty}\left[g_2\big(\Psi_t^v(G_1(U_\infty,V_\infty))\big) - g_2\big(\Psi_t^v(\Psi_1(U_\infty,V_\infty))\big)\right]\mathbb{1}_{(U_\infty,V_\infty)\in[0,1]\times\tilde{\mathcal{V}}^b}\right|. \tag{F.26}$$

To deal with the part outside the bounded set, we follow the framework in Liu and Ying (2018). For $(u,v) \in [0,1] \times \tilde{\mathcal{V}}^b$, i.e., when $v > r_b$, starting from v, let T_v be the time until the process re-enters the bounded set. Note that when $v > r_b$, the total number of patients in the system is of order $O(vN^{3/2})$ as we fix $q = 3/2$. In the deterministic system, under the stability condition, the decreasing rate of this total patient count is $(1-p)C - \lambda \sim O(N)$ since C, λ is of order N. As a result, the time T_v is of order $O((v-r_b)N^{1/2})$ as the decrease is linear with a rate $O(N)$ (also see the proof of Lemma F.2.1 in Appendix F.2.4). Applying the results from the transient analysis, we know that in each period $t \leq T_v$, the difference $\mathbb{E}\left[g_2\big(\Psi_t^v(G_1(U_\infty,V_\infty))\big) - g_2\big(\Psi_t^v(\Psi_1(U_\infty,V_\infty))\big)\right]$ is of order $O(1/N)$. As a result,

$$\mathbb{E}\sum_{t=0}^{T_v}\left[g_2\big(\Psi_t^v(G_1(U_\infty,V_\infty))\big) - g_2\big(\Psi_t^v(\Psi_1(U_\infty,V_\infty))\big)\right] \sim O((v-r_b)N^{-1/2}).$$

With the same argument, we can get the same order for the following term, where g_2 is replaced with g_1

$$\mathbb{E}\sum_{t=0}^{T_v}\left[g_1\big(\Psi_t^u(G_1(U_\infty, V_\infty))\big) - g_1\big(\Psi_t^u(\Psi_1(U_\infty, V_\infty))\big)\right] \sim O((v - r_b)N^{-1/2}).$$

After entering the bounded set $\mathcal{V}^b$, the cumulative difference is of order $O(1/N)$ as proved. Thus, there exist some constants $c_4, c_5 > 0$ such that we can bound (F.25) and (F.26) in the RHS with

$$\begin{aligned} RHS &\leq c_1^*/N + \tilde{c}_1^*/N \\ &+ c_4\mathbb{E}[(V_\infty - r_b)\mathbb{1}_{V_\infty > r_b}] \cdot N^{-1/2} + c_5\mathbb{P}(V_\infty > r_b) \cdot N^{-1}. \end{aligned} \tag{F.27}$$

Then, leveraging the fact that $v_\infty \to 0$ as $N \to \infty$, with $r_b > 1$ sufficiently large, we get

$$\mathbb{E}[(V_\infty - r_b)\mathbb{1}_{V_\infty > r_b}] \leq \mathbb{E}[(V_\infty - v_\infty)\mathbb{1}_{V_\infty > r_b}] \leq \mathbb{E}[(V_\infty - v_\infty)^2\mathbb{1}_{V_\infty > r_b}] \leq \mathbb{E}[(V_\infty - v_\infty)^2].$$

Now, plugging the above back to RHS, we get

$$\begin{aligned} & \mathbb{E}\left[||(U_\infty, V_\infty) - (\mu_\infty, v_\infty)||_2^2\right] = \mathbb{E}\left[(U_\infty - \mu_\infty)^2 + (V_\infty - v_\infty)^2\right] \\ \leq\; & RHS \\ \leq\; & c_1^*/N + \tilde{c}_1^*/N \\ & + c_4\mathbb{E}[(V_\infty - v_\infty)^2] \cdot N^{-1/2} + c_5 \cdot N^{-1}. \end{aligned}$$

In other words,

$$\mathbb{E}[(U_\infty - \mu_\infty)^2] + (1 - c_4{\cdot}N^{-1/2})\mathbb{E}[(V_\infty - v_\infty)^2] = O\left(\frac{1}{N}\right).$$

Since $\lim_{N\to\infty}(1 - c_4{\cdot}N^{-1/2}) = 1$, this concludes the proof. □

F.2.4 Proof for Lemmas

We start by presenting the proof for Lemma F.2.1.

