Research Statement

My research interests lie in the areas of applied probability, with a particular focus on operations management, queueing models and approximations. My other research interests include structural models under financial interconnectedness and online ranking algorithms via probabilistic approach.

Below, I first provide some background on my main research area. I then briefly describe my current research projects and outline some directions of future research.

**Background**

My research deals with multi-server queues that arise in a wide variety of application contexts, including customer contact centers, communication networks, health care and inventory systems. The commonality in all of the above applications is that analyzing such mathematical models helps to understand system performance, such as the steady-state distribution of the number of customers in system and the probability that an arriving customer waits in queue. Exact analysis of these systems is often intractable, but valuable insights can be gained via appropriate asymptotic approximations. The idea is to find a simple and analytically tractable diffusion process that matches the queueing system after proper scaling.

The limit regimes that one often thinks of are the conventional heavy-traffic (CHT) limit regime and the Halfin-Whitt (HW) limit regime. In the CHT regime, the arrival and per-server service rates are scaled up with the number of servers being fixed, whereas in the HW regime, the arrival rate and the number of servers grow large with the service rate staying fixed. The conceptual difference between these two regimes is that a new arrival is almost certain to be delayed in the CHT regime, while the probability of initially being delayed is strictly between zero and one in the HW regime. There is an intermediate limit regime that is often called the non-degenerate slow-down regime where both the service rate and the number of servers grow large as the arrival rate becomes large. Both fluid and diffusion limits for the total number of customers in system can easily be derived with exponential services. However, there is a great deal of motivation to study multi-server queues with general service-time distributions, stemming from applications such as call centers and hospitals. There have been several recent studies using a novel measure-valued representation. There a measure-valued process is used to keep track of the the ages or the residual service times of jobs in the service pool.

**Current work**
**Server-assignment rules:**

Consider a $M/GI/n$ queue with $n$ homogeneous servers in parallel and unlimited waiting space with customers assigned to server in a first-come-first-served (FCFS) order. This stochastic system evolves over a time as consequence of arrivals, service completions and server assignments.

The server-assignment rules takes effect only when a customer arrives and at least one idle server is available. The problem of choosing a good server-assignment rule (under a suitable objective) can be formulated as a stochastic decision process. One can formulate a discrete-time general-state Markov decision problem if letting the discrete times be successive arrival epochs and service completion times, but one can look at a policy in continuous time.

In [5], we develop new rules for assigning idle servers to customers requesting service in a contact center in order to create effective work breaks (long idle periods of certain length) from available idleness. After showing that the standard longest-idle-server-first (LISF) and the random-routing (RR) alternative do not generate breaks often enough, we study the one-parameter rule $D_1$ yielding unannounced work breaks while maintaining work conservation. A sequel [4] to that work provides theoretical support for the performance of the LISF rule and the RR rule by analyzing them in the quality-driven (QD) many-server heavy-traffic (MSHT) regime. The analysis indicates that, with the standard LISF policy, an idle period between two successive service times tends to be a constant value subjected to a Gaussian noise. In contrast, the RR rule produces idle periods which we show to be approximately exponentially distributed. These MSHT limits, along with the $D_1$ rule considered in [5] allow us to gain important insights into how server idle-busy activity patterns can be shaped by server-assignment schemes in the system.

**Ratio-control rules:**

The mainstream literature on diffusion approximations deals with queueing models with a single customer class. Of course, in both call centers and health care settings, as in many service applications contexts, there are multiple customer classes. The classes are differentiated by, for instance, their priority levels, service requirements or impatience. It is plausible to assume FIFS service within each class; however, every time a server becomes free, that server must choose which class to serve next. This gives rise to real-time scheduling control questions.

Of particular interest are a family of ratio-control scheduling policies, as considered in the
paper [2] which showes that fixed-queue-ratio (FQR) controls that schedule (select the next customer to enter service from queue when a server becomes free) aiming to keep the queue lengths at fixed ratios are also effective for achieving delay-based service-differentiation in time-stationary large-scale service systems, delicately balancing the service levels of the different classes.

In [6], we show that, with time-varying (TV) arrival rates, the FQR control may fail badly in stabilizing delay ratios, even though it stabilizes queue ratios well. To provide a remedy, we propose a new head-of-line-delay (HLDR) control. Meanwhile, we establish the first MSHT limits for ratio controls for TV multi-class many-server queues. In particular, we analyze the proposed HLDR rule for a TV multi-class queue with a single pool of exponential servers and multiple customer classes. With class-dependent service, we show that the queueing system can be uniquely characterized by a set of interacting diffusions in the MSHT limit. These MSHT limits show that the HLDR control achieves the desired delay ratios in every sample path.

**Interbank networks:**

In [1] we develop a dynamic model of interbank borrowing and lending activities in which banks are organized into clusters, and adjust their monetary reserve levels so as to meet prescribed capital requirements. Each bank has its own initial monetary reserve level and faces idiosyncratic risks characterized by an independent Brownian motion; whereas system wide, the banks form a hierarchical structure of clusters. We model the interbank transactional dynamics through a set of interacting measure-valued processes. Each individual process describes the intra-cluster borrowing/lending activities, and the interactions among the processes capture the inter-cluster financial transactions. We establish the weak limit of the interacting measure-valued processes as the number of banks in the system grows large. We then use the limiting results to develop asymptotic approximations on two proposed macro-measures, the liquidity stress index and the concentration index, both capturing the dynamics of systemic risk. Numerical examples are used to illustrate the applications of the asymptotics and related sensitivity analysis with respect to various indicators of financial activity.

**Future plans**

**Server-assignment rules:**

While the paper [5] has shown that it is possible to create within-day work breaks from
available idleness, it remains to investigate whether or not these rules would improve the satisfaction of service representatives in real life.

Second, it remains to establish the MSHT (functional weak law of large numbers) FWLLN showing that the sequence of stochastic models converges to the MSHT fluid model as described in [5] under the $D_1$ server-assignment policy. I plan to build on the measure-valued representation approach introduced by [3] to tackle the problem and report results for that in the near future.

**Tandem queues with blocking:**

Blocking may arise in service systems. One such example is hospital emergency department (ED) where patients are held in the ED after they have been admitted to the hospital, because no inpatient beds are available. This practice often results in a number of problems, including ambulance diversions, prolonged waiting times and increased suffering for the waiting patients. Such systems can be conceptualized by a two-station tandem network where customers that are referred from the first station to the second when the latter is saturated (blocked), are forced to wait in the first station while occupying a server there. I plan to develop simple and practical approximations for such queueing network, possibly with time-varying arrivals, and use the results to gain operational insights into the system performance.

**Heavy-traffic approximation for clearing systems:**

A stochastic clearing system (SCS) is somewhat similar to a queueing system. A SCS is characterized by an input process $\{X(t); t \geq 0\}$ with non-decreasing sample paths. For real-world applications, $X(t)$ can be interpreted as the cumulative quantity entering the system up to time $t$. An output mechanism clears the system instantaneously; i.e., removes all the content currently present. Applications include the practice of shipment consolidation that combines several orders/shipments before an aggregate load is dispatched on the same vehicle. A shipment is released either when the cumulative order weight exceeds a certain threshold $w$ or time $T$ has passed since the end of the previous consolidation cycle. An optimal policy (in terms of finding an optimal $w^*$ or $T^*$) has been explicitly characterized with some special distributional assumptions on the input process $\{X(t); t \geq 0\}$. For clearing systems with a general input process, heavy traffic approximation techniques can be useful for obtaining nearly optimal clearing policies.
References


