FRIDAY
JUNE 30

INFORMS

THURSDAY
JUNE 29

21st

Book of abstract

WEDNESDAY
JUNE 28

JUNE 28-30

2023

NANCY - FRANCE

INFORMS-APS2023.EVENT.UNIV-LORRAINE.FR

Contents

Committees

Scientific committee

Urtzi Ayesta (CNRS - IRIT)
Charles Bordenave (CNRS / Université Aix-Marseille, France)
Madalina Deaconu (Inria / Université de Lorraine, France)
Laurent Decreusefond (Telecom Paristech, France)
David A. Goldberg (Cornell University, USA)
Rouba Ibrahim (UC London, UK)
Matthieu Jonckheere (CNRS - LAAS, France)
Ger Koole (Vrije Universiteit Amsterdam, Nederland)
Antoine Lejay (Inria / Université de Lorraine, France)
Pascal Moyal (Université de Lorraine, France)
Mariana Olvera-Cravioto (UNC Chapel Hill, USA)
Ohad Perry (Northwestern University, USA)
Viet Chi Tran (Université Gustave Eiffel, France)
Amandine Veber (CNRS / Université Paris Cité, France)

Organizing committee

Jocelyn Begeot (Université de Lorraine, France)
Nathalie Benito (Université de Lorraine, France)
Coralie Fritsch (Inria / Université de Lorraine, France)
Sara Mazzonetto (Université de Lorraine, France)
Pascal Moyal (Université de Lorraine, France) - Head organizer
Marie-Sophie Nourdin (Université de Lorraine, France)
Edouard Strickler (CNRS / Université de Lorraine, France)

Local crew:

Virgile Brodu (Inria / Université de Lorraine, France)
Victor Dubach (Université de Lorraine, France)
Jules Flin (Université de Lorraine, France)
Mathilde Gaillard (Inria / Université de Lorraine, France)
Mohamed Amine Hazzami (Université de Lorraine, France)
Saad Jamiri (Université de Lorraine, France)
Loïc Jean (Université de Lorraine, France)
Angelo Efoévi Koudou (Université de Lorraine, France)
Rodolphe Loubaton (Inria / Université de Lorraine, France)
Thomas Masanet (Université de Lorraine, France)
Pierre Mercuriali (Inria / Université de Lorraine, France)
Raphael Mignot (Université de Lorraine, France)
Anouk Rago (Inria / Université de Lorraine, France)
Christophe Reype (Inria / Université de Lorraine, France)
Pierrick Siest (Université de Lorraine, France)
Saïd Toubra (Inria / Université de Lorraine, France)
Nicolás Zalduendo (Inria / Université de Lorraine, France)

Sponsors

The organizers gratefully acknowledge the support of the following organizations:

Program at a glance

Wednesday, June 28th

8:30	*Registration opens*
8:50	Welcome Speech
9:00 - 10:00	Plenary talk : Frédéric Chazal
	Coffee break
10:30 - 12:00	Parallel sessions
	Lunch break
13:30 - 15:00	Tutorial lecture : Marc Lelarge & Parallel sessions
	Coffee break
15:30 - 17:00	Parallel sessions & Posters session
	Mini break
17:15 - 18:15	Plenary talk : Beatrice Meini
	Mini break
18:30	APS Business Meeting
19:15	Cocktail at Centre Prouvé

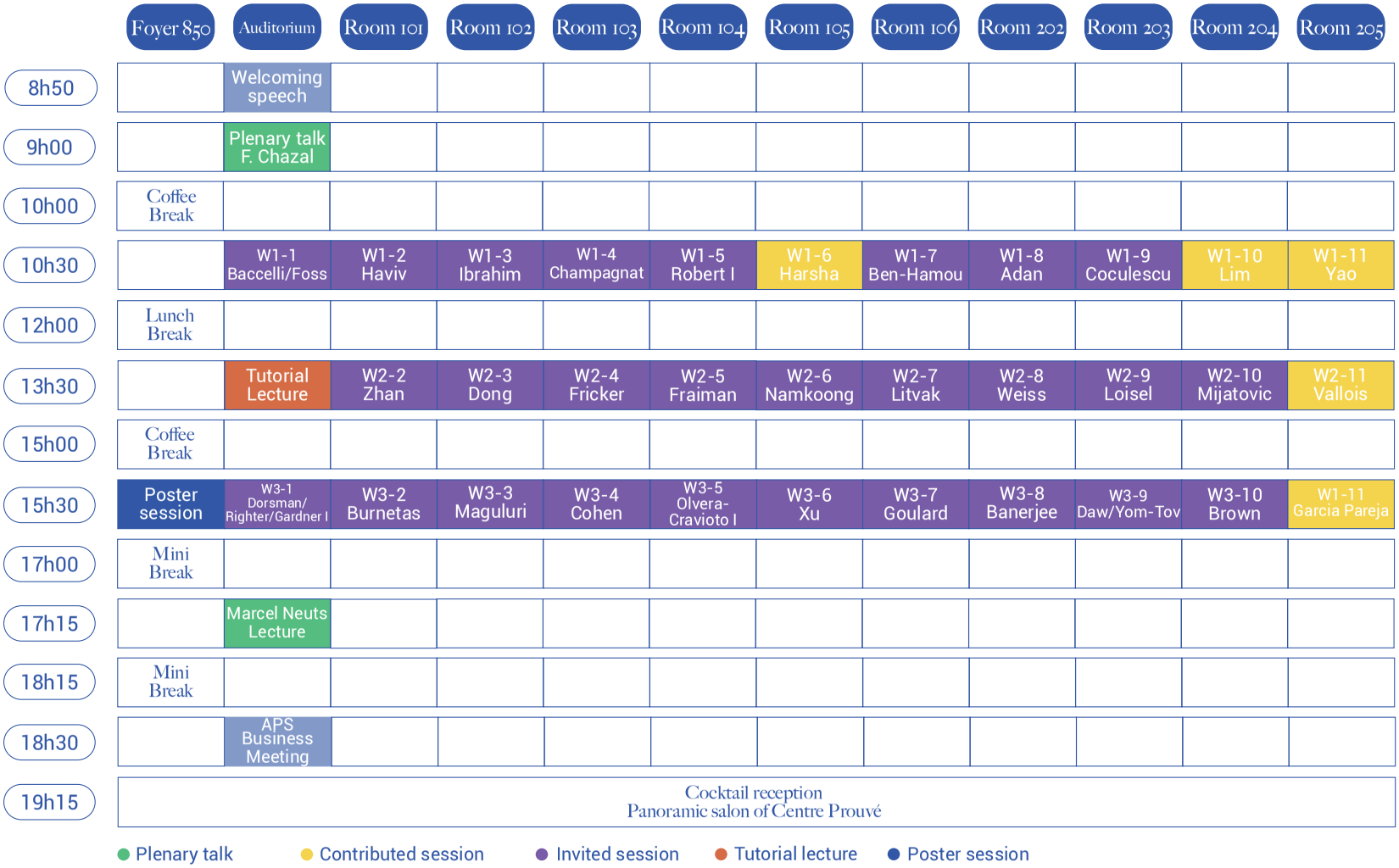
Thursday, June 29th

8:30	*Registration opens*
9:00 - 10:00	Plenary talk : Sylvie Méléard
	Coffee break
10:30 - 12:00	Parallel sessions
	Lunch break
13:30 - 15:00	Parallel sessions
	Coffee break
15:30 - 17:00	Parallel sessions & Posters session
	Mini break
17:15 - 18:15	Plenary talk : Sean Meyn
19:00	Gala dinner

Friday, June 30th

8:30	*Registration opens*
9:00 - 10:00	Plenary talk : Amy Ward
	Coffee break
10:30 - 12:00	Parallel sessions
	Lunch break
13:30 - 15:00	Parallel sessions
	Coffee break
15:30 - 17:00	Parallel sessions & Posters session
	Mini break
17:15	Conclusion

• WEDNESDAY, JUNE 28 •

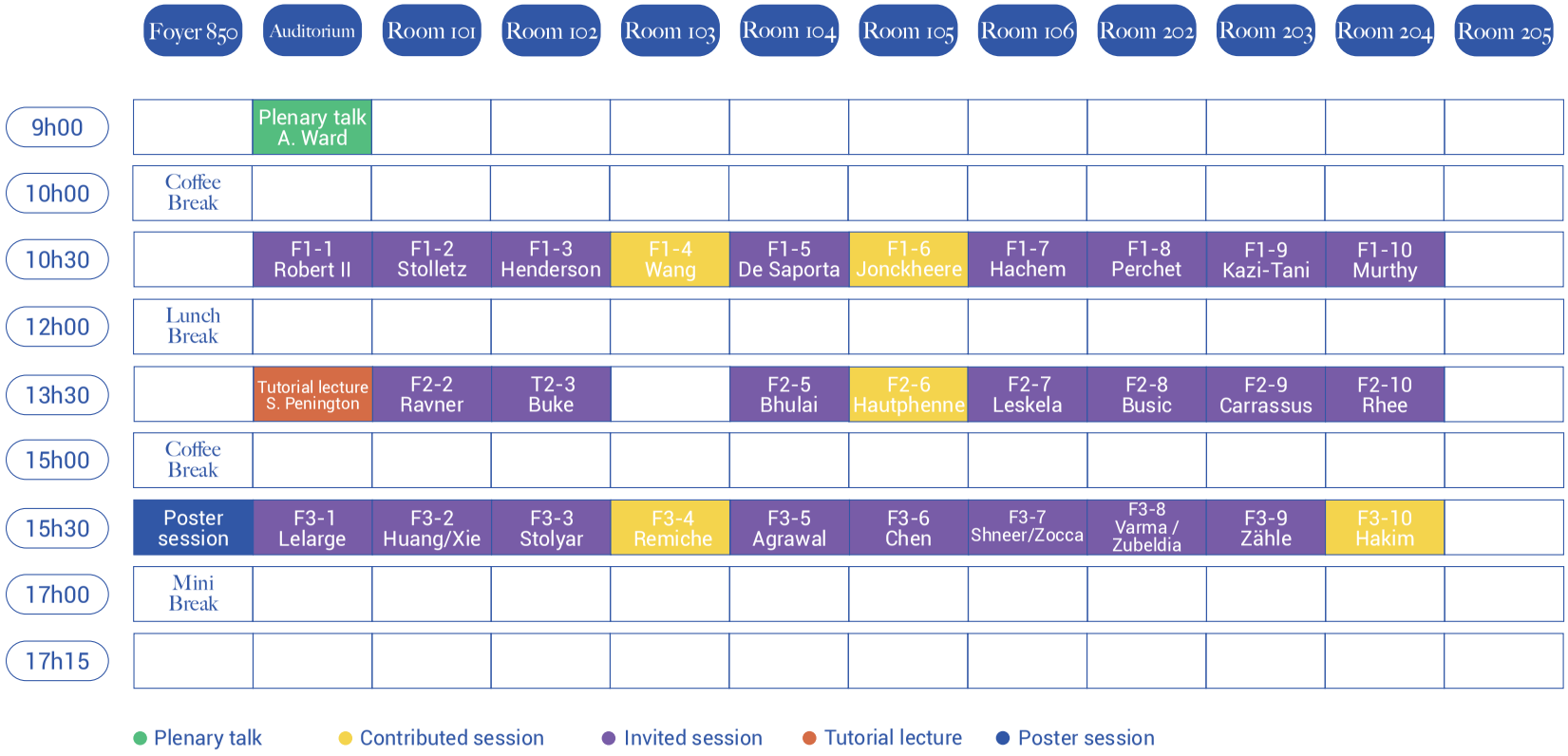


• THURSDAY, JUNE 29 •

	Foyer 850	Auditorium	Room 101	Room 102	Room 103	Room 104	Room 105	Room 106	Room 202	Room 203	Room 204	Room 205
9h00		IMS Medallion Lecture S. Méléard										
10h00	Coffee Break											
10h30		T1-1 Avrachenkov / Ayesta	T1-2 Hasenbein	T1-3 Castellanos	T1-4 Baccelli/ Foss II	T1-5 Coron	T1-6 Shi	T1-7 Olvera-Cravioto II	T1-8 Dorsman/Righter /Gardner II		T1-9 Torrisi	
12h00	Lunch Break											
13h30		T2-1 Ramanan	T2-2 Kanavetas	T2-3 Buke	T2-4 Nazarathy	T2-5 Henry	T2-6 Sanders	T2-7 Avrachenkov	T2-8 Dorsman/Righter /Gardner III		T2-9 Decreusefond I	
15h00	Coffee Break											
15h30	Poster session	T3-1 Blanchet	T3-2 Koole	T3-3 Ayhan	T3-4 Tran	T3-5 Penington	T3-6 Wang	T3-7 Backhausz	T3-8 Comte		T3-9 Decreusefond II	
17h00	Mini Break											
17h15		Applied Probability Trust Lecture S. Meyn										
19h00	Gala Dinner Grands Salons de l'Hôtel de Ville											

● Plenary talk ● Contributed session ● Invited session ● Tutorial lecture ● Poster session

• FRIDAY, JUNE 30 •



List of invited session organizers

- Adan, Ivo (Eindhoven University of Technology, Netherlands) - Random walks and product forms
- Agrawal, Shubhada (Georgia Institute of Technology, United States)- Learning, Optimization and Applied Probability
- Avrachenkov, Kostia (Inria Sophia Antipolis, France)- Network inference I
- Avrachenkov, Kostia (Inria Sophia Antipolis, France)/ Ayesta, Urtzi (CNRS IRIT, France)- Reinforcement learning
- Ayhan, Hayriye (Georgia Institute of Technology, United States) - Control of Queueing Systems
- Baccelli, François (Inria, France) / Foss, Sergey (Heriot-Watt University, United Kingdom) - Infinite-dimensional queueing/communication systems
- Baccelli, François (Inria, France) / Foss, Sergey (Heriot-Watt University, United Kingdom) - Mean-field approach for analysis of complex stochastic systems I
- Backhausz, Agnes (Eötvös Loránd University, Hungary) - Random graphs and matrices
- Banerjee, Sid (Cornell University, United States) - Online decision making
- Ben Hamou, Anna (Sorbonne University, LPSM, France) - Random walks on random graphs
- Bhulai, Sandjai (Vrije Universiteit Amsterdam, Netherlands)- Modern techniques in Markov decision problems
- Brown, David (Duke University, United States) - Advances in Experimentation and Dynamic Models
- Buke, Burak (The University of Edinburgh, United Kingdom)- Many-Server Queues
- Burnetas, Apostolos (National and Kapodistrian University of Athens, Greece) - Strategic queueing III
- Busic, Ana (ENS/Inria, France) - Stochastic matching and applications
- Carassus, Laurence (Ecole Supérieure d'Ingénieurs Léonard De Vinci, France) - Robust finance
- Castellanos, Antonio (University of Chicago, United States) - Queueing Systems with Predictions
- Champagnat, Nicolas / Villemonais, Denis (Inria Nancy/Université de Lorraine) - Quasi-stationary distributions in numerical stochastic methods and statistics
- Chen, Xinyun (CUHK Shenzhen, China) - Learning and queueing theory
- Coculescu, Delia (University of Zurich, Switzerland) - Risk measures in finance and insurance
- Cohen, Asaf (University of Michigan, United States) - Mean-Field Limits and their Applications
- Comte, Céline (CNRS, France) - Online stochastic matching: a diversity of viewpoints
- Coron, Camille (Université Paris Sud, France) - Stochastic models for ecology and evolution
- Daw, Andrew (University of Southern California, United States) / Yom-Tov, Galit (Technion - Israel Institute of Technology, Israel) - Hawkes Processes: New Models and Theory
- De Saporta, Benoîte (Université de Montpellier, France) - Learning while controling: crossroads between stochastic control and statistical learning
- Decreusefond, Laurent (Télécom Paris, France) - Stein method I

