Project Summary

CAREER: Designing Efficient and Equitable Omni-channel Service Systems

Sherwin Doroudi

Overview

Every day, millions of people wait in lines for a variety of services such as food preparation, government processing, and medical procedures. Visits to service systems often involve multiple stages of service, but a new service paradigm allows some users to use technology to self-process their request (e.g., through a mobile ordering and payment application) and skip waiting in line for the first stage of service; e.g., a long queue at the cashier of a busy coffee shop can be bypassed by using the appropriate app.

This PI's research plan seeks to use—and develop novel—queueing-theoretic methodology to study how to best design such **omni-channel service systems**. The key system design questions the PI seeks to answer in any given setting are: (i) when should self-processing opportunities (e.g., mobile apps) be offered at all, (ii) which class of users (e.g., walk-ins) should be prioritized over others at any given time, and (iii) what level of delay information should be disclosed to users. The last question is especially important in settings where impatient users will opt out of waiting for service if they anticipate lengthy delays. These designs will be evaluated in terms of **efficiency** (in terms of average waiting time and throughput) and **equitability** (in terms of the waiting time and throughput experience by each user class). The proposed research consists of two research thrusts: the first will focus on the exact analysis of simpler two-class two-stage models, while the second builds upon the first in examining more complicated networks via a combination of exact analysis, approximation techniques, and simulation.

Intellectual Merit

In investigating these research thrusts, the PI must analyze performance of a variety of Markovian queueing models with rational users (strategic customers). These models present **challenges that preclude exact analysis via preexisting techniques**, necessitating the **development of new queueing-theoretic methodology**. The proposal outlines why these challenges emerge and shows how the PI plans to leverage his prior expertise in both the analysis of queues with strategic behavior and the analysis of large-scale Markov chains—for which he has published two novel methods—to tackle these challenges.

Broader Impacts

By considering equitability, the PI's research plan aims not only to design omni-channel systems that are better for society overall, but also systems whose efficiency does not come at the cost of detrimental treatment to particular user classes. Many common system designs strongly favor mobile users, and will thus indirectly harm lower-income and disabled users for whom mobile applications are often less accessible. **Awareness of equitability is crucial in mitigating the effects of unintended discriminatory treatment.**

The PI's education plan seeks to involve undergraduate students—particularly female students (who represent almost 50% of Industrial and Systems Engineering majors at the University of Minnesota), when possible—in research tied to this project.

1 Overview and Objectives

Every day, millions of people wait in line for services ranging from the preparation of food and beverages in restaurants and coffee shops to the processing of paperwork in government facilities to the execution of medical evaluations in clinics. Service systems frequently involve multiple stages of service; for example, a customer in a coffee shop waits in line to place and pay for their order, and subsequently waits for their order to be prepared. Furthermore, contemporary technology enables some users to experience some stages differently, or skip them altogether, via **"self-processing opportunities."** For example, many coffee shops have developed free **mobile ordering applications**. A customer who has downloaded the application and provided their payment information can make a selection and pay for their coffee (i.e., "self process"), effectively skipping the first line; they need only wait for their order to be prepared. It is often the case that the same staff preparing orders for these **"mobile customers"** is also taking orders from (and preparing orders for) **"walk-in customers"**. Customers in the latter group do not have the application or choose not to use it, and therefore must first wait to submit their order as described above.

Systems that simultaneously serve two or more **classes** of users via different **channels** (typically one physical and online) by leveraging the same service capacity are known as **omni-channel service systems**. As these systems are new, the formal mathematical modeling and analysis of the performance and design of omni-channel service systems is in its infancy.

The proposed long-term research initiative seeks to use queueing-theoretic modeling and analysis with rational user (strategic customer) behavior to understand how to best design efficient and equitable omni-channel service systems. The key system design levers under consideration in this proposal are (i) the question of when self-processing opportunities should be offered in the first place (e.g., should a mobile application be made available?)—my preliminary work suggests that an omni-channel structure is not always beneficial—, (ii) **priority scheduling** (i.e., at any given time, which user class(es) should be prioritized), and (iii) **information disclosure** (i.e., what, if any, real-time estimated delay information should be disclosed to each user class). To a lesser extent, staffing decisions are also under consideration in the context of multi-server systems. System designs will be evaluated in terms of **efficiency** and **equitability**.

Two primary measures of efficiency will be considered. First, designs should seek to minimize the **overall expected waiting time**. They should also seek to maximize **throughput**, the rate at which users are ultimately served. When assuming rational user behavior, users of any kind may opt out of using the service entirely when they anticipate waiting times in excess of their patience thresholds, contributing to a loss of throughput; note that both the priority scheduling and information disclosure design levers can effect the delays anticipated by users.

The proposed project also seeks to develop an understanding of which system designs are equitable. The waiting time and throughput measures both take into account the overall user experience, but they can also be examined individually across each user class (e.g., walk-in vs. mobile customers). Specifically, the proposed work will study the **trade-offs** inherent in helping one class over another via **Pareto analysis**. Understanding equitability is crucial as **simple service designs can indirectly discriminate against populations to whom such applications are inaccessible** (see broader impacts in Section 7). The proposed work will explore how system design can be used to mitigate the effects of such discrimination.

The proposed research activities comprise two primary thrusts. The first thrust involves exploring—primarily through **exact analysis**—the simplest non-trivial family of stochastic Markovian queueing models within the omni-channel service space: **two-class two-stage models**. As the name suggests, such models include two user classes (e.g., the aforementioned walk-in and mobile classes) and two stages of service, with one class skipping the first stage via self-processing. Preliminary findings are promising that such models allow for **computing exact values** of the efficiency measures of interest across a variety of assumptions and system designs. Under rational user behavior (whereby impatient users opt not to wait for service should expected delays exceed their patience threshold), these models become difficult to analyze. **The study of such systems requires pushing the state of the art in the analysis of Markov chains**. Unlike much of the work on queueing with rational users that primarily contribute to advancing game-theoretic and/or optimization techniques rather than those in queueing analysis, **the proposed work will not ignore complexities that emerge in settings with more than one user class and station**.

The second thrust involves exploring—through a combination of exact analysis, simulation, and approximation techniques—a richer family of stochastic queueing **network** models. In such models, I allow for a greater number of user classes and service stations and a variety of (possibly probabilistic) routes taken through these service stations. These models capture the complexities of multi-station government and healthcare facilities, where visitors or patients may need to be served at a number of different stations and the path taken throughout the service facility is only realized over time based on findings from earlier services (e.g., a health-assessment conducted by a triage nurse).

2 PI Qualifications

I am uniquely qualified to carry out the research plan in this proposal because it allows me to combine my expertise in two different areas of queueing theory: (i) the modeling and analysis of queues with rational user (strategic customer) behavior and (ii) the development of novel methods for the analysis of large-scale Markov chains. This synthesis is not only necessary to undertake the timely project proposed here, but also brings together two streams of my past research—which were until now largely separate—enabling me to develop a long-term foundation for a productive research career and leadership role in the intersection of these two areas.

This research plan fits squarely within the domain of rational or strategic queueing (also known as queueing games). Within this area of study, I have published peer-reviewed work [25, 26] characterizing an explicit threshold value given only implicitly in Naor's 1969 paper [68]—the pioneering work in this field—and I have published work on strategic servers in a workshop paper [38] and in the prestigious journal, *Operations Research* [50]. I have also published a technical report on prioritization in queues [36], which is related to one of the key design levers in the proposed research plan.

The bulk of the published works in the strategic queueing literature that rely on exact analysis (for surveys see [55, 53])—including my own contributions—examine game-theoretic and optimization problems built upon simple queueing model. While the problems themselves can be very challenging, the queueing models lying at the heart of these problems are often tractably analyzed via straightforward techniques. My preliminary work on the models proposed here, however, suggests the need for a combination of steady-state and transient queueing analysis alongside **quasi-birth–death process Markov chain analysis**, which often requires tailor-made techniques. I am uniquely suited to this challenge as I have developed and published two such techniques: Recursive Renewal Reward [41, 42] and Clearing Analysis on Phases [37]. Furthermore, I have published work on load-balancing [39] and redundancy [47, 48] involving the analysis of non-trivial Markov chains.

3 Background and Motivation

In many settings, the dominant paradigm for unscheduled service requests remains the traditional one: users walk into a service facility and wait in a **physical queue** until a server can fulfill their request. An emergent alternative paradigm sees users placing their request online—through web and mobile applications—and waiting in a **virtual queue** until their request is fulfilled remotely, as in the case of some online visa application services. A variant of this app-based paradigm is seen in online ride-hailing services (e.g., Uber and Lyft), where the rider's request is placed in a virtual queue (for a typically short duration of time), until they can be assigned a server (driver) that is summoned to pick them up.

The emergent app-based paradigm offers a couple of advantages over its traditional counterpart. First, waiting in a virtual queue can free up the user's time while waiting. Most importantly, in app-based systems the user is—with the help of the application—**self-processing** some of the tasks that would normally require fixed service capacity (e.g., a human server). Moreover, such self-processing tasks can occur in parallel, as well-developed applications can handle many concurrent users at low resource costs. Reduced service requirements allow for the design of systems with less service capacity and/or lower waiting times.

