CanQueue 2021 In-Person Conference

Theme: Behavioural Queueing Science

October 29-30, 2021
McGill University, Montréal

Co-Organizers: Yichuan Ding and Armann Ingolfsson

Host Institution: Desautels Faculty of Management, McGill University

Sponsors:

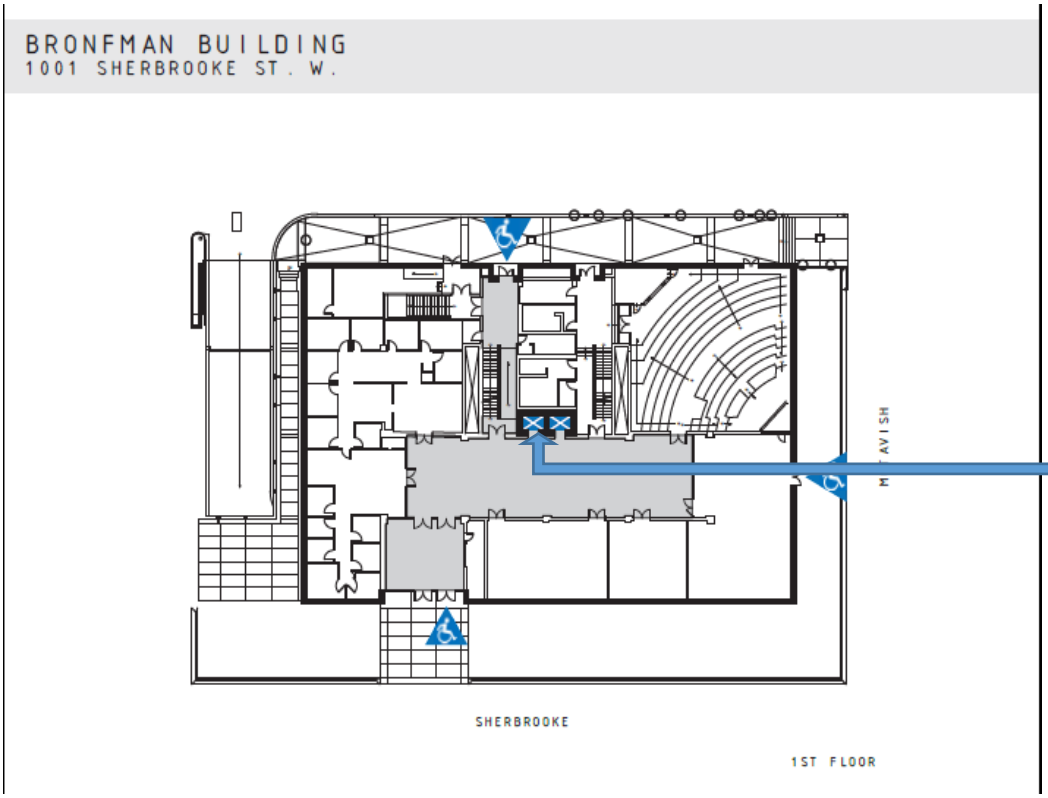
GERAD
GROUP FOR RESEARCH IN DECISION ANALYSIS

McGill | MSRC Management Science Research Centre

Venue: 6th Floor of the Bronfman Building, 1001 Sherbrook St. W., Montreal, H3A 1G5, QC

Enter the building at its right side.

Figure 1: Exterior View of the Bronfman Building



Walk into the center of the building and take the elevator to the **6th**

Figure 2: The First Floor of the Bronfman Building

Schedule

Day 1: October 29, Friday

9:30 – 9:50 **Meet & Greet; Coffee and Snack; Vaccine Passport Check.**
9:50 – 10:00 **Conference Opening**
10:00 – 10:10 **Opening Remarks by Dr. Yolande E. Chan**, Dean of Desautels Faculty of Management, McGill University