Proof. To prove Lemma F.2.1, we first state the result from Bertsimas, D., Gamarnik, D., and Tsitsiklis, J. N. (2001), which will be leveraged to prove the lemma.

Proposition F.2.3 (Bertsimas et al. 2001)**.** *If there exists a function $V : \mathbb{N} \to \mathbb{R}_+$, such that there exists $\gamma > 0$ and $B \geq 0$ for the following to hold:*

- $\mathbb{E}[V(N(t+1)) - V(N(t))|N(t) = n] \leq -\gamma$ *for all n satisfying $V(n) \geq B$;*
- *the DTMC $\{N(t)\}$ has a stationary distribution π and $\mathbb{E}_\pi[V(N(t))] < \infty$,*

then, for the case such that $m = 0, 1, 2, \dots,$ we have that the following inequality always hold

$$\mathbb{P}_\infty(V(N(t) > B + 2v_{max}m) \leq \left(\frac{p_{max}v_{max}}{p_{max}v_{max} + \gamma}\right)^{m+1}, \tag{F.28}$$

where

$$\begin{aligned} p_{max} &= \sup_{n\in\mathbb{N}} \sum_{n'\in\mathbb{N}:V(n')>V(n)} \mathbb{P}(N(t+1) = n'|N(t) = n), \\ v_{max} &= \sum_{n,n':\mathbb{P}(N(t+1)=n'|N(t)=n)>0} |V(n') - V(n)|. \end{aligned}$$

To prove our results, we just need to verify the two conditions hold and $p_{\max}$ and $v_{\max}$ are suitably bounded. In the rest of our proof, we choose the Lyapunov function $V(n) = n$. We also use the time-stationary Assumption F.2.1.

We start by checking the first condition; the second is shown in Section 3.1.1 for proving the stability. For the first condition,

$$\begin{aligned} & \mathbb{E}[V(N(t+1)) - V(N(t))|N(t) = n] \\ = \; & \mathbb{E}[N(t+1) - N(t)|N(t) = n] \\ = \; & \lambda - n + \mathbb{E}[M^B(t)|N(t) = n)] + \mathbb{E}\Big[\mathbb{E}\big[bino(M^{NB}(t), p)|M^B(t) = m, N(t) = n\big]\Big] \\ = \; & \lambda - n + \mathbb{E}[M^B(t)|N(t) = n)] + \mathbb{E}\big[p(n - M^{NB}(t))|N(t) = n\big] \\ = \; & \lambda - (1-p)n + (1-p)\frac{n}{n+\lambda_e}(n + \lambda_e - C)^+. \end{aligned}$$

When $N(t) = n \geq \max\{C - \lambda_e, \lambda_e/\delta\} = B$ for some given constant $\delta > 0$, we can further write the above as

$$\begin{aligned} & \mathbb{E}[V(N(t+1)) - V(N(t))|N(t) = n] \\ = \; & \lambda - (1-p)n + (1-p)n\left(1 - \frac{C}{n+\lambda_e}\right) \\ = \; & \lambda - (1-p)C \cdot \frac{n}{n+\lambda_e} \leq \lambda - \frac{1-p}{1+\delta}C. \end{aligned}$$

Under the stability condition $C > \frac{\lambda}{1-p}$, there exists some constant $\gamma > 0$ such that $\mathbb{E}[V(N(t+1) - V(N(t)|N(t) = n] < -\gamma$ as long as δ is sufficiently small (i.e., B is sufficiently large).