Given the advantages associated with app-based systems, one would expect them to gradually replace traditional service systems across many settings. While this transition is arguably underway, it will likely occur over a long period of time. In the meantime we are increasingly seeing systems that are hybrids of the traditional and emergent app-based systems. Such **omni-channel service systems** (described in Section 1) are the focus of this proposal; they use shared service capacity to serve users in both physical and virtual queues. The omni-channel paradigm described here should not be confused with a superficially similar paradigm that has existed for decades—one where a restaurant servers walk-in customers and call-in orders—as call-in orders are not placed via self-processing; they require processing by staff who must take these orders over the phone rather than attending to other operational tasks.

3.1 Examples of omni-channel service systems

In recent years, many national brands have been launching their own online ordering and payment applications, essentially moving from traditional to omni-channel service systems overnight. Starbucks launched their **mobile-order-and-pay application** in 2014 [10, 65, 71]; other food and beverage brands launching similar apps include Dunkin' Donuts (2016) [4], McDonald's (2017) [29, 71], Chipotle (2009; new app in 2017) [3, 2], Chick-fil-A (2017) [85], and KFC Canada (2018) [72]. Usage of such mobile applications has been growing steadily: Starbucks reports the fraction of transactions conducted via its mobile-order-and-pay application grew from 4% in the second quarter of 2016 to 14% in the fourth quarter of 2018 at a steady rate of 1% per quarter [12, 46]. McDonald's also reports increaed adoption [46], while Dunkin' Donuts plans to open 50 "NextGen" restaurants—that can better serve the needs of mobile customers—by the end of 2019 [86, 87, 64].

It is worth noting that despite the numerous advantages of self-processing applications in the food and beverage industry, their introduction has also introduced complications [11, 18, 78, 79]. In particular, there have been reports of "long lines that are being exacerbated by an uptick in mobile ordering... [causing] customers to walk out" at Starbucks [11, 78, 79], which is a key feature of potential user behavior in the models I propose to study. Such phenomena also illustrate the need for prioritization among customer classes—one of the major design levers in the proposed research plan—in order to mitigate throughput loss due to unsatisfied customers.

Omni-channel service systems are also emerging in the public sector. An online queueing system called "QLess" received significant attention when "The Michigan Secretary of State announced a $1.1 million QLess contract that will eventually deploy the service in 10 of its offices" [88]. This system allows users to join a queue online before arriving at the office. The Houston Permitting Center offers a similar service [8]. Another example announced government mobile app is Mobile Passport which is designed to speed passengers at customs across roughly two dozen airports. Qualified travelers are able to upload their passport information and submit customs declaration forms via their mobile devices in advance [6]. Other travelers have to resort to being processed by human staff or using a limited number of kiosks; self-operated kiosks should not be viewed as self-processing, since travelers must queue to use one.

There have been rapid recent developments in the use of mobile health-care applications, (known generally as mHealth). A recent survey reports "83 percent of physicians in the U.S. already use mobile health technology or mHealth to provide patient care" [13], suggesting great potential in future healthcare service delivery. One mHealth app is an artificially intelligent medical advisor that helps potential patients run self-diagnosis by exchanging messages via a mobile app, with the developers claiming 92% accuracy in the app's diagnostic evaluations [67]. If the patient's body is behaving atypically, the app will reportedly notify the hospital upon arrival, effectively bypassing triage with automated referral to the appropriate doctor; this effectively transforms the healthcare facility into an omni-channel service system. Nonetheless in order for such mHealth apps to be effective, certain accessibility barriers must be overcome [66]; the proposed work seeks to design systems so that in a world where some patients are benefiting from mHealth, those patients without such access can still obtain equitable outcomes.

For companies or service providers who are incapable of developing their own mobile apps but still seek the opportunity to take advantage of omni-channel service, the company skiplino offers a platform that is advertised as a user-friendly way to launch a mobile channel alongside a preexisting traditional channel. Their goal is to use cloud-based smart queue management system to help businesses and service providers to improve overall customer experience [9]. Despite rapidly growing interest in omni-channel service systems in both the public and private sectors, the efficient and equitable design of such systems remains an open problem that the research proposed here seeks to address.

3.2 Prior work on omni-channel service (and retail) systems

Very little analytic academic work has been dedicated to omni-channel service systems. A related area, omni-channel **retail** systems, has received considerably greater attention [28, 20, 32, 43, 19, 40, 44, 21, 60]. Research in the area of omni-channel retail tends to be methodologically distinct from that on omni-channel service. Problems in the former area revolve around inventory and product management and are typically studied using inventory control theory. Meanwhile, problems in the latter area concern service capacity management, which are better addressed via queueing theory [45].

The published work most closely related to the work proposed herein (and to the best of my knowledge, the only analytic work on omni-channel service systems) is Gao and Su's 2018 paper in *Management Science* [45]. This paper investigates the high-level impact of self-processing technologies on customer (user) demand, capacity decisions, revenues. While—like the proposed work in Thrust 1—Gao and Su model an omni-channel service system as a tandem queue, they opt for a stylized representation of user behavior that ignores responses to current queue-length information. This modeling choice greatly simplifies the resulting queueing model, precluding the need for sophisticated and novel queueing analysis—the development of which is one of the goals of this proposal. The main design level considered by Gao and Su is the problem of optimal staffing. Unlike the work proposed here, they do nor consider alternative priority schemes or information disclosure.

Under their stylized model, Gao and Su find that self-order technologies not only reduce waiting times for those customers who opt to use the mobile service (as expected) but could also improve the waiting times of the walk-in customers, especially when customers are highly sensitive to waiting times. My preliminary work shows that more detailed modeling of user response behavior shows that self-processing can actually reduce overall throughput.

The model most closely resembling those proposed here is found in a recent 2019 working paper by Wang et.al. [83]. While their model differs significantly from mine in its details and describes an application that is unrelated to omni-channel service there are several notable similarities with the work I propose in Thurst 1. First, it features two stages of service. Further, their model includes rational user behavior, wherein users can opt out of waiting when anticipating lengthy delays. Finally, their model requires non-trivial Markovian analysis, which they conduct by extending the Recursive Renewal Reward technique that I developed with my co-authors [41, 42].

4 Research Plan — Thrust 1: Two-Class Two-Stage Models

Research Thrust 1 is focused on the exact analysis of a simple, yet rich and difficult to analyze family of omni-channel service systems, and the impact of the design levers on the efficiency and equitability of these systems. This research thrust will be the primary focus of the proposed research plans in Years 1–2, although these investigations will likely continue throughout the entire duration of the project, and form a foundation for the analytic portion of Thrust 2.

4.1 Description of base models for Thurst 1

In describing the model, we assume the setting of a coffee shop with a mobile order-and-pay application.

Arrivals and delay. Walk-in customers (walk-ins) arrive to a physical queue according to a Poisson Process with rate λ_1—where (if they join) they wait for service in the **order-and-pay (OaP) stage**. Once service at this phase is completed (i.e., the customer has placed their order and paid for it), the customer proceeds to the **preparation stage** and waits for their order to be prepared. Mobile customers (mobiles) arrive according to a Poisson Process with rate λ_2 and immediately proceed to the preparation phase. The waiting time or **delay** of a customer, represented by the random variable T in general and T_w (T_m) for walk-ins (mobiles), is their **end-to-end time** in the system from arrival until ultimate service. It is assumed that mobile orders are placed instantaneously and a mobile "arrives" as soon as they place their order; variant models resulting from relaxing one or both of these assumptions are also of potential interest.

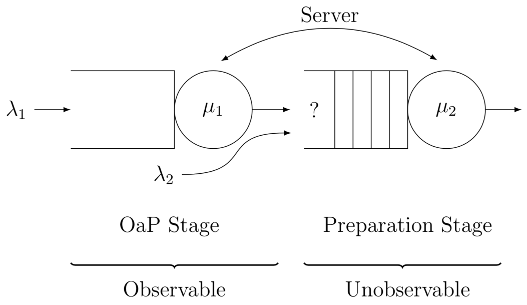

Figure 1: The two-class two-stage coffee shop model where a single server (a floater) can move from serving one queue to another. From the perspective of a walk-in arrival only the queue length at the OaP stage is observable; neither stage is visible to a mobile arrival.

Servers. In general, there are n_1 servers that serve requests at rate μ_1 in the OaP stage (cashiers), n_2 servers that serve requests at rate μ_2 in the preparation stage (baristas), and n_0 servers that can flexibly and instantaneously move between these two roles (floaters), serving OaP and preparation requests at the same rates, μ_1 and μ_2, respectively. The time required to serve a request is exponentially distributed. The **prioritization** design lever involves deciding which class the baristas (and floaters at the preparation stage) are serving and which queue the floaters are focusing on at any given time.

