Stochastic models and data driven simulations for healthcare operations

by

Ross Michael Anderson

B.S., Cornell University (2009)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Author .....
Sloan School of Management
May 15, 2014

Certified by .....
Itai Ashlagi
Assistant Professor of Operations Management
Thesis Supervisor

Certified by .....
David Gamarnik
Professor of Operations Research
Thesis Supervisor

Accepted by .....
Dimitris Bertsimas
Boeing Professor of Operations Research
Co-director, Operations Research Center

Stochastic models and data driven simulations for healthcare operations

by

Ross Michael Anderson

Submitted to the Sloan School of Management
on May 15, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis considers problems in two areas in the healthcare operations: Kidney Paired Donation (KPD) and scheduling medical residents in hospitals. In both areas, we explore the implications of policy change through high fidelity simulations. We then build stochastic models to provide strategic insight into how policy decisions affect the operations of these healthcare systems.

KPD programs enable patients with living but incompatible donors (referred to as *patient-donor pairs*) to exchange kidneys with other such pairs in a centrally organized clearing house. Exchanges involving two or more pairs are performed by arranging the pairs in a *cycle*, where the donor from each pair gives to the patient from the next pair. Alternatively, a so called altruistic donor can be used to initiate a *chain* of transplants through many pairs, ending on a patient without a willing donor. In recent years, the use of chains has become pervasive in KPD, with chains now accounting for the majority of KPD transplants performed in the United States. A major focus of our work is to understand why long chains have become the dominant method of exchange in KPD, and how to best integrate their use into exchange programs. In particular, we are interested in policies that KPD programs use to determine which exchanges to perform, which we refer to as *matching policies*. First, we devise a new algorithm using integer programming to maximize the number of transplants performed on a fixed pool of patients, demonstrating that matching policies which must solve this problem are implementable. Second, we evaluate the long run implications of various matching policies, both through high fidelity simulations and analytic models. Most importantly, we find that: (1) using long chains results in more transplants and reduced waiting time, and (2) the policy of maximizing the number of transplants performed each day is as good as any batching policy. Our theoretical results are based on introducing a novel model of a dynamically evolving random graph. The analysis of this model uses classical techniques from Erdős-Rényi random graph theory as well as tools from queueing theory including Lyapunov functions and Little's Law.

In the second half of this thesis, we consider the problem of how hospitals should design schedules for their medical residents. These schedules must have capacity

to treat all incoming patients, provide quality care, and comply with regulations restricting shift lengths. In 2011, the Accreditation Council for Graduate Medical Education (ACGME) instituted a new set of regulations on duty hours that restrict shift lengths for medical residents. We consider two operational questions for hospitals in light of these new regulations: will there be sufficient staff to admit all incoming patients, and how will the continuity of patient care be affected, particularly in a first day of a patients hospital stay, when such continuity is critical? To address these questions, we built a discrete event simulation tool using historical data from a major academic hospital, and compared several policies relying on both long and short shifts. The simulation tool was used to inform staffing level decisions at the hospital, which was transitioning away from long shifts. Use of the tool led to the following strategic insights. We found that schedules based on shorter more frequent shifts actually led to a *larger* admitting capacity. At the same time, such schedules generally reduce the continuity of care by most metrics when the departments operate at normal loads. However, in departments which operate at the critical capacity regime, we found that even the continuity of care improved in some metrics for schedules based on shorter shifts, due to a reduction in the use of overtime doctors. We develop an analytically tractable queueing model to capture these insights. The analysis of this model requires analyzing the steady-state behavior of the fluid limit of a queueing system, and proving a so called "interchange of limits" result.

Thesis Supervisor: Itai Ashlagi
Title: Assistant Professor of Operations Management

Thesis Supervisor: David Gamarnik
Title: Professor of Operations Research

Acknowledgments

I would like to thank my advisors, David Gamarnik and Itai Ashlagi. They have been very generous with their time, and taught me to be a better problem solver, writer, and presenter. Just as important, they have helped form my own perspective on what problems are interesting and important. Finally, they have been great people to work and socialize with for the last five years.

Thanks to Alvin Roth for serving on my committee and welcoming me into the world of kidney exchange. He introduced me to important people in the field, and has been a fun and interesting person to spend time with.

Chapter 4 is joint work with Yashodhan Kanoria. Yash has been everything one could ask for in a co-author. He is a spectacular problem solver and also willing to do the hard work when writing.

Thank you to everyone I worked with at Brigham and Women's Hospital, especially Danielle Scheurer. Danielle guided David and myself through the complexity of B&W to find a good problem, and was an advocate for our work within the hospital. She also organized funding for our research.

I would like to thank all of our contacts at the APD, NKR, and UNOS. Without their data, much of this research would not have been possible.

Thanks to Dimitris Bertsimas and Patrick Jaillet, the Co-Directors of the ORC. Beyond serving department well, they both were willing to reach out personally to me, both in discussing research and providing general advice. They were both extraordinarily committed to the department and the welfare of all their students, and I am grateful to have been a beneficiary.

I would like to thank Laura Rose and Andrew Carvalho, the administrators for the ORC. I was not the most punctual when it came to turning my paperwork in on time (this document included), but they made sure I never fell through the cracks.

John Tsitsiklis and Rob Freund were both excellent instructors, and I learned quite a lot from them. John has an incredible ability to take something very complex, determine what is important, and share it. I only took half of a class with Rob, but

it was excellent. I was also his TA for a summer and he was a pleasure to work for.

I would like thank all my friends from the ORC. In particular, I'm glad that I have become so close with Adam, my roommate for three years, and other the students from my year, Allison, Andre, Florin, Gonzalo, Maokai, and Vishal. I hope we remain close in the years to come. Thanks to Theo and David, for being both friends and mentors, to Phil, Michael, Mattheiu, Yehua, Nick, Will, Angie, Kris, Fernanda, Nathan, John, Iain, Paul and all of the others for five years of good times. The students of the ORC were the greatest part of MIT for me, both intellectually and socially, and I consider myself lucky to have been a member.

I would like to thank Bobby Kleinberg, David Shmoys, David Williamson, and Robert Bland from Cornell, for being great mentors and instructors. I would also like to thank my roommates Ben, Brian, Bryant, Devin, Matt, Miles and Rob from my time at Cornell. Had I not found my way into this group, I likely never would have made it to MIT.

I would like to thank my family for all their support. One day, I will start paying my cell phone bill and do my own taxes. Hopefully not too soon. I would also like to say that I'm glad that my brother Greg has moved to Boston, and that I can be here with him. You only get one family.

With some exceptions. I would like to thank the Relethfords for being a second family to me for so many years. David is my oldest friend and played a large role in making me who I am today. Our time at 23 Gorham (with Adam and Aileen) was probably too much fun. For the second Thanksgivings, the long car rides through upstate New York, and the many drafts of General Tso's chicken at China 19, I thank all of the Relethfords.

Last, I would like to thank Aileen. The last five years have been the happiest five years of my life, and it has been because of you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we look at a variety of problems arising in two areas of healthcare operations: Kidney Paired Donation (KPD) and scheduling medical residents in hospitals. These application areas will be discussed in detail in the subsequent sections. While these application areas share little from a medical perspective, more abstractly, both can be viewed as moderately large scale systems with complex stochastic dynamics and rich decision spaces, where simple and interpretable solutions are desired. We take a dual approach to these problems, considering them from both an operational and a strategic perspective. From an operational perspective, we build high fidelity simulation tools powered by historical data that can reproduce scenarios of what would have happened under various policies. While this approach allows us to quantitatively estimate the impact of a policy, it does little to provide systematic explanations of the advantages of a policy, particularly when the simulations produce seemingly counter-intuitive results. In order to better understand our problems from a strategic perspective, we design and analyze families of stochastic models. These models capture the key ideas driving the sometimes surprising simulation outcomes, yet are simple enough to reasonably be studied analytically. Our models are based on an innovative blend of models arising in queueing theory and the theory of random graphs, and their analysis is interesting in its own right. As these models are all moderately large scale stochastic systems and we are interested in the long run average behavior, there are some common technical tools we use in their analysis. In

particular, we use the Lyapunov function technique to approximately measure steady state performance. We then consider asymptotic limits of our large scale stochastic systems, and obtain limit theorems that asymptotically characterize the long run performance metrics.

We now discuss our two application areas, namely KPD and scheduling medical residents in hospitals in depth.

1.1 Kidney Paired Donation

As of November 2013, there are more than 98,700 patients in the United States on the cadaver waiting list for kidney transplantation [63]. Many of these patients have a friend or family member willing to be a living kidney donor, but who is biologically incompatible. Kidney Paired Donation (KPD) arose to allow these patients with willing donors (hereby referred to as *patient-donor pairs*) to exchange kidneys, thus increasing the number of living donor transplants and reducing the size of the cadaver waiting list.

In KPD, incompatible patient-donor pairs can exchange kidneys in *cycles* with other such pairs so that every patient receives a kidney from a compatible donor (e.g. see Figure 1-1). Additionally, there are a small number of so-called *altruistic donors*. These individuals are willing to donate their kidney to any patient without asking for anything in return. In KPD, an altruistic donor can be used to initiate a *chain* of transplants with incompatible pairs, ending with a patient on the deceased donor waiting list that has no associated donor (e.g. see Figure 1-2). Often, chains are planned in segments, where instead of immediately ending the chain on a patient from the deceased donor waiting list, the donor from the final pair in a segment is used to begin another segment after new patient-donor pairs arrive. Such donors are called *bridge donors*. Once the segment is executed, the bridge donors and altruistic donors are essentially identical for the purpose of planning future transplants, and are collectively referred to as *non-directed donors* (NDDs).

When donors agree to participate in a kidney exchange, there are no legal con-

Figure 1-1: Left: A cyclic exchange involving two patient-donor pairs. Each pair is represented by a node, where the blue half of the node represents the donor and the red half represents the patient. Right: A cycle with three patient donor pairs.

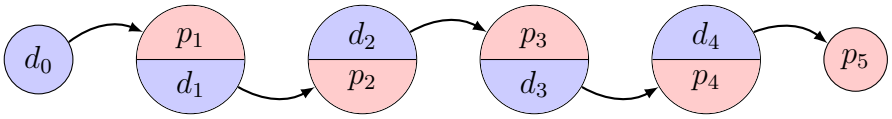

Figure 1-2: A chain exchange involving an altruistic donor, d_0, four patient-donor pairs, and a patient with no donor p_5. Each pair is represented by a node, where the blue half of the node represents the donor and the red half represents the patient.

tractual tools to enforce the exchange. For example, in the two way cycle from Figure 1-1, suppose that d_1 donates to p_2 today with the understanding that d_2 will donate to p_1 tomorrow. Then tomorrow, d_2 either (a) changes his mind about the transplant or (b) becomes too sick to donate. Now p_1 has no recourse, while p_2 has just gotten a kidney in exchange for nothing. Thus p_1 has been *irrevocably harmed*, in that he now has nothing left to trade and will be unable to participate in future exchanges. To prevent irrevocable harm, KPD programs always perform all the transplants in a cycle simultaneously. Consequently, KPD programs rarely organize cyclic exchanges involving more than three patient donor pairs, as it is too difficult logistically to organize more than three simultaneous transplants.

However, simultaneity is not always necessary to prevent irrevocable harm in an exchange. In particular, we need only that for every patient donor pair in the exchange, *the patient receives a kidney at a time no later than the time the donor gives his kidney.* While for cycles, this condition implies simultaneity, for chains initiated by an altruistic donor, it does not. In a chain with nonsimultaneous transplants, it is still possible for a donor to renege on a proposed exchange after his associated patient has received a kidney. While this outcome is very undesirable, as perhaps an

alternative chain could have resulted in more total transplants than the broken chain, it does not leave any patient irrevocably harmed. The benefit of allowing chains to be performed nonsimultaneously is two-fold: chains can be made longer without the logistical complexities of simultaneous surgeries, and patients need not wait for every member of the chain to arrive before they receive their transplant.

In all currently operating KPD programs, the decision of which transplants to perform is decided centrally by the program. We refer to any operational restrictions, e.g., a maximum cycle length or maximum chain segment length (if one is applicable), as *rules*, and the mechanism for deciding which transplants to perform subject to the rules as the *matching policy*. Many KPD programs use a *batching policy* as their matching policy, parametrized by n, where patient donor pairs arrive for n days, and then a set of exchanges are selected to maximize the number of transplants performed that day. We refer to this parameter n as the *match run time*. In the special case when $n = 1$, i.e. the KPD program optimizes for the current pool each day, we refer to this matching policy as the *greedy policy*.

At a high level, we seek to address two problems in KPD. First, for a fixed pool of patient-donor pairs and NDDs, we want to solve optimization problems over the set of feasible transplants. In particular, we focus on the problem of finding the maximum (possibly weighted) number of transplants, organized into cycles and chains, which we refer to as the Kidney Exchange Problem (KEP). An example of a KEP instance is shown in Figure 1-3. We consider several variations and extensions of this problem as well.

Second, we want to explore how the rules and matching policy of a KPD program affect key long run performance metrics, such as number of transplants performed and time patients must wait to receive a transplant. We will focus our analysis on comparing the batching policies (including the greedy policy), as these policies are being used in practice. In particular, we assess the impact of changing the rules restricting the maximum cycle and chain length and changing match run time. Perhaps the more interesting question is deciding the match run time, as it is not obvious how this quantity should be set to minimize the average patient waiting time. There is

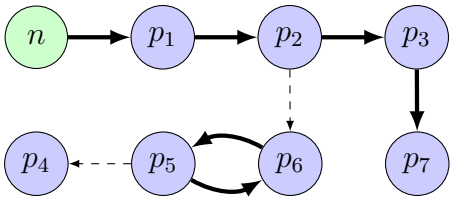

Figure 1-3: An example of a KEP instance. The node labeled n is a non-directed donor, and the remaining nodes p_1 through p_7 correspond to patient-donor pairs. Edges indicate possible transplants from the donor in the source node to the patient in the target node. In the optimal solution for this instance, indicated by the bold edges, we form the chain n, p_1, p_2, p_3, p_7, and the two-cycle with p_5 and p_6, leaving p_4 unmatched.

an intuitive trade-off in setting the match run time described as follows. Optimizing frequently reduces the time patients spend waiting between optimization runs. However, planning to transplant a patient donor pair in a two way cycle the day that they arrive could prevent many transplants in a chain that can only be formed using this pair with another pair that arrives the following day.

We approach this long run performance assessment from two perspectives. First, for several exchange programs, we simulate the dynamics of a KPD pool under various policies using their historical data. Note that these policies typically require solving an optimization problem over the set of feasible transplants periodically, thus it is a prerequisite that we can solve these optimization problems. Using this approach, we found that, surprisingly, the greedy policy was essentially the best among all batching policies in terms of average patient waiting time, and that this was true regardless of any rules restricting the maximum cycle or chain length. This result suggests that there is little to no loss of efficiency for existing KPD programs in using the greedy policy.

This surprising observation motivated our second, analytical approach to long run performance assessment. We designed a dynamic random graph model of a barter exchange system with the goal of explaining this phenomena. In the model, a homogeneous stream of nodes (patient-donor pairs) arrive and directed edges (biologically feasible transplants) are added randomly between the new node and existing nodes, according to a single parameter p indicating the sparseness of the graph. A rule speci-

fies what type of exchanges are allowed (either two-cycles, two-cycles and three-cycles, or the advancement of a single chain), and a policy dictates the strategy that is used to remove nodes (select exchanges). For this model, under all of the aforementioned rules, we show that asymptotically, as $p \to 0$, the greedy policy achieves the minimum possible average patient waiting time, up to constant factors. This result gives us insight as to why greedy was optimal in our (more complex) simulation model of a KPD pool.

In the subsequent sections, we first give some historical perspective on KPD. We then explain our contributions: algorithms for solving the KEP and related optimization problems, simulation results, and the analysis of our dynamic random graph model. Finally, we give some interpretation of the implications of our results for KPD.

1.1.1 Background & Related Literature

Integrating both cycles and chains in KPD was proposed in [67], where both the chains and cycle have unlimited size. As organizing many surgeries simultaneously is very logistically complex, the first implementations of KPD by New England Paired Kidney Exchange and other clearinghouses used only two-way cyclic exchanges. After a short period, clearinghouses have moved to allow three-way exchanges as well.

In [68], it was proposed to relax the requirement of simultaneity to the weaker requirement that every patient-donor pair receive a kidney before they give a kidney. As previously discussed, while this restriction still required all surgeries be performed simultaneously for cycles, it did allow for non-simultaneous chains. Since the first non-simultaneous chain was arranged [65], chain type exchanges have accounted for a majority of the transplants in kidney exchange clearinghouses. Approximately 75 percent of the transplants in National Kidney Registry (NKR) and the Alliance for Paired Donation (APD) are done through chains (these are two of the largest KPD programs in the United States). Chains involving as many as 30 pairs have been performed in practice, capturing significant public interest [70].

Variations of the KEP have been considered in the literature. As noted in [67],

when there is no maximum chain or cycle length, the problem can simply be solved with a single linear program, using the integral network flow polyhedron. The special case where chains are not permitted and only cycles of length two are used can be exactly solved very efficiently, as it is equivalent to the maximum matching problem. In [1], the special case of this problem where only cycles of length two and three are used was shown to be NP-hard. See [16] for a stronger negative result in this special case. However, integer programming techniques have been used by a variety of authors to solve special cases of the KEP without chains or with chains of bounded length, as first proposed in [69]. In [1], by improving the integer programming formulation of [69] and devising good heuristics, the authors are able to solve KEP instances with thousands of patient donor pairs, but without chains. Alternate IP formulations were further explored for this special case in [20]. In [67], a heuristic to produce feasible solutions when using chains and cycles of bounded length was suggested. However, no optimization algorithm was given, opening a major algorithmic challenge to reliably solve large scale instances of the general kidney exchange problem. The technique of [1] was extended in [23] to solve very large instances with bounded cycles and bounded chains. However, the algorithm became impractical when the maximum chain length was larger than four or five, as the formulation required a (column generated) decision variable for every chain with length at most the maximum chain length.

The techniques we use to solve the KEP, as described in this thesis, are similar to techniques used to solve variations of the Traveling Salesman Problem (TSP), a well known and difficult combinatorial optimization problem. The literature on this topic is vast (see [46] and the references within). For several variations of the TSP that are similar in spirit to the KEP, integer programming solvers have been developed [13, 33, 34, 39] (this list is not intended to be comprehensive). In particular, our formulation of the KEP most closely resembles a sparse, directed version of the Prize Collecting TSP (PC-TSP). To the best of our knowledge, this particular variation has not been studied previously.

1.1.2 Algorithmic Results

The primary algorithmic challenge we address is to find an algorithm to solve *real* instances of the KEP without bounding the maximum chain length. Solving this optimization problem is critical to the operations of KPD programs, which form long chains in practice. Previously, these programs either only searched for short chains or relied on heuristics that could lead to suboptimal solutions. Additionally, having fast algorithms to solve the KEP is required to make the simulation of a KPD program practical, which is the subject of the next section. Ultimately, we address this challenge by giving two new algorithms based on integer programming to solve the KEP. One of these algorithms is motivated by an integer programming formulation of the so-called *Prize Collecting Traveling Salesman Problem (PC-TSP)*, another NP-hard problem. The PC-TSP is a variant of the classical Traveling Salesman Problem (TSP), one of the most widely studied NP-hard problems in combinatorial optimization.

We emphasize that in evaluating the success of our algorithm, we use real instances drawn from historical data, as opposed to instances created by randomly generating patient donor pairs or compatibility graphs. We do this because instances encountered in practice are generated by the KPD pool dynamics. As a result, the statistics of compatibility for the patient-donor pairs still waiting to be matched are complicated (a patient-donor pair waiting to be matched is on average more difficult to match than a patient-donor pair randomly selected from historical data, as the easy to match patients are on average matched more quickly). Thus it is difficult to generate representative instances randomly.

Running our algorithms on real data instances (typically with a few hundred nodes) we find that:

- For the vast majority of these instances, both algorithms can solve the problem to optimality in a few seconds.
- There are a few difficult instances where using our PC-TSP based algorithm results in an improvement in run time by several orders of magnitude.

In addition to devising practical algorithms to solve these problems, we obtain several

theoretical results. In particular:

- We prove that the integer programming formulation from our PC-TSP based algorithm is the stronger of the two formulations, as measured by the value of the linear programming relaxation.
- We give a polynomial time algorithm to solve the separation problem for the integer programming formulation from our PC-TSP based algorithm.
- We devise an interesting pathological instance of the KEP that highlights the worst case difficulty of the KEP and potential differences in performance of the various integer programming formulations.

We also provide algorithms to address a variety of extensions to the KEP, including:

- Solving the KEP with a large (but bounded) maximum cycle length and unbounded chains. (While in practice, cycles longer than three are rarely formed due to logistical issues, there is some evidence that there may be benefits to considering longer cycles [7].)
- Solving the KEP with a large (but bounded) maximum chain length and short cycles. (Thus our algorithms are also capable of solving the problem described in [23].)
- Solving a two stage version of the KEP, where a limited set of edges is selected in stage one, then edges fail at random, and cycles and chains are built using only the edges that were selected in phase one and did not fail. This significantly generalizes the results of several other works [17, 24, 61], at the cost of a computationally more expensive algorithm.

In solving all of these problems, a variety of heuristics were developed. The details can be found in Chapter 2.

1.1.3 Simulation Results

We simulated to evolution of the KPD pool for NKR, a large exchange program, over a multi-year horizon. As inputs to the simulation, we considered several factors, most notably including:

- The match run time,

- The maximum size of an allowed cycle,
- The number of altruistic donors (possibly none).

We focused on the following performance metrics: *total number of transplants performed* and *average waiting time to receive a transplant*. Additionally, we looked at these metrics for sub-populations of the exchange pool that are "difficult to match." We identified these sub-populations based on statistical properties of the historical entrants to the NKR KPD program, most notably using *Pair Match Power*, as defined in Chapter 3.

Qualitatively, our results are summarized below:

- *The greedy policy is as good as any batching policy.* Under essentially all combinations of inputs, patient waiting time is minimized or nearly minimized with a match run time of one day (i.e. greedy policy). Additionally, there are no significant increases in the total number of transplants until the match run time is increased to at least three months, which gets only a few percent extra matches. There are several practical reasons why a KPD would not want to wait three months between much runs, as discussed in Section 1.1.5.
- *Disallowing chains results in substantially worse outcomes.* While simulating the NKR KPD pool, eliminating the use of chains reduced the total number of transplants from 264 to 201, and increased average waiting time from 158 days to 214 days.
- *At the current number of altruistic donors, rules changing the maximum cycle length have essentially no effect on long run performance.* Perhaps with fewer altruistic donors in the system, the outcome would have been different, as in [7].
- *There is little room to improve in total transplants relative to the greedy policy.* The number of transplants achieved by the greedy policy is within 15% of the offline solution, an upper bound for any online solution.

Features notably absent in the simulations include:

- *Patient donor pair abandonment*, where a pair permanently leaves the system (typically due to illness). This can happen before a node is matched if it has

waited too long, including while the node is the bridge node in a chain.

- *Edge failures*, where a transplant that has been planned must be canceled due to a discovered incompatibility.
- *Time delays* for transplants to be scheduled and performed, and for checking compatibility (the cause of edge failures).

A discussion of how the inclusion of these features in our simulation model could change our conclusions is briefly given in Section 1.1.5.

1.1.4 Dynamic Random Graph Results

In Section 1.1.3, we observed the surprising result that in essentially all settings, no online policy could generate a lower average patient waiting time than the greedy policy (using a match run time of one). This result was surprising because there was a perceived tradeoff between match run time and average patient waiting time as follows: Increasing the match run time would give us more information before deciding which exchanges to make, potentially enabling smarter matches that reduced average waiting time. However, increasing the match run time would force patients that could be matched immediately to wait longer for an opportunities to be matched. Additionally, we observed that without chains, substantially fewer transplants would be made. In this section, we develop a model that analytically demonstrates why these observations should occur. Our model is a new model of a dynamic random graph that can be analyzed using a combination of tools from random graph theory and queueing theory.

Our model is briefly summarized as follows. In each time period, a new node v arrives (interpreted as a new incompatible pair), and for each other node in the existing graph, a directed edge is added to and from v with probability p, i.i.d. (corresponding to a potential transplant). After each arrival of a new node, there is an option to remove nodes according to one of the three rules. The first two rules are given below:

1. Remove any number of node disjoint directed two-cycles.
2. Remove any number of node disjoint directed two-cycles and three-cycles.

Under the third rule we delete chains. However, we need to maintain an extra part of the state. In every time period, there is a single special node, the bridge donor. The rule is:

3. Remove any directed simple (node disjoint) path starting from the bridge donor. The final node in the path becomes the new bridge donor.

For each rule, we want to compare the greedy policy (where in every time period we remove the maximum number of nodes) to arbitrary policies, or at least batching policies (where we only attempt to remove nodes every n time periods, for some n possibly depending on p). The most relevant performance metric in this model is long run average node waiting time. It turns out that the total number of matches is not actually a relevant metric in our infinite time horizon model, as we will show that all nodes are eventually matched in finite time with probability one.

We discover that the steady state number of nodes in the system under the greedy policy is $\Theta(1/p^2)$ for two-cycles, $\Theta(1/p^{3/2})$ for two-cycles and three-cycles, and $\Theta(1/p)$ for chains, as $p \to 0$. Further, we show that every policy must have a steady state number of nodes in system of at least $\Theta(1/p^2)$ for two-cycles, and $\Theta(1/p)$ for chains, thus we conclude that the greedy policy is optimal up to constant factors for the two-cycle and chain rules. For the rule deleting two and three cycles, we show that every *monotone policy*, which we define in Chapter 4 and includes all batching policies, must have a steady state number of nodes in the system of at least $\Theta(1/p^{3/2})$, so we conclude that the greedy policy is optimal up to constant factors among monotone policies. Finally, we note that by applying "Little's Law," the steady state average waiting time must equal to the steady average number of nodes in system. Thus our conclusions apply to the steady state expected waiting times as well.

1.1.5 Operational & Strategic Insights for KPD

In this section, we discuss how our results relate to several practical aspects of KPD, both at the operational and strategic levels.

At an operational level, our algorithmic results from Section 1.1.2 have two important consequences for KPD. First, our algorithms *enable KPD programs to efficiently*

and optimally search for long chains in their daily operations. Our implementation of our algorithm is currently used to by:

- The APD, a large KPD program,
- Several hospitals that run individual exchange programs, including Northwestern Memorial Hospital (Chicago), Methodist Hospital (San Antonio), Georgetown Medical Center (Washington DC), and Samsung Medical Center (Korea),
- An "inter-exchange program" between the APD and NKR (the largest KPD program in the US as measured by transplants arranged) that finds matches between pairs registered in these two separate KPD programs.

Second, our algorithms *give KPD programs a tool to evaluate the implications of possible future policy changes.* The improvement in solve time from hours or minutes to seconds is of particular importance in this case, as at times it is desirable to simulate many policies, each with hundreds of replications, creating a large burden of computation. Our software has been used to run simulations to analyze a variety of policy decisions for both NKR and the United Network for Organ Sharing (UNOS) pilot exchange program.

At the strategic level, our results inform the discussions around several major questions that have been debated in the KPD community for some time. First, our work supports the proposal that *long chains should be used in kidney exchange.* When long, non-simultaneous chains were first proposed, they were highly controversial [40]. The debate over the importance of long chains in KPD has largely already been resolved, in part due to the success of NKR and the APD (most of their transplants were performed in non-simultaneous chains). Our results provide experimental evidence and theoretical justification to support the observation that long chains have been successful in practice. Specifically, in our (somewhat stylized) random graph model, we found that using chains resulted in a patient waiting time which was an order of magnitude smaller than relying solely on cycles. Likewise, in our simulation experiments, we observed that organizing exchanges based on chains of unbounded length resulted in more transplants and lower patient waiting time than using cycles only.

Second, our results suggest that *there is little loss of efficiency in using the greedy policy*, at least when compared to the batching policies. The optimality result for the greedy policy in our theoretical model does not directly imply this, as in practice compatibility is not homogeneous in KPD, and our result was asymptotic and only up to constant factors. However, the generality of our result under all three rules for making exchanges does suggest that the greedy policy is at least a near optimal way to manage exchange pools. We note that in other dynamic models with heterogeneous patient types [8], batching can improve average waiting time. However, our empirical work suggests that the upside from batching is quite limited, particularly for existing kidney exchange programs (the situation may change if exchange programs become orders of magnitude larger).

Finally, our results tie into a larger question for kidney exchange, at least in the United States. Namely, there has been considerable debate as to *if there should be a single national KPD program*, both in the medical literature and the popular press [71]. In particular, it has been suggested that competing exchanges are incentivized to use a greedy policy so that they can match patients listed in multiple exchanges before their competitors. In [71], a statement was made suggesting that this myopic behavior wastefully uses easy to match pairs, "causing[ing] cherry-picking that undermines optimization... It kind of creates this race to the bottom." In particular, they are suggesting that competition is incentivizing KPD programs to inefficiently organize exchanges, under the assumption that the use of the greedy policy is inefficient. Our result stating that the greedy policy is reasonably efficient suggests that the loss of efficiency due to competition between exchange programs is not substantial. Whether or not KPD should be nationalized in the United States is a very complicated question involving many factors significantly outside the scope of this work. However, our results suggest that from an efficiency perspective, competition may not be as counterproductive as previously thought.

Discussion of Modeling Assumptions

As previously mentioned, there are some important practical features of kidney exchange that our simulations and theoretical models fail to account for, most notably *abandonment*, *edge failures*, and *time delays*. We briefly discuss how the incorporation of these factors into our models might affect our conclusions.

First, we consider *abandonments*, under the assumption that patients will abandon a constant rate, independent of how long they have already been waiting. In such a model, the total number of abandonments would be proportionate to the average waiting time of a patient. *Thus policies that have the least waiting time will have the fewest abandonments.* Without abandonments in our model, we found that the average patient waiting time was made smallest by: (1) using long chains and (2) using a greedy policy. Suppose that by the introducing abandonments into our model, the result still held that using long chains and a greedy policy minimized average patient waiting time. Under this assumption, by the italicized claim above, using long chains and a greedy policy would minimize abandonments as well. Abandonments are clearly undesirable from the perspective of our performance metric of total matches, as patients that abandon cannot be matched. Thus we conclude that adding abandonments to our model would strengthen our arguments in favor of using long chains and using a greedy policy.

A caveat in our argument is that we do not fully consider the tradeoffs involved in using bridge donors vs. conducting exchanges simultaneously when considering the cost of abandonment. It was suggested in [40], another simulation study, that abandonment by bridge donors (who have little incentive not to abandon, as their desired patient has already received a kidney) would result in long non-simultaneous chains producing fewer total long run transplants than short simultaneous chains. The tradeoffs involved in analyzing this problem are complex and somewhat beyond the scope of this work. However, at NKR, the largest KPD program in the United States, abandonment by bridge donors has been rare, (they had no broken chains in 2012 and only 3 broken chains in 2013, out of around 80 total chains), suggesting

that perhaps this is not a first order issue.

Finally, we briefly mention that our assumption of a constant abandonment rate could likely be improved. The rate of abandonments caused by patients becoming too sick to transplant would likely grow over time, as the longer a patient is on dialysis waiting to receive a kidney, the more likely they are to develop further health complications. As our (heuristic) analysis above depended heavily on the abandonment rate being constant, some of the conclusions on the effect of introducing abandonments could change.

Next, we consider *edge failures*. We break the edge failures into two types that are dealt with very differently from an operational perspective. The first is *immediate* failure, where a potential transplant is rejected immediately after it is proposed because of a discovered biological incompatibility or a doctor deeming a donor unfit for their patient. The second type is caused by abandonment, were after an exchange is agreed upon and finalized, a donor or patient becomes unable to participate. In practice, the first type of failure is common at many KPD programs, while the second type of failure tends to be more rare. As the number of edge failures of the second kind are rather small, we believe that their effect should be small and not change our conclusions. Additionally, we mention that NKR has put considerable effort into reducing failures of the first type, and has recently succeeded in making both failure types rare.

At first glance, edge failures appear quite problematic for long chains, as a very long chain is quite likely to have at least one failure, and every transplant after a point of failure cannot be executed. However, the true cost of edge failures depends on the recourse options after edges fail. Under the NKR model, if there is a single *immediate* edge failure in a proposed exchange, all offers are withdrawn and a new set of offers is made. Such a policy is only possible because the NKR match run time is only a day. Under a greedy policy in combination with this offer withdrawal strategy, the effect of immediate edge failures is essentially completely mitigated (and our conclusions will remain unchanged). If instead, the match run time were for example three months, then offer withdrawal recourse is infeasible, as the many iterations required to find

a proposed exchange with no immediate edge failures could take years. If under a three month match run time, the recourse to immediate edge failure is to truncating chains and cancel failed cycles, then forming very long chains (or long cycles) would be of little value.

Finally, we mention that in Chapter 2, we develop a stochastic optimization methodology for the KEP that enables optimizing the expected number of transplants performed after edge failures occur and some recourse is taken. The technique is very general and applicable to a wide range of recourse structures. For example, when proposing exchanges, instead of requiring that the exchanges could all be feasibly executed, one could propose a set of exchanges where some pairs receive or give multiple offers. Then once the immediate failure information is realized, a maximum subset of these exchanges (which did not fail) could be executed in cycles and chains. While understanding the long run implications of adding both edge failures and complex recourse would require further study, it would appear that strategies with complex recourse could benefit from using long chains.

Last, we consider *time delays*. In both our theoretical and simulation models, we assumed that patients were transplanted as soon as they were matched. However, in reality, there are many delays in the system that our models do not account for. As many of our conclusions about greedy policies and using long chains were measured in average patient waiting time, accounting for these delays could affect our conclusions. Sources of delay not accounted for in our models include:

(i) After a set of transplants is proposed, a series of tests must be performed to verify that the donors and patients are biologically compatible.

(ii) For simultaneous exchanges, a date must be found when all the doctors needed are available.

(iii) For non-simultaneous chains, the transplants must be performed in the order that they occur in the chain (so additional delays propagate down the chain). In particular, note that a chain can be resumed from a bridge donor only when all the previous transplants have been executed.

We now discuss how these delays could influence our conclusions. For (i), as the time

required to test for biological compatibility is short and continues to fall as KPD programs improve their operations, we expect this to have a minimal impact on our conclusions. Delay (ii) seems to strengthen our conclusions in favor of using long non-simultaneous chains over cycles (particularly three way cycles), as with chains transplants need not be scheduled simultaneously if doing so is inconvenient. Issue (iii) of delays propagating down long chains and these delays not being accounted for in our models will clearly lead us to overestimate the benefits of long chains. To quantitatively estimate this effect would require further study and likely considerable effort. However, we find it very unlikely that this effect would outweigh the reduction in average waiting time gained by using chains.

1.2 Scheduling Medical Residents in Hospitals

Major hospitals face a difficult challenge of designing shift schedules for their residents that have capacity to treat all incoming patients, provide quality care, and are compliant with regulations restricting shift lengths. Recently, there has been much controversy surrounding the use of long shifts, and the resulting fatigue. In particular, fatigue during long resident shifts has been implicated as a cause of medical errors [10, 37, 57], burnout, depression and other psychological problems [48, 72, 75], and motor vehicle crashes [11]. In order to address these and related issues, the ACGME instituted a new set of regulations on duty hours limiting the duration of shift lengths to 16 hours for the first year residents [50]. The new regulations have forced many academic medical centers to overhaul their shift schedules. Proponents of the long shift are concerned with how this regulation will affect patient *continuity of care*. They argue that reducing shift lengths will result in more patient handoffs between caring practitioners, increasing the chance of miscommunication and accidents [21, 64]. A particularly undesirable type of handoff is a *reassignment*, when a patient is admitted temporarily by one doctor and then is transferred to a resident for a permanent care. Reassignments are dangerous as they greatly increase the risk of losing information that should be used in determining a course of treatment. Ad-

ditionally, long shift advocates additionally argue that reducing residents' hours will force hospitals to increase staffing levels to compensate for lost capacity to admit patients [81, 82]. The impact of shift schedules on *(a) the admitting capacity* and *(b) the number of reassignments* are two main questions we address.

First we approached these problems at an operational level. Working with Brigham and Women's (B&W) hospital, a major academic hospital in the Boston area, we built a high fidelity discrete event simulator that given a resident schedule for the hospital, can provide estimates of (a) capacity and (b) continuity of care using a variety of performance metrics. Additionally, we developed a Markov chain based approximation for the capacity of a hospital's resident schedule. Our approximation was far more accurate than the simple bounds on capacity implied by making a basic rate calculation using the number of residents available with either (i) the number of patients each resident can have in care simultaneously or (ii) the number of patients each resident can admit per day. Using these tools, we were able to inform B&W of the implications for each of the potential schedules they were considering when transitioning to schedules satisfying the 16 hour shift length regulation. In contrast to much of the existing literature on the effects of duty hour regulations which relies on the perceptions of outcomes gathered in surveys (particularly when measuring the effect on quality of care), our approach directly measures the relevant performance metrics.

Our simulations led to a few surprising strategic conclusions. We found that schedules based on shorter more frequent shifts that held total labor hours constant actually led to a larger admitting capacity. At the same time, such schedules generally reduce the continuity of care by most metrics when the departments operate at normal loads. However, as a hospital department approached the critical capacity regime, we found that the continuity of care improved in some metrics for schedules based on shorter shifts, most notably in total reassignments, due to a reduction in the use of overtime doctors.

To better understand these strategic insights, we developed a stylized queueing model of patient flow in a hospital where analytic results on capacity and continuity

of care can be shown. In particular, we prove that a schedule with shorter more frequent shifts has a greater admitting capacity than a schedule with long shifts. Additionally, we show in an asymptotic scaling that the schedule based on shorter more frequent shifts will provide a better continuity of care (measured by the number of reassignments) when the patient load is high.

In the subsequent sections, we first give some background on medical resident scheduling. We then discuss the results from our simulations and our analytic models. Finally, we give some interpretation of our results for medical resident scheduling.

1.2.1 Background & Related Literature

To the best of our knowledge, this paper is the first study to quantitatively measure the impact of duty hour restrictions (in particular, maximum shift lengths) on capacity, i.e. issue (a). There have been some related studies [77, 81, 82], which suggested that duty hour restrictions (reducing the total number of hours residents can work) have caused hospitals to hire PAs and nurse practitioners to decrease the workload of residents. Some previous studies have investigated the impact on continuity [82] after New York state put a cap of 80 hours a week on resident work hours in 1989, [48] when the ACGME created a similar national work hour restrictions in 2003, and most recently [26] and the follow up [27] as well as [75] with the 2011 restrictions. All of these studies have suggested that these regulations have not improved the quality of care. However, these studies often aggregated resident and physician perceptions of the change in the quality of care from before to after regulation, which is a subjective metric [35, 82]. As noted in [35], while there have been many studies on the effect of long shifts and imposing caps on duty hours in the last 30 years, very few studies used randomized control design, and most used perceived outcomes instead of actual outcomes, hence the contradictory results of these studies are inconclusive.

We now mention some other approaches considered in the field of operations research for capacity management in hospitals. A very general survey of capacity management in healthcare is given in [43]. Simulation studies of capacity have been done for medical resident schedules [25, 53] and various other hospital resources

[29, 51, 58, 66, 83]. However, as these simulations are incredibly sensitive to the details of each hospital's operations, the results do not generalize well to other hospitals. There are some recent papers which propose a queueing model of patient flow in a hospital, and then solve for important performance metrics, either analytically [85], asymptotically [22, 85, 86], numerically [74], or with heuristic methods [44]. Although not explicitly about healthcare, in [49, 59], fluid models for queues with time varying arrival rates alternating between overloaded and underloaded periods are considered. These models are more in the spirit of our work than the previously cited models from a technical perspective, as transient behavior and "end of day effects" (see [45]) play a prominent role.

1.2.2 Simulation Results

We constructed a discrete event simulation tool in order to model a variety of resident shift schedules and their implications for the issues (a) and (b) above. In the simulations, a sequence of patient arrival and departure times was given by a historical dataset. The primary caregivers for these patients are the medical residents and the Physician Assistants (PAs), who are organized into teams for the purposes of admitting and caring for patients. Both individuals and teams are restricted in the maximum number of patients they can admit per shift, and the maximum number of patients they can have in care (where often the team limit is smaller than the sum of the individual limits). When all available residents and PAs are at capacity, patients are temporarily admitted by a doctor, then *reassigned* resident or PA for long term care until discharge. There are two channels for reassignment, night floats or an overtime doctor (referred to as a *jeopardy admission*). The latter is far less desirable, as the overtime doctors typically leave soon after the night floats arrive, so patients they admit are transferred from the overtime doctor to the night float and then finally to their caring doctor (a resident or PA), giving additional opportunities for miscommunication. Additionally, overtime doctors are expensive for the hospital.

We ran our tool against the historical data set of patients at B&W in three main areas of services: General Medicine, Cardiology and Oncology. To determine the

impact on admitting capacity, we used the frequency of jeopardy as one performance metric. To determine the impact on continuity, we collected several performance metrics focusing on continuity of care in the critical period of the first 24 hours of a patients time in the hospital, including:

- the number of patients which had to be admitted by one care giver and then permanently transferred to another, referred to as reassignments (e.g. a patient is admitted by a jeopardy doctor or a night float and then is transferred to a resident the following morning).
- the frequency that the admitting doctor remained in the hospital for less than hours after the patient arrived, and
- the frequency that residents must admit another patient within two hours of any admission. (Newly admitted patients take about two hours to "work up," a process requiring much of a resident's attention. When a resident admits two patients within a two hour window, they must work up two patients at the same time, splitting their focus and increasing the chance of error.)

The performance metrics were computed for the existing schedules, which include long (30 hours) shift lengths, as well as for alternative schedules with shorter more frequent shifts, in compliance with the new ACGME regulations. In our comparisons, we only consider schedules where the number of hours that residents will be eligible to admit new patients is the same.

In our findings, the choice of resident schedule had a strong impact on most of the performance metrics. In particular, we found that schedules with shorter (about 10 hours) more frequent shifts are better capable of handling larger volumes of patients. For example, in the General Medicine Service (GMS), a department at B&W, jeopardy levels under the baseline patient load were 33 patients per year under a schedule with short frequent shifts and 155 patients per year under a schedule with long shifts. We also found that for all schedules, the number of jeopardy admissions markedly increases with even modest increases in patient volume (more precisely, the number of jeopardy admissions increases rapidly and non-linearly with an increase in the volume of patients). Continuing our previous example using GMS, when the

patient load increased by 10%, we saw that the short shift schedule had an increase of jeopardy patients to 256 patients per year, while the long shift schedule had a much larger increase to 690 patients per year.

For the continuity of care in the first 24 hours of admission, our findings are more subtle. Schedules based on short more frequent shifts contained gaps in resident coverage, increasing total reassignments during off-peak hours. However, as these schedules lead to an increased capacity (and thus reduced jeopardy instances), they resulted in fewer reassignments during peak hours. Thus the effect of reducing shift lengths on the total number of reassignments was somewhat data dependent and was not uniform across departments. At the baseline level of patient load, we found that in moving from a schedule based on long shifts to a schedule based on short shifts, Cardiology would have an extra 1000 reassignments, Oncology would have about 500 extra reassignments, and for GMS there would be essentially no change in the number of reassignments (however, GMS used both intensive teaching units and extra PAs relative to Oncology and Cardiology, skewing the results, see Chapter 3 for more details). At the same time, under an increased patient load, shorter shifts caused fewer reassignments across all departments. Uninterrupted observation of a patient for the first 6 hours since admission was challenging for all schedules; while the percentage of patients receiving six hours of observation by their admitting doctor was low for schedules with shorter shifts (15–20%, varying by department), it was still only around 50% for schedules using long shifts.

Additionally, we performed some sensitivity analysis where we adjusted the rate that patients arrived to the hospital, and measured how far it could be increased under each resident schedule before the department would be over capacity (e.g. overtime would be required regularly). We call maximum stable arrival rate the *throughput* of the schedule. We then compared the throughput with two natural upper bounds, computed by estimating the maximum rate each team of residents and PAs could treat new patients, and summing this rate up over all teams in the department. The first bound, the *capacity upper bound on throughput*, assumes that each resident is constantly caring for their maximum number of patients in care, and from the average

length of stay, infers the average rate that the resident could admit new patients per day. As this bound ignores the fact that residents are restricted in when they can admit new patients by their schedule, we expect it to exceed the true throughput. The second bound, the *admitting upper bound on throughput*, assumes that each resident admits the maximum number of patients allowed on each shift, and averages over the shift rotation schedule to obtain the average rate the resident could admit new patients per day. As this second bound ignores that the resident may be unable to admit because they have reached their capacity for total patients in care, we again expect it to overestimate the throughput. We find that the minimum of these two upper bounds is quite far from the true throughput. Additionally, we see that often, for resident schedules with similar throughput upper bounds, the true throughput is very different. This difference is most notable when comparing our long shift and short shift schedules. This motivates our *Markov chain throughput upper bound*, where capacity to admit patients and capacity to care for patients are accounted for jointly. We refer to the bound as a Markov chain bound because to compute it requires computing the steady-state solution of a small, finite state Markov chain, which models the number of patients in care for a single team of residents over one cycle of shift rotations. We compute this upper bound, and find that: (1) it is much closer to the true throughput and (2) when used as a qualitative tool, it correctly ranks different schedules by throughputs, unlike our previous bounds on throughput.

To summarize our results, we found that resident schedules have a dramatic impact on the operational performance of hospitals. Specifically, schedules with shorter more frequent shifts are more capable of handling large volumes of patients. Further, such schedules will have comparable continuity of care in the first 24 hours of patient admission to a traditional long shift schedule if the admitting medical staff are at or near capacity, but may moderately decrease the continuity of care otherwise. We developed a simulation tool to evaluate resident schedules, and simple model using a finite state Markov chain to estimate a schedule's throughput. While the study was conducted with Brigham & Women Hospital data, we expect that similar findings apply to other hospitals with large residency programs.

1.2.3 Queueing Model Results

In order to understand the results from Section 1.2.2, we developed an analytically tractable queueing model of patient flow in a hospital. First we briefly describe our model. Patients arrive according to a non-homogenous Poisson process with rate λ_1 in one half of each day (say 10am till 10pm) and rate $\lambda_2 \leq \lambda_1$ in the remaining part of the day. We consider two stylized policies to schedule the residents, which while significantly simplifying the actual polices, capture the salient features of these policies. The formal description of the two policies is given in Section 6.2. Now we just provide a high level description and discuss their main features.

The first policy we analyze in this paper is the *Long Shifts* (LS) policy. According to this schedule two teams of residents with the same number of residents in each team work for the duration of a day every other day, taking a day off after each day on shift. Namely the first team works on days $2n+1, n \geq 0$ and the second team works on days $2n, n \geq 1$.

The second policy we analyze is called *Daily Admitting* (DA). According to the DA policy, two teams of residents each with the same number of residents as for the LS policy work every day during the high load (10am-10pm) half of the day, and a are off-shift for the other half of the day. Both policies organize the residents into two teams that are offset by a day in the rotating schedule, thus providing uniform coverage. The main distinction between the LS and DA policies is that DA is based on adopting *shorter more frequent* shifts.

The arriving patients are assigned to residents according to the following mechanism used both for the LS and DA schedules. Each resident has an upper bound (capacity) on the number of patients he is allowed to have in care. Each arriving patient is assigned to a resident chosen uniformly at random from all residents on shift who have not reached their capacity (in fact the analysis does not depend on how patients are assigned to available residents for these policies). If all the residents are at capacity at the arrival epoch, the patient joins the queue of unassigned patients and is cared for temporarily by a doctor from a back-up supply of care providers.

We assume that we have an infinite supply of these providers, although using them is undesirable (their use corresponds to a night float or jeopardy admission). These patients are subsequently reassigned to a caring resident on the first come first serve basis, as soon as one of them is available. Patients remain in the hospital for a random exponentially distributed amount of time, beginning from when they are assigned to a resident. We make this assumption because in practice, the newly arriving patients without an assigned resident are stabilized but the treatment plan is not determined until they are assigned to a resident.

We now summarize our results. First, we determine the throughput capacity of each policy. Specifically, for a given number of residents, we compute the maximum arrival rate at which patients can arrive before the queueing system becomes unstable, i.e., the number of unassigned patients grows without bound. We show that DA has a greater throughput capacity than LS, independent of the parameters of our model, supporting our hypothesis from the previous section that using shorter, more frequent shifts will increase capacity.

Next, we compare the number of reassignments under LS and DA (where, as in the previous section, reassignments occur when a patient arrives and there is no resident available to immediately treat them). In comparing policies, we are interested in the expected number of reassignments per day in steady state. As direct steady state analysis appears intractable, we instead resort to the method of fluid approximation of the underlying queueing model. We analyze the long term behavior of the fluid model and show that it converges to the unique steady state solution. The steady state fluid solution carries important information about the long-term performance of the underlying stochastic system. In particular, we prove an interchange of limits result, that the steady state number of patients being treated and the number of patients waiting to be reassigned converges to the steady state fluid solution under the appropriate rescaling. We obtain an implicit formula for the number of each type of reassignment per day in the fluid limit that can be solved numerically (reassignments in the first half of the day correspond to jeopardy reassignments, while reassignments in the second half of the day correspond to night float reassignments). Under minor

technical assumptions, we also prove that the number of steady state reassignments per day in the underlying stochastic model converges in the fluid rescaling to a natural function of the steady state fluid solution, thus justifying fluid approximation. Computing the number of reassignments under each policy from the fluid steady state solution, we find that the DA policy leads to *fewer* reassignments than the LS policy when the load is high, and leads to *more* reassignments than the LS policy when the load is low. Again, our model is consistent with our simulation observations. In particular, using short, more frequent shifts will reduce reassignments if and only if the system is heavily loaded.

1.2.4 Operational & Strategic Insights for Scheduling Medical Residents in Hospitals

In this section, we discuss the implications of our results to practical aspects of medical resident scheduling, both at the operational and strategic levels.

At the operational level, we have two primary contributions. First, we produced a high fidelity simulator of the operations at B&W hospital that can measure the implications of using a resident shift schedule through a large number of metrics. This tool was directly used by B&W to inform both resident scheduling decisions and staffing level decisions in 2011. Additionally, we have developed the Markov chain throughput upper bound that provides a way to quickly estimate the capacity of a medical resident schedule. As the estimates generated by this tool do not depend on any low level details of B&W's operations, we would expect it to be of general use in other hospitals.

At the strategic level, we have shown both through simulation and with analytic models that if staffing levels and total hours available to admit patients are held constant, but shorter more frequent shifts are used, then: (a) capacity will increase, and (b) reassignments will decrease if and only if the system is heavily loaded. Since most hospitals tend to operate at high load, our results lead to the conclusion that the hospitals should consider implementing schedules with shorter more frequent shifts, as

it will increase the capacity to admit patients and reduce the number of reassignments. In this sense the new regulation restricting further the length of shifts should not be perceived as an impediment to efficient handling of patients at hospitals.

Chapter 2

Scalable Algorithms for the Kidney Exchange Problem

2.1 Introduction

The Kidney Exchange Problem (KEP) is defined informally as follows: given an edge-weighted directed graph and a maximum cycle length k, find a maximum weight node-disjoint packing of cycles (of length at most k) and chains. KPD programs solve the KEP to maximize the number of transplants performed. While the KEP has been considered by other authors, existing algorithms have proven inadequate to solve real world instances of the KEP while allowing for "long" chains. We consider two algorithms based on integer programming formulations of the KEP. One of the integer programming formulations considered is similar to a formulation for the Prize Collecting Traveling Salesman Problem (PC-TSP). We give constraint generation schemes to solve these IPs efficiently. We demonstrate the power of our algorithms by running them on clinical data from several KPD programs. Finally, we give algorithms to solve several extensions of the KEP that are of operational and/or theoretical interest. These extensions include: (a) problems with a large maximum cycle length, (b) problems with a large but bounded maximum chain length, and (c) two stage stochastic optimization problems to accommodate for edge failure.

We now briefly explain the relationship between our problem and the PC-TSP.

Recall that in the TSP, one is given a list of cities and the cost of traveling between pairs of cities, and the goal is to find a cycle visiting each city exactly once at the minimum cost [12]. In the PC-TSP, again one must find a cycle visiting each city at most once, but now one has the option of skipping some cities entirely and paying a penalty (see [14], or more recently [42]). Qualitatively, the PC-TSP problem is similar to the KEP in that one wants to find long paths in a graph (which the PC-TSP then closes off as a cycle), without the need to visit every node. It differs in three respects: (1) the solution for the KEP contains multiple disjoint large components (paths), while in the PC-TSP solution contains only a single large component (a cycle), (2) short cycles can be added to a solution for the KEP, (3) in the KEP, many edges are missing from the graph (the TSP variant with missing edges is commonly referred to as the sparse TSP). Despite these differences, we will see that our solution to the KEP is similar to the solution for the PC-TSP.

Notation

We introduce some notation. Let $G = (V, E)$ be a directed graph, and let $\mathbf{w} = (w_1, \ldots, w_{|E|})$ be weights on the edges of G. For each $v \in V$, let $\delta^-(v)$ be the edges pointing to v and $\delta^+(v)$ be the edges outgoing from v. Likewise, for a set of nodes $S \subset V$, let $\delta^-(S)$ be the set of edges into S and $\delta^+(S)$ be the set of edges out of S. For every $S \subset V$, let $E(S)$ be the set of all edges with both endpoints in S. For a set of edges $D \subset E$, let $V(D)$ be the set vertices containing the endpoints of each edge in D. Let $\mathcal{C}$ be the set of all simple cycles in G, where each cycle $C \in \mathcal{C}$ is represented by a collection of edges, i.e. $C \subset E$. Let $\mathcal{C}_k$ be the subset of $\mathcal{C}$ consisting of cycles which use k or fewer edges. For each $v \in V$, let $\mathcal{C}_k(v)$ be the cycles from $\mathcal{C}_k$ containing an edge incident to v. Given a cycle C, let $w_C = \sum_{e \in C} w_e$ be the total weight of the cycle according to our weight vector $\mathbf{w}$.

Suppose two formulations of an integer program are given. Without the loss of generality, assume the underlying problem is of the maximization type. The *linear programming relaxation* of an integer program is the value of the optimal solution obtained when the integrality constraints are removed and the problem is solved as

a linear programming problem. Let Z_1 and Z_2 be the optimal solutions to the linear programming relaxations for two different formulations.

Definition 2.1. If $Z_1 \leq Z_2$ for every problem instance, then formulation one is defined to be *at least as strong as* formulation two. We use the notation $Z_1 \preceq Z_2$. If in addition there exists a problem instance such that $Z_1 < Z_2$, then we say that formulation one is *stronger* than formulation two, and use the notation $Z_1 \prec Z_2$.

Very often in practice, the stronger formulations greatly reduce the actual running time of the integer programming problems [12].

Further, suppose that P_1 and P_2 are the polyhedrons for the linear programming relaxations of our two formulations on the same set of variables. If $P_1 \subset P_2$ for every problem instance, then trivially, $Z_1 \preceq Z_2$.

Organization

In Section 2.2, we define the KEP. In Section 2.3, we give two integer programming based algorithms to solve the KEP. In Section 2.4 we analyze the performance of our algorithms, both empirically using large scale instances of the KEP drawn from clinical data, and theoretically using Definition 2.1. In Section 2.5, we show how to solve the aforementioned extensions of the KEP, and discuss their operational and theoretical relevance. Last, in Section 2.6, we give proofs of some of our the theoretical results.

2.2 Problem Statement

An instance of the KEP is described as follows:

- a list of non-directed donors (NDDs),
- a list of patient-donor pairs (where the donor wants to donate to the paired patient but is not compatible with this patient),
- the compatibility information between all donors and patients
- the "weight" or priority, of each potential transplant.

- a bound k on the maximum cycle length

The goal is then to find a set of transplants, organized into cycles (of length at most k) and chains initiated by NDDs that uses each donor and patient at most once and maximizes the sum of the weights of all transplants performed. If all transplants have weight one, then we are simply trying to find the arrangement which maximizes the total number of transplants. In Section 2.5.2, we will show how this definition can be supplemented to include an optional bound on the maximum chain length.

We now formalize this definition of the KEP in graph theoretic terms. We are given a directed graph $G = (V, E)$, a weight vector $\mathbf{w} = (w_1, \ldots, w_{|E|})$, and a nonnegative integer parameter k. The set of nodes V is partitioned into sets N (the NDDs), and P (the pairs of incompatible donors and patients). For $u, v \in V$, a directed edge from u to v in E indicates that the donor in node u is compatible with the patient in node v. As the nodes of N have no patient, they all must have in-degree zero (although there can be nodes in P with in degree zero as well). The values $w_e \in \mathbb{R}$ for each edge $e \in E$ are weights for the edges, indicating the importance of this transplant, and our goal is to find a maximum weight node disjoint cycle and chain packing, where the cycles can use at most k nodes and the chains must originate from nodes in N. See Figure 1-3 for an example.

2.3 Algorithms for the KEP

In this section, we present two algorithms based on integer programming formulations for the KEP. Both of the IP formulations use an exponential number of constraints. Thus special techniques are required to solve even moderate sized instances of these integer programs, which will be described in subsequent sections.

2.3.1 The Edge Formulation

First, we give a straightforward integer programming formulation of the KEP, with a binary variable for every edge, and constraints so that each node is used at most once and no long cycles occur. The objective is to maximize a weighted number of

edges used. Formally, we use decision variables y_e for $e \in E$ and f_v^i (flow in) and f_v^o (flow out) for $v \in V$, and solve:

$$\begin{aligned}
\max \quad & \sum_{e \in E} w_e y_e & & & (2.1)\\
\text{s.t.} \quad & \sum_{e \in \delta^-(v)} y_e = f_v^i & & v \in V & \\
& \sum_{e \in \delta^+(v)} y_e = f_v^o & & v \in V & \\
& f_v^o \leq f_v^i \leq 1 & & v \in P, & (2.2)\\
& f_v^o \leq 1 & & v \in N, & (2.3)\\
& \sum_{e \in C} y_e \leq |C| - 1 & & C \subset \mathcal{C} \setminus \mathcal{C}_k, & (2.4)\\
& y_e \in \{0, 1\} & & e \in E. &
\end{aligned}$$

Note that we introduce some auxiliary variables f_v^i and f_v^o for all $v \in V$ to simplify the formulation, although since they are defined by the equality constraints, they can be eliminated. In words, (2.2) says that for the patient-donor pair nodes, the flow out is at most the flow in, and the flow in is at most one, (2.3) says that for the NDD nodes, the flow out is at most one, and (2.4), the "cycle inequalities," say that for any cycle $|C|$ of length greater than k, the number of edges we can use is at most $|C| - 1$, thus prohibiting long cycles when $\mathbf{y}$ is integral.

The number of constraints in the IP above is exponential in $|E|$, due to (2.4). As a result, for large instances, we cannot simply enumerate all of these constraints and give them directly to the IP solver. Instead, we use a simple recursive algorithm to solve the problem. First, we relax all the constraints in (2.4) and solve the integer program to optimality. Then we check if the proposed solution contains any cycles of length greater than k. If so, we add the violated constraint from (2.4) and resolve. We repeat this procedure until our solution contains no cycles longer than k. This methodology is generally referred to using "lazy constraints." The technique will generally be successful if few constraints from (2.4) need to be generated on a typical

input. In the worst case, we might have to solve exponentially many IPs, and to solve each IP may take an exponential amount of time (see Section 2.4.2 for a specific example of pathological behavior). However, as we will show in Section 2.4.1, this technique is often quite effective in practice. Finally, note that the efficiency of this procedure relies on the fact that we can very quickly detect if any of the constraints from (2.4) are violated for an integer solution, as we can (trivially) find the largest cycle in a degree two graph in linear time.

2.3.2 The Cutset Formulation

The *Cutset* Formulation is closely related to a standard integer programming solution for the TSP and PC-TSP. The name is derived from the Cutset Formulation of the TSP, as in [12].

For each cycle C of length at most k, we introduce a new variable z_C that indicates if we are using the cycle C. We make the natural updates to (2.1) so the objective value does not change when the same edges are used and to (2.2) so that edges cannot be used both in a z_C variable and a y_e variable. Finally, we add (2.6) to prohibit cycles longer than length k, similarly to (2.4). The formulation is:

$$
\begin{aligned}
\max \quad & \sum_{e \in E} w_e y_e + \sum_{C \in \mathcal{C}_k} w_C z_C & & \\
\text{s.t.} \quad & \sum_{e \in \delta^-(v)} y_e = f_v^i & & v \in V \\
& \sum_{e \in \delta^+(v)} y_e = f_v^o & & v \in V \\
f_v^o + \sum_{C \in \mathcal{C}_k(v)} z_C \leq f_v^i + & \sum_{C \in \mathcal{C}_k(v)} z_C \leq 1 & & v \in P, \qquad (2.5) \\
& f_v^o \leq 1 & & v \in N, \\
& \sum_{e \in \delta^-(S)} y_e \geq f_v^i & & S \subset P,\ v \in S \qquad (2.6) \\
& y_e \in \{0,1\} & & e \in E, \\
& z_C \in \{0,1\} & & C \in \mathcal{C}_k.
\end{aligned}
$$

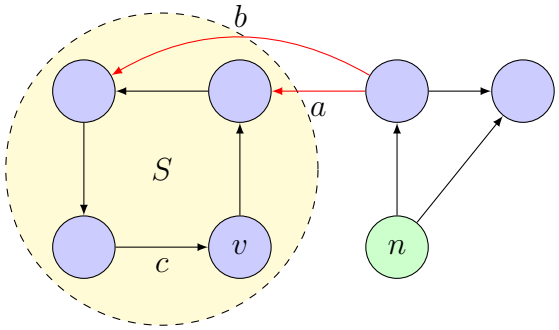

Figure 2-1: An example of a Cutset constraint from (2.6). The graph contains a single NDD in green, labeled n. Observe that if node v is to be involved in any chain (i.e. $f_v^i = 1$), then we must use at least one of the edges a or b that go across the cut separating S from the remaining nodes and NDD.

The constraint (2.6) is very similar to the cutset inequalities for the TSP [12] as adapted to the PC-TSP by several authors (see [42] and the references within). Figure 2-1 provides a clarifying example explaining these constraints. Essentially, they work as follows. Suppose that a chain is reaching some node v, and as a result, f_v^i equals one. Now suppose that we *cut* the graph in two pieces such that half containing v does not contain any of the NDD nodes from N. Since every chain begins at some node in N (and thus does not begin in S), in order for our chain to reach $v \in S$, it must use an edge that begins not in S and ends in S, i.e. and edge $e \in \delta^+(S)$. Thus our constraint requires that whenever there is flow into v, for every way that v can be cut off from the NDDs N, there is at least this much flow over the cut.

Again, the integer programming formulation has exponentially many constraints from (2.6), so we cannot enumerate them and give them all directly to the IP solver. We could simply use the same recursive heuristic ("lazy constraints") from the previous section to obtain a correct algorithm. Instead, our solution still relaxes the constraints (2.6), but more aggressively attempts to find to violated constraints and add them sooner, using a technique called "cutting planes" (or "user cuts"). The method works as follows. The integer programming solver uses the classical branch and bound algorithm to solve the cutset formulation, initially with all of the constraints (2.6) relaxed. However, at every node of the branch and bound tree, the solver checks the fractional solution produced by the LP relaxation for constraints

from (2.6) that are violated, and adds them to the LP as they are encountered. Interrupting the solver to check for violated constraints is a standard feature of many commercial IP solvers (e.g. CPLEX and Gurobi), and is commonly referred to adding user cuts.

To apply this method, we needed an efficient algorithm that given a potentially fractional solution, can either find a violated constraint from (2.6) or determine that none exists. This problem is known in the field of optimization as the *separation problem*.

Theorem 2.1. *The separation problem for* (2.6) *can be solved by solving* $O(|P|)$ *network flow problems.*

A proof of the result is given in Section 2.6. See [12] for more on the separation problem and on the network flow problem. The solution to the separation problem for the Cutset Formulation is very similar to the solutions to the separation problems for the TSP and PC-TSP.

2.4 Algorithm Performance

In this section, we analyze the performance of our two algorithms, both theoretically and empirically using clinical data.

2.4.1 Algorithm Performance on Clinical KPD Data

In this section, we compare the running times of our two algorithms using clinical data from the NKR and APD KPD programs. In Table 2.1, we show the running time of both algorithms on a series of "difficult" but realistic KEP instances encountered in practice. All instances have a maximum cycle length of three. These instances are realistic in that they were taken from the simulations described in Chapter 3, and difficult in that at least one formulation either took a long time solve, or generated a large number of constraints. Thus, these instances represent more of a *worst case* than an *average case* situation for real data.

Instance Info			Running time (s)	
NDDs	Patient-Donor Pairs	Edges	Edge Formulation	Cutset Formulation
3	202	4706	0.18	0.255
10	156	1109	4.425	1.069
6	263	8939	16.186	11.055
5	284	10126	28.063	16.03
6	324	13175	143.432	137.666
6	328	13711	150.877	27.67
6	312	13045	1200*	1200*
10	152	1125	10.388	0.245
3	269	2642	13.896	0.056
10	257	2461	16.206	0.113
7	255	2390	16.7	0.108
6	215	6145	44.101	2.237
10	255	2550	103.112	0.136
1	310	4463	177.582	0.151
11	257	2502	201.154	0.154
6	261	8915	340.312	3.829
10	256	2411	347.791	0.119
6	330	13399	522.619	6.507
10	256	2347	683.949	0.121
7	291	3771	1200*	0.163
8	275	3158	1200*	0.306
4	289	3499	1200*	0.376
3	199	2581	1200*	1.943
7	198	4882	1200*	8.255
2	389	8346	1200*	16.076

Table 2.1: Performance of the Edge and Cutset Formulations, for "difficult" real data KEP instances. Timeouts (optimal solution not found) indicated by *. For instances above the midline, running time for the two algorithms was within an order of magnitude, but for instances below the midline, the Cutset Formulation was at least an order of magnitude faster.

	Instance Info			Edge Formulation		Cutset Formulation	
Instance	NDDs	Paired Nodes	Edges	Time (s)	RAM (GB)	Time (s)	RAM (GB)
APD	47	931	190,820	1.79	1	104	25
NKR	162	1179	346,608	3.074	1	314	37

Table 2.2: Performance of the Edge and Cutset Formulations on very large historical datasets. Performance is measured by running time (in seconds) and RAM consumed (in Gigabytes).

Note that to create snapshots that are representative of actual instances encountered by a KPD program, it is *insufficient* to simply take altruistic donors and donor patient pairs at random from a historical data set, as the patients that are left over after each match run statistically tend to be harder to match than a randomly selected patient (the easy to match patients are more likely to be matched immediately).

Table 2.1 contains the running time of both algorithms on these difficult instances, with a maximum attempted solve time of 20 minutes. In particular, in all instances with the reported running time less than 20 minutes, the optimal solution was found. The instances are separated into two broad groups. In the instances from the top half of the table, the two algorithms took about the same amount of time (to within an order of magnitude). In the instances on the bottom half of the table, the Cutset Formulation was much faster. Looking at the table, we make the following observations in comparing the performance of the two algorithms:

- Both algorithms are able to solve most instances to optimality quickly.
- The Cutset Formulation solves all but one instance to optimality (in fact, this instance was solved after several hours on an independent run).
- The Cutset Formulation is usually faster, although for the easier of these difficult instances, the difference is sometimes negligible.
- On several inputs, the Cutset Formulation is orders of magnitude faster.