Next, for $p_{\max}$ and $v_{\max}$, we get $p_{\max} \leq 1$. For $v_{\max}$, note that

$$N(t+1) - N(t) = M^B(t) + bino(M^{NB}(t), p) + \lambda - N(t) \geq \lambda - C$$

since $M^B(t) \geq N(t) - C$. We also have that $N(t+1) - N(t) \leq \lambda$. Thus, we get

$$v_{\max} = \max\{\lambda, |\lambda - C|\} = \max\{r, |r - r_c|\}N.$$

We set $\tilde{r}_b = \max\{r, |r - r_c|\}$ such that $v_{\max} = \tilde{r}_b N$.

Finally, since $V(N(t)) = N(t)$, leveraging the result given in (F.28) and setting $m = \lceil N^\alpha \rceil$, we have that

$$\mathbb{P}_\infty(N(t) > B + 2\tilde{r}_b N \cdot \lceil N^\alpha \rceil) \quad \leq \quad \left(\frac{\tilde{r}_b N}{\tilde{r}_b N + \gamma} \right)^{\lceil N^\alpha \rceil + 1}, \quad \forall \alpha > 0.$$

Thus, as $N \to \infty$, $\mathbb{P}_\infty(N(t) > B + 2\tilde{r}_b N \cdot \lceil N^\alpha \rceil) \to 0$ for any $\alpha > 0$. Since B does not grow with N, there exists a constant $r_b \geq 1$ such that $\mathbb{P}_\infty(N(t) > r_b N^q) \to 0$, where r_b does not depend on N, q for any $q > 1$. This implies that $N(t) = o(N^q)$ for any $q > 1$ in probability.

□

Lemma F.2.2. *If the capacity C satisfies $\frac{\lambda}{1-p} < C \leq \frac{\lambda}{1-p} + \lambda_e$, there exist some $0 < \delta < u^*$ and $0 < \epsilon < 1$, where $u^* = 1 - \frac{C(1-p)-\lambda}{\lambda_e(1-p)}$ is the blocking probability at equilibrium, for any two states $(u_1(t), v_1(t))$ and $(u_2(t), v_2(t))$ such that $u_1(t)$ and $u_2(t)$ are in the interval $[u^* - \delta, u^* + \delta]$, we have $|\Psi_1^u(u_1(t), v_1(t)) - \Psi_1^u(u_2(t), v_2(t))| \leq (1-\epsilon)|u_1(t) - u_2(t)|$, where this ϵ does not depend on N or u, v. In addition, if $u(t) \in [u^* - \delta, u^* + \delta]$, then $u(t+1) \in [u^* - \delta, u^* + \delta]$.*

Proof. Consider $u(t) > 0$. We first derive the closed form of the deterministic mapping from $u(t)$ to $u(t+1)$. For a state (u, v) of the deterministic system such that $u > 0$, it must satisfy $v = \frac{1}{N^q}\left(\frac{C}{1-u} - \lambda_e\right)$ and $n = vN^q = \frac{C}{1-u} - \lambda$. Thus

$$\begin{aligned} f(u) &= \Psi_1^u(u, v) \\ &= 1 - \frac{C/N^q}{v\big(1 - (1-p)(1-u)\big) + (\lambda + \lambda_e)/N^q} \\ &= 1 - \frac{C}{\left(\frac{C}{1-u} - \lambda_e\right)\big(1 - (1-p)(1-u)\big) + \lambda + \lambda_e} \\ &= 1 - \frac{C(1-u)}{C + \lambda_e(1-p)(1-u)^2 + (\lambda - C(1-p))(1-u)}. \end{aligned}$$

The derivative of f is

$$f'(u) = \frac{C\left[C - \lambda_e(1-p)(1-u)^2\right]}{\left[C + \lambda_e(1-p)(1-u)^2 + (\lambda - C(1-p))(1-u)\right]^2}.$$