Joining and balking. Upon arrival, a customer joins the queue if their **anticipated delay**—given the observed state of the system—is less than their **patience threshold**, $T^{\max}$, and balk otherwise. In reality, anticipated delay should be based on historical experience, but for modeling simplicity we use the expected waiting time given the observed state information **in equilibrium** (see [55]), i.e., $\mathbb{E}[T_w|\mathbf{I}]$ ($\mathbb{E}[T_m|\mathbf{I}]$) for a walk-in (mobile) that observes information **I** upon arriving. Equilibrium joining behavior can involve mixed strategies (e.g., a mobile customer might join with probability 75%) and my preliminary work has identified settings with non-unique equilibria.

Information structure. This model is challenging to analyze—and departs significantly from models studied in the prior work—in the structure of the information, **I**, observed by arrivals: **information observed depends on class and is incomplete**. Denote the system state by (i, j), meaning that there are i customers in the OaP stage's physical queue (including those in service) and j customers in the preparation stage's virtual queue. By default, when a walk-in visits the coffee shop, they observe the OaP stage's physical queue, and hence know the value of i; since the preparation queue is virtual—mobiles and walk-ins who have already placed their order could be waiting anywhere, even outside the coffee shop—a walk-in arrival does not know the value of j. The system from the walk-in's perspective is shown in Figure 1. On the other hand, when a mobile places their order, if we are to assume they are not at the coffee shop, they observe no information at all. These default assumptions can be modified through the use of the **information disclosure** design lever.

Even with the default information structure designed above, it is not true that the walk-in customers are entirely oblivious of the value of j when they enter the system; this is because i **and** j **are correlated**, and uncovering this correlation structure typically involves determining the limiting probability of a quasi-birth–death process Markov chain. Numerical analysis of such chains is possible via matrix analytic methods [69, 62], which can in some situations yield exact solutions [80]. The Recursive Renewal Reward method [41, 42] (which I developed with my coauthors) when appropriately—and significantly—adapted, has proven useful in tackling this correlation structure in my preliminary work on the proposed research.

Impatience. The proposed work considers three different forms of customer patience thresholds. (1) customers are fully patient, i.e., $T^{\max=\infty}$, (2) all have the same patience level $T^{\max}$, or (3) have different patience levels (i.e., $T^{\max}$ can be viewed as a random variable drawn form a given distribution). Moreover, walk-ins and mobiles need not be equally patience; we use $T_w^{\max}$ and $T_m^{\max}$ for walk-ins and mobiles, respectively.

4.2 Prior Work on Rational Queueing in Relation to Thrust 1

Following Naor's pioneering work [68] on joining behavior and observable queues, much work has been devoted to combining game-theoretic and queueing-theoretic analysis to understand queue-joining behaviors [90, 30, 51, 81, 33, 57, 59, 89]. Some work in this area, like the proposed work, allows customers to observe only partial queue-length information upon arrival [35, 61, 75, 56]. D'Auria and Kanta [35] consider strategic behavior in a two-phase (extended to multi-phase by Kim and Kim [61]) setting where arrivals decide to join based on the total number of customers in the system, without knowing which phases(s) others are in. While it may appear to resemble the proposed models, D'Auria and Kanta's setting does not allow one class of customers to bypass the first phase, and the "unobservable" variables they consider are both bounded, which precludes the need to analyze large-scale chains. The unique features of the proposed models make them technically challenging to analyze in ways that have not been addressed by the literature.

4.3 Preliminary work on Thrust 1

Infinite patience. In order to better understand the impact of the **prioritization design lever** on both efficiency and equitability in the two-stage two-system models, I investigated the case when both user classes are fully patient (i.e., $T^{\max=\infty}$) and there is a single flexible floater server $n_0 = 1, n_1 = n_2 = 0$. The resulting model can be viewed as a **priority polling model** (see [63, 82, 23, 76, 77, 74, 24] on polling), where the flexible server must decide which queue (stage) and user class to prioritize at any given time.

When customers are fully patient, all will join the system, so the system is throughput optimal so long as it is stable. Specifically, I aim to understand how the waiting time of each class is affected by prioritizing one class over another. Instead of focusing on minimizing the overall mean waiting time, $\mathbb{E}[T]$, I examine the set of policies that are **Pareto efficient** with respect to the pair of class-specific expected delays, $(\mathbb{E}[T_m], \mathbb{E}[T_w])$; i.e., those policies where the only way to help one class further would necessitate hurting the other.

Through a combination of $M/G/1$ busy-period analysis (see [52]) and an adaptation of the achievable regions method [22, 34], my preliminary work shows that Pareto efficient scheduling policies are one of three policies, and "randomized mixtures" of these policies (i.e., whenever the system is empty of users, we randomly choose between one of these three policies according to a set distribution).

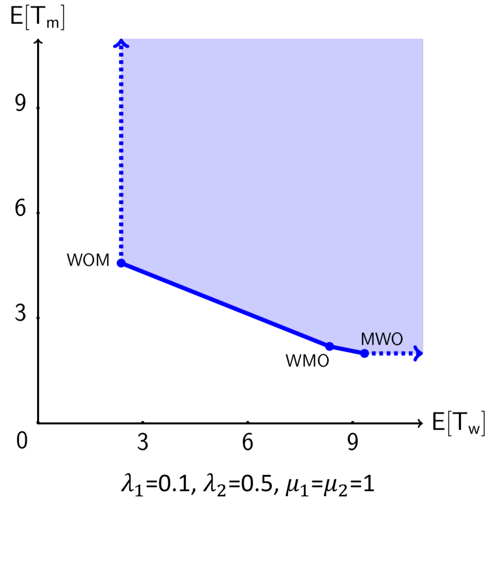

Figure 2: The Pareto frontier of delays with feasible region shaded.

To describe these policies, it will help to categorize all requests into three classes: mobile orders at the preparation stage (M), walk-in customers at the OaP stage (O), and orders from walk-in customers at the preparation stage (W). If we let **XYZ** mean the policy where X is prioritized over Y and Y over Z, the three Pareto efficient policies are **MWO** (where mobile customers are always given priority), **WMO** (where the preparation stage is prioritized, with walk-ins there being prioritized over mobile customers), and **WOM** (where walk-ins are always prioritized). Consistent with the queueing paradigm that one should serve requests closest to completion, W is always prioritized over O. Here we consider **preemptive priority**, meaning that requests will be interrupted (with any progress saved) if a higher priority arrives; our analysis can also accommodate non-preemptive priority. We note that **MWO**, **WMO**, and random mixtures of the two are also optimal with respect to overall delay, $\mathbf{E}[T]$, but if we want to favor walk-ins any more than they are favored under **WMO** (which may be the case depending on equitability constraints), overall efficiency will suffer. We can find closed-form expressions for the mean delays associated with each class under each policy; an example which shows the Pareto frontier of delays and the associated feasible region are shown in Figure 2.

Finite Patience. Next, I examined the impact of impatience, by considering customers with finite patience thresholds (i.e., $T_w^{\max}, T_m^{\max} < \infty$), again in the case of a single flexible floater server ($n_0 = 1, n_1 = n_2 = 0$). Since impatient customers can balk, this is a setting where overall throughput is a key measure of efficiency and class-specific throughput values can be used to assess equitability. In the preliminary work, attention was restricted to the three prioritization policies that were found to be Pareto-optimal in the infinite patience case (i.e., **MWO**, **WMO**, and **WOM**).

This system is not trivial to analyze; in fact, the question of whether $\mathbb{E}[T_w|i]$ is increasing in i—which almost certainly appears to be true as one expects a longer wait when one arrives to a longer queue—remains open. Using a combination of quasi-birth–death process analysis (in order to resolve the aforementioned correlation problem) and busy period analysis, I obtained exact results for the class-based and aggregate efficiency metrics of interest in the case of constant patience threshold values, $T_w^{\max}$ and $T_m^{\max}$. If the thresholds are heterogeneous, an iterative approach that I developed approximates the desired metrics with minimal error as long as the distribution of thresholds has finite support.

In order to examine the impact of the first two design levers (i.e., the questions of whether a mobile application should be offered and how customers should be prioritized), we consider what happens if a traditional coffee shop with only walk-in customers begins to offer a mobile application, which a certain fraction of its existing customers adopt. Figure 3 is a plot (for a particular parameter set) of the loss rate as a function of this fraction—the adoption rate—where the loss rate is the fraction of customers that balk, i.e., $1 - X/(\lambda_1 + \lambda_2)$ where X is the throughput. Other parameter sets I examined tended to show qualitatively similar trends.