We stress again that these instances are the *worst case* instances, in that we only showed results for problems where at least one algorithm had to generate a large number of constraints. These worst case inputs are only a small fraction of all of the simulated inputs, and generally speaking, both algorithms can solve most of these instances to optimality very quickly.

To demonstrate that our algorithms can solve instances even larger than those occurring in current KPD pools, we also ran our algorithms on the entire historical data sets for the KPD programs NKR and APD. Each data set contains around 1000 patients (though arriving over the span of several years), making these instances much larger than the instances described in Table 2.1. The running time for our algorithms on these instances is shown in Table 2.2. We see that:

- Both algorithms can solve both instances.
- The Edge Formulation is much faster.

While the second point may seem surprising, we do have some explanation as to why this is taking place. First and most importantly, these instances are substantially different from the instances that KPDs encounter in practice, in that they do not contain a disproportionately high fraction of hard to match patients. As a result, there is a very large number of two and three cycles, making the number of variables in the Cutset Formulation very large. Second, these instances are both "easy" instances, in that very few of the constraints (2.4) and (2.6) must be added by the algorithm to solve the integer programs, unlike the instances in Table 2.1. As suggested by the results of the previous section, the advantages of using Cutset Formulation over the Edge Formulation depend on the constraints (2.4) and (2.6) being binding in the optimization problem.

For the purposes of comparing algorithms, it would be preferable to have more realistic large scale instances beyond the two described above, but the current historical data does not produce such large scale instances. In an attempt to produce more realistic large scale instances from the historical data set, we experimented with removing fractions of the altruistic donors at random. We found that these did not significantly change in the performance of either algorithm.

2.4.2 Strength of Formulation

In this section, we give a theoretical result comparing our integer programming formulations of the KEP using Definition 2.1. Let Z_{edg} and Z_{cut} be the values of the optimal solutions to the linear programming relaxations of the Edge and Cutset Formulations, respectively.

Theorem 2.2. *The following relationship holds:*

$$Z_{\text{cut}} \prec Z_{\text{edg}}.$$

A proof of the result is given in Section 2.6. It is often the case that integer

programs formulations with stronger linear programming relaxations result in faster running time [12]. This suggests that the Cutset Formulation should solve faster than the Edge Formulation, as we saw was often the case in Section 2.4.1.

In Figure 2-2, we provide an example of a pathological instance of the KEP that the Cutset Formulation can solve orders of magnitude faster than the Edge Formulation. This instance has no altruistic donors, a maximum cycle length of three, and all edges have weight one. Compared to the real instances we considered, this is a very small instance, as it has only 30 nodes and 120 edges. The thirty nodes are arranged in a single large directed cycle, with some additional edges. In particular, each node has an edge pointing to the node two steps ahead and an edge pointing to the node ten steps ahead (along the cycle). For example, node p_2 has edges to p_4 (two ahead) and p_{12} (ten ahead). Also each node and has an incoming edge from the node nine steps ahead (again along the cycle), e.g. node p_2 has an incoming edge from p_{11}. As a result each node has four outgoing and four incoming edges, e.g. the incoming edges to node p_2 are from p_1, p_{30}, p_{11}, and p_{22}, while the outgoing edges are to p_3, p_4, p_{12}, and p_{23} (the incoming edge from p_{22} must be present under our definition because p_{22} has an outbound edge to the node ten ahead, namely p_2, and likewise for the the outgoing edge to p_{23}). This graph has two important properties: there are no cycles of length three or less, but there are a very large number of cycles of length thirty. As a result, there can be no cycles in any feasible solution, and the Cutset Formulation will create no cycle variables (recall the maximum cycle length was assumed to be three). Because there are no non-directed donors or short cycles, the optimal solution is zero. It is easy to see that the linear programming relaxation for the Cutset Formulation is also zero. By considering the cut where $S = P$ and v is arbitrary, as there are no edges going over this cut, the cutset inequality (2.6) implies that the flow into v is zero. As v was arbitrary, we have that the flow in to every node is zero, so the LP relaxation is zero. Thus for this instance, solving the integer programming problem is no harder than solving a single linear programming problem with the Cutset Formulation. Further, given that our proof that the LP was zero only used $|P|$ of the inequalities from (2.6), it is possible to solve the LP

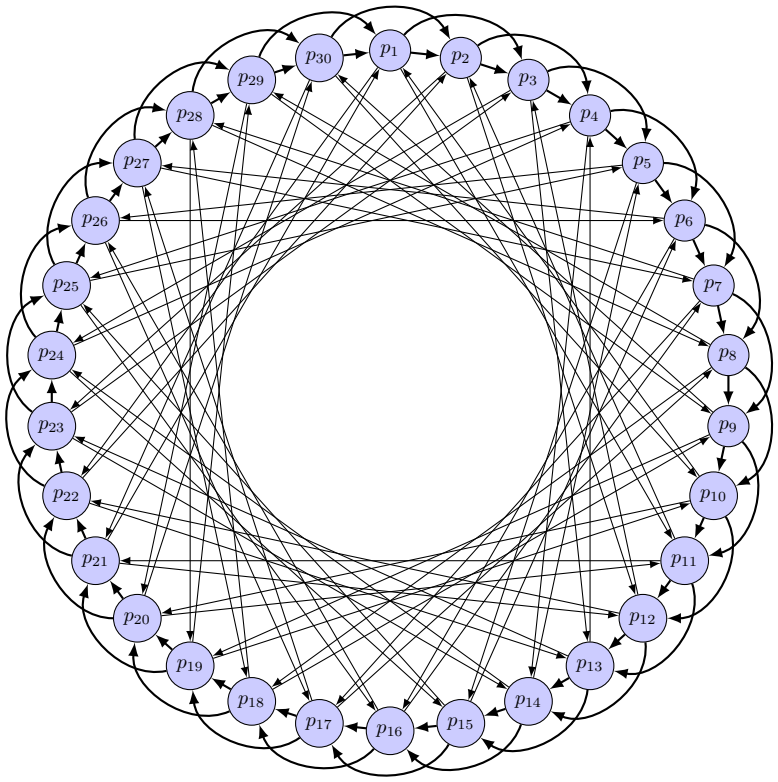

Figure 2-2: A pathological instance of the KEP that is very difficult for the Edge Formulation but is solved trivially by the Cutset Formulation. The optimal solution is zero.

very quickly depending on the strategy used to add violated constraints (in fact, our solver adds all of these constraints in a single round of cut generation). As a result, we are able to solve this instance almost instantly with the Cutset Formulation. In contrast, for Edge Formulation, the linear programming relaxation has the optimal value 29, as it can simply assign value 29/30 to every edge on the length-30 cycle. Worse yet, after two hours of the running time, the best upper bound that the Edge Formulation gives is still 30, as the high redundancy in the structure of the graph results in many possible cycles of length 30. The constraints for each of these cycles needs to be added to obtain an upper bound of 29. For similar reasons, branch and bound is very ineffective.

2.5 Extensions

2.5.1 Long Maximum Cycle Length

In this section, we very briefly discuss how to solve the KEP with a maximum cycle length significantly beyond three or four. The problem can be immediately solved using our Edge Formulation for the KEP with no changes. In fact, for a fixed graph, the Edge Formulation IP will likely become easier as the maximum cycle length increases, as it simply corresponds to using fewer constraints.

However, the Edge Formulation was significantly worse than the Cutset Formulation, both theoretically as defined by strength of formulation, and practically in terms of running time on many instances. We would like to use the Cutset Formulation for these large cycle problems. However, we would quickly run out of memory adding a decision variable for every cycle in the graph, as the number of cycles grows rapidly with the maximum cycle length. Even on a computer with extra memory, adding a large number of variables that would mostly take the value zero would excessively slow down the many linear programs solved internally when solving an IP. One possible solution would be to use the column generation scheme of [1] (see also [23]) to dynamically add variables to model in the middle of the optimization, much like we add the Cutset constraints currently. However, using such a strategy is largely incompatible with using the best commercial integer programming solvers and thus would require significant effort to implement. The investigation of the viability of this approach is a subject for future work.

2.5.2 Bounded Chain Lengths

We show how to adapt the Cutset Formulation to allow for a maximum chain length ℓ. With some additional work, this technique can also be used to adapt the Edge Formulation, although we will not pursue this further. For each NDD $n \in N$ and each edge $e \in E$, we introduce auxiliary edge variables y_e^n and likewise $f_v^{i,n}$ and $f_v^{o,n}$

indicating flow that must begin at n. The formulation becomes:

$$\begin{aligned}
\max \quad & \sum_{e \in E} w_e y_e + \sum_{C \in \mathcal{C}_k} w_C z_C \\
& (\mathbf{y}, \mathbf{z}, \mathbf{f}^i, \mathbf{f}^o) \in P_{\text{cut}} \\
& \sum_{n \in N} y_e^n = y_e && e \in E && (2.7) \\
& \sum_{e \in E} y_e^n \leq \ell && n \in N && (2.8) \\
& \sum_{e \in \delta^-(v)} y_e^n = f_v^{i,n} && v \in V,\ n \in N && (2.9) \\
& \sum_{e \in \delta^+(v)} y_e^n = f_v^{o,n} && v \in V,\ n \in N && (2.10) \\
& f_v^{o,n} \leq f^{i,v} \leq 1 && v \in V,\ n \in N && (2.11) \\
& y_e \in \{0,1\} && e \in E, \\
& z_C \in \{0,1\} && C \in \mathcal{C}_k \\
& y_e^n \in \{0,1\} && e \in E,\ n \in N.
\end{aligned}$$

The new constraints are briefly explained as follows. From (2.7), we have that each edge used (y_e) must be part of a chain beginning at some NDD n. From (2.8), we obtain that each chain can use at most ℓ edges, thus giving the maximum chain length. In (2.9) and (2.10), we just define auxiliary variables denoting if an edge used in a chain starting at n comes into/out of v. Finally, in (2.11), we enforce that the edges used in the chain starting at n are consecutive. The remaining constraints are exactly the same as the PC-TSP constraints with no maximum chain length.

2.5.3 Two Stage Problems

Here we present a general framework for dealing with the possibility that after an edge is selected, it might become ineligible for the matching, an event we refer to as an "edge failure." Edge failures occur commonly in practice for a variety of reasons, e.g. a donor backs out, a patient dies, or a biological incompatibility is discovered.

We propose a two phase system for planning exchanges that anticipates edge failures occurring at random, and plans to maximize the number of transplants performed once the failed edges have been identified and removed. In the first phase, a subset of the edges in the graph are selected to be tested for edge failures. Operational constraints restrict this set, where the basic idea is that it is not practical to check all the edges. Some natural examples of phase one edge sets to test include:

- Use at most m edges in phase one.
- Each node has in degree at most m_i and out degree at most m_o.
- The edges used in phase one must be a feasible solution to the KEP.

The only restriction on the rule used to select phase one edges is that there exists a polyhedron P such that $\mathbf{y} \in P \cap \mathbb{Z}^{|E|}$ iff $\mathbf{y}$ corresponds to a valid set of phase one edges, i.e., the set of phase one edges must be describable as a MIP. After the phase one selections are made, we learn which of the edges that we tested in phase one failed, and in phase two, we solve the regular KEP using only edges that we checked and did not fail in phase one. As we do not know which edges will fail before we make our phase one decision, we use the objective of maximizing the *expected* weight of our phase two KEP solution when picking our phase one solution. Next, we describe the probabilistic framework we use for edge failures, and then the computational technique used to compute our phase one solution.

We assume that there is a family of random variables X_e for $e \in E$, taking the value one if the edge e can be used in the matching (if the edge does not fail) and zero otherwise (if the edge fails). We make no assumptions about the independence structure of the variables X_e. However, we do assume that we can jointly sample the vector of X_e variables.

We now define a two stage stochastic integer optimization problem. We have decision variables y_e for $e \in E$ which indicate the edges we wish to test in stage one. In stage two, we observe our realization $\omega \in \Omega$ of $X_e(\omega)$ for the edges where $y_e = 1$ (the edges we tested), and then we form an optimal cycle and chain packing using only edges that we tested in phase one and where $X_e(\omega) = 1$. We select our phase one edges $\mathbf{y}$, integer and in P, to maximize the expected size of the phase two packing.

This problem can be solved using the method of sample average approximation, as described and mathematically justified in [2, 54, 78]. Suppose that we sample the vector of X_e jointly n times, and let x_e^j for $j = 1, \ldots, n$ be the realization of X_e in the jth sample. Let y_e^j be one if we use edge e in realization j and zero otherwise, and likewise let z_C^j be one if we use cycle C in the jth realization. Let P_{cut}^j be the Cutset polyhedron the variables y_e^j and z_C^j. Our formulation is then as follows:

$$\begin{aligned}
\max \quad & \sum_{j=1}^{n} \left(\sum_{e \in E} c_e y_e^j + \sum_{C \in \mathcal{C}_k} c_C z_C^j \right) && && (2.12)\\
\text{s.t.} \quad & \mathbf{y} \in P, \\
& (\mathbf{y}^j, \mathbf{z}^j) \in P_{\text{cut}}^j, \\
& y_e^j \leq y_e && e \in E, j = 1, \ldots, n, \\
& y_e^j \leq x_e^j && e \in E, j = 1, \ldots, n, \\
& z_C^j \leq y_e && C \in \mathcal{C}_k, e \in C, j = 1, \ldots, n, \\
& z_C^j \leq x_e^j && C \in \mathcal{C}_k, e \in C, j = 1, \ldots, n, \\
& y_e \in \{0, 1\} && e \in E, \\
& y_e^j \in \{0, 1\} && e \in E, j = 1, \ldots, n, \\
& z_C^j \in \{0, 1\} && C \in \mathcal{C}_k, j = 1, \ldots, n.
\end{aligned}$$

This model has a few very attractive features. First, it allows for a general probabilistic model for edge failures, which in practice should be much more accurate than simply i.i.d. edge failures. For example:

- If an edge failed because the donor or receiver became ill or backed out, then all edges involving that donor/receiver would be ruled out simultaneously.
- If an edge failed because a receiver developed a new HLA antibody, then all edges from donors with that HLA antigen incoming to this receiver would fail simultaneously.
- If an edge failed because a doctor or transplant center deemed a donor to be of inadequate quality for the recipient (e.g. the donor was too old), then possibly

other edges pointing to the same doctor/transplant center would fail, but not necessarily all of them, as a highly sensitized recipient may wish to accept a kidney from this donor, while a standard recipient would not.

A salient feature of this model is that we have a great deal of flexibility in choosing P (the set of edges we are allowed to pick in phase one). Our flexibility in choosing P allows us to adapt to various operational constraints of actually running a kidney exchange. Additionally, we can use P to try and influence "agents" (e.g. donors, recipients, doctors, hospitals, and transplant centers) into taking actions that maximize global welfare. For example, if we select more than one incoming edge to a node in phase one, then the receiver, the doctor, the hospital, and the transplant center may be incentivized to reject the worse of the two edges in order to try and get a higher quality donor. One very simple fix is to restrict the edges tested in phase one so that each node has an in-degree of at most one. Then as no one will receive multiple offers, no one will be incentivized to turn down a kidney they otherwise would have accepted.

Finally, note that it is at times desirable to add additional decision variables to the phase one problem. For example, if we were to restrict our phase one solution to be a feasible solution to the KEP, while we could take $P = P_{\text{edg}}$, it is computationally more efficient to use the Cutset Formulation instead. One way of accomplishing this is as follows, add a decision variables $\tilde{y}_e$ for $e \in E$ and $\tilde{z}_C$ for each cycle $C \in \mathcal{C}_k$, let

$$y_e = \tilde{y}_e + \sum_{\substack{C \in \mathcal{C}_k \\ e \in C}} \tilde{z}_C,$$

and then take P to be the PC-TSP polyhedron applied to $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}$, along with the constraint above relating $\mathbf{y}$ to $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}$. Further, note that the Cutset constraints for the P^j_{cut} polyhedrons would automatically be implied by the Cutset constraints from P_{cut} on $(\tilde{\mathbf{y}}, \tilde{\mathbf{z}})$ and thus could be eliminated.

2.6 Proofs

2.6.1 Proof of Cutset Separation

Proof of Theorem 2.1. Let $(\mathbf{y}, \mathbf{z})$ be the point for which we must determine if (2.6) is satisfied. Following the well known procedure for Prize Collecting TSP, first we form a directed weighted graph $\bar{G} = (\bar{V}, \bar{E}, \bar{\mathbf{w}})$ where $\bar{V} = \{s\} \cup V$ where s is an extra node, and $\bar{E} = E \cup \{(s, n) \mid n \in N\}$, and weights $\bar{w}_e$ for $e \in \bar{E}$ are given by

$$\bar{w}_e = \begin{cases} y_e & e \in E, \\ 1 & \text{otherwise}, \end{cases}$$

(the edges with $\bar{w}_e = 1$ each go from the super source to a node in N).

Then for every $v \in P$ where $f_v^i > 0$, we solve the max flow min cut problem with source s and sink v. If we find a cut of weight less than f_v^i, then by taking S to be the set of nodes on the sink side of the cut, we have found a violated constraint. As we are optimizing over all cuts separating v from the super source and then checking all v, we in fact check all the constraints from (2.6). □

2.6.2 Proof of Strength of Formulation

Before proving the result, we introduce two auxiliary integer programming formulations.

The Cycle Formulation

We propose an alternative formulation on the same set of variables as the Cutset Formulation, called the *Cycle* Formulation. In the Cycle Formulation, all of the variables and constraints are the same as the Cutset Formulation except that the constraint (2.6) is replaced by

$$\sum_{e \in C} y_e + \sum_{\substack{D \in \mathcal{C}_k \\ D \neq C}} |D \cap C| z_D \leq |C| - 1 \qquad C \in \mathcal{C}, \tag{2.13}$$

The idea of the constraint is to prevent the edges of C from forming a cycle, unless the variable z_C is used (should the variable exist). Obviously, if $y_e = 1$ for $e \in C$, they would form a cycle, but the constraint prevents this. It would be more intuitive to replace (2.13) with the simpler equation

$$\sum_{e \in C} y_e \leq |C| - 1 \qquad\qquad C \in \mathcal{C}. \tag{2.14}$$

While this would produce a valid formulation of the KEP, we trivially have that this would result in a formulation that is no better (in the sense of Definition 2.1), due to polyhedron containment. In fact, replacing (2.13) by (2.14) results in a formulation that is strictly worse. The example in Figure 2-3 shows an instance of the KEP where if (2.13) is replaced by (2.14), then the LP relaxation of this modified Cycle Formulation is worse than the LP relaxation of the Edge Formulation (without the modification, the LP relaxations of the Edge and Cycle Formulations are the same for this instance).

The Subtour Formulation

Our final formulation, the *Subtour* Formulation, is again on the same set of variables as the Cycle and Cutset Formulations. The name is derived from the Subtour Elimination Formulation of the TSP, as in [12]. In the Subtour Formulation, all of the variables and constraints are the same as the Cutset (and Edge) Formulations except that the constraint (2.6) is replaced by

$$\sum_{e \in E(S)} y_e + \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \subseteq S}} (|D| - 1) z_D + \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \not\subseteq S}} |D \cap E(S)| z_D \leq |S| - 1 \qquad S \subset P. \tag{2.15}$$

The idea of the constraint is that subset S of the nodes, the number of edges used to make chains should be at most $|S| - 1$. Again, in a slightly more intuitive formulation,

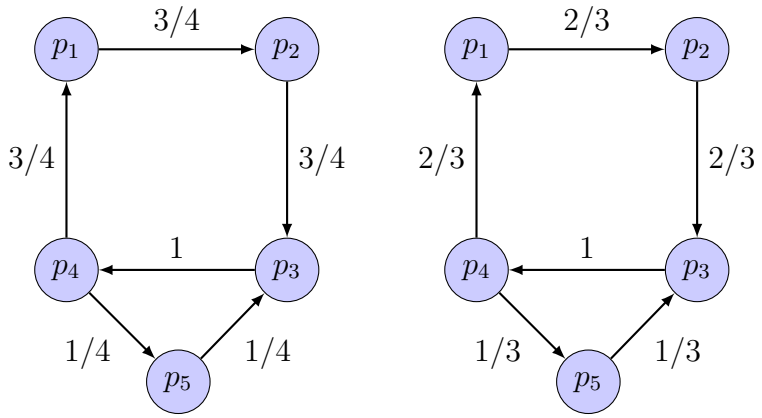

Figure 2-3: The figure above represents two fractional solutions to a single instance of the KEP. The purpose of this example is to demonstrates the necessity of using (2.13) instead of (2.14). In this instance, $P = \{p_1, \ldots, p_5\}$, $N = \emptyset$, the edges are as indicated in the figure above, and all edges have weight one. The numbers next to the edges indicate fractional solutions, namely y_e for the Edge Formulation, and $y_e + \sum_{C \in \mathcal{C}_k, e \in C} z_C$ for the Cycle Formulation. Observe that the solution on the left has greater weight than the solution on the right. The solution on the left is infeasible for the Edge Formulation, as the constraint on the cycle $\{(p_1, p_2), (p_2, p_3), (p_3, p_4), (p_4, p_1)\}$ is violated. The solution on the right is optimal for the Edge Formulation. For the Cycle Formulation, letting the cycle $D = \{(3, 4), (4, 5), (5, 3)\}$, without the second sum from the left hand side of (2.13), we could take $z_D = 1/4$ and $y_e = 3/4$ for $e = (1, 2), (2, 3), (3, 4), (4, 1)$ and then fractional solution on the left would be feasible. This would break the result that $Z_{\text{cyc}} \preceq Z_{\text{edg}}$. However, by including the variable z_D in the constraint against the four cycle, we again have that the solution on the right is optimal for the Cycle Formulation.

we could replace (2.15) by

$$\sum_{e \in E(S)} y_e \leq |S| - 1 \qquad\qquad S \subset P. \tag{2.16}$$

However, this again produces a weaker formulation, as when replacing (2.13) by (2.14). Additionally, observe that for any cycle C, letting $S = V(C)$ be the set of vertices incident to the edges of the cycle (so $|C| = |S|$), then as $C \subset E(S)$,

$$\sum_{e \in C} y_e \leq \sum_{e \in E(S)} y_e \leq |S| - 1 = |C| - 1, \tag{2.17}$$

i.e., (2.16) implies (2.14), so the (weakened) Subtour Formulation polytope is contained in the (weakened) Cycle polytope and thus the (weakened) Subtour Formulation must be at least as strong as the (weakened) Cycle Formulation. Ultimately, we will show a similar result for (2.15) and (2.13).

Proof of Theorem 2.2

Throughout, we let P_{edg}, P_{cyc}, P_{sub}, and P_{cut} be the polyhedrons for the linear programming relaxations of the Edge, Cycle, Subtour and Cutset Formulations of the KEP, respectively. Likewise, we let Z_{edg}, Z_{cyc}, Z_{sub}, and Z_{cut} be the values of the optimal solutions to the linear programming relaxations of these formulations.

Proof of Theorem 2.2. We will instead show

$$Z_{\text{cut}} \prec Z_{\text{sub}} \prec Z_{\text{cyc}} \preceq Z_{\text{edg}}$$

which, as the relations $\prec$ and $\preceq$ are transitive, implies the result.

First, we show that $P_{\text{cut}} \subseteq P_{\text{sub}}$, which immediately implies that $Z_{\text{cut}} \preceq Z_{\text{sub}}$, as the two formulations share the same objective function. It suffices to show that each of the subtour elimination constraints from (2.15) are implied by the entire Cutset Formulation. Fix $S \subset P$, and assume that $\mathbf{y}$ is feasible for the Cutset Formulation.

Fix some $u \in S$. First, we claim that

$$\sum_{\substack{D \in \mathcal{C}_k \\ V(D) \subseteq S}} (|D| - 1) z_D + \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \not\subseteq S}} |D \cap E(S)| z_D$$

$$\leq \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \cap S \neq \emptyset}} (|V(D) \cap S| - 1) z_D \tag{2.18}$$

$$\leq \sum_{\substack{v \in S \\ v \neq u}} \sum_{D \in \mathcal{C}_k(v)} z_D. \tag{2.19}$$

To justify (2.18), observe that for cycles D such that $V(D) \subseteq S$, we immediately have $|D| = |V(D)| = |V(D) \cap S|$, so for these z_D terms, we $|D| - 1 = |V(D) \cap S| - 1$. For D such that $V(D) \not\subseteq S$, we have two cases:

- If $V(D) \cap S = \emptyset$, then $D \cap E(S) = \emptyset$ as well, so these terms can be dropped.
- If D has ℓ vertices in S, where $0 < \ell < |D|$, then at most $\ell - 1$ of the edges of D will have both endpoints in S.

Thus (2.18) has been shown. To justify (2.19), by a simple counting argument, we have that:

- If $u \notin V(D)$, then the term z_D will appear $|V(D) \cap S|$ times in (2.19),
- If $u \in V(D)$, then the term z_D will appear $|V(D) \cap S| - 1$ times in (2.19).

Thus (2.19) has been shown. Applying this inequality, we now have

$$\begin{aligned}
&\sum_{e\in E(S)} y_e + \sum_{\substack{D\in\mathcal{C}_k\\ V(D)\subseteq S}} (|D|-1)z_D + \sum_{\substack{D\in\mathcal{C}_k\\ V(D)\not\subseteq S}} |D\cap E(S)|z_D \\
&\quad\le \sum_{e\in E(S)} y_e + \sum_{\substack{v\in S\\ v\neq u}} \sum_{D\in\mathcal{C}_k(v)} z_D \\
&\quad= \sum_{v\in S} f_v^i - \sum_{e\in\delta^-(S)} y_e + \sum_{\substack{v\in S\\ v\neq u}} \sum_{D\in\mathcal{C}_k(v)} z_D \qquad (2.20)\\
&\quad= f_u^i - \sum_{e\in\delta^-(S)} y_e + \sum_{\substack{v\in S\\ v\neq u}} \left(f_v^i + \sum_{D\in\mathcal{C}_k(v)} z_D \right) \\
&\quad\le \sum_{\substack{v\in S\\ v\neq u}} \left(f_v^i + \sum_{D\in\mathcal{C}_k(v)} z_D \right) \qquad (2.21)\\
&\quad\le |S|-1, \qquad (2.22)
\end{aligned}$$

where (2.20) follows as for a set of nodes S, all edges incoming to a node in S have there other endpoint either in S or outside of S, (2.21) follows from applying (2.6) (multiplied by -1) for the set S and the vertex u, and (2.22) follows from applying the upper bound from flow constraint (2.5) $|S|-1$ times.

Next, we show that $P_{\text{sub}} \subseteq P_{\text{cyc}}$ and thus $Z_{\text{sub}} \preceq Z_{\text{cyc}}$. If suffices to show that for any cycle C, (2.13) is directly implied by (2.15) taking $S = V(C)$. To bound the first term of the left hand side of (2.13), we have

$$\sum_{e\in C} y_e \le \sum_{e\in E(S)} y_e.$$

For the second term, we will partition $D \in \mathcal{C}_k, D \neq C$ into two sets, those where $V(D) \subseteq S$ and $D \neq C$, or those where $V(D) \not\subseteq S$, i.e.,

$$\sum_{\substack{D\in\mathcal{C}_k\\ D\neq C}} |D\cap C|z_D = \sum_{\substack{D\in\mathcal{C}_k\\ V(D)\subseteq S\\ D\neq C}} |D\cap C|z_D + \sum_{\substack{D\in\mathcal{C}_k\\ V(D)\not\subseteq S}} |D\cap C|z_D.$$

For the first sum, we have $|D \cap C| \leq |D| - 1$, as $D \neq C$ (and $D \not\subset C$ since both D and C are simple cycles). Thus

$$\sum_{\substack{D \in \mathcal{C}_k \\ V(D) \subseteq S \\ D \neq C}} |D \cap C| z_D \leq \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \subseteq S \\ D \neq C}} (|D| - 1) z_D \leq \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \subseteq S}} (|D| - 1) z_D$$

For the second sum, as $C \subset E(S)$, we have $|D \cap C| \leq |D \cap E(S)|$ for all $D \in \mathcal{C}_k$, and thus

$$\sum_{\substack{D \in \mathcal{C}_k \\ V(D) \not\subseteq S}} |D \cap C| z_D \leq \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \not\subseteq S}} |D \cap E(S)| z_D.$$

Putting everything together, then applying (2.15) we have

$$\begin{aligned}
&\sum_{e \in C} y_e + \sum_{\substack{D \in \mathcal{C}_k \\ D \neq C}} |D \cap C| z_D \\
&\quad \leq \sum_{e \in E(S)} y_e + \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \subseteq S}} (|D| - 1) z_D + \sum_{\substack{D \in \mathcal{C}_k \\ V(D) \not\subseteq S}} |D \cap E(S)| z_D \\
&\quad \leq |S| - 1 = |C| - 1,
\end{aligned}$$

showing the claim.

To show that $Z_{\text{cyc}} \preceq Z_{\text{edg}}$, consider $(\mathbf{y}^*, \mathbf{z}^*) \in P_{\text{cyc}}$ that is optimal for the Cycle Formulation (the values of f_v^i and f_v^o are implied by $\mathbf{y}^*$). If we let

$$x_e = y_e^* + \sum_{C \in \mathcal{C}_k,\ e \in C} z_C^*,$$

then we claim that $\mathbf{x} \in P_{\text{edg}}$ (again with the values of the flow variables being determined by $\mathbf{x}$). To show this, it suffices to verify (2.2), (2.3) and (2.4) hold for $\mathbf{x}$. To

obtain (2.2), we have

$$\sum_{e\in\delta^+(v)} x_e = \sum_{e\in\delta^+(v)} \left(y_e^* + \sum_{C\in\mathcal{C}_k,\ e\in C} z_C^* \right) \tag{2.23}$$

$$= \sum_{e\in\delta^+(v)} y_e^* + \sum_{C\in\mathcal{C}_k(v)} z_C^* \tag{2.24}$$

where in (2.23) we applied the definition of x_e, and in (2.24), we used that $\mathcal{C}_k(v)$, the set of cycles hitting v, is equal to the disjoint union over all e going out of v of the set of cycles containing e (the union is disjoint as each cycle contains exactly one edge out of v). Likewise, we have

$$\sum_{e\in\delta^-(v)} x_e = \sum_{e\in\delta^-(v)} y_e^* + \sum_{C\in\mathcal{C}_k(v)} z_C^*.$$

Thus (2.5) from the Cycle Formulation implies (2.2) in the Edge Formulation. An analogous argument immediately gives us (2.3) as well. Finally, to obtain (2.4), we have for any cycle C with $|C| > k$,

$$\sum_{e\in C} x_e = \sum_{e\in C} \left(y_e^* + \sum_{\substack{D\in\mathcal{C}_k \\ e\in D}} z_e^* \right)$$

$$= \sum_{e\in C} y_e^* + \sum_{\substack{D\in\mathcal{C}_k \\ D\neq C}} |D\cap C| z_D \tag{2.25}$$

$$\leq |C| - 1, \tag{2.26}$$

where in (2.25), we are counting, and using that $|C| > k$ implies that there is no $D \in \mathcal{C}_k$ such that $D = C$, and in (2.26) we are applying (2.13). Thus we conclude

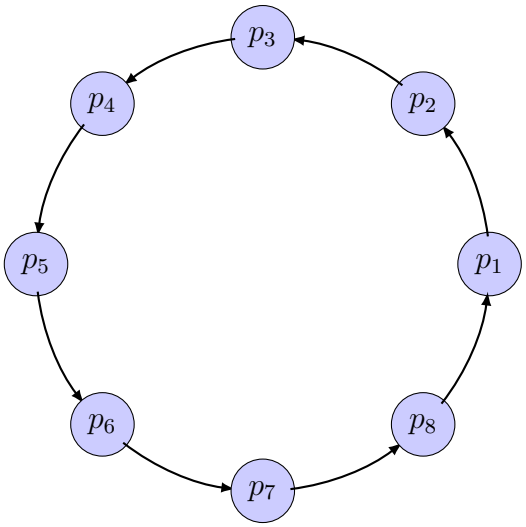

Figure 2-4: Consider the family of problem instances on $n \geq 4$ nodes where $P = \{p_1, \ldots, p_n\}$, $N = \emptyset$, there are n edges forming a single cycle of length n, and $w_e = 1$ for every edge. Above is the instance where $n = 8$. The optimal solution for the IP and the Cutset LP relaxation are both zero, but the Subtour LP relaxation has an optimal solution $n - 1$ (each node has $y_e = (n-1)/n$).

that $\mathbf{x}$ is feasible. Using feasibility, we can obtain the result as follows:

$$\begin{aligned} Z_{\text{edg}} &\geq \sum_{e \in E} c_e x_e \\ &= \sum_{e \in E} c_e \left(y_e^* + \sum_{\substack{C \in \mathcal{C}_k \\ e \in C}} z_e^* \right) \\ &= \sum_{e \in E} c_e y_e^* + \sum_{C \in \mathcal{C}_k} c_C z_e^* \\ &= Z_{\text{cyc}}. \end{aligned}$$

In Figure 2-4, we give a family of problem instances where $Z_{\text{cut}} < Z_{\text{sub}}$. In Figure 2-5 we give an instance where $Z_{\text{sub}} < Z_{\text{cyc}}$. □

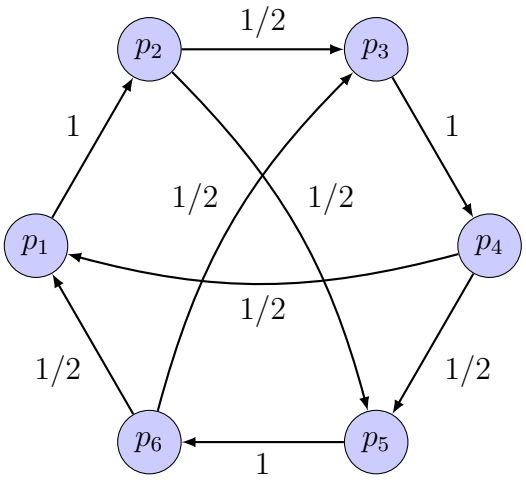

Figure 2-5: In the instance on six nodes above, where $k = 3$, $N = \emptyset$, $P = \{p_1, \ldots, p_6\}$, and each edge has weight one, the IP optimum is zero. Taking y_e to be the edge labels in the figure above, we get a feasible solution to the LP relaxation of $Z_{\text{cyc}} = 6$. However, the LP optimum for the Subtour Formulation is $Z_{\text{sub}} = 5$. We can attain this value by taking $y_{(i,i+1)} = 5/6$ and $y_{(6,1)} = 5/6$. To show that 5 is optimal, we apply the (2.15) taking $S = P$, to obtain that $\sum_{e \in E(P)} y_e \leq 5$, and then observe that $\sum_{e \in E(P)} y_e$ is equal to the objective function.

Chapter 3

Data Driven Simulations for Kidney Exchange

3.1 Introduction

In this chapter, we consider how the types of exchanges performed (e.g. two-cycles, three-cycles, and chains) and the strategy used to select exchanges (e.g. greedy, batching with a match run time of n days) impact aggregate patient outcomes in Kidney Paired Donation (KPD). In particular, we focus on the total transplants performed and the average time patients wait to be transplanted. We quantify these effects by simulating the dynamics of the National Kidney Registry (NKR) KPD pool using historical clinical data. Most importantly, we investigate: (a) the value of forming long chains with altruistic donors, and (b) the trade-off in average patient waiting time when setting the match run time for the batching policies (see Section 1.1 for a discussion of this trade-off).

Organization

This chapter is organized as follows. In Section 3.2, we explain the methods used to perform this simulation analysis. In Section 3.3 we give the results of our simulations. In Section 3.4, we discuss the simulation outcomes and their implications.

3.2 Methods

Data

We simulate the NKR KPD pool over a two year time period from May 24, 2010 to May 24, 2012. We initialize the pool by taking all donors and patients arriving between January 1, 2008 and May 24, 2010, and then removing the donors and patients that NKR had actually matched before May 24th, 2010. This initial pool contains 63 patient-donor pairs, and an additional 410 pairs arrive over the course of the simulation. The dataset contains 75 altruistic donors and 244 patients on the waiting list that have no associated donor. Compatibility between donors and patients is determined primarily by blood type and HLA compatibility rules. Several additional factors were also used including patient preferences (e.g. a patient is unwilling to consider a donor over 60 years old) and previously attempted but failed cross match tests from the NKR database.

Individual Metrics

In this section, we define several metrics designed to quantify the difficulty of matching a patient, donor, or patient-donor pair in kidney exchange. These metrics were first designed by NKR, and we use a slightly modified version.

These metrics are all defined relative to a *population* of patients and donors. Throughout, we take this population to be all donors and patients in a patient-donor pair from the historical NKR dataset. For each patient, the *patient power* is the fraction of donors from the population that are biologically compatible with the patient. Similarly, for each donor, the *donor power* is the fraction of patients from the population that are biologically compatible. For each patient-donor pair, the *pair match power* (PMP) is given by

$$(100 * \text{patient power of patient in pair}) * (100 * \text{donor power of donor in pair}).$$

Finally, we mention that for patients that have more than one willing donor (the

NKR dataset contains several such patients), we need to adapt our definition. We replace the "donor power" term with the following quantity: for a patient with multiple donors, the fraction of patients from the population that at least one donor is compatible with.

Matching Rules and Policies

The *matching rules* specify the types of exchanges that can be made. Typically, we will use the rule that chains of any length and cycles of length two and three are allowed. The *matching policy* is the strategy that is used to determine which exchanges to make (subject to the matching rules).

We will focus on a family of matching policies we refer to as the *batching policies*. Each policy in the family is defined by a single parameter n, referred to as the *match run time*. Exchanges are selected as follows. Patients and donors arrive for a period of n days where no exchanges take place. Then, using all altruistic donors, patient-donor pairs, and waiting list patients currently in the pool, an instance of the KEP (see Chapter 2) is solved find the set of exchanges that maximizes the number of transplants performed (all edges have weight one). The patients and donors that are matched are removed from the pool, and the process is repeated every n days.

An important special case of the batching policies is the *greedy policy*, which every day performs as many exchanges as possible (this is the batching policy with $n = 1$).

Simulation Model

The simulation requires three inputs:

1. A list of patient-donor pairs, altruistic donors, and waiting list patients, each with the date they enter the system,
2. Matching rules to determine which exchanges are allowed,
3. A match run time n to select a batching policy.

The simulation maintains as its state the current date and the current KPD pool of those waiting to be matched. The pool consists of patient-donor pairs, waiting

list patients, and non-directed donors (NDDs). The NDDs are the altruistic donors combined with the bridge donors (see Section 1.1 for more on this distinction). The pool is initialized using NKR's historical data to represent the historical NKR pool for May 24, 2010, and the current date is initialized to May 24th, 2010 as well. The simulation then repeats the following steps until all the patients in the input have arrived:

- Advance the current time n days (e.g. one day for $n = 1$ or one week for $n = 7$).
- Add the patient-donor pairs, altruistic donors, and waiting list patients that have arrived during this n day period to the current pool.
- Solve the KEP to determine a set of transplants to be performed.
- Remove any patient-donor pair that was matched in a cycle from the pool.
- For each chain ending on a waiting list patient, remove every patient-donor pair, waiting list patient, and NDD in the chain from the pool. For each chain ending on a patient-donor pair, remove every patient-donor pair and NDD in the chain from the pool **except the final patient-donor pair in the chain**. The donor from this pair is *converted to an NDD* (sometimes referred to as a bridge donor) and remains in the pool.

System Performance Metrics

To evaluate simulation outcomes, we focus on the following long run performance metrics:

- *Total Matches:* The total number of transplants that were performed during the simulation.
- *Average Waiting Time:* The average over all patient-donor pairs of the following quantity:

$$\min\{\text{match date}, \text{simulation end}\} - \max\{\text{arrival date}, \text{simulation start}\}$$

where for pairs that were not matched by the end of the simulation, their "match date" will be greater than "simulation end".

Intuitively, the quantity *average waiting time* is capturing how much waiting occurred in the simulation window. Importantly, the metric can be used to meaningfully compare outcomes when the number of patient-donor pairs that were matched is different, or even when the number of patient-donor pairs in the simulation is different.

We will also compute these metrics for sub-populations of "hard-to-match" patients. In particular, we consider patient-donor pairs that have a pair match power of less than 20, although this distinction is somewhat arbitrary. For the NKR population of patient-donor pairs, *182 of the 473 patient-donor pairs had a PMP less than 20*, or 38%.

3.3 Simulation Results

Comparison of Batching Policies

In our first experiment we investigate the impact of the match run time on our performance metrics of total transplants and average waiting time. We use the current NKR rule allowing cycles of length up to three, and chains of unbounded length. The results are summarized in Figure 3-1. We see that as the match run time increases from one day to one year, the total number of patient-donor pairs matched increases from 264 to 290. However, we also see that average patient waiting time increases from 158 days to 216 days. Further, we observe for all match run times between a day and a month, the performance is nearly identical. In Figure 3-2, we see that disproportionately many of the additional matches gained by doing batching are for hard-to-match pairs. In particular, when the batch size changes from one day to one year, of the extra 26 transplants gained, 17 are for hard-to-match pairs (pairs with PMP less than 20). The number of hard-to-match pairs matched increases by 33%, while the number of easy-to-match pairs matched increases by only 4%.

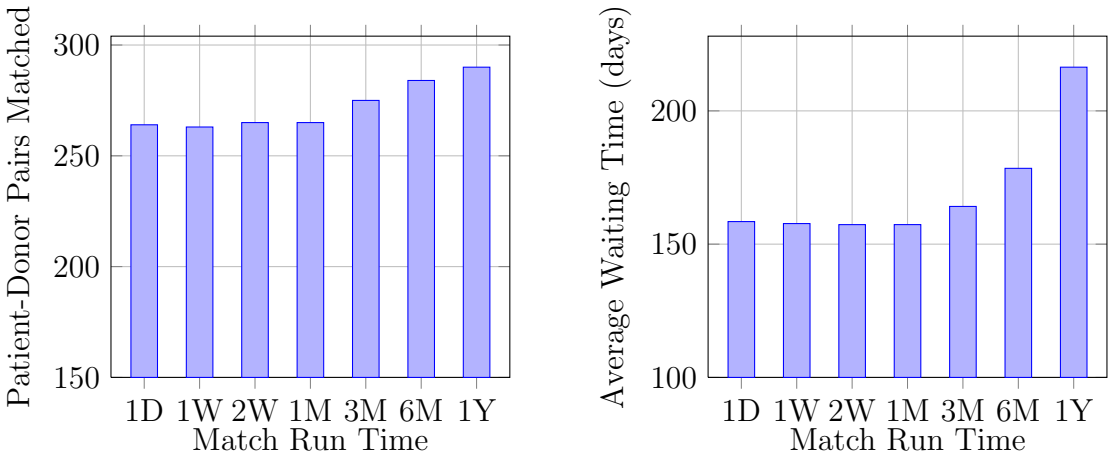

Figure 3-1: Simulation results when using a batching policy and varying the match run time from one day to one year. Left: total patient-donor pairs matched. Right: average waiting time incurred by patient-donor pairs over the course of the simulation.

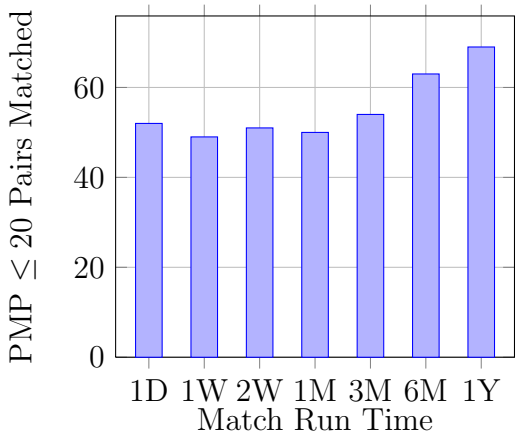

Figure 3-2: The number of hard-to-match patient-donor pairs (pairs with PMP less than 20) matched, as match run time changes from one day to one year.

Measuring the Value of Chains

In this experiment, we consider the implications of a reduced number of altruistic donors. In particular, we run simulations using only 75%, 50%, 25%, and 0% of the altruistic donors from the NKR data set, and see what the impact is on our performance metrics. Throughout, we use the matching rule that cycles can be of length two or three, and that chains can be of any length. However, when we eliminate all of the altruistic donors, we are essentially using the rule that there can be no chains. The only matching policy we use is the greedy policy.

The results are summarized in Figure 3-3. We see that as the number of altruistic donors decreases, the total matches decreases substantially, and the average waiting time increases. In particular, if we eliminate all of the chains, a total of 63 transplants among paired nodes are lost. In addition, another 75 transplants to donors on the waiting list will be lost, as now we cannot end our chains on the waiting list. Thus each altruistic donor on average contributes nearly two transplants when donating through KPD, while by donating directly to the deceased donor waiting list would result in only a single transplant. This quantity should not be interpreted too literally, as we cannot predict what would have happened if the simulation ran over a longer time horizon (perhaps eventually, most of these pairs would have found an exchange in a cycle).

Comparing our other metrics in the cases when we use either all or none of our altruistic donors, we see that the use of chains gives additional positive patients outcomes. Average waiting time for patient-donor pairs decreases by over 50 days when chains are formed. In Figure 3-4, we see that disproportionately many of the additional transplants gained from using the altruistic donors are hard-to-match pairs. In particular, comparing using no altruistic donors and all of the altruistic donors, we increase the number of hard-to-match pairs transplanted by about 100%, while the total number of pairs transplanted increases by only 30%.

Finally, we note that the amount each altruist can contribute depends on the number of patient-donor pairs available to enter an exchange with. As each additional

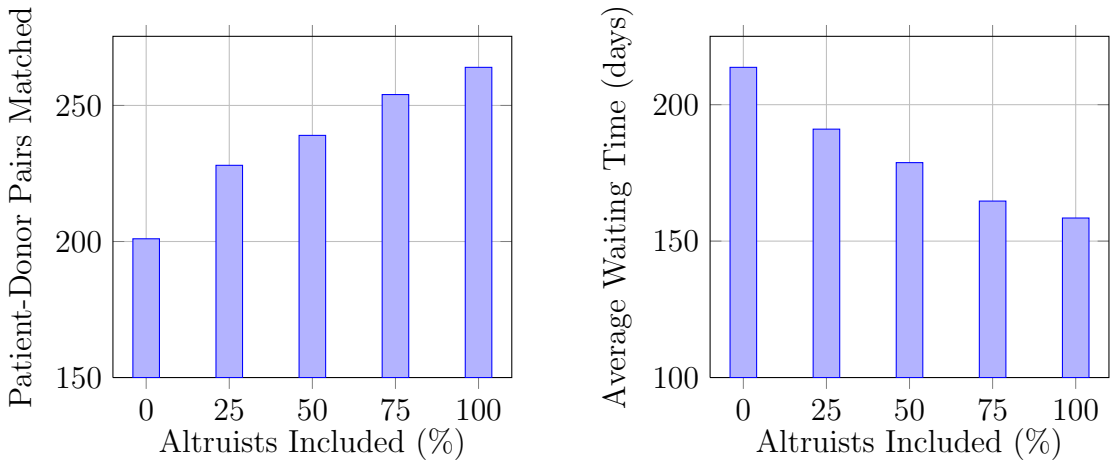

Figure 3-3: Simulation results when using a reduced number of altruistic donors. Left: total patient-donor pairs matched. Right: average waiting time incurred by patient-donor pairs over the course of the simulation.

altruistic donor uses some number of these pairs when forming a chain, fewer remain for subsequent altruists. Thus we would intuitively expect that additional altruistic donors should have diminishing marginal value. While it appears that there may be some evidence of this in Figure 3-3 and Figure 3-4, a more rigorous experimental design would be needed to make any definitive conclusions. We leave this as a topic for future research.

Adjusting the Maximum Cycle Length

In this experiment, we consider the implications of adjusting the maximum cycle length. In particular, we run simulations using a maximum cycle length of zero, two, three, and four. Throughout, we use the matching rule that chains can be of any length, and our matching policy is the greedy policy.

The results are summarized in Figure 3-5. We see that changing the maximum cycle length has very little effect on the total number of matches or average patient waiting time. In Figure 3-6, we see that the number of hard-to-match pairs matched is relatively unaffected by the maximum cycle length as well.

Surprisingly, performance is very slightly better when the maximum cycle length is zero. While such a small improvement is likely to not be statistically significant

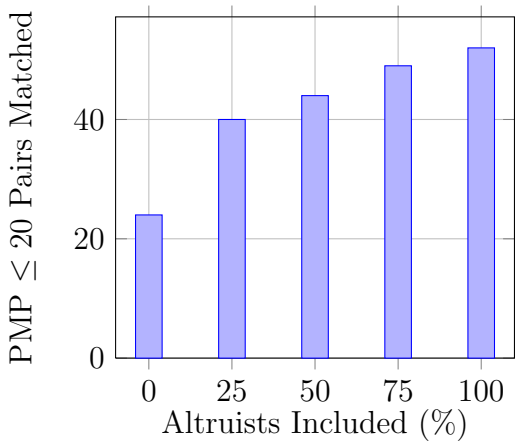

Figure 3-4: The number of hard-to-match patient-donor pairs (pairs with PMP less than 20) matched, as the number of altruistic donors is reduced.

in our experimental setup, a plausible explanation for this behavior is as follows. The most likely two cycle to be formed is between two easy-to-match pairs. Thus by eliminating two-cycles, we gain some easy-to-match pairs to use in chains. These pairs, despite not forming two-cycles with hard-to-match pairs, still link to them. Eventually, when a chain reaches these one of these easy-to-match pairs, we can route the chain to a hard-to-match pair that would otherwise be unreachable, rather than to the easy-to-match pair that would have formed a two-cycle with this pair.

3.4 Discussion

In this section, we summarize the key insights observed in our simulations of the NKR KPD pool. Our findings are as follows:

- *The greedy policy produces nearly as many transplants as any batching policy, and has the lowest average waiting time.*

  First, we note that this result is somewhat surprising, as a priori there is a trade-off in setting the match run time with respect to the average patient waiting time. For a complete discussion, see Section 1.1.

  Second, the batching policies with a match run time between one day and one month all had essentially the same performance. However, as discussed in

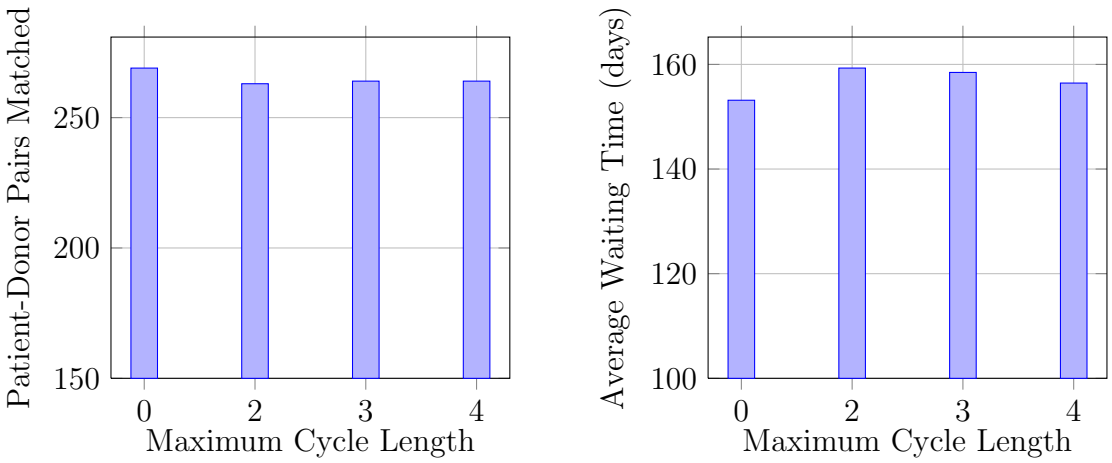

Figure 3-5: Simulation results when varying the maximum cycle length. Left: total patient-donor pairs matched. Right: average waiting time incurred by patient-donor pairs over the course of the simulation.

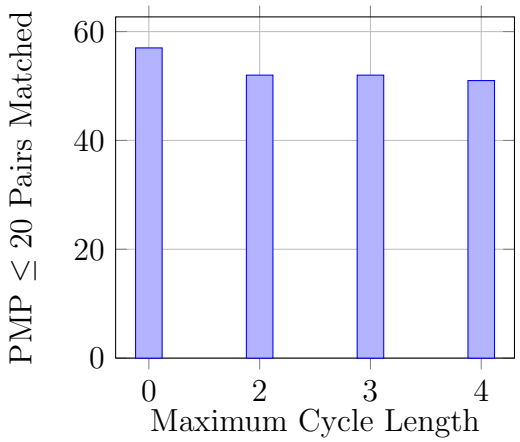

Figure 3-6: The number of hard-to-match patient-donor pairs (pairs with PMP less than 20) matched, as the maximum cycle length is adjusted.

Chapter 1, the greedy policy is of particular interest as it is superior to batching for a variety of practical reasons not captured by our metrics.

- *The use of chains results in significantly more transplants and reduced patient waiting time, and hard-to-match patient-donor pairs are disproportionately the beneficiaries of incorporating chain based exchanges.*
- *According to our performance metrics, there is essentially no benefit in changing the maximum cycle length, given the current number of altruistic donors.*

  We saw that regardless of whether the maximum cycle length was zero, two, three or four, we did approximately the same number of transplants and patients experienced the same average waiting time. If there were fewer altruistic donors, the claim would no longer be true. For (an extreme) example, in the case where there are no altruistic donors, obviously reducing the maximum cycle length to zero would result in the loss of all transplants.

  Despite cycles providing no benefit in our model, we would not recommend eliminating their use completely, as cycles may have benefits not captured by our model. Observe that when we eliminate cycles, since the total number of patient transplants is the same as when using only chains, the chains must on average become longer. As we discuss at some length in Section 1.1.5, our model is missing features (as compared to an actual KPD exchange program) that if incorporated, could potentially make transplanting a patient in a very long chain worse than transplanting a patient in a short cycle (particularly a two cycle).
- *There is very limited room to improve the total number of transplants beyond the level attained by greedy policy.*

  Using a match run time of one year was not seriously considered as an implementable solution. If such a long match run time were used in practice, the rate of patient abandonment (due to sickness) would more than cancel out the increased number of transplants (see Section 1.1.5). Instead, it was supposed to serve as a "upper bound" on the number of transplants that would be attainable for an implementable strategy. We saw that the greedy policy was

able to produce about 90% of the total transplants produced using a match run time of one year, suggesting that there is limited room for any policy to give an improvement over greedy in total matches.

In fact, we can make this notion of an "upper bound" rigorous as follows. Suppose that we take the match run time to be the entire two year time horizon. Then essentially, we have an offline strategy where all of the arrivals are known before any matching decisions are made. Thus we get a bound on the maximum number of transplants achievable by any online matching strategy. It turns out that using a two year match run time gives only 305 transplants, meaning that the greedy policy is within 15% of optimal on total transplants.

Chapter 4

Dynamic Random Graph Models for Kidney Exchange

4.1 Introduction

We consider the problem of efficient operation of a barter exchange platform for indivisible goods. We introduce a dynamic model of barter exchange where in each period one agent arrives with a single item she wants to exchange for a different item. We study a homogeneous and stochastic environment: an agent is interested in the item possessed by another agent with probability p, independently for all pairs of agents. We consider three settings with respect to the types of allowed exchanges: (a) Only two-way cycles, in which two agents swap their items, (b) Two or three-way cycles, (c) (unbounded) chains initiated by altruistic donors who provide an item but expect nothing in return. The goal of the platform is to minimize the average time an agent waits to make an exchange.

In designing a strategy to minimize waiting time, there is a trade-off in determining how quickly feasible swaps should be executed. For example, under a *greedy policy* where swaps are made as soon as they are feasible, agents spend no time waiting to make their exchange beyond the maximum arrival time of any member in the exchange. However, under a *batching policy* where every n days, a set swaps is selected to maximize the number of agents in an exchange, potentially more agents

can be matched. Notice that the greedy policy is essentially equivalent to the batching policy for very small n. We observe the trade-off as n grows: we have the opportunity to match more agents by taking n large, but each agent that is matched must wait some additional time (beyond the maximum arrival time of the agents in their swap) that grows with n for their swap to be executed.

Despite this perceived trade-off, we somewhat surprisingly find that in each of these setting (a), (b) and (c) above, a policy that conducts exchanges in a greedy fashion is near optimal, among a large class of policies that includes batching policies. Further, we find that for small p, allowing three-cycles can greatly improve the waiting time over the two-cycles only setting, and the presence of altruistic donors can lead to a further large improvement in average waiting time. Specifically, we find that a greedy policy achieves an average waiting time of $\Theta(1/p^2)$ in setting a), $\Theta(1/p^{3/2})$ in setting b), and $\Theta(1/p)$ in setting c). Thus, a platform can achieve the smallest waiting times by using a greedy policy, and by facilitating three cycles and chains, if possible.

Our findings are consistent with and provide explanation for empirical and computational observations which compare batching policies in the context of kidney exchange programs.

Organization

This chapter is organized as follows: We describe our model formally in Section 4.2 and state the main results of the chapter in Section 4.3. In Section 4.4, we describe simulation results which support our theoretical findings, and suggest that greedy beats any batching in an absolute, rather than approximate, sense, for each setting we consider. In Section 4.5 we prove our main results for cycles of length two only, in Section 4.6 we prove our results for two and three-cycles (technically the most challenging), and in Section 4.7 we prove our results for chains. We conclude in Section 4.8. Finally, Section 4.9 contains the proofs of some auxiliary results.

Notational Conventions

Throughout, $\mathbb{R}$ ($\mathbb{R}_+$) denotes the set of reals (nonnegative reals). We write that $f(p) = O(g(p))$ where $p \in (0,1]$, if there exists $C < \infty$ such that $|f(p)| \leq Cg(p)$ for all $p \in (0,1]$. We write that $f(p) = \Omega(g(p))$ where $p \in (0,1]$ if there exists $C < \infty$ such that $f(p) \geq Cg(p)$ for all $p \in (0,1]$. Finally, $f(p) = \Theta(g(p))$ if $f(p) = O(g(p))$ and $f(p) = \Omega(g(p)))$. We write that $f(p) = o(g(p))$ where $p \in (0,1]$, if for any $C > 0$, there exists $p_0 > 0$ such that we have $|f(p)| \leq Cg(p)$ for all $p \leq p_0$.

We let Bernoulli(p), Geometric(p), and Bin(n,p), denote a Bernoulli variable with mean p, a geometric variable with mean $1/p$, and a Binomial random variable which is the sum of n independent identically distributed (iid) Bernoulli(p) random variables (r.v.s). We write $X \stackrel{d}{=} \mathcal{D}$ when the random variable X is distributed according to the distribution $\mathcal{D}$. We let ER(n,p) be a *directed* Erdős Réyni random graph with n nodes where every two nodes form a directed edge with probability p, independently for all pairs. We let ER(n,M) be the closely related *directed* Erdős Réyni random graph with n nodes and M directed edges, where the set of edges is selected uniformly at random among all subsets of exactly M directed edges. Unfortunately this notation for the two models makes them nearly indistinguishable. The reader will need to infer from context which model we are referring to, as is common in the literature on random graphs [52]. We let ER(n_L, n_R, p) denote a bipartite *directed* Erdős Réyni random graph with two sides. This graph contains n_L nodes on the left, n_R nodes on the right, and a directed edge between every pair of nodes containing one node from each side is formed independently with probability p. Given a Markov chain $\{X_t\}$ defined on a state space $\mathcal{X}$ and given a function $f\colon \mathcal{X} \to \mathbb{R}$, for $x \in \mathcal{X}$, we use the shorthand

$$\mathbb{E}_x[f(X_t)] \stackrel{\Delta}{=} \mathbb{E}[f(X_t) \mid X_0 = x].$$

4.2 Model

We first state our model for settings with only cyclical exchanges and no chains. Later we augment it to accommodate altruistic/bridge donors and chains.

Consider the following model of a barter exchange where each agent arrives with an item that she wants to exchange for another item. In our simple binary model, each agent is (equally) interested in the items possessed by some of the other agents, and not interested in the items possessed by the rest.

Compatibility graph representation. The state of the system at any time can be represented by a directed graph where each agent is represented by a node, and a directed edge (i, j) exists if agent j wants the item of agent i. Let $\mathcal{G}(t) = (\mathcal{V}(t), \mathcal{E}(t))$ denote the directed graph of compatibilities observed before time t.

Dynamics. Initially the system starts in a state with any finite number of waiting agents. We consider discrete times $t = 0, 1, 2, \ldots$ At each time, one new agent arrives[1]. The new node representing this agent v has an incoming edge from each waiting agent who wants the item of v, and an outgoing edge to each waiting agent whose item v wants.

Stochastic compatibility model. The item of the new agent v is of interest to each of the waiting agents independently with probability p, and independently, the agent v is interested in the item of each waiting agent independently with probability p. Mathematically, there is a directed edge (in each direction) with probability p between the arriving node v and each other node that currently exists in the system, independently for all nodes and directions.

Allocation and policies. An *allocation* in a compatibility graph is a set of disjoint exchanges, namely a set of disjoint cycles and chains. We say that a node that is part of an allocation is *matched*. When an allocation consisting of cycles is executed, the compatibility graph is updated by eliminating the matched nodes and all their incident edges. Immediately after the arrival of a new node, the platform can choose

[1] One can instead consider a stochastic model of arrivals, e.g., Poisson arrivals in continuous time. In our setting, such stochasticity would leave the behavior of the model essentially unchanged, and indeed, each of our main results extend easily to the case of Poisson arrivals at rate 1.

to perform one or more exchanges, based on its chosen *policy*. Here, a policy is a mapping from the history of the system so far to an allocation. An exchange can happen via a cycle, where a *k-way cycle* is a directed cycle in the graph involving k nodes. It can also happen via a chain, which we define below.

Three types of settings (or technologies) are considered, differing by the exchanges permitted in an allocation. In the first two settings, allocations can output only cycles of length at most k, for $k = 2, 3$. These are termed *Cycle Removal*. In the third setting, called *Chain Removal*, allocations consist of only a single chain originating from a bridge node. We augment our model as follows for the chain removal setting.

Altruistic/bridge donors and chains. At the first time period, there is one altruistic donor present in the system, possibly along with other regular agents, and no further altruistic donors arrive to the system later. An altruistic donor is willing to give an item without getting anything in return. We represent an altruistic/bridge donor by a special *bridge* node.[2] Bridge nodes can have only outgoing edges. For a new arrival v, there is an edge from a bridge node to v with probability p, independent of everything else. A *chain* is a directed path that begins with a bridge node. Once a chain is executed by the platform, the last node in a chain becomes a bridge node who can continue the chain in a later period. (All incoming edges to the last node in the chain are eliminated.) Notice that only one bridge donor remains in the system in the system at all times.

One natural policy that will play a key role in our results is the *greedy* policy. The greedy policy attempts to match the maximum number of nodes upon each arrival.

Definition 4.1. The greedy policy for each of the settings is defined as follows:

- Cycle Removal: At the beginning of each time period the compatibility graph does not contain cycles with length at most k. Upon arrival of a new node, if a cycle with length at most k can be formed with the newly arrived node, it is removed, with a uniformly random cycle being chosen if multiple cycles are formed. Clearly, at the beginning of the next time period the compatibility

[2] An example for a bridge node is a non-directed donor in kidney exchange programs.

graph again does not contain any cycles with length at most k. The procedure is described on figure Figure 4-2.

- Chain Removal: There is one bridge node in the system at the beginning of every time period. This bridge node does not have any incoming or outgoing edges. Upon the arrival of a new node at the beginning of a new time interval, the greedy policy identifies an allocation that includes the longest chain originating from the bridge node (breaking ties uniformly at random) and removes these nodes from the system and the last node in the chain becomes a bridge node. Note that such a chain has a positive length if and only if the bridge node has a directed edge from it to the new node. Observe that the new bridge node has in-degree and out-degree zero, so the process can repeat itself. This procedure is described on figure Figure 4-1.

Under each of the settings, the system described above operated under the greedy policy is a Markov chain with a countably infinite number of states, each state corresponding to a compatibility graph, with a bridge node for the second setting, and no bridge nodes for the first setting. Further, this Markov chain is irreducible since an empty graph is reachable from any other state. This raises the question of whether this Markov chain is positive recurrent. If the answer is positive one can further study various performance measures.[3] The performance measure we focus on in this paper is the average (steady state) waiting time, which we define to be the average steady state time interval between the arrival of a node and the time when this node is removed[4]. We also consider policies other than the greedy policy, in general the class of policies under which the system is stationary/periodic and ergodic in the $t \to \infty$ limit. This includes[5] the following class of policies that generalize Markov policies:

[3]The Markov chain turns out to be aperiodic for chain removal and also cycle removal, except for cycle removal with $k = 2$ where it is periodic with period 2. In any case, average (steady state) waiting time, cf. (4.1), is a natural metric for any periodicity.

[4]One may instead consider a cost function that is not linear in waiting time, depending on the intended application. For this first work on dynamic barter exchange, we focus on the simple metric of average waiting time. We remark that our Theorems 4.3 (on chains) and 4.2 (on three-cycles) are scaling results that also hold for any cost function that is bounded above and below by linear functions of waiting time. Similarly, Theorem 4.1 (on two-cycles) leads to a $\Theta(1/p^2)$ scaling result for any cost function of this type.

[5]More precisely, positive recurrent periodic Markov policies (that stabilize the system) lead to a

Definition 4.2. We call a policy a *periodic Markov* policy if it employs τ homogenous first order Markov policies in round robin for some $\tau \in \mathbb{N}$.

In other words, a periodic Markov policy implements a heterogeneous first order Markov chain, where the transition matrices repeat cyclically every τ rounds. Now suppose the resulting Markov chain is irreducible and periodic with period τ'. Without loss of generality, assume that τ is a multiple of τ' (if not, redefine τ as per $\tau \leftarrow \tau\tau'$). Now, clearly the subsequence of states starting with the state at time ℓ and then including states at time intervals of τ, i.e., times $t = \ell, \ell + \tau, \ell + 2\tau, \ldots$ forms an irreducible aperiodic first order Markov chain. If this ℓ-th 'outer' Markov chain is positive recurrent, we conclude that it converges to its unique steady state, leading to a periodic steady state for the original system with period τ. Define

$$W_\ell \equiv \text{Expected number of nodes in the system in the steady state of the } \ell\text{-th outer Markov chain.}$$

Thus, W_ℓ is the expected number of nodes in the system at times that are ℓ mod τ in steady state. Then we define the average waiting time for a periodic Markov policy as

$$W = (1/\tau) \sum_{\ell=0}^{\tau-1} W_\ell \,. \tag{4.1}$$

Note that this is the average number of nodes in the original system over a long horizon in steady state. Recalling Little's law, this is hence identical to the average waiting time for agents who arrive to the system in steady state.

Remark 4.1. We state our results formally for this broad class of periodic Markov policies, though our bounds extend also to other general policies that lead to a stationary/periodic and ergodic system in the $t \to \infty$ limit.

periodic and ergodic system. In any case we are not interested in policies that do not stabilize the system.