At $u^* = 1 - \frac{C(1-p)-\lambda}{\lambda_e(1-p)}$, we have

$$\begin{aligned} f'(u^*) &= \frac{C\left[C - \lambda_e(1-p)\left(\frac{(C(1-p)-\lambda)^2}{\lambda_e^2(1-p)^2}\right)\right]}{\left[C + \lambda_e(1-p)\left(\frac{(C(1-p)-\lambda)^2}{\lambda_e^2(1-p)^2}\right) + (\lambda - C(1-p))\left(\frac{(C(1-p)-\lambda)}{\lambda_e(1-p)}\right)\right]^2} \\ &= \frac{C\left(C - \frac{(C(1-p)-\lambda)^2}{\lambda_e(1-p)}\right)}{C^2} \\ &= 1 - \frac{(C(1-p)-\lambda)^2}{\lambda_e(1-p)C} \end{aligned}$$

Since $\frac{\lambda}{1-p} < C < \frac{\lambda}{1-p} + \lambda_e$, we have $0 < \frac{C(1-p)-\lambda}{\lambda_e(1-p)} \leq 1$ and $0 < C(1-p) - \lambda < C$. It follows that $0 < \frac{(C(1-p)-\lambda)^2}{\lambda_e(1-p)C} < 1$. Thus $f'(u^*) = 1 - \frac{(C(1-p)-\lambda)^2}{\lambda_e(1-p)C} < 1$. In addition, since C, λ, λ_e are of order N, $f'(u^*)$ does not depend on N or q.

Since $f'(u)$ is continuous at u^*, there exist $\delta \in (0, u^*)$ and $\epsilon \in \left(0, \frac{(C(1-p)-\lambda)^2}{\lambda_e(1-p)C}\right)$ such that for every $u \in [u^* - \delta, u^* + \delta]$, $f'(u) < 1 - \epsilon$. In addition, for any $u \in [u^* - \delta, u^* + \delta]$, since

$$|f(u) - u^*| < (1-\epsilon)|u - u^*| < \delta(1-\epsilon) < \delta,$$

$f(u)$ is also in the interval $[u^* - \delta, u^* + \delta]$, i.e. if $u(t) \in [u^* - \delta, u^* + \delta]$, then $u(t+1) \in [u^* - \delta, u^* + \delta]$. □

Lemma F.2.3. *For any given state $(u, v) \in [0,1] \times \mathcal{V}$, and $h : [0,1] \to \mathbb{R}$ that is continuous, twice-differentiable and first derivative of h is $\frac{1}{\gamma}$-Lipschitz. we have that*

$$\left| \mathbb{E}\left[h\big(\Psi_t^u(G_1(u,v))\big) - h\big(\Psi_t^u(\Psi_1(u,v))\big) \right] \right| \leq c_3^* \left(\frac{1}{N}\right) \tag{F.29}$$

for any $t \geq 0$, where c_3^* does not depend (u, v).

Proof. (F.29) is true for $t = 0$ by Lemma F.1.2, since Ψ_0^u is the self-mapping and $h \circ \Psi_0^u : [0,1] \times \mathcal{V} \to \mathbb{R}$ satisfies Assumption F.1.2. For $t > 0$, we show that $h \circ \Psi_t^u$ satisfies Assumption F.1.2 in Lemma F.1.2 by induction. It is easy to verify $h \circ \Psi_1^u$ satisfies Assumption F.1.2 by Lemma F.1.4. Assuming $h \circ \Psi_t^u$ satisfying Assumption F.1.2, for $h \circ \Psi_{t+1}^u$, we have

$$h \circ \Psi_{t+1}^u = (h \circ \Psi_t^u) \circ \Psi_1.$$

Thus $(h \circ \Psi_t^u) \circ \Psi_1$ satisfies Assumption F.1.2 by Lemma F.1.4. Therefore $h \circ \Psi_t^u$ satisfies Assumption F.1.2 for each $t \geq 0$. □