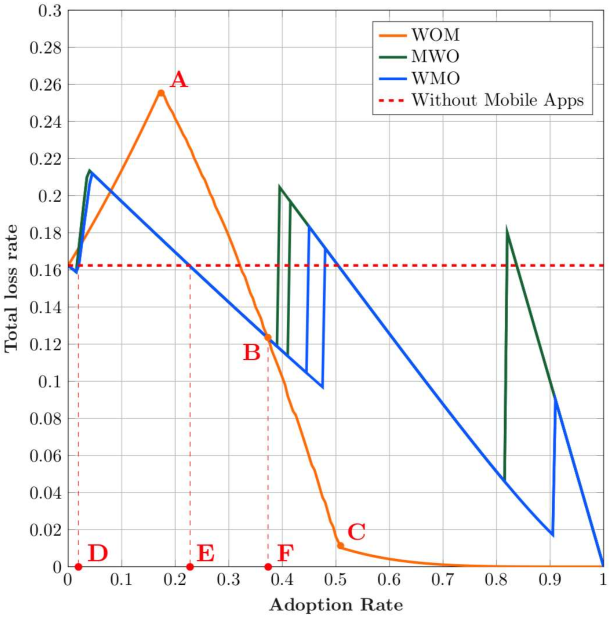

Figure 3: Loss rate as a function of adoption rate where $\lambda_1 + \lambda_2 = 0.05$, $\mu_1 = 0.16$ and $\mu_2 = 0.08$ $T_w^{\max=65}$, $T_m^{\max=100}$.

From the plot, we see that the **MWO** and **WMO** perform similar to one another; in fact, while **WMO** weakly outperforms **MWO**, they coincide **exactly** at the majority of adoption rates due to the discrete nature of queue lengths. Generally, mobile adoption reduces the need for service and improves overall efficiency; however, increased adoption can trigger a discrete "jump" due to a shift in the walk-in queue length threshold for joining. At some adoption rates there are multiple curves associated with the same policy; this is due to equilibria multiplicity.

The **WOM** policy aggressively favors walk-ins, so mobiles do not join until there is sufficient adoption at point **A** reducing the number of walk-ins, which allows a fraction of mobiles to begin joining (they use a mixed strategy). At point **B**, this policy begins to outperform the others (with respect to loss rate) and by point **C** and beyond, all mobiles are now joining the system as fewer walk-ins are interrupting their service.

Most noteworthy is the fact that up to point **D** there is a very modest benefit to offering the app and from point **D** to **E** all three policies are outperformed by the no-app benchmark. Hence, the omni-channel structure is not always beneficial, even though it reduces overall service requirements, justifying our first design lever. This is especially surprising as the mobile users are more patient in this particular parameter set, which would make adoption seem even more positive.

4.4 Proposed work on Thrust 1

The work in this thrust will consist of five primary exploratory tasks, each designed to explore a different feature of this family of models and/or design levers. These tasks form a sequence of problems organized in ascending order of anticipated difficulty, although with the exception of Task 5, they can be pursued in parallel or in any order.

Task 1: analysis of multi-server systems. The first proposed task is to move beyond the $n_0 = 1, n_1 = n_2 = 0$ case in the preliminary work. While considering a single floater allows one to explore the effects of prioritization where they are likely to be strongest, many omni-channel systems feature multiple servers, not all of which are flexible.

The first alternative service structure to be investigated is the case of one cashier and one barista, i.e., $n_0 = 0, n_1 = n_2 = 1$. It is important to note that in this setting prioritization plays a much smaller role as only the barista can make such decisions. Even this seemingly simple case introduces significant analytic challenges, as simultaneous service in both stages complicates the system dynamics. As in our preliminary work, walk-in customers observe the length of the OaP queue i, and then must infer a distribution on the preparation queue, j. A walk-in customer would experience an expected delay of $\mathbb{E}[T_w] = (i+1)/\mu_1 + (j'+1)/\mu_2$, where j' is not the inferred value of j, but rather the value j is expected to take by the time the walk-in customer reaches the preparation phase. Determining j' as a function of i and j requires transient—rather than steady-state—queueing analysis, while determining the distribution of j implied by i requires steady-state analysis. By blending the two approaches, I posit that it is possible to analyze the performance of the system in terms of delay and throughput, at least in the case of constant patience thresholds. An important question that emerges is whether offering the mobile application can still hinder efficiency.

Once the model above has been understood, I propose to investigate other multi-server settings, including those with a mix of flexible and inflexible servers. Understanding different multi-server settings is essential in addressing system design questions where staffing is also a design lever. Specifically, I propose to investigate the advantage flexible servers offer over inflexible servers.

Task 2: prioritization optimization. The second proposed task in this thrust returns to the prioritization design lever, again in the case where there is a single floater. While our preliminary work established that three static priority policies and their mixtures represent the entire set of Pareto-efficient options for prioritization in the case of patient customers, there is no reason to believe this result continues to hold when customers are impatient. That is, there is potentially room for improvement in terms of efficiency, by making the prioritization rule a function of the current joint queue length state, (i, j). This task calls for developing a methodology to find—or characterize the properties of—the optimal dynamic prioritization rule in terms of throughput maximization, or more generally any weighted combination of class-specific delays and throughputs. This task will likely necessitate viewing the prioritization problem as a Markov decision process.

Task 3: analysis of rational abandonment. Balking is not the only rational behavior exhibited by users of queueing systems. Users may renege, i.e., **abandon** the queue, if they revise their estimate of delay to be greater than first anticipated. The literature on rational abandonment [54, 49, 27, 73, 14] typically considers situations where the server may have suffered a break-down, which users eventually guess is the case if they have been waiting in the queue for a long time, leading them to abandon. In our setting, rational abandonment can occur if a walk-in customer waiting in the OaP comes to believe—due to the time that has elapsed—that the unknown queue length of the preparation phase is sufficiently long that they should leave. The assumption is that customers will abandon the queue if they rationally believe that the **remaining delay** they will face exceeds their patience threshold. This task calls for studying the metrics of interest in the presence of this phenomenon, at least in the single floater case.

Of course, each customer in the OaP queue has a different initial estimate on the length of the preparation queue *j* based on what they observed upon entry. However, if one customer reneges then as long as patience thresholds are constant, so should any customers waiting behind the reneging customer in the queue, because earlier arrivals are more informed. Under heterogeneous arrivals, customers waiting behind a reneging customer must update their beliefs on *i* and renege if appropriate.

The partially observable nature of these omni-channel models create a very rich and challenging environment for the study of rational reneging. It may however be the case that the expected value of the preparation queue length, *i*, always decreases in the time a customer spends in the OaP queue, so that the scenarios described above never take place. Addressing this question is a logical first step in addressing this task. If this is indeed the case, I propose to analyze this phenomenon by considering non-exponential service distributions.

Task 4: analysis of information disclosure. This task will focus on the **information disclosure** design lever that was left unexplored in the preliminary work. Although a system designer cannot prevent walk-ins from observing the physical OaP queue length, *i*, they can be informed of the virtual preparation queue length, *j* (e.g., through a digital sign board), similarly mobile customers can be informed of the value of *j* via the app. The goal is to determine whether information disclosure can sometimes lead to more efficient or equitable systems. Of course one could also reveal partial information about *j*.

Disclosing information to mobile customers can actually make analysis easier, even if messy, as under rational queueing behavior it would lead to both queues being bounded, allowing for finite-state Markov chain analysis. Information revealed to walk-in customers, however, introduces a new complication, whereby the threshold which determines whether they join the queue—if it exists—is now two dimensional in nature, as it would depend on both i and j. I am optimistic that in at least some cases, the underlying Markov chain would not be significantly different from those explored in the preliminary work, except over a finite number of states. However, understanding the structural changes in that finite portion in a systematic way may prove challenging.

This task is connected to the literature on **delay announcements** [17, 70], with a crucial twist in that the information revealed can be class-dependent. I intend to collaborate with Dr. Mohammad Delasay—an assistant Professor at Stony Brook University with prior experience on delay announcements—on this task.

Task 5: exploration of further models. The last task in Thrust 1 involves exploring the richer space of models that essentially combine elements of the models explored in the previous tasks, e.g., how should information be disclosed in the presence of rational abandonment or what is the best way to prioritize when one has multiple servers available?

5 Research Plan — Thrust 2: Omni-channel Networks

Thrust 2 will build upon the first research thrust by examining networks with more than two classes and/or stages along with potentially class-dependent probabilistic routing. This thrust is more ambitious and open-ended than the first thrust and will be the focus of Years 3–5, although we anticipate that work on Thrust 1 will continue throughout the duration of the project.

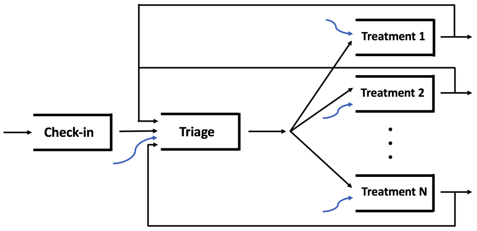

Figure 4: An omni-channel network model of a healthcare clinic.

The models considered in this thrust are general networks with general information structures. As an example, consider the three-class healthcare clinic model depicted in Figure 4. Walk-in patients must wait to check in, then wait for triage, and finally wait for treatment at the appropriate level. Meanwhile, patients using a queue-skipping app can notify the clinic of their arrival, bypassing check-in, and those patients with a self-triaging app can additionally bypass triage. Patients with non-critical conditions may balk if they anticipate a long delay based on the congestion information made visible to them upon arrival. Furthermore, due to the possibility of mis-triage or changing conditions, some patients must be re-triaged even if they originally bypassed triage.