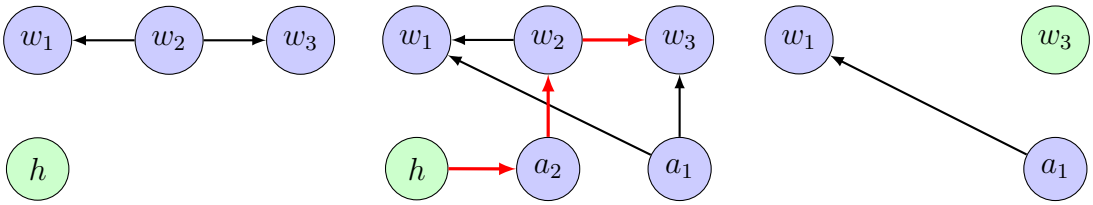

Figure 4-1: An illustration of chain matching under greedy. Initially, h is the head of the chain (the *bridge donor*), and nodes w_1, w_2, and w_3 are waiting to be matched, shown on the left. First, node a_1 arrives, and his good is acceptable by both w_1 and w_3 but no one has a good acceptable by a_1. As h's good is not acceptable by a_1, it is not possible to move the chain. Then node a_2 arrives. His good is acceptable by w_2 and he is able to accept the good from h. The longest possible chain is shown in red in the center above. The chain is formed, h, a_2, and w_2 are removed, and w_3 becomes the new head of the chain (bridge donor). Edges incident to the matched nodes are removed, as well as edges going in to w_3. Note that in this case, the longest chain was not unique; w_1 could have been selected instead of w_3.

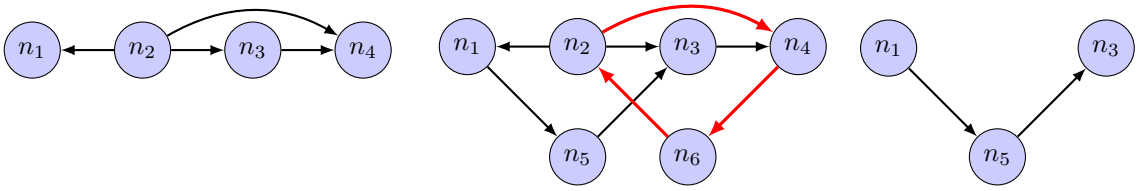

Figure 4-2: An illustration of cycle matching under the greedy policy, with a maximum cycle length of 3. Initially, nodes n_1, n_2, n_3, and n_4 are all waiting, as shown on the left. Node n_5 arrives, but no directed cycles can be formed. Then n_6 arrives, forming the three cycle $n_6 \to n_2 \to n_4 \to n_6$. On the right, the three cycle is removed, along with the edges incident to any node in the three cycle. Note that when n_6 arrives, a six cycle is also formed, but under our assumptions, the maximum length cycle that can be removed is a three cycle.

4.3 Main Results

We consider three different settings: a) two-way cycles only, b) two-way cycles and three-way cycles, and c) unbounded chains initiated by altruistic donors. In each setting we look for a policy that minimizes expected waiting time in steady state.

Two-way cycles only. Our first result considers only 2-way cycles:

Theorem 4.1. *Under the Cycle Removal setting with $k = 2$, the greedy policy (cf. Definition 4.1) achieves an average waiting time of $\ln 2/p^2 + o(1/p^2)$. This is optimal, in the sense that for every periodic Markov policy, cf. Definition 4.2, the average waiting time is at least $\ln 2/(-\ln(1-p^2)) = \ln 2/p^2 + o(1/p^2)$.*

The key fact leading to this theorem is that the prior probability of having a two-cycle between a given pair of nodes is p^2, so an agent needs $\Theta(1/p^2)$ options in order to find another agent with whom a mutual swap is desirable. This result is technically the simplest to establish, but of equal interest in its implications. We prove Theorem 4.1 in Section 4.5.

Two-way cycles and three-way cycles. Our second result considers the case of cycle removals with $k = 3$. Our lower bound in this case applies to a specific class of policies which we now define.

Let $\mathcal{G}$ denote the *global compatibility graph* that includes all nodes that ever arrive to the system, and directed edges representing compatibilities between them.

Definition 4.3. A deterministic policy (under either Chain Removal or Cycle Removal) is said to be *monotone* if it satisfies the following property: Consider any pair of nodes (i, j) and an arbitrary global compatibility graph $\mathcal{G}$ such that the edge (i, j) is present. Let $\bar{\mathcal{G}}$ be the graph obtained from $\mathcal{G}$ when edge (i, j) is removed. Let T_i and T_j be the times of removal of nodes i and j respectively when the compatibility graph is $\mathcal{G}$ and let $T_{ij} = \min(T_i, T_j)$. Then the policy must act in an identical fashion on $\bar{\mathcal{G}}$ and $\mathcal{G}$ for all $t < T_{ij}$, i.e., the same cycles/chains are removed at the same times in each case, up to time T_{ij}. This property must hold for every pair of nodes (i, j) and every possible $\mathcal{G}$ containing the edge (i, j).

A randomized policy is said to be monotone if it randomizes between deterministic monotone policies.

Remark 4.2. Consider the greedy policy for cycle removal defined above. It is easy to see that we can suitably couple the execution of greedy on different global compatibility graphs such that the resulting policy is monotone. The same applies to a batching policy which matches periodically (after arrival of x nodes), by finding a maximum packing of node disjoint cycles and removing them[6].

Note that the class of monotone policies includes a variety of policies in addition to simple batching policies. For instance, a policy that assigns weights to nodes and finds an allocation with maximum weight (instead of simply maximizing the number of nodes matched) is also monotone.

Theorem 4.2. *Under the Cycle Removal setting with $k = 3$, the average waiting time under the greedy policy (cf. Definition 4.1) is $O(1/p^{3/2})$. Furthermore, there exists a constant $C < \infty$ such that, for any monotone policy that is periodic Markov (see Definitions 4.3 and 4.2), the average waiting time is at least $1/(Cp^{3/2})$.*

Theorem 4.2 says that we can achieve a much smaller waiting time with $k = 3$, i.e., two and three-cycle removal, than the removal of two-cycles only (for small p). Further, for $k = 3$ greedy is again near optimal in the sense that no monotone policy can beat greedy by more than a constant factor. Theorem 4.2 is proved in Section 4.6. The proof overcomes a multitude of technical challenges arising from the complex distribution of the compatibility graph at a given time, and introduces several new ideas.

We remark that we could not think of any good candidate policy in our homogeneous model of compatibility that violates monotonicity but should do well on average waiting time. As such, we conjecture (but were unable to prove) that our lower bound on average waiting time applies to arbitrary and not just monotone policies.

The following fact may provide some intuition for the $\Theta(1/p^{3/2})$ scaling of average

[6]Note that such a policy is periodic Markov with a period equal to the batch size.

waiting time[7]: In a static directed Erdős-Rényi graph with (small) edge probability p, one needs the number of nodes n to grow as $\Omega(1/p^{3/2})$ in order to, with high probability, cover a fixed fraction (e.g., 50%) of the nodes with node disjoint two and three cycles[8]. Our rigorous analysis leading to Theorem 4.2 shows that this coarse calculation in fact leads to the correct scaling for average number of nodes in the dynamic system under the greedy policy, and that no monotone policy can do better.

Our result leaves open the case of larger cycles, i.e. $k > 3$, under the greedy, arbitrary monotone and arbitrary general policies. Based on intuition similar to the above, we conjecture that under the Cycle Removal setting with general k, the greedy policy achieves the average waiting time of $\Theta(p^{-\frac{k}{k-1}})$, and furthermore for every policy the average waiting time is lower bounded by $\Omega(p^{-\frac{k}{k-1}})$.

Unbounded chains initiated by altruistic donors. Our final result concerns the performance under the Chain Removal setting.

Theorem 4.3. *Under the Chain Removal setting, the greedy policy (cf. Definition 4.1) achieves an average waiting time of $O(1/p)$. Further, there exists a constant $C < \infty$ such that even if we allow removal of cycles of arbitrary length in addition to chains, for any periodic Markov policy, cf. Definition 4.2, the average waiting time is at least $1/(Cp)$.*

Thus, unbounded chains initiated by altruistic donors allow for a further large reduction in waiting time relative to the case of two-way and three-way cycles, for small p. In fact, as stated in the theorem, removal of cycles of arbitrary length (and chains), with any policy, cannot lead to better scaling of waiting time than that achieved with chains alone. In particular, greedy is near optimal among all periodic Markov policies for chain removal.[9]

[7]Recall that the average number of nodes is the same as the average waiting time, using Little's law.

[8]The expected total number of three cycles is n^3p^3 and the expected number of node disjoint three cycles is of the same order for $n^3p^3 \lesssim n$. We need $n^3p^3 \sim n$ in order to cover a given fraction of nodes with node disjoint three cycles, leading to $n \gtrsim 1/p^{3/2}$. For $n \sim 1/p^{3/2}$, the number of two-cycles is $n^2p^2 \sim 1/p = o(n)$, i.e., very few nodes are part of two-cycles.

[9]One may ask what happens in the setting where chains, two-cycles and three-cycles are all allowed. We argue in Remark 4.3 that, for small p, this setting should be very similar to the setting with chains only.

Theorem 4.3 involves a challenging technical proof presented in Section 4.7.

The intuition for the $\Theta(1/p)$ scaling of waiting time is as follows: Since an agent finds the item of another agent acceptable with probability p, it is not hard to argue that no policy can sustain an expected waiting time that is $o(1/p)$; see our proof of the lower bound in Theorem 4.3 for a formalization of this intuition. On the other hand, under a greedy policy, the chain advances each time a new arrival can accept the item of the bridge donor, which occurs typically at $\Theta(1/p)$ intervals. One might hope that if there are many agents waiting, then typically, the next time there is an opportunity to advance the chain, we will be able to identify a long chain that will eliminate more agents than the number of agents that arrived since the last advancement. Our proof shows that this is indeed the case.

4.4 Computational Experiments

We conducted simulation experiments which measure the average waiting times for nodes under Chain Removal and Cycle Removal with $k = 2$ and $k = 3$. For each of these matching technologies/settings, we simulated the performance of the batching policy with the batch size of x nodes, and compute the results for various values of x. For each scenario, we simulated a time horizon with 3500 arriving nodes, and measured the average number of nodes in the system after the the arrival of the 1000th node. (The first 1000 arrivals serve the role of a "warm-up" period.) 50 trials were conducted for each scenario simulated.

Figure 4-3(a) illustrates that when $p = 0.1$, the greedy policy, which corresponds to the batching policy with the batch size $x = 1$ performs the best among all batch sizes x. In addition, observe the significant difference between average waiting times corresponding to the Chain Removal setting on the one hand and the Cycle Removal setting with $k = 2$ on the other hand[10]. Figures 4-3(b),(c) and (d) provide similar

[10]We see that the difference between waiting times under chain removal and cycle removal with $k = 3$ is less pronounced. One reason for this could be that there are long intervals between consecutive times when a chain can be advanced, leading to a poor constant factor for chain removal. These intervals can be shortened by using non-maximal chains, and this may significantly improve the constant factor.

results for the cases $p = 0.08, 0.06$ and 0.04.

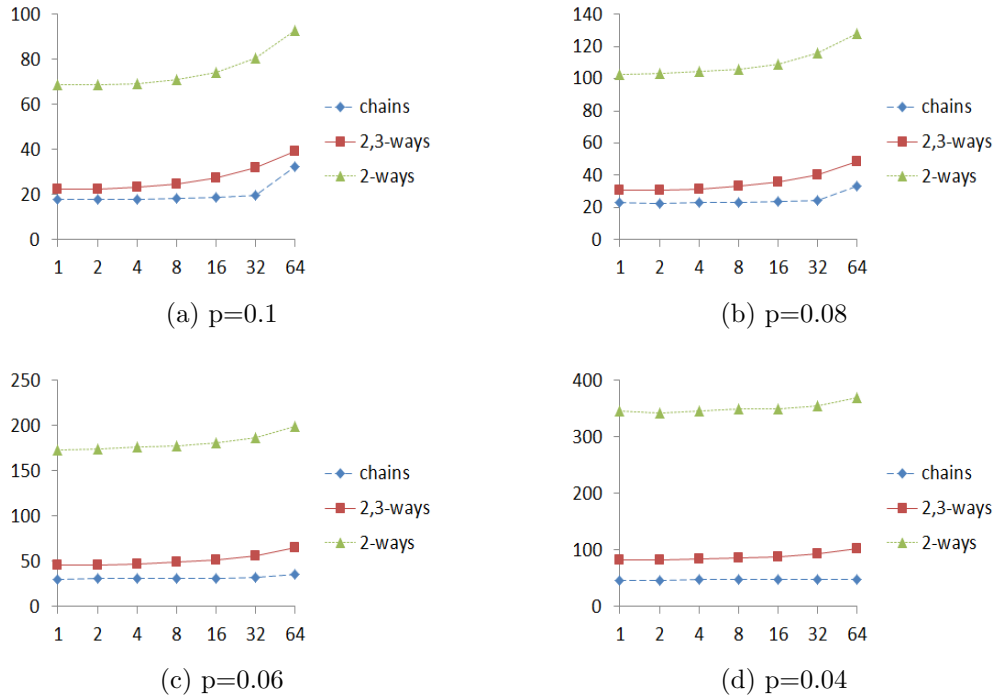

Figure 4-3: Average waiting time under the Chain Removal, Cycle Removal with $k = 2$, and Cycle Removal with $k = 3$, with batching sizes $x = 1, 2, 4, 8, 16, 32, 64$

4.5 Two-way Cycle Removal

In this section we consider Cycle Removal with $k = 2$. The greedy policy corresponding to the Cycle Removal setting when $k = 2$ is simple to characterize, since, as we show below, the underlying process behaves as a simple random walk. We will observe that the random walk has a negative drift when $|\mathcal{V}(t)| \geq \log(2)/p^2$, and obtain a tight characterization of waiting time under greedy using a simple coupling argument. The key idea for the lower bound is that regardless of the implemented policy, the rate at which 2-cycles which will be eventually removed are formed must equal to the half of the rate at which new nodes arrive, which is equal to unity. Further, the probability that we do not form any cycles which will be eventually removed is lower bounded by the probability that we do not form any cycles at all. This probability depends only on the number of nodes in the system, the desired quantity.

Proof of Theorem 4.1. We first compute the expected steady state waiting time under the greedy policy. Observe that for all $t \geq 0$,

$$|\mathcal{V}(t+1)| = \begin{cases} |\mathcal{V}(t)| + 1 & \text{with probability } (1-p^2)^{|\mathcal{V}(t)|}, \\ |\mathcal{V}(t)| - 1 & \text{with probability } 1-(1-p^2)^{|\mathcal{V}(t)|}. \end{cases}$$

Let $\varepsilon > 0$ be arbitrary. If $|\mathcal{V}(t)| > (1+\varepsilon)\ln(2)/p^2$, then there exists a sufficiently small $p = p(\varepsilon)$ such that for all $p > p(\varepsilon)$

$$\mathbb{P}(|\mathcal{V}(t+1)| = |\mathcal{V}(t)| + 1) = (1-p^2)^{|\mathcal{V}(t)|} \leq \frac{1}{2^{1+\varepsilon}}.$$

Let $q = 1/2^{1+\varepsilon} < 1/2$, and let X_t be a sequence of i.i.d. random variables with distribution

$$X_t = \begin{cases} 1 & \text{with probability } q, \\ -1 & \text{with probability } 1-q. \end{cases}$$

Let $S_0 = 0$ and for $t \geq 1$, $S_{t+1} = (S_t + X_t)^+$, so S_t is a Birth-Death process. Letting $r = q/(1-q) < 1$, in steady state $\mathbb{P}(S_\infty = i) = r^i(1-r)$ for $i = 0, 1, \ldots$, so

$$\mathbb{E}[S_\infty] = r/(1-r) = q/(1-2q) = \frac{1}{2^{1+\varepsilon}-2}.$$

We can couple the random walk $|\mathcal{V}(t)|$ with S_t such that $|\mathcal{V}(t)| < (1+\varepsilon)\ln(2)/p^2 + S_t$ for all t. This yields

$$\mathbb{E}[|\mathcal{V}(\infty)|] \leq (1+\varepsilon)\frac{\ln(2)}{p^2} + \mathbb{E}[S_\infty] \leq (1+\varepsilon)\frac{\ln(2)}{p^2} + \frac{1}{2^{1+\varepsilon}-2}.$$

Thus for every $\varepsilon > 0$, we have

$$\lim_{p\to 0} \frac{\mathbb{E}[|\mathcal{V}(\infty)|] - \ln(2)/p^2}{1/p^2} \leq \varepsilon \ln(2).$$

As ε was arbitrary, the result follows.

Now we establish the lower bound on $|\mathcal{V}(\infty)|$. Let v be a newly arriving node at time t, and $\mathcal{W}$ be the nodes currently in system that are waiting to be matched. Let I be the indicator that at the arrival time of v (just before cycles are potentially deleted), no 2-cycles between v and any node in $\mathcal{W}$ exist. Let $\tilde{I}$ be the indicator that at the arrival time of v, no two cycles *that will be eventually removed* that are between v and any node in $\mathcal{W}$ exist (in particular, $\tilde{I}$ depends on the future). Thus $\tilde{I} \geq I$ a.s. Let $\tilde{V}_t$ be the number of vertices in the system before time t such that the cycle which eventually removes them has not yet arrived. We let $\tilde{V}_\infty$ be the distribution of $\tilde{V}_t$ when the system begins in steady state. By stationarity

$$0 = \mathbb{E}[\tilde{V}_{t+1} - \tilde{V}_t] = \mathbb{E}_{\tilde{V}_\infty}[2\tilde{I} - 1],$$

giving $E[\tilde{I}] = 1/2$. Intuitively, in steady state, the expected change in the number of vertices not yet "matched" must be zero. Thus we obtain

$$\frac{1}{2} = \mathbb{E}[\tilde{I}] \geq \mathbb{E}[I] = \mathbb{E}[\mathbb{E}[I \mid |\mathcal{V}(\infty)|]] = \mathbb{E}\left[(1-p^2)^{|\mathcal{V}(\infty)|}\right] \geq (1-p^2)^{\mathbb{E}[\mathcal{V}(\infty)]},$$

by Jensen's inequality. Taking logarithms on both sides and rearranging terms, we get

$$\mathbb{E}[|\mathcal{V}(\infty)|] \geq \frac{\log(1/2)}{\log(1-p^2)} = \frac{\log(2)}{-\log(1-p^2)}.$$

□

4.6 Three-way Cycle Removal

In this section we prove Theorem 4.2. The proof is far more involved than for the case $k = 2$, especially the upper bound, and relies on delicate combinatorial analysis of 3-cycles random graph formed by nodes present in the system in steady state and those arriving over a certain time interval. We consider a time interval of the order $\Theta(1/p^{3/2})$ and assume that the system starts with at least order $\Theta(1/p^{3/2})$ nodes in

the underlying graph. We establish a negative drift in the system and then, as in the case of Chain Removal mechanism, rely on the Lyapunov function technique in order to establish the required upper bound.

For the lower bound, we introduce a novel approach that allows us to prove a matching lower bound (up to constants) for monotone policies by contradiction. The rough idea is as follows: if the steady state expected waiting time is small (in this case smaller than $1/(Cp^{3/2})$ for appropriate C), then a typical new arrival sees a small number of nodes currently in the system, and so typically does not form a two or three-cycle with existing nodes or even the next few arrivals. Thus, the typical arrival typically has a long waiting time, which contradicts our initial assumption of a small expected waiting time.

Preliminaries

We first state a number of propositions and lemmas that will enable our proof of Theorem 4.2. Proofs of these preliminaries are deferred to Section 4.9.

We begin by stating (without proof) the following version of the classical Chernoff bound *(see, e.g. Alon and Spencer 5)*.

Proposition 4.1 (Chernoff bound)**.** *Let $X_i \in \{0,1\}$ be independent with $\mathbb{P}(X_i = 1) = p_i$ for $1 \le i \le n$. Let $\mu = \sum_{i=1}^{n} p_i$.*

(i) For any $\delta \in [0,1]$ we have

$$\mathbb{P}(|X - \mu| \ge \mu\delta) \le 2\exp\{-\delta^2\mu/3\} \tag{4.2}$$

(ii) For any $R > 6\mu$ we have

$$\mathbb{P}(X \ge R) \le 2^{-R} \tag{4.3}$$

Next, we state a straightforward combinatorial bound: In a directed graph, a set $\mathcal{M}$ of node disjoint three-cycles is said to be *maximal* if no three-cycle can be added to $\mathcal{M}$ so that the set remains node disjoint.

Proposition 4.2. *Given an arbitrary directed graph G, let N be the number of three-cycles in a largest in cardinality set of node disjoint three-cycles in G. Then, any maximal set of node disjoint three-cycles consists of at least $N/3$ three-cycles.*

Finally, let $\mathcal{G}_t$ denote the *global compatibility graph* that includes all nodes that ever arrive to the system up to time t, and directed edges representing compatibilities between them. Denote by $\mathcal{W}_t$ the set of nodes out of $0, 1, \ldots, t$ still present in the system at time t. The following is a key property of monotone policies:

Lemma 4.1. *Under any monotone policy, for every two nodes i, j arriving before time t (namely $i, j \leq t$) and every subset of nodes $\mathcal{W} \subset \{0, 1, \ldots, t\}$ containing nodes i and j*

$$\mathbb{P}((i, j) \in \mathcal{G}_t | \mathcal{W}_t = \mathcal{W}) \leq p.$$

In words, pairs of nodes still present in the system at time t are no more likely to be connected at time t than at the time they arrive.

The following corollary follows immediately by linearity of expectations.

Corollary 4.1. *Let $W_t = |\mathcal{W}_t|$ and let E_t be the number of edges between nodes in $\mathcal{W}_t$. Then, under a monotone policy, $\mathbb{E}[E_t | W_t] \leq W_t(W_t - 1)p$.*

Proposition 4.2 and Lemma 4.1 are proved in Section 4.9.

Proof of Theorem 4.2

Proof of Theorem 4.2: the performance of the greedy policy. Suppose at time zero we observe $W \geq C^3/p^{3/2}$ nodes in the system with an arbitrary set of edges between them. Here C is a sufficiently large constant to be fixed later. Call this set of nodes $\mathcal{W}$. Consider the next $T = 1/(Cp^{3/2})$ arrivals, and call this set of nodes $\mathcal{A}$. Wlg, label the times of these arrivals as $1, 2, \ldots, T$, and use the label t for the node that arrives at time t. Let $\mathcal{A}_t \subseteq \{1, 2, \ldots, t-1\}$ be the subset of nodes in $\mathcal{A}$ that have arrived but have not been removed before time t. Similarly define $\mathcal{W}_t$ to be the set

of nodes from $\mathcal{W}$ which are still in the system immediately before time t. Note that, in particular, $\mathcal{W}_1 = \mathcal{W}$.

Let N be the number of three cycles removed during the time period $[0, T]$, that include two nodes from $\mathcal{W}$. Let $\kappa = 1/C^2$ and consider the event

$$\mathcal{E}_1 \equiv \{|\mathcal{A}_{T+1}| - N \geq 2\kappa/p^{3/2}\} \tag{4.4}$$

Introduce the event

$$\begin{aligned}\mathcal{E}_2 \equiv \{&\text{There exists a set of disjoint 2 and 3 cycles in } \mathcal{A} \\ &\text{with cardinality at least } 3/(C^3 p^{3/2}) \ \}.\end{aligned} \tag{4.5}$$

First suppose that the event $\mathcal{E}_1$ does not occur. Then

$$|\mathcal{A}_{T+1}| \leq \frac{2}{C^2 p^{3/2}} + N \leq \frac{1}{16 C p^{3/2}} + N\,, \tag{4.6}$$

for C sufficiently large. Also event $\mathcal{E}_2$ implies that (again for C sufficiently large) at most $9/(C^3 p^{3/2}) \leq 1/(16 C p^{3/2})$ nodes in $\mathcal{A}$ leave due to internal three-cycles or two cycles. Since $T = 1/(C p^{\frac{3}{2}})$, then applying Eq. (4.6), at least $7/(8 C p^{3/2}) - N$ other nodes in $\mathcal{A}$ also leave before $T + 1$. These other nodes belong to cycles of one of the following types:

(i) A three cycle containing another node from $\mathcal{A}$ and a node from $\mathcal{W}$.

(ii) A two cycle with a node from $\mathcal{W}$.

(iii) A three cycle containing two nodes from $\mathcal{W}$. There are exactly N nodes of this type.

Exactly N nodes in $\mathcal{A}$ are removed due to cycles of type (iii) above, so we infer that at least $7/(8 C p^{3/2}) - 2N$ nodes in $\mathcal{A}$ are removed due to cycles of type (i) or (ii) above, meaning that at least $(1/2)(7/(8 C p^{3/2}) - 2N)$ nodes in $\mathcal{W}$ are removed as part of such cycles. Clearly, $2N$ nodes in $\mathcal{W}$ are removed as part of cycles of type (iii).

It follows that

$$|\mathcal{W}_{T+1}| \leq |\mathcal{W}| - 2N - \frac{1}{2}\left(\frac{7}{8Cp^{3/2}} - 2N\right) \leq |\mathcal{W}| - \frac{7}{16Cp^{3/2}} - N\,. \tag{4.7}$$

Combining Eqs. (4.6) and (4.7), we deduce that

$$|\mathcal{W}_{T+1}| + |\mathcal{A}_{T+1}| \leq |\mathcal{W}| - \frac{3}{8Cp^{3/2}}\,. \tag{4.8}$$

We also have that the number of nodes in the system increases by at most $T = 1/(Cp^{3/2})$. We will show that $\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \leq \varepsilon = 1/9$. Before establishing this claim we show how this claim implies the result. We have

$$\mathbb{E}[|\mathcal{W}_{T+1}| + |\mathcal{A}_{T+1}| - |\mathcal{W}|] \leq \varepsilon T - (1-\varepsilon)\frac{3}{8Cp^{3/2}} = -\frac{2}{9Cp^{3/2}}\,, \tag{4.9}$$

i.e., the number of nodes decrease by at least $2/(9Cp^{3/2})$ in expectation. We now apply Proposition A.4 to the embedded Markov chain observed at times which are multiples of T. Namely, let $T_i = i \cdot T$, and take $X_i = \mathcal{G}(T_i) = (\mathcal{V}(T_i), \mathcal{E}(T_i))$ and define $V(X_i) = |\mathcal{V}(T_i)|$. If we let $\mathcal{D}_i$ be the set of nodes that are deleted in some cycle during the time interval $[T_i, T_{i+1})$, we obtain a decomposition

$$V(X_{i+1}) = |\mathcal{V}(T_{i+1})| = |\mathcal{V}(T_i)| + T - |\mathcal{D}_i| = V(X_i) + T - |\mathcal{D}_i|. \tag{4.10}$$

Since $T > 0$ is deterministic it is trivially independent from $\mathcal{G}(T_i)$. Thus the assumptions on decomposing V from (A.8) are satisfied. The assumption that $\{\mathcal{G} \mid V(\mathcal{G}) < n\}$ is finite for every n is satisfied as there are only finitely many graphs with n nodes. We take $\alpha = C^3/p^{3/2}$ making $\mathcal{B} = \{\mathcal{G} \mid |\mathcal{V}(\mathcal{G})| \leq C^3/p^{3/2}\}$. We can take $C_1 = 1$ as T is deterministic. We can take $C_2 = 3$, as trivially $|\mathcal{D}_i| \leq 3T$ since each newly arriving node can be in at most one three cycle (thus making $\tilde{D}_k = D_k$ in Proposition A.4). Finally, we can take $\lambda = 2/9$, as by (4.9),

$$\mathbb{E}[T - |\mathcal{D}_i|] \leq -\frac{2}{9Cp^{3/2}} = -\frac{2}{9}\mathbb{E}[T].$$

Thus by applying Proposition A.4, we obtain that

$$\begin{aligned}\mathbb{E}[|\mathcal{V}(T_\infty)|] &\leq \max\left\{\alpha, \frac{\max\{1, C_2-1\}^2 C_1}{\lambda}\mathbb{E}[A_k]\right\}\left(2+\frac{2}{\lambda}\right)\\ &= \max\left\{\frac{C^3}{p^{3/2}}, \frac{4}{2/9}\frac{1}{Cp^{3/2}}\right\}\left(2+\frac{2}{2/9}\right)\\ &= \frac{11C^3}{p^{3/2}},\end{aligned}$$

for C sufficiently large. Finally, since the embedded chain is observed over deterministic time intervals, the bound above applies to the steady-state bound. We conclude

$$\mathbb{E}[|\mathcal{V}(\infty)|] \leq \frac{11C^3}{p^{3/2}}.$$

It remains to bound $\mathbb{P}(\mathcal{E}_1)$ and $\mathbb{P}(\mathcal{E}_2)$ to complete the proof. We do this below. We claim that $\mathbb{P}(\mathcal{E}_2) \leq \varepsilon/4$. We first show that there is likely to be a maximal set of node disjoint three-cycles in $\mathcal{A}$ of size less than $2/(3C^3p^{3/2})$. This will imply, using Proposition 4.2, that the maximum number of node disjoint three-cycles in $\mathcal{A}$ is at most $2/(C^3p^{3/2})$. Reveal the graph on $\mathcal{A}$ and simultaneously construct a maximal set of node disjoint three-cycles as follows: Reveal node 1. Then reveal node 2. Then reveal node 3 and whether it forms a three-cycle with the existing nodes. If it does remove this three-cycle. Continuing, at any stage t if a three-cycle is formed, choose uniformly at random such a three-cycle and remove it.

Since this process corresponds to a monotone policy (cf. Definition 4.3), then using Corollary 4.1, the residual graph immediately before step t contains no more than $2\binom{t-1}{2}p$ edges in expectation, as the number of nodes is no more than $t-1$. It follows that the conditional probability of three-cycle formation at step t is no more than $\mathbb{E}[\text{Number of three-cycles formed}] = 2\binom{t-1}{2}p^3$. It follows that we can set up a coupling such that the total number of three-cycles removed (this is a maximal set of edge disjoint three-cycles resulting from our particular greedy policy) is no more than $Z = \sum_{t=1}^{T} X_t$ where $X_t \sim \text{Bernoulli}(2\binom{t-1}{2}p^3)$ are independent. Now $\mathbb{E}[Z] = 2\binom{T}{3}p^3 \leq 1/(3C^3p^{3/2})$. Using Proposition 4.1 (i), we obtain that $\mathbb{P}(Z \geq 2/(3C^3p^{3/2})) < \varepsilon/8$, for large enough p, establishing the desired bound on the number of node disjoint three

cycles. We have shown that the probability of having more than $2/(C^3p^{3/2})$ node disjoint three cycles in $\mathcal{A}$ is less than $\varepsilon/8$.

Let Z' be the number of two cycles internal to $\mathcal{A}$. Then $Z' \sim \text{Bin}(\binom{T}{2}, p^2)$. Hence, $\mathbb{E}[Z'] \leq 1/(C^2p)$ and $\mathbb{P}(Z' \geq 1/(C^3p^{3/2})) \leq \varepsilon/8$ for sufficiently small p using Proposition 4.1 (ii). It follows that the probability of having more than $1/(C^3p^{3/2})$ node disjoint two cycles in $\mathcal{A}$ is less than $\varepsilon/8$. Now $\mathbb{P}(\mathcal{E}_2) \leq \varepsilon/4$ follows by union bound.

We now show $\mathbb{P}(\mathcal{E}_1) \leq 3\varepsilon/4$. To prove this, we find it convenient to define two additional events. Denote by $\mathcal{N}(\mathcal{S}_1, \mathcal{S}_2)$ the (directed) neighborhood of the nodes in $\mathcal{S}_1$ in the set of nodes $\mathcal{S}_2$, i.e., $\mathcal{N}(\mathcal{S}_1, \mathcal{S}_2) = \{j \in \mathcal{S}_2 : \exists i \in \mathcal{S}_1 \text{ s.t. } (i,j) \in \mathcal{E}\}$. Abusing notation, we use $\mathcal{N}(i, \mathcal{S})$ to denote the neighborhood of node i in $\mathcal{S}$. Further, we find it convenient to define $\mathcal{B}_t = \mathcal{N}(t, \mathcal{A}_t)$. Define

$$\mathcal{E}_{3,t} \equiv \{|\mathcal{A}_t| \geq \kappa/p^{3/2}, \text{ and } |\mathcal{B}_t| < \kappa/(2p^{1/2})\}, \tag{4.11}$$

and $\mathcal{E}_3 = \cup_{0 \leq t \leq T} \mathcal{E}_{3,t}$. Define

$$\mathcal{E}_{4,t} \equiv \{|\mathcal{A}_t| \geq \kappa/p^{3/2}, \text{ and } |\mathcal{N}(\mathcal{B}_t, \mathcal{W}_t)| < C^3\kappa/(8p)\}, \tag{4.12}$$

and let $\mathcal{E}_4 = \cup_{0 \leq t \leq T} \mathcal{E}_{4,t}$. We make use of

$$\begin{aligned} \mathcal{E}_1 \;&\subseteq\; (\mathcal{E}_4^c \cap \mathcal{E}_1) \cup \mathcal{E}_4 \;\subseteq\; (\mathcal{E}_4^c \cap \mathcal{E}_1) \cup \mathcal{E}_3 \cup (\mathcal{E}_4 \cap \mathcal{E}_3^c) \\ \Rightarrow\; \mathbb{P}(\mathcal{E}_1) \;&\leq \mathbb{P}(\mathcal{E}_4^c \cap \mathcal{E}_1) + \mathbb{P}(\mathcal{E}_3) + \mathbb{P}(\mathcal{E}_4 \cap \mathcal{E}_3^c) \end{aligned}$$

Reveal the edges between t and $\mathcal{A}_t$ when node t arrives. The existence of each edge is independent of the other edges and the current revealed graph. Thus we can bound the probability of the event $\mathcal{E}_{3,t}$ using Proposition 4.1 (i) by $2\exp(-1/(12C^2p^{1/2}))$ for large enough C. It follows that for sufficiently small p, we have

$$\mathbb{P}(\mathcal{E}_3) \leq 2T\exp(-1/(12C^2p^{1/2})) \leq \varepsilon/4\,. \tag{4.13}$$

We now bound $\mathbb{P}(\mathcal{E}_4^c \cap \mathcal{E}_1)$. Let N_t be the number of three cycles removed before time t of type (iii) (recall that type (iii) three cycles include two nodes from $\mathcal{W}$). Define $Z_t \equiv |\mathcal{A}_t| - N_t$. Define

$$\mathcal{E}_{5,t} \equiv \{\text{Node } t \text{ is part of a three-cycle of type (i)}\}.$$

Note that

- If $\mathcal{E}_{5,t}$ then $|\mathcal{A}_{t+1}| = |\mathcal{A}_t| - 1, N_{t+1} = N_t$ if such a three cycle is removed and $|\mathcal{A}_{t+1}| = |\mathcal{A}_t|, N_{t+1} = N_t + 1$ if a three cycle of type (iii) is removed instead. In either case, we have $Z_{t+1} = Z_t - 1$.
- With probability one we have $|\mathcal{A}_{t+1}| \leq |\mathcal{A}_t| + 1$ and $N_{t+1} \geq N_t$. It follows that $Z_{t+1} \leq Z_t + 1$.

Now suppose $Z_t \geq \kappa/p^{3/2}$ and $\mathcal{E}_{4,t}^c$. Clearly $Z_t \geq \kappa/p^{3/2} \Rightarrow |\mathcal{A}_t| \geq \kappa/p^{3/2}$ and hence $\mathcal{E}_{4,t}^c \Rightarrow |\mathcal{N}(\mathcal{B}_t, \mathcal{W}_t)| \geq C^3\kappa/(8p) = C/(8p)$. Revealing the edges between from $\mathcal{N}(\mathcal{B}_t, \mathcal{W}_t)$ to t, we see that

$$\mathbb{P}(\mathcal{E}_{5,t} | Z_t \geq \kappa/p^{3/2}, \mathcal{E}_{4,t}^c) \geq 1 - (1-p)^{C/(4p)} \geq 3/4, \tag{4.14}$$

for large enough C and small enough p, independent of everything so far. So, informally, if $\mathcal{E}_{4,t}^c$ then Z_t is a bounded above by a random walk with a downward drift whenever $Z_t \geq \kappa/p^{3/2}$. We now formalize this.

Define the random walk $(\widetilde{Z}_t)_{t\geq 1}$ as follows: Let $\widetilde{Z}_1 = 0$. Whenever $\widetilde{Z}_t = 0$, we have $\widetilde{Z}_{t+1} = 1$, else

$$\widetilde{Z}_{t+1} = \begin{cases} \widetilde{Z}_t + 1 & \text{w.p. } 1/4 \\ \widetilde{Z}_t - 1 & \text{w.p. } 3/4 \end{cases} \tag{4.15}$$

So $(\widetilde{Z}_t)_{t=1}^{T+1}$ is a downward biased random walk reflected upwards at 0.

Proposition 4.3. *There exists $C < \infty$ such that for any $T \in \mathbb{N}$ and $\nu > 0$, we have $\mathbb{P}(\widetilde{Z}_{T+1} \geq \nu) \leq CT\exp(-\nu/C)$.*

The proof is omitted, as this is a standard result for random walks with a negative

drift. Using Proposition 4.3, we have that for sufficiently small p,

$$\mathbb{P}(\widetilde{Z}_{T+1} \geq \kappa/(2p^{3/2})) \leq \varepsilon/4\,.$$

Let τ be the first time at which event $\mathcal{E}_{4,t}$ occurs for $t \leq T$, and let $\tau = T+1$ if $\mathcal{E}_4$ does not occur. We now show that the following claim holds:

Claim 4.1. *We can couple Z_t and $\widetilde{Z}_t$ such that for all $t < \tau$, whenever $Z_t \geq \kappa/p^{3/2}$ we have $\widetilde{Z}_{t+1} - \widetilde{Z}_t \geq Z_{t+1} - Z_t$.*

Proof of Claim. If $\mathcal{E}_{5,t}$ occurs, then (see above) we know that $Z_{t+1} = Z_t - 1$ and $\widetilde{Z}_{t+1} - \widetilde{Z}_t \geq -1$ holds by definition of $\widetilde{Z}$. Hence, it is sufficient to ensure that $\widetilde{Z}_{t+1} = \widetilde{Z}_t + 1$ whenever $\mathcal{E}_{5,t}^c$ occurs. But this is easy to satisfy since Eq. (4.14) implies that

$$\mathbb{P}(\mathcal{E}_{5,t}^c | Z_t \geq \kappa/p^{3/2}, \mathcal{E}_{4,t}^c) \leq 1/4\,,$$

whereas $\mathbb{P}(\widetilde{Z}_{t+1} = \widetilde{Z}_t + 1) = 1/4$. This completes the proof of the claim. □

The following claim is an immediate consequence:

Claim 4.2. *We have $Z_t \leq \widetilde{Z}_t + \lceil \kappa/p^{3/2} \rceil$ for all $t \leq \tau$.*

Proof of Claim. The claim follows from Claim 4.1 and a simple induction argument. □

It follows that

$$\begin{aligned}\mathbb{P}(Z_{T+1} \geq 2\kappa/p^{3/2}, \tau = T+1) &\leq \mathbb{P}(\widetilde{Z}_{T+1} \geq \kappa/p^{3/2}, \tau = T+1) \\ &\leq \mathbb{P}(\widetilde{Z}_{T+1} \geq \kappa/p^{3/2}) \\ &\leq \varepsilon/4\,.\end{aligned}$$

Thus we obtain

$$\mathbb{P}(\mathcal{E}_4^c \cap \mathcal{E}_1) \leq \varepsilon/4\,. \tag{4.16}$$

Finally, we bound $\mathbb{P}(\mathcal{E}_4 \cap \mathcal{E}_3^c)$. For any $\mathcal{S} \subseteq \mathcal{A}$, let $\mathcal{W}_{\sim\mathcal{S}} \subseteq \mathcal{W}$ be the set of waiting nodes that would have been removed before $T+1$ if (hypothetically) the nodes in $\mathcal{S}$ had no incident edges in either direction, but we left all other compatibilities unchanged. Define the event $\mathcal{E}_6$ as follows: for all $\mathcal{S} \subseteq \mathcal{A}$ such that $|\mathcal{S}| = \kappa/(2p^{1/2})$, the bound

$$|\mathcal{N}(\mathcal{S}, \mathcal{W}\backslash\mathcal{W}_{\sim\mathcal{S}})| \geq C/(8p) \tag{4.17}$$

holds.

Claim 4.3. *The event $\mathcal{E}_6$ occurs whp.*

Before proving the claim, we show that it implies $\mathbb{P}(\mathcal{E}_4 \cap \mathcal{E}_3^c) \leq \varepsilon/4$. Suppose that $\mathcal{E}_6$ and $\mathcal{E}_3^c$ occur. Consider any t such that $|\mathcal{A}_t| \geq \kappa/p^{3/2}$. Since $\mathcal{E}_3^c$, we have that $|\mathcal{B}_t| \geq \kappa/(2p^{1/2})$. Take any $\mathcal{S} \subseteq \mathcal{B}_t$ such that $|\mathcal{S}| = \kappa/(2p^{1/2})$. Notice that for our monotone greedy policy, cf. Remark 4.2, the set of waiting nodes that are removed before time t must be a subset of $\mathcal{W}_{\sim\mathcal{S}}$, i.e., we have we have $\mathcal{W}_t \supseteq \mathcal{W}\backslash\mathcal{W}_{\sim\mathcal{S}}$. Since $\mathcal{E}_6$ occurs, it follows that $|\mathcal{N}(\mathcal{S}, \mathcal{W}_t)| \geq C/(8p) \Rightarrow |\mathcal{N}(\mathcal{B}_t, \mathcal{W}_t)| \geq C/(8p)$. Thus we have $\mathcal{E}_4^c$. This argument just established that

$$\begin{aligned} &\mathcal{E}_6 \cap \mathcal{E}_3^c \subseteq \mathcal{E}_4^c \cap \mathcal{E}_3^c \\ \Rightarrow\ &\mathcal{E}_6^c \cap \mathcal{E}_3^c \supseteq \mathcal{E}_4 \cap \mathcal{E}_3^c\,. \end{aligned}$$

It follows that $\mathbb{P}(\mathcal{E}_4 \cap \mathcal{E}_3^c) \leq \mathbb{P}(\mathcal{E}_6^c \cap \mathcal{E}_3^c) \leq \mathbb{P}(\mathcal{E}_6^c) \leq \varepsilon/4$ using Claim 4.3, as required.

Proof of Claim 4.3. Consider any $\mathcal{S} \subseteq \mathcal{A}$ such that $|\mathcal{S}| = \kappa/(2p^{1/2})$. Clearly, since each node in $\mathcal{A}$ can eliminate at most 2 nodes in $\mathcal{W}$, we have $|\mathcal{W}_{\sim\mathcal{S}}| \leq 2|\mathcal{A}\backslash\mathcal{S}| \leq 2|\mathcal{A}| = 2/(Cp^{3/2})$. It follows that $|\mathcal{W}\backslash\mathcal{W}_{\sim\mathcal{S}}| \geq C^3/p^{3/2} - 2/(Cp^{3/2}) \geq C^3/(2p^{3/2})$ for large enough C. Now notice that by definition $\mathcal{W}_{\sim\mathcal{S}}$ is a function only of the edges between nodes in $\mathcal{W} \cup (\mathcal{A}\backslash\mathcal{S})$, and is independent of the edges coming out of $\mathcal{S}$. Thus, for each node $i \in \mathcal{W}\backslash\mathcal{W}_{\sim\mathcal{S}}$ independently, we have that each node in $\mathcal{S}$ has an edge to i independently w.p. p. We deduce $i \in \mathcal{N}(\mathcal{S}, \mathcal{W}\backslash\mathcal{W}_{\sim\mathcal{S}})$ w.p. $1-(1-p)^{\kappa/(2p^{1/2})} \geq \kappa p^{1/2}/3$

for small enough p, i.i.d. for each $i \in \mathcal{W} \backslash \mathcal{W}_{\sim \mathcal{S}}$. It follows from Proposition 4.1 (i) that

$$|\mathcal{N}(\mathcal{S}, \mathcal{W} \backslash \mathcal{W}_{\sim \mathcal{S}})| \;<\; \frac{C^3}{2p^{3/2}} \cdot \frac{\kappa p^{1/2}}{3} \cdot \frac{3}{4} \;=\; \frac{\kappa C^3}{8p} \;=\; \frac{C}{8p}$$

occurs w.p. at most $2\exp\big\{-(1/4)^2 \cdot C/(6p) \cdot (1/3)\big\} \leq \exp(-C/(300p))$ for small enough p. Now, the number of candidate subsets $\mathcal{S}$ is $\binom{1/(Cp^{3/2})}{\kappa/p^{1/2}} \leq (1/(Cp^{3/2}))^{\kappa/p^{1/2}} \leq \exp(1/p^{\varepsilon+1/2})$ for small enough p. It follows from union bound that $|\mathcal{N}(\mathcal{S}, \mathcal{W} \backslash \mathcal{W}_{\sim \mathcal{S}})| < \frac{C}{8p}$ for one (or more) of these subsets $\mathcal{S}$ with probability at most $\exp(-C/(300p)) \cdot \exp(1/p^{\varepsilon+1/2}) \leq \exp(-C/(400p)) \xrightarrow{p \to 0} 0$. Thus, whp, $|\mathcal{N}(\mathcal{S}, \mathcal{W} \backslash \mathcal{W}_{\sim \mathcal{S}})| < \frac{C}{8p}$ occurs for no candidate subset $\mathcal{S}$, i.e., event $\mathcal{E}_6$ occurs whp. □

Proof of Theorem 4.2: lower bound for monotone policies. Denote by m the expected steady state number of nodes in the system, which by Little's Law equals the expected steady state waiting time. Suppose $m \leq 1/(Cp^{3/2})$, where C is any constant larger than 36. Fix a node i, and reveal the number of nodes W in the system when i arrives. Notice $W \leq 3m$ occurs with probability at least $1 - 1/3 = 2/3$ in steady state by Markov's inequality. Assume that $W \leq 3m$ holds. Let $\mathcal{W}$ denote the nodes waiting in the system when i arrives, and let $\mathcal{A}$ be the nodes that arrive in the next $3m$ time slots after node i arrives. Now, if node i leaves the system within $3m$ time slots of arriving, then i must form a two or three cycle with nodes in $\mathcal{A} \cup \mathcal{W}$. The probability of forming such a cycle is bounded above by

$$\mathbb{E}[\text{Number of two cycles between } i \text{ and } \mathcal{A} \cup \mathcal{W} | W] \tag{4.18}$$

$$+\, \mathbb{E}[\text{Number of three cycles containing } i \text{ and two nodes from } \mathcal{A} \cup \mathcal{W} | W]\,. \tag{4.19}$$

Clearly,

$$\mathbb{E}[\text{Number of two cycles between } i \text{ and } \mathcal{A} \cup \mathcal{W} | W] \leq 6mp^2 \leq 1/C \tag{4.20}$$

for p sufficiently small. To bound the other term we notice that

$$\begin{aligned}&\mathbb{E}[\text{Number of three cycles containing } i \text{ and two nodes from } \mathcal{A}\cup\mathcal{W}|W]\\&= p^2\cdot\mathbb{E}[\text{Number of edges between nodes in } \mathcal{A}\cup\mathcal{W}|W]\end{aligned}\tag{4.21}$$

We use Corollary 4.1 to bound the expected number of edges between nodes in $\mathcal{W}$ at the time when i arrives by $W(W-1)p$ and notice that other compatibilities (j_1, j_2) for $\{j_1, j_2\} \not\subseteq \mathcal{W}$ are present independently with probability p. Hence, we have

$$\mathbb{E}[\text{Number of edges between nodes in } \mathcal{A}\cup\mathcal{W}|W] \le |\mathcal{W}\cup\mathcal{A}|(|\mathcal{W}\cup\mathcal{A}|-1)p \le 6m(6m-1)p\,.$$

Using Eq. (4.21) we infer that

$$\begin{aligned}&\mathbb{E}[\text{Number of three cycles containing } i \text{ and two nodes from } \mathcal{A}\cup\mathcal{W}|W]\\&\le 6m(6m-1)p^3 \le 36/C^2 \le 1/C\end{aligned}\tag{4.22}$$

for $C > 36$. Using Eqs. (4.20) and (4.22) in (4.19), we deduce that the probability of node i being removed within $3m$ slots is no more than $2/C$.

Combining, the unconditional probability that node i stays in the system for more than $3m$ slots is at least $(2/3)(1-2/C) > 1/3$ for large enough C. This violates Markov inequality, implying that our assumption, $m \le 1/(Cp^{3/2})$, was false. This establishes the stated lower bound. □

The following conjecture results if we assume that n_t concentrates, and that typical number of edges in a compatibility graph at time t with n_t nodes is close to what it would have been under an ER(n_t, p) graph.

Conjecture 4.1. *For cycle removal with $k = 3$, the expected waiting time in steady state under a greedy policy scales as $\sqrt{\ln(3/2)}/p^{3/2}+o(1/p^{3/2})$, and no periodic Markov policy (including non-monotone policies) can achieve an expected waiting time that scales better than this.*

Here the constant $\sqrt{\ln(3/2)}$ results from requiring (under our assumptions) that

a newly arrived node forms a triangle with probability 1/3. (Thus the "drift" is zero, as with probability 1/3, we form a triangle which gives a net loss of two nodes, and with probability 2/3, we form no triangle and gain a node.)

Our simulation results, cf. Figure 4-3, are consistent with this conjecture: the predicted expected waiting time for greedy from the leading term $\sqrt{\ln(3/2)}/p^{3/2}$ is $W = 80$ for $p = 0.04$, $W = 43$ for $p = 0.06$, $W = 28$ for $p = 0.08$ and $W = 20.1$ for $p = 0.1$. If proved, this conjecture would be refinement of Theorem 4.2. A proof would require a significantly more refined analysis for both the upper bound and the lower bound.

4.7 Chain Removal

In this section we prove Theorem 4.3. At any time there is one bridge donor in the system. Under a greedy policy, the chain advances when the newly arrived agent can accept the item of the bridge donor. The basic idea to show that greedy achieves $O(1/p)$ waiting time is to show that when there are more than C/p waiting nodes just after we move the chain forward, then, on average, the next time the chain moves forward, it will remove more nodes than were added in the interim. This "negative drift" in number of nodes is crucial in establishing the bound (following which we again use Proposition A.4 to infer a bound on the expected waiting time). The lower bound proof is based on the idea that the waiting time for a node must be at least the time for the node to get an in-degree of one. The lower bound is again proved by contradiction.

Preliminaries

We will need a simple result on the tails of geometric random variables below.

Lemma 4.2. *There exist p_0 and κ_0 such that for all $p < p_0$ and all $\kappa > \kappa_0$, if*

$X \sim \text{Geometric}(p)$, *then*

$$\mathbb{E}\left[X\mathbb{I}_{\{X<\frac{1}{p\sqrt{\kappa}} \text{ or } X>\frac{\kappa}{p}\}}\right] \leq \frac{2}{\kappa p}.$$

The proof is in Section 4.9. Additionally, we need Corollary B.2 from Appendix B showing that there will be a long path in a bipartite directed Erdős-Rényi random graph with high probability. We show this using a result of [3] (see [55] for a more recent reference).

Proof of Theorem 4.3

We introduce the following notation. Let $\mathcal{G}(t) = (\mathcal{V}(t), \mathcal{E}(t), h(t))$ be the directed graph at time t describing the compatibility graph at time t. Here $h(t)$ is a special node not included in $\mathcal{V}(t)$ that is the head of the chain, which can only have out-going edges. We denote by $\mathcal{G}(\infty) = (\mathcal{V}(\infty), \mathcal{E}(\infty), h(\infty))$ the steady-state version of this graph (which exists as we show below).

According to the greedy policy, whenever $h(t)$ forms a directed edge to a newly arriving node, a largest possible chain starting from $h(t)$ is made. Thus before the new node arrives, $h(t+1)$ will always have an in degree and out degree of zero (as explained in Section 4.2), and we can only advance the chain when a newly arriving node has an in edge from $h(t)$. We refer to these periods between chain advancements as *intervals*. Let τ_i for $i = 1, 2, \ldots,$ denote the length of the ith interval, so that $\tau_i \sim \text{Geometric}(p)$. Let $T_0 = 0$ and $T_i = \sum_{j=1}^{i} \tau_j$ for $i = 1, 2, \ldots,$ be the time at the end of the ith interval. Additionally, let $\mathcal{A}_i$ be the set of nodes that arrived during the ith interval $[T_{i-1}, T_i]$, so that $|\mathcal{A}_i| = \tau_i$, and let $\mathcal{W}_i$ be the set of nodes that were "waiting" at the start of the i-th interval, namely at time T_{i-1}. Thus, right before the chain is advanced, every node in the graph is either in $\mathcal{W}_i$, $\mathcal{A}_i$ or it is $h(t)$ itself.

The intuition for the upper bound on waiting time for the greedy policy in Theorem 4.3 is as follows. We will use the Lyapunov function argument (Proposition A.4) to argue that for some C, if there are at least C/p nodes in the graph at the start of an interval $[T_i, T_i + \tau_{i+1}]$, then the number of vertices deleted in an interval

is on average greater than the number of vertices that arrive in that interval, i.e. we have a negative drift on the number of vertices. We lower bound the number of nodes removed in the ith interval by the length of a longest path in the bipartite graph formed by putting nodes in $\mathcal{A}_i$ (the newly arrived nodes) to the left part of the graph, and putting nodes $\mathcal{W}_i$ (the nodes from the previous interval) to the right part of the graph, and maintaining only edges between these two parts (thus in particular preserving the bipartite structure). We bound the expected size of a longest path on this subgraph using Corollary B.2. Observe, that the length of a longest path on our bipartite graph is at most $2|\mathcal{A}_i|$. This will enable us to truncate the downward jumps when applying Proposition A.4.

Proof of Theorem 4.3: performance of the greedy policy. We apply Proposition A.4, taking as our Markov chain $X_i = \mathcal{G}(T_i)$, and our Lyapunov function $V(\cdot)$ to be $V(\mathcal{G}(T_i)) \triangleq |\mathcal{V}(T_i)|$. For a constant $C > 0$ to be specified later, we let α from Proposition A.4 be $\alpha = C/p$. Thus our finite set of exceptions is $\mathcal{B} = \{\mathcal{G} = (\mathcal{V}, \mathcal{E}, h) \colon |\mathcal{V}| \leq C/p\}$, the directed graphs with at most C/p nodes. Obviously our state space is countable and $\mathcal{B}$ is finite. Let $\mathcal{P}_i$ be the path of nodes that are removed from the graph in the ith interval. Thus

$$|\mathcal{V}(T_i)| = |\mathcal{V}(T_{i-1})| + |\mathcal{A}_i| - |\mathcal{P}_i|.$$

By taking $A_i = |\mathcal{A}_i| = \tau_i$ and $D_i = |\mathcal{P}_i|$, we have that $V(\cdot)$ satisfies the form of (A.8) and the independence assumptions on A_i and D_i. As $\tau_i \sim$ Geometric(p), $\mathbb{E}[|\mathcal{A}_i|^2] \leq 2/p^2 = 2\mathbb{E}[|\mathcal{A}_i|]^2$, so we can take $C_1 = 2$. We set $C_2 = 2$, and so to apply Proposition A.4, we must find $\lambda > 0$ such that for every graph $\mathcal{G} \notin \mathcal{B}$,

$$\mathbb{E}_{\mathcal{G}}\left[|\mathcal{A}_i| - \min\{|\mathcal{P}_i|, 2|\mathcal{A}_i|\}\right] \leq -\lambda \mathbb{E}[|\mathcal{A}_i|], \tag{4.23}$$

where $\mathbb{E}_{\mathcal{G}}[\cdot]$ denote the expectation conditioned on the event $\mathcal{G}_{i-1} = \mathcal{G}$. We create an auxiliary bipartite graph $\mathcal{G}'_i = (\mathcal{A}_i, \mathcal{W}_i, \mathcal{E}'_i)$, where $\mathcal{A}_i$ are the nodes on the left, $\mathcal{W}_i$ are the nodes on the right, and $\mathcal{E}'_i$ is the subset of $\mathcal{E}(T_i)$ consisting of edges (u, v)

such that either $u \in \mathcal{A}_i, v \in \mathcal{W}_i$ or vice versa (thus ensuring $\mathcal{G}_i'$ is bipartite). We let $v_i' \in \mathcal{A}_i$ be the node newly arrived at T_i that $h(T_i - 1)$ connected to. Finally, we let $\mathcal{P}_i'$ be the longest path in $\mathcal{G}_i'$ starting at v_i'. Trivially, $|\mathcal{P}_i'| \leq |\mathcal{P}_i|$. Observe that $\mathcal{G}_i'$ is a $\mathrm{ER}(|\mathcal{A}_i|, |\mathcal{W}_i|, p)$ bipartite random graph. Thus we apply Corollary B.2 to show $|\mathcal{P}_i'|$ is appropriately large with high probability. In particular, given arbitrary $\varepsilon > 0$ and $\kappa > \kappa_0 > 1$, where κ_0 is to be specified later, find C and p_0 according to Corollary B.2. Then $\mathcal{G}(T_{i-1}) \notin B$ implies $|\mathcal{W}_i| = |\mathcal{V}(T_{i-1})| \geq C/p$. Then if $p < p_0$ and $a \in \left[\frac{1}{p\sqrt{\kappa}}, \frac{\kappa}{p}\right]$, then by Corollary B.2,

$$\mathbb{P}\left(|\mathcal{P}_i'| < 2|\mathcal{A}_i|(1-\varepsilon) \,\middle|\, |\mathcal{A}_i| = a\right) \leq \varepsilon. \tag{4.24}$$

We define the events E_i and F_i by

$$E_i = \left\{\mathcal{A}_i \notin \left[\frac{1}{p\sqrt{\kappa}}, \frac{\kappa}{p}\right]\right\} \qquad F_i = \left\{|\mathcal{P}_i'| < 2|\mathcal{A}_i|(1-\varepsilon)\right\}.$$

We define $Z_i \triangleq 2|\mathcal{A}_i|(1-\varepsilon)\mathbb{I}_{E_i^c \cap F_i^c}$. Thus $Z_i \leq |\mathcal{P}_i'| \leq |\mathcal{P}_i|$ from the definition of the event F_i, and $Z_i \leq 2|\mathcal{A}_i|$ by construction. We now use this to get an upper bound (4.23) as follows. First, we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{G}}\left[|\mathcal{A}_i| - \min\{|\mathcal{P}_i|, 2|\mathcal{A}_i|\}\right] &\leq \mathbb{E}_{\mathcal{G}}\left[|\mathcal{A}_i| - Z_i\right] \\ &= \mathbb{E}\left[|\mathcal{A}_i|\mathbb{I}_{E_i}\right] + \mathbb{E}\left[|\mathcal{A}_i| - Z_i \,\middle|\, E_i^c\right]\mathbb{P}(E_i^c), \end{aligned} \tag{4.25}$$

where in (4.25), we used that Z_i is zero on E_i. Now noting that for all $a \in \left[\frac{1}{p\sqrt{\kappa}}, \frac{\kappa}{p}\right]$, i.e. in the event E_i^c, we have

$$\mathbb{P}\left(Z_i = 0 \,\middle|\, |\mathcal{A}_i| = a\right) = \mathbb{P}\left(F_i \,\middle|\, |\mathcal{A}_i| = a\right) \leq \varepsilon,$$

by (4.24), and therefore

$$\mathbb{P}\left(Z_i = 2(1-\varepsilon)|\mathcal{A}_i| \,\middle|\, |\mathcal{A}_i| = a\right) = \mathbb{P}\left(F_i^c \,\middle|\, |\mathcal{A}_i| = a\right) \geq 1 - \varepsilon.$$

as well. We now compute that

$$
\begin{aligned}
&\mathbb{E}\left[|\mathcal{A}_i| - Z_i \,\middle|\, E_i^c\right] \\
&= \sum_{a \in \left[\frac{1}{p\sqrt{\kappa}}, \frac{\kappa}{p}\right]} \mathbb{E}\left[|\mathcal{A}_i| - Z_i \,\middle|\, |\mathcal{A}_i| = a\right] \mathbb{P}\left(|\mathcal{A}_i| = a \,\middle|\, E_i^c\right) \\
&= \sum_{a \in \left[\frac{1}{p\sqrt{\kappa}}, \frac{\kappa}{p}\right]} \mathbb{E}\left[|\mathcal{A}_i| - Z_i \,\middle|\, F_i \cap |\mathcal{A}_i| = a\right] \mathbb{P}\left(F_i \,\middle|\, |\mathcal{A}_i| = a\right) \mathbb{P}\left(|\mathcal{A}_i| = a \,\middle|\, E_i^c\right) \\
&\qquad + \sum_{a \in \left[\frac{1}{p\sqrt{\kappa}}, \frac{\kappa}{p}\right]} \mathbb{E}\left[|\mathcal{A}_i| - Z_i \,\middle|\, F_i^c \cap |\mathcal{A}_i| = a\right] \mathbb{P}\left(F_i^c \,\middle|\, |\mathcal{A}_i| = a\right) \mathbb{P}\left(|\mathcal{A}_i| = a \,\middle|\, E_i^c\right) \\
&\leq \sum_{a \in \left[\frac{1}{p\sqrt{\kappa}}, \frac{\kappa}{p}\right]} \mathbb{E}\left[|\mathcal{A}_i| \,\middle|\, F_i \cap |\mathcal{A}_i| = a\right] \cdot \varepsilon \cdot \mathbb{P}\left(|\mathcal{A}_i| = a \,\middle|\, E_i^c\right) \\
&\qquad + \sum_{a \in \left[\frac{1}{p\sqrt{\kappa}}, \frac{\kappa}{p}\right]} \mathbb{E}\left[|\mathcal{A}_i| - 2(1-\varepsilon)|\mathcal{A}_i| \,\middle|\, F_i^c \cap |\mathcal{A}_i| = a\right] \cdot (1-\varepsilon) \cdot \mathbb{P}\left(|\mathcal{A}_i| = a \,\middle|\, E_i^c\right) \\
&\leq \varepsilon \mathbb{E}\left[|\mathcal{A}_i| \,\middle|\, E_i^c\right] + (1-\varepsilon)\mathbb{E}\left[(-1+2\varepsilon)|\mathcal{A}_i| \,\middle|\, E_i^c\right] \\
&= (-1 + 4\varepsilon - 2\varepsilon^2)\mathbb{E}\left[|\mathcal{A}_i| \,\middle|\, E_i^c\right] \\
&\leq (-1+4\varepsilon)\mathbb{E}\left[|\mathcal{A}_i| \,\middle|\, E_i^c\right].
\end{aligned}
$$

Now combining this with (4.25), we have

$$
\begin{aligned}
\mathbb{E}\left[|\mathcal{A}_i|\mathbb{I}_{E_i}\right] + \mathbb{E}\left[|\mathcal{A}_i| - Z_i \,\middle|\, E_i^c\right]\mathbb{P}(E_i^c) &\leq \mathbb{E}\left[|\mathcal{A}_i|\mathbb{I}_{E_i}\right] + (-1+4\varepsilon)\mathbb{E}\left[|\mathcal{A}_i| \,\middle|\, E_i^c\right]\mathbb{P}(E_i^c) \\
&= \mathbb{E}\left[|\mathcal{A}_i|\mathbb{I}_{E_i}\right] + (-1+4\varepsilon)\mathbb{E}\left[|\mathcal{A}_i|\mathbb{I}_{E_i^c}\right]\mathbb{P}(E_i^c) \\
&\leq \frac{2}{\kappa p} + (-1+4\varepsilon)\left(\frac{1}{p} - \frac{2}{\kappa p}\right) \qquad (4.26) \\
&\leq -\frac{1}{p} + \frac{4\varepsilon}{p} + \frac{4}{\kappa p}, \\
&= -\frac{1}{p}\left(1 - 4\varepsilon - \frac{4}{\kappa}\right),
\end{aligned}
$$

where in (4.26) we used Lemma 4.2 twice. Now, we let $\delta \triangleq 4\varepsilon + 4/\kappa$, and observe that we can make δ arbitrarily and the inequality will still hold for sufficiently small

p by our choice of ε and κ_0. As we have

$$\mathbb{E}_{\mathcal{G}}\left[|\mathcal{A}_i| - \min\{|\mathcal{P}_i|, 2|\mathcal{A}_i|\}\right] \leq -\frac{1}{p}(1-\delta) = -\mathbb{E}[|\mathcal{A}_i|](1-\delta),$$

we can apply Proposition A.4 with $\lambda = (1-\delta)$ to obtain that

$$\begin{aligned}\mathbb{E}[|\mathcal{V}(T_\infty)|] &\leq \max\left\{\alpha, \frac{\max\{1, C_2-1\}^2 C_1}{\lambda}\mathbb{E}[A_k]\right\}\left(2+\frac{2}{\lambda}\right) \\ &= \max\left\{\frac{C}{p}, \frac{2}{1-\delta}\frac{1}{p}\right\}\left(2+\frac{2}{1-\delta}\right)\end{aligned}$$

Finally, recall that we are working with the "embedded Markov chain" as we are only observing the process at times T_i. We can relate the actual Markov chain to the embedded Markov chain as follows:

$$\mathbb{E}[|\mathcal{V}(\infty)|] = \lim_{t\to\infty}\frac{1}{t}\sum_{s=0}^{t}|\mathcal{V}(s)| \tag{4.27}$$

$$= \lim_{n\to\infty}\frac{1}{T_n}\sum_{s=0}^{T_n}|\mathcal{V}(s)| \tag{4.28}$$

$$= \lim_{n\to\infty}\frac{n}{T_n}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{s=T_{i-1}}^{T_i}|\mathcal{V}(s)|$$

$$= p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(|\mathcal{V}(T_{i-1})|(T_i - T_{i-1}) + \frac{(T_i - T_{i-1})(T_i - T_{i-1}+1)}{2}\right) \tag{4.29}$$

$$\leq p\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|\mathcal{V}(T_{i-1})|(T_i - T_{i-1}) + \frac{1}{n}\sum_{i=1}^{n}(T_i - T_{i-1})^2\right). \tag{4.30}$$

Here (4.27) follows from the positive recurrence of $\mathcal{G}(t)$. We have (4.28) as $a_n \to a$ implies that for every subsequence a_{n_i}, we have $a_{n_i} \to a$ as well, and using that as $T_n \to \infty$ a.s. We obtain the left term in (4.29) by observing that T_n is the sum of n independent Geometric(p) random variables and then applying the SLLN. For the right term of (4.29), we simply use that $|\mathcal{V}(s+1)| = |\mathcal{V}(s)| + 1$ for $s \in [T_{i-1}, T_i - 1]$, and then the identity $\sum_{i=1}^{n} i = n(n+1)/2$.