List of speakers

Abdelhakmi Anas
Adan Ivo
Afèche Philipp
Agarwal Anish
Agrawal Shubhada
Ali Mohammadi Yeganeh
Allmeier Sebastian
Amanihamedani Alireza
Anton Elene
Aouad Ali
Araya Valdivia Ernesto
Arlotto Alessandro
Asanjarani Azam
Ata Baris
Atar Rami
Avramidis Thanos
Ayhan Hayriye
Baars Justin
Baccelli Francois
Banerjee Siddhartha
Barigou Karim
Baron Opher
Bastani Hamsa
Bayati Mohsen
Bekker Rene
Ben-Hamou Anna
Bhambay Sanidhay
Bhamidi Shankar
Bharti Bharti
Bhulai Sandjai
Bignozzi Valeria
Bindiganavile Ramadas Vinay Kumar
Blaszczyszyn Bartlomiej
Blomerus Wessel
Bodas Shreehari
Bodas Tejas
Borkar Vivek
Borst Sem
Boucherie Richard
Bountali Olga
Braun Alexander
Briend Simon
Britton Tom
Brodu Virgile
Brown David
Buke Burak
Burnetas Apostolos
Caldentey Rene
Campillo Fabien
Cappelletti Daniele
Carassus Laurence
Cardinaels Ellen
Castellanos Antonio
Castiel Eyal
Chazal Frédéric
Che Ethan
Chehrazi Naveed
Cheikhi David
Chen Xinyun
Chen Yilun
Chenchen Xing
Cheung Wang Chi
Chick Stephen
Christianen Mark
Clarkson Jake
Cohen Asaf
Comte Céline
Contat Alice
Corujo Rodriguez Josué
Costa Manon
Coupier David
Coutin Laure
Curien Nicolas
Dai Jim
Das Bikramjit
Davydov Michel
Daw Andrew
De Turck Koen
Decreusefond Laurent
Deep Vikas
Deka Prabhanka
Dendievel Sarah
Denis Laurent
Deo Anand
Diallo Aoudi Mohamed Habib Aliou
Dieleman Nanne
Dimitrakopoulos Yiannis
Dimitriou Ioannis
Dong Jing
Dongyuan Zhan
Doroudi Sherwin
Dreveton Maximilien
Dwivedi Raaz
Economou Antonis
Ecoto Geoffrey
El Masmari Soukaina
Ertel Sebastian
Fan Zhou
Feng Yifan
Ferragut Andres
Flint Ian
Foss Sergey
Fourneau Jean-Michel
Fritsch Coralie
Ganassali Luca
Ganguly Ankan
Gao Xuefeng
Garcia Pareja Celia
Gardner Kristen
Garlati Elena
Gast Nicolas
Gaujal Bruno
Ger Koole
Gerencsér Balázs
Ghosh Soumyadip
Goffard Pierre-Olivier
Goldsztajn Diego
Gonzalez Casanova Adrian
Gonzalez Cazares Jorge
Grigorova Miryana
Gupta Hritika
Gösgens Martijn
Hakim Arief
Halconruy Hélène
Hanukov Gabi
Hartung Lisa
Hassin Refael
Haugh Martin
Hautphenne Sophie
Haviv Moshe
Henderson Shane
Heydenreich Markus
Hillairet Caroline
Hillas Lisa
Holmgren Cecilia
Hong Guiyu
Honnappa Harsha
Hu Yue
Huang Junfei
Huang Zhiyuan
Hyytiä Esa
Ibrahim Rouba
Jacko Peter
Jahnel Benedikt
Janicka Agnieszka
Jhunjhunwala Prakirt
Jiang Jiashuo
Johlke Anika
Jonckheere Matthieu
Joshi Purva
Juneja Sandeep

Kalvit Anand
Kanavetas Odysseas
Karabag Oktay
Karunarathne Wathsala
Kazi-Tani Nabil
Keppo Jussi
Kerimov Süleyman
Kerner Yoav
Khayyati Siamak
Khezeli Ali
Khodadadian Sajad
Kirchner Matthias
Kobeissi Ziad
Kohlenberg Angela
Koopmans Camiel
Kraakman Yanna
Krätschmer Volker
Kumar Rachitesh
Lagos Guido
Laurence Lucie
Lauriere Mathieu
Le Corre Thomas
Le Quellennec Orlane
Lebeau Hugo
Lee Junghyun
Legros Benjamin
Lelarge Marc
Leskelä Lasse
Levering Nikki
Lewis Mark
Li Hannah
Li Shuangning
Li Siqiao
Liberopoulos George
Lim Andrew
Lim Eunji
Lin Feng
Litvak Nelly
Loeser Eva
Logothetis Dimitrios
Loisel Stéphane
Lomys Niccolò
Loubaton Philippe
Loubaton Rodolphe
Lu Yingdong
Luo Yuwei
Lykouris Thodoris
Maggiori Andreas
Maguluri Siva Theja
Mai Xiaoyi
Maillard Antoine
Maillard Odalric-Ambrym
Maillard Pascal
Mandjes Michel
Manou Athanasia
Mao Cheng
Marek Oheim
Marguet Aline
Mariucci Ester
Mastropietro Daniel
Mathien Joffrey
Matias Catherine
Mcarthur Harry
Meini Beatrice
Méléard Sylvie
Mertz Laurent
Meyn Sean
Miao Sentao
Miguelez Fernando
Mijatovic Aleksandar
Minabutdinov Aleksei
Mohamed Hanène
Mohammadi Narges
Moriarty John
Mouchtaki Omar
Moyal Pascal
Mukherjee Debankur
Mukhopadhyay Arpan
Murthy Karthyek
Müller Noela
Nagy Oliver
Namkoong Hongseok
Neufeld Ariel
Nieto Benoît
Nourdin Ivan
Olvera-Cravioto Mariana
Oz Binyamin
Pain Michel
Pakyari Reza
Pal Chandan
Paraggio Paola
Peccati Giovanni
Penington Sarah
Perarnau Guillem
Pergamenchtchikov Serguei
Perry Ohad
Pfaffelhuber Peter
Pham Huyên
Pijnappel Tom
Popovic Lea
Privault Nicolas
Qian Pengyu
Qu Yanlin
Ramanan Kavita
Ravner Liron
Ray Rounak
Rebuffi Louis-Sébastien
Reid Mirabel
Reiffenhäuser Rebecca
Remiche Marie-Ange
Reveillac Anthony
Rhee Chang-Han
Righter Rhonda
Robledo Francisco
Rochlin Igor
Rosazza Gianin Emanuela
Saha Subhamay
Sanders Jaron
Sankararaman Abishek
Sarhangian Vahid
Sarkovic Andjela
Schmocker Philipp
Schwarz Justus Arne
Scully Ziv
Seddik Mohamed El Amine
Senen-Cerda Albert
Sentenac Flore
Shadmi Yonatan
Shah Abhin
Shi Pengyi
Shneer Seva
Si Nian
Singh Simrita
Singh Vikas Vikram
Skerman Fiona
Snitkovsky Ran
Spieksma Floske
Srikant R
Stegehuis Clara
Stephan Ludovic
Stolyar Alexander
Strickler Edouard
Sun Maotong
Sundaresan Rajesh
Swan Yvik
Tercieux Olivier
Tourniaire Julie
Toyoizumi Hiroshi
Trottner Lukas
Vallois Pierre
Van Der Hoorn Pim
Van Der Mei Rob
Van Der Sar Erica
Van Eekelen Wouter
Van Santvoort Mike
Van Werde Alexander
Vande Velde Sylvie
Varma Sushil
Vignoud Gaetan
Villemonais Denis
Vo Thi Phuong Thuy
Vuong Christophe
Véber Amandine
Walraevens Joris
Walton Neil
Wang Jinting
Wang Qiong
Wang Tiandong
Wang Weina
Wang Xingyu
Wang Zhe
Wang Zhengli
Wang Zicheng
Ward Amy
Weber Stefan
Weber Thomas
Weedage Lotte
Wei Zhaoxuan
Weiss Gideon
Wellalage Achini
Wichelhaus Cornelia
Wiesel Johannes
Xie Yao
Xu Huifu
Xu Jiaming
Xu Kuang
Yang Xiaotang
Yang Yunfang
Yasodharan Sarath

Chapter 1

Plenary talks

All plenaries take place in Auditorium.

1.1 Frédéric Chazal (Inria Saclay, France)
Wednesday, June 28, 9:00 - 10:00

Clustering of discrete measures via mean measure quantization: application to Topological Data Analysis

In Topological Data Analysis, robust topological information from data is commonly represented by a discrete measure in the Euclidean plane, the so-called persistence diagram, that needs to be converted into a vector to be used in classical machine learning pipelines. In this talk, after a brief introduction to persistent homology, we will introduce a fast, learnt, unsupervised vectorization method for measures in Euclidean spaces and use it for clustering of distributions of discrete measures. The algorithm is simple and efficiently discriminates important space regions where meaningful differences to the "mean" measure arise. Applied to persistence diagrams, we will show that it is proven to be able to separate clusters of persistence diagrams. We will illustrate the strength and robustness of our approach on a few synthetic and real data sets.

1.2 Beatrice Meini (Univ. Pisa, Italy) - Marcel Neuts Lecture
Wednesday, June 28, 17:15 - 18:15

Matrix-analytic Methods from an algebraic and computational point of view

Matrix-Analytic Methods (MAM) were introduced by Marcel Neuts in the '80s and have been applied to a variety of stochastic models, primarily queues but also specific models for dams and inventories.

As Neuts wrote in the review paper Matrix-analytic methods in queueing theory, European Journal of Operation Research, 1984, "[MAM] originated in the search for algorithmic methods and has led to results that are well-suited for computer implementation."

Indeed, MAM allow to compute efficiently the stationary probability distribution of an infinite dimensional Markov chain which has a Toeplitz structure. More specifically, the transition matrix has an infinite number of block rows and columns (representing transitions between levels) and finite dimensional blocks (representing transitions between phases). Moreover, the block entries of the transition matrix are constant along each block diagonal, except for the first block rows or columns. According to MAM, the computation of the stationary probability distribution is ultimately reduced to the solution of a suitable nonlinear matrix equation.

Starting from the '90s, many numerical methods have been developed for the numerical solution of such nonlinear matrix equations, and the research is still active.

In this talk, we review MAM from an algebraic point of view, by interpreting the methodology

introduced by Neuts in terms of infinite dimensional structured linear systems. Also the resulting nonlinear matrix equation can be interpreted by means of a suitable infinite dimensional linear system. This interpretation allows introducing efficient numerical methods, based on Cyclic Reduction, for solving the nonlinear matrix equation. We recall classical fixed point iterations, Newton's method, the cyclic reduction and logarithmic reduction algorithms and we present the more recent advances, based on the Quasi-Toeplitz matrix arithmetic, which allow to treat the case of infinite number of phases, like in the queueing models represented by random walks in the quarter plane.

1.3 Sylvie Méléard (École Polytechnique, France) - IMS Medallion Lecture
Thursday, June 29, 9:00 - 10:00

Multiscale eco-evolutionary models: from individuals to populations

Motivated by recent biological experiments, we emphasize the effects of small and random populations on long time population dynamics. We will quantify such effects on macroscopic approximations. The individual behaviors are described by the mean of a stochastic measure-valued process. We study different long time asymptotic behaviors depending on the assumptions on mutation size and frequency and on horizontal transmission rate. In some cases, simulations indicate that these models should exhibit surprising asymptotic behaviors such as cyclic behaviors. We explore these behaviors on a simple model where population and time sizes are on a log-scale. Explicit criteria are given to characterize the possible asymptotic behaviors. The impact of the time and size scales on macroscopic approximations is also investigated, leading to Hamilton-Jacobi equations.

1.4 Sean Meyn (Univ. Florida, USA) - APT Plenary Lecture
Thursday, June 29, 17:15 - 18:15

Who is Q? A beginner's guide to reinforcement learning

One theoretical foundation of reinforcement learning is optimal control, usually the Markovian variety known as Markov decision processes (MDPs). The MDP model consists of a state process, an action (or input) process, and a one-step cost function that is a function of state and action. The goal is to obtain a policy (function from states to actions) that is optimal in some predefined sense. Chris Watkins introduced the Q-function in the 1980s as part of a methodology for reinforcement learning. Given its importance for over three decades, it is not surprising that the question of the true meaning of Q was a hot topic for discussion during the Simons Institute's Fall 2020 program on Theory of Reinforcement Learning. In this lecture we discover the truth about Q's origins, and what has happened since.
WeâĂŹve all heard about the magic of Q-consider alpha-zero and chatGPT. As we review the foundations of the reinforcement universe, you may share the speaker's amazement that Q-learning is ever successful! This invites many research questions: why does Q-learning result in successful solutions for decision and control? How can we create new approaches to reinforcement learning that are efficient in terms of training, and also provide some estimate of policy performance?
The lecture draws on Chapters 5 and 9 of the new monograph, Control Systems and Reinforcement Learning, as well as recent papers on Convex Q-Learning and Logistic Q-Learning.

1.5 Amy Ward (University of Chicago, USA)
Friday, June 30, 9:00 - 10:00

Queues with (Non-Exponential) Reneging

The study of scheduling problems has a long history in the academic literature. However, many classic models used to study scheduling problems do not incorporate customer impatience. Furthermore, many of the ones that do assume the time a customer is willing to wait for service is exponentially distributed. The issue is that that assumption can lead to poor scheduling

decisions.
In this talk, we discuss the interplay between customer impatience and
scheduling decisions when managing heterogeneous customer classes in (1) a many server queue, and (2) a matching model. To do this, we develop measure-valued fluid models that approximate the evolution of the underlying discrete-event model. A key insight from our analysis is that the scheduling decisions depend critically on the patience time distribution. We end by showing how our fluid model can be used to simplify scheduling problems in which distributional knowledge is not known in advance, and must be learned.

This talk will cover results in the following papers (all available here):

1. Puha and Ward (2022, MOR), Fluid Limits for Multiclass Many Server Queues with General Reneging Distributions and Head-of-Line Scheduling

2. Zhong, Birge, and Ward (2023 Working Paper), Learning the Scheduling Policy in Time-Varying Multiclass Many Server Queues with Abandonment

3. Aveklouris, Puha, and Ward (2023 Working Paper), A Fluid Approximation for a Matching Model with General Reneging Distributions

4. Aveklouris, Devalve, and Ward (2023 Working Paper), Matching Impatient and Heterogeneous Demand and Supply

Chapter 2

Tutorial lectures

2.1 Marc Lelarge (Inria Paris, France)

Wednesday, June 28, 13:30 - 15:00

Phase transition for statistical estimation: algorithms and fundamental limits

In this tutorial, we study various statistical problems such as community detection on graphs, Principal Component Analysis (PCA), sparse PCA, and Gaussian mixture clustering in a Bayesian framework. Using a statistical physics point of view, we show that there exists a critical noise level above which it is impossible to estimate better than random guessing. Below this threshold, we compare the performance of existing polynomial-time algorithms to the optimal one and observe a gap in many situations: even if non-trivial estimation is theoretically possible, computationally efficient methods do not manage to achieve optimality. This tutorial will present how we adapted the tools and techniques from the mathematical study of spin glasses to study high-dimensional statistics.

2.2 Canceled : Sarah Penington (University of Bath, UK)

Friday, June 30, 13:30 - 15:00

Branching random walks with selection

The N-particle branching random walk is a model for a population in which each individual has a fitness value, given by a random displacement from the fitness of its parent, and only the individuals with highest fitness survive to produce new offspring. I will discuss results and open conjectures about the long-term behaviour in this model when N, the total number of individuals, is large.

Chapter 3

Abstracts of invited talks and contributed talks

3.1 Wednesday 28th, 10:30 - 12:00

3.1.1 Invited session W1-1 (Baccelli & Foss): Mean-field approach for analysis of complex stochastic systems I (Auditorium)

Wednesday 28th, 10:30 - 12:00 - chaired by François Baccelli & Sergey Foss

Davydov Michel: Propagation of chaos for replica mean-field models of intensity-based neural networks

Davydov Michel[1]

1 - Inria de Paris (France)

Abstract: Neural computations arising from myriads of interactions between spiking neurons can be modeled as network dynamics with punctuate interactions. However, most relevant dynamics do not allow for computational tractability. To circumvent this difficulty, the Poisson Hypothesis regime replaces interaction times between nodes by Poisson processes. This approximation is often conjectured or numerically validated but not proven. We show that the Poisson Hypothesis holds at the limit of an infinite number of replicas in the replica-mean-field model, which consists of randomly interacting copies of the network of interest.

Gast Nicolas: The bias of mean field approximation

Gast Nicolas[1,2]

1 - Inria Grenoble-Rhône Alpes (France)

2 - Laboratoire d'Informatique de Grenoble (France)

Abstract: Mean field approximation provides a good tool to study a stochastic process. It is constructed by assuming that the particles that compose a system are independent. In this talk, I will talk about tools that are used to compute the approximation error made when using such an approximation. I will introduce the notion of generators and how to compare them. This will lead us to talk about Poisson equation and Stein's method.

Ramanan Kavita: Limits of empirical measures of interacting jump processes on sparse graphs

Ramanan Kavita[1], Ganguly Ankan, Cocomello Juniper

1 - Brown University (United States)

Abstract: We provide characterizations of empirical measure and marginal dynamics of interacting jump processes on large sparse random graphs as solutions to a path-dependent nonlinear evolution equation. We also show how this leads to principled approximations that shed insight into the behavior of several concrete models on large random graphs. This is joint work with J. Cocomello and A. Ganguly.