Task 6: Generalizing findings from Thrust 1. The first task associated with this research thrust is to explore analytically (and if possible, via exact analysis), which findings obtained from Thrust 1 apply in a network context and to what extent. There will also be an exploration of how methodology developed in Thrust 1 can be adapted to the more general network setting. While the curse of dimensionality may prove prohibitive in some cases, there are special cases, such as when at most one queue can be unbounded, that should be amenable to exact analysis.

Task 7: Studying non-stationary demand. Realistic omni-channel service systems, like most service systems, will typically feature time-varying demand [58, 84]. Specifically, demand for each channel (i.e., the arrival rate for each user class) should vary according to a non-homogeneous Poisson process, that can be further treated as “fluid” if certain assumptions hold (see [31]). This task will examine how such a system can be analyzed through fluid approximation techniques. One challenge associated with studying non-stationary demand in this setting is that new arrivals must estimate unobservable queue lengths based on both observable information and past arrival patterns.

Task 8: Developing a simulation tool. The proposal for Thrust 2 also involves the development of a simulation tool for use by both practitioners and other researchers. In order to ensure reasonable runtimes, the simulation tool will be integrated with analytic results and will incorporate the features that are addressed through exact analysis (Task 6) and fluid approximation techniques (Task 7). The simulation will then be used to simulate omni-channel networks in order to gain greater insights regarding which designs perform best. Once sufficiently polished, the simulation tool will be freely shared with the public for use by researchers and practitioners alike.

6 Education Plan — Integration with Research

The proposed research, if successful, will provide content that can be integrated into two existing courses (one each at the undergraduate and graduate levels) and a new course I plan to develop in the near future (all at the University of Minnesota). Additionally, there will be opportunities for undergraduate students (particularly female students where possible) to become involved in research.

My teaching at the PhD level will train University of Minnesota PhD students in techniques that will equip them to solve problems involving strategic queueing. This ensures that my own future graduate students are qualified to carry out the research proposed here, and serves as an avenue for recruiting interested new students to join the proposed project. My goal is to create a synergistic cycle whereby my research will improve my teaching and vice-versa.

Undergraduate Teaching. I have taught an undergraduate course on **Production and Inventory Control** (IE 4551) at the University of Minnesota over the past three Spring semesters; I plan to continue teaching this course regularly. I assign articles on emerging and contemporary aspects of inventory, supply chain, and service systems throughout the semester to provide a qualitative supplement to the mostly quantitative material featured in the course. In the coming years, I plan to assign one carefully selected reading assignment on omni-channel service systems. Eventually **I plan to write my own short article on the subject of omni-channel service systems for a general audience, sharing my high-level research findings**, which I would then assign to the class.

Students also work on group projects that are carried out in a few weeks. While students are free to pick their own topics, I will encourage topics in the omni-channel domain. The experience of reading and conducting projects on omni-channel service systems will help introduce students to the area, allowing them to determine if they would be willing to pursue an undergraduate research project in the area. *Evaluation plan:* The success of the introduction of new teaching materials will be measured via anonymous course surveys, in line with my usual procedure of soliciting continuous feedback from undergraduate students. In addition, interest in undergraduate research resulting from these new materials will be informally assessed.

Undergraduate Research. In the 2017–2018 academic year I was a senior thesis advisor for a female undergraduate student whose research on "A Study of a Two-Server Problem with Varied Customer Priority" served as very preliminary work in studying omni-channel service systems via discrete event simulations. The project was a learning opportunity for her to gain first-hand experience with research and for me to better understand how to mentor undergraduate students and find exciting problems for them that are suited to their level of experience.

My work in general, and the work proposed here in particular, typically requires experience with analytic skills that undergraduate students are very unlikely to have been trained in. However, as most of the models I focus on (including those in Thrust 1 of this proposal) are often simple to describe and understand—even if difficult to analyze—I have found that coding, running, and analyzing the outcomes of simple stochastic simulations is within the skill set of a typical Industrial Engineering undergraduate student at the University of Minnesota. Therefore, I propose to recruit one undergraduate student per year to pursue an undergraduate research opportunity and/or write a senior thesis under my mentorship. I will also budget to have one undergraduate student attend a national conference (either the IISE Annual Conference or INFORMS Annual Meeting) each year, even if the student will not be presenting work at the conference. Attending will give the students greater exposure to both the industrial engineering community in both academia and industry, and potentially help them with determining their future plans. *Evaluation Plan:* The success of these undergraduate research initiatives will be tracked by monitoring the placement of involved students (in graduate programs and/or industry positions) in the years following graduation.

Graduate Teaching and Research. I have taught **Stochastic Processes and Queueing Systems** and designed and taught a very successful follow-up elective, **Modeling & Analysis of Queuing Systems.** When I next teach the latter course, I plan to include any new methodological developments resulting from the proposed research. This course also helped me recruit two PhD students, and ensured that all three of my students were technically competent in queueing theory; this has been invaluable to the early development of my career.

I also plan to develop and teach a new PhD course on **queueing games**. Although my courses cover queueing theory and several existing courses at the university cover game theory, no course covers the rich intersection of these areas. Therefore, there is a need for such a course, especially for students working in the area of **revenue management** and those studying problems with strategic behavior and delays through the University of Minnesota's NSF-funded **Initiative on the Sharing Economy** [5]. This course would allow me to share the models described in this proposal along with any potential research findings; it would help in recruiting interested students to join my research group and become involved in the proposed research. *Evaluation plan:* Surveys will be used about two years after students have completed the new course to determine the relevance of the material to the students' graduate careers. PhD student mentoring will be evaluated by tracking student milestones within the program (completion of the qualifying and preliminary exams and publication of papers) and eventual career placements upon graduation.

7 Broader Impacts

The proposed research and integrated education plan will work toward meeting the following Broader Impacts goals, which are outlined in the NSF CAREER solicitation [7].

Improved well-being of individuals in society: The omni-channel system designs I am proposing to study will be evaluated in terms of both **efficiency** and **equitability**. By considering equitability alongside efficiency, we ensure that we are not merely improving system design in terms of how it is experienced by the aggregate population, but also ensuring—to the extent possible—that the overall efficiency does not come at the cost of prohibiting the access of or imposing excessive delays on one user class.

Designing systems with equitability in mind is especially important when we consider that **lower-income users often lack access to the technology and resources that facilitate self-processing** (e.g., mobile phones, data plans, credit cards, etc.) and **users with disabilities may be unable to use applications lacking in accessibility options**. Indeed, recent surveys on the digital divide have shown that lower-income and disabled Americans are less likely to own smartphones [15, 16]. Moreover, simple service designs (such as those featuring first-come-first-serve scheduling) tend to allow a form of "line skipping" that favors mobile users, effectively allowing for indirect—even if unintended—discrimination toward lower income and/or disabled users. The proposed work seeks to identify when and **how such detrimental treatment of walk-in users can be mitigated** without imposing significant negative externalities on mobile users. Awareness of such concerns is especially important for government and healthcare facilities. *Evaluation Plan:* Findings will be disseminated through academic publications, and also to system design practitioners. This includes making simulation tools that measure projected system equitability available to practitioners and tracking their use in government and industry.

Full participation of women in STEM: In 2018, nearly 50% of incoming students in Industrial & Systems Engineering were women, one of the highest percentages in the College of Science and Engineering [1]. I plan to serve as a research mentor for one talented undergraduate student in each year, who will be working on simulation problems related to this proposal (see Section 6). I will give preference to students who are women when possible, continuing my recent trend of mentoring students who are women (one undergraduate student in academic year 2017–2018, one senior PhD student in Summer 2018, and one visiting masters student intern in Fall 2018). *Evaluation Plan:* Involvement of women in the research project—and their career outcomes—will be tracked by monitoring their eventual career and further study placements.

Increased partnerships between academia, industry, and others: I have had the opportunity to collaborate with **Microsoft Research** in the past, resulting in several publications [25, 26, 36]. I am currently engaged in research-oriented industry collaborations with the third party logistics company **C.H. Robinson** and the international private public transportation provider **TransDev**. One of my long-term career goals is to eventually reach out to potential public or private sector collaborators in an attempt to use our findings to inform real-world system design. *Evaluation Plan:* The success of such collaborations will be evaluated on the publication of papers with industry collaborators.

I have not received any prior support from NSF as a PI or co-PI.

References

[1] 2018 University of Minnesota Industrial & Systems Engineering Magazine. http://www.isye.umn.edu/magazine/pdf. Accessed: 2019-07-14.

[2] Chipotle Mexican Grill launches iPhone app. https://www.mediapost.com/publications/article/112233/chipotle-mexican-grill-launches-iphone-app.html. Accessed: 2019-07-13.

[3] Chipotles New Mobile App Makes Ordering Real Food On-The-Go Even Easier. https://ir.chipotle.com/news-releases?item=122385. Accessed: 2019-07-13.