We now considering each sum from (4.30) independently. For the first sum, observ-

ing that $|\mathcal{V}(T_{i-1})|(T_i - T_{i-1})$ is a function of our positive recurrent Markov chain $\mathcal{G}(T_i)$, we have that there exists a random variable X^* such that $|\mathcal{V}(T_{i-1})|(T_i - T_{i-1}) \Rightarrow X^*$ and the average value of $|\mathcal{V}(T_{i-1})|(T_i - T_{i-1})$ converges to $\mathbb{E}[X^*]$ a.s. The convergence in distribution $|\mathcal{V}(T_{i-1})|(T_i - T_{i-1}) \Rightarrow X^*$ implies the existence of $|\tilde{\mathcal{V}}(\tilde{T}_{i-1})|(\tilde{T}_i - \tilde{T}_{i-1})$ that converges to X^* a.s. Putting these together, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} |\mathcal{V}(T_{i-1})|(T_i - T_{i-1}) &= \mathbb{E}[X^*] && (4.31)\\
&= \mathbb{E}\left[\lim_{i\to\infty} |\tilde{\mathcal{V}}(\tilde{T}_{i-1})|(\tilde{T}_i - \tilde{T}_{i-1})\right] \\
&\leq \liminf_{i\to\infty} \mathbb{E}\left[|\mathcal{V}(T_{i-1})|(T_i - T_{i-1})\right] && (4.32)\\
&= \liminf_{i\to\infty} \mathbb{E}\left[|\mathcal{V}(T_{i-1})|\right] \mathbb{E}[T_i - T_{i-1}] && (4.33)\\
&= \frac{1}{p}\mathbb{E}[|\mathcal{V}(T_\infty)|]. && (4.34)
\end{aligned}$$

Here we have (4.31) by the ergodic theorem for Markov chains, (4.32) by Fatou's lemma, (4.33) by the independence of $T_i - T_{i-1}$ from $\mathcal{V}(T_{i-1})$, and (4.34) by Theorem 2 from [80] (alternatively, (4.34) can be shown a little extra work using a simpler result from [47]).

For the second sum, as $T_i - T_{i-1} = \tau_i$ are i.i.d. Geometric(p), by the SLLN,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (T_i - T_{i-1})^2 = \mathbb{E}[\tau_1^2] = \frac{2-p}{p^2} \leq \frac{2}{p^2}$$

Thus

$$\mathbb{E}[|\mathcal{V}(\infty)|] \leq p\left(\frac{1}{p}\mathbb{E}[|\mathcal{V}(T_\infty)|] + \frac{2}{p^2}\right) = \mathbb{E}[|\mathcal{V}(T_\infty)|] + \frac{2}{p},$$

showing the result, as we have for the embedded process that $\mathbb{E}[|\mathcal{V}(T_\infty)|] = \Omega(1/p)$. □

Finally, we mention that in moving from the "embedded Markov chain" back to the original Markov chain we make use of the fact that τ_i is light tailed in the sense that $\mathbb{E}[\tau_i^2] = O((\mathbb{E}[\tau_i])^2)$, to obtain a bound of $O(1/p)$ of the steady state expected

number of nodes in the system.

The proof for the lower bound in Theorem 4.3 is based on the following key idea: the waiting time for a node must be at least the time for the node to get an in-degree of one. Using this, if the steady state average waiting time is $w = o(1/p)$, then by Little's Law when a typical vertex v arrives there are only $o(1/p)$ vertices in system, so v is likely not to have any in edges connecting with any of these existing nodes. After w steps, the number of newly arrived nodes is $w = o(1/p)$, so v is likely not to connect to any of these nodes either. Then the idea is to show that v will be in the system for greater than w steps with high enough probability (i.e. with probability at least $1/3$), contradicting that the expected waiting time v is w.

Proof of Theorem 4.3: lower bound. Let $C = 24$. We will show that the expected steady state waiting time w is at least $1/(Cp)$ for all p, giving the result. Assume for contradiction that there exists p such that $w \leq 1/(Cp)$. By Little's law we have that $w = \mathbb{E}[|\mathcal{V}(\infty)|] \leq 1/(Cp)$ as well. Let i be a node entering at steady state, and let W_i be the waiting time of node i. Let $\mathcal{W}$ be the set of nodes in the system when i arrives, so $\mathcal{W} \stackrel{d}{=} |\mathcal{V}(\infty)|$, and define the event $E_1 = \{|\mathcal{W}| \leq 3w\}$. By Markov's inequality, $\mathbb{P}(E_1) \geq 2/3$. Note that i cannot leave the system until it has an in degree of at least one. Let $\mathcal{A}$ be the first $3w$ arrivals after i, and let the event E_2 be the event that either a node from $\mathcal{W}$ or a node from $\mathcal{A}$ has an edge pointing to i. We have

$$\mathbb{P}(E_2) = \mathbb{P}(\text{Bin}(|\mathcal{W}| + 3w, p) \geq 1),$$

making

$$\mathbb{P}(E_2 \mid E_1) \leq \mathbb{P}(\text{Bin}(6w, p) \geq 1) \leq \mathbb{P}(\text{Bin}(6/(Cp), p) \geq 1) \leq \frac{6}{C} = \frac{1}{4}$$

using the definition of E_1, then that $w \leq 1/(Cp)$, then Markov's inequality, and finally that $C = 24$. Thus

$$w = \mathbb{E}[W_i] \geq 3w\mathbb{P}(E_2^c) \geq 3w\mathbb{P}(E_2^c|E_1)\mathbb{P}(E_1) \geq 3w(1 - 1/4)(2/3) = 3w/2 > w$$

providing the contradiction. □

Remark 4.3. Consider, instead, a setting where two and three-cycles can be removed in addition to chains. Theorem 4.3 tells us that under any policy, we still have a lower bound of $1/(Cp)$ on the expected waiting time (in fact, this holds for arbitrarily long cycles). It also tells us that under a greedy policy that executes only chains (ignoring opportunities to conduct two and three-cycles), the expected waiting time is $O(1/p)$. Further, it is not hard to see that this policy misses two and three cycle opportunities for $O(p)$ fraction of nodes. As such, we conjecture that a greedy policy that executes two and three-cycles in addition to chains will have almost identical performance to greedy with chains only, and in particular, the expected waiting time will still be $O(1/p)$.

4.8 Conclusion

Overcoming the rare coincidence of wants is a major obstacle in organizing a successful barter marketplace. In this chapter, we studied how the policy adopted by the clearinghouse affect agents' waiting times in a thin marketplace. We investigated this question for a variety of settings determined by the feasible types of exchanges, which are largely driven by the technology adopted by the marketplace. We also studied how the feasible types of exchanges affect the waiting times. Our study of such marketplaces is motivated in part by questions arising in the design of kidney exchange programs.

We studied these questions in a dynamic model with a stylized homogenous stochastic demand structure. The market is represented by a compatibility graph: agents are represented by nodes, and each directed edge, which represents that the source agent has an item that is acceptable to the target agent, exists a priori with probability p. Exchanges take place in the form of cycles and chains, where chains are initiated by an altruistic donor who is willing to give away his item without asking anything in return. The key technical challenge we face is that in our dynamic setting, the compatibility graph between agents present at a particular time has a complicated

distribution that depends on the feasible exchanges and the policy employed by the clearinghouse.

We analyzed the long run average time agents spend waiting to exchange their item, for small p, in a variety of settings depending on the feasible exchanges, chains, 2-way cycles, 2 and 3-way cycles, or chains. Our main finding is that regardless of the setting, the greedy policy which attempts to match upon each arrival, is approximately optimal (minimizes average waiting time) among a large class of policies that includes batching policies. Under the greedy policy, with cycles of length two, the steady state average waiting time is of order $\Theta(1/p^2)$, while allowing both length two and three cycles leads to a steady state average waiting time of $\Theta(1/p^{3/2})$. Finally, exchanges based on chains lead to a steady state average waiting time of $\Theta(1/p)$. Thus, three-way cycles and chains lead to large improvement in waiting times relative to two-cycles only. Simulations in each setting support these findings, showing that greedy beats batching for small p and also for moderate values of p.

We do not model competition between clearinghouses. In the presence of such competition, where the same agents may participate in multiple clearinghouses, there is an incentive for clearinghouses to complete exchanges at the earliest to avoid agents completing an exchange in a different clearinghouse. One may worry that such an incentive may lead clearinghouses to hurt social welfare when they adopt a greedy-like policy. However, our work suggests that this is not the case, since greedy may be near optimal also from the users' perspective (we find that it approximately minimizes expected waiting times).

Though we motivated our work primarily in the context of centralized marketplaces, our results also have implications for decentralized marketplaces. One implication is that while organizing longer cycles and chains may require a centralized marketplace, our results imply that this may be an option worth considering (only two-way exchanges are typically possible in a decentralized marketplace). Another implication is more subtle: the fact that greedy is near optimal (for two-cycles only) suggests that decentralized marketplaces typically cannot improve outcomes by hiding possible matches; simply informing participants of available matches should be

nearly optimal.

Although we do not model important details of kidney exchange clearinghouses, our findings are consistent with computational experiments that show that the greedy policy is optimal among batching policies. In addition, our findings can serve as foundations for the importance of using chains and cycles of size more than 2 when the kidney exchange pool contains many "hard-to-match" patient-donor pairs, as explained in Section 1.1.3.

Our work raises several further questions and we describe here a few of these. Allowing for heterogeneous agents or goods may lead to different qualitative results in some settings. For example, if Bob is a very difficult-to-please agent who is willing to accept only Alice's item but they are not both part of any single exchange, it may be beneficial to make Alice wait for some time in the hope of finding an exchange that can allow Bob to get Alice's item (note that such a policy is not monotone). In particular, when chains or cycles of more than two agents are permitted, some waiting may improve efficiency in the presence of heterogeneity (some evidence for this is given by Ashlagi et al. [8]).

In kidney exchange, the existence of easy-to-match and hard-to-match pairs (again see Section 1.1.3) creates the following problem. Many hospitals are internally matching their easy-to-match pairs, and enrolling their harder-to-match pairs to centralized multi-hospital clearinghouses [6]. An important question is how much waiting times of hard-to-match pairs will improve as the percentage of easy-to-match pairs grows.

Allowing for agents' departures and outside options are other issues worth exploring. Designing "good" mechanisms that make it safe for agents to participate in barter exchanges is an important direction (see a concurrent work by Akbarpour et al. [4] who study how to make it safe for agents to report their "deadlines" in a homogenous sparse marketplace that conducts pairwise exchanges).

4.9 Proofs of Preliminary Results

Proof of Proposition 4.2. We prove the result by contradiction. Assume that a maximal set of node disjoint three-cycles $\mathcal{W}$ contained fewer than $N/3$ three-cycles. Then there must be a three-cycle X from a largest set of node disjoint three-cycles such that for every three-cycle $Y \in \mathcal{W}$, X and Y have no nodes in common. This yields a contradiction, as we could then add X to $\mathcal{W}$ to make a larger set of node disjoint three-cycles, thus making $\mathcal{W}$ not maximal. □

Proof of Lemma 4.1. We assume that the removal policy is deterministic. The proof for the case of randomized policies follows immediately. Fix any two nodes i, j which arrive before time t (namely $i, j \leq t$). Given any directed graph $\mathcal{G}$ on nodes $0, 1, \ldots, t$ (that is nodes arriving up to time t) such that the edge (i, j) belongs to $\mathcal{G}$, denote by $\bar{\mathcal{G}}$ the same graph $\mathcal{G}$ with edge (i, j) deleted. Let $\mathcal{W}$ be any subset of nodes $0, 1, \ldots, t$ containing i and j. Recall that we denote by $\mathcal{G}_t$ the directed graph generated by nodes $0, 1, \ldots, t$ and by $\mathcal{W}_t$ the set of nodes observed at time t. Note that, since the policy is deterministic, graph $\mathcal{G}_t$ uniquely determines the set of nodes $\mathcal{W}_t$.

We have

$$\mathbb{P}(\mathcal{W}_t = \mathcal{W}) = \sum_{\mathcal{G}} \mathbb{P}(\mathcal{G}) + \sum_{\mathcal{G}} \mathbb{P}(\bar{\mathcal{G}}),$$

where the first sum is over graphs $\mathcal{G}$ containing edge (i, j) such that the set of nodes observed at time t is $\mathcal{W}$ when $\mathcal{G}_t = \mathcal{G}$, and the second sum is over graphs $\mathcal{G}$ containing edge (i, j), such that when $\mathcal{G}_t = \bar{\mathcal{G}}$, the set of nodes observed at time t is $\mathcal{W}$. Note, however that by our monotonicity assumption, if $\mathcal{G}_t = \mathcal{G}$ implies $\mathcal{W}_t = \mathcal{W}$, then $\mathcal{G}_t = \bar{\mathcal{G}}$ also implies $\mathcal{W}_t = \mathcal{W}$. Thus

$$\mathbb{P}(\mathcal{W}_t = \mathcal{W}) \geq \sum_{\mathcal{G}} (\mathbb{P}(\mathcal{G}) + \mathbb{P}(\bar{\mathcal{G}})),$$

where the sum is over graphs $\mathcal{G}$ containing edge (i, j) such that $\mathcal{G}_t = \mathcal{G}$ implies $\mathcal{W}_t = \mathcal{W}$. At the same time note that $\mathbb{P}(\bar{\mathcal{G}}) = \mathbb{P}(\mathcal{G})(1 - p)/p$ since it corresponds to

the same graph except edge (i,j) deleted. We obtain

$$\mathbb{P}(\mathcal{W}_t = \mathcal{W}) \geq \sum_{\mathcal{G}} \mathbb{P}(\mathcal{G})(1 + (1-p)/p) = \sum_{\mathcal{G}} \mathbb{P}(\mathcal{G})/p.$$

We recognize the right-hand side as $\mathbb{P}(\mathcal{W}_t = \mathcal{W} | (i,j) \in \mathcal{G}_t)$. Now we obtain

$$\begin{aligned}\mathbb{P}((i,j) \in \mathcal{G}_t | \mathcal{W}_t = \mathcal{W}) &= \mathbb{P}(\mathcal{W}_t = \mathcal{W} | (i,j) \in \mathcal{G}_t)\mathbb{P}((i,j) \in \mathcal{G}_t)/\mathbb{P}(\mathcal{W}_t = \mathcal{W}) \\ &\leq \mathbb{P}((i,j) \in \mathcal{G}_t) \\ &\leq p,\end{aligned}$$

and the claim is established. □

Proof of Lemma 4.2. By the memoryless property of the geometric distribution, for all $t > 0$,

$$\mathbb{E}[X \mid X > t] = t + \mathbb{E}[X] = t + \frac{1}{p}. \tag{4.35}$$

Thus for all sufficiently large κ we have

$$\begin{aligned}\mathbb{E}\left[X \mathbb{I}_{X > \frac{\kappa}{p}}\right] &= (1-p)^{\frac{\kappa}{p}} \left(\frac{\kappa}{p} + \frac{1}{p}\right) && (4.36)\\ &\leq e^{-\kappa} \frac{1+\kappa}{p} && (4.37)\\ &\leq \frac{1}{2\kappa p}, && (4.38)\end{aligned}$$

where (4.36) follows from (4.35), (4.37) follows as $(1-p)^{1/p} \leq e^{-1}$ for all p (take logarithms), and finally (4.38) holds provided $\kappa \geq \kappa_0$ for appropriately large κ_0.

For the remaining term, we have that for all sufficiently large κ and sufficiently

small p,

$$\begin{aligned}
\mathbb{E}\left[X\mathbb{I}_{X\le\frac{1}{\sqrt{\kappa p}}}\right] &= \mathbb{E}[X] - \mathbb{E}\left[X\mathbb{I}_{X>\frac{1}{\sqrt{\kappa p}}}\right] \\
&= \frac{1}{p} - (1-p)^{\frac{1}{\sqrt{\kappa p}}}\left(\frac{1}{\sqrt{\kappa p}} + \frac{1}{p}\right) && (4.39)\\
&\le \frac{1}{p} - \left(\frac{1-p}{e}\right)^{\frac{1}{\sqrt{\kappa}}}\left(\frac{1}{\sqrt{\kappa p}} + \frac{1}{p}\right) && (4.40)\\
&\le \frac{1}{p} - (1-p)^{\frac{1}{\sqrt{\kappa}}}\left(1-\frac{1}{\sqrt{\kappa}}\right)\left(\frac{1}{\sqrt{\kappa p}} + \frac{1}{p}\right) && (4.41)\\
&= \frac{1}{p} - (1-p)^{\frac{1}{\sqrt{\kappa}}}\frac{1}{p} + (1-p)^{\frac{1}{\sqrt{\kappa}}}\frac{1}{\kappa p}\\
&\le \frac{2}{\sqrt{\kappa}} + \frac{1}{\kappa p} && (4.42)\\
&\le \frac{3}{2\kappa p} && (4.43)
\end{aligned}$$

where (4.39) follows from (4.35). To obtain (4.40), by Taylor's theorem, $(1-p)^{1/p} = e^{-1}(1-p/2)+o(p)$ as $p\to 0$, thus for sufficiently small p, we have $(1-p)^{1/p}\ge e^{-1}(1-p)$. In (4.41) we use that $e^{-x}\ge 1-x$, in (4.42), we use that for all x sufficiently small, $1\ge(1-x)^n\ge 1-2xn$, and (4.43) follows by taking κ sufficiently large. Thus the result is shown. □

Chapter 5

Data Driven Simulations for Scheduling Medical Residents in Hopsitals

5.1 Introduction

In 2011, the ACGME instituted a new set of regulations on duty hours that restrict shift lengths for medical residents. We consider two operational questions for hospitals in light of these new regulations: will there be sufficient staff to admit all incoming patients, and how will the continuity of patient care be affected, particularly in a first day of a patients hospital stay? To address these questions, we built a discrete event simulation tool using historical data from a major academic hospital, and compared several policies using both long and short shifts. Using our simulation tool, we will find that schedules based on shorter more frequent shifts actually lead to a larger admitting capacity. At the same time, such schedules generally reduce the continuity of care by most metrics when the departments operate at normal loads. However, in departments which operate at the critical capacity regime, we found that the continuity of care improved in some metrics for schedules based on shorter shifts, due to a reduction in the use of overtime doctors. In contrast to much of the existing

literature on the effects of duty hour regulations, our approach directly measures the relevant performance metrics, rather than relying on the perceptions of outcomes gathered in surveys.

In our simulations, we find that the relationship between how residents are scheduled and the capacity of the hospital to admit patients is quite complex. In particular, the capacity is not well estimated by simply considering the number of residents in conjunction with any one of: (a) the number of hours residents are available to admit patients, (b) the number of patients each resident is allowed to admit per shift, (c) the number of patients each resident is allowed to have in care simultaneously. Instead, these constraints interact in highly non-trivial ways. We observe that the capacity of two schedules can differ greatly even when the same number of residents are used and long run averages are constant for (a), (b), and (c). We develop the *Markov chain throughput upper bound* as a simple model that can quickly compute the interaction between these constraints to estimate the capacity of a schedule.

Organization

The chapter is organized as follows. In Section 5.2, we explain how our simulation model works. In Section 5.3, we give the simulation results. In Section 5.4 we discuss our results and the conclusions we draw for scheduling medical residents in hospitals. In Section 5.5, we give the details of how our *Markovian chain throughput upper bound* is computed. Finally, in Section 5.6, we give some additional details on the statistical models used to scale up or down the historical data set.

5.2 Materials and Methods

5.2.1 Simulation Model of Patient Flow and Assignment to Doctors

We now describe a model for evaluating a hospital departments ability to admit and treat patients under a particular staffing schedule. As input, the model takes the

historical data of every patient treated by the department over some period of time (e.g. a year), including the time they arrive and the time they leave the hospital. Another input to the model is a list of teams of admitters. Each team is either a *resident team* or a *physicians assistant (PA) team*. Patients are admitted to specific members of teams. Each team follows a *schedule* specifying when they are eligible to admit patients. Schedules give typically a four to nine day rotation, and specify on which days of the rotation and at what times the team on *shift* i.e. eligible to admit patients. Further, each team is subject to some additional constraints:

- *(C1)* maximum number of patients in care allowed for a team,
- *(C2)* maximum number of patients admitted per shift allowed for a team,
- *(C3)* maximum number of patients in care allowed for an individual resident or PA,
- *(C4)* maximum number of patients admitted per shift allowed for an individual resident or PA.

Most of the resident constraints are dictated by ACGME work hour restrictions, while most of the PA constraints are based on hospital policy as a practical measure. For example, if a team of two admitters had a constraint of type *(C3)* saying each admitter could only have 10 patients in care and a constraint of type *(C1)* saying the team could only have 18 patients in care, then it would be possible for one admitter to have 10 patients and the other 8, but having 11 and 7 or 10 and 9 would not be allowed. The model further specifies a selection rule that determines which of the eligible admitters on shift will be assigned to an arriving patient. Throughout, we use the following rule: when a patient arrives, we look at all teams on shift with an eligible admitter, and select one such team uniformly at random. We then select a team member uniformly at random from the eligible admitters in this team. We do this as hospital data suggested that there was not a systematic procedure determining which patients were assigned to which admitters.

Given these inputs, our simulation model works as follows. We define the set of *event times* to be the times of patient arrivals, patient departures, and the beginning of admitter shifts. For each admitter, we maintain a list of current patients in care.

We also maintain a list of each patient that has arrived but has not yet been assigned to an admitter. We then iterate through the event times chronologically and update the lists. There are three possibilities, based on the type of the event.

- When the next event is a patient arrival, we find the admitting teams that are on shift. From these teams, we find the members that satisfy constraints *C1-4*. Assuming there is such an admitter, we use the selection rule to choose who receives this patient. However, if there is no available admitter, i.e. every admitter that is currently on shift violates at least one of *C1-4*, then that patient is added to the list of patients that have arrived but not yet been assigned to an admitter. These patients are admitted temporarily by either the night float service or the jeopardy service (to be transferred to a resident team or PA team later).
- When the next event is a patient departure, if the patient has been assigned to an admitter, we remove that patient from list of patients in care by the admitter. Otherwise, the patient must be in the list of patients not yet assigned to an admitter, and we remove them from this list.
- When the next event is the beginning of a new shift, we take the patients in the list of patients not yet assigned to an admitter, and attempt to assign them to an admitter on shift satisfying constraints *C1-4* using the selection rule. The patients that have been waiting longest are assigned first.
- If at the beginning of a new shift, there is insufficient capacity to admit all patients waiting for a reassignment, the remaining patients are "dropped" from the simulation. In reality, the hospital is forced to declare a state of emergency census and the capacity of each team is increased. Under a properly functioning schedule, this happens very infrequently. Thus the number of dropped patients in a simulation is a good indicator of the system operating at an unsustainable load. Note that the load on the admitting staff is to some degree shielded from extreme demands by the number of beds in the hospital. No matter how fast patients are actually arriving, from the perspective of the physicians there can

only be as many inpatients as there are beds.[1]

Note that while in principle, an admitter could take a new patient from the list of unassigned patients after another patient departed, we do not allow for this in our model. In fact, in practice at B&W, once a resident or team of residents has reached capacity on one of the constraints *C1-4*, the resident does not admit any new patients for the remainder of the shift, even if patients depart and the resident is eligible again. This occurs as the admission nurses assign the patients to admitters manually, and are only notified when an admitter becomes ineligible. As there can only be patients unassigned to an admitter when all the admitting teams are capped, we know that the team which just experienced a departure is capped as well, so our simulator actually operates as B&W does.

This specifies the dynamics of our model. We use it to collect statistics on the performance metrics of interest, e.g. the fraction of patients who do not receive an admitter immediately upon arrival, or the percentage of time that a particular team is capped.

5.2.2 Physician's Assistants and Admitting

The *physician's assistant (PA) teams* for B&W each provide 2 PAs in the hospital 24 hours a day, 7 days a week (each team consists of around 6 individuals, but there are always 2 present). These teams can have up to 15 patients in care simultaneously, and have no other formal restrictions. B&W employed one such team for GMS in the period 2007-2011. For 2011-2012, they added a second team for GMS. The Cardiology department only had a team for 2011-2012, and the Oncology department only had a team for 2010-2012. B&W felt that admitting a patient to a PA service provided the same quality of care as admitting to a resident.

[1] Additionally, in most academic medical centers there are boarders, which are patients that do not have an inpatient bed, but are admitted to an inpatient team and stay in the emergency room.

5.2.3 Resident Admitting Shifts, Schedules, and Policies

Here we describe the scheduling system used to determine when residents are on shift, and the total number of residents and PAs required for each department. As far as the residents are concerned, the system consists of three layers, *shifts*, *schedules*, and *policies*. Residents are put in *teams* typically of size 2-3, and teams are organized into *groups*, each with 2-4 teams. Shifts specify when a resident is on shift or off shift for a 24 hour period from 7am-7am. A schedule applies to a group of teams of residents. It takes a sequence of shifts (usually 4 or 6) and assigns each resident to rotate through these shifts. It also specifies the offset between residents on the same team and the offset between teams on the rotation. Finally, a policy assigns schedules to every group of team of residents, and further specifies an offset between groups in a department. Generally, offsets are chosen so that the residents provide coverage uniformly throughout the rotation. A policy also specifies the number of PA teams a department will be assigned.

As an example, we now describe in detail the policy by GMS for the 2010-11 academic year. The shifts used were the *Long Shift* (L) that ran from 7am to noon the following day, the *Short Shift* (S) that ran from 7am to 2pm, and the *Off Shift* (O) where residents were not admitting new patients. At the schedule level, we have two groups of teams. For the first group, there were four teams with two residents each, called GMS A-D. They all followed the four day rotating schedule Long Shift, Off, Short Shift, Off (LOSO). Within a team, there was no offset between the two residents, but between teams, there was a one day stagger, ensuring that every day there was a team on Long Shift. The second group of teams formed the *intensive teaching unit* (ITU). It had two teams, ITU 1 and ITU 2, and each team had three residents. The residents followed the 6 day rotating schedule Long Shift, Off, Off, Short Shift, Short Shift, Off (LOOSSO). Within a team, each resident was offset by a day on the rotating schedule, and between teams, there was a three day offset on the rotating schedule, again insuring that every day there was one resident on Long Shift. The *policy* for GMS was simply to use the GMS A-D group and the ITU 1-2 group.

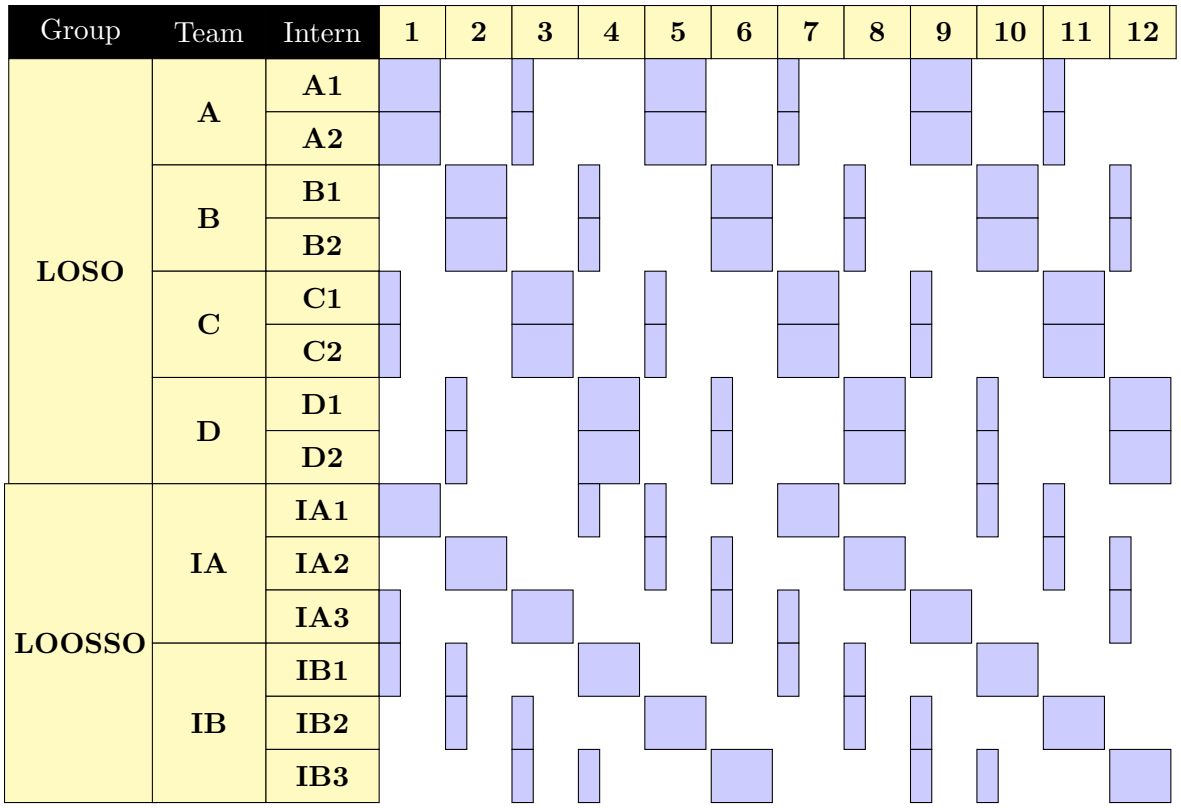

Figure 5-1: The GMS policy for the 2010-11 academic year. The GMS A-D teams have a 4 day schedule and the ITU 1-2 teams have a 6 day rotating schedule, so the system has an overall 12 day rotating schedule.

There was no need to specify any offset at the policy level as both groups of residents have schedules which are symmetric. Figure 5-1 provides a diagram explaining all the offsets between residents for the policy. Additionally, the policy allocated a single team of 2 PAs to GMS (giving a PA capacity of 15 patients in care). There are several additional details describing shift structure. The number of hours a resident must spend at the hospital is longer than the shift, as it takes the resident approximately two hours on average to treat a newly admitted patient. For some shifts, if the shift falls on a weekend, the shift is canceled. Some shifts have different lengths in the first and second semesters, to give more experienced residents more hours. Under some policies where there are no residents admitting patients between 5am and 7am, patients arriving at this time wait until 7am and then are admitted by the doctor starting a new shift. This mechanism is called "Early Admit." There is a maximum number of patients a resident can admit while on this shift. This is set as realistically

Name	Admitting Hours	In hospital hours	Weekends	Early Admit	Max Admits	(Onc)
L	7am-5am (22 hours)	7am-Noon (29 hours)	Yes	No	5+2	(5+0)
S	7am-2pm (7 hours)	7am-2pm+ (7+ hours)	No	No	2+1	(1+0)
M	2pm-10pm (8 hours)	2pm-Midnight (10 hours)	Sometimes	No	3+1	(3+0)
M'	10am-6pm (8 hours)	10am-8pm (10 hours)	Sometimes	No	3+1	(3+0)
D	7am-6pm (11 hours)	7am-8pm (13 hours)	Yes	Yes	4+1	N/A
D'	7am-7pm (12 hours)	7am-9pm (14 hours)	Yes	Yes	4+1	N/A
N	6pm-5am (11 hours)	6pm-7am (13 hours)	Yes	No	4+1	N/A
O	None	Variable	Variable	No	0	0

Table 5.1: The above shifts were used to construct schedules for residents.

each new admission takes two to three hours of work, and residents are not supposed to take patients that force them to stay beyond the end of their shift. The maximum number of admissions per shift is typically of the form "5+2," meaning that the resident can do five new admissions, but may do a total of seven admissions including taking reassignment patients (see Section 5.2.4). If there are no reassignments, then the resident can only admit 5 patients, but if the resident must admit more than 2 reassignments, then the resident can still only admit a maximum of seven patients for the day. Also this maximum is dependent on the department. Specifically, doctors from Oncology can admit fewer patients per shift as their admissions take more time. All of these details are summarized for each type of shift in Table 5.1.

As previously described, a *schedule* assigns a rotating sequence of shifts to a group of resident teams, and gives an offset within each team and between teams in the rotating schedule. The schedules considered in this project are summarized in Table 5.2. Here average hours/week indicates the number of hours spend admitting. This translates the 4 and 6 day schedules into 7 day averages accounting for weekends. The offset information is omitted, but under each schedule, the coverage provided by the residents is uniform throughout the rotation. We now give a brief discussion of the motivation for each schedule.

1. **LOSO**– This was the default schedule that was in place for all three B&W departments prior to the new regulations. By using long shifts, the schedule provides long periods of patient observation following admission and reduces the total number of daily shift changes, providing continuous ongoing care for patients.

2. **LOOSSO**– This schedule was only used by the *Intensive Teaching Unit* teams for GMS. As these teams are spending additional time on educational aspects of patient care, they have a reduced capacity of 18 patients in care for 3 residents (recall that LOSO has a capacity of 20 patients in care for 2 residents).
3. **MMMO**– This schedule was designed as an alternative to LOSO that with the goal of increasing capacity and reducing jeopardy. When the project began, due to a recent restructuring, several B&W departments found themselves in a perpetual state of emergency census (when teams must operate above their capacity) and jeopardy. Interestingly, we observed that by having many short frequent shifts instead of occasional long shifts, the fundamental capacity of the system is increased, an issue further explored in Section 5.5.
4. **M'M'M'O**– This schedule is nearly the same as MMMO, except the resident shifts are shifted four hours earlier. This schedule was proposed as an alternative to MMMO with the goals of improving resident quality of life (by allowing them to get home from work at a reasonable hour) and increasing the overlap between the hours when a resident was admitting and the hours when a resident was in the hospital attending to other responsibilities such as treating and discharging existing patients (these activities tend to take place in the morning).
5. **D'OSO**– This schedule essentially just takes the old LOSO schedule and makes it regulations compliant by significantly reducing the length of the long shift. Like M'M'M'O, this schedule has shifts ending during peak hours, which can prevent residents from leaving on time. To try and ease the residents' departures, the *flex admitter* is introduced for this schedule, as described in Section 5.2.4. Also notice that the *team in care capacity* of D'OSO is 15, while for LOSO it was 20.
6. **DOOONO**– This schedule was designed to provide both day and night admitting coverage by residents without using long shifts. The schedule also reduces the number of patients in care per resident (a team cap of 20 for 3 residents, as opposed to a team cap of 20 for 2 residents under LOSO). Additionally, under DOOONO residents only can admit 2 days out of every 6 and ten patients every

Name	Teams	Residents/Team	Team in Care Cap
LOSO	4	2	20
LOOSSO	2	3	18
MMMO	4	2	20
M'M'M'O	4	2	20
D'OSO	4	2	15
DOOONO	2	3	20
DOSOOONOO	3	3	20

Table 5.2: The following schedules were used to manage resident teams. Included are both actual policies implemented by B&W as well as alternatives considered as part of this research.

six days, whereas under LOSO residents can admit 2 out of every 4 days and 10 patients every four days.

7. **DOSOOONOO**– This 9 day rotating schedule is essentially a variant of DOOONO proposed for the cardiology department that provides uniform coverage using 9 total residents, as opposed to the 6 residents required for DOOONO.

Finally, we describe the policies. Recall that a *policy* was simply a list of schedules assigned to *groups*, with an offset between each group saying where to start in the rotating schedule. Additionally, a policy must specify the capacity of the PA teams.

For each policy, we define the *throughput* as follows. Suppose that there were infinitely many patients waiting to be treated initially, but each patient only begins recovering when they are assigned to a resident or PA. The *throughput* of the policy is average rate that patients leave for the entire department (in patients per day). The throughput of a policy is a good indicator of how a policy will perform under an increasing patient load. Intuitively, as the average number of patients arriving per day increases towards the throughput, the likelihood of jeopardy and dropped patients increases.

The throughput is difficult to determine exactly, but there are two natural upper bounds that we can easily observe. First, suppose that the residents could ignore constraints (C2) and (C4), which restrict the number of patients they can admit per day. Then as we are assuming that there are infinitely many patients waiting, the residents would always have the maximum number of patients in care as allowed by

(C1) and (C3). For example, if an resident always has ten patients in care simultaneously, and the average patient stays for four days, then on average $10/4 = 2.5$ patients will depart per day that were under the care of that resident. In reality, the resident cannot maintain ten patients in care at all times because of constraints (C2) and (C4) restricting the number of admissions per day, so we are overestimating the throughput. We call this estimate the *capacity upper bound on throughput*. For our second upper bound, we instead assume that the residents can ignore constraints (C1) and (C3) restricting the number of patients they have in care, but must obey the constraints (C2) and (C4) restricting the number of patients they can admit per day. For example, under LOSO, on an "L" day, the resident admits seven patients, on an "S" day the resident admits 3 patients, and otherwise the resident does not admit any patients. Thus the resident admits ten patients every four days, or $10/4 = 2.5$ patients per day. However, there is a small error in this computation, as the "S" shift only occurs on week days, so really on average the resident admits $(7 + 3 \cdot 5/7)/4 = 2.25$ patients per day. Again in reality, the resident cannot always admit all these patients, as sometimes the resident will reach ten patients in care and be forced to stop admitting. Thus we have obtained a second upper bound on the throughput, which we refer to as the *admitting upper bound on throughput*.

As the constraints (C1)-(C4) only affect individual residents and teams of residents, we can approximate the throughput of a policy by approximating the throughput of each team of residents and the PAs, and then summing the results up. The throughput of the PAs is equal to the capacity upper bound on throughput, as PAs only have a constraint on the number of patients in care, not on the number of patients admitted per day. For example, for GMS, if the PAs have the capacity to treat 15 patients simultaneously and patients stay for an average of four days, then the PAs add $15/4 = 3.75$ to the policy throughput. To give a very simple example of how to compute the throughput of a policy, suppose that GMS put both group 1 and group 2 on the schedule LOSO, that GMS had a PA service that could treat 15 patients simultaneously, and that the average length of stay was four days. Then the PA service would have a throughput of exactly 3.75 patients per day. There would

be 8 teams each with two residents on LOSO, totaling 16 residents. Each resident has a capacity upper bound of 2.5 patients per day, giving the residents a total capacity upper bound of $16 \cdot 2.5 = 40$ patients per day, and making the policy capacity upper bound 43.75 patients per day. Similarly, each resident has an admitting upper bound of 2.25 patients per day, making the admitting upper bound for all residents $16 \cdot 2.25 = 36$ patients per day, and thus making the admitting upper bound for the policy equal to 39.75. As a result, we expect the throughput of the proposed policy to be less than 39.75.

Using these two upper bounds, we can conclude that the true throughput must be less than the minimum of these two upper bounds. One might suspect that the true throughput would be equal to the minimum of these two upper bounds, or at least that for two policies for which the minimum of these upper bounds is the same, that the throughput should be the same. In particular, when we apply our upper bounds for the policies Initial and Daily Admitting, we get the same bound on the throughput, so we suspect that the policies have the same throughput. However, it turns out that this is not the case. In Section 5.5, we derive a much more accurate approximation of the throughput where (C1)-(C4) are all used, which we refer to as the *Markov chain throughput upper bound*. When all four constraints are combined, the constraints (C2) and (C4) on the number of patients that can be admitted per day cause the residents to not always have 10 patients in care simultaneously, and constraints (C1) and (C3) on the number of patients that a resident can have in care simultaneously cause the resident to be unable to admit the maximum number of patients per shift, thus making both upper bounds too large. While the upper bounds to provide a good first order test to compare schedules, we will see in later sections that the Markov chain throughput upper bound correctly predicts that policies using MMMO should experience fewer dropped patients than policies using LOSO.

Select policies are given in Table 5.3 for GMS, Table 5.4 for Cardiology, and Table 5.5 for Oncology, along with the capacity upper bound on throughput, the admitting upper bound on throughput, and the Markov chain throughput upper bound. A detailed breakdown of throughput for each schedule is given in Section 5.5, along with

Name	Group 1	Group 2	PA Capacity	Throughput Upper Bound (Patients/Day)		
				Capacity	Admitting	Markov
Initial	LOSO	LOOSSO	15	30.4	33.2	25.9
Daily Admitting	MMMO	LOOSSO	15	30.4	36.7	27.7
Preserve Day	D'OSO	D'OSO	30	33.7	35.3	28.7
Corkscrew	DOOONO	DOOONO	30	24.7	26.7	19.8
Hybrid-15	DOOONO	D'OSO	15	25.9	27.7	20.9
Hybrid-30	DOOONO	D'OSO	30	29.2	31.0	24.2

Table 5.3: A sample of policies considered for GMS.

Name	Group 1	PA Capacity	Throughput Upper Bounds (Patients/Day)		
			Capacity	Admitting	Markov
Initial	LOSO	0	17.0	18.3	13.7
Daily Admitting	MMMO	0	17.0	21.7	15.3
Triploid-0	DOSOOONOO	0	12.8	12.4	10.8
Triploid-15	DOSOOONOO	15	16.0	15.6	14.0
Triploid-30	DOSOOONOO	30	19.2	18.8	17.2

Table 5.4: A sample of policies considered for Cardiology.

some additional explanation for the calculations. In particular, schedules creating teams where the *(C1)* constraint (team patients in care capacity) is not dominated by the *(C3)* constraint (individual patients in care capacity), e.g. LOOSSO, require some additional approximation.

Notice that policies using the schedules D'OSO, DOOONO and DOSOOONOO have additional PA capacity. This is necessary to offset the reduction in the number of patients residents can have in care and the number they can admit under these schedules, i.e. the reduced throughput.

Name	Group 1	PA Capacity	Throughput Upper Bound (Patients/Day)		
			Capacity	Admitting	Markov
Initial	LOSO	15	15.0	13.8	12.2
Daily Admitting	MMMO	15	15.0	18.7	13.9
Daily Admitting Shifted	M'M'M'O	15	15.0	18.7	13.9

Table 5.5: A sample of policies considered for Oncology.

5.2.4 Patient Flow for Reassigned Patients

Here we describe what happens to the patients that arrive when all admitting teams are at capacity and are forced to wait. Such patients are referred to as *reassignments*, as they are admitted by one doctor and then permanently transferred to another. At B&W, three different groups provide this temporary care: the *night floats*, the *flex admitters*, and the *jeopardy service*. Further, the flex admitters have two modes of admission, *flex forward* and *flex backward*. We now briefly discuss medical and financial implications of admitting patients through each of these mechanisms.

The *night floats* are residents and doctors that arrive in the evening and leave the following morning. The exact hours when the night floats arrive and depart tends to vary quite widely depending on the policy. The night floats primary responsibility is cross coverage (providing care for patients that have already been admitted, but whose doctors have left the hospital for the night). Night floats typically are covering for many patients simultaneously, and as a result a large fraction of their time is spent simply keeping existing patients stable. However, when all PAs and residents on shift have reached capacity and can no longer take new patients, the night floats watch over newly arriving patients temporarily. The following morning when there is a shift rotation, these patients are handed off to a resident team or PA service. These patients are referred to as *night float reassignments*. In our model, there is no limit to the number of patients the night floats can admit.

Admissions by night floats are considered quite undesirable. When a patient is admitted, the admitter spends typically around two hours with the patient while determining a course of treatment. This can be a very substantial fraction of the total time and attention they will receive from their doctor. Thus when a patient is admitted by one doctor and then permanently transferred to another, information can be lost. Additionally, there are underlying issues with the night float system that compound these problems. For all patients treated by night floats, both reassignments and patients in cross coverage, it has been historically difficult to enforce protocols for the night floats to document their observations and communicate them with the

resident receiving the patient the following morning. In the case of reassignments, these lapses in communication are particularly dangerous as many critical decisions regarding the course of treatment are made soon after the patient arrives, making this a particularly bad time for a mistake. Thus it is desirable to have schedules with low night float admissions rates.

Next we consider the *jeopardy service*. If all the admitters on shift have reached capacity before the night floats arrive, a state of jeopardy is declared. Doctors are called in for overtime to temporarily admit and treat new patients until the night floats arrive. These patients are handed off to the night floats, and then handed off again to a resident or PA the following morning.

Admissions by jeopardy are significantly worse than an admission directly to the night floats, both from a financial and medical perspective. As jeopardy relies on doctors being paid overtime, jeopardy admissions are much more expensive than other types of admissions. Additionally, patients are now handed off twice before they reach their caring doctor, doubling the opportunity for a communication failure. It was the expectation of B&W that jeopardy should occur infrequently.

Finally, the *flex admitters* were a newly created service for the 2011-12 year. They admit new patients from 4-8pm daily, but only when resident and PA service capacity is exhausted. The flex admitter is used only when the schedule D'OSO is used, namely, by GMS under the policies Preserved Day and Hybrid. For the purposes of our model, the flex admitter is not constrained in the number of patients they can temporarily hold.

The flex admitter has two modes of admission, *flex forward* and *flex backward*. The modes differentiate which doctor will ultimately treat the patient. When the flex admitter takes the patient, if there is a resident whose shift ended within the last two hours that is not yet at capacity (and on a team that is not yet at capacity), the patient will be handed back to this doctor the following day, resulting in a *flex backward* admission. When there are no such residents available, the patient will be handed off to the night floats at 8pm and then handed off again to a resident or PA the following morning, resulting in a *flex forward*.

The flex forward is somewhat analogous to a jeopardy admission, but it is slightly better in that no overtime pay is necessary. However, the real motivation of this mechanism is to reduce jeopardy admissions in exchange for flex backward admissions. Flex backward admissions are desirable as the resident can observe the patient with the flex admitter for a portion of the admitting process and have a face to face discussion with the flex admitter about the patient before the resident leaves for the day. Thus the resident will know more about this patient than a resident beginning shift the following day, significantly reducing information lost due to miscommunication.

5.3 Results

We now discuss the result of simulating the policies described above in our model. First, we list the performance metrics that we compute in our simulation studies, as well as the motivation and goals behind the selecting the schedules and policies that we simulated. Then, we look at simulations of these policies and compare their performance.

5.3.1 Performance Metrics

We now briefly summarize how we measure the quality of each policy

1. *Reassignments*– As described in Section 5.2.4, a reassignment occurs when all residents and PAs are at capacity when a patient arrives, so that patient must be temporarily admitted by another doctor and then transferred to a resident or PA the following day. The four types of reassignments are listed below from least to most desirable:
   (a) *Jeopardy reassignment*– These patients are admitted by doctors working overtime before the night floats arrive. They are handed off to the night floats, and then handed off again to a resident or PA the following morning. Using jeopardy requires the hospital to pay overtime wages.

(b) *Flex forward reassignment*– These patients are admitted by the flex admitter. These patients are also handed off to night floats and then to a resident or PA the following morning.

(c) *Night float reassignment*– These patients are temporarily admitted by the night floats and then reassigned to a resident or PA the following morning.

(d) *Flex backward reassignment*– These patients are admitted by the flex admitter with the support of a resident in the final two hours of their shift. The patient is watched temporarily by the night floats, and then returned to the same resident the following morning.

2. *Jeopardy Days*– This is simply the number of days in the year that a jeopardy reassignment occurred. As usually only a single jeopardy doctor is required to deal with all jeopardy admissions for a day, the cost of jeopardy is driven by the number of days of occurrence, not the number of patients.
3. *Dropped Patients*– These are patients that should have been reassigned, but at the time of reassignment to a resident or PA, the capacity was exhausted (as described in Section 5.2.1). Such patients are a sign that the policy is chronically over capacity.
4. *Observation < 6 hours*– This is the fraction of patients admitted by the residents that were viewed continuously for less than 6 hours following admission by the resident before the residents shift ended. Here we are only accounting time *on shift* in the observation hours. Note that while residents spend typically 80 hours a week in the hospital, only 40 of these are hours on shift. Thus depending on how the remaining time in the hospital is allocated, this statistic could significantly underestimate the actual frequency of at least six hours of observation. For example, under LOSO, on an "L" day, residents would remain in the hospital until noon the following day, so effectively there would always be less than 6 hours of observation.
5. *Shift runs late*– Here, we indicate the frequency that a resident must admit a patient in the final two hours of a shift (among all days with any shift other than "off"). This figure is important as admitting a patient requires two to three

hours of work, so the resident will be forced to work past the end of their shift. This can ultimately lead to residents violating restrictions on hours worked per week, so it is important to anticipate with what frequency such admissions will occur. Historically, both and B&W and at other hospitals, there have been problems getting residents out of the hospital when their shifts end during peak admitting hours.

6. *Interarrival < 2 hours*– Here we measure the frequency that a residents successive admission times are separated by less than two hours. We use two hours as it takes about two hours to admit a new patient. Thus we are measuring how often residents are forced to deal with two or more patients simultaneously. From a quality of care perspective, it is desirable that this number be low.

5.3.2 Assessing Policies at Historical Patient Arrival Rate

First, we compare the performance of the policies for the 2009-10 historical patient arrival data for each of our three departments, Oncology, Cardiology, and GMS.

In Table 5.6, we give the performance of these policies in the Oncology department. In all of our simulations, the Oncology department has four teams each of two residents, and a PA service with a capacity of 15 patients in care that admits patients 24 hours a day, 7 days a week. We consider three policies from Table 5.5, *Initial*, *Daily Admitting*, and *Daily Admitting Shifted*. We see that under all three policies, we do not drop patients, implying that they provide sufficient admitting capacity to treat incoming patients. We see that Initial has some jeopardy, but both Daily Admitting and Daily Admitting Shifted have none. However, we see that Initial uses significantly fewer night float admissions than Daily Admitting and Daily Admitting Shifted. We see that Initial is more likely to provide 6 hours of continuous observation than either Daily Admitting or Daily Admitting Shifted. As previously remarked in Section 5.3.1, the policy Initial actually will observe nearly all patients for at least 6 hours, as this figure only accounts for time spent on shift, not time in the hospital. Although the hours in hospital not on shift have not been specified for Daily Admitting or Daily Admitting Shifted, if these policies were implemented these

Schedule	Initial	Daily Admitting	Daily Admitting Shifted
Night Float Admissions	242	676	1298
Jeopardy Days	20	0	0
Jeopardy Admissions	43	0	0
Dropped Patients	0	0	0
Observation < 6 Hours (%)	47.0	83.7	95.4
Shift Runs Late (%)	15.9	13.0	18.5
Interarrival < 2 Hours (%)	38.0	19.3	12.8

Table 5.6: Performance of the Oncology Department under three different policies for scheduling residents. This simulation is based on 2009-10 historical data. There were 3246 patients of the course of the year in the simulation.

hours would likely be before the shift began, so that the residents could be present in the morning for rounds and to discharge patients.

Finally, we see that under Initial it is much more likely that the patient interarrival time is less than two hours than under either Daily Admitting or Daily Admitting Shifted. This is occurring as Daily Admitting and Daily Admitting Shifted have more residents on hand admitting simultaneously at peak hours when patients are arriving more frequently.

In comparing Daily Admitting and Daily Admitting Shifted directly, we see that Daily Admitting Shifted puts significantly more load on the night floats. This occurs as the peak arrival process is between 2pm and 10pm, so half of the peak arrivals are missed by Daily Admitting Shifted and are forced onto the night floats. We also see that Daily Admitting Shifted has a higher fraction of patients with less than 6 hours of observation before the end of the admitters shift. This occurs as we only meet the 6 hour threshold when a patient arrives in the first two hours of the shift, and there are far fewer arrivals between 10am and noon than between 2pm and 4pm. We see an increase in the likelihood that a shift will run late under M'M'M'O, as this schedule has residents end their shifts at the peak arrival time of the day. However, this policy will provide residents with a higher quality of life, as they are released from the hospital at an earlier hour of the day.

Next, in Table 5.7, we look at the performance of the policies from Table 5.4, *Initial*, *Daily Admitting*, and *Triploid*. Under the first two policies, a group of 8

Schedule	Initial	Daily Admitting	Triploid-0	Triploid-15	Triploid-30
Night Float Admissions	445	1313	473	125	5
Jeopardy Days	17	0	210	63	1
Jeopardy Admissions	24	0	760	144	3
Dropped Patients	0	0	156	10	0
Observation < 6 Hours (%)	44.8	84.8	70.5	61.3	60.4
Shift Runs Late (%)	16.6	14.9	25.7	29.7	31.7
Interarrival < 2 Hours (%)	40.9	18.9	41.4	37.5	32.3

Table 5.7: Performance of the Cardiology Department under five different policies for scheduling residents. This simulation is based on 2009-10 historical data. There were 3257 patients of the course of the year in the simulation.

residents and no PAs are used. Under the Triploid policy, nine residents are used, but individually residents are constrained to admit patients more slowly and care for fewer total patients. To offset this, we consider 3 levels of PA staffing, no PAs (as under the first two policies), referred to as *Triploid-0*, one PA team (giving a capacity of 15 patients), *Triploid-15* and two PA teams (giving a capacity of 30 patients), referred to as *Triploid-30*.

The comparison between Initial and Daily Admitting for Cardiology is very similar to the relation of these policies for Oncology. We see that Initial puts less load on the night floats than Daily Admitting, but that Daily Admitting puts less (in fact no) load on jeopardy than Initial. Again, Daily Admitting is less likely to have 6 continuous hours of observation than Initial, but is also less likely to have to admit twice in a two hour window than Initial.

Looking at the three levels of PA staffing for the Triploid policy, it is immediate from the dropped patient and jeopardy statistics that Triploid-0, i.e. not increasing PA staffing, is infeasible. Under Triploid-15, we still have the occasional dropped patient and quite a bit of jeopardy, suggesting that we are at the edge of the capacity of the system. However, in Triploid-30, the schedule actually performs quite well. Between the PAs and the residents, there is sufficient capacity to admit essentially all arriving patients and the night floats and jeopardy services are rarely needed. The policy performance is comparable to Initial and Daily Admitting in other metrics as well, although it is benefiting from having a greater staffing level.

Schedule	Initial	Daily Admitting	Hybrid-15	Hybrid-30	Preserved Day	Corkscrew
Night Float Reassignments	1450	1523	1585	812	1395	803
Jeopardy Days	67	7	72	12	3	149
Jeopardy Reassignments	242	16	252	25	7	775
Flex Forward Reassignments	0	0	527	79	17	0
Flex Backward Reassignments	0	0	267	311	304	0
Dropped Patients	10	0	478	38	28	316
Observation $<$ 6 Hours (%)	56.6	78.4	66.1	59.5	73.0	61.2
Shift Runs Late (%)	21.9	24.5	20.3	26.1	32.3	19.5
Interarrival $<$ 2 Hours (%)	40.5	29.5	38.8	37.8	24.8	41.6

Table 5.8: Performance of GMS under two different policies for scheduling residents. This simulation is based on 2009-10 historical data. There were 7218 patients of the course of the year in the simulation.

Finally, in Table 5.8 we compare several the performance of the policies from Table 5.3 for the GMS department. In these policies, GMS used two groups of teams of residents, and has a PA patient in care capacity of either 15 under Initial, Daily Admitting, and Hybrid-15, and 30 under Hybrid-30, Preserved Day and Corkscrew. Under Initial, we see that GMS is at the brink of instability, with some dropped patients and many days of jeopardy. Comparing Initial with Daily Admitting, we see that Daily Admitting has about 100 more night float admissions, but over 200 fewer jeopardy admissions, giving an overall improvement on reassignments. This is due to the hybridization effect of MMMO reducing jeopardy admissions and LOOSSO providing night coverage.

Looking at the other schedules, we see that Preserved Day has less capacity than Initial and that Corkscrew has less capacity than Preserved Day, to the point where the dropped patients necessitate increasing the PA capacity to 30. Preserved Day provides inadequate night float coverage, and corkscrew lacks capacity, but hybridized (in Hybrid 30) together they produce a schedule that takes the better properties from both components. We see that even with the hybridization, we cannot reduce the PA level back to 15 without dropping patients.

5.3.3 Performance Analysis under Increased Patient Volume

In this section, we consider performance of the policies under an increased and decreased patient volume. We look only at dropped patients, jeopardy reassignments, and total reassignments. Here, the total number of reassignments is the sum of the

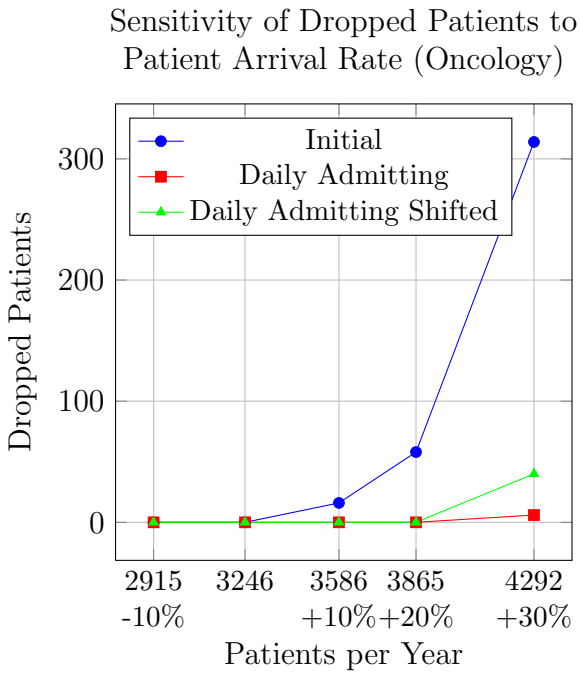

Figure 5-2: Sensitivity of dropped patients to patient arrival rate. Simulations based on 2009-10 historical data for Oncology.

jeopardy reassignments, night float reassignments, and when appropriate the flex forward reassignments (we do not count the flex backward reassignments as from a quality of care perspective, these reassignments arent particularly harmful, see Section 5.2.4).

First, we consider the Oncology department under an increasing patient load, as shown in Figure 5-2 and Figure 5-3. We observe first that the Initial policy begins to break down (i.e. experience dropped patients) with only a 10% increase in patient volume, and completely fails with a 20% increase, while Daily Admitting and Daily Admitting Shifted experience no dropped patients until a 30% increase in patient volume. Next, we observe that when the load is increased by 20%, Daily Admitting causes fewer total reassignments than Initial, despite the fact that Daily Admitting is reassigning most patients arriving off peak hours.

For Cardiology, in Figure 5-4 and Figure 5-5, we observe that Initial and Daily Admitting both have almost no dropped patients, even when the patient arrival rate is increased by 20% (results for Triploid policies are in the appendix). We also see

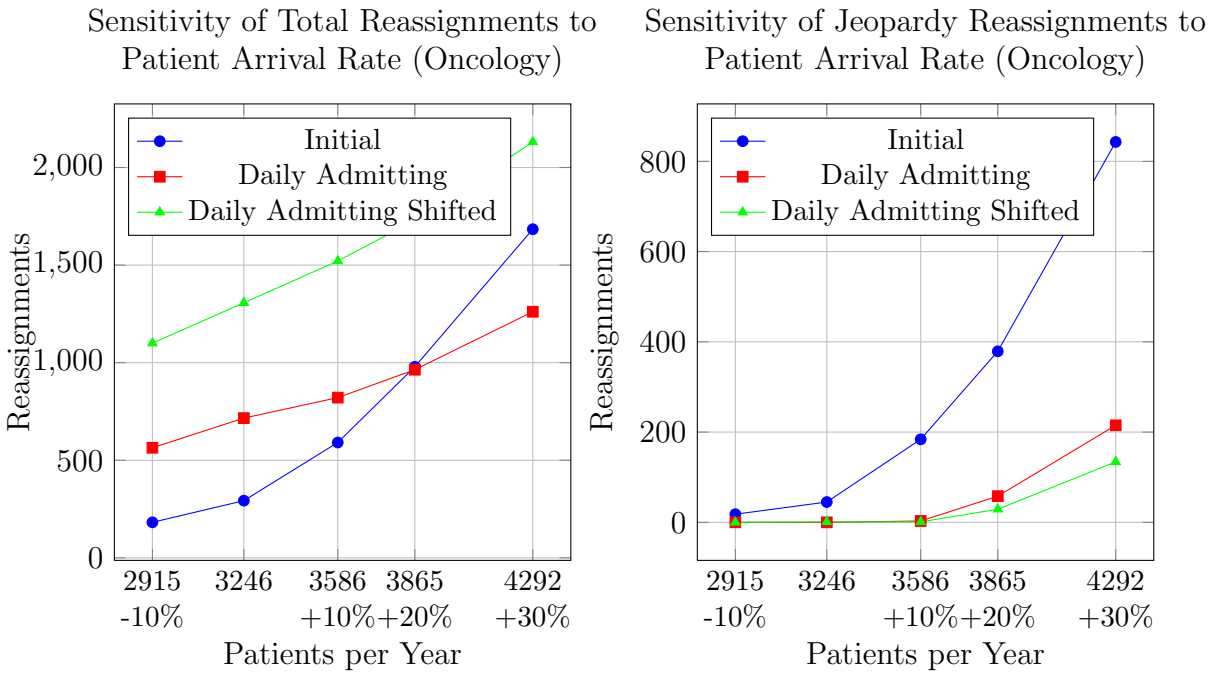

Figure 5-3: Performance of Oncology Department with various policies under changing patient volume. This simulation is based on 2009-10 historical data.

that unlike Oncology, Daily Admitting is always causing more reassignments than the other policies. This is both due the fact that Cardiology is not supported by any PAs under Daily Admitting (as it was with Oncology), and because the system is not critically loaded to the point of Initial failing.

Finally, in Figure 5-6, Figure 5-7 and Figure 5-8 we see that for the GMS department, all policies are right at the edge of their capacity, giving very poor performance with only a 10% increase in the number of patients in the system. One should keep in mind that Initial, Daily Admitting, and Hybrid 15 are using fewer PAs than the other policies (thus we present there results in separate figures).

5.4 Discussion

5.4.1 Key Insights

Here we summarize general insights revealed through our simulation analysis on the influence of the resident policy and the arrival rate of patients on the number dropped

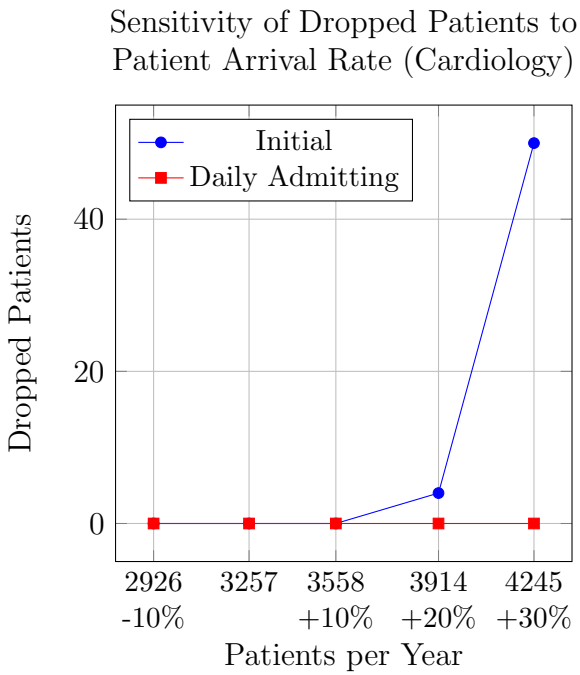

Figure 5-4: Sensitivity of dropped patients to patient arrival rate. Simulations based on 2009-10 historical data for Cardiology.

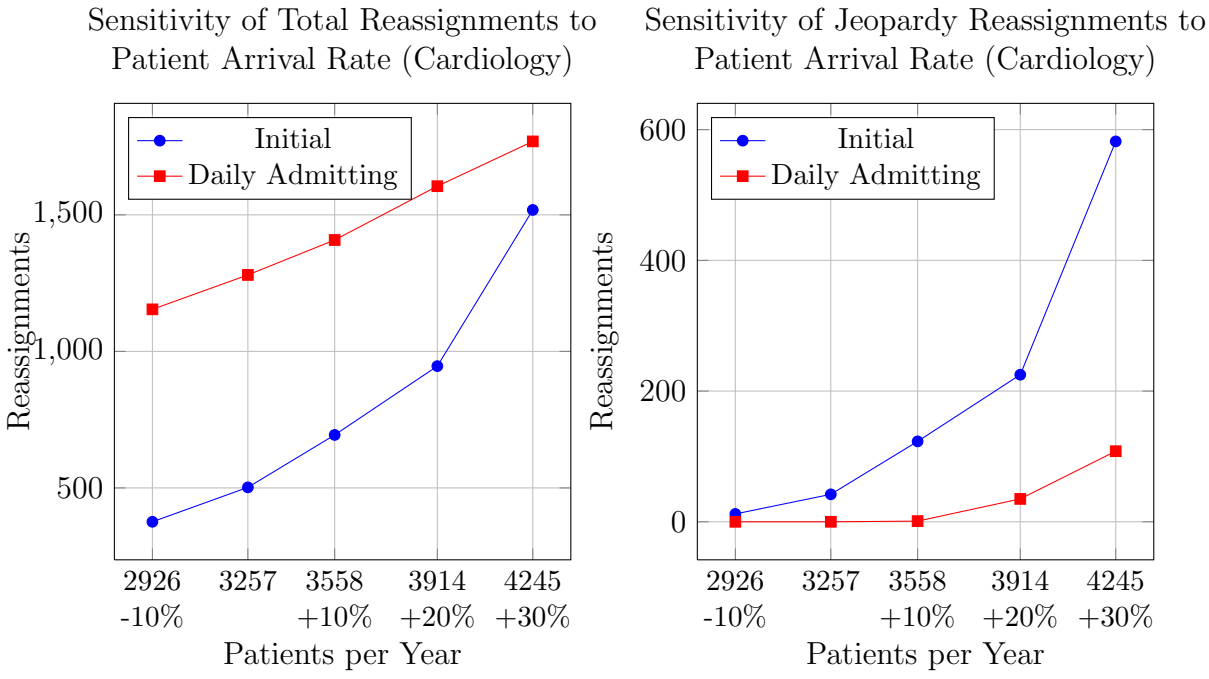

Figure 5-5: Performance of Cardiology Department with various policies under changing patient volume. This simulation is based on 2009-10 historical data.

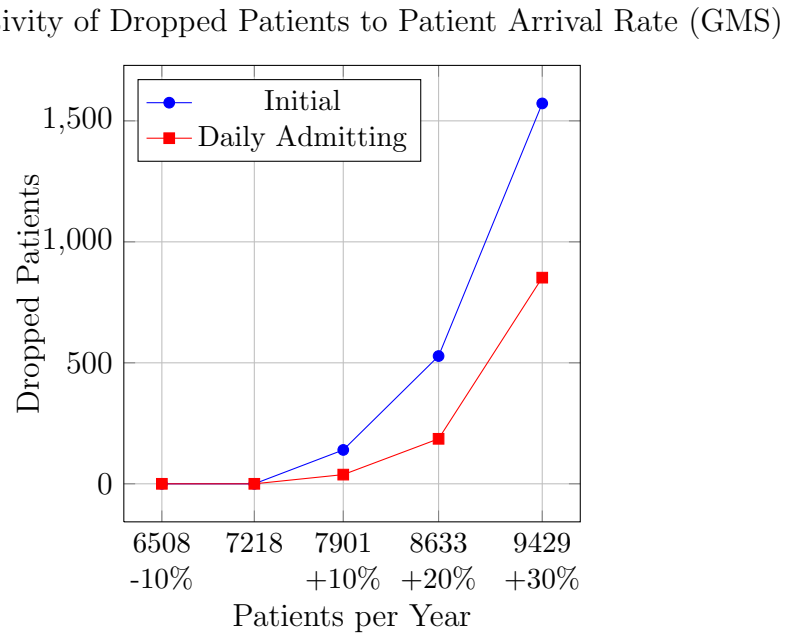

Figure 5-6: Sensitivity of dropped patients to patient arrival rate. Simulations based on 2009-10 historical data for GMS.