Baccelli Francois & Foss Sergey: Migration-Contagion Processes

Baccelli Francois[1], Foss Sergey[2], Shneer Seva[2]

1 - Inria (France)
2 - Heriot-Watt University (United Kingdom)

Abstract: Consider the following migration process based on a closed network of N queues with $K = K(N)$ customers. Each station is an infinite-server queue with rate-μ exponential service times. Upon departure, a customer is routed independently and uniformly at random to another station. In addition to migration, these customers are subject to an SIS (Susceptible, Infected, Susceptible) dynamics. That is, customers are in one of two states: I for infected or S for susceptible. Customers can swap their state either from I to S or from S to I only in stations. More precisely, at any station, each susceptible customer becomes infected with the instantaneous rate αY if there are Y infected customers in the station, whereas each infected customer recovers and becomes susceptible with rate β. We let N and K tend to infinity, with assuming that $K/N \to \eta$, where η is a positive constant representing the customer density. The main problem of interest is about the set of parameters of such a system for which there exists a stationary regime where the epidemic survives in the limiting system. The latter limit will be referred to as the thermodynamic limit. We establish several structural properties (monotonicity and convexity) of the system, which allow us to give the structure of the phase transition diagram of this thermodynamic limit w.r.t. η. The analysis of this SIS model reduces to that of a wave-type PDE for which we found no explicit solution. This plain SIS model is one among several companion stochastic processes that exhibit both migration and contagion. Two of them are discussed in the talk as they provide variants to the plain SIS model as well as some bounds and approximations. These two variants are the DOCS (Departure On Change of State) and the AIR (Averaged Infection Rate), which both admit closed-form solutions. The AIR system is a classical mean-field model where the infection mechanism based on the actual population of infected customers is replaced by a mechanism based on some empirical average of the number of infected customers in all stations. The latter admits a product-form solution. DOCS features accelerated migration in that each change of SIS state implies an immediate departure. This model leads to another wave-type PDE that admits a closed-form solution. In this text, the main focus is on the closed systems and their limits. The open systems consisting of a single station with Poisson input are instrumental in the analysis of the thermodynamic limits and are also of independent interest.

3.1.2 Invited session W1-2 (Haviv): Strategic queueing I (Room 101)

Wednesday 28th, 10:30 - 12:00 - chaired by Moshe Haviv

Haviv Moshe: Intermediate priorities and the balancing of the C-mu rule

Haviv Moshe[1,2]

1 - Chinese University of Hong Kong, Shenzhen (China)
2 - Hebrew University of Jerusalem (Israel)

Abstract: The $C\mu$-rule is well-known to be socially optimal in the sense that it minimizes the overall mean waiting costs due to queueing. Yet, this rule is blind to fairness. In particular, it is possible that those with a low cost of wait per unit of time end up suffering more than those with a corresponding high parameter. We suggest here a fairer scheme which minimizes the overall cost under the constraint that this anomaly does not exist. It is based on partitioning customers' classes into leagues, such that absolute priority is granted among the leagues a-la the $C\mu$-rule, while within leagues relative or accumulated priorities are used. Towards that end we firstly revisited some results on such priority schemes and derived some new ones.

Hassin Refael: Optimal and Self Selection of Service Type in a Queueing System where Long Service Postpones the Need for the Next Service

Hassin Refael[1], Wang Jiesen[2]

1 - Tel Aviv University [Tel Aviv] (P.O. Box 39040, Tel Aviv 6997801 Israel)
2 - Tel Aviv University (Tel Aviv Israel)

Abstract: We study a make-to-order system with a finite set of customers. Production is stochastic with a nonlinear dependence between the ordered quantity and production rate. Customers may have to queue until their turn arrives and therefore their order decisions interact. Specifically, while being served, customers are aware of the queue length and choose one of two order quantities (or, service types). The time to the next replenishment (their activity time) is stochastic and depends on the order quantities.

A customer is inactive during service and wheile waiting in the queue. We refer to the service type with a greater ratio of expected activity to service time as more efficient. Customers wish to maximize their proportion of activity time, while the system is interested in maximizing the steady-state average number of active customers, which is referred to as system efficiency. We show that choosing the more efficient service is not always socially optimal. We demonstrate that if we constrain the strategy to be a threshold strategy depending on the number of inactive customers, then the resulting maximal system efficiency is close to optimal. We also study individuals' optimal selection of the service type, and observe that individuals and the manager have opposite incentives: when the queue is long, individuals tend to choose the long service, while the manager prefers the short service in this case. This makes the system hard to regulate. However, we show that simply removing the less efficient service yields an efficient behavior.

Clarkson Jake: On the Price of Information in Queueing

Clarkson Jake[1], Avrachenkov Konstantin[1], Altman Eitan[1]

1 - Centre Inria d'Université Côte d'Azur (France)

Abstract: Consider an M/M/1 queue where joining attains a known reward, but a known cost is paid for each time unit spent waiting for service. In the 1960s, Naor showed that an arrival optimally joins the queue if and only if its length is less than a known threshold. Yet often the queue length cannot be observed without some effort, for example, access of a website or travel to the queue location itself. Therefore, we assume that the queue length can only be learned by paying an additional known cost. Any arrival hence has three options: join blindly, balk blindly, or pay for queue-length information to make the optimal joining decision using Naor's threshold. The value of knowing the queue length depends on the joining strategies of other arrivals. In a recent paper, Hassin and Roet-Green prove that a unique Nash equilibrium always exists and classify regions where the equilibrium probabilities are non-zero. We simplify some of these findings and obtain closed-form expressions for the equilibrium probabilities in certain cases. Further, Hassin and Roet-Green show that social welfare is maximised by minimising the cost of learning the queue length. We compare the effect on social welfare of lowering this queue-length cost with increasing the queue-joining reward. We show that the former dominates and that the latter can even have a detrimental effect.

Bountali Olga: Strategic Double-Booking and Its Impact on Healthcare Operations

Bountali Olga[1]

1 - Rotman School of Management (Canada)

Abstract: During the COVID-19 vaccination process, a significant mass of patients booked double (or multiple) appointments for their vaccines with the hope of receiving treatment faster. This led to many unfulfilled appointments (a.k.a. no-shows) worldwide, left capacity under-utilized, and hindered the efficiency of the vaccination process during a very crucial period. We introduce a queuing model with strategic patients to capture the single- vs. double-booking decisions and examine their impact of system performance and patient outcomes. We use a benchmark representative of transparency, where a central mechanism allows patients to only single book, and quantify the corresponding loss/gain induced by double-booking. We further explore potential interventions for central planners and policy makers to mitigate the negative effects of double-booking.

3.1.3 Invited session W1-3 (Ibrahim): Scheduling in service systems (Room 102)

Wednesday 28th, 10:30 - 12:00 - chaired by Rouba Ibrahim

Daw Andrew: Closing the Service: Service Success Heterogeneity and Its Impact on Routing

Castellanos Antonio[1], Daw Andrew[2], Yom-Tov Galit[3]

1 - Booth School of Business [Chicago] (United States)
2 - University of Southern California (United States)
3 - Technion - Israel Institute of Technology [Haifa] (Israel)

Abstract: Motivated by data from message-based contact centers, we investigate the impact of service-level heterogeneity on system-level decision making. In particular, we show that how the service is closed can have significant effects on real-time routing decisions. Through a static planning problem from the literature stylized to include salient elements of our context, we prove that while the classic lightest load policy may be optimal if services are homogenous, service heterogeneity may cause that policy to fail

even in settings that are otherwise the same. Furthermore, we show that the manner in which services are closed will either cause or prevent this heterogeneity. Finally, within a class of closure policies that justify lightest load routing as optimal, we inspect closure as a decision and determine the optimal closure policy to balance service success and service duration.

Dong Jing: Optimal routing under demand surge: the value of future arrival rate information

Dong Jing[1], Chen Jinsheng[1], Shi Pengyi[2]

1 - Columbia University (United States)
2 - Purdue University (United States)

Abstract: With the growing availability of advanced demand forecasting models, an important operations question is how to utilize this information to make better resource allocation decisions, especially when facing demand surges. In this work, we design surge routing policies that explicitly incorporate future arrival rates for multi-class multi-pool parallel-server systems with partial flexibility. Our proposed policy is easy to interpret and achieves near-optimal performance. The analysis quantifies the benefit of having access to future arrival rate information.

Scully Ziv: Performance of the Gittins Policy in the G/G/1 and G/G/k, With and Without Setup Times

Hong Yige[1], Scully Ziv[2]

1 - Carnegie Mellon University, Computer Science Department (United States)
2 - Cornell University, Operations Research and Information Engineering (United States)

Abstract: How should we schedule jobs to minimize mean queue length? In the preemptive M/G/1 queue, we know the optimal policy is the Gittins policy, which uses any available information about jobs' remaining service times to dynamically prioritize jobs. For models more complex than the M/G/1, optimal scheduling is generally intractable. This leads us to ask: beyond the M/G/1, does Gittins still perform well? Recent results indicate that that Gittins performs well in the M/G/k, meaning that its additive suboptimality gap is bounded by an expression which is negligible in heavy traffic. But allowing multiple servers is just one way to extend the M/G/1, and most other extensions remain open. Does Gittins still perform well with non-Poisson arrival processes? Or if servers require setup times when transitioning from idle to busy? We give the first analysis of the Gittins policy that can handle any combination of (a) multiple servers, (b) non-Poisson arrivals, and (c) setup times. Our results thus cover the G/G/1 and G/G/k, with and without setup times, bounding Gittins's suboptimality gap in each case. Each of (a), (b), and (c) adds a term to our bound, but all the terms are negligible in heavy traffic, thus implying Gittins's heavy-traffic optimality in all the systems we consider. Another consequence of our results is that Gittins is optimal in the M/G/1 with setup times at all loads.

Zhong Yueyang: Learning the Scheduling Policy in Time-Varying Multiclass Many Server Queues with Abandonment

Zhong Yueyang[1], Birge John[1], Ward Amy[1]

1 - The University of Chicago Booth School of Business (United States)

Abstract: We consider a learning variant of a canonical scheduling problem in a multiclass many server queue with abandonment (specifically, the $M_t/M/N+M$ and the $GI/GI/N+GI$ queues). The objective is to minimize the long-run average class-dependent expected linear holding and abandonment costs when the class-dependent model parameters (arrival rates, service rates and abandonment rates) are a priori unknown. The difficulty is that even when parameters are known, characterizing an optimal scheduling policy appears intractable. Fortunately, the simple $c\mu/\theta$ rule, that prioritizes classes in accordance with a static ranking that depends on the costs, the service rates, and the abandonment rates, is asymptotically optimal as the arrival rates and number of servers become large under certain conditions. Then, our task is to learn the service and abandonment rates well enough to determine an optimal static priority ranking for the classes, and we can benchmark our performance by defining the regret relative to the $c\mu/\theta$ rule. We propose a Learn-Then-Schedule algorithm, which is composed of a learning phase during which point estimates of the mean service and patience times are formed, and an exploitation phase during which the $c\mu/\theta$ rule with empirical mean estimates as a surrogate for actual parameters is followed. It is shown that the smallest achievable regret for static priority scheduling policies in T periods is $\omega(\log T)$, and we prove that our proposed algorithm achieves a regret upper bound of $O(\log T)$ which matches the lower bound.

3.1.4 Invited session W1-4 (Champagnat & Villemonais): Quasi-stationary distributions in numerical stochastic methods and statistics (Room 103)

Wednesday 28th, 10:30 - 12:00 - chaired by Nicolas Champagnat & Denis Villemonais

Corujo Rodriguez Josué: Convergence of the empirical measure induced by a Moran type particle system

Corujo Rodriguez Josué[1]

1 - Institut de Recherche Mathématique Avancée (France)

Abstract: We will discuss some results on the convergence of the empirical distribution induced by a mutation-selection Moran type particle system, which is a continuous-time Markov chain. For this model, it is assumed that the particles can be of different types (among a discrete set) and they interact according to two mechanisms: a mutation process where they mutate independently of each other according to an irreducible rate matrix, and a Moran type reproduction process, where two individuals are chosen, one dies and the other is duplicated. Our results include a uniform in time bound for the convergence of the empirical measure to its deterministic limit, and the asymptotic normality with zero mean and explicit variance, for the approximation error between the empirical distribution and its limit, when the number of particles tend towards infinity. We explore the interpretation of this Moran model as a particle process whose empirical probability measure approximates a quasi-stationary distribution, in the same spirit as the Fleming-Viot particle systems. This is a work in collaboration with Bertrand Cloez (INRAE Montpellier).

Fritsch Coralie: Quasi-stationary behavior of the Crump-Young model of chemostat

Cloez Bertrand[1], Fritsch Coralie[2,3]

1 - INRAE - MISTEA, Montpellier (France)
2 - Institut Élie Cartan de Lorraine (France)
3 - Inria Nancy - Grand Est (France)

Abstract: The Crump-Young model consists of two fully coupled stochastic processes modeling the substrate and micro-organisms dynamics in a chemostat. Substrate evolves following an ordinary differential equation whose coefficients depend of micro-organisms number. Micro-organisms are modeled though a pure jump process whose the jump rates depend on the substrate concentration. It is known that this model extincts almost-surely in the sense that micro-organism population vanishes. However, despite its simplicity, the long-time behavior of this process is not well understood. The existence (but not the uniqueness) of quasi-stationary distribution (QSD) as well as some regularity properties of these QSD were proved in the literature. Nevertheless, the long-time behavior of the process before extinction was, until now, unknown. In this work, we prove that there exist a unique QSD and that, conditionally on the non-extinction, the Crump-Young model converges exponentially fast to this QSD.Convergence to quasi-stationary distribution is usually proved though Hilbert techniques. However, our process is not reversible making these techniques difficult to deal with. To overcome this problem, we use recent results [Champagnat-Villemonais 17 & 20, Bansaye-Cloez-Gabriel-Marguet 19] which are a generalization of usual techniques to prove convergence to stationary distribution.The proof is mainly based on hitting time estimates and Lyapunov functions bounds.Due to the deterministic part, the dynamics of the Crump-Young model is highly degenerated. The proof then consists of technical sharp estimates.

Strickler Edouard: Wasserstein contraction and convergence of penalized Markov processes

Strickler Edouard[1,2], Villemonais Denis, Champagnat Nicolas

1 - CNRS (France)
2 - Faculté des Sciences et Technologies [Université de Lorraine] (France)

Abstract: We consider a Markov process living in some space E, and killed (penalized) at a rate depending on its position. In the last decade, several conditions have been given [Champagnat-Villemonais, Bansaye-Cloez-Gabriel-Marguet] ensuring that the law of the process conditioned on survival converges to a quasi-stationary distribution. When one of these conditions is satisfied, the convergence holds exponentially fast in total variation distance. In this talk, we will present very simple examples of penalized Markov process whose conditional law cannot converge in total variation, and we will give a sufficient condition implying contraction and convergence of the conditional law in Wasserstein distance. We then apply this criterion to a collection of examples.

Zalduendo Vidal Nicolas: The multi-type bisexual Galton-Watson branching process

Zalduendo Vidal Nicolas[1,2], Fritsch Coralie[1,2], Villemonais Denis[1,2]

1 - Faculté des Sciences et Technologies [Université de Lorraine] (France)
2 - Inria Nancy - Grand Est (France)

Abstract: The bisexual Galton-Watson process [Daley, 68] is an extension of the classical Galton-Watson process, but taking into account the mating of females and males, which form couples that can accomplish reproduction. Properties such as extinction conditions and asymptotic behavior have been studied in the past years, but multi-type versions have only been treated in some particular cases. In this work we deal with a general multi- dimensional version of Daley's model, where we consider different types of females and males, which mate according to a "mating function". We consider that this function is superadditive, which in simple words implies that two groups of females and males will form a larger number of couples together rather than separate. One of the main difficulties in the study of this process is the absence of a linear operator that is the key to understand its behavior in the asexual case, but in our case it turns out to be only concave. To overcome this issue, we use a concave Perron-Frobenius theory [Krause '94] which ensures the existence of eigen-elements for some concave operators. Using this tool, we find a necessary and sufficient condition for almost sure extinction as well as a law of large numbers. Finally, we study the convergence of the process in the long-time through the identification of a supermartingale.

3.1.5 Invited session W1-5 (Robert): Stochastic Chemical Reaction Networks (Room 104)

Wednesday 28th, 10:30 - 12:00 - chaired by Philippe Robert

Popovic Lea: Stationary distributions for reaction networks with discreteness-induced transitions

Popovic Lea[1], Gallinger Cameron[1]

1 - Concordia University, Montreal (Canada)

Abstract: The phenomenon of discreteness-induced transitions is highly stochastic dependent dynamics observed in a family of autocatalytic chemical reaction networks. These reaction networks describe the behaviour of several different species interacting with each other, and the counts of species concentrate in different extreme possible values, occasionally switching between them. This phenomenon is only observed under some regimes of rate parameters in the network, where stochastic effects of small counts of species takes effect. The dynamics for networks in this family is ergodic with a unique stationary distribution. While an analytic expression for the stationary distribution in the special case of symmetric autocatalytic behaviour was derived by Bibbona, Kim, and Wiuf, not much is known about it in the general case. Here we provide a candidate distribution when the autocatalytic rates are different, inspired by the Moran model with genic selections. We show that the this distribution is stationary when autocatalytic rates are equal, and that it is close to stationary when they are not equal.