10:10 – 11:10 **Session 1: Advances in Queueing Theory**
Ding, Likang *The relationship between expected service times and service rates for state-dependent queues*
Tavakoli, Javad *Making queueing systems Markovian and solving them by randomization*

11:10 – 11:30 **Coffee Break**

11:30– 12:30 **Keynote 1 by Dr. Galit B. Yom-Tov:** ***The human side of queueing systems***

12:30 – 2:00 **Lunch**

2:00 – 3:30 Session 2: Healthcare Applications I
Baron, Opher *On pooled versus dedicated service in the presence of triage errors*
Jin, Yiwen *The cost of task switching: evidence from the emergency department*
Jouini, Oualid *Appointment-driven queueing systems with non-punctual customers*

3:30 – 3:50 **Coffee Break**

3:50 – 5:20 **Session 3: Service Operations**
Asgari, Arash *Model validation for the call centre of a non-profit organization*
Castellanos, Antonio *Silent abandonment in contact centers: estimating customer patience from uncertain data*
Qi, Wei *Stall economy: the value of mobility in retail on wheels*

6:00 – 8:00 **Dinner**

Day 2: October 30, Saturday

9:30 – 11:00 **Session 4: Healthcare Applications II**
Ding, Yichuan *Joint appointment and reentry scheduling: mitigating onsite overcrowding in outpatient services*
Roet-Green, Ricky *On queue-length information in a tandem queueing system*
van Dijk, Nico *On operating theater – intensive care unit – step-down unit systems with overflow and COVID-19*

11:00 – 11:20 Coffee Break

11:20 – 12:20 Keynote 2 by Dr. Laurens Debo: *The human face of queuing: my research journey in queuing land*

12:20 – 2:00 Lunch

2:00 – 3:30 Session 5: Queueing Games

Farajollahzadeh, Setareh	*Potty parity: process flexibility via unisex restrooms*
Yawo, Kobara	*Intensive care unit / step-down unit queuing game with the length of stay decision*
Snitkovsky, Ran	*Stochastic approximation of symmetric Nash equilibria in queueing games*

3:30 – 3:50 Coffee Break

3:50 – 4:50 Session 6: Fluid Queues

Furman, Eugene	*Optimal capacity planning for cloud service providers with periodic, time-varying demand*
Zhao, Yiqiang	*Open problems for Gaussian fluid queues*

4:50 – 5:00 Conference Closing; discuss location for CanQueue 2022

Day 1

Session 1 (10:10 – 11:10): Advances in Queueing Theory

The relationship between expected service times and service rates for state-dependent queues
By Likang Ding, University of Alberta
Abstract: Queueing models are formulated using service rates, but empirical research focuses on expected service times. We analyze the relationship between service times and service rates in state-dependent queueing models. Our work provides a way to translate findings for expected service times to service rates. We investigate two faulty generalizations: that service rates are the inverse of expected service times and that monotone expected service times translate to monotone service rates in an opposite manner. Improved understanding of the relationship between service times and service rates could lead to improved decision making, for example, improved staffing.

Making queueing systems Markovian and solving them by randomization
By Javad Tavakoli, UBC Okanagan
Abstract: Queueing systems have one or more queues that are changed by events, such as arrivals. We assume the one of the events, say event 1, is generated by a renewal process, while the rest of the system is Markovian. To solve the problem, we embed the process at the points where event 1 occurs,

and we determine the transition probabilities of the embedded Markov chain by randomization/uniformization.

Coffee Break (11:10 – 11:30)

Keynote (11:30 – 12:30)

Emotional load: The human side of queueing systems
By Galit B. Yom-Tov, Technion
Abstract: Service delivery relies on interactions of agents with customers. Research has long acknowledged that emotions are an important factor of these interactions, and the industry standard is for agents to serve customers with a smile. In recent years, my research group has promoted the understanding of emotional load – a type of workload that results from the emotional behavior of customers and the agents' need to carefully respond to such behavior. Emotional load affects agent behavior in many areas including online services, healthcare, and the courts. Technological advances make data for researching such constructs more available and new NLP tools provide models that can capture aspects of human emotional behavior. In my talk I will focus on technologically mediated services, specifically chat services, where agent and customer behavior within an interaction are fully observable for analysis. In a series of studies we developed and validated sentiment analysis tools for chat interactions and analyzed large-scale databases of service companies. We explored the relationships between customer emotions, employee behavior, and service effectiveness, as well as the operational implications thereof. I will demonstrate that accounting for emotional load creates more accurate queuing models and that this load has a fundamental role in operational decisions.