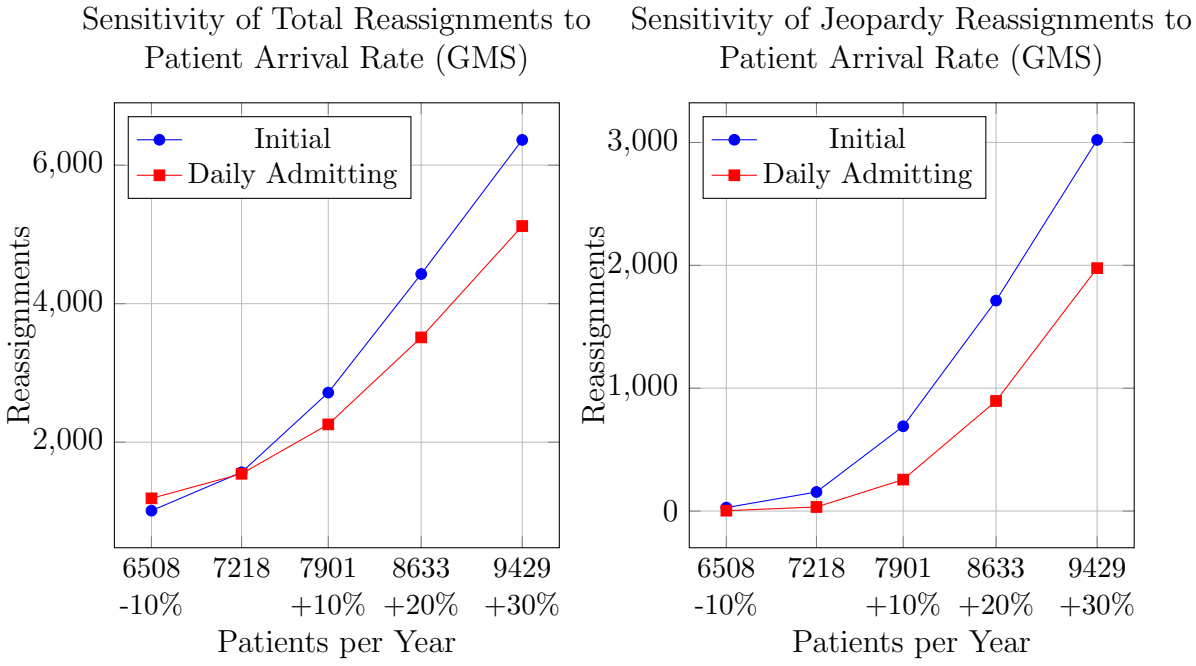

Figure 5-7: Performance of GMS with various policies under changing patient volume. This simulation is based on 2009-10 historical data.

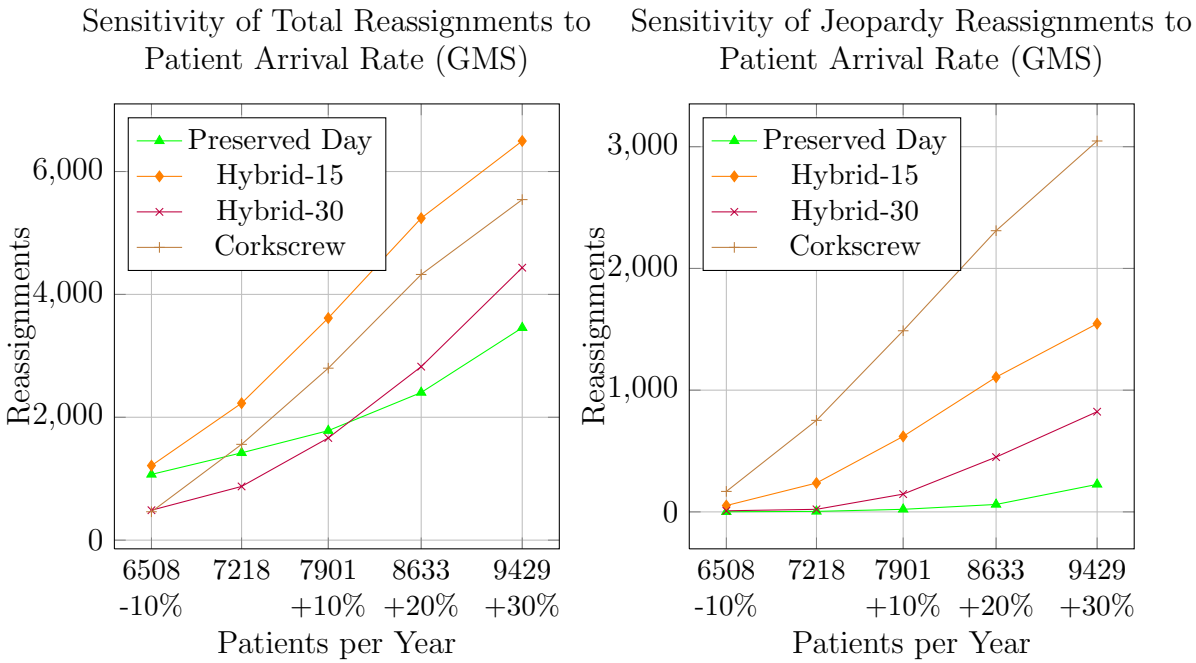

Figure 5-8: Performance of GMS with various policies under changing patient volume. This simulation is based on 2009-10 historical data.

patients and reassignments. In comparing policies, it is important to remember that there are two distinct groups of policies playing under different rules: the first group comparing Initial and Daily Admitting policies, and the second group comparing Preserved Day, Hybrid, Corkscrew, and Triploid policies. For the first group, the constraints on the total number of patients in care are less restrictive than for the second group. However, the first group also generally was assigned lower PA staffing levels. As many variables are being changed at once, it is difficult to draw conclusions from comparisons made between these two groups of policies. For all three departments, the Initial policy and the Daily Admitting policy actually hold these variables relatively constant, and thus provide the best tools for comparison.

We now give three key insights:

1. *Policies with more frequent shorter shifts have a greater capacity than policies with long infrequent shifts and thus are superior under heavy patient loads.*

   We can directly observe this behavior in comparing Initial and Daily Admitting for all three departments. We see that as the patient load increases, more patients are dropped under the Initial policy than the Daily Admitting policy.

As an additional indicator, we see that the Initial policy always results in more instances of jeopardy than Daily Admitting. Note that beyond comparing long shifts versus shorter more frequent shifts, the two policies use equivalent resources. Each resident can admit up to ten patients every four days and have up to ten patients in care simultaneously. Additionally, the requirements for number of residents, PAs, and admitting hours per four day rotation are all about the same. The difference in capacity arises because admitting more frequently allows residents to better maintain close to 10 patients in care at all times. A mathematical justification of this phenomenon is further explored in Section 5.5 and again in Chapter 6.

2. *As the patient load increases, the number of reassignments and dropped patients increases rapidly.*

   This can be observed for all policies, although the effect is most noticeable for Oncology under the Initial policy, and under all policies in GMS. This kind of non-linear performance degradation as the system approaches capacity is typical in capacitated systems.

3. *Schedules in which the admitting hours of doctors are aligned with the arrival rate of patients result in significantly fewer reassignments.*

   This can most clearly be seen in comparing the policies Daily Admitting and Daily Admitting Shifted for the Oncology department. Daily admitting focuses all admitting capacity on the window 2pm-10pm, while Daily Admitting Shifted focuses all admitting capacity on the window 10am-6pm. Under Daily Admitting, we have 676 total reassignments, while under Daily Admitting shifted, we have 1298 total reassignments. Additionally, observing that the number of dropped patients under these policies is almost the same, we see that capacity and alignment are two independent issues.

   We also see that for GMS, the Preserved Day Schedule results in many more reassignments than Hybrid-30, as Preserved Day fails to cover the latter half of peak admitting hours as well.

5.4.2 Assumptions and Limitations of the Model

We now identify some of the assumptions and limitations of our study. First, we look at the alignment between the measured outcomes optimized for by our choice of shift schedules, and actual hospital goals. Next, we briefly discuss possible sources of modeling error. Finally, we consider several studies contradicting the basic assumptions motivating shorter shifts, and discuss some plausible interpretations of seemingly contradictory evidence.

The goals of medical residency programs include (a) delivering high quality patient care, (b) providing residents with a good educational experience, (c) providing residents with a safe work environment and healthy work life balance, and (d) providing services economically. Our study does not consider the effect of shift schedules on (b), although a recent study found that residents perceived that the post regulation shift schedules left less time for education [27]. For (d), schedules with a greater capacity correspond to requiring fewer total interns to treat a fixed number of patients, suggesting that shorter more frequent shifts are financially more viable. Furthermore, jeopardy admissions are costlier than night float admissions. Therefore, since schedules with shorter more frequent shifts reduce jeopardy admissions while increasing night float admissions, such schedules are more cost effective. However, as we do not explicitly account for the increased expense of additional night floats required, a more careful analysis is warranted. Additionally, we are ignoring indirect changes in expenses, e.g. if quality of care is increased, some expensive accidents may be avoided. In [62], a more comprehensive estimate of the cost for restricting duty hours was estimated, but at a less granular scale (i.e. costs for the national residency program, not a single hospital). Regarding (a) and (c), in light of the studies referenced in the introduction, one might assume that a schedule relying on shorter shifts that allows residents to sleep more regularly and have more sleep hours in total should improve the quality of care (due to a reduction in errors caused by fatigue), reduce occupational hazards for residents, and improve resident quality of life. This opinion was held by some residents prior to implementing the 2011 regulations [26],

although there was a strong concern that a loss in continuity of care would cause more accidents than those prevented by reducing fatigue. Our study addresses (a) by quantifying the change in continuity of care when using schedules with shorter more frequent shifts, focusing on the first 24 hours of a patient's stay in the hospital (e.g. as measured by reassignments and frequency of six hours of observation after admission), but makes no further attempt to quantify the benefits for (a) and (c) in using shorter shifts due to reduced fatigue.

Next, we briefly discuss potential sources of modeling error. In particular, we note that our simulation model is an extreme simplification of how hospitals actually operate. As a result, our conclusions may only be valid under certain assumptions implicit in our model. We quickly give two such examples. First, we assume that the only bottleneck in admitting patients is the availability of doctors. However, in an actual hospital setting, there are many potential bottlenecks to admitting patients, including beds, nurses, and specialized equipment. We made this assumption because the practitioners in B&W suggested that this was correct up to first order in B&W in 2009-11. However, if the patient load increased 20% as in some of our sensitivity analysis, most likely another resource would become a bottleneck. Second, we considered the distribution of arrival times and departure times exogenous. However, adjusting schedules would likely have some effect on both. A small but non-negligible fraction of patient arrivals are scheduled by doctors at their own will, so changing their hours would likely change their incentives when scheduling patients. Additionally, patients may adapt to avoid waiting times if under some schedules, arriving a certain time of day often resulted in long waits. Departure times are determined by when doctors choose to discharge patients, so again, altering their schedules could have unintended consequences. Thus, we suggest using some caution when applying these results, particularly against expecting certain quantitative outcomes.

Several recent studies of the aftermath of the 2011 regulation have challenged some of the assumptions we adopted on a number of grounds. In [27, 76], the authors find (by issuing surveys to residents) that after the change in regulation, residents are on average not getting much more total sleep, some residents are violating new duty

hour restrictions, more medical accidents take place, and the rates of depression and burnout are not lower. There are many possible explanations as to why the authors may have reached these somewhat counter-intuitive conclusions. For example, if the duty hour restrictions were not enforced, we would not expect any improvement in resident quality of life. Instead, if duty hours were enforced but hospitals were not properly prepared to adapt operationally, residents might experience additional stress on the job and have more errors due to general disorganization in the new system (particularly in the handoff protocols). Finally, as these studies were based on surveying residents and physicians perceptions, they are subject to bias in the perceptions of the participants over something which is an increasingly political issue. These difficulties, whether perceived or actual, may improve with time as hospitals adapt their operations to improve handoffs, and more senior residents, who may feel victimized by the changes in regulations (which increased their work load to reduce the workload on first year residents) graduate from the residency program, as suggested in [35] after the 2003 regulations. Additionally, one should note that there are a few new studies challenging the notion that fatigue results in worse medical outcomes [30, 84]. However, these studies were performed as retrospective analyses, with many potential confounding factors that could lead to erroneous conclusions.

5.5 The Markov Chain Throughput Upper Bound

In this section, we give a probabilistic analysis of the long run average rate that a resident or team of residents can treat patients, as used to compute the *Markov chain throughput upper bound* in Section 5.2.3. Using this tool, we give mathematical justification for why a schedule that has short, frequent, and evenly spaced shifts such as MMMO should have a greater capacity than a schedule with fewer, longer, unevenly spaced shifts such as LOSO. First, we consider a simple model where a single resident that only has the capacity to treat one patient at a time must treat an infinite incoming stream of patients. We consider two schedules for the resident and compute the long run average number of patients treated under each schedule. In this

special case, we prove that short evenly spaced shifts have a greater capacity than long shifts. Then, we generalize the model and provide an algorithm to compute the maximum rate that the residents and PAs of an entire department can treat patients.

In the basic model, we have an infinite number of time periods $t = 1, 2, \ldots$, each representing a single day. The resident has a four day rotating schedule that specifies if the resident is *on shift* or *off shift*. We consider two schedules for the resident. The first, which we refer to as *Consecutive* (C), is On, On, Off, Off, and the second, which we refer to as *Spread* (S) is On, Off, On, Off. At the start of each day, if the resident is on shift and does not currently have a patient in care, the resident takes this new patient. Then each day, regardless of whether or not the resident is on shift, the patient in care leaves with probability p (a parameter of the model, $0 < p < 1$), and stays otherwise. We now determine the long run average number of patients that our resident will treat under schedules (S) and (C). Under policy (S), for each day $t = 1, 2, \ldots$, we define $D_{\mathrm{S}}(t)$ be one if a patient departed in time period t and zero otherwise. Likewise, under policy (C), for each time period $t = 1, 2, \ldots$, we define $D_{\mathrm{C}}(t)$ to be one if a patient departed in time period t and zero otherwise. Using this notation, the long run average number of patients treated by our resident under each policy, denoted $D_{\mathrm{S}}(\infty)$ and $D_{\mathrm{C}}(\infty)$, is given by

$$D_{\mathrm{C}}(\infty) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{C}}(t)$$
$$D_{\mathrm{S}}(\infty) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{S}}(t).$$

By the law of large numbers, we can compute the long run average behavior by instead directly computing the average behavior in a four day cycle, namely,

$$D_{\mathrm{C}}(\infty) = \mathbb{E}\left[\frac{1}{4}\sum_{t=1}^{4} D_{\mathrm{C}}(t)\right] = \frac{1}{4}\sum_{t=1}^{4} \mathbb{E}[D_{\mathrm{C}}(t)] = \frac{1}{4}\left(p + p + (1-p)p + (1-p)^2 p\right),$$
$$D_{\mathrm{S}}(\infty) = \mathbb{E}\left[\frac{1}{4}\sum_{t=1}^{4} D_{\mathrm{S}}(t)\right] = \frac{1}{4}\sum_{t=1}^{4} \mathbb{E}[D_{\mathrm{S}}(t)] = \frac{1}{4}\left(p + (1-p)p + p + (1-p)p\right).$$

We briefly justify this calculation. Any day the resident is on shift, the probability of discharging a patient will be p, independent of the past. Likewise, if a resident was on shift the previous day, and is off shift today, then the probability of discharging a patient can be computed as $\mathbb{P}(\text{no discharge yesterday})\mathbb{P}(\text{discharge today}) = (1-p)p$, independent of events occurring greater than one day ago. By induction, it is simple to show that if a resident last was on shift n days ago, then the probability of discharge today is given by $(1-p)^n p$, and is independent of events occurring earlier than n days previous. Observe that the terms $\mathbb{E}[D_{\mathrm{C}}(1)]$ and $\mathbb{E}[D_{\mathrm{C}}(2)]$ are equal to the probability of discharging a patient while on shift, and the terms $\mathbb{E}[D_{\mathrm{C}}(3)]$ and $\mathbb{E}[D_{\mathrm{C}}(4)]$ are equal to the probability of discharging one day after being on shift and two days after being on shift, respectively. This justifies the calculation for $D_{\mathrm{C}}(\infty)$. The computation of $D_{\mathrm{S}}(\infty)$ is justified similarly.

We observe that for every possible value of p (the probability patients depart each day),

$$D_{\mathrm{S}}(\infty) - D_{\mathrm{C}}(\infty) = \frac{1}{4}(1-p)p - \frac{1}{4}(1-p)^2 p = \frac{1}{4}(p^2 - p^3) > 0,$$

i.e. $D_{\mathrm{S}}(\infty) > D_{\mathrm{C}}(\infty)$.

From this calculation, we can conclude the resident will in the long run treat more patients under (S) than under (C).

We now extend the model so that residents can treat multiple patients simultaneously, but must obey a capacity on the maximum number of patients in care and the maximum number of patients admitted per day (according to a schedule). Again, we have an infinite number of time periods $t = 1, 2, \ldots$, each corresponding to a day. For now, we will restrict ourselves to the schedules from Section 5.2.3 where the (C1) constraint (team patients in care capacity) is dominated by (C3) constraint (individual patients in care capacity), i.e. LOSO and MMMO. We will let θ denote such a schedule. Let c_θ denote the maximum number of patients in care for a resident. Let $a_\theta(t)$ be the maximum number patients that can be admitted on day t under policy θ. For example, under LOSO, $a_\theta(1) = 7, a_\theta(2) = 0, a_\theta(3) = 3, a_\theta(4) = 0, a_\theta(5) = 7, \ldots$. For $t = 0, 1, \ldots,$, let $X_\theta(t)$ be the number of patients in care in time period t. Finally,

let $D_\theta(t)$ be the number of patients that depart in period t. We will assume that $D_\theta(t)$ has the distribution $\text{Bin}(X_\theta(t-1), p)$, i.e. each patient in care departs with probability p, independently of the other patients, as in the previous section. We assume that an initial condition $X_\theta(0)$ is given, and then the dynamics of our model are given by

$$X_\theta(t) = \min\{X_\theta(t-1) + a_\theta(t) - D_\theta(t), c\}$$

for $t = 1, 2, \ldots$. When $a_\theta(t)$ is on an n day rotating schedule, then $X_\theta(t)$ is a Markov chain with period n on the state space $\{0, 1, \ldots, c\}$. It is easy to see that the process $X_\theta(nt)$, $t = 0, 1, 2, \ldots$, will be a time homogeneous Markov chain on the same state space with a single recurrent class. Thus we can solve for the long run behavior of $X_\theta(nt)$ by computing the stationary distribution, i.e. solving a system of $c + 1$ linear equations. Once the stationary distribution is known, it is straight forward to compute the expected number of departures per n day rotation (and thus per day).

To improve the accuracy of our results, we make the following enhancement to the model. For shifts that do not occur on weekends, we make $a_\theta(t)$ random, taking the value zero with probability 2/7, and the maximum number of admissions otherwise. While it would be more accurate to make the day of the week as well as the day in the n day cycle both part of the state, it would increase the number of states from nc_θ to potentially $7nc_\theta$ with little practical gain.

Finally, we discuss the case when constraint (C1), resident team capacity, is not dominated by constraint (C3), individual resident capacity. When this happens, on our particular problem instances the (C1) constraint is much more restrictive than the (C3) constraint. Therefore it is a reasonable approximation to simply ignore the (C3) constraint and view the entire team as a single resident with the combined admitting schedule of all residents on the team. This technique works very well for D'OSO where the team capacity is 15 patients but the individual capacity is ten patients, and there are two residents on a team with no offset in their schedules (i.e. they both have "D"' on the same days). However, for other schedules where the residents are staggered, e.g. LOOSSO, things are a little more complicated. We would to say

Schedule	Residents	c_θ	$a_\theta(1)$	$a_\theta(2)$	$a_\theta(3)$	$a_\theta(4)$	$a_\theta(5)$	$a_\theta(6)$	$a_\theta(7)$	$a_\theta(8)$	$a_\theta(9)$
LOSO	1	10	7	0	3*	0					
MMMO	1	10	4	4*	4	0					
LOOSSO	3	20	10	7	7	3*	3*	3*			
D'OSO	2	15	10	0	6*	0					
DOOONO	3	20	10	5	5	0	5	5			
DOSOOONOO	3	20	5	5	8	3	3	0	5	5	5

Table 5.9: The values of c_θ and $a_\theta(t)$ used for each schedule for GMS and Cardiology. Values of $a_\theta(t)$ marked with * take the value zero with probability 2/7, to approximate that some shifts are skipped on weekends.

Schedule	Residents	c_θ	$a_\theta(1)$	$a_\theta(2)$	$a_\theta(3)$	$a_\theta(4)$
LOSO	1	10	5	0	1*	0
MMMO	1	10	3	3*	3	0
M'M'M'O	1	10	3	3*	3	0

Table 5.10: The values of c_θ and $a_\theta(t)$ used for each schedule for Oncology. Values of $a_\theta(t)$ marked with * take the value zero with probability 2/7, to approximate that some shifts are skipped on weekends.

that the capacity is $[7,0,0,3,3,0] + [0,7,0,0,3,3] + [3,0,7,0,0,3] = [10,7,7,3,6,6]$. However, on the first day, when it falls on weekend, we only lose part, not all of the capacity. While this could be correctly accounted for, instead, for simplicity, when this situation occurs, we just assume that the admitting capacity equals ten, regardless of whether it is a weekday or weekend.

We use this technique to make the Markov chain throughput upper bound for residents and teams of residents. We omit the details of the calculations but summarize the inputs to the calculation in Table 5.9 for GMS and Cardiology and Table 5.10 for Oncology, and the results of the calculation in Table 5.11 for GMS, in Table 5.12 for Cardiology, and in Table 5.13 for Oncology. Note that the "Residents" column indicates the number of individuals that computation is for. Using these quantities, it is straightforward to compute the Markov chain throughput upper bound for each policy, as you only need to multiply the estimate for a resident (or team of resident) by the number of residents (teams of residents), and then add in the PA capacity, as given in Table 5.14 (multiplied by the capacity of the PA service divided by 15).

Another improvement to the model would be to replace the infinite stream of patients waiting to be treated by a random number of patients arriving each day.

		Throughput Upper Bound (Patients/Day)		
Schedule	Residents	Capacity	Admitting	Markov
LOSO	1	2.25	2.29	1.78
MMMO	1	2.25	2.71	2.00
LOOSSO	3	4.50	5.79	4.17
D'OSO	2	3.37	3.57	2.75
DOOONO	3	4.50	5.00	3.26

Table 5.11: The long run average number of departures per day under each schedule for GMS.

		Throughput Upper Bound (Patients/Day)		
Schedule	Residents	Capacity	Admitting	Markov
LOSO	1	2.13	2.29	1.71
MMMO	1	2.13	2.71	1.91
DOSOOONOO	3	4.26	4.14	3.61

Table 5.12: The long run average number of departures per day under each schedule for Cardiology.

		Throughput Upper Bound (Patients/Day)		
Schedule	Residents	Capacity	Admitting	Markov
LOSO	1	1.58	1.43	1.23
MMMO	1	1.58	2.04	1.44

Table 5.13: The long run average number of departures per day under each schedule for Oncology.

Department	Average Patient Stay (days)	Throughput of Capacity 15 PA Team (patients/day)
GMS	4.44	3.37
Cardiology	4.70	3.19
Oncology	6.32	2.37

Table 5.14: Average patient length of stay by department and corresponding PA throughput.

Then, as in reality, the resident would occasionally be idle. Analyzing such a model is significantly more complicated than the above analysis, and is the subject of Chapter 6. However, the a key finding that chapter is essentially as follows: Suppose that each day the average arrival rate of patients is λ. Then the number of patients waiting to be treated will increase to infinity as $t \to \infty$ if and only if λ exceeds the throughput of the policy.

5.6 Statistical Analysis of Patient Flows

5.6.1 Patient Arrival and Departure Data

The patient data we use is from the B&W Oncology, Cardiology and General Medicine departments. Recall that a resident or PA is assigned to a patient, and in turn each patient is assigned to a bed. At B&W, these events occur simultaneously. For about 75% of patients, we have this exact time. The remaining 25% of patients were transfers from other hospitals. For these patients, we use as a proxy the time the patient reached the bed, as for transfer patients these times are relatively close. For all patients, we know the time the patient was discharged. Each of the three departments has specialist teams that instead of taking any new arriving patient, only take new patients with a particular ailment, and take all such patients. For example, General Medicine has two PA teams, "REN-PA" for patients with renal failure and "PUL-PA" for patients with chronic pulmonary disease. As such patients are treated separately in an independent system, we excluded them from our model.

5.6.2 A Statistical Model for Patient Arrivals

In order to simulate a schedule with a patient loads above the historical patient load, we need to create additional patients to augment the historical data. For each of the three departments we considered, we used the same procedure described below. We assume that the patient arrival process is a non-stationary Poisson process, where for each hour of the day, the rate is constant. We let λ_{hn} denote the arrival rate for hour

h for day n of the simulation. Let $d(n)$ be the day of the week for simulation day n and let $m(n)$ be the month of the year for simulation day n. Further, we assume that for each hour of the day h, each day of the week d, and each month of the year m, there exists constants λ_h, λ_d and λ_m such that

$$\lambda_{hn} = \lambda_h + \lambda_{d(n)} + \lambda_{m(n)}.$$

We produce estimates of our unknowns $\hat{\lambda}_h$ of λ_h for each hour h, $\hat{\lambda}_d$ of λ_d for each day d, and $\hat{\lambda}_m$ of λ_m for each month m by solving the regularized least squares problem below. Let a_{hn} denote the actual number of arrivals we observed in hour h of day n. We choose a small constant $\varepsilon > 0$ to be our regularizer, and then solve

$$\begin{aligned} \min_{\hat{\lambda}_h, \hat{\lambda}_d, \hat{\lambda}_m, \hat{\lambda}_{hn}} \quad & \sum_{n=1}^{365} \sum_{h=1}^{24} (a_{hn} - \hat{\lambda}_{hn})^2 + \varepsilon \left(\sum_{h=1}^{24} \hat{\lambda}_h + \sum_{d=1}^{7} \hat{\lambda}_d + \sum_{m=1}^{12} \hat{\lambda}_m \right) \\ \text{subject to} \quad & \hat{\lambda}_h + \hat{\lambda}_{d(n)} + \hat{\lambda}_{m(n)} = \hat{\lambda}_{hn} \qquad h = 1, \ldots, 24,\ n = 1, \ldots, 365. \end{aligned}$$

The regularization is necessary only for technical reasons and the solution was relatively insensitive to the choice of ε. The optimization problem can be solved efficiently by most commercial packages for scientific computing (e.g. SciPy).

To assure the reader that the patient arrival rate does depend on all three of these variables and allow the reader the visualize these relationships, we plot the average number of patients per hour, the average number of patients per day of week, and the number of patients per month of the year using the 2009-10 data for Oncology (Figure 5-9), Cardiology (Figure 5-10) and GMS (Figure 5-11).

Given these estimators, to increase the arrival rate by 5%, we would generate a non stationary Poisson process with a fixed rate in each hour of the day h and each day of the simulation n of $.05 \cdot \lambda_{hn}$. We would then add these artificial arrival times to the historical data.

We use a simpler technique to reduce the arrival rate by 5%. We simply select 5% of the patients at random from the historical data, and remove them from the simulation.

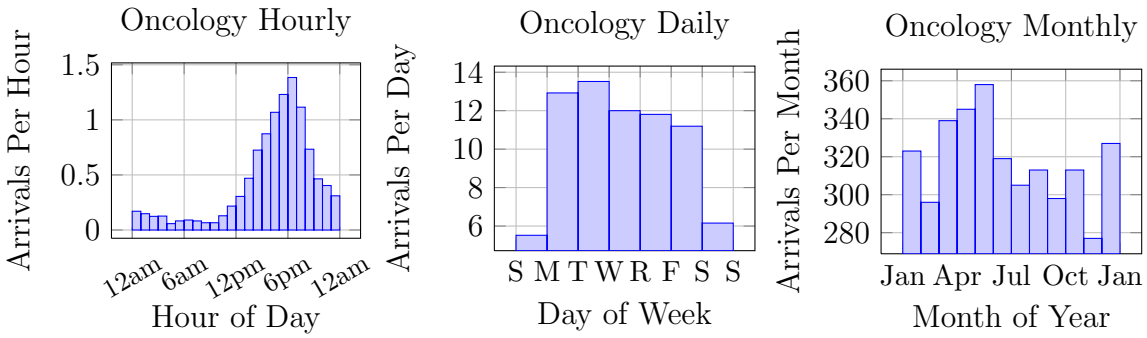

Figure 5-9: For the Oncology Department in 2009-10, the average number of arrivals by hour of day, day of week, and month of year.

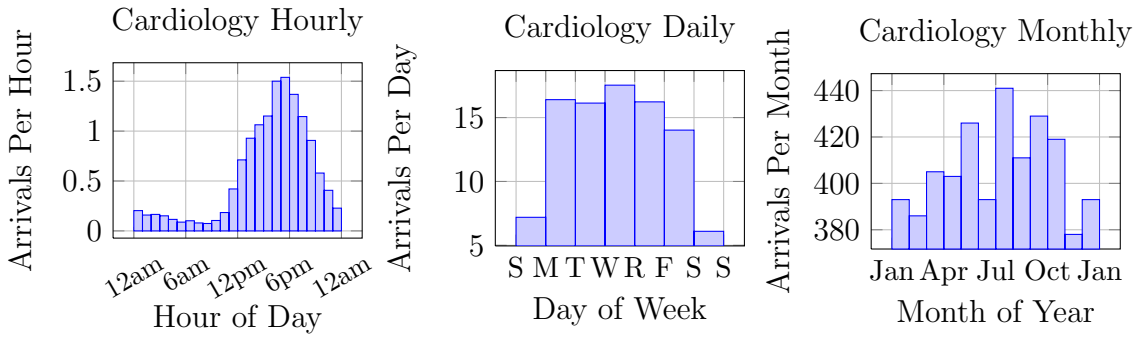

Figure 5-10: For the Cardiology Department in 2009-10, the average number of arrivals by hour of day, day of week, and month of year.

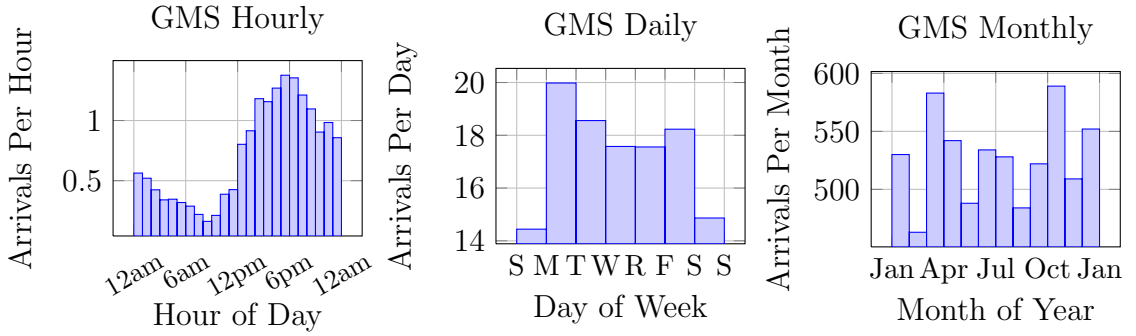

Figure 5-11: For the GMS in 2009-10, the average number of arrivals by hour of day, day of week, and month of year.

5.6.3 A Statistical Model for Patient Departures

For the synthesized patient data from the previous section, we need a statistical model to determine their length of stay in the hospital. Here it is problematic to assume any parametric form. In Figure 5-12, we see a sinusoidal pattern with a period of one day in the length of stay. Looking closely, splitting our patients by the hour of the day that they arrived, we see that while the length of stay is dependent on the arrival time, the time of day of departure is not. Specifically, regardless of when a patient arrived, they will typically leave around noon. However, the number of days a patient stays in the hospital appears to depend on the time of day arrived. For example, a patient arriving late at night is unlikely to depart the following day at noon.

To systematically test the relationship between the arrival hour of day, day of week, and month of year with the patient length of stay, we used a regression model similar to the model used in the previous section. We found that the hour of day and day of week were strongly correlated to the patient length of stay, but that the month of year was not particularly relevant.

To actually generate the random length of stay each synthesized patients, instead of fitting our data to some parametric distribution and then sampling, as we did in the previous section, we instead draw from the empirical distribution in our data. Specifically, if we had a synthesized arrival in hour h and day of week d, we looked at all patients that arrived in the same hour of the day on the same day of the week, and picked one of their length of stays uniformly at random. Because we had hundreds of data points for every hour day pair and were generating relatively few points, this was a reasonable approach. Had we not ruled out the month of year dependence, we would have had insufficient data points for this method to give reliable results.

Finally, we describe the statistical test used to determine that the month of the year was not an important factor in determining the length of stay for patients, but that the hour of the day and the day of the week were important. We assumed that when a patient arrived on day n during hour h, they would have an average length of stay of ℓ_{hn} hours Using the same notation as in the previous section, we assumed

that there were constants ℓ_h, $\ell_{d(n)}$ and $\ell_{m(n)}$ such that

$$\ell_{hn} = \ell_h + \ell_{d(n)} + \ell_{m(n)}.$$

Then for each patient $i = 1, \ldots, k$, where k is the number of patients treated in the department for a year, let a_i be the actual length of stay for that patient, h_i be the hour of day of the patient arrival, n_i be the day of the year of the patient arrival, m_i be the month of year of patient arrival, and d_i be the day of week of the patient arrival. We then computed our estimates by solving

$$\begin{aligned} \min_{\hat{\ell}_h, \hat{\ell}_d, \hat{\ell}_m, \hat{\ell}_{hn}} \quad & \sum_{i=1}^{k} (a_{h_i n_i} - \hat{\ell}_{h_i n_i})^2 + \varepsilon \left(\sum_{h=1}^{24} \hat{\ell}_h + \sum_{d=1}^{7} \hat{\ell}_d + \sum_{m=1}^{12} \hat{\ell}_m \right) \\ \text{subject to} \quad & \hat{\ell}_{h_i} + \hat{\ell}_{d(n_i)} + \hat{\ell}_{m(n_i)} = \hat{\ell}_{h_i n_i} \qquad i = 1, \ldots, k, \end{aligned}$$

Running the test, we observed that the influence of the month term in determining $\hat{\ell}_{hn}$ was an order of magnitude smaller than the day of week term or hour of day term and could safely be ignored. The hour of day term was more important than the day of week term. There was some discrepancy in the day of week terms was between weekdays and weekends.

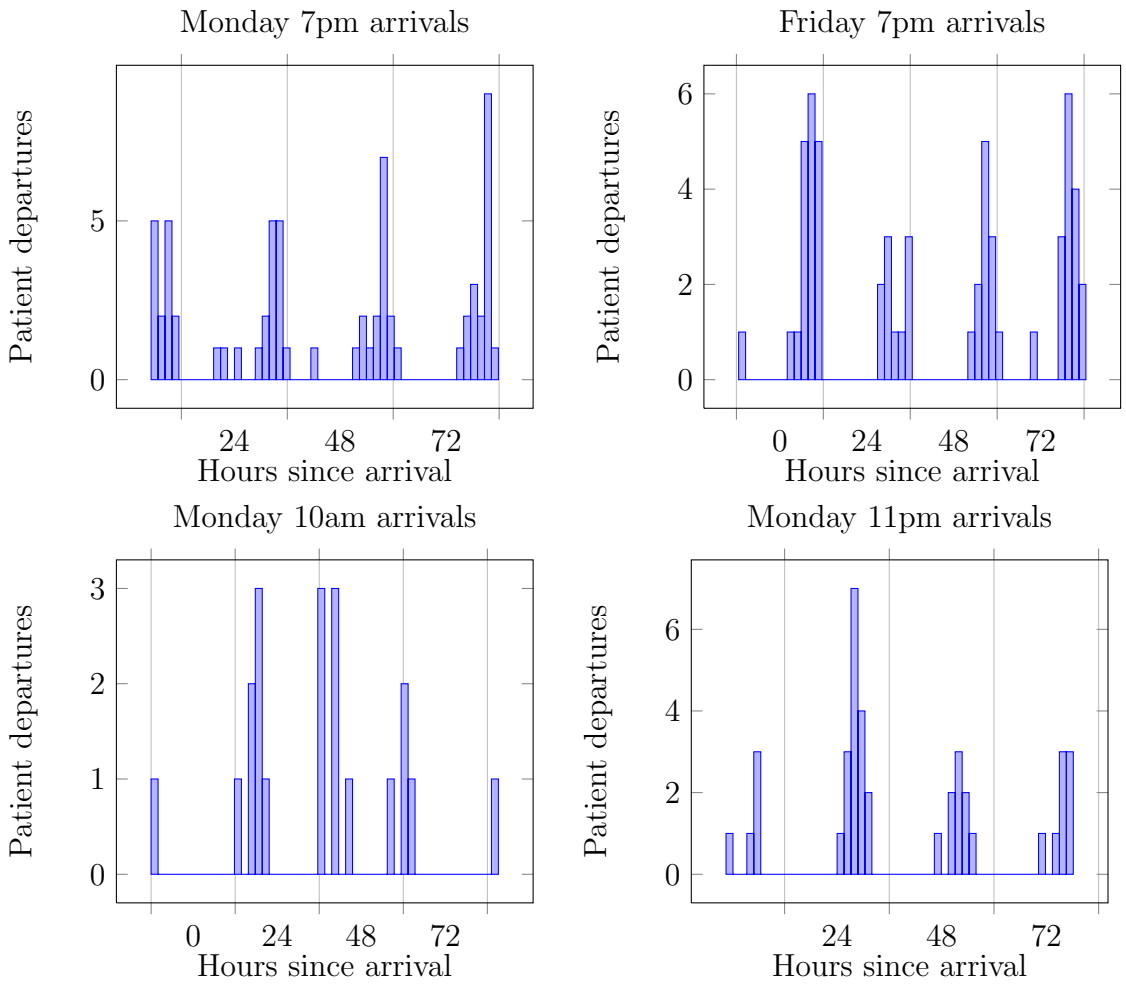

Figure 5-12: The distribution for patients' length of stay is heavily influenced by the time of day the patient arrives, as patients all tend to leave the hospital at midday. Patient length of stays exceeding 100 hours not shown above.

Chapter 6

Queuing and Fluid Models for Scheduling Medical Residents in Hospitals

6.1 Introduction

Major hospitals face a difficult challenge of designing shift schedules for their residents that satisfy demand, provide quality care, and are compliant with regulations restricting shift lengths. Motivated by empirical work conducted by the authors at the Brigham and Women's (B&W) Hospital in Boston, we analyze the impact of shift lengths on two key performance metrics. The first metric is admitting capacity—the largest patient arrival rate sustainable by a given shift schedule. The second metric is the number of reassigned patients—the number of patients admitted temporarily by one doctor and then permanently transferred to a resident.

We build a queueing model to compare two shift scheduling policies that are representative of the alternatives encountered in hospitals: one where residents work long shifts on alternating days, called Long Shifts (LS), and another where residents admit patients daily in short shifts, called Daily Admitting (DA). We determine the admitting capacity for our queueing model under each policy. Then we construct a

fluid model—a large scale approximation of the underlying queueing model. We show that for each policy, the fluid model has a unique steady state solution. Finally, we establish an interchange of limits between the stochastic and fluid models in steady state. We use these results to compare the key performance metrics under the two policies.

Our analysis shows that the DA policy has a greater capacity to admit patients than the LS policy for all parameter choices. Furthermore, we numerically establish the existence of a threshold value, such that the number of reassigned patients is smaller for the DA policy than for the LS policy if and only if the arrival rate of patients is greater than the threshold value. Since most hospitals operate at near critical loads, our two findings lead to the conclusion that schedules which rely on shorter more frequent shifts than those found in practice would increase admitting capacity and reduce the number of reassigned patients.

Organization

The remainder of this chapter is organized as follows. The queueing model and its fluid limit are described Section 6.2 and the main results are stated there. In Section 6.3 we numerically solve for the steady state behavior of the fluid model and discuss the performance implications for our queuing model. Then we give some concluding remarks in Section 6.4. The proofs of the main results are in the following sections. In Section 6.5, we exactly characterize the stability of our queueing model under each policy using a simple linear Lyapunov function type argument. In Section 6.6, we use quadratic Lyapunov functions to bound the expected steady state queue length. In Section 6.7, we prove the existence of the fluid limits, applying the results in [60]. Then in Section 6.8, we show that the fluid limit has a consistent periodic long run behavior under each policy, where the solution in each period is characterized by a simple system of differential equations. In Section 6.9, we prove that the long run solution to the fluid model approximates the steady state queue lengths of the underlying queueing model. Justifying this requires an argument for an "interchange of limits." As in [38], we use our moment bound from Section 6.6 to show tightness

of the rescaled stationary distributions, and then we follow the technique of [31] and similarly [79] to prove the interchange of limits. In Section 6.10, we use the result of Section 6.9 to show that the long run number of daily reassignments converges in the fluid rescaling converges to a natural function of the fluid limit. Finally, we have two rather technical sections: Section 6.11, where we show several elementary properties of the solution to a differential equation, and Section 6.12, where we use another Lyapunov function argument to distinguish between the null recurrent and transient cases in our queueing model.

Summary of Notation

We conclude with a summary of the mathematical notation used in the chapter. Throughout, $\mathbb{R}$ ($\mathbb{R}_+$) denotes the set of (nonnegative) reals, and likewise, $\mathbb{Z}$ ($\mathbb{Z}_+$) denotes the set of (nonnegative) integers. For a vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_p = (\sum_{i=1}^n x_i^p)^{1/p}$ is the ℓ_p-norm. The ℓ_1 ball of radius r is denoted $B_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^2 \mid \|\mathbf{x} - \mathbf{y}\|_1 < r\}$. For $x, y \in \mathbb{R}$, $x \wedge y = \min\{x, y\}$ and $(x)^+ = \max\{x, 0\}$. We define $f(t^-)$ as $\lim_{\tau \nearrow t} f(\tau)$ when the limit exists. We let $\text{Exp}(\mu)$, $\text{Pois}(\lambda)$, and $\text{Bin}(n, p)$, denote an exponential random variable with mean $1/\mu$, a Poisson random variable with mean λ, and a Binomial random variable with mean np and variance $np(1-p)$, respectively (these moments characterize the distributions). If the sequence of random vectors $\mathbf{X}^n$, $n = 1, 2, \ldots$, converges weakly (in distribution) to $\mathbf{X}$ as $n \to \infty$, we say $\mathbf{X}^n \Rightarrow \mathbf{X}$. For a stochastic process $X(t)$ in either discrete or continuous time, $\mathbb{E}_x[X(t)]$ denotes $\mathbb{E}[X(t) \mid X(0) = x]$. A sequence of continuous time vector valued stochastic processes $\mathbf{X}^n(t)$ on a common probability space Ω converges almost surely (a.s.) and uniformly on compact sets (u.o.c.) to a deterministic function $\mathbf{x}(t)$ if for every $t > 0$ and almost every $\omega \in \Omega$,

$$\lim_{n \to \infty} \sup_{0 \le s \le t} \{\|\mathbf{X}^n(s, \omega) - \mathbf{x}(s)\|_1\} = 0,$$

where $\|\cdot\|_1$ is the 1-norm for vectors. See [19] for more details. As in [28] sections 11.2-11.3, for functions $f \colon \mathbb{R}^n \to \mathbb{R}$, we let $\|f\|_L = \sup_{\mathbf{x}, \mathbf{y} \in \mathbb{R}^n} |f(\mathbf{x}) - f(\mathbf{y})| / \|\mathbf{x} - \mathbf{y}\|_1$

denote Lipschitz semi-norm (when f is a Lipschitz function, the value of this norm is the smallest Lipschitz constant that f satisfies). We let $\|f\|_{\mathrm{BL}} = \|f\|_L + \|f\|_\infty$. This quantity is a true norm.

6.2 Model, Assumptions and Main Results

We begin by introducing our model of residents admitting and treating a flow of incoming patients. The patients are assumed to arrive according to a non-homogeneous Poisson process. For each $k \in \mathbb{Z}_+$, the process has rate λ_1 over the time intervals, $[k, k+\frac{1}{2})$ and rate $\lambda_2 < \lambda_1$ over the time intervals $[k+\frac{1}{2}, k+1)$. Let let $\lambda = (\lambda_1+\lambda_2)/2$ denote the average arrival rate and

$$\lambda(t) = \begin{cases} \lambda_1 & t \in [k, k+\frac{1}{2}), \\ \lambda_2 & t \in [k+\frac{1}{2}, k+1). \end{cases}$$

The intervals $[k, k+1)$ represent, for example, 24 hour cycles, where $[k, k+\frac{1}{2})$ is the portion of the day, say from 10am to 10pm, in which the vast majority of patients arrive (see Figure 5-9). The residents are combined into two teams, A and B, identical in size, which are eligible to admit patients (are on shift) according a schedule to be described below. Each team has capacity $c > 0$ bounding the maximum number of patients the team can have in care. Each arriving patient is assigned to one of the residents on a team, chosen uniformly at random, provided that at least one of the teams on shift has not reached its capacity c. Note, that this is equivalent to saying that each resident has in care capacity c/N, where N is the number of residents on each the team. If each on shift teams has reached its capacity, the patient joins a single queue and is cared for by one of the back-up doctors until one of the residents is available, at which point the waiting terminates, using the First-In-First-Out assignment policy. The availability occurs either when one of the assigned patients leaves the hospital freeing the capacity of one of the teams, or when one of the teams with load less than c begins a shift.

At any time, each team is in one of two states, *on shift* or *off shift*, as specified by a policy. Patients remain assigned to a team until they leave the hospital. The durations of hospital stays are assumed to be i.i.d. and exponentially distributed with rate μ. That the random length of treatment time each patient requires begins accumulating at the moment of assignment to a team, and continues to accumulate when team is off shift. How this corresponds to actual practices is explained in the introduction.

We consider two scheduling policies controlling when each team is on shift, *Long Shifts* (LS) motivated by LOSO, and *Daily Admitting* (DA), motivated by MMMO (see the introduction for descriptions of LOSO and MMMO). LS is a two day rotating schedule. Team A is on shift on odd days, i.e. $[2k+1, 2k+2)$ for all $k \in \mathbb{Z}_+$, and off shift otherwise. Similarly, team B is on shift for even days, i.e. $[2k, 2k+1)$ for all $k \in \mathbb{Z}_+$, and off otherwise. In DA, both teams A and B are on shift every day for the first half of each day, i.e. $[k, k+\frac{1}{2})$ for all $k \in \mathbb{Z}_+$, and off otherwise, effectively creating a single team with double the capacity.

To state our results, it will be convenient to introduce the following quantities describing the dynamics of our model. For $t \geq s \geq 0$, let $A(s,t)$ denote the number of patients that arrive in the time interval $[s,t]$ according to our non-homogeneous Poisson process. For each policy $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, let $Q_\theta(t)$ denote the number of patients in the queue not yet assigned to a team plus the number of patients assigned to the teams that are on shift at time t. For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, let $R_\theta(t)$ denote the total number of patients currently assigned to teams which are off shift at time t. Further, for each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$ we introduce the random vector $\mathbf{S}_\theta(t) = (Q_\theta(t), R_\theta(t))$, which we take to be the state of our system. The processes $Q_\theta(t)$, $R_\theta(t)$, and $\mathbf{S}_\theta(t)$ are assumed to be right-continuous with left limits. For every $s < t$, we let $D_\theta^{\mathrm{on}}(s,t)$ denote the total number of patients which departed in the time interval $[s,t]$ from the teams which were on shift in this period. We define $D_\theta^{\mathrm{off}}(s,t)$ analogously.

Under the policy LS, for each $k \in \mathbb{Z}_+$ and each $t \in [0,1)$, Q_{LS} and R_{LS} satisfy

$$Q_{\text{LS}}(k+t) = Q_{\text{LS}}(k) + A(k,k+t) - D^{\text{on}}_{\text{LS}}(k,k+t), \tag{6.1}$$

$$R_{\text{LS}}(k+t) = R_{\text{LS}}(k) - D^{\text{off}}_{\text{LS}}(k,k+t). \tag{6.2}$$

On $[k,k+1)$, we see that in distribution $Q_{\text{LS}}(t)$ behaves exactly as the total number of customers in system for an $M(t)/M/c$ queue with arrival rate $\lambda(t)$, service rate μ, and initial value $Q_{\text{LS}}(k)$. Similarly, on $[k,k+1)$, $R_{\text{LS}}(t)$ behaves as an $M/M/c$ system with no arrivals, service rate μ, and initial value $R_{\text{LS}}(k)$. At each time $k \in \mathbb{Z}_+$ a transition occurs: the off shift team switches to on shift and vice versa, and patients that are waiting can get assigned to the team rotating on shift. In terms of Q_{LS} and R_{LS}, this can be described as follows:

$$\begin{aligned} Q_{\text{LS}}(k) &= (Q_{\text{LS}}(k^-) - c)^+ + R_{\text{LS}}(k^-), \\ R_{\text{LS}}(k) &= Q_{\text{LS}}(k^-) \wedge c. \end{aligned}$$

We define operator $\Gamma : \mathbb{R}^2_+ \times \mathbb{R}_+ \to \mathbb{R}^2_+$ by

$$\Gamma(q,r;\kappa) \stackrel{\Delta}{=} (q-\kappa)^+ + r, q \wedge \kappa), \tag{6.3}$$

and note that we can equivalently write

$$\mathbf{S}_{\text{LS}}(k) = \Gamma(\mathbf{S}_{\text{LS}}(k^-); c). \tag{6.4}$$

The equations (6.1), (6.2) and (6.4) along with a distribution of the initial states $\mathbf{S}_{\text{LS}}(0)$ completely determine the distribution of $\mathbf{S}_{\text{LS}}(t)$ for all $t \in \mathbb{R}_+$. It is immediate that on integer times, the process $\mathbf{S}_{\text{LS}}(k)$ is a two dimensional Markov chain on the countable state space $\mathbb{Z}_+ \times \{0,1,\ldots,c\}$. Moreover, without loss of generality we can restrict the state space to the set,

$$\mathcal{S}_{\text{LS}} \stackrel{\Delta}{=} \{0,\ldots,c\}^2 \cup \{(q,c) \mid q \in \mathbb{Z}_+\}, \tag{6.5}$$

as this set is the image of $\Gamma(\cdot; c)$ when we take the domain to be $\mathbb{Z}_+ \times \{0, 1, \ldots, c\}$. Thus $\mathbf{S}_{\mathrm{LS}}(k) \in \mathcal{S}_{\mathrm{LS}}$ for all $k \geq 1$ with probability one.

We now describe similar relations for the policy DA. For every $k \in \mathbb{Z}_+$ and every $t \in [0, \frac{1}{2})$,

$$\begin{aligned} Q_{\mathrm{DA}}(k+t) &= Q_{\mathrm{DA}}(k) + A(k, k+t) - D^{\mathrm{on}}_{\mathrm{DA}}(k, k+t), \\ R_{\mathrm{DA}}(k+t) &= 0, \end{aligned} \tag{6.6}$$

where on $[k, k+\frac{1}{2})$, the process $Q_{\mathrm{DA}}(t)$ has the same distribution as an $M/M/2c$ queue with an arrival rate of λ_1, a service rate of μ and an initial value of $Q_{\mathrm{DA}}(k)$. At time $k+\frac{1}{2}$, both teams move off shift, resulting in the transition

$$\begin{aligned} Q_{\mathrm{DA}}(k+\tfrac{1}{2}) &= (Q_{\mathrm{DA}}((k+\tfrac{1}{2})^-) - 2c)^+, \\ R_{\mathrm{DA}}(k+\tfrac{1}{2}) &= Q_{\mathrm{DA}}((k+\tfrac{1}{2})^-) \wedge 2c. \end{aligned}$$

Equivalently, in terms of Γ,

$$\mathbf{S}_{\mathrm{DA}}(k+\tfrac{1}{2}) = \Gamma(\mathbf{S}_{\mathrm{DA}}((k+\tfrac{1}{2})^-); 2c). \tag{6.7}$$

On $[k+\frac{1}{2}, k+1)$,

$$\begin{aligned} Q_{\mathrm{DA}}(k+\tfrac{1}{2}+t) &= Q_{\mathrm{DA}}(k+\tfrac{1}{2}) + A(k+\tfrac{1}{2}, k+\tfrac{1}{2}+t), \\ R_{\mathrm{DA}}(k+\tfrac{1}{2}+t) &= R_{\mathrm{DA}}(k+\tfrac{1}{2}) - D^{\mathrm{off}}_{\mathrm{DA}}(k+\tfrac{1}{2}, k+\tfrac{1}{2}+t). \end{aligned} \tag{6.8}$$

Now on $[k+\frac{1}{2}, k+1)$, $Q_{\mathrm{DA}}(k+\frac{1}{2}+t)$ changes according to a Poisson process with arrival rate λ_2, and $R(t)$ is an $M/M/2c$ queue with no arrivals and service rate μ. At integer times, we have a second shift change, this time leading to

$$\begin{aligned} Q_{\mathrm{DA}}(k+1) &= Q_{\mathrm{DA}}((k+1)^-) + R_{\mathrm{DA}}((k+1)^-), \\ R_{\mathrm{DA}}(k+1) &= 0. \end{aligned}$$

Equivalently, in terms of Γ,

$$\mathbf{S}_{\mathrm{DA}}(k+1) = \Gamma(\mathbf{S}_{\mathrm{DA}}((k+1)^{-}); 0). \tag{6.9}$$

Equations (6.6), (6.7), (6.8) and (6.9) along with the distribution over the initial states $\mathbf{S}_{\mathrm{DA}}(0)$ determine the distribution of $\mathbf{S}_{\mathrm{DA}}(t)$ for all $t \in \mathbb{R}_+$. Again on integer times, the process $\mathbf{S}_{\mathrm{DA}}(k)$ is a Markov chain on the countable state space. However, now the state space is the one-dimensional set

$$\mathcal{S}_{\mathrm{DA}} \triangleq \mathbb{Z}_+ \times \{0\}, \tag{6.10}$$

as $R_{\mathrm{DA}}(k) = 0$ for all $k \in \mathbb{Z}_+$.

We are now ready to discuss our main results. We define the stochastic process $\mathbf{S}_\theta(t)$ to be *stable* if the embedded discrete time process $\mathbf{S}_\theta(k)$ for $k \in \mathbb{Z}_+$ is positive recurrent, and *unstable* otherwise. Normally we define stability for this type of problem as positive Harris recurrence of the process $\mathbf{S}_\theta(t)$, $t \in \mathbb{R}_+$. However, it is easy to see that in our case these definitions are equivalent, and further that the former definition is much easier to work with.

Observe that under both policies, $R_\theta(k)$ is bounded hence $Q_\theta(k)$ is the only potential source of instability. Thus when $\mathbf{S}_\theta(t)$ is unstable, with probability one, the number of patients waiting to be assigned to a resident team will grow without bound. Note that in reality, when a hospital has a large number of patients waiting, it reroutes incoming patients to other hospitals to reduce congestion, so instability would actually correspond to the hospital frequently being forced to turn patients away—clearly a very undesirable situation.

We now discuss conditions under which $\mathbf{S}_\theta(t)$ is stable. Before formally stating our results, we provide some intuition. Let $L_c(t)$ and $L_{2c}(t)$ be the number of patients in system for an $M(t)/M/c$ queue and $M(t)/M/2c$ queue both driven by the arrival process $A(0,t)$, respectively. For LS, it is not difficult to see that we can couple $\mathbf{S}_{\mathrm{LS}}(t)$

with $L_c(t)$ and $L_{2c}(t)$ such that surely, for every t,

$$L_{2c}(t) \leq Q_{\mathrm{LS}}(t) + R_{\mathrm{LS}}(t) \leq L_c(t).$$

The inequalities hold as the process $\mathbf{S}_{\mathrm{LS}}(t)$ has capacity between c and $2c$ at all times t. Similarly, we can couple $\mathbf{S}_{\mathrm{DA}}(t)$ and $L_{2c}(t)$ such that

$$L_{2c}(t) \leq Q_{\mathrm{DA}}(t) + R_{\mathrm{DA}}(t).$$

Recall from basic queueing theory that the process $L_c(t)$ is positive recurrent iff $\lambda < c\mu$ and $L_{2c}(t)$ is positive recurrent iff $\lambda < 2c\mu$. In light of our coupling, we thus expect the maximum throughput (the largest $\lambda = (\lambda_1 + \lambda_2)/2$ such that $\mathbf{S}_\theta(t)$ is stable) of $\mathbf{S}_{\mathrm{LS}}(t)$ to lie between $c\mu$ and $2c\mu$, and likewise we expect the maximum throughput of $\mathbf{S}_{\mathrm{DA}}(t)$ to be at most $2c\mu$. This suggests that we need to determine to what extent each policy can utilize the $2c$ total capacity available to treat patients, or conversely how much forced idling is caused under each policy by a team's inability to admit new patients when off shift. To this end, we let

$$\rho_{\mathrm{LS}} \triangleq \frac{\lambda}{c(1 - e^{-\mu}) + c\mu}, \tag{6.11}$$

$$\rho_{\mathrm{DA}} \triangleq \frac{\lambda}{2c(1 - e^{-\mu/2}) + c\mu}. \tag{6.12}$$

We intend to show that $\mathbf{S}_\theta(t)$ is positive recurrent iff $\rho_\theta < 1$ for each θ. These values of ρ_θ imply that

$$\lambda^*_{\mathrm{LS}} \triangleq \frac{\lambda}{\rho_{\mathrm{LS}}} = c(1 - e^{-\mu}) + c\mu, \qquad \lambda^*_{\mathrm{DA}} \triangleq \frac{\lambda}{\rho_{\mathrm{DA}}} = 2c(1 - e^{-\mu/2}) + c\mu,$$

give the maximum throughput of $\mathbf{S}_{\mathrm{LS}}(t)$ and $\mathbf{S}_{\mathrm{DA}}(t)$, respectively. Thus our main stability result is as follows.

Theorem 6.1. *For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, the process $\mathbf{S}_\theta(t)$ is positive recurrent when $\rho_\theta < 1$, null recurrent when $\rho_\theta = 1$, and transient when $\rho_\theta > 1$. Namely, the process*

$\mathbf{S}_\theta(t)$ *is stable iff* $\rho_\theta < 1$. *Furthermore,* $\rho_{\text{DA}} < \rho_{\text{LS}}$. *In particular, the Daily Admitting policy has a greater maximum throughput.*

The intuition behind the result is that the queue will be stable as long as conditional on the queue being large, the expected number of arrivals per day is less than the expected number of departures per day. Independent of the initial queue length, we expect λ arrivals per day. For the sake of argument, assume the initial queue length were infinite, so the resident teams are only idle when they are off shift and have completed caring for their initial patients.

Under the policy LS, in a single day the team on shift has c patients in care at all times each recovering at rate μ, producing $\text{Pois}(c\mu)$ departures. Thus the expected number of departures from the team on shift is $c\mu$. For the team off shift, as we assumed there were infinitely many patients initially in care, we begin with all c capacity utilized. Again patients depart at rate μ, but now when they leave they are not replaced. The probability a patient will depart is $\mathbb{P}(\text{Exp}(\mu) \leq 1) = 1 - e^{-\mu}$. As whether or not each patient departs is independent, we have $\text{Bin}(c, 1-e^{-\mu})$ departures, so the expected number of departures from the team off shift is $c(1 - e^{-\mu})$. Thus the expected change for the number of patients in the system is given by

$$-\gamma_{\text{LS}} \triangleq \lambda - c\mu - c(1 - e^{-\mu}).$$

Recalling that we expect the system to be stable when $\gamma_{\text{LS}} > 0$, we see from (6.11) that this is equivalent to $\rho_{\text{LS}} < 1$. Performing a similar computation for DA, we see that in a single day there are $\text{Pois}(c\mu)$ on shift departures and $\text{Bin}(2c, 1 - e^{-\mu/2})$ off shift departures, giving an expected change in the number of patients in system of

$$-\gamma_{\text{DA}} \triangleq \lambda - c\mu - 2c(1 - e^{-\mu/2}).$$

Thus the queue should be stable if $\gamma_{\text{DA}} > 0$, or equivalently from (6.12), when $\rho_{\text{DA}} < 1$.

To show $\rho_{\text{DA}} < \rho_{\text{LS}}$, it suffices to show that $2 - 2e^{-\mu/2} > 1 - e^{-\mu}$, which follows since

$$1 - 2e^{-\mu/2} + e^{-\mu} = (1 - e^{-\mu/2})^2 > 0. \tag{6.13}$$

As $c\mu + 2c(1 - e^{-\mu/2}) < 2c\mu$, we see that DA still results in fewer expected departures than an $M/M/2c$ queue. However, if we are willing to consider schedules with more shift changes per day, we can achieve an expected number of departures arbitrarily close to our "upper bound" of $2c\mu$ by generalizing the policy DA. Given $k > 0$ integer and even, consider the schedule where both teams are on shift for $[i/k, (i+1)/k)$ for all i even (the case of $i = 2$ is simply the policy DA). This divides the day into k equally sized pieces, where for $k/2$ such pieces both teams are on shift, and for the remaining $k/2$ periods both teams are off shift. We see immediately that independent of k, each team still spends half of each day on shift. In this half day on shift, our two teams' $2c$ capacity will again have Pois$(c\mu)$ departures. Now in each of our off shift periods, the probability of a patient leaving is $\mathbb{P}(\text{Exp}(\mu) \leq 1/k) = 1 - e^{-\mu/k}$, so we have Bin$(2c, 1 - e^{-\mu/k})$ off shift departures in each of our $k/2$ off shifts, or Bin$(kc, 1 - e^{-\mu/k})$ off shift departures per day. Thus the expected off shift departures per day is $kc(1 - e^{-\mu/k})$. Letting $k \to \infty$, we see through Taylor expansion that our off shift departures tend to $c\mu$, giving $2c\mu$ total departures as with the $M/M/2c$ queue. While in practice, we cannot have arbitrarily short shifts, we do see a general trend that shorter shifts increase capacity.

The stability property however is not the only relevant performance measure. An important quantity to look at is the number of *patient reassignments* (i.e. the number of arriving patients forced to wait due to the non-availability of residents, as discussed in the introduction). For each policy $\theta \in \{\text{LS}, \text{DA}\}$, we can easily verify that $\mathbf{S}_\theta(k)$ is irreducible and aperiodic on $\mathcal{S}_\theta$. Thus under the condition $\rho_\theta < 1$, there exists a unique steady state distribution for $\mathbf{S}_\theta(t)$, and we denote this random vector by $\mathbf{S}_\theta(\infty)$. Analyzing $\mathbf{S}_\theta(\infty)$ directly appears to be intractable. Instead, we resort to the method of fluid approximation, which we now define.

Given the parameters of our queueing model λ_1, λ_2, μ and c, we consider a sequence

of approximate models $n = 1, 2, \ldots$, where we change the parameters so that in the nth model, $\lambda_1^n = \lambda_1 n$, $\lambda_2^n = \lambda_2 n$, $\mu^n = \mu$, and $c^n = cn$. In words, the rate of patient recovery is fixed, but the patient arrival rates and patient capacity scale up linearly. For each policy $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, we let $Q_\theta^n(t), R_\theta^n(t)$ and $\mathbf{S}_\theta^n(t)$ be the corresponding processes. We let $\mathcal{S}_\theta^n$ be the set $\mathcal{S}_\theta$ as defined in (6.5) and (6.10) for the processes $\mathbf{S}_\theta^n(t)$. The process associated with fluid rescaling is defined as $\mathbf{S}_\theta^n(t)/n$. We immediately note by (6.11) and (6.12) that ρ_θ does not change with n, so the stability criteria for each $\mathbf{S}_\theta^n(t)$ is the same. Thus for $\rho_\theta < 1$ the sequence $\mathbf{S}_\theta^n(\infty)/n$ is well defined. Our next main result is that as $n \to \infty$, the sequences $\mathbf{S}_\theta^n(t)/n$ and $\mathbf{S}_\theta^n(\infty)/n$ converge meaningfully to some deterministic process $\mathbf{s}_\theta(t) = (q_\theta(t), r_\theta(t))$, and its unique fixed point $\lim_{k\to\infty} \mathbf{s}_\theta(k)$, respectively. We now provide details.

For LS, we define the process $\mathbf{s}_{\mathrm{LS}}(t) = (q_{\mathrm{LS}}(t), r_{\mathrm{LS}}(t))$ on $\mathbb{R}_+ \times [0, c]$ inductively on intervals $[k, k+1)$. For each interval, consider the system of ordinary differential equations (ODEs)

$$\dot{q}_{\mathrm{LS}}(t) = \lambda(t) - \mu(q_{\mathrm{LS}}(t) \wedge c), \tag{6.14}$$

$$\dot{r}_{\mathrm{LS}}(t) = -\mu r_{\mathrm{LS}}(t). \tag{6.15}$$

At integer times $k \geq 1$, the process jumps as did $\mathbf{S}_{\mathrm{LS}}(k)$. Specifically, we let

$$\mathbf{s}_{\mathrm{LS}}(k) = \Gamma(\mathbf{s}_{\mathrm{LS}}(k^-); c). \tag{6.16}$$

In analogy with $\mathcal{S}_{\mathrm{LS}}$, we will show that at integer times $k \geq 1$ this process is actually restricted to

$$\mathcal{T}_{\mathrm{LS}} \triangleq [0, c]^2 \cup \mathbb{R}_+ \times \{c\}. \tag{6.17}$$

We now give a similar construction for $\mathbf{s}_{\mathrm{DA}}(t) = (q_{\mathrm{DA}}(t), r_{\mathrm{DA}}(t))$ on $\mathbb{R}_+ \times [0, 2c]$. Again for each interval $[k, k + \frac{1}{2})$, we let $r_{\mathrm{DA}}(t) = 0$ and define $q_{\mathrm{DA}}(t)$ by

$$\dot{q}_{\mathrm{DA}}(t) = \lambda_1 - \mu(q_{\mathrm{DA}}(t) \wedge 2c). \tag{6.18}$$

At times $k+\frac{1}{2}$, $k \in Z_+$, we let

$$\mathbf{s}_{\mathrm{DA}}(k+\tfrac{1}{2}) = \Gamma(\mathbf{s}_{\mathrm{DA}}((k+\tfrac{1}{2})^-); 2c).$$

For each interval $[k+\frac{1}{2}, k+1)$, $q_{\mathrm{DA}}(t)$ and $r_{\mathrm{DA}}(t)$ are defined by the following ODEs:

$$\begin{aligned}\dot{q}_{\mathrm{DA}}(t) &= \lambda_2,\\ \dot{r}_{\mathrm{DA}}(t) &= -\mu r_{\mathrm{DA}}(t).\end{aligned}$$

Again at integer times $k \geq 1$, we define

$$\mathbf{s}_{\mathrm{LS}}(k) = \Gamma(\mathbf{s}_{\mathrm{DA}}(k^-); 0).$$

We let

$$\mathcal{T}_{\mathrm{DA}} \stackrel{\Delta}{=} \mathbb{R}_+ \times \{0\}. \tag{6.19}$$

We will show that this is the set of possible values $\mathbf{s}_{\mathrm{DA}}(k)$ can take for integer $k \geq 1$.

Proposition 6.1. *For every $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, and every $\mathbf{s}_\theta(0) \in \mathcal{T}_\theta$, $\mathbf{s}_\theta(t)$ exists and is uniquely defined for all $t \in \mathbb{R}_+$. Further, for all integer $k \geq 1$, $\mathbf{s}_\theta(k) \in \mathcal{T}_\theta$.*

The result is shown in Section 6.8. We now formally relate $\mathbf{S}_\theta(t)$ to $\mathbf{s}_\theta(t)$.

Theorem 6.2. *For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, if $\mathbf{S}^n_\theta(0)/n \to \mathbf{s}_\theta(0)$ a.s., then*

$$\lim_{n\to\infty} \frac{\mathbf{S}^n_\theta(t)}{n} = \mathbf{s}_\theta(t),$$

a.s. and u.o.c.

While this theorem allows us to approximate $\mathbf{S}_\theta(t)$ by the simpler process $\mathbf{s}_\theta(t)$, we have not established any relationship between $\mathbf{S}_\theta(\infty)$ and $\mathbf{s}_\theta(k)$ as $k \to \infty$. We do this next, but first we need some definitions.