[4] Dunkin Donuts On-the-Go Mobile Ordering Now Available Nationwide. https://news.dunkindonuts.com/news/dunkin-donuts-on-the-go-mobile-ordering-now-available-nationwide. Accessed: 2019-07-13.

[5] Initiative on the Sharing Economy. http://sharingeconomy.umn.edu/. Accessed: 2019-07-14.

[6] Mobile Passport. https://mobilepassport.us/. Accessed: 2019-07-14.

[7] NSF Faculty Early Career Development Program (CAREER). Program Solicitation NSF 17-537. https://www.nsf.gov/pubs/2017/nsf17537/nsf17537.htm. Accessed: 2019-07-14.

[8] Online Queuing Service. https://www.hpceservices.org/node/118. Accessed: 2019-07-14.

[9] Skiplino: Queue Management System. https://skiplino.com/. Accessed: 2019-07-14.

[10] Starbucks Company Timeline. https://www.starbucks.com/about-us/company-information/starbucks-company-timeline. Accessed: 2019-07-13.

[11] Starbucks mobile ordering is working too well. https://www.forbes.com/sites/retailwire/2017/04/12/starbucks-mobile-ordering-is-working-too-well/#6e03818cea28. Accessed: 2019-07-14.

[12] Starbucks Stories & News: Press. https://stories.starbucks.com/press. Accessed: 2019-07-13.

[13] What is mobile health technology? https://www.athenahealth.com/knowledge-hub/mobile-health-technology/what-is-mobile-health-technology. Accessed: 2019-07-14.

[14] P. Afèche and V. Sarhangian. Rational abandonment from priority queues: Equilibrium strategy and pricing implications. *Working paper*, 2015. SSRN.

[15] M. Anderson. Digital divide persists even as lower-income Americans make gains in tech adoption. *Pew Research Center*, 2019.

[16] M. Anderson and A. Perrin. Disabled Americans are less likely to use technology. *Pew Research Center*, 2017.

[17] M. Armony, N. Shimkin, and W. Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, 57(1):66–81, 2009.

[18] L. Baertlein. Starbucks' mobile order push meets resistance from ritual seekers. https://www.reuters.com/article/us-starbucks-mobileorder/starbucks-mobile-order-push-meets-resistance-from-ritual-seekers-idUSKBN1GX0KA. Accessed: 2019-07-14.

[19] A. Bayram and B. Cesaret. Ship-from-store operations in omni-channel retailing. In *IIE Annual Conference. Proceedings*, pages 1181–1186. Institute of Industrial and Systems Engineers (IISE), 2017.

[20] D. R. Bell, S. Gallino, and A. Moreno. How to win in an omnichannel world. *MIT Sloan Management Review*, 56(1):45–53, 2014.

[21] D. R. Bell, S. Gallino, and A. Moreno. Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Science*, 64(4):1629–1651, 2018.

[22] D. Bertsimas. The achievable region method in the optimal control of queueing systems; formulations, bounds and policies. *Queueing Systems*, 21:337, 1995.

[23] M. A. A. Boon, R. D. van der Mei, and E. M. M. Winands. Applications of polling systems. *Surveys in Operations Research and Management Science*, 16:67–82, 2011.

[24] M. A. A. Boon, R. D. van der Mei, and E. M. M. Winands. Waiting times in queueing networks with a single shared server. *Queueing Systems*, 74:403–429, 2013.

[25] C. Borgs, J.T. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu. Pricing and queueing. *ACM SIGMETRICS Performance Evaluation Review*, 40(3):71–73, 2012.

[26] C. Borgs, J.T. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu. The optimal admission threshold in observable queues with state dependent pricing. *Probability in the Engineering and Informational Sciences*, 28(1):101–119, 2014.

[27] A. Brandt and M. Brandt. On the two-class m/m/1 system under preemptive resume and impatience of the prioritized customers. *Queueing Systems*, 47(2):147–168, 2004.

[28] E. Brynjolfsson, Y. J. Hu, and M. S. Rahman. Competing in the age of omnichannel retailing. *MIT Sloan Management Review*, 54(4):23–29, 2013.

[29] A. Carman. McDonald's is now testing mobile ordering in the US. https://www.theverge.com/2017/3/15/14933638/mcdonalds-mobile-app-order-ahead. Accessed: 2019-07-13.

[30] H. Chen and M.Z. Frank. State dependent pricing with a queue. *IIE Transactions*, 33(10):847–860, 2001.

[31] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag, New York, 2001.

[32] S. Chopra. How omni-channel can be the future of retailing. *Decision*, 43(2):135–144, 2016.

[33] S. Cui and S. Veeraraghavan. Blind queues: The impact of consumer beliefs on revenues and congestion. *Management Science*, 62(12):3656–3672, 2016.

[34] M. Dacre, K. Glazebrook, and J. Nio-Mora. The achievable region approach to the optimal control of stochastic systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 61(4):747–791, 1999.

[35] B. D'Auria and S. Kanta. Pure threshold strategies for a two-node tandem network under partial information. *Operations Research Letters*, 43:467–470, 2015.

[36] S. Doroudi, M. Akan, M. Harchol-Balter, J. Karp, C. Borgs, and J.T. Chayes. Priority pricing in queues with a continuous distribution of customer valuations. *Technical report CM-CS-13-109, Computer Science Department, Carnegie Mellon University*, 2013.

[37] S. Doroudi, B. Fralix, and M. Harchol-Balter. Clearing analysis on phases: Exact limiting probabilities for skip-free, unidirectional, quasi-birth-death processes. *Stochastic Systems*, 6(2):420–458, 2017.

[38] S. Doroudi, R. Gopalakrishnan, and A. Wierman. Dispatching to incentivize fast service in multi-server queues. *Performance Evaluation Review*, 39(3):43–45, 2011.

[39] S. Doroudi, E. Hyytiä, and M. Harchol-Balter. Value driven load balancing. *Performance Evaluation*, 79:306–327, 2014.

[40] S. Gallino, A. Moreno, and I. Stamatopoulos. Channel integration, sales dispersion, and inventory management. *Management Science*, 63(9):2813–2831, 2017.

[41] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. In *ACM SIGMETRICS Performance Evaluation Review*, volume 41, pages 153–166. ACM, 2013.

[42] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Queueing Systems*, 77(2):177–209, 2014.

[43] F. Gao and X. Su. Omnichannel retail operations with buy-online-and-pick-up-in-store. *Management Science*, 63(8):2478–2492, 2016.

[44] F. Gao and X. Su. Online and offline information for omnichannel retailing. *Manufacturing & Service Operations Management*, 19(1):84, 98 2017.

[45] F. Gao and X. Su. Service operations with online and offline self-order technologies. *Management Science*, 64(8):3595–3608, 2018.

[46] T. Garcia. McDonalds, like Starbucks, sees opportunities in mobile-order-and-pay. https://www.marketwatch.com/story/mcdonalds-like-starbucks-sees-opportunities-in-mobile-order-and-pay-2018-10-24. Accessed: 2019-07-13.

[47] K. Gardner, S. Zbarsky, S. Doroudi, M. Harchol-Balter, E. Hyytiä, and Scheller-Wolf A. Reducing latency via redundant requests: Exact analysis. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):347–360, 2015.

[48] K. Gardner, S. Zbarsky, S. Doroudi, M. Harchol-Balter, E. Hyytiä, and Scheller-Wolf A. Queueing with redundant requests: exact analysis. *Queueing Systems*, 83(3-4):227–259, 2016.

[49] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

[50] R. Gopalakrishnan, S. Doroudi, A.R. Ward, and A. Wierman. Routing and staffing when servers are strategic. *Operations Research*, 64(4):1033–1050, 2016.

[51] P. Guo and P. Zipkin. Analysis and comparison of queues with different levels of delay information. *Management Science*, 53(6):962–970, 2007.

[52] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.

[53] R. Hassin. *Rational Queueing*. CRC Press, 2016.

[54] R. Hassin and M. Haviv. Equilibrium strategies for queues with impatient customers. *Operations Research Letters*, 17(1):41–45, 1995.

[55] R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, USA, 2003.

[56] R. Hassin and R. Roet-Green. The armchair decision: On queue-length information when customers travel to a queue. *Working paper*, 2018. SSRN.

[57] M. Hu, L. Yang, and J. Wang. Effcient ignorance: Information heterogeneity in a queue. *Management Science*, 2017. Articles in Advance.

[58] O. B. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996.

[59] C. Jin, L. Debo, and S. Iravani. Observational learning in large-scale congested service systems. *Working paper*, 2017.

[60] M. Jin, G. Li, and T.C.E. Cheng. Buy online and pick up in-store: Design of the service area. *European Journal of Operational Research*, 268(2):613–623, 2018.

[61] B. Kim and J. Kim. Equilibrium strategies for a tandem network under partial information. *Operations Research Letters*, 44:532–534, 2016.

[62] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, Philadelphia, 1999.