Cappelletti Daniele: Two-dimensional weakly reversible reaction networks are positive recurrent

Agazzi Andrea[1], Anderson Dave[2], Cappelletti Daniele[3], Mattingly Jonathan[4]

1 - Universita' degli studi di Pisa (Italy)
2 - University of Wisconsin-Madison (United States)
3 - Politecnico di Torino (Italy)
4 - Duke university (United States)

Abstract: Stochastic reaction networks are continuous-time Markov chains typically used in biology, epidemiology, and population dynamics. The goal is to keep track of the abundance of the different reactants over time. What makes them special from a mathematical point of view is the fact that their qualitative dynamics is described by a finite set of allowed transformation rules, referred to as "reaction graph". A long-standing conjecture is that models with a reaction graph composed by a union of strongly connected components are necessarily positive recurrent, meaning that each single state is positive recurrent. In my talk I will discuss why the conjecture makes intuitive sense and why it is difficult to prove it. I will then show how my collaborators and I adapted Forster-Lyapunov techniques to prove the conjecture in two dimensions.

Laurence Lucie: Scaling methods for stochastic chemical reaction networks.

Laurence Lucie[1]

1 - Sorbonne Université-INRIA (France)

Abstract: In this talk we investigate stochastic chemical reaction networks with scaling methods. This approach is used to study the stability properties of the associated Markov processes, but also to investigate the transient behavior of these networks. It may give insight on the impact of complex features of these networks such as their polynomial rates,leading to the coexistence of multiple timescales. Several examples are discussed.

Loeser Eva: Fluid Limit for a Stochastic Model of Enzymatic Processing with General Distributions

Loeser Eva[1], Williams Ruth[1]

1 - University of California [San Diego] (United States)

Abstract: In this talk, we consider a stochastic chemical reaction system arising as a model for enzymatic processing in a cell. This can also be thought of as a multiserver multiclass queue with reneging operating under the random order of service discipline. Stochastic primitives for the model such as production/interarrival times, processing/service times, and lifetimes are assumed to be generally distributed. We establish a fluid limit for a measure valued process that keeps track of the remaining lifetime for each entity in the system. We prove uniqueness for fluid model solutions under mild conditions and study the asymptotic behavior of fluid model solutions as time goes to infinity. This talk is based on joint work with Ruth Williams.

3.1.6 Contributed session W1-6 (Harsha): Markov control and decision making (Room 105)

Wednesday 28th, 10:30 - 12:00 - chaired by Honnappa Harsha

Chehrazi Naveed: Dynamic Filtration and the Curse of Dimensionality

Chehrazi Naveed[1]

1 - Washington University in St. Louis (United States)

Abstract: Many practical optimal control problems suffer from the curse of dimensionality. Using existing classical methods to reduce the dimensionality of a problem often leads to other complications. For example, methods that are based on state-space aggregation break the underlying Markovian structure of the problem. Similarly, methods that are based on value function approximation break the Hamilton-Jacobi-Bellman (HJB) optimality condition. In this work, we show that if instead of aggregating over the state space we aggregate over the problem's information filtration, we can reduce the dimensionality of the problem by an arbitrary degree without comprising its underlying Markovian structure or breaking the validity of the HJB equation. We illustrate our approach for an important and open problem of managing inventory when inventory records are inaccurate. In our formulation of this problem, the inventory level and its record have separate dynamics. The former is driven by the demand and unobservable loss process while the latter is driven by the demand and transaction error process. We analytically characterize the dynamics of the retailer's belief on the shelf inventory level and prove that the optimal restocking policy is a threshold policy. The presence of transaction errors breaks the retailer's ability to calculate the shelf inventory distribution by tracking a small set of sufficient statistics: it remains a function of the entire observed history. Hence, computing the optimal restocking policy, even though we obtain its exact analytical form, remains numerically intractable. We show that by dynamically adjusting the retailer's information set, we can reduce the dimensionality of the problem by an arbitrary degree. This effectively projects the dynamics of the inventory level onto a manifold that is adaptively adjusting to the information revealed by the inventory record process. We show that the resulting implementable policy, which is optimal for the chosen dynamic filtration, also exhibits significant robustness to model misspecification. To our knowledge, this paper is the first paper that uses a dynamically-adjusted information filtration to tackle the curse of dimensionality.

Mouchtaki Omar: Beyond IID: Data-Driven Decision-Making in Heterogeneous Environments

Besbes Omar[1], Ma Will[1], Mouchtaki Omar[1]

1 - Columbia University (United States)

Abstract: How should one leverage historical data when past observations are not perfectly indicative of the future, e.g., due to the presence of unobserved confounders which one cannot "correct" for? Motivated by this question, we study a data-driven decision-making framework in which historical samples are generated from unknown and different distributions assumed to lie in a heterogeneity ball with known radius and centered around the (also) unknown future (out-of-sample) distribution on which the performance of a decision will be evaluated. This work aims at analyzing the performance of central data-driven policies but also near-optimal ones in these heterogeneous environments. We first establish, for a general class of policies, a new connection between data-driven decision-making and distributionally robust optimization with a regret objective. We then leverage this connection to quantify the performance that is achievable by Sample Average Approximation (SAA) as a function of the radius of the heterogeneity ball: for any integral probability metric, we derive bounds depending on the approximation parameter, a notion which quantifies how the interplay between the heterogeneity and the problem structure impacts the performance of SAA. When SAA is not rate-optimal, we design and analyze problem-dependent policies achieving rate-optimality. We compare achievable guarantees for three widely-studied problems –newsvendor, pricing, and ski rental– under heterogeneous environments. Our work shows that the type of achievable performance varies considerably across different combinations of problem classes and notions of heterogeneity.

Neufeld Ariel: Markov Decision Processes under Model Uncertainty

Neufeld Ariel[1]

1 - Nanyang Technological University (Singapore)

Abstract: In this talk we introduce a general framework for Markov decision problems under model uncertainty in a discrete-time infinite horizon setting. By providing a dynamic programming principle we obtain a local-to-global paradigm, namely solving a local, i.e., a one time-step robust optimization problem leads to an optimizer of the global (i.e. infinite time-steps) robust stochastic optimal control problem, as well as to a corresponding worst-case measure. Moreover, we apply this framework to portfolio optimization involving data of the S&P 500. We present two different types of ambiguity sets; one is fully data-driven given by a Wasserstein-ball around the empirical measure, the second one is described by a parametric set of multivariate normal distributions, where the corresponding uncertainty sets of the parameters are estimated from the data. It turns out that in scenarios where the market is volatile or bearish, the optimal portfolio strategies from the corresponding robust optimization problem outperforms the ones without model uncertainty, showcasing the importance of taking model uncertainty into account. This talk is based on joint work with Julian Sester and Mario Sikic.

Honnappa Harsha: Offline Estimation of Controlled Markov Chains: Minimaxity and Sample Complexity

Honnappa Harsha[1], Banerjee Imon[2], Rao Vinayak[1]

1 - Department of Statistics, Purdue University (United States)
2 - Department of Statistics, Purdue University (United States)

Abstract: This talk will present our recent advances analyzing nonparametric estimators of transition kernels of controlled Markov chains, with applications in reinforcement learning. We consider an offline setting with a fixed dataset, collected using a so-called logging policy. We first consider a discrete controlled Markov chain setting, where the natural nonparametric estimator is a simple counting estimator. Second, in the setting of continuous state, but discrete action, controlled Markov chains, we consider a histogram estimator. We develop sample complexity bounds for these estimators and establish conditions for minimaxity. Our statistical bounds crucially depend on the logging policy through its mixing properties. We show that achieving a particular statistical risk bound involves a subtle and interesting trade-off between the strength of the mixing properties and the number of samples. We demonstrate the validity of our results under various examples, such as ergodic Markov chains, weakly ergodic inhomogeneous Markov chains, and controlled Markov chains with non-stationary Markov, episodic, and greedy controls. Lastly, we use these sample complexity bounds to establish concomitant ones for offline evaluation of stationary Markov control policies.

3.1.7 Invited session W1-7 (Ben Hamou): Random walks on random graphs (Room 106)

Wednesday 28th, 10:30 - 12:00 - chaired by Anna Ben Hamou

Nagy Oliver: Linking the mixing times of random walks on static and dynamic random graphs

Nagy Oliver[1]

1 - Universiteit Leiden (Netherlands)

Abstract: I will present a framework that, subject to mild general conditions, ties together mixing properties of non-backtracking random walks on top of random graphs generated according to the configuration model in two different settings: in one case the graph does not change in time, whereas in the other case one works with a dynamic version of the random graph in which the edges are randomly rewired, but the degrees are preserved. This framework is then used to study an entire class of rewiring dynamics parametrized by the range of the rewiring relative to the position of the random walk. All the graph dynamics in this class exhibit the trichotomy we found earlier, namely, no cut-off, one-sided cut-off or two-sided cut-off. For graph dynamics that interpolate between the two extreme cases in this class of dynamics, this trichotomy splits further into six different sub-regimes. Joint work with Frank den Hollander, Luca Avena, Hakan Guldas (all three Leiden U.), and Remco van der Hofstad (TU Eindhoven). These results were published in https://doi.org/10.1016/j.spa.2022.07.009.

Stephan Ludovic: Tensor completion via nomadic walks

Stephan Ludovic[1], Zhu Yizhe[2]

1 - Ecole Polytechnique Fédérale de Lausanne (Switzerland)
2 - University of California [Irvine] (United States)

Abstract: We study the problem of tensor completion, in which the goal is to recover a tensor T from a randomly sampled subset of its entries. In a very sparse regime, corresponding to the conjectured algorithmic phase transition, we relate this problem to the study of a nomadic walk matrix, that counts weighted walks on a random bipartite Erdos-Rényi graph. We precisely characterize the spectrum of this matrix, and show how its top eigenvectors relate to specific quantities of interest of T. This also yields provable guarantees of reconstruction in a sparse regime than previous works.

Ben-Hamou Anna: Mixing time of the non-backtracking random walk on two-community random graphs

Ben-Hamou Anna[1]

1 - Sorbonne Université - UFR de Mathématiques (France)

Abstract: In this talk, we consider the non-backtracking random walk on sparse random graphs generated according to a variant of the classical configuration model which incorporates a two-community structure. The strength of the bottleneck is measured by a parameter which roughly corresponds to the fraction of edges that go from one community to the other. We establish a threshold on this parameter that separates two different regimes for the mixing time of the walk: a regime with slow mixing without cutoff and a regime with fast mixing and cutoff.

Sarkovic Andjela: Cutoff for random walk on random graphs with a community structure

Hermon Jonathan[1], Šarković Anđela[2], Sousi Perla[2]

1 - University of British Columbia (Canada)
2 - University of Cambridge (United Kingdom)

Abstract: We consider a variant of the configuration model with an embedded community structure, where every vertex has an internal and an outgoing number of half edges. We pick a uniform matching of the half edges subject to the constraints that internal edges in each community are matched to each other and the proportion of half edges between communities i and j being Q(i,j). We prove that a simple random walk on the resulting graph G=(V,E) exhibits cutoff if and only if the product of the Cheeger constant of Q times log|V| diverges.

3.1.8 Invited session W1-8 (Adan): Random walks and product forms (Room 202)

Wednesday 28th, 10:30 - 12:00 - chaired by Ivo Adan

Boucherie Richard: Blocking policies for networks of finite capacity quasi-reversible nodes

Boucherie Richard[1]

1 - Mathematics of Operations Research [University of Twente] (Netherlands)

Abstract: We consider a network of multiple nodes. Customers of different types arrive according to Poisson processes and follow a fixed type-dependent route along some of the nodes. We consider nodes with finite capacity under two blocking policies: the jump-over protocol, under which an arriving customer that finds all capacity occupied jumps over the nodes to the next node on its route, and the push-out protocol under which a customer that finds all capacity occupied pushes a customer in service out of the node and this customer immediately continues its route to the next node on its route. We show that a node that is quasi-reversible for unlimited capacity remains quasi-reversible for finite capacity.

Dimitriou Ioannis: On some reflected autoregressive processes with dependencies

Dimitriou Ioannis[1]

1 - University of Ioannina, Department of Mathematics (Greece)

Abstract: Motivated by queueing applications, we study various reflected autoregressive processes with dependencies. Specifically, we study cases where the interarrival and service times are proportionally dependent with/without additive delay, as well as cases cases where interarrival times depends on whether the service duration of the previous arrival exceeds or not a threshold. We provide expressions for the Laplace transform of the waiting time of a customer in the system in terms of an infinite product of known Laplace transforms. We also study an integer-valued reflected autoregressive process that can be used to model a novel retrial queueing system with orbital searching time to depend on whether the last busy period starts with an empty or a non empty orbit queue. For such a model the generating function of the stationary orbit queue length is given as an infinite product of known generating functions. Some additional generalizations are also discussed.

Walraevens Joris: A discrete-time random walk with jumps in the quarter plane motivated by road traffic

Walraevens Joris[1], Sasaninejad Sara[1], Wittevrongel Sabine[1]

1 - Ghent University (Belgium)

Abstract: We analyze a particular two-dimensional process that is motivated by road traffic. Suppose that vehicles arrive to a bottleneck where they have to queue. We assume two types of vehicles, say regular and freight vehicles. The queue then typically dissolves at a lower speed if freight vehicles are present, due to the speed and acceleration characteristics of these vehicles. We translate this to a queueing model as follows: we have customers (the vehicles) of two different classes arriving in a infinite-capacity single-server queueing system. When no customers of the second class (freight vehicles) are present at the beginning of a service time, the service time follows a particular distribution. On the other hand, if at least one customer of that second class is present, the service time follows another distribution. The position of the first customer of the second class (if any) together with the total number of customers behind this customer at departure times forms a two-dimensional Markov chain in the quarter plane. We construct and solve the functional equation of the joint probability generating function of the stationary vector of this process. The end solution consists of infinite sums of infinite products, which is a common result for this type of processes.

Adan Ivo: On the Generating Function Approach for the Analysis of Queueing Models

Adan Ivo[1], Down Doug[2], Van Eijndhoven Stef

1 - Eindhoven University of Technology (Netherlands)
2 - McMaster University (Canada)

Abstract: The stationary analysis of queueing models is often studied through functional equations that the probability generating function satisfies. It is commonly argued that the solution of these equations must be unique. In this talk, we suggest that one must be cautious in making such a conclusion.

3.1.9 Invited session W1-9 (Coculescu): Risk measures in finance and insurance (Room 203)

Wednesday 28th, 10:30 - 12:00 - chaired by Delia Coculescu

Bignozzi Valeria: Fair valuation under parameter uncertainty

Bignozzi Valeria[1], Tsanakas Andreas, Scognamiglio Salvatore

1 - University of Milano-Bicocca (Italy)

Abstract: Fair valuation of insurance liabilities is generally performed via a two-step approach, combining quadratic hedging with application of a risk measure on the residual liability, to obtain a cost-of-capital margin. This procedure requires the knowledge of the joint distribution of the liability and the assets in the market, which is often not available in practice and estimated from historical data. Focusing on a framework where we only have uncertainty about the parameters of the distribution we propose to consider the estimated parameters as random variables themselves and present a bootstrapping approach to evaluate this extra source of risk. This approach is implemented using a fairly general neural network algorithm.

Grigorova Miryana: Optimal stopping: Bermudan strategies meet non-linear evaluations

Grigorova Miryana[1], Quenez Marie-Claire[2], Yuan Peng[1]

1 - University of Warwick (United Kingdom)

2 - Université Paris Cité - UFR Mathématiques [Sciences] (France)

Abstract: We address an optimal stopping problem over the set of Bermudan-type strategies (which we understand in a more general sense than the stopping strategies for Bermudan options in finance) and with non-linear (possibly non-convex) operators (non-linear evaluations) assessing the rewards, under general assumptions on the non-linear operators. This problem can be seen as a problem of risk minimization with general dynamic (possibly non-convex) risk measures over the set of Bermudan stopping times. We provide a characterization of the value family V in terms of a suitably defined non-linear Snell envelope of the the pay-off family. We establish a Dynamic Programming Principle. We provide an optimality criterion in terms of a non-linear martingale property of V on a stochastic interval. We investigate the non-linear martingale structure and we show that, under suitable conditions, the first time when the value family coincides with the pay-off family is optimal. The reasoning simplifies in the case where there is a finite number, say n, of pre-described stopping times, where n does not depend on the scenario/state of nature. We provide examples of non-linear operators from the stochastic control, mathematical finance, and risk measurement literature, which enter into our framework. Based on a joint work with Marie-Claire Quenez (Paris) and Peng Yuan (Warwick).