Suppose we are given a discrete time dynamical system on a state space $\mathcal{X} \subset \mathbb{R}^n$

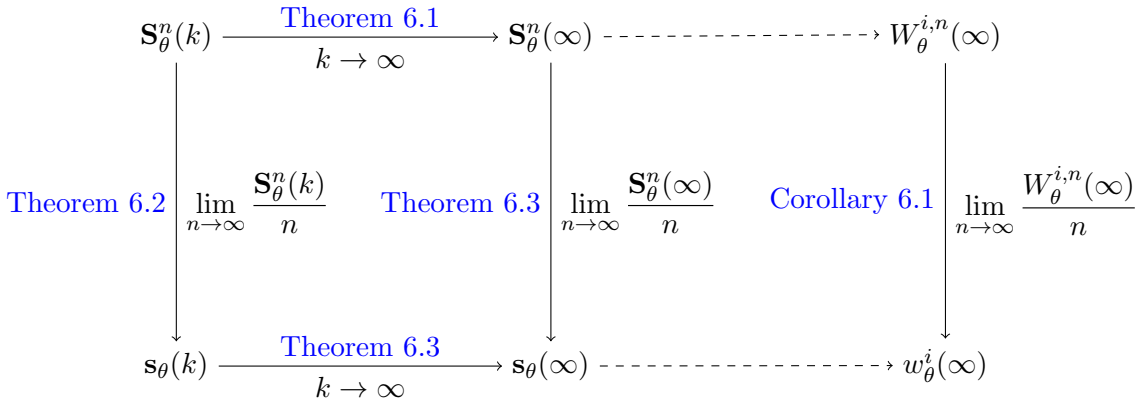

Figure 6-1: A diagram explaining how each of our theorems relate our stochastic process and the fluid limit, for finite times, at steady state, and then finally for the steady state number of reassignments.

defined by $\mathbf{f}\colon \mathcal{X} \to \mathcal{X}$, i.e. $\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k)$ for all k. A point $\mathbf{x}^*$ is defined to be *attractive* if for all $\mathbf{x}_0 \in \mathcal{X}$,

$$\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}^*.$$

Note that there can be at most one attractive point. We now state our next result relating $\mathbf{s}_\theta(\infty)$ and $\mathbf{S}_\theta(\infty)$.

Theorem 6.3. *For each policy $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, the sequence $\mathbf{s}_\theta(k)$ has a unique attractive point $\mathbf{s}_\theta(\infty) \in \mathcal{T}_\theta$ iff $\rho_\theta < 1$. Moreover, when $\rho_\theta < 1$, the following convergence in probability takes place:*

$$\lim_{n\to\infty} \frac{\mathbf{S}_\theta^n(\infty)}{n} = \mathbf{s}_\theta(\infty).$$

Notice that condition for the existence of an attractive point for $\mathbf{s}_\theta(t)$ is exactly the same as the stability condition for $\mathbf{S}_\theta(t)$. In the second claim of Theorem 6.3, we are essentially justifying an interchange of limits, as informally we are "equating" $\lim_{n\to\infty}\lim_{k\to\infty} \mathbf{S}_\theta^n(k)/n$ with $\lim_{k\to\infty}\lim_{n\to\infty} \mathbf{S}_\theta^n(k)/n$, as shown in the left half of Figure 6-1.

We now use this result to approximate the steady state number of reassignments

in the queueing model. For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$ and each $k \in \mathbb{Z}_+$, let $W_\theta^1(k)$ (resp. $W_\theta^2(k)$) be the number of arriving patients during $[k, k+\frac{1}{2})$ (resp. $[k+\frac{1}{2}, k+1)$) that are forced to wait a nonzero amount of time before assignment to a resident, i.e. the number of reassignments. Similarly, when $\rho_\theta < 1$, we define $W_\theta^1(\infty)$ (resp. $W_\theta^2(\infty)$) to be the steady state number patients forced to wait during $[0, \frac{1}{2})$ (resp. $[\frac{1}{2}, 1)$). Next we define variables for the fluid approximations of these quantities. Let $b_\theta(t)$ be

$$b_{\mathrm{LS}}(t) = \mathbb{I}_{\{q_{\mathrm{LS}}(t) \geq c\}}, \tag{6.20}$$

$$b_{\mathrm{DA}}(t) = \begin{cases} \mathbb{I}_{\{q_{\mathrm{DA}}(t) \geq 2c\}} & t \in [k, k+\frac{1}{2}), \\ 1 & t \in [k+\frac{1}{2}, k+1), \end{cases} \tag{6.21}$$

i.e. $b_\theta(t)$ is the indicator that the on shift teams are saturated. We define

$$w_\theta^1(k) = \int_k^{k+\frac{1}{2}} b_\theta(t)\lambda_1 dt, \tag{6.22}$$

$$w_\theta^2(k) = \int_{k+\frac{1}{2}}^{k+1} b_\theta(t)\lambda_2 dt. \tag{6.23}$$

When $\rho_\theta < 1$, we let $w_\theta^1(\infty) = w_\theta^1(0)$ and $w_\theta^2(\infty) = w_\theta^2(0)$ assuming the fluid system begins in steady state, i.e. $\mathbf{s}_\theta(0) = \mathbf{s}_\theta(\infty)$. We next argue that $w_\theta^1(\infty)$ and $w_\theta^2(\infty)$ asymptotically describe the steady state number reassignments. Let $W_\theta^{1,n}(k)$ and $W_\theta^{2,n}(k)$ be the number of reassignments for $\mathbf{S}_\theta^n(t)$ from our fluid approximation. Then

Corollary 6.1. *For policy LS, assuming $\lambda_1, \lambda_2 \neq c\mu$, and for policy DA, assuming $\lambda_1 \neq 2c\mu$, the following convergence in probability takes place:*

$$\lim_{n\to\infty} \frac{W_\theta^{1,n}(\infty)}{n} = w_\theta^1(\infty),$$
$$\lim_{n\to\infty} \frac{W_\theta^{2,n}(\infty)}{n} = w_\theta^2(\infty).$$

The case when $\lambda_j = c\mu$ for either $j = 1, 2$ presents some annoying technical difficulties. As realistically we will never have exact equality, we do not pursue this

issue further. This sequence of results justifies approximating $W_\theta^1(\infty)$ and $W_\theta^2(\infty)$ by $w_\theta^1(\infty)$, $w_\theta^2(\infty)$, respectively. The result of Corollary 6.1 are summarized in the right half of Figure 6-1.

6.3 Numerical Results

In this section, we numerically solve for the steady state solution of the fluid model of each policy. We then compare the cost of the reassignments in a single day starting at steady state under each policy as we vary the average arrival rate. We relate our numerical observations to our empirical observations from Figure 5-3.

Throughout this section, we use the following parameters in our model: $\mu = 1/2$, $c = 40$, $\lambda_1 = 9\lambda/5$, and $\lambda_2 = \lambda/5$. Our choice of μ and c imply that $\lambda^*_{\text{LS}} \approx 35.7388$ and $\lambda^*_{\text{DA}} \approx 37.6959$. The value of c and the ratio of λ_1 to λ_2 were chosen to be representative of a department from a large hospital such as B&W. The value of μ must be chosen more carefully. In light of Remark 6.1, we set μ to control the relative sizes of the average length of stay and length of time between shifts. At B&W under the policy LOSO, there is a long shift every four days and the average patient length of stay is four days. Thus in our model we set the average length of stay $(1/\mu)$ to be two days as the policy LS has a long shift every two days.

In Figure 6-2, we fix λ at 34 and observe the steady state behavior of our two policies in the fluid limit over the course of a day. Notice that $\lambda < \lambda^*_{\text{LS}} < \lambda^*_{\text{DA}}$, so under both policies the fluid model is stable, but heavily loaded. We see that for both policies, under these particular parameters, the number of patients being treated by the teams on shift plus the number of patients waiting, $q_\theta(t)$, increases over the first half of day. For LS, the capacity of 40 for the teams on shift (as indicated by the dotted black line) is exceeded, and resulting in some reassignments. For DA however, as both teams are working during the first half of the day, we stay below the capacity of 80 patients and have no reassignments. In the second half of the day, under LS we see that the backlog of patients subsides and we return below 40 patients by the end of the day. For DA, as both teams are off shift during the second half of the day,

we see a jump at time 1/2 between q_{DA} and r_{DA} and then small backlog of arrivals accumulate in the second half of the day.

In Figure 6-3, we show the number of reassignments for each policy in $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$ as we vary λ. The dotted vertical lines indicate λ^*_{LS} and λ^*_{DA}, the largest patient arrival rates such that LS and DA are stable. We see that our observation from Figure 6-2, that LS had many reassignments in $[0, \frac{1}{2})$ while DA had no reassignments in this period, is typical when the system is heavily loaded (for λ near λ^*_{LS}). We also see in Figure 6-3 that when λ is low, both polices cause no reassignments in the first half of the day, and only DA causes reassignments in the second half the day. As λ increases towards λ^*_{LS}, we see LS begin to reassign nearly all patients, while DA continues to only reassign patients arriving in the second half of the day. Finally, for very large λ, we eventually see DA reassigning some patients during the first half of the day. While for these particular parameter settings, we only see DA reassignments in the first half of the day for λ so large that LS is unstable, this does not hold for all parameter settings. Interestingly, we see that under DA for λ near λ^*_{DA}, the number of reassignments does not approach λ, while it does for LS. This is occurring as under these parameters, we have more patients leaving than arriving in the second half of the day, creating some spare capacity during the start of the first half of the following day.

Comparing Figure 6-3 with our empirical observations from Figure 5-3 we see that the relationship between LS and DA is qualitatively similar to the relationship between the B&W policies LOSO (labeled Initial in the figure) and MMMO (labeled Daily Admitting in the figure). Most importantly, we have preserved the property that shorter more frequent shifts are the better policy when the patient load is heavy.

6.4 Conclusion

We have developed a queueing model to determine the effect of long shifts in medical resident schedules on the hospital's capacity to admit patients and the quality of care delivered. Our model was motivated by the empirical work from Chapter 5

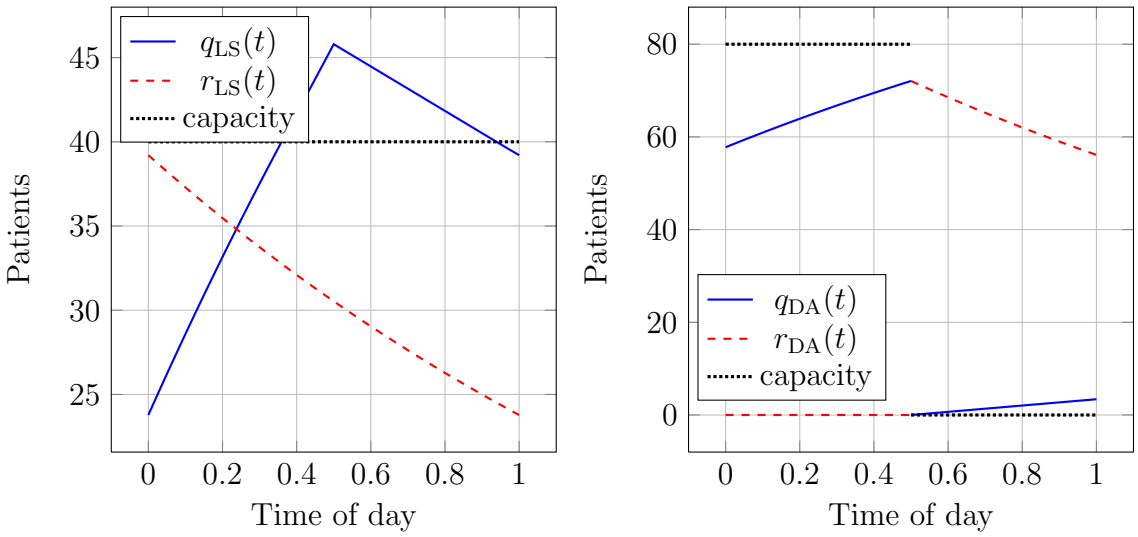

Figure 6-2: Steady state number of patients in system in the fluid limit under each policy. Parameter values: $\lambda = 34$, $\lambda_1 = 9\lambda/5$, $\lambda_2 = \lambda/5$, $\mu = 1/2$, $c = 40$.

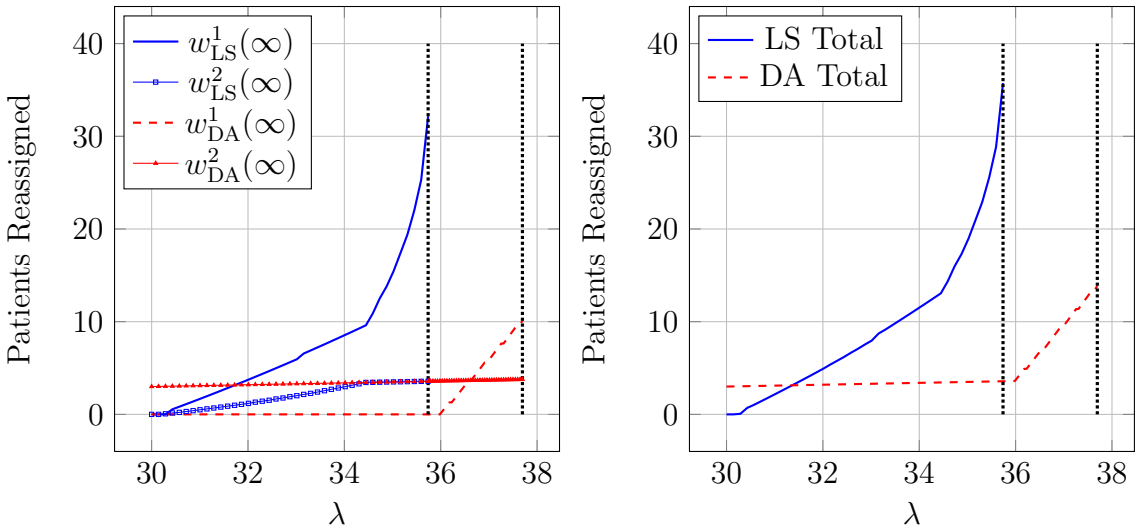

Figure 6-3: Steady state number of patient reassignments in the fluid limit for different λ under each policy. Parameter values: $\lambda_1 = 9\lambda/5$, $\lambda_2 = \lambda/5$, $\mu = 1/2$, $c = 40$.

on scheduling medical residents for B&W hospital. In this paper, we compared the stylized schedules *Long Shifts* (LS), where residents worked 24 hour shifts on alternating days, and *Daily Admitting* (DA), where residents worked every day but only during peak arrival hours. We used Lyapunov function techniques to characterize the stability of our queueing model under each policy. We found that DA has a greater capacity to admit patients than LS for all parameter choices. To analyze the long-run performance of our queueing model, we first considered the associated fluid model, which is a deterministic system with periodic dynamics. We showed that under each policy, when the queueing model is stable, the fluid model had a unique periodic steady state solution. We showed that our queueing model under the fluid rescaling converges to the fluid model on finite time intervals. Then we used an interchange of limits argument to show that the steady state queue lengths under the fluid rescaling converge to the unique steady state solution of the fluid model. We use these results to approximate the steady state number of reassignments in our queueing model by the steady state behavior of the fluid model. Numerically solving for the steady state of the fluid model under various parameter choices, we found evidence suggesting the existence of a threshold value on the arrival rate such that DA causes fewer reassignments than LS iff the arrival rate exceeds the threshold value. These results substantiate the main empirical findings in Chapter 5. The issue of resident schedules is currently quite pertinent, as new regulations restrict residents to a maximum shift length of 16 hours [50]. Our work contributes to understanding the implication of the new regulation. As hospitals tend to operate in heavily loaded regimes, we find that schedules relying on shorter more frequent shifts could increase capacity and reduce reassignments.

6.5 Stability Conditions for Two Schedules. Proof of Theorem 6.1

In this section, we prove Theorem 6.1, characterizing the stability of $\mathbf{S}_\theta(t)$. We show these results using the method of Lyapunov functions, using Proposition A.1 and Proposition A.2. The statement of these results and theirs proofs can be found in Appendix A, Section A.1.

Let the Lyapunov function $V\colon \mathcal{S}_\theta \to \mathbb{R}_+$ be defined by $V(q,r) = q + r$ for the Markov chain $\{\mathbf{S}_\theta(k)\}$. We first analyze the drift under the policy LS, namely $\mathbb{E}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))]$. Let $A \triangleq A(0,1)$, $D^{\mathrm{on}}_{\mathrm{LS}} \triangleq D^{\mathrm{on}}_{\mathrm{LS}}(0,1)$, and $D^{\mathrm{off}}_{\mathrm{LS}} \triangleq D^{\mathrm{off}}_{\mathrm{LS}}(0,1)$ denote the number of arrivals and departures in a single day under LS.

Lemma 6.1. *We have*

$$-\gamma_{\mathrm{LS}} = \lim_{q\to\infty} \mathbb{E}_{(q,c)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))] = \inf_{(q,r)\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{(q,r)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))]. \tag{6.24}$$

Additionally, there exists a constant $C_{\mathrm{LS}} > 0$ depending only on μ such that

$$\sup_{(q,r)\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{(q,r)}[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^2] \le \lambda^2 + \lambda + C_{\mathrm{LS}}(c^2 + c). \tag{6.25}$$

Proof. First, observe that the value of the Lyapunov function does change at time 1:

$$V(\mathbf{S}_{\mathrm{LS}}(1)) = V(\Gamma(\mathbf{S}_{\mathrm{LS}}(1^-); c)) = V(\mathbf{S}_{\mathrm{LS}}(1^-)), \tag{6.26}$$

as applying Γ does not change the number of patients in system. Thus for $\ell = 1, 2$,

$$\begin{aligned}\mathbb{E}_{(q,r)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^\ell\right] &= \mathbb{E}_{(q,r)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1^-)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^\ell\right] \\ &= \mathbb{E}_{(q,r)}\left[(A - D^{\mathrm{on}}_{\mathrm{LS}} - D^{\mathrm{off}}_{\mathrm{LS}})^\ell\right]. \end{aligned} \tag{6.27}$$

Let $\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \overset{d}{=} \mathrm{Pois}(c\mu)$ and $\tilde{D}^{\mathrm{off}}_{\mathrm{LS}} \overset{d}{=} \mathrm{Bin}(c, 1 - e^{-\mu})$ such that $\tilde{D}^{\mathrm{on}}_{\mathrm{LS}}$, $\tilde{D}^{\mathrm{off}}_{\mathrm{LS}}$, and A are independent. As $D^{\mathrm{on}}_{\mathrm{LS}}(0,t)$ for $0 \le t < 1$ has the distribution of the departure process for

an $M(t)/M/c$ queue, we can couple $D^{\text{on}}_{\text{LS}}$ with $\tilde{D}^{\text{on}}_{\text{LS}}$ such that regardless of $\mathbf{S}_{\text{LS}}(0)$,

$$D^{\text{on}}_{\text{LS}} \leq \tilde{D}^{\text{on}}_{\text{LS}}. \tag{6.28}$$

For the off shift departures, as the patient length of stay is exponential, each of the $r = R(0) \leq c$ patients in care will depart in the interval $[0, 1)$ with probability $1 - e^{-\mu}$ independently of other patients. Thus $D^{\text{off}}_{\text{LS}} \stackrel{d}{=} \text{Bin}(r, 1 - e^{-\mu})$, so trivially it can be coupled with $\tilde{D}^{\text{off}}_{\text{LS}}$ such that

$$D^{\text{off}}_{\text{LS}} \leq \tilde{D}^{\text{off}}_{\text{LS}}, \tag{6.29}$$

with equality when $r = c$. From (6.28) and (6.29) we obtain that for any initial condition $\mathbf{S}(0) = (q, r) \in \mathcal{S}_{\text{LS}}$

$$A - D^{\text{on}}_{\text{LS}} - D^{\text{off}}_{\text{LS}} \geq A - \tilde{D}^{\text{on}}_{\text{LS}} - \tilde{D}^{\text{off}}_{\text{LS}}.$$

Taking expectations and then the infimum over all $(q, r) \in \mathcal{S}_{\text{LS}}$, we obtain that

$$\inf_{(q,r)\in\mathcal{S}_{\text{LS}}} \mathbb{E}_{(q,r)}[V(\mathbf{S}_{\text{LS}}(1)) - V(\mathbf{S}(0))] \geq \lambda - c\mu - c(1 - e^{-\mu}) = -\gamma_{\text{LS}}.$$

Observe that for any realization where $Q(0) = q$ and $D^{\text{on}}_{\text{LS}} \leq \tilde{D}^{\text{on}}_{\text{LS}} \leq q - c$, we also have $Q_{\text{LS}}(t) \geq c$ for all $t \in [0, 1^-)$, and thus $D^{\text{on}}_{\text{LS}} = \tilde{D}^{\text{on}}_{\text{LS}}$. As a result,

$$\mathbb{E}_{(q,c)}[D^{\text{on}}_{\text{LS}}] \geq \mathbb{E}_{(q,c)}\left[D^{\text{on}}_{\text{LS}}\mathbb{I}_{\{\tilde{D}^{\text{on}}_{\text{LS}}\leq q-c\}}\right] = \mathbb{E}\left[\tilde{D}^{\text{on}}_{\text{LS}}\mathbb{I}_{\{\tilde{D}^{\text{on}}_{\text{LS}}\leq q-c\}}\right]. \tag{6.30}$$

Since almost surely

$$\lim_{q\to\infty} \tilde{D}^{\text{on}}_{\text{LS}}\mathbb{I}_{\{\tilde{D}^{\text{on}}_{\text{LS}}\leq q-c\}} = \tilde{D}^{\text{on}}_{\text{LS}},$$

by monotonicity of expectation and then the Monotone Convergence Theorem, we

obtain that

$$\liminf_{q\to\infty} \mathbb{E}_{(q,c)}[D_{\mathrm{LS}}^{\mathrm{on}}] \geq \lim_{q\to\infty} \mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on}} \leq q-c\}}] = \mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on}}] = c\mu.$$

Combining this inequality with (6.28), we obtain $\lim_{q\to\infty} \mathbb{E}_{(q,c)}[D_{\mathrm{LS}}^{\mathrm{on}}] = c\mu$ and thus

$$\lim_{q\to\infty} \mathbb{E}_{(q,c)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))] = \lambda - c(1-e^{-\mu}) - c\mu = -\gamma_{\mathrm{LS}}.$$

Lastly, to show (6.25), using independence, (6.28), and (6.29),

$$\begin{aligned}
&\sup_{(q,r)\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{(q,r)}[(A - D_{\mathrm{LS}}^{\mathrm{on}} - D_{\mathrm{LS}}^{\mathrm{off}})^2] \\
&\quad \leq \mathbb{E}[A^2] + \mathbb{E}[(\tilde{D}_{\mathrm{LS}}^{\mathrm{on}})^2] + \mathbb{E}[(\tilde{D}_{\mathrm{LS}}^{\mathrm{off}})^2] + 2\mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on}}]\mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{off}}] \\
&\quad = \lambda^2 + \lambda + c^2\mu^2 + c\mu + c(1-e^{-\mu})e^{-\mu} + (c(1-e^{-\mu}))^2 + 2c^2\mu(1-e^{-\mu}).
\end{aligned}$$

□

We now give an analogous result to the previous lemma for DA. As the proof is nearly the same, some details have been omitted. As before, let $D_{\mathrm{DA}}^{\mathrm{on}} \stackrel{\Delta}{=} D_{\mathrm{DA}}^{\mathrm{on}}(0, \frac{1}{2})$ and $D_{\mathrm{DA}}^{\mathrm{off}} \stackrel{\Delta}{=} D_{\mathrm{DA}}^{\mathrm{off}}(\frac{1}{2}, 1)$, give the number of departures under policy DA in a single day (note that there is no one on shift during $[\frac{1}{2}, 1)$ and no one off shift during $[0, \frac{1}{2})$ under DA).

Lemma 6.2. *We have*

$$-\gamma_{\mathrm{DA}} = \lim_{q\to\infty} \mathbb{E}_{(q,0)}[V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0))] = \inf_{(q,r)\in\mathcal{S}_{\mathrm{DA}}} \mathbb{E}_{(q,r)}[V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0))]. \tag{6.31}$$

Additionally, there exists a constant $C_{\mathrm{DA}} > 0$ depending only on μ such that

$$\sup_{(q,r)\in\mathcal{S}_{\mathrm{DA}}} \mathbb{E}_{(q,r)}[(V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0)))^2] \leq \lambda^2 + \lambda + C_{\mathrm{DA}}(c^2 + c). \tag{6.32}$$

Proof. Again, for any initial state $\mathbf{S}_{\mathrm{DA}}(0) = (q, 0)$, $V(\mathbf{S}_{\mathrm{DA}}(1)) = V(\mathbf{S}_{\mathrm{DA}}(1^-))$ and

$V(\mathbf{S}_{\mathrm{DA}}(\frac{1}{2})) = V(\mathbf{S}_{\mathrm{DA}}(\frac{1}{2}^-))$, as applying Γ at times $\frac{1}{2}$ and 1 does not change the number of patients in system. Thus for $\ell = 1, 2$,

$$\mathbb{E}_{(q,0)}[(V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0)))^\ell] = \mathbb{E}_{(q,0)}\left[(A - D^{\mathrm{on}}_{\mathrm{DA}} - D^{\mathrm{off}}_{\mathrm{DA}})^\ell\right]. \tag{6.33}$$

Let $\tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \stackrel{d}{=} \mathrm{Pois}(c\mu)$ and $\tilde{D}^{\mathrm{off}}_{\mathrm{DA}} \stackrel{d}{=} \mathrm{Bin}(2c, 1 - e^{-\mu/2})$ such that $\tilde{D}^{\mathrm{on}}_{\mathrm{DA}}$, $\tilde{D}^{\mathrm{off}}_{\mathrm{DA}}$, and A are independent. As $D^{\mathrm{on}}_{\mathrm{DA}}(0, t)$ for $0 \le t < \frac{1}{2}$ has the distribution of the departure process for an $M(t)/M/2c$ queue, we can couple $D^{\mathrm{on}}_{\mathrm{DA}}(0, \frac{1}{2}^-)$ with $\tilde{D}^{\mathrm{on}}_{\mathrm{DA}}$ such that regardless of $\mathbf{S}_{\mathrm{DA}}(0)$,

$$D^{\mathrm{on}}_{\mathrm{DA}} \le \tilde{D}^{\mathrm{on}}_{\mathrm{DA}}. \tag{6.34}$$

For the off shift departures, as the patient length of stay is exponential and thus memoryless, each of the $R_{\mathrm{DA}}(\frac{1}{2}) \le 2c$ patients in care will depart in the interval $[\frac{1}{2}, 1)$ with probability $1 - e^{-\mu/2}$ independently of other patients. Thus $D^{\mathrm{off}}_{\mathrm{DA}} \stackrel{d}{=} \mathrm{Bin}(R_{\mathrm{DA}}(\frac{1}{2}), 1 - e^{-\mu/2})$, so it can be coupled with $\tilde{D}^{\mathrm{off}}_{\mathrm{DA}}$ such that

$$D^{\mathrm{off}}_{\mathrm{DA}} \le \tilde{D}^{\mathrm{off}}_{\mathrm{DA}}, \tag{6.35}$$

with equality when $R_{\mathrm{DA}}(\frac{1}{2}) = 2c$. From (6.34) and (6.35) we obtain that for any $(q, 0) \in \mathcal{S}_{\mathrm{DA}}$,

$$A - D^{\mathrm{on}}_{\mathrm{DA}} - D^{\mathrm{off}}_{\mathrm{DA}} \ge A - \tilde{D}^{\mathrm{on}}_{\mathrm{DA}} - \tilde{D}^{\mathrm{off}}_{\mathrm{DA}},$$

and thus by taking expectations

$$\inf_{(q,r)\in\mathcal{S}_{\mathrm{DA}}} \mathbb{E}_{(q,r)}[V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0))] \ge \lambda - c\mu - 2c(1 - e^{-\mu/2}) = -\gamma_{\mathrm{DA}}.$$

As before, to complete showing the three term equality in (6.31), it suffices to show $\lim_{q\to\infty} \mathbb{E}_{(q,0)}[V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0))] = -\gamma_{\mathrm{DA}}$. Given $Q_{\mathrm{DA}}(0) = q$, for any realization such that $D^{\mathrm{on}}_{\mathrm{DA}} \le \tilde{D}^{\mathrm{on}}_{\mathrm{DA}} \le q - 2c$, we have $Q_{\mathrm{DA}}(t) \ge 2c$ for all $t \in [0, \frac{1}{2}^-)$, and thus $D^{\mathrm{on}}_{\mathrm{DA}} = \tilde{D}^{\mathrm{on}}_{\mathrm{DA}}$. Further, $Q_{\mathrm{DA}}(\frac{1}{2}^-) \ge 2c$ ensures $R_{\mathrm{DA}}(\frac{1}{2}) = 2c$ and thus $D^{\mathrm{off}}_{\mathrm{DA}} = \tilde{D}^{\mathrm{off}}_{\mathrm{DA}}$. As

a result,

$$\mathbb{E}[D_{\mathrm{DA}}^{\mathrm{on}}] \geq \mathbb{E}\left[D_{\mathrm{DA}}^{\mathrm{on}}\mathbb{I}_{\{\tilde{D}_{\mathrm{DA}}^{\mathrm{on}}\leq q-2c\}}\right] = \mathbb{E}\left[\tilde{D}_{\mathrm{DA}}^{\mathrm{on}}\mathbb{I}_{\{\tilde{D}_{\mathrm{DA}}^{\mathrm{on}}\leq q-2c\}}\right],$$
$$\mathbb{E}[D_{\mathrm{DA}}^{\mathrm{off}}] \geq \mathbb{E}\left[D_{\mathrm{DA}}^{\mathrm{off}}\mathbb{I}_{\{\tilde{D}_{\mathrm{DA}}^{\mathrm{on}}\leq q-2c\}}\right] = \mathbb{E}\left[\tilde{D}_{\mathrm{DA}}^{\mathrm{off}}\mathbb{I}_{\{\tilde{D}_{\mathrm{DA}}^{\mathrm{on}}\leq q-2c\}}\right].$$

We can apply the Monotone Convergence Theorem as before but now on both $\tilde{D}_{\mathrm{DA}}^{\mathrm{on}}\mathbb{I}_{\{\tilde{D}^{\mathrm{on}}\leq q-2c\}}$ and $\tilde{D}_{\mathrm{DA}}^{\mathrm{off}}\mathbb{I}_{\{\tilde{D}^{\mathrm{on}}\leq q-2c\}}$ to obtain the desired limit. The rest of the proof is exactly as in the previous lemma. □

Proof of Theorem 6.1. Suppose $\rho_\theta < 1$, i.e. $\gamma_\theta > 0$. From (6.24) of Lemma 6.1, we obtain that

$$-\gamma_{\mathrm{LS}} = \lim_{q\to\infty} \mathbb{E}_{(q,c)}[V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0))].$$

Similarly, from (6.31) of Lemma 6.2, we

$$-\gamma_{\mathrm{DA}} = \lim_{q\to\infty} \mathbb{E}_{(q,0)}[V(\mathbf{S}_{\mathrm{DA}}(1)) - V(\mathbf{S}_{\mathrm{DA}}(0))].$$

Recall by (6.5) that for every $(q,r) \in \mathcal{S}_{\mathrm{LS}}$, when $q \geq c$, we must have $r = c$, and by (6.10) for every $(q,r) \in \mathcal{S}_{\mathrm{DA}}$ we have $r = 0$. Thus the sets

$$B_\theta = \left\{(q,r) \in \mathcal{S}_\theta \;\middle|\; \mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))] > -\gamma_\theta/2\right\},$$

are finite. Observe that for both LS and DA, (A.1) is satisfied by (6.25) and (6.32), respectively. Applying Proposition A.1, taking $B = B_\theta$ and $\gamma = \gamma_\theta/2$, we conclude that $\{\mathbf{S}_\theta(k)\}$ is positive recurrent.

Now suppose instead that $\rho_\theta \geq 1$, i.e. $\gamma_\theta \leq 0$. In the setting of Proposition A.2, for both θ we take $B_\theta = \{(0,0)\}$ and observe that (A.2) is trivially satisfied by taking $y = (q,r)$ for any nonzero $(q,r) \in \mathcal{S}_\theta$. The condition in (A.3) is satisfied by (6.25) for

LS and (6.32) for DA. Finally, (A.4) is satisfied as by (6.24) and (6.31), we have

$$\inf_{(q,r)\in\mathcal{S}_\theta} \mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))] = -\gamma_\theta \geq 0.$$

Thus from Proposition A.2 we conclude that $\{\mathbf{S}_\theta(k)\}$ is either null recurrent or transient.

It remains to show that $\{\mathbf{S}_\theta(k)\}$ is null recurrent when $\rho_\theta = 1$ and transient when $\rho_\theta > 1$. To do so, we use another Lyapunov function argument, Theorem 3.2 from [56], (see also [32] section 3.6). However, as the statement of this theorem is rather technical, we defer this part of the proof to Section 6.12.

Finally, that $\rho_{\text{DA}} < \rho_{\text{LS}}$ follows from (6.13). □

Remark 6.1. Note that as $\mu \to 0$, $(1 - e^{-\mu/2})^2 \to 0$, so by (6.13) we see that $\rho_{\text{LS}} - \rho_{\text{DA}} \to 0$ as well. Intuitively, if patients take many days to recover, the amount of forced idle time due to not being able to admit patients while off shift will be negligible. Conversely, when μ is large,

$$\rho_{\text{LS}} \approx \frac{\lambda}{c\mu + c}, \qquad \rho_{\text{DA}} \approx \frac{\lambda}{c\mu + 2c}.$$

In this regime, nearly all patients recover in each off shift. When c is also large, we see DA has a larger stability region, due to the off shifts for each team being shorter. While this regime isn't particularly relevant in the hospital setting, where $\mu \approx 1/4$, it exposes another interesting and relevant parameter, namely the length of the time between shifts relative to the recovery rate μ.

6.6 Uniform Bounds for Stationary Performance Measures

In this section, we consider the sequence of systems under the fluid rescaling $\mathbf{S}^n_\theta(t)/n$ with the assumption that $\rho_\theta < 1$, and give bounds on the expected stationary queue lengths that are independent of n. We will again use the Lyapunov function technique,

namely Proposition A.3. The statement of this result and a proof can be found in Appendix A, Section A.2.

First, we need a property of sample paths of the $M(t)/M/m$ queue. For every initial queue length $q \in \mathbb{Z}_+$, we create a separate queue length process with the same arrival and service rates on a common probability space Ω. Let $f\colon \mathbb{Z}_+ \times \mathbb{R}_+ \times \Omega \to \mathbb{Z}_+$ map an initial queue length q, a time t, and a realization $\omega \in \Omega$ to the number of patients in the $M(t)/M/m$ system length at time t. The queues are coupled such that they share a single common Poisson process determining arrival times, and a single independent common Poisson process determining potential departure times (which only result in departures when there are patients in care). The relationship between the arrival process, the potential departure process, and the actual departures is the same as the relationship between $A(0,t)$, $\tilde{D}_{\text{on}}(0,t)$ and $D_{\text{on}}(0,t)$ from Section 6.5.

Lemma 6.3. *For every realization $\omega \in \Omega$, every time $t \in \mathbb{R}_+$, and all initial queue lengths $q, r \in \mathbb{Z}_+$ such that $q \geq r$, $f(\cdot, t, \omega)$ satisfies*

$$0 \leq f(q,t,\omega) - f(r,t,\omega) \leq q - r.$$

Namely, $f(\cdot, t, \omega)$ is monotone increasing and 1-Lipschitz continuous with respect to the ℓ_1 norm in the initial queue length.

Proof. Fix ω, and consider $q, r \in \mathbb{Z}_+$, $q > r$. Let $\tau_0 = 0$ and for $i = 1, 2, \ldots$, let τ_i be the time of the ith event from our processes driving arrivals and departures. It suffices to prove that

$$0 \leq |f(q,\tau_i,\omega) - f(r,\tau_i,\omega)| \leq q - r, \tag{6.36}$$

for all τ_i, as the queue length can only change at the times of these events. Trivially the claim holds at τ_0. Suppose the claim holds until τ_i. At time τ_{i+1}:

1. Suppose the event was an arrival. Then $f(q,\tau_{i+1},\omega) = 1 + f(q,\tau_i,\omega)$ and

$f(r, \tau_{i+1}, \omega) = 1 + f(r, \tau_i, \omega)$, so

$$f(q, \tau_{i+1}, \omega) - f(r, \tau_{i+1}, \omega) = f(q, \tau_i, \omega) - f(r, \tau_i, \omega),$$

thus (6.36) holds.

2. Suppose the event was a potential departure. By our inductive hypothesis, we must be in one of the two cases below:
   (a) $f(q, \tau_i, \omega) = f(r, \tau_i, \omega)$. Then by our coupling, the system under initial condition q and under initial condition r must both either have an actual departure or have no departure at τ_{i+1}. As the change in queue lengths will be the same, by the same reasoning as when we have an arrival, we continue to satisfy (6.36).
   (b) $f(q, \tau_i, \omega) > f(r, \tau_i, \omega)$. Now either both systems have a departure, or only the system with initial queue length q has a departure (as it has more active servers), which with the inductive hypothesis implies

$$f(q, \tau_{i+1}, \omega) - f(r, \tau_{i+1}, \omega) \leq f(q, \tau_i, \omega) - f(r, \tau_i, \omega) \leq q - r.$$

   When both systems experience an actual departure, $f(q, \tau_{i+1}, \omega) - f(r, \tau_{i+1}, \omega) = f(q, \tau_i, \omega) - f(r, \tau_i, \omega) \geq 0$ where the inequality is by the inductive hypothesis. When only the system under initial condition q has a departure, we still have

$$f(q, \tau_{i+1}, \omega) - f(r, \tau_{i+1}, \omega) = f(q, \tau_i, \omega) - 1 - f(r, \tau_i, \omega) \geq 0,$$

   where the inequality holds by initial assumption for this case. Thus (6.36) holds.

□

Next we establish a few properties of Γ, defined in (6.3).

Lemma 6.4. *For all $\kappa \geq 0$ and all $(q,r),(q',r') \in \mathbb{R}^2_+$, the operator $\Gamma(\cdot;\kappa)$ satisfies*

$$n\Gamma\left(\frac{(q,r)}{n};\kappa\right) = \Gamma(q,r;n\kappa), \tag{6.37}$$

$$\|\Gamma(q,r;\kappa) - \Gamma(q',r';\kappa)\|_1 \leq \|(q,r)-(q',r')\|_1. \tag{6.38}$$

Namely, with respect to the ℓ_1 norm, $\Gamma(\cdot;\kappa)$ is 1-Lipschitz continuous. Further, each component of $\Gamma(q,r;\kappa)$ increases monotonically in both q and r.

Proof. The first property, follows immediately by definition of Γ, as

$$n\Gamma\left(\frac{(q,r)}{n};\kappa\right) = n\left(\frac{q}{n}\wedge\kappa, \left(\frac{q}{n}-\kappa\right)^+ - \frac{r}{n}\right) = \left(q\wedge n\kappa, (q-n\kappa)^+ - r\right) = \Gamma(q,r;n\kappa).$$

To show the second part, we find that

$$\begin{aligned}\|\Gamma(q,r;\kappa) - \Gamma(q',r';\kappa)\|_1 &= |(q-\kappa)^+ + r - (q'-\kappa)^+ - r'| + |(q\wedge\kappa) - (q'\wedge\kappa)| \\ &\leq |(q-\kappa)^+ - (q'-\kappa)^+| + |r-r'| + |(q\wedge\kappa)-(q'\wedge\kappa)|.\end{aligned}$$

Without loss of generality, suppose $q \geq q'$. Now by considering the exhaustive cases $q' \geq \kappa$, $q > \kappa > q'$, and $\kappa \geq q$, the Lipschitz continuity follows trivially. The monotonicity property is an immediate consequence of (6.3). □

Corollary 6.2. *For $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, fix any $\mathbf{s}, \bar{\mathbf{s}} \in \mathcal{S}_\theta$. Assume that the processes $\mathbf{S}_\theta(t)$ has $\mathbf{S}_\theta(0) = \mathbf{s}$ and let $\tilde{\mathbf{S}}_\theta(t)$ be a version of $\mathbf{S}_\theta(t)$ with instead $\tilde{\mathbf{S}}_\theta(0) = \tilde{\mathbf{s}}$. There is a coupling between these processes such that for all $t \geq 0$,*

$$\|\mathbf{S}_\theta(t) - \tilde{\mathbf{S}}_\theta(t)\|_1 \leq \|\mathbf{s}-\tilde{\mathbf{s}}\|_1.$$

Moreover, $\mathbf{s} \geq \tilde{\mathbf{s}}$ componentwise implies $\mathbf{S}_\theta(t) \geq \tilde{\mathbf{S}}_\theta(t)$ componentwise for all t.

Proof. It suffices to show the claim for all $t \in (0,1]$, as then the result follows by induction. For $\theta = \mathrm{LS}$, for all $t \in (0,1)$, $Q_{\mathrm{LS}}(t)$ and $\tilde{Q}_{\mathrm{LS}}(t)$ are $M(t)/M/c$ queues with the same arrival and service rates. Thus by Lemma 6.3, we obtain that $Q_{\mathrm{LS}}(t)$ is monotone increasing and 1-Lipschitz in q. Likewise, $R_{\mathrm{LS}}(t)$ and $\tilde{R}_{\mathrm{LS}}(t)$ are $M(t)/M/c$

queues with arrival rate zero and the same service rate, so again by Lemma 6.3, we obtain that $R_{\mathrm{LS}}(t)$ is monotone increasing and 1-Lipschitz in r. As $Q_{\mathrm{LS}}(t)$ does not depend on r and $R_{\mathrm{LS}}(t)$ does not depend on q, the claim holds for $t \in (0,1)$. For $t = 1$,

$$\begin{aligned}\left\|\mathbf{S}_{\mathrm{LS}}(1) - \tilde{\mathbf{S}}_{\mathrm{LS}}(1)\right\|_1 &= \left\|\Gamma(\mathbf{S}_{\mathrm{LS}}(1^-); c) - \Gamma(\tilde{\mathbf{S}}_{\mathrm{LS}}(1^-); c)\right\|_1 \\ &\leq \left\|\mathbf{S}_{\mathrm{LS}}(1^-) - \tilde{\mathbf{S}}_{\mathrm{LS}}(1^-)\right\|_1 \qquad (6.39) \\ &\leq \|\mathbf{s} - \tilde{\mathbf{s}}\|_1 , \qquad (6.40)\end{aligned}$$

where (6.39) follows from Lemma 6.4 and (6.40) follows from our analysis of the case $t \in (0,1)$. Monotonicity follows as each component of $\mathbf{S}_\theta(1)$ is the composition of monotone increasing functions and thus monotone increasing in every input, again by Lemma 6.4 and our analysis of the case $t \in (0,1)$. For $\theta = \mathrm{DA}$, the proof is very similar. □

We also need a simple uniform bound on a sequence of Poisson random variables.

Lemma 6.5. *If $X_n \stackrel{d}{=} \mathrm{Pois}(\gamma n)$, then for all $k > 2\gamma(e-1)$ and all $n = 1, 2, \ldots,$*

$$\mathbb{P}(X_n \geq kn) \leq e^{-kn/2}.$$

Proof. We have

$$\mathbb{P}(X_n \geq kn) \leq \exp(-kn)\mathbb{E}[\exp(X_n)] = \exp(-kn)\exp(\gamma n(e-1)) \leq \exp(-kn/2).$$

□

We now analyze our system in the fluid scaling. As before let $A^n \stackrel{\Delta}{=} A^n(0,1)$, $D^{\mathrm{on},n}_{\mathrm{LS}} \stackrel{\Delta}{=} D^{\mathrm{on},n}_{\mathrm{LS}}(0,1^-)$, $D^{\mathrm{on},n}_{\mathrm{DA}} \stackrel{\Delta}{=} D^{\mathrm{on},n}_{\mathrm{DA}}(0,\frac{1}{2}^-)$, $D^{\mathrm{off},n}_{\mathrm{LS}} \stackrel{\Delta}{=} D^{\mathrm{off},n}_{\mathrm{LS}}(0,1^-)$, and $D^{\mathrm{off},n}_{\mathrm{DA}} \stackrel{\Delta}{=} D^{\mathrm{on},n}_{\mathrm{DA}}(\frac{1}{2},1^-)$. Using Lemma 6.5, we show that when $\rho_\theta < 1$, the convergence as q goes to infinity in Lemma 6.1 is uniform in n.

Lemma 6.6. *We have*

$$\lim_{k\to\infty} \sup_{n>0} \mathbb{E}_{(nk,nc)} \left[\frac{V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))}{n} \right] = -\gamma_{\mathrm{LS}}.$$

Proof. From (6.27), we have that for all k and n,

$$\begin{aligned}\mathbb{E}_{(nk,nc)} \left[\frac{V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))}{n} \right] &= \mathbb{E}_{(nk,nc)} \left[\frac{A^n - D^{\mathrm{on},n}_{\mathrm{LS}} - D^{\mathrm{off},n}}{n} \right] \\ &= \lambda - c(1-e^{-\mu}) - \mathbb{E}_{(nk,nc)} \left[\frac{D^{\mathrm{on},n}_{\mathrm{LS}}}{n} \right].\end{aligned}$$

where the second equality follows as $A^n \stackrel{d}{=} \mathrm{Pois}(n\lambda)$ and $D^{\mathrm{off},n}_{\mathrm{LS}} \stackrel{d}{=} \mathrm{Bin}(nc, 1-e^{-\mu})$ (see discussion (6.29)). As in (6.28) we have $\tilde{D}^{\mathrm{on},n}_{\mathrm{LS}} \stackrel{d}{=} \mathrm{Pois}(nc\mu)$ coupled with $D^{\mathrm{on},n}_{\mathrm{LS}}$ such that $\tilde{D}^{\mathrm{on},n}_{\mathrm{LS}} \geq D^{\mathrm{on},n}_{\mathrm{LS}}$. Thus $\mathbb{E}[D^{\mathrm{on},n}_{\mathrm{LS}}] \leq nc\mu$ for all n and all k, so it suffices to show that

$$\liminf_{k\to\infty} \sup_{n>0} \mathbb{E}_{(nk,nc)} \left[\frac{D^{\mathrm{on},n}_{\mathrm{LS}}}{n} \right] \geq c\mu.$$

As in (6.30), we find that

$$\liminf_{k\to\infty} \sup_{n>0} \frac{1}{n} \mathbb{E}_{(nk,nc)}[D^{\mathrm{on},n}_{\mathrm{LS}}] \geq \lim_{k\to\infty} \sup_{n>0} \frac{1}{n} \mathbb{E}[\tilde{D}^{\mathrm{on},n}_{\mathrm{LS}} \mathbb{I}_{\{\tilde{D}^{\mathrm{on},n}_{\mathrm{LS}} \leq (k-c)n\}}].$$

Now

$$\begin{aligned}
&\lim_{k\to\infty}\sup_{n>0}\left|\frac{1}{n}\mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n}\mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n}\le(k-c)n\}}]-c\mu\right| \\
&\quad=\lim_{k\to\infty}\sup_{n>0}\left|\frac{1}{n}\mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n}\mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n}\le(k-c)n\}}-\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n}]\right| \\
&\quad=\lim_{k\to\infty}\sup_{n>0}\frac{1}{n}\mathbb{E}[\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n}\mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n}\ge(k-c)n\}}] \\
&\quad\le\lim_{k\to\infty}\sup_{n>0}\frac{1}{n}\sqrt{\mathbb{E}[(\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n})^2]\mathbb{E}[\mathbb{I}_{\{\tilde{D}_{\mathrm{LS}}^{\mathrm{on},n}\ge(k-c)n\}}]} \\
&\quad\le\lim_{k\to\infty}\sup_{n>0}\frac{1}{n}\sqrt{((nc\mu)^2+nc\mu)\exp(-kn/2)\exp(cn)} \qquad (6.41)\\
&\quad=\lim_{k\to\infty}\sqrt{((c\mu)^2+c\mu)\exp(-k/2)\exp(c)} \qquad (6.42)\\
&\quad=0.
\end{aligned}$$

Here (6.41) follows from Lemma 6.5, and (6.42) follows as when k, is large, the supremum is attained by taking $n=1$. □

Lemma 6.7. *We have*

$$\lim_{k\to\infty}\sup_{n>0}\mathbb{E}_{(nk,0)}\left[\frac{V(\mathbf{S}_{\mathrm{DA}}^n(1))-V(\mathbf{S}_{\mathrm{DA}}^n(0))}{n}\right]=-\gamma_{\mathrm{DA}}.$$

The proof is very similar to previous lemma and omitted. We can give the uniform moment bounds for $\mathbf{S}_\theta^n(\infty)$.

Lemma 6.8. *For each policy $\theta\in\{\mathrm{LS},\mathrm{DA}\}$, when $\rho_\theta<1$, there exists a constant M_θ depending on λ, c and μ such that for every $n>0$, $\mathbb{E}\left[V(\mathbf{S}_\theta^n(\infty))\right]\le M_\theta n$.*

Proof. As $\rho_\theta<1$, we have $\gamma_\theta>0$. By Lemma 6.6, there exists $\bar{k}_{\mathrm{LS}}$ such that for all $k>\bar{k}_{\mathrm{LS}}$ and all n,

$$\frac{1}{n}\mathbb{E}_{(nk,nc)}[V(\mathbf{S}_{\mathrm{LS}}^n(1))-V(\mathbf{S}_{\mathrm{LS}}^n(0))]\le-\frac{\gamma_{\mathrm{LS}}}{2}.$$

For any $\bar{q} > q$,

$$0 \leq \mathbb{E}_{(\bar{q},r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1))] - \mathbb{E}_{(q,r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1))] \leq \bar{q} - q,$$

by the monotonicity and 1-Lipschitz of Corollary 6.2. Thus

$$\begin{aligned}
&\mathbb{E}_{(\bar{q},r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))] - \mathbb{E}_{(q,r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))] \\
&\quad = \mathbb{E}_{(\bar{q},r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1))] - \mathbb{E}_{(q,r)}[V(\mathbf{S}^n_{\mathrm{LS}}(1))] + q - \bar{q} \\
&\quad \leq 0.
\end{aligned}$$

As a result, we obtain that for every n, for all $q > \bar{k}_{\mathrm{LS}} n$,

$$\frac{1}{n}\mathbb{E}_{(q,nc)}[V(\mathbf{S}^n_{\mathrm{LS}}(1)) - V(\mathbf{S}^n_{\mathrm{LS}}(0))] \leq -\frac{\gamma_{\mathrm{LS}}}{2}. \tag{6.43}$$

We define $\bar{k}_{\mathrm{DA}}$ analogously using Lemma 6.7 and γ_{DA}. Let

$$b_\theta \triangleq \max\left\{\bar{k}_\theta, \frac{1}{\gamma_\theta}\left(\lambda^2 + \lambda + C_\theta(c^2 + c)\right)\right\},$$

where C_θ is as defined by (6.25) for LS and (6.32) for DA. Thus by Lemma 6.1 and Lemma 6.2, for all $\mathbf{s} \in \mathcal{S}^n_\theta$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{s}}[(V(\mathbf{S}^n_\theta(1)) - V(\mathbf{S}^n_\theta(0)))^2] &\leq (\lambda^n)^2 + \lambda^n + C_\theta((c^n)^2 + c^n) \\
&= n^2\lambda^2 + n\lambda + C_\theta(n^2c^2 + nc) \\
&\leq b_\theta \gamma_\theta n^2.
\end{aligned} \tag{6.44}$$

We let

$$
\begin{aligned}
B_\theta^n &\triangleq \{(q,r) \in \mathcal{S}_{\mathrm{LS}} \mid q + r \leq 2nb_\theta\}, \qquad (6.45)\\
U(q,r) &\triangleq (q+r)^2,\\
\alpha_\theta^n &\triangleq 2nb_\theta,\\
\beta_\theta^n &\triangleq n^2 b_\theta(4\lambda + \gamma_\theta),\\
\gamma_\theta^n &\triangleq n\gamma_\theta/2.
\end{aligned}
$$

and apply Proposition A.3 for each n, using U as our Lyapunov function and $f(q,r) = q + r$. We observe that (A.6) holds trivially. For $(q,r) \in \mathcal{S}_\theta^n$ we have

$$
\begin{aligned}
\mathbb{E}_{(q,r)}[U(\mathbf{S}_\theta^n(1)) - U(\mathbf{S}_\theta^n(0))] &= \mathbb{E}_{(q,r)}[(q + r + A^n - D_\theta^{\mathrm{on},n} - D_\theta^{\mathrm{off},n})^2] - (q+r)^2\\
&= 2(q+r)\mathbb{E}_{(q,r)}[A^n - D_\theta^{\mathrm{on},n} - D_\theta^{\mathrm{off},n}]\\
&\quad + \mathbb{E}_{(q,r)}\left[(A^n - D_\theta^{\mathrm{on},n} - D_\theta^{\mathrm{off},n})^2\right]\\
&= 2(q+r)\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0))]\\
&\quad + \mathbb{E}_{(q,r)}[(V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0)))^2]\\
&\leq 2(q+r)\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0))] + n^2\gamma_\theta b_\theta, \qquad (6.46)
\end{aligned}
$$

where (6.46) is a consequence of (6.44). Now, for $(q,r) \in B_\theta^n$, using (6.45) with (6.46), we see that

$$
\begin{aligned}
2(q+r)\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0))] + n^2\gamma_\theta b_\theta &\leq 4nb_\theta\mathbb{E}_{(q,r)}[A^n(0,1^-)] + n^2\gamma_\theta b_\theta\\
&\leq n^2 b_\theta(4\lambda + \gamma_\theta)\\
&= \beta_\theta^n,
\end{aligned}
$$

showing that (A.7) holds. Finally, for $\mathbf{s} \in \mathcal{S}_\theta^n \setminus B_\theta^n$,

$$\begin{aligned}
\mathbb{E}_{(q,r)}[U(\mathbf{S}_\theta^n(1)) - U(\mathbf{S}_\theta^n(0))] &\leq 2(q+r)\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta^n(1)) - V(\mathbf{S}_\theta^n(0))] + n^2\gamma_\theta b_\theta \quad (6.47)\\
&\leq -\gamma_\theta n(q+r) + n^2\gamma_\theta b_\theta \quad (6.48)\\
&\leq -\frac{\gamma_\theta}{2}n(q+r) - \frac{\gamma_\theta}{2}n(2nb_\theta) + n^2\gamma_\theta b_\theta \quad (6.49)\\
&= -\frac{\gamma_\theta n}{2}f(q,r),
\end{aligned}$$

where (6.47) follows from (6.46), (6.48) follows as $\mathbf{s} \in \mathcal{S}_\theta^n \setminus B_\theta^n$ so we can apply (6.43), and (6.49) follows as $\mathbf{s} \in \mathcal{S}_\theta^n \setminus B_\theta^n$ ensures $q + r > 2nb_\theta$. This shows that (A.5) is satisfied for each n. Thus for every n we can apply Proposition A.3 to obtain that

$$\mathbb{E}[f(\mathbf{S}^n(\infty))] \leq \alpha^n + \frac{\beta^n}{\gamma^n} = n\left(2b_\theta\left(4\lambda/\gamma_\theta + 1\right)\right),$$

showing the result. □

6.7 Fluid Model Approximations. Proof of Theorem 6.2

In this section, we establish Theorem 6.2. First, we introduce the following additional notation to be used throughout the section. Let u_θ^i be the time of the ith shift change under policy θ, i.e. for $i = 0, 1, 2, \ldots$, $u_{\text{LS}}^i \triangleq i$ and $u_{\text{DA}}^i \triangleq i/2$. Let c_θ^i be the number of residents on shift during $[u_i, u_{i+1})$, i.e. $c_{\text{LS}}^i \triangleq c$ for $i = 0, 1, 2, \ldots$, $c_{\text{DA}}^i \triangleq 2c$ for even i, and $c_{\text{DA}}^i \triangleq 0$ for odd i.

To prove the result, we will invoke a theorem from [60] that shows the convergence of multidimensional Markovian queueing processes to its fluid limit under the so called "uniform acceleration." Consider a Markov process $\mathbf{X}(t)$ on state space $\mathbb{Z}_+^m$ with transition rates that depend both on the current state and the time. The process $\mathbf{X}(t)$ is driven by a finite set of independent rate one exogenous Poison process $E_i(t)$, $i = 1, \ldots, k$. The events from these processes trigger a "jump" $\mathbf{v}_i \in \mathbb{Z}^m$ in $\mathbf{X}(t)$. For each process i, there is a rate function $\alpha_i(\mathbf{x}, t)\colon \mathbb{R}_+^m \times \mathbb{R}_+ \to \mathbb{R}_+$ that depends both on

the state $\mathbf{x}$ and the time t. Assume that for each i and $\bar{t} \in \mathbb{R}_+$, $\alpha_i(\cdot, \bar{t})$ is γ_i-Lipschitz in $\mathbf{x}$ where γ_i does not depend on $\bar{t}$. We define $\mathbf{X}(t)$ by

$$\mathbf{X}(t) \triangleq \mathbf{X}(0) + \sum_{i=1}^{k} \mathbf{v}_i E_i \left(\int_0^t \alpha_i(\mathbf{X}(\tau), \tau) d\tau \right),$$

In Theorem 9.2 from [60], it is shown that this procedure uniquely defines the process $\mathbf{X}(t)$. Next, we consider a deterministic process $\mathbf{x}(t)$ on $\mathbb{R}_+^m$ defined by

$$\mathbf{x}(t) \triangleq \mathbf{x}(0) + \sum_{i=1}^{k} \mathbf{v}_i \int_0^t \alpha_i(\mathbf{x}(\tau), \tau) d\tau.$$

Again the existence and uniqueness of such an $\mathbf{x}(t)$ is shown in Theorem 11.4 from [60]. To approximate $\mathbf{X}(t)$ by $\mathbf{x}(t)$, we consider a sequence of processes $\mathbf{X}^n(t)$, $n = 1, 2, \ldots$, defined by

$$\mathbf{X}^n(t) \triangleq \mathbf{X}^n(0) + \sum_{i=1}^{k} \mathbf{v}_i E_i \left(n \int_0^t \alpha_i \left(\frac{\mathbf{X}^n(\tau)}{n}, \tau \right) d\tau \right),$$

i.e. $\mathbf{X}^1(t)$ is the original process. A special case of their result is as follows.

Proposition 6.2 ([60], Theorem 2.2)**.** *If* $\mathbf{X}^n(0)/n \to \mathbf{x}(0)$ *a.s., then*

$$\lim_{n \to \infty} \frac{\mathbf{X}^n(t)}{n} = \mathbf{x}(t),$$

a.s. and u.o.c.

We now return to our model. On the intervals $[u_\theta^i, u_\theta^{i+1})$ between shift changes, our processes $\mathbf{S}_\theta(t)$ and $\mathbf{s}_\theta(t)$ are of the form of $\mathbf{X}(t)$ and $\mathbf{x}(t)$ from the theorem. Specifically, we can take $\mathbf{v}_1 = (1, 0)$ and $\alpha_1((q, r), t) = \lambda(u_\theta^i + t)$ so that the arrival

process $A(u_\theta^i, u_\theta^i + t) = E_1(t)$ for $t \in [u_\theta^i, u_\theta^{i+1}]$. Similarly, we take $\mathbf{v}_2 = (-1, 0)$ and

$$\alpha_2((q,r),t) \triangleq (c_\theta^i \wedge q)\mu = \begin{cases} (c \wedge q)\mu & \theta = \text{LS}, \\ (2c \wedge q)\mu & \theta = \text{DA}, \ i \text{ even}, \\ 0 & \theta = \text{DA}, \ i \text{ odd}, \end{cases}$$

then $D^{\text{on}}(u_\theta^i, t) = E_2(t)$ for $t \in [u_\theta^i, u_\theta^{i+1}]$. Finally, we take $\mathbf{v}_3 = (0, -1)$ and $\alpha_3((q,r),t) \triangleq r\mu$ so that $D^{\text{off}}(u_\theta^i, t) = E_3(t)$ for $t \in [u_\theta^i, u_\theta^{i+1}]$. We satisfy the Lipschitz condition on $\alpha_i(\cdot, t)$ as α_1 does not depend on the state and both α_2 and α_3 are μ-Lipschitz in (q,r) independent of t.

Thus the proposition immediately yields that if $\mathbf{S}_\theta^n(u_\theta^i)/n \to \mathbf{s}_\theta(u_\theta^i)$ a.s., then $\mathbf{S}_\theta^n(t)/n \to \mathbf{s}_\theta(t)$ u.o.c. From this point, the primary difficulty in proving Theorem 6.2 is showing that $\mathbf{S}_\theta(t)$ jumping at each shift change does not ruin the convergence. We can now prove the main result of the section.

Proof of Theorem 6.2. For each policy $\theta \in \{\text{LS}, \text{DA}\}$, we will show by induction on i that $\mathbf{S}_\theta^n(t)/n \to \mathbf{s}_\theta(t)$ a.s. and uniformly on $[0, u_\theta^i]$. The case of $i = 0$ holds by the assumption of the theorem.

Suppose the claim holds for i. We notice that

$$\sup_{0 \le \tau \le u_\theta^{i+1}} \left\| \frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau) \right\|_1 =$$
$$\max \left\{ \sup_{0 \le \tau \le u_\theta^i} \left\| \frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau) \right\|_1, \sup_{u_\theta^i \le \tau < u_\theta^{i+1}} \left\| \frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau) \right\|_1, \left\| \frac{\mathbf{S}_\theta^n(u_\theta^{i+1})}{n} - \mathbf{s}_\theta(u_\theta^{i+1}) \right\|_1 \right\}.$$

By the definitions of u_θ^i and c_θ^i, it follows immediately that when applying Γ for the shift change at time u_{i+1}, we use $\Gamma(\cdot; c_\theta^i)$. Recalling that $\mathbf{S}_\theta^n$ and $\mathbf{s}_\theta$ are a.s. RCLL,

and applying (6.37) and then (6.38) from Lemma 6.4,

$$\begin{aligned}\left\|\frac{\mathbf{S}_\theta^n(u_\theta^{i+1})}{n} - \mathbf{s}_\theta(u_\theta^{i+1})\right\|_1 &= \left\|\frac{\Gamma(\mathbf{S}_\theta^n((u_\theta^{i+1})^-); nc_\theta^i)}{n} - \Gamma(\mathbf{s}_\theta((u_\theta^{i+1})^-); c_\theta^i)\right\|_1 \\ &= \left\|\Gamma\left(\frac{\mathbf{S}_\theta^n((u_\theta^{i+1})^-)}{n}; c_\theta^i\right) - \Gamma\left(\mathbf{s}_\theta((u_\theta^{i+1})^-); c_\theta^i\right)\right\|_1 \\ &\leq \left\|\frac{\mathbf{S}_\theta^n((u_\theta^{i+1})^-)}{n} - \mathbf{s}_\theta((u_\theta^{i+1})^-)\right\|_1.\end{aligned}$$

For $\tau \in [u_\theta^i, u_\theta^{i+1}]$, we let $\bar{\mathbf{S}}_\theta^n(t)$ and $\bar{\mathbf{s}}_\theta(t)$ be the continuous extension of $\mathbf{S}_\theta^n(t)$ and $\mathbf{s}_\theta(t)$, respectively, from $[u_\theta^i, u_\theta^{i+1})$ to $[u_\theta^i, u_\theta^{i+1}]$, i.e.,

$$\bar{\mathbf{S}}_\theta^n(\tau) \triangleq \begin{cases}\mathbf{S}_\theta^n(\tau) & \tau < u_\theta^{i+1}, \\ \mathbf{S}_\theta^n((u_\theta^{i+1})^-) & \tau = u_\theta^{i+1},\end{cases} \qquad \bar{\mathbf{s}}_\theta(\tau) \triangleq \begin{cases}\mathbf{s}_\theta(\tau) & \tau < u_\theta^{i+1}, \\ \mathbf{s}_\theta((u_\theta^{i+1})^-) & \tau = u_\theta^{i+1}.\end{cases}$$

Thus we obtain that

$$\sup_{0\leq\tau\leq u_\theta^{i+1}}\left\|\frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau)\right\|_1 = \max\left\{\sup_{0\leq\tau\leq u_\theta^i}\left\|\frac{\mathbf{S}_\theta^n(\tau)}{n} - \mathbf{s}_\theta(\tau)\right\|_1, \sup_{u_\theta^i\leq\tau\leq u_\theta^{i+1}}\left\|\frac{\bar{\mathbf{S}}_\theta^n(\tau)}{n} - \bar{\mathbf{s}}_\theta(\tau)\right\|_1\right\}.$$

By induction, the first term in the above maximum goes to zero a.s. and in particular $\mathbf{S}_\theta^n(u_\theta^i)/n \to \mathbf{s}_\theta(u_\theta^i)$ a.s. By Proposition 6.2, the second term converges to zero a.s. as well, as previously discussed. □

Remark 6.2. The result from [60] is actually stronger than what we suggested. It implies that a.s. and u.o.c., as $n \to \infty$, $\|\mathbf{S}_\theta^n(t)/n - \mathbf{s}_\theta(t)\|_1 \leq O(\log n)$. This can be generalized to our case inductively in the same manner, but we do not pursue this further.

6.8 Long Run Behavior of the Fluid Model

In this section, for each policy $\theta \in \{\text{LS}, \text{DA}\}$, we show that the fluid limit at integer times $\{\mathbf{s}_\theta(k)\}$, which is a deterministic discrete time dynamical system, has a simple long run behavior. To show this, we need to recall definitions from Section 6.2. For

a set $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X} \neq \emptyset$, a function $\mathbf{f}\colon \mathcal{X} \to \mathcal{X}$, and an initial condition $\mathbf{x}_0 \in \mathcal{X}$, let $\mathbf{x}_k$, $k \in \mathbb{Z}_+$ be defined by $\mathbf{f}(\mathbf{x}_k) = \mathbf{x}_{k+1}$. Recall from Section 6.2 that a point $\mathbf{x}^* \in \mathcal{X}$ is *attractive* if for every $\mathbf{x}_0 \in \mathcal{X}$, $\lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}^*$. As previously mentioned, such an $\mathbf{x}^*$ must be unique. Further, when $\mathbf{f}$ is continuous on $\mathcal{X}$, it immediately follows that $\mathbf{f}(\mathbf{x}^*) = \mathbf{x}^*$, i.e. $\mathbf{x}^*$ is a *fixed point* of $\mathbf{f}$. First, we give a known (e.g. [18] page 183) criterion for identifying attractive points.

Proposition 6.3. *Suppose $\mathcal{X}$ is nonempty and compact, $\mathbf{f}\colon \mathcal{X} \to \mathcal{X}$ is continuous on $\mathcal{X}$, and for every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,*

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_p < \|\mathbf{x} - \mathbf{y}\|_p,$$

for some $p \geq 1$. Then there exists a unique attractive point $\mathbf{x}^* \in \mathcal{X}$.

We now give a sufficient condition to ensure that in finite time $\{\mathbf{x}_k\}$ will reach a bounded set, such as the compact set in the previous theorem.

Proposition 6.4. *If there is a function $V\colon \mathcal{X} \to \mathbb{R}_+$, $\gamma > 0$ and $B \subset \mathcal{X}$ such that for all $\mathbf{x} \in \mathcal{X} \setminus B$,*

$$V(\mathbf{f}(\mathbf{x})) - V(\mathbf{x}) \leq -\gamma,$$

then for all $\mathbf{x}_0 \in \mathcal{X} \setminus B$, there exists $m \leq \lceil V(\mathbf{x}_0)/\gamma \rceil$ such that $\mathbf{x}_m \in B$.