[63] H. Levy and M. Sidi. Polling systems : Applications, modeling, and optimization. *IEEE Transactions on Commmunications*, 38:1750–1760, 1990.

[64] I. Liffreing. How Dunkin' Donuts turns to voice to boost mobile orders. https://digiday.com/marketing/dunkin-donuts-plans-boost-mobile-orders/. Accessed: 2019-07-13.

[65] Molla, R. Starbuckss mobile payments system is so popular in the U.S., it has more users than Apples or Googles: New estimates predict that it will stay that way. https://www.vox.com/2018/5/22/17377234/starbucks-mobile-payments-users-apple-pay-google. Accessed: 2019-07-13.

[66] D. Muoio. AI triage tools won't empower consumers without mastering convenient, personalized service. https://www.mobihealthnews.com/content/ai-triage-tools-wont-empower-consumers-without-mastering-convenient-personalized-service. Accessed: 2019-07-14.

[67] M. Murgia. How smartphones are transforming healthcare. https://www.ft.com/content/1efb95ba-d852-11e6-944b-e7eb37a6aa8e. Accessed: 2019-07-14.

[68] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, 1969.

[69] M.F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Dover Publications, 1981.

[70] S. Nirenberg, A. Daw, and J. Pender. The impact of queue length rounding and delayed app information on disney world queues. In *2018 Winter Simulation Conference*, pages 3849–3860, 2018.

[71] R. Pucci. Mobile Order and Pay Competition: Starbucks vs. McDonalds. https://www.paymentsjournal.com/mobile-order-pay-competition/. Accessed: 2019-07-13.

[72] B. Shankar. KFC adds mobile ordering to its Android and iOS app. https://mobilesyrup.com/2018/11/19/kfc-adds-mobile-orders-android-ios-app/. Accessed: 2019-07-13.

[73] N. Shimkin and A. Mandelbaum. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems*, 47:117–146, 2004.

[74] M. Sidi, H. Levy, and S. W. Fuhrmann. A queuing network with a single cyclically roving server. *Queueing Systems*, 11:121–144, 1992.

[75] D. F. Silva, B. Zhang, and H. Ayhan. Admission control strategies for tandem markovian loss systems. *Queueing Systems*, 90:35, 2018.

[76] H. Takagi. *Analysis of Polling Systems*. Cambridge: MIT Press, 1986.

[77] T. Takine, H. Takagi, and T. Hasegawa. Sojourn times in vacation and polling systems with bernoulli feedback. *Journal of Applied Probability*, 28:422–432, 1991.

[78] K. Taylor. One of Starbucks' biggest strengths is becoming a huge problem for the chain. https://www.businessinsider.com/starbucks-mobile-ordering-problems-2017-1. Accessed: 2019-07-14.

[79] K. Taylor. We went to Starbucks every day for a week to see if the coffee giant has fixed an annoying problem. https://www.businessinsider.com/starbucks-attempts-to-fix-mobile-ordering-review-2017-8. Accessed: 2019-07-14.

[80] B. Van Houdt and J.S.H. van Leeuwaarden. Triangular M/G/1-Type and Tree-Like Quasi-Birth-Death Markov Chains. *INFORMS Journal on Computing*, 23(1):165–171, 2011.

[81] S. Veeraraghavan and L. Debo. Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management*, 11(4):543–562, 2009.

[82] V. M. Vishnevskii and O. V. Semenova. Mathematical methods to study the polling systems. *Automation and Remote Control*, 67:173–220, 2006.

[83] J. Wang, H. Abouee Mehrizi, O. Baron, and O. Berman. Staffing tandem queues with impatient customers–application in financial service operations. *Rotman School of Management Working Paper*, (3116815), 2019.

[84] W. Whitt. Time-varying queues. *Queueing Models and Service Management*, 1(2):79–164, 2018.

[85] R. Williams. Chick-fil-A boosts conversions with in-app payment update. https://www.mobilemarketer.com/news/chick-fil-a-boosts-conversions-with-in-app-payment-update/509852/. Accessed: 2019-07-13.

[86] R. Williams. Dunkin' Brands expands blueprint for mobile ordering, loyalty growth. https://www.mobilemarketer.com/news/dunkin-brands-expands-blueprint-for-mobile-ordering-loyalty-growth/516743/. Accessed: 2019-07-13.

[87] R. Williams. Dunkin' plans to drive visibility to mobile ordering after Q1 growth. https://www.mobilemarketer.com/news/dunkin-plans-to-drive-visibility-to-mobile-ordering-after-q1-growth/522293/. Accessed: 2019-07-13.

[88] C. Wood. Online queue system brings rave reviews for the DMV. https://www.govtech.com/gov-experience/Online-Queue-System-Brings-Rave-Reviews-for-the-DMV.html. Accessed: 2019-07-14.

[89] L. Yang and L. Debo. Referral priority program: Leveraging social ties via operational incentives. *Management Science*, 65(5):2231–2248, 2018.

[90] U. Yechiali. On optimal balking rules and toll charges in the GI/M/1 queuing process. *Operations Research*, 19(2):349–370, 1971.

BIOGRAPHICAL SKETCH — Sherwin Doroudi

(a) Professional Preparation

Undergraduate Institution(s)	Location	Major	Degree & Year
California Institute of Technology	Pasadena, CA	Economics	B.S., 2011

Graduate Institution(s)	Location	Major	Degree & Year
Carnegie Mellon University	Pittsburgh, PA	IA (OM)*	M.S., 2012
Carnegie Mellon University	Pittsburgh, PA	IA (OM)*	Ph.D., 2016

* IA (OM) is short for "Industrial Administration (Operations Management)"

(b) Appointments

Industrial and Systems Engineering Department University of Minnesota Assistant Professor	2016–Present

(c) Publications

(i) Publications most closely related to proposed project

- S. Doroudi, B. Fralix, and M. Harchol-Balter. Clearing analysis on phases: Exact limiting probabilities for skip-free, unidirectional, quasi-birth-death processes. *Stochastic Systems*, 6(2):420–458, 2016.** https://doi.org/10.1109/5.771073

- R. Gopalakrishnan, S. Doroudi, A.R. Ward, and A. Wierman. Routing and staffing when servers are strategic. *Operations Research*, 64(4):1033–1050, 2016. https://doi.org/10.1287/opre.2016.1506

- C. Borgs, J.T. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu. The optimal admission threshold in observable queues with state dependent pricing. *Probability in the Engineering and Informational Sciences*, 28(1):101–119, 2014. https://doi.org/10.1017/S0269964813000351

- A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Queueing Systems*, 77(2):177–209, 2014. *** https://doi.org/10.1007/s11134-014-9409-7

- A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. In *ACM SIGMETRICS Performance Evaluation Review*, volume 41, pages 153–166. ACM, 2013. *** https://doi.org/10.1145/2494232.2465760

** Accepted and published in 2017 for an issue dated 2016.
*** *Queueing Systems* invited the authors to submit an expanded version of the *SIGMETRICS* paper, which was accepted for publication.

(ii) Other significant publications

- K. Gardner, S. Zbarsky, S. Doroudi, M. Harchol-Balter, E. Hyytiä, and A. Scheller-Wolf. Queueing with redundant requests: exact analysis. *Queueing Systems*, 83(3-4):227–259, 2016.*** https://doi.org/10.1007/s11134-016-9485-y

- K. Gardner, S. Zbarsky, S. Doroudi, M. Harchol-Balter, E. Hyytiä, and A. Scheller-Wolf. Reducing latency via redundant requests: Exact analysis. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):347–360, 2015. https://doi.org/10.1145/2745844.2745873

- S. Doroudi, E. Hyytiä, and M. Harchol-Balter. Value driven load balancing. *Performance Evaluation*, 79:306–327, 2014. https://doi.org/10.1016/j.peva.2014.07.019

- C. Borgs, J.T. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu. Pricing and queueing. *ACM SIGMETRICS Performance Evaluation Review*, 40(3):71–73, 2012. https://doi.org/10.1145/2425248.2425266

- S. Doroudi, R. Gopalakrishnan, and A. Wierman. Dispatching to incentivize fast service in multi-server queues. *ACM SIGMETRICS Performance Evaluation Review*, 39(3):43–45, 2011. https://doi.org/10.1145/2160803.2160855

*** *Queueing Systems* invited the authors to submit an expanded version of the *SIGMETRICS* paper, which was accepted for publication.

(d) Synergistic Activities

- **2019 INFORMS Applied Probability Society Conference Program Committee Member.** Organized and moderated an invited session on "Analysis of Computing and Service Systems."

- **2019 IISE Conference Session Organizer**

- **UMN ISyE Analytics Seminar Coordinator** (2018–2019). Strengthened existing and developed new connections between UMN's ISyE group and analytics experts in industry, by inviting them to give talks on campus aimed at masters students.

- Developed and published the **Clearing Analysis on Phases (CAP)** method for analyzing Markov chains (2017).

- Developed and published **Recursive Renewal Reward (RRR)** method for analyzing Markov chains (2013) that has been used by other researchers.