Hillairet Caroline: Hawkes processes and application to cyber-insurance derivatives

Hillairet Caroline[1]

1 - ENSAE Paris, CREST (France)

Abstract: In this talk, we provide an expansion formula for the valuation of reinsurance contracts (such that Stop-Loss contracts) whose payoff depends on a cumulative loss indexed by a Hawkes process. This methodology can be applied to cyber-insurance contracts, as the times of occurrence of cyber-claims exhibit self-exciting behavior. The expansion formula involves the addition of jumps at deterministic times to the Hawkes process in the spirit of the well-known integration by parts formula for Poisson functional. From the actuarial point of view, these processes can be seen as “stressed” scenarios. Based on Joint works with Anthony Réveillac and Mathieu Rosenbaum.

Rosazza Gianin Emanuela: Capital allocation rules and generalized collapse to the mean

Rosazza Gianin Emanuela[1], Centrone Francesca[2]

1 - University of Milano-Bicocca (Italy)

2 - Università del Piemonte Orientale (Italy)

Abstract: In the context of risk measures, a key problem consists in deciding how to share the capital to be allocated among different business lines (or sub-units) of a risky activity. This problem implies the definition and the axiomatic study of capital allocation rules (CARs) induced by risk measure. See Denault (2001) and Kalkbrener (2005), among others. A well-known capital allocation principle is the so-called Gradient allocation, whose importance is also recognized in applications and that, intrinsically, assumes Gateaux differentiability of the underlying coherent risk measure. In this case, a capital allocation rule "collapses to a mean" in the sense that it reduces to the expectation w.r.t. the Gateaux derivative. Under conditions weaker than differentiability, the gradient allocation has been generalized by means of the subdifferential capital allocation (defined in terms of elements of the subdifferential of the risk measure), also in the convex and quasi-convex setting. The main goal of this paper is to investigate under

which conditions a capital allocation rule collapses to a mean or, better, to the subdifferential/gradient allocation. In particular, in a law invariance and in a general framework we make a set of assumptions under which a capital allocation rule (with respect to a coherent or (quasi)convex risk measure) reduces to the trivial expectation allocation w.r.t. a suitable probability measure. Our results extend the well known result for Gateaux differentiable risk measures and can be seen as a generalized collapse to the mean (involving a probability measures not necessarily coinciding with the one given a priori).

3.1.10 Contributed session W1-10 (Lim): Simulation and Inference (Room 204)

Wednesday 28th, 10:30 - 12:00 - chaired by Eunji Lim

Dwivedi Raaz: Counterfactual inference for sequential experiments

Dwivedi Raaz[1], Tian Katherine[2], Tomkins Sabina[3], Klasnja Predrag[3], Murphy Susan[2], Shah Devavrat[4]

1 - Harvard and MIT (United States)
2 - Harvard University (United States)
3 - University of Michigan (United States)
4 - MIT (United States)

Abstract: We consider after-study statistical inference for sequentially designed experiments wherein multiple units are assigned treatments for multiple time points using treatment policies that adapt over time. Our goal is to provide inference guarantees for the counterfactual mean at the smallest possible scale – mean outcome under different treatments for each unit and each time – with minimal assumptions on the adaptive treatment policy. Without any structural assumptions on the counterfactual means, this challenging task is infeasible due to more unknowns than observed data points. To make progress, we introduce a latent factor model over the counterfactual means that serves as a non-parametric generalization of the non-linear mixed effects model and the bilinear latent factor model considered in prior works. For estimation, we use a non-parametric method, namely a variant of nearest neighbors, and establish a non-asymptotic high probability error bound for the counterfactual mean for each unit and each time. Under regularity conditions, this bound leads to asymptotically valid confidence intervals for the counterfactual mean as the number of units and time points grows to infinity. Finally, we also introduce an improved variant of nearest neighbors, which we call a doubly robust nearest neighbor estimator that provides a (near-)quadratic improvement in the error when the policies are non-adaptive.

Bayati Mohsen: Adaptive Experiment Design with Panel Data

Xiong Ruoxuan[1], Athey Susan[2], Bayati Mohsen[2], Imbens Guido[2]

1 - Emory University [Atlanta, GA] (United States)
2 - Stanford Graduate School of Business [Stanford] (United States)

Abstract: In this talk, we study the design and analysis of experiments conducted on a set of units over multiple time periods where the starting time of the treatment may vary by unit. The design problem involves selecting an initial treatment time for each unit in order to most precisely estimate both the instantaneous and cumulative effects of the treatment. In this talk, we consider an adaptive experimental design setting, where both the decision to continue the experiment and treatment assignment decisions are updated after each period's data is collected. We propose a new algorithm, the Precision-Guided Adaptive Experiment (PGAE) algorithm, that addresses the challenges at both the design stage and at the stage of estimating treatment effects, ensuring valid post-experiment inference accounting for the adaptive nature of the design. Using realistic settings, we demonstrate that our proposed solutions can reduce the opportunity cost of the experiments by over 50%, compared to non-adaptive design benchmarks.

Lu Yingdong: An analytical approach to Hamiltonian Monte Carlo

Ghosh Soumyadip[1], Lu Yingdong[1], Nowicki Tomasz[1]

1 - IBM [Yorktown] (United States)

Abstract: Through an analytical approach, we identify Hamiltonian Monte Carlo method as iterations of linear operators on function spaces of probability densities. First, this approach enables us to show the sequence of probability densities produced by HMC algorithms converge in Lp space (weakly when p is between (1,2) and strongly when p>=2). Then, coupled with probabilistic arguments, we demonstrate that the convergence is geometric (in both Lp spaces and Wasserstein space) for HMC implementations in Euclidean spaces under strong convexity conditions on the target distributions. Finally, this analytical approach allows us to design a new algorithm, ADHMC, that can utilize a much larger class of auxiliary distributions, including asymmetric and multi-mode distributions. In addition to guarantee its converge in theory, we also demonstrate that ADHMC outperforms its traditional HMC version in practice.

Lim Eunji: The limiting behavior of isotonic and convex regression estimators when the model is misspecified

Lim Eunji[1]

1 - Adelphi University (United States)

Abstract: We study the asymptotic behavior of the least squares estimators when the model is possibly misspecified. We consider the setting where we wish to estimate an unknown function f^* from noisy observations $(X_1, Y_1), ..., (X_n, Y_n)$; our estimator is the minimizer of sum of squared errors over a set G of functions. We provide sufficient conditions on the metric entropy of G, under which our proposed estimator converges to g^* as the number of observations increases to infinity, where g^* is the minimizer of $||g - f^*|| = E[g(X) - f^*(X)]^2$ over g in G. As corollaries of our theorem, we establish the consistency of our estimator as the number of observations increases to infinity when G is the set of monotone functions or the set of convex functions. We also make a connection between the convergence rate of our estimator and the metric entropy of G. As special cases of our finding, we compute the convergence rate of our estimator when G is the set of bounded monotone functions or the set of bounded convex functions.

3.1.11 Contributed session W1-11 (Xie): Optimization, games (Room 205)

Wednesday 28th, 10:30 - 12:00 - chaired by Yao Xie

Keppo Jussi: The Wisdom of Strategically Diverse Crowds

Jia Yanwei[1], Keppo Jussi[2], Satopaa Ville[3]

1 - Columbia University (United States)
2 - National University of Singapore (Singapore)
3 - INSEAD (France)

Abstract: We introduce strategic diversity as a new source of heterogeneity in predictions made by a group of agents. Our model describes agents who have private information and different external motivations, ranging between conformity and contrarianism. We find that the agents' strategic tendencies become more extreme, i.e., conformists (contrarians) get closer to (deviate away from) the crowd average when the crowd becomes, on average, more conformal. Both a randomly selected individual and the group's consensus perform worse as the level of strategic diversity in the group increases. Therefore, unlike other types of diversity such as information diversity, strategic diversity is not beneficial in crowd forecasting. If strategic diversity exists, however, a weakly contrarian crowd leads to the best performance for both a randomly selected agent and the group's consensus. In contrast to the simple average, the median is more robust to the perturbation caused by strategic motives. We also discuss the implications of such diversity in information aggregation.

Saha Subhamay: Partially Observable Discrete-time Discounted Markov Games with General Utility

Saha Subhamay[1], Bhabak Arnab[1]

1 - Indian Institute of Technology, Guwahati (India)

Abstract: In this talk, we will discuss about partially observable zero sum games where the state process is a discrete time Markov chain. We will consider a general utility function in the optimization criterion. We will show the existence of value for both finite and infinite horizon games and will also establish the existence of optimal polices. The main step involves converting the partially observable game into a completely observable game which also keeps track of the total discounted accumulated reward/cost. This is a joint work with Arnab Bhabak.

Pal Chandan: Discrete-time zero-sum games for Markov chains with risk-sensitive average cost criterion

Pal Chandan[1], Ghosh Mrinal K[2], Golui Subrata[1], Pradhan Somnath[3]

1 - Indian Institute of Technology, Guwahati (India)
2 - Indian Institute of Science (India)
3 - Queen's University (Canada)

Abstract: We study zero-sum stochastic games for controlled discrete time Markov chains with risk-sensitive average cost criterion with countable/compact state space and Borel action spaces. The payoff function is nonnegative and possibly unbounded for countable state space case and for compact state

space case it is a real-valued and bounded function. For countable state space case, under a certain Lyapunov type stability assumption on the dynamics we establish the existence of the value and a saddle point equilibrium. For compact state space case we establish these results without any Lyapunov type stability assumptions. Using the stochastic representation of the principal eigenfunction of the associated optimality equation, we completely characterize all possible saddle point strategies in the class of stationary Markov strategies. Also, we present and analyze an illustrative example.

Xie Yao: Sinkhorn distributionally robust optimization with application to robust hypothesis tests

Xie Yao[1]

1 - Georgia Institute of Technology (United States)

Abstract: We present a study of distributionally robust optimization (DRO) using Sinkhorn divergence, which is a type of Wasserstein distance with entropic regularization. Our work introduces a convex reformulation of the problem that enables computationally tractable solutions for a wider range of loss functions. This approach offers a more generalizable worst-case distribution than the Wasserstein divergence-induced uncertainty set by extending to the entire space. We propose an efficient first-order algorithm with a bisection search to solve the dual reformulation while providing performance guarantees. Furthermore, we show how this methodology can be used to solve robust hypothesis tests, a fundamental problem in statistics. We provide numerical examples using both synthetic and real data to demonstrate our approach's competitive performance and computational efficiency. This is joint work with Jie Wang and Rui Gao.

3.2 Wednesday 28th, 13:30 - 15:00

3.2.1 Invited session W2-2 (Zhan): Strategic queueing II (Room 101)

Wednesday 28th, 13:30 - 15:00 - chaired by Dongyuan Zhan

Afèche Philipp: Pricing In A Two-sided Market With Time-sensitive Customers And Suppliers

Afèche Philipp[1], Akan Mustafa[2]

1 - UNIVERSITY OF TORONTO (Canada)

2 - Carnegie Mellon University (United States)

Abstract: We consider a firm that matches stochastically arriving and time-sensitive customers and suppliers. We characterize and compare the structure and performance of the profit-maximizing and socially optimal pricing policies.

Zhong Yueyang: Behavior-Aware Queueing: When Strategic Customers Meet Strategic Servers

Zhong Yueyang[1], Gopalakrishnan Raga[2], Ward Amy[1]

1 - The University of Chicago Booth School of Business (United States)

2 - Queen's University Smith School of Business (Canada)

Abstract: Service system design is often informed by queueing theory. Traditional queueing theory assumes that customers are indefinitely patient and servers work at constant speeds. That is reasonable in computer science and manufacturing contexts. However, customers and servers in service systems are people, and, in contrast to jobs and machines, systemic incentives created by design decisions influence their behavior. First, we study the behavior of strategic servers whose choice of work speed depends on managerial decisions regarding (i) how many servers to staff and how much to pay them, and (ii) whether and when to turn away customers. We develop a game-theoretic many-server Markovian queueing model with a finite or infinite buffer in which the work speeds emerge as the solution to a non-cooperative game. In an asymptotic regime in which demand becomes large and the utility function becomes concave, we establish existence, uniqueness, and monotonicity properties of underloaded, critically loaded, and overloaded equilibria for various regions in the design space. We then extend our model to also incorporate strategic customers' joining decisions, which endogenously induce a finite buffer. By comparing equilibria where strategic individuals maximize their own utility with those that maximize social welfare or net profit, we characterize the managerial costs of ignoring strategic behavior.

Zur (Schwarz) Elisheva: Priority queues with arrival rate uncertainty

Zur (Schwarz) Elisheva[1], Oz Binyamin[1], Skir Gregory[1]

1 - The Hebrew University of Jerusalem (Israel)

Abstract: We study an unobservable single-server queue with arrival-rate uncertainty. Customers arrive according to a Poisson process with a random arrival rate and decide whether to pay for priority. We characterize the equilibrium customer behavior under different pricing and information disclosure policies. The model is analyzed under two different assumptions regarding the customers' degree of rationality. The first assumes that customers are fully rational and take into account the RASTA (rate-biased ASTA) phenomenon. The second assumes that customers are naive and base their decision on the PASTA phenomenon even though the arrival process is not Poisson.

Dongyuan Zhan: Advance Selling and Upgrading in Priority Queues

Wang Yaolei, Cao Ping, Xie Jingui, Dongyuan Zhan[1]

1 - UCL (United Kingdom)

Abstract: We study advance selling and upgrading in a priority queue setting that emerges in the amusement park industry. Customers choose to buy the fast-track or regular tickets depending on both their heterogeneous waiting costs and the demand/congestion level. At the park entrance, customers can choose tickets based on the observed congestion levels. Advance ticket purchases are cheaper, but may pose congestion uncertainty. Upgrading options allow customers to purchase regular tickets in advance and upgrade on-site if the congestion is high. Our objective is to find the optimal pricing scheme to maximize revenue. We find that upgrading is not necessary if all customers purchase tickets online, but it may generate more revenue if there exist some offline customers who cannot purchase in advance.

3.2.2 Invited session W2-3 (Dong): Efficient scheduling and resource allocation (Room 102)

Wednesday 28th, 13:30 - 15:00 - chaired by Jing Dong

Arlotto Alessandro: On the constant regret in dynamic resource allocation problems

Vera Alberto[1], Arlotto Alessandro[2], Gurvich Itai[3], Levin Eli[4]

1 - Amazon (United States)
2 - Duke university [Durham] (United States)
3 - Northwestern University [Chicago, Ill. USA] (United States)
4 - Yale University [New Haven] (United States)

Abstract: In this talk, I will review and compare recent constant-regret results in dynamic resource allocation problems. In such problems, requests of different types and with different rewards arrive in each of T periods and, if served, they consume a certain amount of a subset of available resources. Every time a request arrives, a decision maker must decide whether to serve that request without the possibility of looking to the future or of changing past decisions. The goal of the decision maker is to maximize the sum of the expected rewards of the resources that are served. (Based on joint work with A. Vera, I. Gurvich and E. Levin.)

Ibrahim Rouba: Size-based scheduling in service systems

Ibrahim Rouba[1], Dong Jing

1 - University College London (School of Management, University College London, 1 Canada Square, London E14 5AB, U.K. United Kingdom)

Abstract: Size-based scheduling policies, such as shortest-remaining-processing-time (SRPT) and shortest-job-first (SJF), have been extensively studied, for decades, yet almost exclusively in single-server queues with infinitely patient jobs and under exact job-size information. Motivated by applications to service systems, such as call centers or healthcare facilities, we analyze the performance of size-based scheduling in multiserver queues with abandonment and inexact job-size information. In particular, we demonstrate that system performance under size-based scheduling with noisy service times is asymptotically equivalent, in the many-server heavy-traffic limit, to performance under a two-class priority queue where customers with short service times (below a threshold) are served without wait, and customers with long service times (above a threshold) eventually abandon without service.

Sarhangian Vahid: Dynamic transfer policies for parallel queues

Chan Timothy, Park Jangwon, Sarhangian Vahid[1]

1 - UNIVERSITY OF TORONTO (Canada)

Abstract: Service systems with multiple locations, such as hospital networks, are often vulnerable to disproportionate congestion caused by spatial mismatches in customer demand and available capacity. The imbalance in congestion motivates transferring customers between different locations. To gain insights into the structure of optimal transfer policies, we consider a system of parallel single-server queues and investigate dynamic policies for transferring customers between queues at the beginning of discrete intervals. We consider minimizing the total expected cost incurred over a finite horizon, comprised of holding costs, as well as variable and fixed transfer costs. By analyzing associated discrete-time fluid control problem, we provide insights on the structure of optimal transfer policies and various cost trade-offs. The discrete-time setting enables us to capture a key trade-off with "excessive" idleness, which arises due to the inability to re-distribute customers over the duration of a period. When applied to systems of two-queues, we recover a natural threshold-type structure akin to state-dependent (s,S) policies in inventory control. This result states that each queue has a pair of optimal "re-order" and "order-up-to" levels such that when one queue falls below its re-order level, it is optimal to transfer just enough customers from the other to replenish it to its order-up-to level.