Proof. Let $n = \lceil V(\mathbf{x}_0)/\gamma \rceil + 1$ and assume for contradiction that $\mathbf{x}_0, \ldots, \mathbf{x}_n$ are all not in B. Then

$$V(\mathbf{x}_n) = V(\mathbf{x}_0) + \sum_{k=1}^{n} V(\mathbf{x}_k) - V(\mathbf{x}_{k-1}) \leq V(\mathbf{x}_0) - n\gamma < 0,$$

contradicting the non-negativity of V. □

Finally, we give a criteria for instability.

Proposition 6.5. *Suppose* $V\colon \mathcal{X} \to \mathbb{R}_+$ *is a continuous function such that* $\sup\{V(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\} = \infty$*, and for all* $\mathbf{x} \in \mathcal{X}$*,*

$$V(\mathbf{f}(\mathbf{x})) - V(\mathbf{x}) \geq 0.$$

Then an attractive point does not exist.

Proof. Assume for contradiction there were an attractive point $\mathbf{x}^*$. Let $\mathbf{x}_0 \in \mathcal{X}$ be such that $V(\mathbf{x}_0) > V(\mathbf{x}^*)$. Such a point exists as have assumed that the supremum of V is infinite. However, as $\mathbf{x}^*$ is attractive and V is continuous,

$$V(\mathbf{x}^*) = \lim_{n\to\infty} V(\mathbf{x}_n) = V(\mathbf{x}_0) + \lim_{n\to\infty} \sum_{k=1}^{n} V(\mathbf{x}_k) - V(\mathbf{x}_{k-1}) \geq V(\mathbf{x}_0),$$

contradicting $V(\mathbf{x}_0) > V(\mathbf{x}^*)$. □

We now consider the differential equation that controls the evolution of the state for the fluid model.

Lemma 6.9. *Given parameters* $(\gamma, m, \mu, x(0)) \in \mathbb{R}_+^4$*, the differential equation*

$$\dot{x}(t) = \gamma - (x(t) \wedge m)\mu, \tag{6.50}$$

has a unique solution. The solution $x(t)$ *is monotone in* t *and satisfies* $x(t) \geq 0$ *for all* $t \geq 0$*. Further, if* $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ *is defined by* $g(x(0)) = x(\frac{1}{2})$*, then* g *is strictly increasing and 1-Lipschitz. Finally, if we let*

$$\tilde{x} \stackrel{\Delta}{=} \begin{cases} m & \gamma \geq m\mu, \\ m - (\gamma - m\mu)/2 & \gamma < m\mu, \end{cases}$$

then $x(0) \geq \tilde{x}$ *implies that:*

(a) $x(t) \geq m$ *for* $t \in [0, \frac{1}{2})$*,*

(b) $x(\frac{1}{2}) = x(0) + (\gamma - m\mu)/2$*,*

(c) For all $y > x(0)$*,* $g(y) - g(x(0)) = y - x(0)$*.*

On the other hand, if $x(0) < \tilde{x}$, then each of (a), (b) and (c) above are violated. In particular,

(a') There exists $s \in [0, \frac{1}{2})$ such that $x(s) < m$,

(b') $x(\frac{1}{2}) > x(0) + (\gamma - m\mu)/2$,

(c') For all $y > x(0)$, $g(y) - g(x(0)) < y - x(0)$.

The proof is rather lengthy but no difficult, so it is deferred to Section 6.11. We now use this to analyze the fluid limits of LS and DA. Let $g^1_{\mathrm{LS}}(q)$ and $g^2_{\mathrm{LS}}(q)$ be the function g from Lemma 6.9 when the parameters (λ_1, c, μ, q) and (λ_2, c, μ, q) are used, respectively. Let $\mathbf{f}_{\mathrm{LS}} \colon \mathcal{T}_{\mathrm{LS}} \to \mathcal{T}_{\mathrm{LS}}$ be given by

$$\mathbf{f}_{\mathrm{LS}}(q, r) \triangleq \Gamma(g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)), re^{-\mu}; c). \tag{6.51}$$

Similarly, let $g_{\mathrm{DA}}(q)$ be the function from Lemma 6.9 using parameters $(\lambda_1, 2c, \mu, q)$, and let

$$\mathbf{h}^1_{\mathrm{DA}}(q, r) \triangleq \Gamma(g_{\mathrm{DA}}(q), 0; 2c), \tag{6.52}$$

$$\mathbf{h}^2_{\mathrm{DA}}(q, r) \triangleq \Gamma\left(q + \frac{\lambda_2}{2}, re^{-\mu/2}; 0\right), \tag{6.53}$$

$$\mathbf{f}_{\mathrm{DA}}(\mathbf{s}) \triangleq \mathbf{h}^2(\mathbf{h}^1(\mathbf{s})). \tag{6.54}$$

We now prove Proposition 6.1.

Proof of Proposition 6.1. To prove existence and uniqueness of $\mathbf{s}_\theta(t)$, it suffices to show that for every $k \in \mathbb{Z}_+$, $\mathbf{s}_\theta(t)$ exists and is uniquely defined for all $0 \le t \le k$. For $k = 0$, we are given $\mathbf{s}_\theta(0)$ in the statement of the proposition. Suppose the claim holds for k. We now consider cases on θ.

LS – By induction $\mathbf{s}_{\mathrm{LS}}(k)$ is uniquely defined. As $q_{\mathrm{LS}}(t)$ solves the differential equation on $[k, k + \frac{1}{2})$ as used to define $g^1_{\mathrm{LS}}(q)$ with initial condition $q_{\mathrm{LS}}(k)$, and likewise solves the differential equation on $[k + \frac{1}{2}, 1)$ used to define $g^2_{\mathrm{LS}}(q)$ with initial condition $q_{\mathrm{LS}}(k+\frac{1}{2})$, we obtain by Lemma 6.9 that $q_{\mathrm{LS}}(t)$ is uniquely determined on $[k, k+1)$. As $\dot{r}_{\mathrm{LS}}(t) = -\mu r_{\mathrm{LS}}(t)$ and by induction $r_{\mathrm{LS}}(k)$ is uniquely defined,

we obtain that for $t \in [k, k+1)$, $r(t) = r(k)e^{-\mu(t-k)}$, uniquely defining $\mathbf{s}_{\mathrm{LS}}(t)$ on that interval as well. Thus we immediately obtain that for every $\mathbf{s}_{\mathrm{LS}}(k) \in \mathcal{T}_{\mathrm{LS}}$,

$$\mathbf{s}_{\mathrm{LS}}(k+1) = \mathbf{f}_{\mathrm{LS}}(\mathbf{s}_{\mathrm{LS}}(k)),$$

showing the hypothesis.

DA – The argument is similar. Briefly, for all $\mathbf{s}_{\mathrm{DA}}(k) \in \mathcal{T}_{\mathrm{DA}}$,

$$\begin{aligned}
\mathbf{s}_{\mathrm{DA}}(k+\tfrac{1}{2}) &= \mathbf{h}^1(\mathbf{s}_{\mathrm{DA}}(k)),\\
\mathbf{s}_{\mathrm{DA}}(k+1) &= \mathbf{h}^2(\mathbf{s}_{\mathrm{DA}}(k+\tfrac{1}{2})),\\
\mathbf{s}_{\mathrm{DA}}(k+1) &= \mathbf{f}_{\mathrm{DA}}(\mathbf{s}_{\mathrm{DA}}(k)),
\end{aligned}$$

and at intermediate times in $t \in (k, k+\frac{1}{2})$ and $t \in (k+\frac{1}{2}, k+1)$, $\mathbf{s}_{\mathrm{DA}}(t)$ is the unique solution to a linear ODE either with constant coefficients or of the type from Lemma 6.9.

Finally, we must show that $\mathbf{s}_\theta(k) \in \mathcal{T}_\theta$, $k \geq 1$. For LS, we must show that $r_{\mathrm{LS}}(k) < c$ implies $q_{\mathrm{LS}}(k) \leq c$. By definition

$$\mathbf{s}_{\mathrm{LS}}(k) = \Gamma(\mathbf{s}_{\mathrm{LS}}(k^-); c) = \begin{cases} (q_{\mathrm{LS}}(k^-) - c + r_{\mathrm{LS}}(k^-), c) & q_{\mathrm{LS}}(k^-) \geq c, \\ (r_{\mathrm{LS}}(k^-), q_{\mathrm{LS}}(k^-)) & q_{\mathrm{LS}}(k^-) < c. \end{cases}$$

Suppose $r_{\mathrm{LS}}(k) < c$. Note that $r_{\mathrm{LS}}(k) < c$ only in the second case. As $q_{\mathrm{LS}}(k) = r_{\mathrm{LS}}(k^-) \leq r_{\mathrm{LS}}(k-1) \leq c$, we obtain that $q_{\mathrm{LS}}(k) \leq c$, showing the claim. For DA, must simply show that $r_{\mathrm{DA}}(k) = 0$, which holds as

$$r_{\mathrm{DA}}(k) = \Gamma(\mathbf{s}_{\mathrm{DA}}(k^-); 0) = (q(k^-) + r(k^-), 0).$$

□

Next we give some simple structural properties of the functions g^i_{LS} and $\mathbf{f}_\theta$ that will be needed in the analysis of the long run behavior of the fluid limits.

Lemma 6.10. *There exists a unique* $\tilde{q}_{\text{LS}}$ *such that when* $q_{\text{LS}}(0) \geq \tilde{q}_{\text{LS}}$,

(a) $q_{\text{LS}}(t) \geq c$ *for* $t \in [0, 1)$,

(b) $q_{\text{LS}}(1^-) = q_{\text{LS}}(0) + \lambda - c\mu$,

(c) For $\bar{q} > q_{\text{LS}}(0)$, $g^2_{\text{LS}}(g^1_{\text{LS}}(\bar{q})) - g^2_{\text{LS}}(g^1_{\text{LS}}(q_{\text{LS}}(0))) = \bar{q} - q_{\text{LS}}(0)$.

and when $q_{\text{LS}}(0) < \tilde{q}_{\text{LS}}$, *(a), (b) and (c) are violated. In particular,*

(a') There exists $s \in [0, 1)$ *such that* $q_{\text{LS}}(s) < c$,

(b') $q_{\text{LS}}(1^-) > q_{\text{LS}}(0) + \lambda - c\mu$,

(c') For $\bar{q} > q_{\text{LS}}(0)$, $g^2_{\text{LS}}(g^1_{\text{LS}}(\bar{q})) - g^2_{\text{LS}}(g^1_{\text{LS}}(q_{\text{LS}}(0))) < \bar{q} - q_{\text{LS}}(0)$.

Proof. For $i = 1, 2$, let $\tilde{q}^i_{\text{LS}}$ be $\tilde{x}$ from Lemma 6.9 when used to create g^i_{LS}. It is easy to see that properties (a), (b) and (c) will hold when for all $i = 1, 2$, properties (a), (b) and (c) from the application Lemma 6.9 to create g^i_{LS} hold, i.e. we have both $q_{\text{LS}}(0) \geq \tilde{q}^1_{\text{LS}}$ and $q_{\text{LS}}(\frac{1}{2}) \geq \tilde{q}^2_{\text{LS}}$.

Similarly, it is easy to see that properties (a'), (b') and (c') will hold if there exists $i \in \{1, 2\}$ such that that properties (a'), (b') and (c') from the application of Lemma 6.9 to create g^i_{LS} hold, i.e. if either $q_{\text{LS}}(0) < \tilde{q}^1_{\text{LS}}$ or $q_{\text{LS}}(\frac{1}{2}) < \tilde{q}^2_{\text{LS}}$.

As $q_{\text{LS}}(\frac{1}{2}) = g^1_{\text{LS}}(q_{\text{LS}}(0))$ and by Lemma 6.4 g^1_{LS} is strictly increasing, there is a threshold q^* such that $q_{\text{LS}}(\frac{1}{2}) \geq \tilde{q}^2_{\text{LS}}$ iff $q_{\text{LS}}(0) \geq q^*$. Thus by taking $\tilde{q}_{\text{LS}} = \max\{q^*, q^1_{\text{LS}}\}$, we will have both $q_{\text{LS}}(0) \geq \tilde{q}^1_{\text{LS}}$ and $q_{\text{LS}}(\frac{1}{2}) \geq \tilde{q}^2_{\text{LS}}$ iff $q_{\text{LS}}(0) \geq \tilde{q}_{\text{LS}}$. This gives the result. □

Lemma 6.11. *For* $\theta \in \{\text{LS}, \text{DA}\}$, $\mathbf{f}_\theta$ *is 1-Lipschitz with respect to the* ℓ_1 *norm and both outputs of the function* $\mathbf{f}_\theta$ *are monotonically increasing in both inputs.*

Proof. For LS, the functions g^1_{LS} and g^2_{LS} are monotonically increasing and 1-Lipschitz by Lemma 6.9. Similarly $re^{-\mu}$ as a function of r is increasing and 1-Lipschitz, and by Lemma 6.4, $\Gamma(\cdot; c)$ is 1-Lipschitz and each component is monotonically increasing in every input. Thus $\mathbf{f}_{\text{LS}}(q, r)$ is a composition of monotone increasing 1-Lipschitz functions and thus monotone increasing and 1-Lipschitz.

The argument is similar for DA. The functions g_{DA}, $q + \lambda_2/2$ as a function of q, $re^{-\mu/2}$ as a function of r, $\Gamma(\cdot; 2c)$ and $\Gamma(\cdot; 0)$ are all 1-Lipschitz and monotonically increasing in every input (by Lemma 6.9 for g_{DA} and by Lemma 6.4 for $\Gamma(\cdot, 2c)$

and $\Gamma(\cdot, 0)$). Therefore $\mathbf{f}_{\mathrm{DA}}(q, r)$ is a composition of 1-Lipschitz monotone increasing functions and thus 1-Lipschitz and monotone increasing. □

We can now analyze the long run behavior of the fluid limits. First we will show that when $\rho_\theta < 1$, $\mathbf{f}_\theta$ restricted to some $T_\theta \subset \mathcal{T}_\theta$ has an attractive fixed point using contractive mapping (Proposition 6.3). Then we will use a Lyapunov function argument to show that the attractive point over T_θ is in fact attractive over all of $\mathcal{T}_\theta$ (Proposition 6.4). Finally, we will use another Lyapunov function argument to show that $\mathbf{f}_\theta$ has no attractive points when $\rho_\theta \geq 1$ (Proposition 6.5).

Proposition 6.6. *The process $\{\mathbf{s}_\theta(k)\}$ has a unique attractive fixed point iff $\rho_\theta < 1$.*

Proof. Assume $\rho_{\mathrm{LS}} < 1$. Recall $\tilde{q}_{\mathrm{LS}}$ from Lemma 6.10, and let $\tilde{q}_{\mathrm{DA}}$ be $\tilde{x}$ from Lemma 6.9 as applied to create g_{DA}. Let

$$T_\theta \triangleq \mathcal{T}_\theta \setminus \{(q, r) \mid q > \tilde{q}_\theta\}. \tag{6.55}$$

We now check the assumptions of Proposition 6.3 are satisfied by $\mathbf{f}_\theta$ restricted to T_θ. We can immediately verify by definition that $\mathbf{f}_\theta$ is the composition of continuous functions and thus continuous (recall that g^1_{LS}, g^2_{LS}, and g_{DA} are continuous by Lemma 6.9 and $\Gamma(\cdot, \kappa)$ is continuous for all κ by Lemma 6.4). Noting that $\tilde{q}_{\mathrm{LS}} \geq c$ by part (a) of Lemma 6.10, we see that $T_{\mathrm{LS}} = [0, c] \times [0, c] \cup \{(q, c) \mid 0 \leq q \leq \tilde{q}_{\mathrm{LS}}\}$ and thus it is a nonempty compact set. Likewise $T_{\mathrm{DA}} = \{(q, 0) \mid 0 \leq q \leq \tilde{q}_{\mathrm{DA}}\}$ where by Lemma 6.9 we see that $\tilde{q}_{\mathrm{DA}} \geq 2c$, thus T_{DA} is a nonempty compact set as well. We still need to check that $\mathbf{f}_\theta \colon T_\theta \to T_\theta$ and that $\mathbf{f}_\theta$ is contractive on T_θ.

To show $\mathbf{f}_{\mathrm{LS}} \colon T_{\mathrm{LS}} \to T_{\mathrm{LS}}$, we have that for any $(q, r) \in T_{\mathrm{LS}}$,

$$\begin{aligned}
\mathbf{f}_{\mathrm{LS}}(q, r) &\leq \mathbf{f}_{\mathrm{LS}}(\tilde{q}_{\mathrm{LS}}, c) && (6.56)\\
&= \Gamma\left(\tilde{q}_{\mathrm{LS}} + \lambda - c\mu, ce^{-\mu}; c\right) && (6.57)\\
&= \left(\tilde{q}_{\mathrm{LS}} + \lambda - c\mu - c + ce^{-\mu}, c\right) && (6.58)\\
&\leq (\tilde{q}_{\mathrm{LS}}, c), && (6.59)
\end{aligned}$$

where the inequalities are componentwise. Here (6.56) holds by Lemma 6.11. We obtain (6.57) by Lemma 6.10 part (b). Then (6.58) holds as we have $\tilde{q}_{\mathrm{LS}} + \lambda - c\mu = q_{\mathrm{LS}}(1^-) \geq c$ by part (a) of Lemma 6.10. Finally (6.59) holds as $\rho_{\mathrm{LS}} < 1$ implies $\lambda - c\mu - c + ce^{-\mu} = -\gamma_{\mathrm{LS}} < 0$. Thus we obtain that $\mathbf{f}_{\mathrm{LS}} \colon T_{\mathrm{LS}} \to T_{\mathrm{LS}}$.

To show $\mathbf{f}_{\mathrm{DA}} \colon T_{\mathrm{DA}} \to T_{\mathrm{DA}}$, we first compute that

$$\mathbf{h}_1(\tilde{q}_{\mathrm{DA}}, 0) = \Gamma\left(\tilde{q}_{\mathrm{DA}} + \frac{\lambda_1}{2} - c\mu, 0; 2c\right) \tag{6.60}$$

$$= \left(\tilde{q}_{\mathrm{DA}} + \frac{\lambda_1}{2} - c\mu - 2c, 2c\right). \tag{6.61}$$

Here (6.60) follows from part (b) of Lemma 6.9 on g_{DA}, and (6.61) follows from part (a) of the lemma. Thus for all $(q, 0) \in T_{\mathrm{DA}}$,

$$\mathbf{f}_{\mathrm{DA}}(q, 0) \leq \mathbf{f}_{\mathrm{DA}}(\tilde{q}_{\mathrm{DA}}, 0) \tag{6.62}$$

$$= \mathbf{h}_2\left(\tilde{q} + \frac{\lambda_1}{2} - c\mu - 2c, 2c\right)$$

$$= \Gamma\left(\tilde{q} + \frac{\lambda_1}{2} - c\mu - 2c + \frac{\lambda_2}{2}, 2ce^{-\mu/2}; 0\right)$$

$$= \left(\tilde{q} + \lambda - c\mu - 2c + 2ce^{-\mu/2}, 0\right)$$

$$\leq (\tilde{q}_{\mathrm{DA}}, 0). \tag{6.63}$$

where (6.62) holds by the monotonicity of $\mathbf{f}_{\mathrm{DA}}$ and (6.63) holds as $\rho_{\mathrm{DA}} < 1$. Again the inequalities are componentwise. Thus we obtain that $\mathbf{f}_{\mathrm{DA}} \colon T_{\mathrm{DA}} \to T_{\mathrm{DA}}$.

We show that $\mathbf{f}_{\mathrm{LS}}$ is contractive on T_{LS} with respect to the $\|\cdot\|_1$ norm. Consider $(q, r), (q', r') \in T_{\mathrm{LS}}$, such that $(q, r) \neq (q', r')$. If $q \neq q'$, then by part (c') of Lemma 6.10

$$|g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)) - g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q'))| < |q - q'|. \tag{6.64}$$

Similarly, when $r \neq r'$, then

$$|re^{-\mu} - r'e^{-\mu}| < |r - r'|. \tag{6.65}$$

Thus we obtain that

$$\begin{aligned}
\|\mathbf{f}_{\mathrm{LS}}(q,r) - \mathbf{f}_{\mathrm{LS}}(q',r')\|_1 &= \|\Gamma(g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)), re^{-\mu}; c) - \Gamma(g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q')), r'e^{-\mu}; c)\|_1 \\
&\leq |g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)) - g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q'))| + |re^{-\mu} - r'e^{-\mu}| \qquad (6.66) \\
&< |q - q'| + |r - r'| \qquad (6.67) \\
&= \|(q,r) - (q',r')\|_1.
\end{aligned}$$

Here (6.66) holds by Lemma 6.4 and (6.67) follows from (6.64) if $q \neq q'$ and from (6.65) if $r \neq r'$.

Showing that $\mathbf{f}_{\mathrm{DA}}$ is contractive on T_{DA} with respect to the $\|\cdot\|_1$ norm is very similar to the LS case. Briefly, we observe that when $q < \tilde{q}_{\mathrm{DA}}$, that $\mathbf{h}^1_{\mathrm{DA}}$ is strictly contractive by (c') of Lemma 6.9. It is easy to see that $\mathbf{h}^2_{\mathrm{DA}}$ is non-expansive for all $(q,r) \in \mathbb{R}^2_+$. Thus $\mathbf{f}_{\mathrm{DA}}$ on T_{DA} is the composition of a contractive function and a non-expansive function and hence contractive.

Thus the assumptions of Proposition 6.3 are satisfied by $\mathbf{f}_\theta$ on T_θ when $\rho_\theta < 1$. This implies that once $\mathbf{s}_\theta(k)$ enters T_θ it will converge to a unique fixed point. For the case $\rho_\theta < 1$, it remains to show that for $\mathbf{s}_\theta(0) \notin T_\theta$, we reach T_θ in finite time.

To this end, we apply Proposition 6.4 using the Lyapunov function $V(q,r) \stackrel{\Delta}{=} q+r$, the set of exceptions as T_θ, and γ to be γ_θ (we have $\gamma_\theta > 0$ as $\rho_\theta < 1$). We now show that the drift condition is satisfied. For LS, $(q,r) \notin T_{\mathrm{LS}}$ implies $q \geq \tilde{q}_{\mathrm{LS}} \geq c$, thus we must have $r = c$. Thus,

$$\begin{aligned}
V(\mathbf{f}_{\mathrm{LS}}(q,c)) - V(q,c) &= V(\Gamma(g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)), ce^{-\mu}; c)) - (q+c) \\
&= g^2_{\mathrm{LS}}(g^1_{\mathrm{LS}}(q)) + ce^{-\mu} - (q+c) \qquad (6.68) \\
&= q + \lambda - c\mu + ce^{-\mu} - (q+c) \qquad (6.69) \\
&= -\gamma_{\mathrm{LS}}.
\end{aligned}$$

Here (6.68) follows from the same argument justifying (6.26), and (6.69) follows from

part (b) of Lemma 6.10. For DA, $(q,0) \notin T_{\mathrm{DA}}$ implies $q \geq \tilde{q}_{\mathrm{DA}}$. We obtain

$$\mathbf{h}_1(q,0) = \left(q + \frac{\lambda_1}{2} - c\mu - 2c, 2c\right),$$

just as we justified (6.60) and (6.61). Thus for such q,

$$\begin{aligned} V(\mathbf{f}_{\mathrm{DA}}(q,0)) - V(q,0) &= V\left(\mathbf{h}_2\left(q + \frac{\lambda_1}{2} - c\mu - 2c, 2c\right)\right) - q \\ &= V(\Gamma(q + \lambda - c\mu - 2c, 2ce^{-\mu/2}; 0)) - q \\ &= -\gamma_{\mathrm{DA}}. \end{aligned}$$

Thus the assumptions of Proposition 6.4 are satisfied, establishing the claim in the case when $\rho_\theta < 1$.

Finally, we show that $\{\mathbf{s}_\theta(k)\}$ has no attractive point when $\rho_\theta \geq 1$ by applying Proposition 6.5. We again take $V(q,r) = q + r$, and immediately verify that it is continuous and unbounded on $\mathcal{T}_\theta$. For LS, we compute that for any $\mathbf{s}_{\mathrm{LS}}(0)$,

$$\begin{aligned} V(\mathbf{s}_{\mathrm{LS}}(1)) - V(\mathbf{s}_{\mathrm{LS}}(0)) &= q_{\mathrm{LS}}(1^-) - q_{\mathrm{LS}}(0) + r_{\mathrm{LS}}(1^-) - r_{\mathrm{LS}}(0) && (6.70) \\ &= \int_0^1 \lambda(t) - \mu(q_{\mathrm{LS}}(t) \wedge c)dt - r_{\mathrm{DA}}(0)(1 - e^{-\mu}) && (6.71) \\ &\geq \lambda - c\mu - c(1 - e^{-\mu}) && (6.72) \\ &= -\gamma_{\mathrm{LS}} \\ &\geq 0, && (6.73) \end{aligned}$$

where (6.70) follows similarly to (6.26), (6.71) follows from the definition of $\dot{q}_{\mathrm{LS}}(t)$, (6.72) follows as $r_{\mathrm{LS}}(0) \leq c$ and $q_{\mathrm{LS}}(t) \wedge c \leq c$, and finally (6.73) follows as $\rho_{\mathrm{LS}} \leq 1$.

For DA, we compute that for any $\mathbf{s}_{\mathrm{DA}}(0)$ that

$$
\begin{aligned}
V(\mathbf{s}_{\mathrm{DA}}(1)) - V(\mathbf{s}_{\mathrm{DA}}(0)) &= q_{\mathrm{DA}}(1^-) - q_{\mathrm{DA}}(0) + r_{\mathrm{DA}}(1^-) && (6.74)\\
&= \frac{\lambda_2}{2} + q_{\mathrm{DA}}(\tfrac{1}{2}) - q_{\mathrm{DA}}(0) + r_{\mathrm{DA}}(\tfrac{1}{2})e^{-\mu/2} \\
&= \frac{\lambda_2}{2} + q_{\mathrm{DA}}(\tfrac{1}{2}) + r_{\mathrm{DA}}(\tfrac{1}{2}) - q_{\mathrm{DA}}(0) - r_{\mathrm{DA}}(\tfrac{1}{2})(1 - e^{-\mu/2}) \\
&\geq \frac{\lambda_2}{2} + q_{\mathrm{DA}}(\tfrac{1}{2}) + r_{\mathrm{DA}}(\tfrac{1}{2}) - q_{\mathrm{DA}}(0) - 2c(1 - e^{-\mu/2}) && (6.75)\\
&= \frac{\lambda_2}{2} + q_{\mathrm{DA}}\left(\tfrac{1}{2}^-\right) - q_{\mathrm{DA}}(0) - 2c(1 - e^{-\mu/2}) && (6.76)\\
&= \frac{\lambda_2}{2} + \int_0^{\frac{1}{2}} \lambda_1 - \mu(q_{\mathrm{DA}}(t) \wedge 2c)dt - 2c(1 - e^{-\mu/2}) && (6.77)\\
&\geq \lambda - \int_0^{\frac{1}{2}} 2c\mu \, dt - 2c(1 - e^{-\mu/2}) \\
&= -\gamma_{\mathrm{DA}} \\
&\geq 0, && (6.78)
\end{aligned}
$$

where (6.74) follows similarly to (6.26), (6.75) follows as $r_{\mathrm{DA}}(\frac{1}{2}) \leq 2c$, (6.76) follows as by Γ, $q_{\mathrm{DA}}(\frac{1}{2}^-) + r_{\mathrm{DA}}(\frac{1}{2}^{-1}) = q_{\mathrm{DA}}(\frac{1}{2})$, (6.77) follows from the definition of $\dot{q}_{\mathrm{DA}}(t)$, and finally (6.78) holds as $\rho_{\mathrm{DA}} \geq 1$. Thus we see by Proposition 6.5 that $\rho_\theta \geq 1$ implies that $\{\mathbf{s}_\theta(k)\}$ has no attractive point, completing the proof of Proposition 6.6. □

Finally, we give a result providing a uniform bound on the distance moved towards the fixed point in each iteration of the fluid model. This result will be useful in the proof of Theorem 6.3. Let $V_\theta \colon \mathcal{T}_\theta \to \mathbb{R}_+$ be given by

$$V_\theta(\mathbf{s}) \triangleq \|\mathbf{s} - \mathbf{s}_\theta(\infty)\|_1. \tag{6.79}$$

Corollary 6.3. *For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, when $\rho_\theta < 1$, for every $z > 0$, there exists $\gamma > 0$ such that*

$$\inf_{\mathbf{s} \in \mathcal{T}_\theta \setminus B_z(\mathbf{s}_\theta(\infty))} V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}) \leq -\gamma.$$

Proof. Recall from in the proof of Proposition 6.6 that for each θ, we defined sets T_θ

such that that $\mathbf{f}_\theta$ is contractive on T_θ and for all $\mathbf{s} \in \mathcal{T}_\theta \setminus T_\theta$,

$$V(\mathbf{f}_\theta(\mathbf{s})) - V(\mathbf{s}) = -\gamma_\theta < 0.$$

Suppose $\mathbf{s} = (q, r) \notin T_\theta$. Trivially $\mathbf{s} \geq \mathbf{s}_\theta(\infty)$ componentwise as $\mathbf{s}_\theta(\infty) \in T_\theta$. By Lemma 6.11, as $\mathbf{s} \geq \mathbf{s}_\theta(\infty)$ componentwise, we have $\mathbf{f}_\theta(\mathbf{s}) \geq \mathbf{f}_\theta(\mathbf{s}_\theta(\infty)) = \mathbf{s}_\theta(\infty)$ componentwise as well. Thus for $\mathbf{s} \notin T_\theta$, letting $(q', r') = \mathbf{f}_\theta(\mathbf{s})$,

$$\begin{aligned} V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}) &= \|\mathbf{f}_\theta(\mathbf{s}) - \mathbf{s}_\theta(\infty)\|_1 - \|\mathbf{s} - \mathbf{s}_\theta(\infty)\|_1 \\ &= q' - q_\theta(\infty) + r' - r_\theta(\infty) - (q - q_\theta(\infty) + r - r_\theta(\infty)) \\ &= V(\mathbf{f}_\theta(\mathbf{s})) - V(\mathbf{s}) \\ &\leq -\gamma_\theta. \end{aligned}$$

For all $\mathbf{s} \in T_\theta$, $\mathbf{s} \neq \mathbf{s}_\theta(\infty)$, as $\mathbf{f}_\theta$ is contractive on T_θ, we have

$$\begin{aligned} V_\theta(\mathbf{f}_\theta(\mathbf{s})) &= \|\mathbf{f}_\theta(\mathbf{s}) - \mathbf{s}_\theta(\infty)\|_1 \\ &= \|\mathbf{f}_\theta(\mathbf{s}) - \mathbf{f}_\theta(\mathbf{s}_\theta(\infty))\|_1 \\ &< \|\mathbf{s} - \mathbf{s}_\theta(\infty)\|_1 \\ &= V_\theta(\mathbf{s}). \end{aligned}$$

Thus for $\mathbf{s} \in T_\theta$, $V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}) \leq 0$, holding with equality only when $\mathbf{s} = \mathbf{s}_\theta(\infty)$. Fix z from the statement of the Lemma. As $V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s})$ is a composition of continuous functions and thus continuous and as T_θ is compact (as shown in the proof of Proposition 6.6), we have that given our z there exists $\varepsilon > 0$ such that

$$\inf_{\mathbf{s} \in T_\theta \setminus B_z(\mathbf{s}_\theta(\infty))} V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}) \leq -\varepsilon.$$

Thus by taking $\gamma = \min\{\varepsilon, \gamma_\theta\}$, we obtain the result. □

6.9 Interchange of Limits. Proof of Theorem 6.3

In this section, we prove Theorem 6.3, showing that the rescaled steady state distributions $\mathbf{S}_\theta^n(\infty)/n$ converge in probability to the fixed point $\mathbf{s}_\theta(\infty)$ of the fluid limit at integer times.

Recall that a set of random vectors $\{\mathbf{X}_n\}$ is defined to be *tight* if for every ε there exists k such that for every n, $\mathbb{P}(\|\mathbf{X}_n\|_1 > k) \leq \varepsilon$. As a direct consequence of Lemma 6.8 and Markov's inequality, we obtain:

Corollary 6.4. *For each policy $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, when $\rho_\theta < 1$, the set of random vectors $\{\mathbf{S}_\theta^n(\infty)/n\}$ is tight.*

By Prokhorov's theorem, this implies that $\{\mathbf{X}_n\}$ is relatively compact. That is, for every subsequence $\mathbf{X}_{n_i}$ there exists a random vector $\mathbf{X}$ and a subsubsequence $\mathbf{X}_{n_{i_j}}$ such that $\mathbf{X}_{n_{i_j}} \Rightarrow \mathbf{X}$ (see [15]). Thus for every subsequence n_i there is a subsubsequence n_{i_j} and a random vector $\bar{\mathbf{S}}_\theta$ such that as $j \to \infty$

$$\frac{\mathbf{S}_\theta^{n_{i_j}}(\infty)}{n_{i_j}} \Rightarrow \bar{\mathbf{S}}_\theta.$$

Thus to show Theorem 6.3, it is sufficient to show that for every sequence n_i, the resulting $\bar{\mathbf{S}}_\theta$ equals $\mathbf{s}_\theta(\infty)$ with probability one, as convergence in distribution to a constant implies convergence in probability.

Proof of Theorem 6.3. First, we claim that for $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$,

$$\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta) \stackrel{d}{=} \bar{\mathbf{S}}_\theta, \tag{6.80}$$

where $\mathbf{f}_{\mathrm{LS}}$ and $\mathbf{f}_{\mathrm{DA}}$ are defined by (6.51) and (6.54), respectively. By Proposition 11.3.2 from [28], we can equivalently check that the Lévy-Prokhorov distance between these variables is zero, i.e. that for all $g\colon \mathcal{T}_\theta \to \mathbb{R}$ such that $\|g\|_{\mathrm{BL}} \leq 1$, we have

$$\mathbb{E}[g(\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta)) - g(\bar{\mathbf{S}}_\theta)] = 0.$$

Here, we use the three term estimate as devised by [31], Chapter 4, Theorem 9.10, in a similar continuous time interchange of limits argument. See [79] for similar but less terse argument. Assume for every n that $\mathbf{S}_\theta^n(0) \stackrel{d}{=} \mathbf{S}_\theta^n(\infty)$. Now for every n, we have

$$\begin{aligned}\left|\mathbb{E}[g(\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta)) - g(\bar{\mathbf{S}}_\theta)]\right| \leq & \left|\mathbb{E}\left[g(\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta)) - g\left(\mathbf{f}_\theta\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right)\right)\right]\right| \\ & + \left|\mathbb{E}\left[g\left(\mathbf{f}_\theta\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right)\right) - g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right)\right]\right| \\ & + \left|\mathbb{E}\left[g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right) - g\left(\bar{\mathbf{S}}_\theta\right)\right]\right|.\end{aligned}$$

As $\mathbf{f}_\theta$ and g are continuous and g is bounded, $g \circ \mathbf{f}_\theta$ is a bounded continuous function. For the first term, $\mathbf{S}_\theta^n(0)/n \stackrel{d}{=} \mathbf{S}_\theta^n(\infty)/n \Rightarrow \bar{\mathbf{S}}_\theta$, so we can apply the Continuous Mapping Theorem and then the Bounded Convergence Theorem to obtain that $\mathbb{E}[g\left(\mathbf{f}_\theta\left(\mathbf{S}_\theta^n(0)/n\right)\right)] \to \mathbb{E}[g(\mathbf{f}_\theta(\bar{\mathbf{S}}_\theta))]$ as $n \to \infty$ along n_{i_j}. By stationarity $\mathbf{S}_\theta^n(1)/n \stackrel{d}{=} \mathbf{S}_\theta^n(0)/n \stackrel{d}{=} \mathbf{S}_\theta^n(\infty)/n$, implying that the third term converges to zero along n_{i_j} by a similar argument.

Finally we bound the second term. Let $h_\theta^n \colon \mathcal{T}_\theta \to \mathbb{R}$ and $h_\theta \colon \mathcal{T}_\theta \to \mathbb{R}$ be given by

$$\begin{aligned}h_\theta^n(\mathbf{s}) &\stackrel{\Delta}{=} \mathbb{E}_{\lfloor n\mathbf{s} \rfloor}\left[g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right)\right], \\ h_\theta(\mathbf{s}) &\stackrel{\Delta}{=} g(\mathbf{f}_\theta(\mathbf{s})),\end{aligned}$$

so that

$$\left|\mathbb{E}\left[g\left(\mathbf{f}_\theta\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right)\right) - g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right)\right]\right| = \left|\mathbb{E}\left[h_\theta\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right) - h_\theta^n\left(\frac{\mathbf{S}_\theta^n(0)}{n}\right)\right]\right|. \quad (6.81)$$

We need some properties of h_θ^n and h to make an estimate. First, we claim that for all $\mathbf{s}$, $h_\theta^n(\mathbf{s}) \to h_\theta(\mathbf{s})$ as $n \to \infty$. As a consequence of Theorem 6.2, we have that for every $\mathbf{s} \in \mathcal{T}_\theta$, if $\mathbf{S}_\theta^n(0) = \lfloor n\mathbf{s} \rfloor$ so that $\mathbf{S}_\theta^n(0)/n \to \mathbf{s}$ a.s., then $\mathbf{S}_\theta^n(1)/n \to \mathbf{f}_\theta(\mathbf{s})$ a.s. as well. By the continuity of g, it follows from that $g(\mathbf{S}_\theta^n(1)/n) \to g(\mathbf{f}_\theta(\mathbf{s}))$ a.s. as well. Noting that $g(\mathbf{S}_\theta^n(1)/n)$ is bounded, we can apply the Bounded Convergence

Theorem to obtain that for $\mathbf{s} \in \mathcal{T}_\theta$,

$$\lim_{n\to\infty} h_\theta^n(\mathbf{s}) = h_\theta(\mathbf{s}). \tag{6.82}$$

We can now bound (6.81) with a coupling argument. By the Skorokhod Representation Theorem, let Ω be a common probability space for $\{\mathbf{S}_\theta^n(0)\}$ and $\bar{\mathbf{S}}_\theta$ such that for $\omega \in \Omega$, $\mathbf{S}_\theta^n(0,\omega) \to \bar{\mathbf{S}}_\theta(\omega)$ a.s. Now we have

$$\begin{aligned}\left|h_\theta\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right) - h_\theta^n\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right)\right| \leq \quad & \left|h_\theta\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right) - h_\theta\left(\bar{\mathbf{S}}_\theta(\omega)\right)\right| \\ & + \left|h_\theta\left(\bar{\mathbf{S}}_\theta(\omega)\right) - h_\theta^n\left(\bar{\mathbf{S}}_\theta(\omega)\right)\right| \\ & + \left|h_\theta^n\left(\bar{\mathbf{S}}_\theta(\omega)\right) - h_\theta^n\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right)\right|.\end{aligned}$$

We claim each of these terms converges to zero a.s. The first term converges to zero as h_θ is a continuous function and $\mathbf{S}_\theta^n(0,\omega)/n \to \bar{\mathbf{S}}_\theta(\omega)$ a.s. The second term converges to zero by (6.82). Let $\tilde{\mathbf{S}}_\theta^n(t)$ be another version of the process $\mathbf{S}_\theta^n(t)$ with the initial condition $\tilde{\mathbf{S}}_\theta^n(0) = \lfloor n\bar{\mathbf{S}}_\theta \rfloor$ that is coupled to $\mathbf{S}_\theta^n(t)$ as in Corollary 6.2. Then

$$\begin{aligned}&\left|h_\theta^n\left(\bar{\mathbf{S}}_\theta(\omega)\right) - h_\theta^n\left(\frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right)\right| \\ &\quad = \left|\mathbb{E}\left[g\left(\frac{\tilde{\mathbf{S}}_\theta^n(1)}{n}\right) - g\left(\frac{\mathbf{S}_\theta^n(1)}{n}\right) \,\middle|\, \tilde{\mathbf{S}}_\theta^n(0) = \lfloor n\bar{\mathbf{S}}_\theta(\omega)\rfloor,\, \mathbf{S}_\theta^n(0) = \mathbf{S}_\theta^n(0,\omega)\right]\right| \\ &\quad \leq \frac{\|g\|_{\mathrm{BL}}}{n}\mathbb{E}\left[\left\|\tilde{\mathbf{S}}_\theta^n(1) - \mathbf{S}_\theta^n(1)\right\|_1 \,\middle|\, \tilde{\mathbf{S}}_\theta^n(0) = \lfloor n\bar{\mathbf{S}}_\theta(\omega)\rfloor,\, \mathbf{S}_\theta^n(0) = \mathbf{S}_\theta^n(0,\omega)\right] \\ &\quad \leq \frac{\|g\|_{\mathrm{BL}}}{n}\left\|\lfloor n\bar{\mathbf{S}}_\theta(\omega)\rfloor - \mathbf{S}_\theta^n(0,\omega)\right\|_1 \qquad (6.83) \\ &\quad \leq \|g\|_{\mathrm{BL}}\left(\left\|\bar{\mathbf{S}}_\theta(\omega) - \frac{\mathbf{S}_\theta^n(0,\omega)}{n}\right\|_1 + \frac{1}{n}\right),\end{aligned}$$

showing that the third term converges to zero a.s. as well. Here (6.83) follows from Corollary 6.2. Finally, as h_θ^n and h_θ are bounded by one, the Bounded Convergence Theorem implies that the right hand side of (6.81) converges to zero. Thus we have shown (6.80).

Now we show that $\bar{\mathbf{S}}_\theta = \mathbf{s}_\theta(\infty)$ a.s. We assume the conclusion is false to show

a contradiction. By assumption, there exists some $\tilde{\mathbf{s}}_\theta \neq \mathbf{s}_\theta(\infty)$ and $\varepsilon > 0$ such that $\mathbf{s}_\theta(\infty) \notin B_\varepsilon(\tilde{\mathbf{s}}_\theta)$ and

$$\mathbb{P}(\bar{\mathbf{S}}_\theta \in B_\varepsilon(\tilde{\mathbf{s}}_\theta)) > 0.$$

Let N be such that

$$\mathbb{P}(\|\bar{\mathbf{S}}_\theta\|_1 > N) < \mathbb{P}(\bar{\mathbf{S}}_\theta \in B_\varepsilon(\tilde{\mathbf{s}}_\theta)). \tag{6.84}$$

Let $z = \|\tilde{\mathbf{s}} - \mathbf{s}_\theta(\infty)\|_1 - \varepsilon$ and let $Z = B_z(\mathbf{s}_\theta(\infty))$ be the largest ball around $\mathbf{s}_\theta(\infty)$ disjoint from $B_\varepsilon(\tilde{\mathbf{s}}_\theta)$. We now use V_θ from (6.79), and let

$$-d \triangleq \inf_{\mathbf{s} \notin Z} V_\theta(\mathbf{f}_\theta(\mathbf{s})) - V_\theta(\mathbf{s}).$$

Note that $d > 0$ by Corollary 6.3. Let $n \in \mathbb{Z}_+$ be such that

$$nd > \sup_{\|\mathbf{s}\|_1 < N} V_\theta(\mathbf{s}),$$

(the supremum is bounded as $\{\mathbf{s} \in \mathcal{T}_\theta | \|\mathbf{s}\|_1 < N\}$ is compact and V_θ is continuous). Let $\mathbf{f}_\theta^{(m)}$ be the function $\mathbf{f}_\theta$ composed with itself m times. We now claim that for all $\mathbf{s} \in \mathcal{T}_\theta$,

$$\mathbf{f}^{(n)}(\mathbf{s}) \notin Z \text{ implies that } \|\mathbf{s}\|_1 > N. \tag{6.85}$$

We show the contrapositive using Proposition 6.4. We take our bounded set of exceptions as Z, use the Lyapunov function V_θ, and drift $-d$. The drift condition is satisfied as $d > 0$. Thus $\|\mathbf{s}\|_1 < N$ implies that there exists m with $0 \leq m \leq n$ such that $\mathbf{f}_\theta^{(m)}(\mathbf{s}) \in Z$. Notice that Corollary 6.3 implies that $\mathbf{f}_\theta(Z) \subset Z$. Thus $\|\mathbf{s}\|_1 < N$ in fact implies that $\mathbf{f}_\theta^{(n)}(\mathbf{s}) \in Z$, showing (6.85).

Thus we have the inequalities

$$\mathbb{P}(\|\bar{\mathbf{S}}_\theta\|_1 > N) < \mathbb{P}(\bar{\mathbf{S}}_\theta \in B_\varepsilon(\tilde{\mathbf{s}}_\theta)) \tag{6.86}$$

$$= \mathbb{P}(\mathbf{f}_\theta^{(n)}(\bar{\mathbf{S}}_\theta) \in B_\varepsilon(\tilde{\mathbf{s}}_\theta)) \tag{6.87}$$

$$\leq \mathbb{P}(\mathbf{f}_\theta^{(n)}(\bar{\mathbf{S}}_\theta) \notin Z) \tag{6.88}$$

$$\leq \mathbb{P}(\|\bar{\mathbf{S}}_\theta\|_1 > N), \tag{6.89}$$

where (6.86) follows from (6.84), (6.87) follows from (6.80), (6.88) follows as Z and $B_\varepsilon(\tilde{\mathbf{s}}_\theta)$ are disjoint, and (6.89) follows from (6.85). Thus we have obtained a contradiction, which shows that $\bar{\mathbf{S}}_\theta$ equals $\mathbf{s}_\theta(\infty)$ with probability one. This completes the proof. □

6.10 Convergence of Reassignments in the Fluid Limit

Before proving Corollary 6.1, we need a simple monotonicity result for the fluid limit $q_\theta(t)$.

Lemma 6.12. *Under the assumptions of Corollary 6.1, for $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, $q_\theta(t)$ is monotone on $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$. Further, for any $t^* \in [0,1]$ such that $q_{\mathrm{LS}}(t^*) = c$, (resp. $t^* \in [0, \frac{1}{2})$ such that $q_{\mathrm{DA}}(t^*) = 2c$), $q_\theta(t)$ is strictly monotone in a neighborhood of t^*. Consequently, there exists ε_0 such that for every $\varepsilon < \varepsilon_0$ there exists δ such that $|q_\theta(t) - q_\theta(t^*)| < \delta$ implies $|t - t^*| < \varepsilon$.*

Proof. The monotonicity on $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$ follows as on each such interval, $q_\theta(t)$ is the solution to the differential equation of the type from Lemma 6.9. For LS, for any t^* such that $q_{\mathrm{LS}}(t^*) = c$, we have $\dot{q}_{\mathrm{LS}}(t^*) = \lambda(t^*) - c\mu$. As we have assumed that $\lambda_1, \lambda_2 \neq c\mu$, we have that $\dot{q}_{\mathrm{LS}}(t^*) \neq 0$. Noting that $q_{\mathrm{LS}}(t)$ is continuously differentiable, it follows that $q_{\mathrm{LS}}(t)$ is strictly monotone. A similar argument applies for DA. □

Finally, we show the convergence of the number reassignments (the number of

patients forced to wait per day), completing the commutative diagram in Figure 6-1.

Proof of Corollary 6.1. We first show that $W_{\mathrm{LS}}^{1,n}(\infty)/n \to w_{\mathrm{LS}}^1(\infty)$ in probability. Noting that the limit is a constant, it sufficient to show convergence in distribution. By Proposition 11.3.3 from [28], $W_{\mathrm{LS}}^{1,n}(\infty)/n \Rightarrow w_{\mathrm{LS}}^1(\infty)$ iff for all $g\colon \mathbb{R}_+ \to \mathbb{R}$ such that $\|g\|_{\mathrm{BL}} \leq 1$, we have

$$\lim_{n\to\infty} \mathbb{E}\left[g\left(\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}\right)\right] = g(w_{\mathrm{LS}}^1(\infty)).$$

For each n we take each $\mathbf{S}_{\mathrm{LS}}^n(0) \overset{d}{=} \mathbf{S}_{\mathrm{LS}}^n(\infty)$. Using Theorem 6.3 and the Skorokhod Representation Theorem, we put the $\mathbf{S}_{\mathrm{LS}}^n(0)$ on a common probability space such that $\mathbf{S}_{\mathrm{LS}}^n(0)/n \to \mathbf{s}_{\mathrm{LS}}(\infty)$ a.s. We use this process to generate the $W_{\mathrm{LS}}^{1,n}(\infty)$ and $w_{\mathrm{LS}}^1(\infty)$ all on a common probability space.

It follows from Lemma 6.12 that there can be at most one time $t^* \in [0, \frac{1}{2})$ such that $q_{\mathrm{LS}}(t^*) = c$, and that $q_{\mathrm{LS}}(t)$ must be strictly monotone at t^*. Let ε_0 be from the lemma and fix some $\varepsilon \in (0, \varepsilon_0)$. Let δ be from Lemma 6.12 such that $|q_{\mathrm{LS}}(t) - q_{\mathrm{LS}}(t^*)| < \delta$ implies $|t - t^*| < \varepsilon$. Recall from Theorem 6.2 that when $\mathbf{S}_{\mathrm{LS}}^n(0)/n \to \mathbf{s}_{\mathrm{LS}}(0)$ a.s., then $\sup_{0\leq t\leq \frac{1}{2}} \|\mathbf{S}_{\mathrm{LS}}^n(t)/n - \mathbf{s}_{\mathrm{LS}}(t)\|_1 \to 0$ a.s. As $|Q_{\mathrm{LS}}^n(t)/n - q_{\mathrm{LS}}(t)| \leq \|\mathbf{S}_{\mathrm{LS}}^n(t)/n - \mathbf{s}_{\mathrm{LS}}(t)\|_1$, we also have $\sup_{0\leq t\leq \frac{1}{2}} |Q_{\mathrm{LS}}^n(t)/n - q_{\mathrm{LS}}(t)| \to 0$ a.s. As a result, we also have convergence in probability. In particular, for our δ,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{0\leq t\leq \frac{1}{2}} \left|\frac{Q^n(t)}{n} - q(t)\right| > \delta\right) = 0.$$

Let E_δ^n be the event

$$E_\delta^n \triangleq \left\{\sup_{0\leq t\leq \frac{1}{2}} \left|\frac{Q^n(t)}{n} - q(t)\right| > \frac{\delta}{2}\right\},$$

i.e. $\lim_{n\to\infty} \mathbb{P}(E_\delta^n) = 0$ for all δ. Let $\bar{E}_\delta^n$ denote the complement of this event. We

have that

$$\begin{aligned}
&\left|\mathbb{E}\left[g\left(\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}\right)\right] - g(w_{\mathrm{LS}}^{1}(\infty))\right| \\
&\quad\leq \quad \left|\mathbb{E}\left[g\left(\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}\right) - g(w_{\mathrm{LS}}^{1}(\infty)) \;\middle|\; E_\delta^n\right]\right|\mathbb{P}(E_\delta^n) \\
&\qquad + \left|\mathbb{E}\left[g\left(\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n}\right) - g(w_{\mathrm{LS}}^{1}(\infty)) \;\middle|\; \bar{E}_\delta^n\right]\right|\mathbb{P}(\bar{E}_\delta^n) \\
&\quad\leq 2\|g\|_{\mathrm{BL}}\mathbb{P}(E_\delta^n) + \|g\|_{\mathrm{BL}}\mathbb{E}\left[\left|\frac{W_{\mathrm{LS}}^{1,n}(\infty)}{n} - w_{\mathrm{LS}}^{1}(\infty)\right| \;\middle|\; \bar{E}_\delta^n\right], \qquad (6.90)
\end{aligned}$$

where in (6.90), we are using both that g is bounded by $\|g\|_{\mathrm{BL}}$ and has Lipschitz constant at most $\|g\|_{\mathrm{BL}}$. Letting $n \to \infty$, we see the first term go to zero. For the second term, we consider two cases:

1. Suppose that $\inf_{t\in[0,\frac{1}{2})} |q_{\mathrm{LS}}(t) - c| = \gamma > 0$. We can assume without loss of generality that $\delta < \gamma$, as we can always take δ smaller without interfering in the convergence of our first term, and doing so will only increase the proposed infimum. As $\delta < \gamma$, we ensure that conditional on $\bar{E}_\delta^n$, for every time $t \in [0, \frac{1}{2})$ that

$$\begin{aligned}
\left|\frac{Q_{\mathrm{LS}}^n(t)}{n} - c\right| &\geq \left||q_{\mathrm{LS}}(t) - c| - \left|\frac{Q_{\mathrm{LS}}^n(t)}{n} - q_{\mathrm{LS}}(t)\right|\right| \qquad (6.91) \\
&= |q_{\mathrm{LS}}(t) - c| - \left|\frac{Q_{\mathrm{LS}}^n(t)}{n} - q_{\mathrm{LS}}(t)\right| \qquad (6.92) \\
&\geq \gamma/2,
\end{aligned}$$

   where (6.91) is the reverse triangle inequality and (6.92) follows by our choice of δ. We have two further cases:

   (a) If $q_{\mathrm{LS}}(t) < c$ and thus $Q_{\mathrm{LS}}(t)/n < c$ for all t, then both $W_{\mathrm{LS}}^{1,n}(\infty)$ and $w_{\mathrm{LS}}^{1}(\infty)$ will be zero, so we will have that (6.90) converges to zero as $n \to \infty$.

   (b) Similarly, if $q_{\mathrm{LS}}(t) > c$ and thus $Q_{\mathrm{LS}}(t)/n > c$ for all t, then $W_{\mathrm{LS}}^{1,n}(\infty) =$

$A^n(0,\frac{1}{2})$ and $w^1_{\mathrm{LS}}(\infty) = \lambda_1/2$. Thus

$$\begin{aligned}\mathbb{E}\left[\left|\frac{W^{1,n}_{\mathrm{LS}}(\infty)}{n} - w^1_{\mathrm{LS}}(\infty)\right| \,\middle|\, \bar{E}^n_\delta\right] &= \frac{1}{\mathbb{P}\left(\bar{E}^n_\delta\right)}\mathbb{E}\left[\left|\frac{W^{1,n}_{\mathrm{LS}}(\infty)}{n} - w^1_{\mathrm{LS}}(\infty)\right| \mathbb{I}_{\bar{E}^n_\delta}\right] \\ &= \frac{1}{\mathbb{P}\left(\bar{E}^n_\delta\right)}\mathbb{E}\left[\left|\frac{A^n(0,\frac{1}{2})}{n} - \frac{\lambda_1}{2}\right| \mathbb{I}_{\bar{E}^n_\delta}\right] \\ &\leq \frac{1}{\mathbb{P}\left(\bar{E}^n_\delta\right)}\mathbb{E}\left[\left|\frac{A^n(0,\frac{1}{2})}{n} - \frac{\lambda_1}{2}\right|\right].\end{aligned}$$

The above converges to zero almost surely since $\mathbb{E}[A^n(0,\frac{1}{2})/n] \to \lambda_1/2$ as $n \to \infty$. Thus (6.90) converges to zero as $n \to \infty$.

2. Suppose instead that $q_{\mathrm{LS}}(t)$ crosses c. Suppose $\lambda_1 > c\mu$, so by Lemma 6.12 $q_{\mathrm{LS}}(t)$ is monotonically increasing. Let t^* be the time such that $q_{\mathrm{LS}}(t^*) = c$. Then

$$w^1_{\mathrm{LS}}(\infty) = \int_{t^*}^{\frac{1}{2}} \lambda_1 dt = (\tfrac{1}{2} - t^*)\lambda_1.$$

We claim that conditional on $\bar{E}^n_\delta$,

$$|t - t^*| \geq \varepsilon \text{ implies that } |Q^n_{\mathrm{LS}}(t)/n - c| \geq \delta/2. \tag{6.93}$$

We will show the contrapositive. We have that when $|Q^n_{\mathrm{LS}}(t)/n - c| < \delta/2$, then

$$|q_{\mathrm{LS}}(t) - c| \leq \left|q_{\mathrm{LS}}(t) - \frac{Q^n_{\mathrm{LS}}(t)}{n}\right| + \left|\frac{Q^n_{\mathrm{LS}}(t)}{n} - c\right| < \delta.$$

Now by Lemma 6.12, $|q_{\mathrm{LS}}(t) - c| < \delta$ implies $|t - t^*| < \varepsilon$, showing the claim. Next, we claim that then conditional on $\bar{E}^n_\delta$, for all $t > t^* + \varepsilon$, $Q^n_{\mathrm{LS}}(t)/n > c$.

Assume not for contradiction. Then

$$0 \leq c - \frac{Q^n_{\mathrm{LS}}(t)}{n} - \frac{\delta}{2} \tag{6.94}$$

$$\leq c - q_{\mathrm{LS}}(t) + \left| \frac{Q^n_{\mathrm{LS}}(t)}{n} - q_{\mathrm{LS}}(t) \right| - \frac{\delta}{2}$$

$$\leq c - q_{\mathrm{LS}}(t) \tag{6.95}$$

$$< 0, \tag{6.96}$$

giving a contradiction. Here (6.94) holds by (6.93) in conjunction with $Q^n_{\mathrm{LS}}(t)/n < c$, (6.95) holds as we are assuming $\bar{E}^n_\delta$, and finally (6.96) holds as $\lambda_1 > c\mu$ and Lemma 6.12 implies that $q_{\mathrm{LS}}(t)$ is increasing in t, $q_{\mathrm{LS}}(t^*) = c$, and $t > t^*$.
By an analogous argument, we can show that for all $t < t^* - \varepsilon$, $Q^n_{\mathrm{LS}}(t)/n < c$. As the number of reassignments $W^{1,n}_{\mathrm{LS}}(\infty)$ is the number of arrivals such that $Q^n_{\mathrm{LS}}(t) \geq cn$ at the time t of arrival, we thus have that all arrivals $A^n(t^* + \varepsilon, \frac{1}{2})$, will be reassignments, none of the arrivals $A^n(0, t^* - \varepsilon)$ will be reassignments, and the remaining arrivals are to be determined. This implies

$$A^n(t^* + \varepsilon, \tfrac{1}{2})\mathbb{I}_{\bar{E}^n_\delta} \leq W^{1,n}_{\mathrm{LS}}(\infty)\mathbb{I}_{\bar{E}^n_\delta} \leq A^n(t^* - \varepsilon, \tfrac{1}{2})\mathbb{I}_{\bar{E}^n_\delta}.$$

Thus

$$\begin{aligned}
\mathbb{E}\left[\left| \frac{W^{1,n}_{\mathrm{LS}}(\infty)}{n} - w^1_{\mathrm{LS}}(\infty) \right| \,\middle|\, \bar{E}^n_\delta \right] &\leq \mathbb{E}\left[\left| \frac{A^{1,n}_{\mathrm{LS}}(t^* + \varepsilon)}{n} - w^1_{\mathrm{LS}}(\infty) \right| \right. \\
&\qquad \left. + \left| \frac{A^{1,n}_{\mathrm{LS}}(t^* - \varepsilon)}{n} - w^1_{\mathrm{LS}}(\infty) \right| \,\middle|\, \bar{E}^n_\delta \right] \\
&\leq \mathbb{E}\left[\left| \frac{A^{1,n}_{\mathrm{LS}}(t^* + \varepsilon)}{n} - w^1_{\mathrm{LS}}(\infty) \right| \right. \\
&\qquad \left. + \left| \frac{A^{1,n}_{\mathrm{LS}}(t^* - \varepsilon)}{n} - w^1_{\mathrm{LS}}(\infty) \right| \right] \Big/ \mathbb{P}(\bar{E}^n_\delta) \\
&\leq 2\varepsilon\lambda_1 / \mathbb{P}(\bar{E}^n_\delta).
\end{aligned}$$

Letting $n \to \infty$, the above converges to $2\varepsilon\lambda_1$. As ε was arbitrary, the above

term and thus (6.90) must converge to zero as $n \to \infty$.

It is not hard to see that if instead $\lambda_1 < c\mu$, then as $q_{\text{LS}}(t)$ will be decreasing, we will obtain a similar bound of the form

$$A^n(0, t^* - \varepsilon)\mathbb{I}_{\bar{E}^n_\delta} \leq W^{1,n}_{\text{LS}}(\infty)\mathbb{I}_{\bar{E}^n_\delta} \leq A^n(0, t^* + \varepsilon)\mathbb{I}_{\bar{E}^n_\delta},$$

After taking expectations, we could again show the convergence of (6.90) to zero. Thus we conclude that $W^{1,n}_{\text{LS}}(\infty)/n \Rightarrow w^1_{\text{LS}}(\infty)$.

Showing that $W^{2,n}_{\text{LS}}(\infty) \Rightarrow w^2_{\text{LS}}(\infty)$, $W^{1,n}_{\text{DA}}(\infty)/n \Rightarrow w^1_{\text{DA}}(\infty)$, and $W^{2,n}_{\text{DA}}(\infty)/n \Rightarrow w^2_{\text{DA}}(\infty)$ is very similar and the details are omitted. □

6.11 Proof of Lemma 6.9

Here we give a series of lemmas about the differential equation

$$\dot{x}(t) = \gamma - \mu(x(t) \wedge m),$$

with $x(0) \in \mathbb{R}_+$, $\gamma, \mu, m > 0$, that will ultimately allow us to prove Lemma 6.9. For convenience, we let $g(x, t)$ equal $x(t)$ when $x(0) = x$ (making the function $g(x)$ defined in Lemma 6.9 equal to $g(x, \frac{1}{2})$).

Lemma 6.13. *The differential equation given by $\dot{x}(t)$ with initial condition $x(0) \in \mathbb{R}_+$ has a unique solution $x(t)$ with $x(t) \geq 0$ for all $t \in [0, \frac{1}{2}]$. Further, for every $x \in \mathbb{R}_+$, $g(x, t)$ is either strictly increasing in t for all t, strictly decreasing in t for all t, or equal to $g(x, 0)$ for all t.*

Proof. The differential equations

$$\dot{y}(t) = \gamma - m\mu,$$
$$\dot{z}(t) = \gamma - z(t)\mu,$$

with an initial condition $y(s) \in \mathbb{R}_+$ and $z(s) \in \mathbb{R}_+$ both have unique solutions for all

$t > 0$ given by

$$y(t) = y(s) + (t-s)(\gamma - m\mu),$$
$$z(t) = \frac{\gamma}{\mu} + \exp(-\mu(t-s))\left(z(s) - \frac{\gamma}{\mu}\right),$$

respectively. We claim that these two differential equations in combination determine the path of $x(t)$. Given our formula for $\dot{x}(t)$, we observe that $x(t) \geq m$ and $x(t) = y(t)$ implies $\dot{x}(t) = \dot{y}(t)$. Suppose that for some s we have $x(s) \geq m$ and let $y(s) = x(s)$. Then for all $t \geq s$ such that $y(t) \geq m$, we will have $x(t) = y(t)$. Analogously, we observe that $x(t) < m$ and $x(t) = z(t)$ implies that $\dot{x}(t) = \dot{z}(t)$. Suppose that for some s we have $x(s) < m$, and let $z(s) = x(s)$. Then for all $t \geq s$ such that $z(t) \leq m$, we will have $x(t) = z(t)$. Thus we can show that $x(t)$ has a unique solution on $[0, \frac{1}{2})$ for all initial conditions by showing that there is a clean exchange at the boundary $\{t \mid x(t) = m\}$. In particular, it suffices to show that we cross the boundary at most one time, which follows from the monotonicity claim in the second part of the Lemma.

First however, we need to analyze the long run behavior of $z(t)$. We can immediately see from $\dot{z}(t)$ that if $z(0) = \gamma/\mu$, then $\dot{z}(t) = 0$ so $z(t) = \gamma/\mu$ for all t. Similarly, when $z(t) < \gamma/\mu$, z will be strictly increasing at t, and when $z(t) > \gamma/\mu$, z will be strictly decreasing at t. Further, from the solution for $z(t)$, we see that if at any time s, $z(s) < \gamma/\mu$ then for all times $t > s$, we will still have $z(t) < \gamma/\mu$. Namely, z will approach γ/μ but never reach it. Likewise, when $z(s) > \gamma/\mu$, we will have $z(t) > \gamma/\mu$ for all $t > s$.

We can now finish the Lemma by considering three cases:

1. Suppose $\gamma > m\mu$, or equivalently $\gamma/\mu > m$. If $x(0) < m$, then as the attractive point of $z(t)$ is greater than m, we will have $x(t)$ strictly increasing until either time $\frac{1}{2}$ or s such that $z(s) = m$, if $s < \frac{1}{2}$. There is nothing left the prove in the first case, so we consider the second. Once $x(t) \geq m$, as $\gamma \geq m\mu$, we have $\dot{x}(t) = \dot{y}(t) = \gamma - m\mu > 0$, so $x(t)$ will increase strictly and never again fall before m. Thus in all cases, $x(t)$ is strictly increasing for all t.
2. Suppose $\gamma < m\mu$. Then if $x(t) \geq m$, we will have $\dot{x}(t) = \dot{y}(t) = \gamma - m\mu < 0$, so

$x(t)$ will be strictly decreasing. Again there are two possibilities, either there is a time $s < \frac{1}{2}$ such that $x(s) = m$, or $x(s)$ will not reach m before time $\frac{1}{2}$. As $x(t)$ is uniquely defined in the second case, we need only consider the first case further. Once we reach m, the dynamics of $x(t)$ will be that of $z(t)$. Recall that $z(t)$ will strictly decrease towards the fixed point γ/μ for all $t > s$ when $z(s) < \gamma/\mu$. Thus for all $x > \gamma/\mu$, $g(x,t)$ is strictly decreasing in t. When $x(0) = \gamma/\mu$, then by our previous analysis of $z(t)$ we have that $g(x(0),t) = \gamma/\mu$ for all t. Finally, when $x(0) < \gamma/\mu$, we know that $x(t)$ will be strictly increasing for all t towards γ/μ. Thus in all cases on $x(0)$ the criteria of the Lemma are met.

3. Suppose $\gamma = m\mu$. Then for all $x \geq m$, $\dot{x}(t) = \gamma - m\mu = 0$, so $g(x,t) = g(x,0)$. For all $x < m$, by our analysis of $z(t)$, we know that $x(t)$ will be strictly increasing towards $\gamma/\mu = m$ but never reach it.

Thus we can conclude that $x(t)$ has a unique solution for all $t \in [0, \frac{1}{2})$. □

Lemma 6.14. *When $x > y$, we have $g(x,t) > g(y,t)$ for all t.*

Proof. We will make a coupling argument. By Lemma 6.13, we know $g(x,t)$ is either strictly increasing in t, strictly decreasing in t, or constant. Assume for contradiction that there is a time s such that $g(y,s) \geq g(x,s)$. We now consider cases:

1. Suppose that $g(x,t)$ is strictly increasing in t. As $g(y,t)$ is continuous in t, and at time s, $g(y,s) \geq g(x,s) > g(x,0)$, by the Intermediate Value Theorem there must be some time r with $0 < r \leq s$ such that $g(y,r) = g(x,0)$. But as $\dot{x}(t)$ is not a function of t, only $x(t)$, we thus obtain that $g(y,s) = g(x,s-r) < g(x,s)$, giving a contradiction.
2. Suppose that $g(x,t)$ is constant. As $g(y,t)$ is continuous in t, and at time s, we have $g(y,s) \geq g(x,s) = g(x,0)$, by the Intermediate Value Theorem there is a time $r \leq s$ such that $g(y,r) = g(x,0)$. But then $g(y,t)$ is constant at r, contradicting that $g(y,t)$ must either be strictly increasing, strictly decreasing, or constant for all t.
3. Suppose that $g(x,t)$ is strictly decreasing and $g(y,t)$ is either strictly decreasing

or constant. Then by taking $\bar{g}(x,t) \triangleq -g(y,t)$ and $\bar{g}(y,t) \triangleq -g(x,t)$, we can apply cases one and two to $\bar{g}(x,t)$ to show the claim.