Joint work with Anat Rafaeli, Daniel Altman, Shelly Ashtar, Nitzan Carmeli, and Marcelo Olivares.

Lunch (12:30 – 2:00)

Session 2 (2:00 – 3:30): Healthcare Applications I

On pooled versus dedicated service in the presence of triage errors
By Opher Baron, University of Toronto
Abstract: We model and ED that serves acute (type 1) and non acute (type 2) patients under differentiated service level (SL) requirements in the presence of, possibly inaccurate, triage. We evaluate the performance measures of the optimal pooled system with either (i) a first come first serve (FCFS) or a (ii) priority (PR) and the optimal (iii) dedicated system. We discuss the implications of different systems primitives on the optimal system design. For the benchmark case with no triage errors, we show that the capacity of the pooled server with PR may well be dictated by the, less demanding, SL requirement of the type 2 customers. Triage errors change the pattern of arrival rates to servers and the pattern of service rates to customers. We discuss the impact of these changes. For the dedicated system we show that dedicated systems with triage errors may be better than without errors simultaneously in that

servers may face a lower utilization (even though one capacity increases), customers are served faster (one type strictly faster), and costs are lower! We establish similar observations for the pooled system.

The cost of task switching: evidence from the emergency department
By Yiwen Jin, UBC
Abstract: Emergency department physicians treat patients with different symptoms and constantly switch between tasks. We empirically explore the impact of task switching where we construct an instrumental variable to overcome the selection bias in the estimation. We find that task switching increases the average pick-to-pick time by 3.4 to 16 percent or 0.8 to 3.1 minutes per patient. It also affects physicians' routing decisions, yet has little impact on quality. Our counterfactual analysis shows that eliminating the switch cost can reduce the average waiting time by 25.3 to 48.3 percent and the average waiting census by 21.6 to 43.1 percent.

Appointment-driven queueing systems with non-punctual customers
By Oualid Jouini, Centrale Supélec
Abstract: We consider a single server queueing system where a finite number of customers arrive over time to receive service. Arrivals are driven by appointments. However, customers are not necessarily punctual and may arrive either earlier or later than their scheduled appointment times or may not show up at all. We develop both exact and approximate approaches for characterizing the distribution of the number of customers seen by each arrival. We show how this can be used to obtain the distribution of waiting time for each customer. We also illustrate how our approach can be used to support individualised appointment scheduling.

Coffee Break (3:30 – 3:50)

Session 3 (3:50 – 5:20): Service Operations

Model validation for the call centre of a non-profit organization
By Arash Asgari, University of Alberta
Abstract: We report progress on developing several models that reproduce the empirically observed average wait times and abandonment proportions for the current operations of a call centre. So far, we have investigated stationary and transient versions of the Erlang A model and a fluid model, with limited success. Important real-world features ignored in these models include a customer call-back option and variability in the number of agents on duty. We plan to investigate the incorporation of these features using discrete-event simulation.

Silent abandonment in contact centers: estimating customer patience from uncertain data
By Antonio Castellanos, Technion
Abstract: Contact centers are one of the favorite channels of communication with companies. However, they face operational challenges – common proxies for customer experience are subject to information uncertainty. A main source of such is silent abandonment by customers. These customers leave the system while waiting for a reply but give no indication for doing so. As a result, agent capacity is wasted. In two case studies we show that up to 70% of the abandoning customers abandon silently, and that such behavior reduces system efficiency by up to 15%. We develop methodologies to identify silent abandonment and to estimate customer patience. We show how accounting for silent abandonments in

a queueing model improves the estimation accuracy of key measures of performance. Finally, we suggest strategies to operationally cope with the phenomenon.