Shi Pengyi: Refined Mean-Field Approximation for Discrete-Time Queues to Improve Itinerary Completion in Coordinated Care Networks

Shi Pengyi[1], Pan Yang[2]

1 - Purdue University [West Lafayette] (United States)
2 - ETH Zürich (Switzerland)

Abstract: Coordinated care is a burgeoning model of healthcare delivery that involves comprehensive diagnosis and treatment planning based on collaboration between multiple co-located medical services. The care model necessitates an itinerary of patient visits to different services within a care network over the course of several days. Achieving a rapid completion of patient diagnosis and treatment plans is critical to care delivery. We develop a new discrete-time queueing network that incorporates several salient features in coordinated care to optimize capacity allocation and improve itinerary completion. We characterize the itinerary completion time with a doubly-stochastic phase-type distribution and develop a refined mean-field approximation to tackle the curse of dimensionality. We explicitly quantify the convergence rate for this approximation as $O(1/N)$ with N being the system size. Not only is this convergence better than that proven in prior work, but our approximation shows a significant improvement in performance prediction accuracy when the system size is small, compared to the conventional (unrefined) mean-field approximation. This accuracy makes our approximation appealing to support decision-making in practice. Using real data, we show that our approach can efficiently solve networks as large as with 26 stations of services and significant improvements can be achieved over the historical practice as well as intuitive rule-of-thumb heuristics. We demonstrate hat appointment allocation is a multifaceted problem and that ignoring any of those facets can lead to poor performance, highlighting the significance of our integrated approach.

3.2.3 Invited session W2-4 (Fricker): Mean-field approach for analysis of complex stochastic systems II (Room 103)

Wednesday 28th, 13:30 - 15:00 - chaired by Christine Fricker

Allmeier Sebastian: Bias of Multiscale Mean Field Models and Averaging Methods

Allmeier Sebastian[1], Gast Nicolas[1]

1 - INRIA Grenoble-Rhône Alpes (France)

Abstract: The mean field approximation is a powerful technique which has been used in many settings to study large-scale stochastic systems. In the case of two-timescale systems, the approximation is obtained by a combination of scaling arguments and the use of the averaging principle. In this talk, I will present how to analyze the approximation error of this average' mean field approach for a two-timescale model (X, Y), where the slow component X describes a population of interacting particles which is fully coupled with a rapidly changing environment Y.

We show that under relatively mild conditions the 'average' mean-field approximation has a bias of order $O(1/N)$, where N is a scaling factor. We also derive a steady state bias correction term that reduces the error to order $O(1/N^2)$. This refined 'average' mean-field approximation allows the computation of

an accurate approximation even for small scaling factors, i.e., $N \approx 10$–50. We illustrate the developed framework and accuracy results through an application to a random access CSMA model.

Castiel Eyal: Time scale separation in Heavy traffic for Queue Based CSMA

Castiel Eyal[1], Sem Borst[2], Laurent Miclo[3], Florian Simatos[4], Philip Whiting[5]

1 - Georgia Institute of Technology [Atlanta] (United States)
2 - Eindhoven University of Technology
3 - Toulouse School of Economics
4 - ISAE-SUPAERO
5- Macquarie University

Abstract: In wireless networks, the analysis of the Queue Based CSMA algorithm is a challenging problem due to the complex interactions between scheduling decisions and the queuing network. Queues are placed on an interference graph, with activation/deactivation rates adapted as a function of the queue length at each station. For suitably chosen rates, the schedule and queue lengths operate on different time scales and the schedule homogenizes towards an equilibrium adapted to the current queue lengths. This talk will focus on a simple interference graph and present a method to gain explicit error bounds on the time scale separation and use them to prove a deterministic heavy traffic convergence with an unusual scaling. We will use time scale separation to establish a state space collapse for queue lengths and investigate the influence of the idleness due to the distributed nature of the algorithm.

El Masmari Soukaina: The asymptotic behavior of a large-scale storage system with loss, duplication and admitting policies

El Masmari Soukaina[1], El Kharroubi Ahmed

1 - University Hassan II - Casablanca (Morocco)

Abstract: Nowadays, with the huge volume of data generated everyday, one of the crucial aspects of the research interest is to establish effective solutions that guarantee a good data storage management. In this talk we will present interesting results coming from probabilistic tools and the queueing theory that aim to study different models associ- ated to the unreliable storage systems with duplication mechanism and the admitting policy. Resolution of the Skorokhod problem is one of the impor- tant tools used to solve the stochastic differential equations associated to the models considered, in addition to the renormalisation techniques that gave some suitable scaling serving in obtaining interesting conclusions on the system behavior. In fact the investigation has first been done on a system assumed to be with infinite capacity where the explicit expression of the fluid limits associated to the model are obtained that helps in giving a good statistical performance assured in terms of increasing file reliability, to be followed by some results concerning the case of the storage system with finite capacity [2]. This last exhibits the use of an averaging principle and a Skorokhod problem in its study [3]. The study of similar storage systems have been already studied by many authors with quite different models see for example [4], [1], [5], [6], and also see the Santiago Duran's thesis where the study of such unreliable storage system relies on the mean field theory, see [7] Keywords: Skorokhod problem, Fluid limits, Large-scale storage system. References [1] Feuillet, M., and Robert, P. , Scaling analysis of a transient stochastic network. Advances in Applied Probability, 46(2), 516-535 (2014). [2] El Kharroubi, A., El Masmari, S. Fluid limits of a loss storage network, Queueing Syst 101, 137-164 (2022). [3] Kharroubi, A. E., Masmari, S. E. (2023). On a loss storage network with finite capacity. arXiv preprint arXiv:2302.05237. [4] Philippe Robert. Stochastic Networks and Queues. Vol. 52. Stochastic Mod- elling and Applied Probability Series. New-York: Springer, June 2003, pp. xix+398. [5] P.J. Hunt and T.G. Kurtz, title Large loss networks, journal Stochastic Processes and their Applications, volume 53, year 1994, issn 0304-4149, [6] Wen Sun., Mathieu Feuillet., Philippe Robert. (2016) Analysis of large unreliable stochastic networks. Annals of Applied Probability, 26, 2959-3000. [7] Duran, Santiago. duran 2020 resource, Resource allocation with observable and unobservable environments. Universit ÌĄe Paul Sabatier-Toulouse III

Mohamed Hanène: Large scale analysis of a stochastic model for a system with two types of particles

Mohamed Hanène[1], Fricker Christine, Rigonat Alessia

1 - Université Paris Nanterre - Département de Mathématiques et Informatique (France)

Abstract: This talk deals with the large-scale analysis of a system with M particles moving between the N nodes of a network according to markovian dynamics, N and M being large with the same order. Each node has a finite but large capacity cN and is subject to the arrival of another type of particles, called external particles, which leave the node after service. The external particles are more numerous and

compete with the M particles of the system for the capacity of the nodes. This is an original simple model suited to free-floating car-sharing system dynamics. The remaining capacity seen by the M particles is then random, with transition rates of order N. Therefore, although the main ingredient is a mean-field approach, a stochastic averaging principle allows to obtain the large-scale behavior of the system. It is described at different time scales by convergence results, especially the mean-field limit and the explicit long term limit. It turns out that, for this M-particles system behavior, a phase transition occurs between a saturated and an unsaturated regime. The critical value depends only on parameters of the network and the external particles.

3.2.4 Invited session W2-5 (Fraiman): Algorithms on random graphs (Room 104)

Wednesday 28th, 13:30 - 15:00 - chaired by Nicolas Fraiman

Briend Simon: Finding adam under polynomial connectivity assumption

Briend Simon[1,2]

1 - Universitat Pompeu Fabra [Barcelona] (Spain)
2 - Université Paris-Saclay (France)

Abstract: We study the problem of root finding in growing random graphs under general hypothesis on connection probabilities. A simple algorithm is introduced, that runs in polynomial time and that outputs a set of vertices containing the original vertex with high probability. We present its performance in the particular case of a "general graph" version of the uniform attachment tree. Unlike in the tree, each new vertex connects to several ancestors (and not only one) chosen uniformly at random. We then explain why this method can be extended to a wide range of growing graphs models and give other examples of such models.

Perarnau Guillem: Short Synchronizing Words for Random Automata

Chapuy Guillaume[1], Perarnau Guillem[2,3]

1 - Université Paris Cité (France)
2 - Universitat Politècnica de Catalunya (Spain)
3 - Centre Recerca Matemàtica (Spain)

Abstract: We prove that a random deterministic finite automaton on n states has a synchronizing word of length $O(n^{1/2}\log n)$ with high probability, confirming conjectures of several authors based on numerical simulations. The best theoretical result prior to this work was the bound $O(n\log^3 n)$ given by Cyril Nicaud. Additionally, our proof provides a quasi-quadratic algorithm to find such synchronizing word.

Skerman Fiona: Is it easier to count communities than find them?

Skerman Fiona[1], Rush Cynthia, Wein Alex, Yang Dana

1 - Uppsala Universitet [Uppsala] (Sweden)

Abstract: Random graph models with planted community structure have been studied extensively in the literature. For the problems of finding algorithms to both detect and recover community structure, an interesting landscape of statistical and computational phase transitions has emerged. A natural unanswered question is: might it be possible to infer properties of the community structure (for instance, the number and sizes of communities) even in situations where actually finding those communities is believed to be computationally hard? We show the answer is no. In particular, we consider certain hypothesis testing problems between models with different community structures, and we show (in the low-degree polynomial framework) that testing between two options is as hard as finding the communities. In addition, our methods give the first computational lower bounds for testing between two different 'planted' distributions, whereas previous results have considered testing between a planted distribution and an i.i.d. 'null' distribution. Joint work with Cynthia Rush, Alex Wein and Dana Yang.

Stegehuis Clara: Is this network problem hard? Theory vs practice

Stegehuis Clara[1]

1 - University of Twente [Netherlands] (Netherlands)

Abstract: While many graph-based problems are in theory NP complete, for many such problems there exist algorithms that run extremely quickly on large-scale real-world networks. This shows a disparity between theory and practice: while some theoretical examples exist on which any algorithm can take extremely long, these graphs do usually not appear in real-world networks. Therefore, it is often possible to show that these problems can be solved efficiently on random graph with realistic network properties. In this talk, we focus on algorithms that list all cliques in a graph. We show that this problem can take an exponential time even on many classes of random graphs with realistic network properties. However, here again a disparity between theory and practice arises: while we can show an exponential lower bound on the running time of this problem with a max-clique based algorithm, this lower bound is dominated by a linear term for all practical purposes.

3.2.5 Invited session W2-6 (Namkoong): Frontiers in sequential learning (Room 105)

Wednesday 28th, 13:30 - 15:00 - chaired by Hongseok Namkoong

Caldentey Rene: Diffusion Approximations for a Class of Sequential Experimentation Problems

Caldentey Rene[1], Araman Victor[2]

1 - The University of Chicago (United States)
2 - American University of Beirut (Lebanon)

Abstract: We consider a decision maker who must choose an action in order to maximize a reward function that depends on an unknown parameter. The decision maker can delay taking the action in order to experiment and gather additional information on the unknown parameter. We model the decision maker's problem using a Bayesian sequential experimentation framework and use dynamic programming and diffusion-asymptotic analysis to solve it. For that, we scale our problem in a way that both the average number of experiments that is conducted per unit of time is large and the informativeness of each individual experiment is low. Under such regime, we derive a diffusion approximation for the sequential experimentation problem, which provides a number of important insights about the nature of the problem and its solution

Kalvit Anand: Diffusion limits of multi-armed bandit experiments under optimism-based policies

Kalvit Anand[1], Zeevi Assaf[1]

1 - Columbia University (United States)

Abstract: Our work provides new results on the arm-sampling behavior of the classical UCB family of optimism-based multi-armed bandit algorithms, leading to several important insights. Among these, it is shown that arm-sampling rates under UCB are asymptotically deterministic, regardless of the problem complexity. This discovery facilitates new sharp asymptotic characterizations that among other things, reveal profound distinctions between UCB and Thompson Sampling such as an "incomplete learning" phenomenon characteristic of the latter.

Zhang Kelly: Inference after Using RL to Learn Across Users for Digital Interventions

Zhang Kelly[1]

1 - Harvard University [Cambridge] (United States)

Abstract: Online reinforcement learning (RL) algorithms are a key tool for personalizing decision-making for digital interventions, as these algorithms can use previously collected data from users to learn and improve future treatment decisions. We provide methods to perform a variety of statistical analyses using data collected by RL and other adaptive sampling algorithms for digital interventions. In this work, we focus on data collected by online RL algorithms that can learn across users, i.e., use the data of multiple users to learn and inform treatment decisions. This data type is important since highly noisy outcomes (as is the case for digital interventions) mean RL algorithms that learn using the data of multiple users can significantly reduce noise and learn faster. At the same time, this data type is challenging to develop inferential theory for because online RL algorithms that combine data across induce dependence between the collected user data trajectories. We develop a general inferential approach for this non-i.i.d. data type that allows one to, for example, construct confidence intervals for time-varying treatment effects and perform off-policy analyses.

Namkoong Hongseok: Adaptive Experimentation at Scale

Namkoong Hongseok[1], Che Ethan[1]

1 - Columbia Business School (United States)

Abstract: Standard bandit algorithms that assume continual reallocation of measurement effort are challenging to implement due to delayed feedback and infrastructural/organizational difficulties. Motivated by practical instances involving a handful of reallocation epochs in which outcomes are measured in batches, we develop a new adaptive experimentation framework that can flexibly handle any batch size. Our main observation is that normal approximations universal in statistical inference can also guide the design of scalable adaptive designs. By deriving an asymptotic sequential experiment, we formulate a dynamic program to leverage prior information on average rewards. To (approximately) solve the DP, we propose a simple iterative planning method, Residual Horizon Optimization, and demonstrate it significantly improves statistical power over standard adaptive policies, even when compared to Bayesian bandit algorithms (e.g., Thompson sampling) that require full distributional knowledge of individual rewards. Overall, we expand the scope of adaptive experimentation to settings that are difficult for standard adaptive policies, including problems with a small number of reallocation epochs, low signal-to-noise ratio, and unknown reward distributions.

3.2.6 Invited session W2-7 (Litvak): Spatial network models (Room 106)

Wednesday 28th, 13:30 - 15:00 - chaired by Nelly Litvak

Bindiganavile Ramadas Vinay Kumar: Community detection on block models with geometric kernels

Bindiganavile Ramadas Vinay Kumar[1], Avrachenkov Konstantin[1], Leskela Lasse[2]

1 - INRIA Sophia Antipolis (France)
2 - Department of Mathematics [Aalto] (Finland)

Abstract: We consider the community recovery problem on a graph of N nodes where every node has two independent labels: a location label and a community label. A geometric kernel maps the locations of pairs of nodes to probabilities. Edges are drawn between pairs of nodes based on their communities and the value of the kernel. Given the graph so generated along with the location labels, the latent communities of the nodes is to be inferred. Similar models have been investigated in [1,2]. In this talk, we will look into the fundamental limits for recovering the communities in such models. Further, we propose a spectral algorithm and characterize its performance in terms of the accuracy of community recovery. This talk is based on a joint work with K. Avrachenkov (INRIA, Sophia Antipolis) and L. Leskela (Aalto University, Finland). [1] Abbe, E., Baccelli, F., & Sankararaman, A. (2021). Community detection on Euclidean random graphs. Information and Inference: A Journal of the IMA, 10(1), 109-160. [2] Avrachenkov, K., Bobu, A., & Dreveton, M. (2021). Higher-order spectral clustering for geometric graphs. Journal of Fourier Analysis and Applications, 27(2), 22.

Dreveton Maximilien: Higher-Order Spectral Clustering for Geometric Graphs

Dreveton Maximilien[1]

1 - EPFL (Switzerland)

Abstract: The present paper is devoted to clustering geometric graphs. While the standard spectral clustering is often not effective for geometric graphs, we present an effective generalization, which we call higher-order spectral clustering. It resembles in concept the classical spectral clustering method but uses for partitioning the eigenvector associated with a higher-order eigenvalue. We establish the weak consistency of this algorithm for a wide class of geometric graphs which we call the Soft Geometric Block Model. A small adjustment of the algorithm provides strong consistency. We also show that our method is effective in numerical experiments even for graphs of modest size.

Van Der Hoorn Pim: Convergence of discrete graph curvature for random geometric graphs

Van Der Hoorn Pim[1], Cunningham William, Krioukov Dmitri, Lippner Gabor, Trugenberger Carlo

1 - Eindhoven University of Technology (Netherlands)

Abstract: Curvature is a fundamental geometric characteristic of smooth spaces. In recent years different notions of curvature have been developed for combinatorial discrete objects such as graphs. However, the connections between such discrete notions of curvature and their smooth counterparts remain unclear.