BUDGET JUSTIFICATION

Section A: Senior Personnel

The request is for 1 month of PI Dr. Doroudi's summer salary for the duration of the grant. This is based on a 9 month salary, and adjusted for 3% inflation annually. The PI will be conducting researching, supervising students, and writing to communicate results.

Section C: Fringe Benefits

Fringe is calculated at 36% of PI Dr. Doroudi's requested salary.

Section E: Travel

Travel funds are being requested to travel to the INFORMS Annual Meeting during each year, and the INFORMS Applied Probability Society Conference (which is held every two years) during Years 1, 3, and 5. The INFORMS Annual Meeting is domestic, whereas the INFORMS APS Conference is expected to be international in Years 1 and 5 and domestic in year 3. The PI plans to travel to these conferences with one graduate student in Years 1–2 and with two graduate students in Years 3–5. Based on past experience the expense of travel to a domestic conference is estimated at \$1800 for the PI and \$800 for students. For an international conference the estimates are \$3000 and \$2000 respectively. Lower registration costs and the opportunity to apply for travel grants make student travel less expensive than PI travel. The plan is for the PI and all traveling students to each deliver at least one presentation on the proposed research at these conferences, with the goal of disseminating recent research findings.

An additional \$800 per year is being request to fund student travel for one undergraduate student, in order to allow undergraduate researchers to become involved with the greater research and industry communities even if they will not be presenting research at the conference. Undergraduate students will be encouraged to either attend the IISE Annual Conference or the INFORMS Annual meeting. There are many opportunities for student involvement at these conferences. Undergraduate students will be encouraged to volunteer at the conference, join and participate in student chapters, and submit work in progress for presentation in contributed sessions.

Requested travel funds were calculated on the basis of the above as follows:

- Year 1: $\$1800 + \$800 + \$800 = \$3400$ domestic
and $\$3000 + \$2000 = \$5000$ international.

- Year 2: $\$1800 + \$800 + \$800 = \$3400$ domestic.

- Year 3: $\$1800 + \$800 + \$800 + \$800 + \$1800 + \$800 + \$800 = \$7600$ domestic.

- Year 4: $\$1800 + \$800 + \$800 + \$800 = \$4200$ domestic.

- Year 5: $\$1800 + \$800 + \$800 + \$800 = \$4200$ domestic and $\$3000 + \$2000 + \$2000 = \$7000$ international.

Student Costs

The request is to fund 1 Ph.D. student for one academic semester and one summer semester as a 50% (full time) research assistant in each of Years 1–2 and 2 Ph.D. students for one academic semester and one summer semester as 50% (full time) research assistants in each of Years 3–5. A student salary in Year 1 is $18,426 ($11,080 for the academic semester and $7,307 for the summer semester), adjusted 3% annually for inflation thereafter. Student fringe, which includes tuition (in the academic semester) and health benefits (in both the academic and summer semesters) ranges from $11,003–$11,325 per year throughout the 5-year period.

In years 1–2 the PI and one graduate student will be investigating Research Thrust 1 focusing on the exact analysis of two-class two-stage omni-channel service models, and the development of methodological techniques toward that aim.

In years 3–5, the PI and two graduate students will be investigating Research Thrust 2 focusing on combining exact analysis, approximations, and simulations to study omni-channel service networks. One student will continue to develop the ideas in Research Thurst 1 and will attempt to generalize some of these to Research Thurst 2. The other student will primarily focus on developing a robust simulation tool for modeling omni-channel service networks. Both students will make use of approximation techniques such as fluid analysis.

Undergraduate student researchers will not be compensated with funds awarded from this proposal, and will instead apply for funds through the Undergraduate Research Opportunities Program at the University of Minnesota and/or will be earn course credit for conducting research.

Indirect Costs

Indirect costs are calculated at 54% of modified direct total costs (MDTC).

CURRENT AND PENDING SUPPORT - PI Dr. Doroudi

Pending Support

CAREER: Designing Efficient and Equitable Omni-channel Service Systems
[THIS PROPOSAL]

- **Source of Support:** NSF
- **Total Award Amount:** $520,175 requested
- **Total Award Period Covered:** 9/1/2020–8/31/2025
- **Location of Project:** Minneapolis, MN
- **Person Months Per Year Committed to Project:** 1 (Summer)

Submission Planned in Near Future

LEAP-HI: On-Demand Transit: Contracting, Planning and Performance Evaluation

- **Source of Support:** NSF
- **Total Award Amount:** Request to be determined (estimated at ≈ $1,300,00)
- **Total Award Period Covered:** dates to be determined (a 3-year duration)
- **Location of Project:** Minneapolis, MN
- **Person Months Per Year Committed to Project:** 1 (Summer)

Current Support

University of Minnesota College of Science & Engineering New Faculty Startup Funds

- **Source of Support:** University of Minnesota
- **Total Award Amount:** $350,000 (with no indirect costs)
- **Total Award Period Covered:** 8/29/2016–indefinite (but 2/3 of funds in excess of 15% of the original $350,000 will be forfeited after 9/1/2020)
- **Location of Project:** Minneapolis, MN
- **Person Months Per Year Committed to Project:** No commitment, can be used flexibly for summer salary.

FACILITIES, EQUIPMENT AND OTHER RESOURCES

Facilities

This proposed project does not require any special facilities. Graduate students in the University of Minnesota's Industrial & Systems Engineering Department no longer work in lab spaces. The department provides shared office spaces to all students (currently in the Shepherd Labs building), so no special space is needed for students to work. The PI has a faculty office in Lind Hall.

Equipment

This project does not require any special equipment. Although this project will require computing resources, it is not anticipated that there will be any need for high performance computing. Therefore, the standard computing resources that individuals already have access to (personal computers, shared computers provided by the University of Minnesota's computer lab, and the PI's office computers) should be sufficient to carry out the proposed activities. University of Minnesota faculty, staff, and students also have access to resources such as the Virtual Online Linuix Environment, where they can remotely run computational tasks, as needed, so long as they comply with the data management plan.

Other Resources

Dr. Mohammad Delasay, Assistant Professor at Stony Brook University's College of Business, intends to collaborate with the PI (and students working under the PI) on some of the research proposed under Thurst 1. He brings expertise in the area of queueing systems with delay announcements, making him an ideal collaborator on the proposed project. Dr. Delasay will not be paid by the PI, the University of Minnesota, or any funds that may be awarded as a result of this proposal. Dr. Delasay may be a co-author on one or more of the products produced as a result of this proposed research.

DATA MANAGEMENT PLAN

Types of data

The data associated with the proposed research will be generated by one of two modes. The first mode is running computer code to find exact numerical values or approximations associated with the relevant system measurements (e.g., mean waiting time, throughput, etc.) under different input parameters and system designs. These will be based on formulas that will be derived and algorithms that will be developed as a result of the proposed research. The second mode is generated by running discrete event stochastic simulations (again, under different input parameters and system designs). The desired summary statistics related to the relevant system measurements are then calculated on the basis of each simulation run's results. Where appropriate, tables and graphs will be created on the basis of the data described above.

The data for this project will include the code that is written (for exact calculations, approximations, and simulations), input parameters and system design choices, outputs (measurements), and the data files generated from simulation runs. The project will not be using any preexisting existing data nor will any data be collected apart from the data generation procedures described above. When possible, however, input parameters may be chosen to reflect values that deemed "realistic" by the research team. Moreover, no human and animal subjects are involved and there will be no sensitive data associated with this project that will warrant any high data security measures.

Data and metadata standards

Data will be generated by the PI, graduate and undergraduate students working with the PI on the project, and where appropriate, external collaborators. During the course of the project all computer code used and data generated will be saved in the University of Minnesota Google Team Drive with access limited to only the research group. This is in accordance with the recommend procedures provided by the University of Minnesota Data Management Guide. Moreover, all generated data files will be logged in a Google Sheet tabulating the following information: file name, google folder location/link, date created, person responsible, related code, problem parameters used, description of results, and other comments. The Google Team Drive will be managed by the PI.

When publishing any work using any of the data associated with this project, all relevant files for the work to be published, including computer code, generated data sets, and graphs will be stored in the Data Repository of the University of Minnesota (DRUM). DRUM allows University of Minnesota researchers to publish research data openly as well as allow research material for access by request. Moreover DRUM allows classification of the different data files during storage.

Policies for access and sharing and provisions for appropriate protection/privacy

The results of the research will be shared through tables and figures presented in publications. Raw data will be made available to researchers on request using DRUM. Any simulation tools developed for researchers and practitioners will be listed on the PI's academic website and shared as open source software through DRUM.

Policies and provisions for re-use, re-distribution

Other researchers may use the data, results, and tools produced through the proposed research as long as they cite the relevant work. As mentioned above, if a simulation tool is created for use by other researchers and practitioners, it will be shared with the greater research and practitioner communities through DRUM.

Plans for archiving and preservation of access

By default, DRUM allows for data access for a duration of 10 years. Longer preservation times can be made by request.