4. Finally, suppose that $g(x,t)$ is strictly decreasing and $g(y,t)$ is strictly increasing. Under our assumption that $g(y,s) \geq g(x,s)$, again by the Intermediate Value Theorem, there must be some time $0 < r \leq s$ such that $g(x,r) = g(y,r)$. But as $\dot{x}(t)$ depends only $x(t)$, not t, we would then have that for all $t \geq r$, $g(x,t) = g(y,t)$. This creates a contradiction, as we have assumed that $g(x,t)$ is strictly decreasing in t and $g(y,t)$ is strictly increasing in t.

□

Lemma 6.15. *For $x \geq \tilde{x}$ as defined in Lemma 6.9, $g(x,t) \geq m$ for $0 \leq t \leq \frac{1}{2}$ and $g(x,\frac{1}{2}) = x + (\gamma - c\mu)/2$. For $x < \tilde{x}$, there exist $0 \leq s < t \leq \frac{1}{2}$ such that for $\tau \in (s,t)$, $g(x,\tau) < m$, and $g(x,\frac{1}{2}) < x + (\gamma - c\mu)/2$.*

Proof. We compute for $0 \leq t \leq \frac{1}{2}$ that

$$\begin{aligned} g(\tilde{x},t) &= \tilde{x} + \int_0^t \gamma - \mu(m \wedge g(\tilde{x},\tau))d\tau \\ &\geq \tilde{x} + \int_0^t \gamma - m\mu d\tau \\ &= \tilde{x} + t(\gamma - m\mu). \end{aligned} \tag{6.97}$$

We first consider $g(\tilde{x},t)$ in cases:

1. Suppose that $\gamma \geq m\mu$. Then $\tilde{x} = m$, and as $\gamma - m\mu \geq 0$, we obtain from (6.97) that $g(\tilde{x},t) \geq m$ for all $t \geq 0$.
2. Suppose that $\gamma < m\mu$. Then $\tilde{x} = m - (\gamma - m\mu)/2$, so by (6.97) for all $t \geq 0$,

$$g(\tilde{x},t) \geq m + (\gamma - m\mu)\left(t - \frac{1}{2}\right) \geq m.$$

We thus conclude that $g(\tilde{x},t) \geq m$ for all $0 \leq t \leq \frac{1}{2}$. As a result, we can now make

the exact computation that for all $x \geq \tilde{x}$,

$$\begin{aligned} g(x, \tfrac{1}{2}) &= x + \int_0^{\frac{1}{2}} \gamma - \mu(m \wedge g(x,t))dt \\ &= x + \int_0^{\frac{1}{2}} \gamma - \mu m dt \qquad (6.98) \\ &= x + \frac{1}{2}(\gamma - m\mu). \end{aligned}$$

where (6.98) holds as $g(x,t) \geq g(\tilde{x},t) \geq m$ by Lemma 6.14 and then the above analysis, making $m \wedge g(x,t) = m$ for all t.

We now consider $x < \tilde{x}$, and find s and t such that for all $\tau \in (s,t)$, $g(x,\tau) < m$, as in the statement of the lemma. We consider cases:

1. Suppose $\gamma > m\mu$ and thus $\tilde{x} = m$. Then $g(x,0) = x < \tilde{x} = m$ so obviously we can take $s = 0$ and t small to show the claim.
2. Suppose instead that $\gamma < m\mu$ and thus $\tilde{x} = m - (\gamma - m\mu)/2$. For $x < m$, again the claim obviously holds as then $g(x,t) < m$ for all $t \leq \frac{1}{2}$. For x such that $m \leq x < \tilde{x}$, observe that as $\tilde{x} = m - (\gamma - m\mu)/2$,

$$0 \leq t^* \triangleq \frac{x-m}{m\mu - \gamma} < \frac{\tilde{x} - m}{m\mu - \gamma} = \frac{1}{2},$$

and thus

$$g(x, t^*) = x + (\gamma - m\mu)\frac{x-m}{m\mu - \gamma} = m.$$

As $\dot{x}(t) = \gamma - \mu(m \wedge x(t)) \leq \gamma - m\mu < 0$ by our assumptions, we can take $s = t^*$ and $t = \frac{1}{2}$.

Finally, using s and t from the statement of the lemma, we show that $g(x(0), \frac{1}{2}) >$

$x(0) + (\gamma - c\mu)/2$ when $x(0) < \tilde{x}$. We compute that

$$\begin{aligned} g(x, \tfrac{1}{2}) &= x + \int_0^{\frac{1}{2}} \gamma - \mu(g(x,\tau) \wedge m) d\tau \\ &\geq x + \int_0^s \gamma - \mu m d\tau + \int_s^t \gamma - \mu g(x,\tau) d\tau + \int_t^{\frac{1}{2}} \gamma - \mu m d\tau \\ &= x + (\gamma - m\mu)(\tfrac{1}{2} - (t-s)) + \int_s^t \gamma - \mu g(x,\tau) d\tau \\ &> x + (\gamma - m\mu)/2. \end{aligned}$$

The final inequality is strict as $g(x,\tau) < m$ for all $\tau \in (s,t)$. □

Lemma 6.16. *For all $x > y \geq \tilde{x}$, where $\tilde{x}$ is defined in Lemma 6.9,*

$$g(x, \tfrac{1}{2}) - g(y, \tfrac{1}{2}) = x - y,$$

and for all $x > y$, $y < \tilde{x}$,

$$0 < g(x, \tfrac{1}{2}) - g(y, \tfrac{1}{2}) < x - y.$$

Proof. For $x > y \geq \tilde{x}$, by Lemma 6.15, we have that

$$g(x, \tfrac{1}{2}) - g(y, \tfrac{1}{2}) = x + (\gamma - m\mu)/2 - y - (\gamma - m\mu)/2 = x - y.$$

For $x > y$, $y < \tilde{x}$, let s and t be from Lemma 6.15 such that for all $\tau \in (s,t)$, $g(y,\tau) < m$. Then

$$g(x, \tfrac{1}{2}) - g(y, \tfrac{1}{2}) = x - y + \mu \int_0^{\frac{1}{2}} (g(y,\tau) \wedge m) - (g(x,\tau) \wedge m) d\tau.$$

As $x > y$, by Lemma 6.14 we have $g(x,\tau) > g(y,\tau)$ for all τ, and thus $g(x,\tau) \wedge m \geq$

$g(y,\tau) \wedge m$ for all τ, making the integrand nonpositive. Thus

$$\begin{aligned} g(x,\tfrac{1}{2}) - g(y,\tfrac{1}{2}) &= x - y + \mu \int_s^t (g(y,\tau) \wedge m) - (g(x,\tau) \wedge m) d\tau \\ &= x - y + \mu \int_s^t g(y,\tau) - (g(x,\tau) \wedge m) d\tau \\ &\leq x - y + \mu \int_s^t g(y,\tau) - g(x,\tau) d\tau \\ &< x - y. \end{aligned}$$

That $g(x,\frac{1}{2}) - g(y,\frac{1}{2}) > 0$ follows immediately from Lemma 6.14. □

Proof of Lemma 6.9. The Lemma follows immediately from Lemma 6.13, Lemma 6.14, Lemma 6.15, and Lemma 6.16. □

6.12 Null Recurrence and Transience

We distinguish between the null recurrent and transient cases using Proposition A.5 from Appendix A, Section A.4. As before, we will use the Lyapunov function $V(q,r) \stackrel{\Delta}{=} q + r$.

Lemma 6.17. *For each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, the following limit exists, is finite, and is non-zero:*

$$F_\theta \stackrel{\Delta}{=} \lim_{\substack{q\to\infty \\ (q,r)\in\mathcal{S}_\theta}} \mathbb{E}_{(q,r)} \left[(V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0)))^2 \right]. \tag{6.99}$$

Further,

$$\sup_{\mathbf{s}\in\mathcal{S}_\theta} \mathbb{E}_{\mathbf{s}} \left[(V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0)))^4 \right] < \infty. \tag{6.100}$$

Finally, when $\rho_\theta = 1$, as $q \to \infty$, we have for r such that $(q,r) \in \mathcal{S}_\theta$ that

$$\mathbb{E}_{(q,r)}[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))] = O(\exp(-q/2)). \tag{6.101}$$

Proof. First consider $\theta = \text{LS}$. Recall from (6.27) that for any $\ell \geq 0$,

$$\mathbb{E}_{(q,c)}\left[\left(V(\mathbf{S}_{\text{LS}}(1)) - V(\mathbf{S}_{\text{LS}}(0))\right)^{\ell}\right] = \mathbb{E}_{(q,c)}\left[\left(A + D_{\text{LS}}^{\text{on}} - D_{\text{LS}}^{\text{off}}\right)^{\ell}\right].$$

Further, recall that we coupled $D_{\text{LS}}^{\text{on}}$ with $\tilde{D}_{\text{LS}}^{\text{on}}$ that had distribution $\text{Pois}(c\mu)$ such that $D_{\text{LS}}^{\text{on}} < \tilde{D}_{\text{LS}}^{\text{on}}$, and $D_{\text{LS}}^{\text{off}}$ with $\tilde{D}_{\text{LS}}^{\text{off}}$ that had distribution $\text{Bin}(c, 1 - e^{-\mu})$ such that $D_{\text{LS}}^{\text{off}} \leq \tilde{D}_{\text{LS}}^{\text{off}}$. Using these couplings, we can show (6.100) just as we showed (6.25).

Recall that under our coupling, we have $D_{\text{LS}}^{\text{off}} = \tilde{D}_{\text{LS}}^{\text{off}}$ for initial (q, r) such that $r = c$. Likewise, we have that $D_{\text{LS}}^{\text{on}}$ was equal to $\tilde{D}_{\text{LS}}^{\text{on}}$ under an initial condition (q, r) for those realizations where $\tilde{D}_{\text{LS}}^{\text{on}} < q - c$. This led to (6.30), namely that for $q > c$,

$$\mathbb{E}_{(q,c)}\left[D_{\text{LS}}^{\text{on}}\right] \geq \mathbb{E}\left[\tilde{D}_{\text{LS}}^{\text{on}} \mathbb{I}_{\{\tilde{D}_{\text{LS}}^{\text{on}} < q-c\}}\right].$$

We now can show (6.101). Assuming $\rho_\theta = 1$ and thus $\gamma_\theta = 0$, we have

$$\begin{aligned}
\mathbb{E}_{(q,c)}[V(\mathbf{S}_{\text{LS}}(1)) - V(\mathbf{S}_{\text{LS}}(0))] &= \mathbb{E}\left[A - D_{\text{LS}}^{\text{off}} - \tilde{D}_{\text{LS}}^{\text{on}}\right] + \mathbb{E}_{(q,c)}\left[\tilde{D}_{\text{LS}}^{\text{on}} - D_{\text{LS}}^{\text{on}}\right] \\
&\leq -\gamma_\theta + \mathbb{E}[\tilde{D}_{\text{LS}}^{\text{on}} \mathbb{I}_{\{\tilde{D}_{\text{LS}}^{\text{on}} \geq q-c\}}] \\
&\leq \sqrt{\mathbb{E}\left[\left(\tilde{D}_{\text{LS}}^{\text{on}}\right)^2\right] \mathbb{E}\left[\mathbb{I}_{\{\tilde{D}_{\text{LS}}^{\text{on}} \geq q-c\}}\right]} \\
&\leq \sqrt{((c\mu)^2 + c\mu)\mathbb{E}\left[\exp\left(\tilde{D}_{\text{LS}}^{\text{on}} - q + c\right)\right]} \\
&\leq \exp(-q/2)\sqrt{((c\mu)^2 + c\mu)\exp(c)\exp(c\mu(e-1))} \\
&= O\left(\exp(-q/2)\right),
\end{aligned}$$

as $q \to \infty$. Here we use that for any random variable X, $\mathbb{I}_{\{X \geq t\}} \leq \exp(X - t)$.

It remains to show (6.99). We will show that

$$F_{\text{LS}} = \mathbb{E}[(A - \tilde{D}_{\text{LS}}^{\text{on}} - \tilde{D}_{\text{LS}}^{\text{off}})^2].$$

For any $q > 0$, observe that

$$\mathbb{E}_{(q,c)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^2\right] = \mathbb{E}_{(q,c)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} > q-c\}}\right] + \mathbb{E}_{(q,c)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \leq q-c\}}\right].$$

For the second term, we have that

$$\begin{aligned}
&\lim_{q\to\infty} \mathbb{E}_{(q,c)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \leq q-c\}}\right] \\
&\quad = \lim_{q\to\infty} \mathbb{E}\left[\left(A - \tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - \tilde{D}^{\mathrm{off}}_{\mathrm{LS}}\right)^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} \leq q-c\}}\right] \qquad (6.102) \\
&\quad = \mathbb{E}\left[\left(A - \tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - \tilde{D}^{\mathrm{off}}_{\mathrm{LS}}\right)^2\right] \qquad (6.103) \\
&\quad = F_{\mathrm{LS}},
\end{aligned}$$

where (6.102) follows by our coupling and (6.103) follows from the Monotone Convergence Theorem. For the first term, we have

$$\begin{aligned}
&\lim_{q\to\infty} \mathbb{E}_{(q,c)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^2 \mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} > q-c\}}\right] \\
&\quad \leq \lim_{q\to\infty} \sqrt{\mathbb{E}_{(q,c)}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^4\right] \mathbb{E}_{(q,c)}\left[\mathbb{I}_{\{\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} > q-c\}}\right]} \\
&\quad \leq \lim_{q\to\infty} \sqrt{\mathbb{P}\left(\tilde{D}^{\mathrm{on}}_{\mathrm{LS}} - c > q\right)} \sqrt{\sup_{\mathbf{s}\in\mathcal{S}_{\mathrm{LS}}} \mathbb{E}_{\mathbf{s}}\left[(V(\mathbf{S}_{\mathrm{LS}}(1)) - V(\mathbf{S}_{\mathrm{LS}}(0)))^4\right]} \\
&\quad = 0,
\end{aligned}$$

where in the final equality we use (6.100). This shows (6.99) and thus the Lemma in the case of LS. The case of DA is similar. □

We now complete the proof of Theorem 6.1 by showing that for each $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$, $\{\mathbf{S}_\theta(k)\}$ is null recurrent when $\rho_\theta = 1$ and transient when $\rho_\theta > 1$.

Proof of Theorem 6.1. We have already established in section Appendix 6.5 that when $\rho_\theta \geq 1$, $\{\mathbf{S}_\theta(k)\}$ is either null recurrent or transient, and when $\rho_\theta < 1$, $\{\mathbf{S}_\theta(k)\}$ is positive recurrent. Thus it suffices to show that $\{\mathbf{S}_\theta(k)\}$ is recurrent when $\rho_\theta = 1$

and transient otherwise. We proceed using Proposition A.5. We must check that the assumptions of the proposition are satisfied by $\{\mathbf{S}_\theta(k)\}$ for $\theta \in \{\mathrm{LS}, \mathrm{DA}\}$. By (6.100) from Lemma 6.17, (A.23) is satisfied. It is obvious that (A.22) is satisfied. Finally, to check (A.21), we verify the sufficient condition (A.26). Recalling Corollary 6.2, it is immediate that for all $z > 0$,

$$\inf_{\mathbf{s} \in \mathcal{S}_\theta} \mathbb{P}(V(\mathbf{S}_\theta(1)) \geq z \mid \mathbf{S}_\theta(0) = \mathbf{s}) = \mathbb{P}(V(\mathbf{S}_\theta(1)) \geq z \mid \mathbf{S}_\theta(0) = (0,0)) > 0.$$

Now suppose $\rho_\theta = 1$. We will show that (A.24) holds. For a constant $b_\theta > 0$, we will take B_θ of the form $\{(q,r) \in \mathcal{S}_\theta \mid q < b_\theta\}$. By (6.101) from Lemma 6.17, as $q \to \infty$, we have

$$\mathbb{E}_{(q,r)}\left[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))\right] = O(\exp(-q/2)),$$

and by (6.99) from Lemma 6.17, as $q \to \infty$,

$$\frac{\mathbb{E}_{(q,r)}\left[\left(V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))\right)^2\right]}{2V(q,r)} = \Theta\left(\frac{1}{q}\right).$$

Thus by taking b_θ sufficiently large and using that $\exp(-q/r) = o(1/q)$, we have (A.24) for all $\mathbf{s} \in \mathcal{S}_\theta \setminus B_\theta$, showing that $\{\mathbf{S}_\theta(k)\}$ is null recurrent.

Alternatively, suppose that $\rho_\theta > 1$. We now must show that (A.25) holds. We use b_θ to define B_θ in the same way. By Lemma 6.1 and Lemma 6.2 we have

$$\lim_{\substack{q \to \infty \\ (q,r) \in \mathcal{S}_\theta}} \mathbb{E}_{(q,r)}\left[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))\right] = -\gamma_\theta > 0,$$

and again by (6.99) we have

$$\frac{\mathbb{E}_{(q,r)}\left[V(\mathbf{S}_\theta(1)) - V(\mathbf{S}_\theta(0))\right]}{2V(\mathbf{S}_\theta(0))} = \Theta\left(\frac{1}{q}\right).$$

Thus by taking b_θ sufficiently large and $\varepsilon = 1$, we have (A.25), showing $\{\mathbf{S}_\theta(k)\}$ is transient, completing the proof. □

Appendix A

Lyapunov Functions

Throughout, suppose $\{X_k\}$ is a discrete time irreducible Markov chain on a countable state space $\mathcal{X}$. We now give a series of results characterizing the recurrence of $\{X_k\}$ and, for a function $V\colon \mathcal{X} \to \mathbb{R}_+$, the moments of $V(X_k)$. The tools used to obtained these results are commonly referred to as Lyapunov-Foster functions, or test functions. These results are used extensively in Chapters 4 and 6.

A.1 Positive Recurrence

In this section, we give sufficient conditions to prove whether or not $\{X_k\}$ is positive recurrent. Distinguishing between the null recurrent and transient cases is addressed in Section A.4 with similar techniques.

First we give a condition for the positive recurrence of $\{X_k\}$ due to [36], see [9] for a modern reference. We use $\mathbb{E}_x$ to denote the expectation operator conditional on $X_0 = x$.

Proposition A.1 (Foster et al. 36)**.** *If there exists a function $V\colon \mathcal{X} \to \mathbb{R}$, $\gamma > 0$, and a finite set $B \subset \mathcal{X}$ such that for all $x \in B$,*

$$\mathbb{E}_x[V(X_1) - V(X_0)] < \infty, \tag{A.1}$$

and for all $x \in \mathcal{X} \setminus B$,

$$\mathbb{E}_x[V(X_1) - V(X_0)] \leq -\gamma,$$

then $\{X_k\}$ *is positive recurrent.*

A function V satisfying these properties is usually called a *Lyapunov function*. Lyapunov functions can also be used to prove $\{X_k\}$ is not positive recurrent when the drift is nonnegative. The following is a special case of Proposition 5.4 from [9].

Proposition A.2. *Suppose there exists a Lyapunov function* $V \colon \mathcal{X} \to \mathbb{R}$, *a finite set* $B \subset \mathcal{X}$ *and a state* $y \in \mathcal{X} \setminus B$ *satisfying*

$$\sup_{x \in B} V(x) < V(y), \tag{A.2}$$

$$\sup_{x \in \mathcal{X}} \mathbb{E}_x \left[(V(X_1) - V(X_0))^2 \right] < \infty, \tag{A.3}$$

$$\inf_{x \in \mathcal{X} \setminus B} \mathbb{E}_x[V(X_1) - V(X_0)] \geq 0. \tag{A.4}$$

Then $\{X_k\}$ *is either null recurrent or transient.*

A.2 Moment Bounds

Now, suppose that $\{X_k\}$ is positive recurrent, and let X_∞ denote the unique steady state distribution. We now give a bound on the first moment of $f(X_\infty)$ for any function f. The result below is new, but similar to existing results from Gamarnik and Zeevi [38], Glynn and Zeevi [41].

Proposition A.3. *Suppose that* X_t *is positive recurrent and that there exist* $\alpha, \beta, \gamma > 0$, *a set* $B \subset \mathcal{X}$ *and functions* $U \colon \mathcal{X} \to \mathbb{R}_+$ *and* $f \colon \mathcal{X} \to \mathbb{R}_+$ *such that for* $x \in \mathcal{X} \setminus B$,

$$\mathbb{E}_x[U(X_1) - U(X_0)] \leq -\gamma f(x), \tag{A.5}$$

and for $x \in B$,

$$f(x) \leq \alpha, \tag{A.6}$$

$$\mathbb{E}_x[U(X_1) - U(X_0)] \leq \beta. \tag{A.7}$$

Then

$$\mathbb{E}[f(X_\infty)] \leq \alpha + \frac{\beta}{\gamma}.$$

Proof. For every $z > 0$ let $U_z \colon \mathcal{X} \to \mathbb{R}_+$ and $f_z \colon X \to \mathbb{R}_+$ be given by

$$U_z(x) \triangleq \min\{U(x), z\}, \qquad\qquad f_z(x) \triangleq f(x)\mathbb{I}_{\{U(x)<z\}}.$$

Trivially for all sufficiently large z, we have $f_z(x) = f(x)$ and $U_z(x) = U(x)$ for all $x \in B$, so (A.6) and (A.7) are satisfied by f_z and U_z for all large z. We claim that (A.5) is satisfied as well. Suppose that x is such that $U(x) \geq z$. Then

$$\mathbb{E}_x[U_z(X_1) - U_z(X_0)] = \mathbb{E}_x[U_z(X_1)] - z \leq 0 = f_z(x) = -\gamma f_z(x).$$

Alternatively, when x is such that $U(x) < z$, using that $U_z(x) \leq U(x)$ for all x,

$$\begin{aligned}
\mathbb{E}_x[U_z(X_1) - U_z(X_0)] &= \mathbb{E}_x[U_z(X_1) - U(X_0)] \\
&\leq \mathbb{E}_x[U(X_1) - U(X_0)] \\
&\leq -\gamma f(x) \\
&= -\gamma f_z(x).
\end{aligned}$$

Thus for all z and all $x \in \mathcal{X} \setminus B$, $\mathbb{E}_x[U_z(X_1) - U_z(X_0)] \leq -\gamma f_z(x)$ in analogy with (A.5). As for all z, U_z is bounded by construction, $\mathbb{E}[U_z(X_\infty)]$ is finite. By stationarity

and the finiteness of $\mathbb{E}[U_z(X_\infty)]$,

$$\begin{aligned}
0 &= \mathbb{E}\left[\mathbb{E}_{X_\infty}\left[U_z(X_1) - U_z(X_0)\right]\right] \\
&= \sum_{x\in B} \mathbb{P}(X_\infty = x)\mathbb{E}_x[U_z(X_1) - U_z(X_0)] + \sum_{x\notin B} \mathbb{P}(X_\infty = x)\mathbb{E}_x[U_z(X_1) - U_z(X_0)] \\
&\leq \beta\mathbb{P}(X_\infty \in B) - \gamma \sum_{x\notin B} \mathbb{P}(X_\infty = x) f_z(x) \\
&\leq \beta - \gamma\mathbb{E}[f_z(X_\infty)\mathbb{I}_{\{X_\infty \notin B\}}],
\end{aligned}$$

or rearranging terms,

$$\mathbb{E}[f_z(X_\infty)\mathbb{I}_{\{X_\infty \notin B\}}] \leq \frac{\beta}{\gamma}.$$

Thus

$$\begin{aligned}
\mathbb{E}[f_z(X_\infty)] &= \mathbb{E}[f_z(X_\infty)\mathbb{I}_{\{X_\infty \in B\}}] + \mathbb{E}[f_z(X_\infty)\mathbb{I}_{\{X_\infty \notin B\}}] \\
&\leq \alpha\mathbb{P}(X_\infty \in B) + \frac{\beta}{\gamma} \\
&\leq \alpha + \frac{\beta}{\gamma}.
\end{aligned}$$

Now by Fatou's lemma, as $f_z(X_\infty) \to f(X_\infty)$ almost surely as $z \to \infty$,

$$\mathbb{E}[f(X_\infty)] = \mathbb{E}\left[\lim_{z\to\infty} f_z(X_\infty)\right] \leq \liminf_{z\to\infty} \mathbb{E}[f_z(X_\infty)] \leq \alpha + \frac{\beta}{\gamma},$$

giving the result. □

Remark A.1. Observe that we need not assume that B is bounded.

A.3 Moment Bound with Unbounded Downward Jumps

When applying Proposition A.3 in queuing problems, it is common to take $f(\cdot)$ as the number of customers in system and $U(\cdot)$ as the square of the number of customers in system. However, when analyzing systems where jumps downward are not easily bounded, e.g., a queue with abandonment, it can be difficult to show that (A.5) holds by naïvely evaluating the equation using this $f(\cdot)$ and $U(\cdot)$. In this section, we specialize our analysis to a family of Markov chains with a certain queuing structure, and then give a steady state moment bound that applies in the presence of large downward jumps.

Given an irreducible aperiodic Markov chain $\{X_k\}$ on a countable statespace $\mathcal{X}$, suppose there exists a nonnegative function $V\colon \mathcal{X} \to \mathbb{R}_+$ which admits the following decomposition

$$V(X_k) = V(X_{k-1}) + A_k - D_k, \tag{A.8}$$

where the $A_k \geq 0$ is an i.i.d. sequence such that A_k is independent from state X_k, while the $D_k \geq 0$ may depend on X_k and A_k, but not on k directly. Specifically, D_k is a function of X_k and A_k. A_k and D_k are interpreted as the number of arrivals and departures in the time period of length k, respectively. Assume in addition that $\mathcal{B}(\alpha) \triangleq \{x \in \mathcal{X} \mid V(x) \leq \alpha\} \subset \mathcal{X}$ is finite for every α. Note that as $V(x) \geq 0$, we have that $D_k \leq V(x) + A_k$ a.s.

Proposition A.4. *Suppose $\mathbb{E}[A_k^2]$ is finite and C_1 satisfies $\mathbb{E}[A_k^2] \leq C_1 \mathbb{E}[A_k]^2 < \infty$. Suppose there exists $\alpha, \lambda, C_2 > 0$ such that for every $x \notin \mathcal{B}(\alpha)$*

$$\mathbb{E}\left[A_k - \tilde{D}_k | X_k = x\right] \leq -\lambda \mathbb{E}[A_k], \tag{A.9}$$

where $\tilde{D}_k$ is defined to be $\min\{D_k, C_2 A_k\}$. Then X_k is positive recurrent with the

unique stationary distribution X_∞ *and*

$$\mathbb{E}[V(X_\infty)] \leq \max\left\{\alpha, \frac{\max\{1, C_2 - 1\}^2 C_1}{\lambda}\mathbb{E}[A_k]\right\}\left(2 + \frac{2}{\lambda}\right).$$

The reason for introducing a truncated downward jump process $\tilde{D}_k$ as opposed to using just D_k is that in general the statement of the proposition is not true. Namely, there exists a process such the assumptions of the proposition above hold true when D_k replaces $\tilde{D}_k$ in (A.9) and $\mathbb{E}[V(X_\infty)] = \infty$, as shown by Example A.1 below.

Proof. First, we apply Proposition A.1 to X_k using the same $V(x)$, $\mathcal{B}$, and $\gamma = \lambda\mathbb{E}[A_k]$ as in the statement of Proposition A.4. For $x \notin \mathcal{B}$, we have

$$\mathbb{E}_x[V(X_1) - V(X_0)] = \mathbb{E}_x[A_0 - D_0] \leq \mathbb{E}_x\left[A_0 - \tilde{D}_0\right] \leq -\lambda\mathbb{E}_x[A_0] = -\gamma,$$

where in the inequalities we use $\tilde{D}_k \leq D_k$ and then (A.9). For all x, we have

$$\mathbb{E}_x[V(X_1) - V(X_0)] = \mathbb{E}_x[A_0 - D_0] \leq \mathbb{E}_x[A_0] < \infty.$$

Thus as $\mathcal{B}$ is bounded, we can apply Proposition A.1 to obtain positive recurrence of X_k. Let X_∞ be the steady state version of the Markov chain X_k.

Next, we apply Proposition A.3 taking $U(x) = V^2(x)$ and $f(x) = V(x)$. We let

$$\alpha' = \max\left\{\alpha, \frac{\max\{1, C_2 - 1\}^2 C_1}{\lambda}\mathbb{E}[A_k]\right\}, \tag{A.10}$$

thus making our set of exceptions from Proposition A.3 $\mathcal{B}' = \{x \in \mathcal{X} \mid V(x) \leq \alpha'\}$.

We have for $x \in \mathcal{B}'$ that

$$\begin{aligned}
\mathbb{E}_x[U(X_1) - U(X_0)] &= \mathbb{E}_x\left[(V(X_0) + A_0 - D_0)^2 - V(X_0)^2\right] && \text{(A.11)}\\
&\leq \mathbb{E}_x\left[(V(X_0) + A_0)^2 - V(X_0)^2\right] && \text{(A.12)}\\
&= 2V(x)\mathbb{E}[A_0] + \mathbb{E}[A_0^2]\\
&\leq 2\alpha'\mathbb{E}[A_0] + C_1\mathbb{E}[A_0]^2 && \text{(A.13)}\\
&\leq 2\alpha'\mathbb{E}[A_0] + \alpha'\lambda\mathbb{E}[A_0] && \text{(A.14)}\\
&= \alpha'(2+\lambda)\mathbb{E}[A_0]\\
&\triangleq \beta' && \text{(A.15)}
\end{aligned}$$

where (A.11) follows from (A.8), (A.12) follows as $V(X_1) \geq 0$, and (A.13) follows from (A.10) and the definition of C_1, and (A.14) follows again from (A.10) and as $\max\{1, C_2 - 1\}^2 \geq 1$ by definition. We have for $x \notin \mathcal{B}'$, that

$$\begin{aligned}
\mathbb{E}_x[U(X_1) - U(X_0)] &= \mathbb{E}_x\left[(V(X_0) + A_0 - D_0)^2 - V(X_0)^2\right] && \text{(A.16)}\\
&\leq \mathbb{E}_x\left[(V(X_0) + A_0 - \tilde{D}_0)^2 - V(X_0)^2\right] && \text{(A.17)}\\
&= 2V(x)\mathbb{E}_x[A_0 - \tilde{D}_0] + \mathbb{E}_x\left[(A_0 - \tilde{D}_0)^2\right]\\
&\leq -2V(x)\lambda\mathbb{E}[A_0] + \mathbb{E}\left[\max\{1, C_2 - 1\}^2 A_0^2\right] && \text{(A.18)}\\
&\leq -(V(x) + \alpha')\lambda\mathbb{E}_x[A_0] + \max\{1, C_2 - 1\}^2 C_1\mathbb{E}\left[A_0\right]^2 && \text{(A.19)}\\
&\leq -(V(x) + \alpha')\lambda\mathbb{E}_x[A_0] + \alpha'\lambda\mathbb{E}\left[A_0\right] && \text{(A.20)}\\
&= -V(x)\lambda\mathbb{E}_x[A_0]
\end{aligned}$$

where (A.16) follows from (A.8), (A.17) follows as $V(X_1) \geq 0$ and $\tilde{D}_0 \leq D_0$, (A.18) follows from (A.9) and as $\tilde{D}_k \leq C_2 A_k$ a.s. implies that $|A_k - D_k| \leq \max\{1, C_2 - 1\}A_k$ a.s., (A.19) follows from (A.10) and the definition of C_1, and finally (A.20) follows again from (A.10). Thus by taking $\gamma' = \lambda\mathbb{E}_x[A_0]$, we can now apply Proposition A.3

with α', β' and γ' to obtain that

$$\begin{aligned}\mathbb{E}[V(\infty)] &\leq \alpha' + \frac{\beta'}{\gamma'} = \alpha' + \frac{\alpha'(2+\lambda)\mathbb{E}[A_0]}{\lambda \mathbb{E}[A_0]} = \alpha'\left(1 + \frac{2+\lambda}{\lambda}\right) \\ &= \max\left\{\alpha, \frac{\max\{1, C_2 - 1\}^2 C_2}{\lambda}\mathbb{E}[A_k]\right\}\left(1 + \frac{2+\lambda}{\lambda}\right)\end{aligned}$$

showing the result.

□

Last, we give a quick counter example showing that without some assumptions beyond simply having negative drift, we may not even have a finite first moment.

Example A.1. Consider the following random walk X_t on the nonnegative integers parametrized by some $\gamma \in (0, 1)$. From state 0, we always go up to state 1. For every other state $k = 1, 2, \ldots$, with probability $(1+\gamma)/(k+1)$, we go to state 0, and with the remaining probability, $(k-\gamma)/(k+1)$, we go up to state $k+1$. This walk has the property that for all $k \geq 1$,

$$\mathbb{E}_k[X_1 - X_0] = (k+1) \cdot \frac{k-\gamma}{k+1} + 0 \cdot \frac{1+\gamma}{k+1} - k = -\gamma,$$

and thus is positive recurrent and has some stationary distribution $\pi_k = \mathbb{P}(X_\infty = k)$. However, we will show that $\mathbb{E}[X_\infty] = \infty$. A direct computation of the steady-state equations gives that $\pi_0 = \pi_1$, and for $n \geq 2$,

$$\pi_n = \frac{n-\gamma}{n+1}\pi_{n-1} = \pi_0 \prod_{k=2}^{n} \frac{k-\gamma}{k+1} = \frac{\pi_0}{\Gamma(2-\gamma)} \frac{\Gamma(n+1-\gamma)}{\Gamma(n+2)},$$

where Γ is the Gamma function. Using the identity

$$\lim_{n\to\infty} \frac{\Gamma(n+\alpha)}{\Gamma(n)n^\alpha} = 1$$

for all $\alpha \in \mathbb{R}$, we have that

$$\pi_n = \Theta\left(\frac{1}{n^{1+\gamma}}\right).$$

Thus there exists $c > 0$ and $\ell \in \mathbb{Z}_+$ such that

$$\mathbb{E}[X_\infty] = \sum_{n=0}^{\infty} n\pi_n \geq \sum_{n=\ell}^{\infty} \frac{c}{n^\gamma} = \infty,$$

showing the claim.

A.4 Null Recurrence

We give a sufficient condition to distinguish between the null recurrent and transient cases from [56], Theorem 3.2, (see also Section 3.6 from [32]). We do not present the theorem in full generality.

Proposition A.5. *Given a finite set $B \subset \mathcal{X}$ and Lyapunov function $V \colon \mathcal{X} \to \mathbb{R}_+$, assume that*

$$\mathbb{P}\left(\limsup_{k\to\infty} V(X_k) = \infty\right) = 1, \tag{A.21}$$

$$\inf_{x\in\mathcal{X}} \mathbb{E}_x\left[(V(X_1) - V(X_0))^2\right] > 0, \tag{A.22}$$

$$\sup_{x\in\mathcal{X}} \mathbb{E}_x\left[(V(X_1) - V(X_0))^4\right] < \infty. \tag{A.23}$$

If for all $x \in \mathcal{X} \setminus B$,

$$\mathbb{E}_x\left[V(X_1) - V(X_0)\right] \leq \frac{\mathbb{E}_x\left[(V(X_1) - V(X_0))^2\right]}{2V(x)}, \tag{A.24}$$

then $\{X_k\}$ is recurrent. Alternatively, if there exists $\varepsilon > 0$ such that for all $x \in \mathcal{X} \setminus B$,

$$\mathbb{E}_x\left[V(X_1) - V(X_0)\right] \geq (1+\varepsilon)\frac{\mathbb{E}_x\left[(V(X_1) - V(X_0))^2\right]}{2V(x)}, \tag{A.25}$$

then $\{X_n\}$ is transient.

Remark A.2. As noted in [56], a sufficient condition for (A.21) is that for every

$z \geq 0$,

$$\inf_{x \in \mathcal{X}} \mathbb{P}(V(X_1) \geq z \mid X_0 = x) > 0. \tag{A.26}$$

Appendix B

Random Graphs

In this chapter, we prove a result showing that a directed bipartite Erdős-Rényi random graph with high probability. The result is used to prove Theorem 4.3. Throughout, we use the notation introduced in Chapter 4.

B.1 Results

We begin by stating a result on long chains in a static Erdős-Rényi random graph. The following result was first shown by [3] and refined in a series of papers, see [55] for a historical account and the most tight result.

Proposition B.1 (Krivelevich et al. 55)**.** *Fix any $\varepsilon > 0$ and any $\delta > 0$. There exist C and n_0 such that for all $c > C$ and all $n > n_0$ the following occurs: Consider an* $\mathrm{ER}(n, c/n)$ *directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and let D be the length of the longest directed cycle. We have*

$$\mathbb{P}(D > (1 - (2 + \delta)ce^{-c})n) > 1 - \varepsilon\,.$$

In words, (for large c) we have a cycle containing a large fraction of the nodes with high probability. From this, we can easily obtain a similar result about the longest path starting from a specific node.

Corollary B.1. *Fix any $\varepsilon > 0$. There exist C and n_0 such that for all $c > C$ and all $n > n_0$ the following occurs: Consider a set $\mathcal{V}$ of n vertices including a fixed vertex $v \in \mathcal{V}$, and draw an $\mathrm{ER}(n, c/n)$ directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Let P_v denote the length of a longest path starting at v. Then*

$$\mathbb{P}(P_v < n(1-\varepsilon)) \leq \varepsilon.$$

The proof is deferred to Section B.2. We extend the result above to the case of bipartite random graphs.

Corollary B.2. *Fix any $\kappa > 1$ and $\varepsilon > 0$. Then there exists $p_0 > 0$ and $C > 0$ such that the following holds: Consider any $c_L \in [1/\sqrt{\kappa}, \kappa]$, any $c_R > C$, and any $p < p_0$. Let $\mathcal{L}$ be a set of c_L/p vertices and let $\mathcal{R}$ be a set of c_R/p vertices. Fix a vertex $v \in \mathcal{L}$. Draw $\mathcal{G} = (\mathcal{L}, \mathcal{R}, \mathcal{E})$ as an $\mathrm{ER}(c_L/p, c_R/p, p)$ bipartite random graph. We have*

$$\mathbb{P}\left(P_v < 2\frac{c_L}{p}(1-\varepsilon)\right) \leq \varepsilon,$$

where again, P_v is the length of a longest path starting at v.

Again, the proof is in Section B.2. The requirement $c_L \in [1/\sqrt{\kappa}, \kappa]$ here will correspond to p times the 'typical' interval between successive times when the chain advances under greedy. These intervals are distributed i.i.d. Geometric(p), and hence typically lie in the range $[1/(p\sqrt{\kappa}), \kappa/p]$ for large κ, as stated in Lemma 4.2 below. The $1/\sqrt{\kappa}$ term in the lower bound of this 'typical' range is a somewhat arbitrary choice we make that facilitates a proof of Theorem 4.3 (a variety of other decreasing functions of κ would work as well).

B.2 Proofs

We first prove Corollary B.1. The proof follows relatively easily from Proposition B.1. The idea is as follows. A sufficient condition to form a long chain from a node v is for v to be a member of the long cycle that will occur with high probability according to

the proposition. Note that with constant probability e^{-c}, v will be isolated and thus not be part of the cycle, but we can make this probability small by taking C large.

Proof of Corollary B.1. Given ε from the statement of Corollary B.1, let $\bar{C}$ and $\bar{n}_0$ be values guaranteed to exist from Proposition B.1 applied when $\delta = 1$ and the probability of a long chain existing is at least $1 - \varepsilon/2$.

There exists C^* such that for all $c > C^*$, $3ce^{-c} < \varepsilon/2$ as the function $f(x) = xe^{-x}$ is strictly decreasing for $x > 1$. We claim that given our ε, Corollary B.1 holds by taking $C = \max\{\bar{C}, C^*\}$ and $n_0 = \bar{n}_0$.

Given our $\mathrm{ER}(n, c/n)$ graph where $n > n_0$ and $c > C$ and a fixed node v, let A be the event it contains a cycle of length at least $(1 - 3ce^{-c})n$, and let $B \subset A$ be the event that that v is in the cycle. Observe that it suffices to prove that $P(B) > 1 - \varepsilon$ to show the result, as $3ce^{-c} < \varepsilon$ by our assumption that $c > C \geq C^*$ and the definition of C^*. Thus we compute that

$$\mathbb{P}(B) = P(B|A)\mathbb{P}(A) \geq (1 - \varepsilon/2)(1 - \varepsilon/2) \geq 1 - \varepsilon,$$

showing the result, where $\mathbb{P}(A) \geq 1 - \varepsilon/2$ follows from Proposition B.1 and $P(B|A) \geq 1 - \varepsilon/2$ follows as the cycle is equally likely to pass through every node, so when the cycle hits $1 - \varepsilon/2$ fraction of the nodes, it has this probability of hitting v. □

Next, we prove Corollary B.2. The idea of the proof is as follows. First, we show with a simple calculation that a constant fraction of the nodes in $\mathcal{R}$ will have both in and out degree one, as $p \to 0$. We consider paths which only use this subset of nodes from $\mathcal{R}$. Such a path is equivalent to a path in a modified graph on the set of nodes $\mathcal{L}$ where there is an edge between two nodes u and v if and only if there is a path of length two between them via an intermediate node in $\mathcal{R}$ which has in and out degree one. Such a graph behaves (approximately) as an Erdős-Rényi graph on the nodes of $\mathcal{L}$, with the number of edges proportionate to $|\mathcal{R}|$. Thus by ensuring that $|\mathcal{R}|$ is sufficiently large, we can apply Corollary B.1 to obtain the result.

Proof of Corollary B.2. Fix $\kappa > 1$ and $\varepsilon > 0$ from the statement of the corollary. For

C and p_0 to be chosen later, let $c_L \in [1/\sqrt{\kappa}, \kappa]$, $c_R > C$, and $p < p_0$ be arbitrary. Given our graph $\mathcal{G} = (\mathcal{L}, \mathcal{R}, \mathcal{E})$ that is $\mathrm{ER}(c_L/p, c_R/p, p)$, consider the subgraph $\mathcal{G}' = (\mathcal{L}, \mathcal{R}', \mathcal{E}')$ of $\mathcal{G}$ where $\mathcal{R}'$ is the set of vertices in $\mathcal{R}$ with in degree one and out degree one in $\mathcal{G}$, and $\mathcal{E}'$ are the edges in $\mathcal{E}$ such that both endpoints are in $\mathcal{G}'$. From this graph, we create a new directed non-bipartite digraph $\mathcal{G}'' = (\mathcal{L}, \mathcal{E}'')$ where and there is an edge from $u \in \mathcal{L}$ to $v \in \mathcal{L}$ iff there is at least one node $r \in \mathcal{R}'$ such that $(u, r) \in \mathcal{E}'$ and $(r, v) \in \mathcal{E}'$. Observe that a path of length k in $\mathcal{G}''$ gives a path of length $2k$ in $\mathcal{G}$ by following the two edges in $\mathcal{G}'$ for each edge in the path on $\mathcal{G}''$, so it suffices to find a path of length $(1 - \varepsilon)c_L/p$ in $\mathcal{G}''$.

For any vertex $r \in \mathcal{R}$, let I_r be the indicator variable that r has an in degree of one and an out degree of one. Note that these variables are independent. Further, we have

$$\mu(p) \triangleq \mathbb{P}(I_r = 1) = \mathbb{P}(\mathrm{Bin}(|\mathcal{L}|, p) = 1)^2 = \left(\frac{c_L}{p} p (1-p)^{\frac{c_L}{p} - 1}\right)^2 \to c_L^2 \exp(-2c_L),$$

as $p \to 0$. As each of the I_r are independent, we have that

$$|\mathcal{R}'| \stackrel{d}{=} \mathrm{Bin}\left(\frac{c_R}{p}, \mu(p)\right).$$

Letting

$$A_1(\delta_1) = \left\{(1 - \delta_1)\frac{c_R}{p}\mu(p) < |\mathcal{R}'| < (1 + \delta_1)\frac{c_R}{p}\mu(p)\right\}$$

we have by Proposition 4.1 that for all p,

$$\mathbb{P}(A_1(\delta_1)) \geq 1 - 2\exp\left(-\delta_1^2 \frac{c_R}{p}\frac{\mu(p)}{3}\right).$$

We can view the edges of $\mathcal{G}''$ as being generated by the following process: for each $r \in \mathcal{R}'$, pick a source and then a destination uniformly at random from $\mathcal{L}$ and add an edge from the source to the destination unless either:

- the source and destination are the same node,

- an edge between the source and destination already exists in the graph.

Thus $|\mathcal{E}''|$ is the number of non empty bins if we throw $|\mathcal{R}'|$ balls into $(c_L/p)^2$ bins and then throw out the c_L/p bins that correspond to self edges. (Alternatively, we can think of this process as throwing c_R/p balls, but each ball "falls through" only with probability $1-\mu(p)$. This problem was studied extensively in [73], but here we need only a coarse analysis). Trivially, $|\mathcal{E}''| \leq |\mathcal{R}'|$. We now show that typically, the number of nonempty bins is almost equal to the number of balls thrown. For each $r \in \mathcal{R}'$, let X_r be the indicator that there is $\ell \in \mathcal{L}'$ such that $(\ell, r) \in \mathcal{E}'$ and $(r, \ell) \in \mathcal{E}'$. It is easy to see that the X_r are i.i.d. Bernoulli(p/c_L). For each $\{r, s\} \subset \mathcal{R}'$, let $Y_{\{rs\}}$ be the indicator that the nodes r and s are "colliding" on both their source and destination choices in $\mathcal{L}'$, i.e. there is $\ell, m \in \mathcal{L}'$, $\ell \neq m$, such that $(\ell, r), (\ell, s), (r, m), (s, m) \in \mathcal{E}'$. It is easy to see that $\mathbb{P}(Y_{\{rs\}} = 1) \leq p^2/c_L^2$ for each $\ell, m \in \mathcal{L}'$. We have

$$|\mathcal{E}''| \geq |\mathcal{R}'| - \sum_{r \in \mathcal{R}'} X_r - \sum_{\{r,s\} \subset \mathcal{R}'} Y_{\{rs\}}.$$

We compute that for any fixed $\mathcal{R}'$

$$\mathbb{E}\left[\sum_{r \in \mathcal{R}'} X_r + \sum_{\{r,s\} \subset \mathcal{R}'} Y_{\{rs\}}\right] \leq |\mathcal{R}'|\frac{p}{c_L} + \binom{|\mathcal{R}'|}{2}\frac{p^2}{c_L^2} \leq |\mathcal{R}'|\frac{p}{c_L} + \left(|\mathcal{R}'|\frac{p}{c_L}\right)^2$$

Letting

$$A_2(\delta_2) = \left\{\sum_{r \in \mathcal{R}'} X_r + \sum_{\{r,s\} \subset \mathcal{R}'} Y_{rs} \leq \delta_2 |\mathcal{R}'|\right\},$$

we have that

$$\mathbb{P}(A_2(\delta_2)) \geq 1 - \frac{p}{\delta_2 c_L} - |\mathcal{R}'|\delta_2^{-1}\left(\frac{p}{c_L}\right)^2$$

Letting

$$B(\delta_1, \delta_2) = \left\{ (1-\delta_1)(1-\delta_2)\frac{c_R}{p}\mu(p) < |\mathcal{E}''| < (1+\delta_1)\frac{c_R}{p}\mu(p) \right\},$$

we have that $B(\delta_1, \delta_2) \supset A_1(\delta_2) \cap A_2(\delta_2)$, and thus by taking complements and then applying the union bound,

$$\begin{aligned}\mathbb{P}(B(\delta_1, \delta_2)) &\geq 1 - \mathbb{P}(A_1(\delta_1)^c) - \mathbb{P}(A_2(\delta_2)^c) \\ &\geq 1 - 2\exp\left(-\delta_1^2 \frac{c_R}{p}\mu(p)/3\right) - \frac{p}{\delta_2 c_L} - (1+\delta_1)c_R\mu(p)\frac{p}{\delta_2^2 c_L^2},\end{aligned}$$

thus giving us a high probability bound on the size of $|\mathcal{E}''|$ as $p \to 0$.

For our fixed ε, let $\tilde{C}$ and $\tilde{n}_0$ be C and n_0 from Corollary B.1 such that for any $c > \tilde{C}$ and $n > \tilde{n}_0$, given a node in graph $\mathrm{ER}(n, c/n)$, there exists a path with length at least $n(1-\varepsilon/2)$ with probability at least $1 - \varepsilon/2$. We now specifiy p_0 from the corollary to be such that for all $c_L \in [1/\sqrt{\kappa}, \kappa]$, we have $c_L/p_0 > \tilde{n}_0$, i.e. $p_0 < 1/(n_0\sqrt{\kappa})$.

Let $\tilde{\mathcal{G}} = (\mathcal{L}, \tilde{\mathcal{E}})$ be an $\mathrm{ER}(c_L/p, \tilde{C}p/c_L)$ directed random graph. We now couple $\mathcal{G}''$ (a directed $\mathrm{ER}(n, M)$ graph, where M is random but independent of the edges selected) and $\tilde{\mathcal{G}}$ (a directed $\mathrm{ER}(n, p)$ graph) in the standard way so that when $|\tilde{\mathcal{E}}| \leq |\mathcal{E}''|$, then $\tilde{\mathcal{E}} \subset \mathcal{E}''$ and when $|\mathcal{E}''| \leq |\tilde{\mathcal{E}}|$, then $\mathcal{E}'' \subset \tilde{\mathcal{E}}$. Thus if $\tilde{\mathcal{G}}$ has a long path and $|\tilde{\mathcal{E}}| < |\mathcal{E}''|$, then $\mathcal{G}''$ will have at least as long a path as well, as it will contain more edges on the same nodes. Let $\tilde{P}$ be the length of a longest path starting at v in $\tilde{\mathcal{G}}$. Letting

$$A_3 = \left\{ \tilde{P} > \left(1 - \frac{\varepsilon}{2}\right)\frac{c_L}{p} \right\}$$

and recalling that $p_0 < 1/(n_0\sqrt{\kappa})$ implies that $c_L/p > \tilde{n}_0$, we have by Proposition B.1

$$\mathbb{P}(A_3) \geq 1 - \frac{\varepsilon}{2}$$

We now need to show that $\mathcal{G}''$ will have more edges than $\tilde{\mathcal{G}}$ with high probability for

all c_R sufficiently large (which we can control by choice of C from the statement of the corollary). We have $|\tilde{\mathcal{E}}| \sim \text{Bin}((c_L/p-1)c_L/p, \tilde{C}p/c_L)$, thus by Proposition 4.1, if

$$A_4(\delta_4) = \left\{\tilde{C}(c_L/p-1)(1-\delta_4) < |\tilde{\mathcal{E}}| < \tilde{C}(c_L/p-1)(1+\delta_4)\right\}$$

then

$$\mathbb{P}(A_4(\delta_4)) \geq 1 - 2\exp\left(-\delta_4^2\tilde{C}\left(\frac{c_L}{p}-1\right)/3\right)$$

Now, for any fixed choice of $\delta_1, \delta_2, \delta_4$, there exists C sufficiently large such that if $c_R > C$ then for all $p < p_0$ and all $c_L \in [1/\sqrt{\kappa}, \kappa]$,

$$\tilde{C}\left(\frac{c_L}{p}-1\right)(1+\delta_4) < (1-\delta_1)(1-\delta_2)\frac{c_R}{p}\mu(p)$$

(recall that $\mu(p)$ converges to a constant depending only c_L uniformly over $[1/\sqrt{\kappa}, \kappa]$ as $p \to 0$). For such c_R, we have

$$\{|\mathcal{E}''| > |\tilde{\mathcal{E}}|\} \subset B(\delta_1, \delta_2) \cap A_4(\delta_4),$$

as B makes $|\mathcal{E}''|$ big and A_4 ensures that $|\tilde{\mathcal{E}}|$ is small. Putting everything together, we have that

$$\left\{P > 2\frac{c_L}{p}(1-\varepsilon)\right\} \supset B(\delta_1, \delta_2) \cap A_3 \cap A_4(\delta_4),$$

so by taking complements and then applying the union bound, we obtain

$$\mathbb{P}\left(P > 2\frac{c_L}{p}(1-\varepsilon)\right) \geq 1 - \mathbb{P}(B(\delta_1,\delta_2)^c) - \mathbb{P}(A_3^c) - \mathbb{P}(A_4(\delta_4)^c) = 1 - \frac{\varepsilon}{2} - O(p),$$

showing the result. □

Bibliography

[1] D. J. Abraham, A. Blum, and T. Sandholm. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 295–304. ACM, 2007.

[2] S. Ahmed, A. Shapiro, and E. Shapiro. The sample average approximation method for stochastic programs with integer recourse. *SIAM Journal of Optimization*, 12:479–502, 2002.

[3] M. Ajtai, J. Komlós, and E. Szemerédi. The longest path in a random graph. *Combinatorica*, 1(1):1–12, 1981.

[4] M. Akbarpour, S. Li, and S. O. Gharan. Dynamic matching market design. *arXiv preprint arXiv:1402.3643*, 2014.

[5] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2004.

[6] I. Ashlagi and A. E. Roth. Free riding and participation in large scale, multi-hospital kidney exchange, 2011.

[7] I. Ashlagi, D. Gamarnik, M. A. Rees, and A. E. Roth. The need for (long) chains in kidney exchange. Technical report, National Bureau of Economic Research, 2012.

[8] I. Ashlagi, P. Jaillet, and V. H. Manshadi. Kidney exchange in dynamic sparse heterogenous pools. *arXiv preprint arXiv:1301.3509*, 2013.

[9] S. Asmussen. *Applied probability and queues*. Springer Verlag, 2003. ISBN 0387002111.

[10] N. Ayas, L. Barger, B. Cade, D. Hashimoto, B. Rosner, J. Cronin, F. Speizer, and C. Czeisler. Extended work duration and the risk of self-reported percutaneous injuries in interns. *JAMA: the journal of the American Medical Association*, 296 (9):1055, 2006.

[11] L. Barger, B. Cade, N. Ayas, J. Cronin, B. Rosner, F. Speizer, C. Czeisler, et al. Extended work shifts and the risk of motor vehicle crashes among interns. *New England Journal of Medicine*, 352(2):125, 2005.

[12] D. Bertsimas and R. Weismantel. *Optimization over integers*, volume 13. Dynamic Ideas Belmont, 2005.

[13] J.-F. Bérubé, M. Gendreau, and J.-Y. Potvin. A branch-and-cut algorithm for the undirected prize collecting traveling salesman problem. *Networks*, 54(1): 56–67, 2009.

[14] D. Bienstock, M. X. Goemans, D. Simchi-Levi, and D. Williamson. A note on the prize collecting traveling salesman problem. *Mathematical programming*, 59 (1):413–420, 1993.

[15] P. Billingsley. *Convergence of probability measures*. Wiley New York, 1968.

[16] P. Biró, D. F. Manlove, and R. Rizzi. Maximum weight cycle packing in directed graphs, with application to kidney exchange programs. *Discrete Mathematics, Algorithms and Applications*, 1(04):499–517, 2009.

[17] A. Blum, A. Gupta, A. Procaccia, and A. Sharma. Harnessing the power of two crossmatches. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 123–140. ACM, 2013.

[18] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Verlag, 2006. ISBN 0387295704.

[19] H. Chen and D. Yao. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*. Springer Verlag, 2001. ISBN 0387951660.

[20] M. Constantino, X. Klimentova, A. Viana, and A. Rais. New insights on integer-programming models for the kidney exchange problem. *European Journal of Operational Research*, 2013.

[21] R. Cook, M. Render, and D. Woods. Gaps in the continuity of care and progress on patient safety. *British Medical Journal*, 320(7237):791, 2000.

[22] F. de Véricourt and O. Jennings. Nurse-to-patient ratios in hospital staffing: a queuing perspective. *ESMT Working Paper No. 08-005*, 2008.

[23] J. P. Dickerson, A. D. Procaccia, and T. Sandholm. Optimizing kidney exchange with transplant chains: Theory and reality. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 711–718. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[24] J. P. Dickerson, A. D. Procaccia, and T. Sandholm. Failure-aware kidney exchange. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 323–340. ACM, 2013.

[25] R. Dittus, R. Klein, D. DeBrota, M. Dame, and J. Fitzgerald. Medical resident work schedules: design and evaluation by simulation modeling. *Management Science*, 42(6):891–906, 1996. ISSN 0025-1909.

[26] B. C. Drolet, L. B. Spalluto, and S. A. Fischer. Residents' perspectives on acgme regulation of supervision and duty hoursa national survey. *New England Journal of Medicine*, 363(23), 2010.

[27] B. C. Drolet, D. A. Christopher, and S. A. Fischer. Residents' response to duty-hour regulationsa follow-up national survey. *New England Journal of Medicine*, 366(24), 2012.

[28] R. Dudley. *Real analysis and probability*. Cambridge University Press, 2002.

[29] E. El-Darzi, C. Vasilakis, T. Chaussalet, and P. Millard. A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, 1(2):143–149, 1998. ISSN 1386-9620.

[30] P. I. Ellman, M. G. Law, C. Tache-Leon, T. B. Reece, T. S. Maxey, B. B. Peeler, J. A. Kern, C. G. Tribble, and I. L. Kron. Sleep deprivation does not affect operative results in cardiac surgery. *The Annals of thoracic surgery*, 78(3):906–911, 2004.

[31] S. Ethier and T. Kurtz. *Markov processes: Characterization and convergence*. Wiley New York, 1986.

[32] G. Fayolle, V. Malyshev, and M. Menshikov. *Topics in the constructive theory of countable Markov chains*. Cambridge University Press, 1995. ISBN 0521461979.

[33] M. Fischetti, J. J. Salazar Gonzalez, and P. Toth. A branch-and-cut algorithm for the symmetric generalized traveling salesman problem. *Operations Research*, 45(3):378–394, 1997.

[34] M. Fischetti, J. J. S. Gonzalez, and P. Toth. Solving the orienteering problem through branch-and-cut. *INFORMS Journal on Computing*, 10(2):133–148, 1998.

[35] K. E. Fletcher, W. Underwood, S. Q. Davis, R. S. Mangrulkar, L. F. McMahon, and S. Saint. Effects of work hour reduction on residents lives: a systematic review. *Jama*, 294(9):1088–1100, 2005.

[36] F. Foster et al. On the stochastic matrices associated with certain queuing processes. *The Annals of Mathematical Statistics*, 24(3):355–360, 1953.

[37] D. Gaba and S. Howard. Fatigue among clinicians and the safety of patients. *New England Journal of Medicine*, 347(16):1249–1255, 2002.

[38] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability*, 16(1):56–90, 2006. ISSN 1050-5164.

[39] M. Gendreau, G. Laporte, and F. Semet. A branch-and-cut algorithm for the undirected selective traveling salesman problem. *Networks*, 32(4):263–273, 1998.

[40] S. E. Gentry, R. A. Montgomery, B. J. Swihart, and D. L. Segev. The roles of dominos and nonsimultaneous chains in kidney paired donation. *American Journal of Transplantation*, 9(6):1330–1336, 2009.

[41] P. Glynn and A. Zeevi. Bounding stationary expectations of Markov processes. In *Markov processes and related topics: A Festschrift for Thomas G. Kurtz. Selected papers of the conference, Madison, WI, USA, July*, pages 10–13, 2006.

[42] M. X. Goemans. Combining approximation algorithms for the prize-collecting tsp. *arXiv preprint arXiv:0910.0553*, 2009.

[43] L. Green. Capacity planning and management in hospitals. *Operations Research and Health Care*, pages 15–41, 2005.

[44] L. Green, J. Soares, J. Giglio, and R. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006. ISSN 1553-2712.

[45] L. Green, P. Kolesar, and W. Whitt. Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. *Production and Operations Management*, 16(1):13–39, 2007. ISSN 1937-5956.

[46] G. Gutin and A. P. Punnen. *The traveling salesman problem and its variations*, volume 12. Springer, 2002.

[47] P. Holewijn and A. Hordijk. On the convergence of moments in stationary markov chains. *Stochastic Processes and Their Applications*, 3(1):55–64, 1975.

[48] M. Hutter, K. Kellogg, C. Ferguson, W. Abbott, and A. Warshaw. The impact of the 80-hour resident workweek on surgical residents and attending surgeons. *Annals of surgery*, 243(6):864, 2006.

[49] R. Ibrahim and W. Whitt. Wait-Time Predictors for Customer Service Systems With Time-Varying Demand and Capacity. *Oper. Res., forthcoming. Columbia University, NY, NY*, 2010.

[50] J. Iglehart. The ACGME's Final Duty-Hour Standards Special PGY-1 Limits and Strategic Napping. *New England Journal of Medicine*, 363(17):1589–1591, 2010.

[51] S. Jacobson, S. Hall, and J. Swisher. Discrete-event simulation of health care systems. *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 211–252, 2006.

[52] S. Janson, T. Luczak, and V. Kolchin. *Random graphs*. Cambridge Univ Press, 2000.

[53] R. Klein, R. Dittus, D. DeBrota, and M. Dame. Using discrete event simulation to evaluate housestaff work schedules. In *Proceedings of the 22nd conference on Winter simulation*, pages 738–742. IEEE Press, 1990. ISBN 0911801723.

[54] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[55] M. Krivelevich, E. Lubetzky, and B. Sudakov. Longest cycles in sparse random digraphs. *Random Structures & Algorithms*, 2012.

[56] J. Lamperti. Criteria for the recurrence or transience of stochastic process, part I. *J. Math. Anal. and Appls.*, 1, 1960.

[57] C. Landrigan, J. Rothschild, J. Cronin, R. Kaushal, E. Burdick, J. Katz, C. Lilly, P. Stone, S. Lockley, D. Bates, et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. *New England Journal of Medicine*, 351(18):1838, 2004.

[58] E. Litvak and I. Joint Commission Resources. *Managing Patient Flow in Hospitals: Strategies and Solutions*. Joint Commission Resources, 2010. ISBN 1599403722.

[59] Y. Liu and W. Whitt. A fluid model for a large-scale service system experiencing periods of overloading, 2010.

[60] A. Mandelbaum, W. Massey, and M. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30(1):149–201, 1998. ISSN 0257-0130.

[61] M. Molinaro and R. Ravi. Kidney exchanges and the query-commit problem, 2013.

[62] T. K. Nuckols and J. J. Escarce. Cost implications of acgmes 2011 changes to resident duty hours and the training environment. *Journal of general internal medicine*, 27(2):241–249, 2012.

[63] OPTN. Optn waiting list. `http://optn.transplant.hrsa.gov/latestData/rptData.asp`, 2013. Accessed: 2013-11.

[64] L. Petersen, T. Brennan, A. O'Neil, E. Cook, and T. Lee. Does housestaff discontinuity of care increase the risk for preventable adverse events? *Annals of internal medicine*, 121(11):866, 1994. ISSN 0003-4819.

[65] M. A. Rees, J. E. Kopke, R. P. Pelletier, D. L. Segev, M. E. Rutter, A. J. Fabrega, J. Rogers, O. G. Pankewycz, J. Hiller, A. E. Roth, et al. A nonsimultaneous, extended, altruistic-donor chain. *New England Journal of Medicine*, 360(11): 1096–1101, 2009.

[66] M. Rossetti, G. Trzcinski, and S. Syverud. Emergency department simulation and determination of optimal attending physician staffing schedules. In *wsc*, pages 1532–1540. IEEE, 1999.

[67] A. E. Roth, T. Sönmez, and M. U. Ünver. Kidney exchange. *Quarterly Journal of Economics*, 119:457–488, 2004.

[68] A. E. Roth, T. Sönmez, M. U. Ünver, F. L. Delmonico, and S. L. Saidman. Utilizing list exchange and nondirected donation through chain kidney paired donations. *American Journal of Transplantation*, 6:2694–2705, 2006.

[69] A. E. Roth, T. Sönmez, and M. U. Ünver. Efficient kidney exchange: Coincidence of wants in markets with compatibility-based preferences. *The American economic review*, pages 828–851, 2007.

[70] K. Sack. 60 Lives, 30 Kidneys, All Linked. *The New York Times*, 2012.

[71] K. Sack. Experts Recommend Single Registry to Oversee Kidney Transplant Donations. *The New York Times*, 2012.

[72] J. Samkoff and C. Jacques. A review of studies concerning effects of sleep deprivation and fatigue on residents' performance. *Academic Medicine*, 66(11):687, 1991. ISSN 1040-2446.

[73] E. Samuel-Cahn. Asymptotic distributions for occupancy and waiting time problems with positive probability of falling through the cells. *The Annals of Probability*, 2(3):515–521, 1974.

[74] T. Schoenmeyr, P. Dunn, D. Gamarnik, R. Levi, D. Berger, B. Daily, W. Levine, and W. Sandberg. A model for understanding the impacts of demand and capacity on waiting time to enter a congested recovery room. *Anesthesiology*, 110 (6):1293, 2009. ISSN 0003-3022.

[75] S. Sen, H. R. Kranzler, J. H. Krystal, H. Speller, G. Chan, J. Gelernter, and C. Guille. A prospective cohort study investigating factors associated with depression during medical internship. *Archives of general psychiatry*, 67(6):557–565, 2010.

[76] S. Sen, H. R. Kranzler, A. K. Didwania, A. C. Schwartz, S. Amarnath, J. C. Kolars, G. W. Dalack, B. Nichols, and C. Guille. Effects of the 2011 duty hour reforms on interns and their patients: a prospective longitudinal cohort study. *JAMA internal medicine*, 173(8):657–662, 2013.

[77] R. Steinbrook. The debate over residents' work hours. *New England Journal of Medicine*, 347(16):1296, 2002.

[78] C. Swamy and D. B. Shmoys. Sampling-based approximation algorithms for multi-stage stochastic optimization. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 357–366. IEEE, 2005.

[79] J. Tsitsiklis and K. Xu. On the Power of (even a little) Centralization in Distributed Processing. *ACM Sigmetrics, San Jose*, 2011.

[80] R. Tweedie. The existence of moments for stationary markov chains. *Journal of Applied Probability*, pages 191–196, 1983.

[81] D. Weinstein. Duty hours for resident physicians–tough choices for teaching hospitals. *New England Journal of Medicine*, 347(16):1275, 2002.

[82] E. Whang, M. Mello, S. Ashley, and M. Zinner. Implementing resident work hour limitations: lessons from the New York State experience. *Annals of surgery*, 237 (4):449, 2003.

[83] K. White Jr. A survey of data resources for simulating patient flows in healthcare delivery systems. In *Proceedings of the 37th conference on Winter simulation*, pages 926–935. Winter Simulation Conference, 2005. ISBN 0780395190.

[84] A. Yaghoubian, A. H. Kaji, B. Ishaque, J. Park, D. K. Rosing, S. Lee, B. E. Stabile, and C. de Virgilio. Acute care surgery performed by sleep deprived residents: are outcomes affected? *Journal of Surgical Research*, 163(2):192–196, 2010.

[85] G. Yom-Tov. Queues in Hospitals: Semi-Open Queueing Networks in the QED Regime. *IE PhD thesis proposal*, 2008.

[86] G. Yom-Tov and A. Mandelbaum. Time-varying QED queues with reentrant customers in support of healthcare staffing. *The Technion, Haifa, Israel*, 2010.