In particular, it is not rigorously known if any notion of graph curvature converges to any traditional notion of curvature of smooth spaces. In this talk I will address this challenge by considering Ollivier-Ricci curvature on random geometric graphs on smooth and complete Riemannian manifolds. I will identify proper settings for which the Ollivier-Ricci curvature converges to the Ricci curvature of the manifold. These results create a bridge between the discrete structure of graphs and continuous geometry underneath them. The results hold for different notions of graph distances, including the rescaled shortest path distance, and for different graph densities, with the average degree ranging from nearly logarithmic to nearly linear in the graph size.

Weedage Lotte: Resource and location sharing in spatial networks

Weedage Lotte[1], Stegehuis Clara[1]

1 - University of Twente (Netherlands)

Abstract: Infrastructure sharing is a tool in wireless communications to enhance internet coverage and capacity by increasing redundancy and over-provisioning resources. In this settings, resources of different operators are shared and their cell towers can be collocated. In earlier work, we showed, based on simulations, that resource and location sharing benefits the network in terms of throughput and coverage. However, the underlying mathematical theory is underdeveloped. In this talk, we introduce a mathematical model for location sharing where base stations of different operators can be collocated on the same node. We model cell towers and users as two independent Poisson Point Processess. Every user connects to all cell towers within a certain radius of that user. To quantify the benefit of location sharing, we define a measure which resembles the strength of a link. This strength depends on the number of resources at a cell tower, the degree of this cell tower and the average link distance. We show results on the performance of this strength under different conditions, and derive an optimal radius in which users should connect to optimize the link strength.

3.2.7 Invited session W2-8 (Weiss): Applications of Dynamic Matching (Room 202)

Wednesday 28th, 13:30 - 15:00 - chaired by Gideon Weiss

Chen Yilun: Bridging online and offline matching: Simulate-Optimize-Assign-Repeat (SOAR)

Chen Yilun[1], Kanoria Yash[2], Kumar Akshit[2], Zhang Wenxin[2]

1 - SDS, CUHK Shenzhen (China)
2 - Columbia Business School (United States)

Abstract: We consider a feature-based dynamic matching problem faced by centralized platforms in a highly heterogeneous market. Specifically, a set of heterogeneous supply units, each characterized by i.i.d. supply feature vector, is available initially. In each period, a customer arrives with an i.i.d. demand weight vector describing her type, and requests to consume a supply unit. The platform seeks a dynamic matching policy that assigns supply units to customers to maximize the expected average matching utility. We propose and analyze a simple, simulation-based matching policy, dubbed Simulate-Optimize-Assign-Repeat (SOAR). We prove that SOAR enjoys a surprisingly universal (near) optimality guarantee. Indeed, it achieves the optimal regret scaling under various assumptions on the demand and supply distributions and for different matching utility functions. Our result is premised on a novel connection between the expected matching utility under SOAR and the offline optimum of the matching instance, which may be of wider applicability and independent interest. Extensive numerical simulations support the robustness of the performance of SOAR.

Kohlenberg Angela: The Cost of Impatience in Dynamic Matching: Scaling Laws and Operating Regimes

Kohlenberg Angela[1], Gurvich Itai[1]

1 - Northwestern University [Evanston] (United States)

Abstract: We study matching queues with abandonment. The simplest of these is the two-sided queue with servers on one side and customers on the other, both arriving dynamically over time and abandoning if not matched by the time their patience elapses. We identify non-asymptotic and universal scaling laws for the matching loss due to abandonment, which we refer to as the "cost of impatience." The scaling laws characterize the way in which this cost depends on the arrival rates and the (possibly different) mean patience of servers and customers. Our characterization reveals four operating regimes defined by the

relationship between the traffic intensity (heavy or light) and the level of "inventory" that can be held (high or low). Traffic intensity is a measure of customer impatience relative to demand-supply mismatch, while inventory level is, similarly, a relative measure of server impatience. The more patient the servers, the more "inventory" is held in the system. When the inventory level is low, the abandonment rate is given by an insightful expression where only the minimum of the two patience rates appears. Considering the trade-off between abandonment and costly capacity, we characterize the scaling of the optimal safety capacity as a function of costs, arrival rates, and patience parameters. Known results for the single-sided queue with abandonment (the $M/M/1+M$ queue) are recovered as a special case. In the two-sided queue, the ability to hold "inventory" of patient servers means that the optimal safety capacity is of smaller magnitude than that prescribed in the single-sided queue.

Qian Pengyu: The Competition for Partners in Matching Markets

Qian Pengyu[1], Kanoria Yash[2], Min Seung-Ki[3]

1 - Purdue University [West Lafayette] (United States)
2 - Columbia University (United States)
3 - Korea Advanced Institute of Science and Technology (South Korea)

Abstract: We study the competition for partners in two-sided matching markets with heterogeneous agent preferences, with a focus on how the equilibrium outcomes depend on the connectivity in the market. We model random partially connected markets, with each agent having an average degree d in a random (undirected) graph, and a uniformly random preference ranking over their neighbors in the graph. We formally characterize stable matchings in large markets random with small imbalance and find a threshold in the connectivity d at $\log^2 n$ (where n is the number of agents on one side of the market) which separates a “weak competition” regime, where agents on both sides of the market do equally well, from a “strong competition” regime, where agents on the short (long) side of the market enjoy a significant advantage (disadvantage). Numerical simulations confirm and sharpen our theoretical predictions, and demonstrate robustness to our assumptions.We leverage our characterizations in two ways: First, we derive prescriptive insights into how to *design* the connectivity of the market to trade off optimally between the average agent welfare achieved and the number of agents who remain unmatched in the market. For most market primitives, we find that the optimal connectivity should lie in the weak competition regime or at the threshold between the regimes. Second, our analysis uncovers a new conceptual principle governing whether the short side enjoys a significant advantage in a given matching market, which can moreover be applied as a diagnostic tool given only basic summary statistics for the market. Counterfactual analyses using data on centralized high school admissions in a major USA city show the practical value of both our design insights and our diagnostic principle.

Weiss Gideon: On A Conjecture for FCFS Parallel Service Systems

Weiss Gideon[1]

1 - The University of Haifa (Israel)

Abstract: We consider a service system with several types of customers and several types of servers subject to a bipartite compatibility graph, under the FCFS policy. The design and staffing of such systems is complicated by the fact that stability, waiting times, and utilization of servers can only be determined if the matching rates are known. Thus calculation of these matching rates is of great importance but it is intractable in almost all scenarios. Such matching rates are known for the FCFS binary infinite matching model. It was conjectured (see QS 100 anniversary issue) that if the arrival rates of customers and the number of servers of each type are large, matching rates will converge to these infinite matching rates. This conjecture is supported by simulation results. We will discuss attempts to prove this conjecture, and classify easier and harder cases as generality of the model increases.

3.2.8 Invited session W2-9 (Loisel): Actuarial mathematics and quantitative risk management (Room 203)

Wednesday 28th, 13:30 - 15:00 - chaired by Stéphane Loisel

Barigou Karim: Gaussian Process-Based Mortality Monitoring using Multivariate Cumulative Sum Procedures

Barigou Karim[1], Loisel Stéphane[2], Salhi Yahia[2]

1 - Université Laval [Québec] (Canada)
2 - Université Claude Bernard Lyon 1 (France)

Abstract: This paper presents a multivariate cumulative sum (MCUSUM) procedure to detect changes in mortality intensity, which plays a critical role in risk management for insurance companies and pension funds. The MCUSUM algorithm is built on Gaussian process-based non-parametric mortality forecasts and tracks differences between predicted and realized mortality rates in real-time. Unlike univariate methods, the MCUSUM accounts for interdependence between age-groups and provides a more comprehensive analysis of mortality trends. The efficiency of the MCUSUM method is demonstrated through a comparison to univariate control charts and a case study of recent mortality data in France, Japan, Canada, and the USA.

Toyoizumi Hiroshi: Optimal Dividend Strategy with Poisson Jump Revenue and Constant Negative Running Cost

Toyoizumi Hiroshi[1]

1 - Department of Applied Mathematics [University of Waseda] (Japan)

Abstract: We discuss the optimal dividend strategy of the business with the cash level fluctuating by the fixed size revenue of Poisson jumps and constant negative running cost, which is the dual of the classic Cramer-Lundberg risk model. We show that the value function of this problem is the unique viscosity solution of the first-order variational inequality with delay, and the barrier strategy is optimal. Iteratively solving the delayed differential equation, we can obtain the barrier value in this optimal strategy and its appropriate level of working capital.

Marek Oheim: Risk Class Management in Life Insurance Markets

Marek Oheim[1]

1 - RPTU Kaiserslautern-Landau (Germany)

Abstract: Grouping customers of different risk types into contracts is important for the stability and the robustness of an insurance market. In order to analyze the effects of these risk coalitions, we set up a market model which is inspired by a paper of Sass and Seifried (2014). Afterwards, we provide studies on the equilibrium premiums of classic life insurance products in monopolistic and competitive insurance markets. In our model we can show that the market specifications have a high impact on the optimal choice of different contracts in the insurance market. Regarding markets with capacity constraints serve as a window for a deeper understanding of the model. Further extensions of the model and its applications to other insurance types complete the talk.

References:

J. Sass, F.T. Seifried (2014). Insurance markets and unisex tariffs: Is the European Court of Justice improving or destroying welfare? Scandinavian Actuarial Journal. 2014, 228-254.

Loisel Stéphane: Quickest detection problem and actuarial applications

Loisel Stéphane[1]

1 - ISFA, Univ. Lyon 1 (France)

Abstract: In this talk we present theoretical and practical results concerning the quickest detection problem, concerning abrupt changes in some parameters of dynamic population models. We start with the case of the abrupt change in the intensity of a Poisson process, then explain how to extend the result to a birth-and-death process, and then present applications on simulated and real-world datasets.

3.2.9 Invited session W2-10 (Mijatovic): Simulation, Estimation and Monte Carlo Algorithms for Levy Processes (Room 204)

Wednesday 28th, 13:30 - 15:00 - chaired by Aleks Mijatovic

Mariucci Ester: Nonparametric density estimation of small jumps of Lévy processes

Mariucci Ester[1], Duval Céline[2], Jalal Taher[1]

1 - Université de Versailles Saint-Quentin-en-Yvelines (France)

2 - Université Lille Nord (France) (France)

Abstract: We consider the problem of estimating the density of the process associated with the small jumps of a pure jump Lévy process, from discrete observations of one trajectory. We discuss results both from low and high frequency observations, for Lévy processes possibly of infinite variation. In a low frequency setting, we propose an estimator obtained via a spectral approach which achieves the minimax

rate with respect to the L_2 loss. In a high frequency setting we propose two different estimators, a kernel estimator and one based on a spectral approach, which achieve the same rate of convergence. This rate depends on the sampling scheme and on the Blumenthal-Getoor index of the process. Finally, we discuss the optimality of these results. This is a joint work in collaboration with Céline Duval, and Taher Jalal.

Gonzalez Cazares Jorge: Simulation of the drawdown and its duration in Lévy models via stick-breaking Gaussian approximation

Gonzalez Cazares Jorge[1,2]

1 - The Alan Turing Institute (United Kingdom)
2 - Department of Statistics [Warwick] (United Kingdom)

Abstract: We develop a computational method for expected functionals of the drawdown and its duration in exponential Lévy models. It is based on a novel simulation algorithm for the joint law of the state, supremum and time the supremum is attained of the Gaussian approximation of a general Lévy process. We bound the bias for various locally Lipschitz and discontinuous payoffs arising in applications and analyse the computational complexities of the corresponding Monte Carlo and multilevel Monte Carlo estimators. Monte Carlo methods for Lévy processes (using Gaussian approximation) have been analysed for Lipschitz payoffs, in which case the computational complexity of our algorithm is up to two orders of magnitude smaller when the jump activity is high. At the core of our approach are bounds on certain Wasserstein distances, obtained via the novel SBG coupling between a Lévy process and its Gaussian approximation. Numerical performance, based on the implementation in the dedicated GitHub repository, exhibits a good agreement with our theoretical bounds. This is joint work with Aleksandar Mijatovic.

Lin Feng: Fast exact simulation of the first passage of a tempered stable subordinator

Lin Feng[1]

1 - University of Warwick [Coventry] (United Kingdom)

Abstract: In this paper we construct a fast exact simulation algorithm for the simulation of the first passage time, undershoot and overshoot of a tempered stable subordinator over an arbitrary non-increasing absolutely continuous function. We prove that the running time of our algorithms have finite exponential moments and provide explicit bounds on their expected running time. These algorithms are applied to obtain Monte Carlo estimators for the price of barrier options and for the solutions of fractional partial differential equations.

Mijatovic Aleksandar: Geometrically Convergent Simulation of the Extrema of Lévy Processes

Mijatovic Aleksandar[1,2]

1 - The Alan Turing Institute (United Kingdom)
2 - Department of Statistics [Warwick] (United Kingdom)

Abstract: We develop a novel approximate simulation algorithm for the joint law of the position, the running supremum and the time of the supremum of a general Levy process at an arbitrary finite time. We prove that the error decays geometrically in L^p (for any $p \geq 1$) as a function of the computational cost, in contrast with the polynomial decay for the approximations available in the literature. We establish a central limit theorem and construct non-asymptotic and asymptotic confidence intervals for the corresponding Monte Carlo estimator. We prove that the multilevel Monte Carlo estimator has optimal computational complexity (i.e. of order ϵ^{-2} if the mean squared error is at most ϵ^2) for locally Lipschitz and barrier-type functions of the triplet and develop an unbiased version of the estimator. We illustrate the performance of the algorithm with numerical examples. This is joint work with Jorge Gonzalez Cazares and Geronimo Uribe Bravo.

3.2.10 Contributed session W2-11 (Vallois): Health and healthcare modeling (Room 205)

Wednesday 28th, 13:30 - 15:00 - chaired by Pierre Vallois

Wang Zicheng: Clonal Diversity at Cancer Recurrence

Leder Kevin[1], Wang Zicheng[1]

1 - University of Minnesota (United States)

Abstract: Despite initial success, many cancer therapies eventually fail due to mutation driven drug resistance. In this work we are interested in the status of the tumor at the time of cancer recurrence. In particular we investigate clonal diversity at the time of cancer recurrence, an important factor to consider when treating recurrent cancers. To model cancer dynamics involving drug-sensitive and drug-resistant cells, we employ a two-type branching process. During treatment, drug-resistant cells develop from drug-sensitive cells. We investigate the clonal diversity of drug-resistant cells at the time of cancer recurrence, defined as the first time the population size of drug-resistant cells exceeds a specified proportion of the initial population of drug-sensitive cells. We examine two clonal diversity indices: the number of clones and the Simpson's Index. We obtain the expectation of these two indices and use them to develop statistics that estimate model parameters. We then look at these two indices conditioned on early recurrence in the special case of deterministically decaying sensitive population. We demonstrate that the recurrence time is a useful indicator of clonal diversity, providing new insights into treating recurrent cancers.

Jacko Peter: Bandit Procedures for Designing Patient-Centric Clinical Trials

Jacko Peter[1,2], Villar Sofia S.[3]

1 - Lancaster University Management School (United Kingdom)
2 - Berry Consultants (United Kingdom)
3 - MRC Biostatistics Unit, University of Cambridge (United Kingdom)

Abstract: Multi-armed bandit problems (MABPs) are a special type of optimal control problem that has been studied in the fields of operations research, statistics, machine learning, economics, and others. It is a framework well suited to model resource allocation under uncertainty in a wide variety of contexts. The use of bandit models to optimally design clinical trials is one of the typical motivating application, for designing the so-called patient-centric trials, which would take into account the benefit of the in-trial patients. Nevertheless, the resulting theory has had little influence on the actual design of clinical trials. Contrary to similar learning problems arising for instance in digital marketing where interventions can be tested on millions of users at negligible cost, clinical trials are about "small data", as recruiting patients is remarkably expensive and (in many cases) ethically challenging. Due to the focus on small sizes, we do not resort to the use of the normal distribution to approximate a binomial distribution which is a common practice for large samples either "for simplicity" or "for ease of computation". We evaluate and compare the performance of a variety of operations research and machine learning procedures for the finite-horizon MABP, including the traditional and still dominant clinical trial design choice - equal fixed randomization - and interpret them in the context of designing clinical trials. Our results illustrate how bandit approaches could offer significant advantages, mainly in terms of allocating more patients to better interventions, but still pose important inferential challenges, particularly in terms of their resulting lower statistical power, potential for bias in estimation and existence of closed-form test distributions or asymptotic theory. We illustrate some promising modifications to bandit procedures to address power and bias issues, and we reflect upon the open challenges that remain for an increased uptake of bandit models in clinical trials. Keywords: bandit problem, dynamic resource allocation, index rules, clinical trials

De Turck Koen: Limits of arrival processes into vaccination centers with unpunctual patients.