Stall economy: the value of mobility in retail on wheels
By Wei Qi, McGill University
Abstract: Urban open space emerges as a new territory to embrace retail innovations. Selling products in public spaces with autonomous wheeled vending stalls can potentially become ubiquitous in our future cities. Transition into such a "stall economy" paradigm is being spurred by the recent global pandemic, but has been scarcely studied. This paper provides models, algorithms, and managerial insights to understand how to deploy and operate wheeled stalls in cities to scale up the stall economy. In a broader sense, this work demonstrates an expanded scope of retail operations reshaped by the pandemic and big data.

Dinner (6:00 - 8:00)

The dinner will take place at l'Academie Crescent, 2100 Rue Crescent, Montreal, QC H3G 2B8

View in Google maps

You can get there independently or walk with us. We will leave from the side door of Bronfman Building at 5:45 PM.

Day 2

Session 4 (9:30 – 11:00): Healthcare Applications II

Joint appointment and reentry scheduling: mitigating onsite overcrowding in outpatient services
By Yichuan Ding, McGill University
Abstract: In this paper, we study outpatient scheduling problem in the presence of both walk-in arrivals and patient reentry. The model is based on the new growing care coordination model – the integrated practice unit – that have physician consults, imaging facilities, and other services co-located in the same place, such that patients can finish a sequence of visits within a day. This business model facilitates patient care and improves patient experiences and outcomes, but presents unique challenges in scheduling and patient flow management. We develop a novel, iterative algorithm to efficiently solve the joint appointment time and reentry time scheduling problem, with provable bounds. Our algorithm also deals with possible endogeneity between the schedule and system parameters, which is often overlooked in the literature. Collaborating with a large teaching hospital in China, we demonstrate that our algorithm can reduce patient delay and mitigate overcrowding in comparison to algorithms that ignore patient reentry. Meanwhile, the simplicity and adaptability of our algorithm allows implementation – an ongoing effort we have with our healthcare partner.

On queue-length information in a tandem queueing system
By Roet-Green, Ricky, University of Rochester

Abstract: ED congestion is the main reason that drives patients to leave without being seen before being admitted and also to leave against medical advice after admission while waiting for treatment. Our paper suggests that providing queue-length information to patients might help reduce these phenomena. We model the patient flow as a tandem-queueing system which consists of an admission queue followed by a treatment queue, and patients must complete both services in this order to receive service reward. As in the ED example, we assume patients can observe each queue upon arrival to it and may decide to balk or renege from the system at any time.

On operating theater - intensive care unit - step-down unit systems with overflow and COVID-19
By Nico van Dijk, University of Amsterdam
Abstract: The availability of an intensive care (ICU) bed is of considerable interest, both with and without COVID patients. By integrating an operating theater – intensive care unit (OT-ICU) system with a medium care or step-down unit care (MC or SDU), the availability can be greatly enhanced. Different (artificial) protocols are studied to conclude product form expressions. These are of special interest as covering serial overflow with distinguishable (non) COVID patients. The expressions can be used to provide bounds on the ICU congestion probability for both COVID and non-COVID (operated) patients and allocating ICU and SDU capacities. Numerical support is provided based on realistic (Dutch hospital) data.

Coffee Break (11:00 – 11:20)

Keynote (11:20 – 12:20)

The human face of queuing: my research journey in queuing land
By Laurens Debo**,** Dartmouth College
Abstract: Queues are part of everyone's everyday life. Originally developed to help managers of telephone, computer and manufacturing systems, queuing theory has ignored for a long time the humans *inside* a queue. These humans may matter as they may decide whether to join or not, stay or not and to return or not. Great progress has been made in the past two decades understanding system performance via introducing queuing games, behavioral queuing and queuing experiments in which (human) decision makers are the main focus. I will summarize my journey in this area, identify pitfalls and present some research opportunities. Foremost, I hope to spark enthusiasm among researchers to leverage queuing formalism to explore new frontiers that improve understanding of queues in everyone's everyday life.

Lunch (12:20 – 2:00)

Session 5 (2:00 – 3:30): Queueing Games

Potty parity: process flexibility via unisex restrooms
By Setareh Farajollahzadeh, Toronto University
Abstract: We study the problem of inequitable access to public restrooms by women and the LGBTQ community. Individuals choose to enter a restroom based on their gender identity and the expected or observed wait time. We analytically show the benefits of having unisex restrooms from three angles: (1)

improving wait time parity, (2) enhancing utility parity among users, and (3) increasing safety perception. In addition, we identify optimal restroom designs based on location and restroom types.

Intensive care unit / step-down unit queuing game with the length of stay decision
By *Kobara Yawo,* University of Western Ontario
Abstract: Length of stay (LOS) competition between two servers in tandem without queue between them is investigated using queuing games. This typifies the relationship between the intensive care unit (ICU) and the step-down unit (SDU) of a hospital. We determined the service time each server should take to have an equilibrium quality of life service for the system under four scenarios: (i) both servers cooperate; (ii)the servers do not cooperate and make decisions simultaneously; (iii) the servers do not cooperate but the first server, the ICU is the leader; (iv) the servers do not cooperate, the second server the SDU is the leader. The payoff of the first server, the ICU, is expressed as the difference between the service benefit and the penalty of waiting in the queue, while that of the second server, the SDU is the difference between the service benefit and the penalty of overstaying the ICU due to the SDU being busy. Analytical results showed that the SDU is profitable only when the queueing cost is greater than the service benefit. Numerical results illustrated graphically suggest that the optimal burden of care under the one station with one server ICU system is equivalent to that of the ICU/SDU system under simultaneous decision and SDU Stackelberg but it dominates all four power scenarios.

Stochastic approximation of symmetric Nash equilibria in queueing games
By Ran Snitkovsky, Columbia University
Abstract: We suggest a novel stochastic approximation algorithm to compute a symmetric Nash equilibrium strategy in a general queueing game with a finite action space. The algorithm involves a single simulation of the queueing process with dynamic updating of the strategy at regeneration times. Under mild assumptions regarding the regenerative structure of the process the algorithm converges to a symmetric equilibrium strategy almost surely. This yields a powerful tool that can be used to approximate equilibrium strategies in a broad range of strategic queueing models in which direct analysis is impracticable.

Coffee Break (3:30 – 3:50)

Session 6 (3:50 – 4:50): Fluid Queues

Optimal capacity planning for cloud service providers with periodic, time-varying demand
By Eugene Furman, University of Toronto
Abstract: We analyze a cloud computing system where a provider wants to determine the optimal number of servers and the optimal retrial interval when all servers are busy. Servers in this setting represent components of a computer network and customers are jobs in the cloud. By modeling the system as a fluid queue and using a calculus-of-variations approach, we derive the optimal amount of service capacity in anticipation of time-varying dynamics as well as the optimal retrial interval. Using data collected from a real cloud service provider, we quantify the cost savings using our approach.

Open problems for Gaussian fluid queues
By Yiqiang Zhao, Carleton University

Abstract: In this talk, we consider isonormal Gaussian processes as the input process for Gaussian queueing systems. In terms of Ornstein-Uhlenbeck semigroup, Malliavin derivative, and Mehler's formula, we present a relationship between the covariance function of the queue length process and the covariance function of the input Gaussian process. As applications, properties of two specific isonormal Gaussian processes, fractional Brownian motion (fBm), and sub-fBm, are presented, which lead to answers to two open problems raised by Krzysztof Dębicki and Michel Mandjes 10 years ago (Queueing Systems, 68, 267-273, 2011).

This talk is based on joint work with Prof. Hossein Jafari.