De Turck Koen[1], Fiems Dieter[1]

1 - Department of TELIN (Ghent University) (Belgium)

Abstract: We study the arrivals into a vaccination center where the time scale of unpunctuality is large as compared to the scheduling interval. We show that in the limit, the process converges to a Poisson process. This can be shown by either classical techniques based on the characteristic function as well as by a Chen-Stein approach, where the latter also provides relatively easy to compute error bounds on the distribution.

We then input the pre-limit arrival process into both a finite as well an infinite server system. The infinite-server case is well-explained by the Poisson limit of the arrival process, but simulation results show that the finite-server system in its entirety does not converge to an equivalent finite-server system with Poisson inputs and consistently shows considerably better performance. We attempt to shed light on this surprising result by taking into account the (negative) correlation structure of the pre-limit arrival process and looking at the large-deviations limits.

Vallois Pierre: Probabilistic Cellular Automata modeling of intercellular interactions in airways: complex pattern formation in patients with Chronic Obstructive Pulmonary Disease

Vallois Pierre[1]

1 - Institut Élie Cartan de Lorraine (France)

Abstract: The chronic obstructive pulmonary disease (COPD) is a highly prevalent lung disease, in which unusual interactions between fibrocytes and CD8+ T lymphocytes in the peribronchial area could induce chronic inflammation and tissue remodeling. We considered a probabilistic cellular automata type model where the two types of cells follow simple local interaction rules taking into account cell death, proliferation, migration and infiltration. A rigorous mathematical analysis carried out within the framework of a streamlined model makes it possible to estimate with precision the parameters of the model using multiscale experimental data obtained in control and disease conditions. The simulation of the model is simple to be implemented. In simulations, two distinct patterns emerged, which can be analyzed quantitatively. In particular, we show that the change in fibrocyte density in the COPD condition is mainly the consequence of their infiltration into the lung during exacerbations, suggesting possible explanations for experimental observations in normal and COPD tissue. Our integrated approach combining probabilistic cellular automata type model and experimental findings will provide further insights into COPD in future studies.

3.3 Wednesday 28th, 15:30 - 17:00

3.3.1 Invited session W3-1 (Dorsman, Righter & Gardner): Load balancing I (Auditorium)

Wednesday 28th, 15:30 - 17:00 - chaired by Jan-Pieter Dorsman, Rhonda Righter & Kristen Gardner

Comte Céline: Reinforcement Learning in Product-Form Queueing Networks

Comte Céline[1], Jonckheere Matthieu[1], Senen-Cerda Albert[2], Sanders Jaron[2]

1 - LAAS-CNRS (France)

2 - Eindhoven University of Technology [Eindhoven] (Netherlands)

Abstract: We introduce a reinforcement-learning gradient-based algorithm to optimize performance in product-form queueing networks. More specifically, our goal is to maximize the undiscounted return in a queueing network whose equilibrium distribution is given, up to a multiplicative constant, as a product of functions that are differentiable with respect to a vector of control parameters. Our main contribution is a novel expression of the gradient of the undiscounted return in the form of a covariance vector for which we can build an unbiased estimator. This estimator is used to optimize the undiscounted return through a stochastic gradient algorithm. The performance of this algorithm is compared numerically to that of classical reinforcement-learning methods on several examples, such as load-balancing in multiserver systems. Our current focus is on (i) proving that this algorithm converges to a (global or local) optimum, and (ii) generalizing our results to other Markov decision processes, provided that we can build an estimator of the gradient of the logarithm of a stationary measure with respect to the vector of control parameters. This is a joint ongoing work with Matthieu Jonckheere (LAAS-CNRS), Jaron Sanders (Eindhoven University of Technology), and Albert Senen Cerda (Eindhoven University of Technology).

Cardinaels Ellen: Heavy-Traffic Analysis of Redundancy Systems without Complete Resource Pooling

Cardinaels Ellen[1], Borst Sem[1], Van Leeuwaarden Johan[2]

1 - Eindhoven University of Technology (Netherlands)

2 - Tilburg University [Tilburg] (Netherlands)

Abstract: The complete resource pooling (CRP) condition has been a common assumption in the heavy-traffic analysis of parallel-server systems. Systems operating under this condition, even when they are subject to highly intricate job-server assignment constraints, typically achieve the same asymptotic performance as a fully flexible single-resource system. In contrast, we examine the asymptotic behavior of parallel-server systems operating under redundancy scheduling policies beyond the typical CRP condition. Building on existing product-form distributions, we derive an insightful closed-form expression for the joint probability generating function of the queue lengths. The heavy-traffic analysis when exploiting the structure in this generating function reveals that the joint distribution of the scaled queue lengths can asymptotically be written in terms of a linear combination of independent and exponentially distributed random variables. Hence, even without the CRP condition, a broad range of systems involving job-server assignment constraints exhibit similar behavior and a partial state space collapse can be observed.

Anton Elene: Novel coupling techniques for the stationary distribution of redundancy-d models with general scheduling policies

Anton Elene[1,2], Gardner Kristen[3]

1 - University of the Basque Country/Euskal Herriko Unibertsitatea (Spain)
2 - Eindhoven University of Technology (Netherlands)
3 - Amherst College (United States)

Abstract: In this talk, we characterize the stationary distribution of redundancy-d systems for a broad family of scheduling policies. Redundancy has gained considerable attention as a dispatching paradigm that promises the potential for significant response time improvements. The redundancy-d policy is of particular interest in the theoretical community. Much of the prior work on redundancy-d -assuming exponentially distributed service times and independent and identically distributed copies- has focused on either deriving response time results under first-come first-served (FCFS) scheduling or stability conditions under a wider range of scheduling policies including processor sharing (PS) and random order of service (ROS). In this talk, we consider the family of scheduling policies that are work-conserving, non-anticipating, non-size based, and non-class based. We introduce a novel stochastic coupling technique that modifies the standard coupled sample paths approach by resampling the system state at key moments in time. This approach enables us to obtain, the first proof that all scheduling policies within the family under consideration admit the same stationary distribution of the number of jobs in the system. We further extend our results to the cancel-on-start redundancy model and show that the stationary distribution of the number of jobs in the system coincides under ROS and FCFS, for a large family of assignment rules including both assign-longest-idle server (ALIS) and random-assignment-to-idle-server (RAIS).

Gardner Kristen: New directions in product forms

Dorsman Jan-Pieter[1], Gardner Kristen[2]

1 - University of Amsterdam (Netherlands)
2 - Amherst College (United States)

Abstract: In recent years, due to a surge of interest in redundancy based systems, a number of models involving different compatibilities between jobs and servers in queueing systems, or between agents and resources in matching systems, have been studied. Under appropriate assumptions, e.g. the service rate functions satisfy order-independent conditions, the stationary distributions of many of these seemingly different systems have been shown to have product forms. While multiple broad frameworks of such product-form models have been proposed, the setup of a framework which captures all such models remains an open problem.

In this presentation, we closely study this open problem. In doing so, we broaden the space of known product form systems by generalizing along several dimensions the recently studied pass-and-swap queue, which itself is a generalization of the order-independent queue. We also identify several important questions that remain open and present a roadmap for future study.

3.3.2 Invited session W3-2 (Burnetas): Strategic queueing III (Room 101)

Wednesday 28th, 15:30 - 17:00 - chaired by Apostolos Burnetas

Burnetas Apostolos: Price and Capacity Competition between a Make-to-Order and a Make-to-Stock Firm with Strategic Customers

Benioudakis Myron[1,2], Deligiannis Michalis[1], Burnetas Apostolos[3], Liberopoulos George[1]

1 - University of Thessaly (Greece)
2 - Athens University of Economics and Business (Greece)
3 - National and Kapodistrian University of Athens (Greece)

Abstract: We propose a model of two firms competing in a market of time-sensitive strategic customers. One firm is a make-to-order producer operating on a first-come-first-served basis, while the other is a make-to-stock supplier with a fixed order cost, negligible replenishment lead time, and immediate delivery. The products sold by the two firms are substitutes but not identical, and customers choose which product to buy based on their valuations, selling prices, and estimated waiting time at the make-to-order producer. We consider the competition between the two firms in a stylized model of queueing and inventory control policies under strategic customer behavior. The two firms determine their selling prices and the production capacity and ordering policies, respectively. We analyze Nash equilibrium strategies between the two firms and the customers under several competition frameworks, explore the social welfare

under equilibrium, and discuss coordination possibilities. A practical application is a market of custom-made and ready-made products or new and second-hand products, where the high price and delay in acquiring the custom-made product may drive some customers to buy the less expensive ready-made product.

Liberopoulos George: Price and Capacity Optimization in Service Systems with Time-Sensitive, Mixed-Risk Customers

Deligiannis Michalis[1], Benioudakis Myron[1,2], Liberopoulos George[1], Burnetas Apostolos[3]

1 - Department of Mechanical Engineering, University of Thessaly (Greece)
2 - Athens University of Economics and Business (Greece)
3 - National and Kapodistrian University of Athens (Greece)

Abstract: We study a model of a profit-maximizing firm providing service to time-sensitive customers who must decide whether to join the queue or not. Customers arrive one at a time and cannot observe the system state, relying on their expected utility to decide. If they join, they exhibit mixed-risk behavior towards the total waiting time. They are risk-seekers if the waiting time is below a critical threshold, and risk-averse if it is above it. This behavior affects the customers' perception of waiting time and is due to two opposing forces: impatience linked to the service's value and frustration related to waiting. Each joining customer's utility is the value of service minus the price and waiting cost per unit of perceived delay. The firm must determine the service price and/or service capacity to maximize its expected profit given the customers' arrival rate. We identify the customers' symmetric aggregate Nash equilibrium join/balk strategy, derive the firm's optimal service price and/or capacity, and perform a sensitivity analysis of the firm's performance with respect to the customers' critical threshold.

Oz Binyamin: Strategic flexibility in service systems

Kerner Yoav[1], Oz Binyamin[2], Shneer Seva[3]

1 - Ben-Gurion University of the Negev (Israel)
2 - The Hebrew University of Jerusalem (Israel)
3 - Heriot-Watt University [Edinburgh] (United Kingdom)

Abstract: We study a multi-server system. Customers may join any unobservable queue in front of each server or join all of them simultaneously. In the latter case, they will be served by any available server and removed immediately from the all other queues upon service completion. Equilibrium and socially optimal strategies are studied.

Manou Athanasia: Heterogeneous Strategic Customers in an M/M/1 queue: Price of Ignoring Heterogeneity

Economou Antonis[1], Kanavetas Odysseas[2], Manou Athanasia[1]

1 - Department of Mathematics, National and Kapodistrian University of Athens (Greece)
2 - Mathematical Institute, Leiden University (Netherlands)

Abstract: We consider the unobservable M/M/1 queue with strategic customers. We assume that customers decide whether to join or balk based on their expected utility that depends on the reward from service, the fee, and the waiting cost. Customers are heterogeneous in their waiting cost rate and their reward from service. Specifically, we assume that a customer's reward from service is increasing in his/her waiting cost rate. We determine equilibrium customer behavior and calculate the expected revenue of the administrator and the expected welfare under equilibrium. We also calculate the optimal fee imposed by a revenue-maximizing administrator. We explore the effect of customer heterogeneity on the administrator revenue, the social welfare and the optimal fee. Finally, we determine the price of ignoring heterogeneity by comparing the performance of the system when the administrator takes into account customer heterogeneity with the performance when he/she ignores heterogeneity.

3.3.3 Invited session W3-3 (Maguluri): Stochastic processing networks (Room 102)

Wednesday 28th, 15:30 - 17:00 - chaired by Siva Theja Maguluri

Borst Sem: Load Balancing with Sparse Dynamic Graphs

Goldsztajn Diego[1], Borst Sem[1], Van Leeuwaarden Johan[2]

1 - Eindhoven University of Technology (Netherlands)
2 - Tilburg University (Netherlands)

Abstract: Load balancing algorithms play a crucial role in distributing service requests in large-scale parallel-resource systems such as data centers and cloud networks. Due to the massive size of these systems, implementation complexity of load balancing algorithms has emerged as a critical concern besides conventional performance metrics such as delay. In this talk we examine novel load balancing algorithms which use random graphs to guide local load exchanges between queues, focusing on sparse topologies to ensure scalability. We establish fluid limits which explicitly depend on the probability law of the random graphs through the generating function of its limiting degree distribution. The fluid limits demonstrate that these local algorithms match the performance of their global counterparts when the random graph is resampled at a sufficiently high rate. We further show that performance degrades if queues are isolated with a positive probability and that performance improves as the degree distribution becomes more concentrated. Note: Based on joint work with Diego Goldsztajn (TU/e) and Johan van Leeuwaarden (Tilburg University).

Dai Jim: Asymptotic steady-state independence for generalized Jackson networks in multi-scale heavy traffic

Dai Jim[1], Glynn Peter, Xu Yaosheng

1 - Cornell University [New York] (United States)

Abstract: We prove that under a multi-scale heavy traffic condition, the stationary distribution of the scaled queue length process in any generalized Jackson network has a product-form limit. Each component in the product-form has an exponential distribution, corresponding to the Brownian approximation of a single station queue. Each "single station" can be constructed precisely and its parameters have a good intuitive interpretation. The proof employs a basic adjoint relationship (BAR)-approach that has been advanced recently by Braverman, Dai, and Miyazawa (2017, 2023). This is a joint work with Peter Glynn at Stanford and Yaosheng Xu at Amazon.

Jhunjhunwala Prakirt: Heavy Traffic Joint Queue Length Distribution without Resource Pooling

Jhunjhunwala Prakirt[1]

1 - Georiga Institute of Technology (United States)

Abstract: This work studies the heavy-traffic joint distribution of queue lengths in a well-known stochastic processing network (SPN), viz., an input-queued switch, operating under the MaxWeight scheduling policy. An input-queued switch serves as a representative of SPNs that do not satisfy the so-called Complete Resource Pooling (CRP) condition, and consequently exhibit a multidimensional State Space Collapse (SSC). Except in special cases, only the mean queue lengths of such non-CRP system is known in the literature. In this work, we develop the Transform method to study the input-queued switch. The key challenge is in solving an implicit functional equation involving the Laplace transform of the heavy-traffic limiting distribution. Under a conjecture on uniqueness of the solution of the functional equation, we obtain an exact joint distribution of the heavy-traffic limiting queue-lengths in terms of a non-linear combination of i.i.d. exponentials.

Varma Sushil: Power-of-d Choices Load Balancing in the Sub-Halfin Whitt Regime

Varma Sushil Mahavir[1], Castro Francisco[2], Maguluri Siva Theja[1]

1 - Georgia Institute of Technology [Atlanta] (United States)

2 - UCLA Anderson School of Management (United States)

Abstract: We consider the load balancing system under Poisson arrivals, exponential services, and homogeneous servers. Upon arrival, a job is to be routed to one of the servers, where it is queued until service. We consider the Power-of-d choices routing algorithm, which chooses the queue with minimum length among d randomly sampled queues. We study this system in the many-server heavy-traffic regime where the number of servers goes to infinity simultaneously when the load approaches the capacity. In particular, we consider a sequence of systems with n servers, where the arrival rate of the nth system is $\lambda = n - n^{1-\gamma})$ for some $\gamma \in (0, 0.5)$. This is known as the sub-Halfin-Whitt regime. It was shown in [Liu Ying (2020)] that under Power-of-d choices routing with $d \geq n^{\gamma} \log n$, the queue length behaves similarly to that of JSQ, and there are asymptotically zero queueing delays. The focus of this paper is to characterize the behavior when d is below this threshold. We obtain high probability bounds on the queue lengths for various values of d and large enough n. In particular, we show that when d grows polynomially in n, but slower than in [Liu Ying (2020)], i.e., if d is $\theta(n^{\gamma} \log n)^{1/m}$ for some integer $m > 1$, then the asymptotic queue length is m with high probability. This finite queue length behavior is similar to that of JSQ in the so-called nondegenerate slowdown regime (where $\gamma = 1$). Moreover, if d grows

polylog in n, i.e. slower than any polynomial, but is at least $\Omega(\log(n)^3)$, the queue length blows up to infinity asymptotically similar to that under JSQ in the so-called super slow down regime ($\gamma > 1$). We obtain these results by deploying the iterative state space collapse approach that was first developed by [Liu Gong Ying (2022)]. We first establish a weak state-space collapse on the queue lengths. We then bootstrap from this weak collapse to iteratively narrow down the region of the collapse. After enough steps, this inductive refinement of the collapse provides the bounds that we seek. These sequences of collapses are established using Lyapunov drift arguments.

3.3.4 Invited session W3-4 (Cohen): Mean-Field Limits and their Applications (Room 103)

Wednesday 28th, 15:30 - 17:00 - chaired by Asaf Cohen

Kobeissi Ziad: The tragedy of the commons: a Mean-Field Game approach to the reversal of travelling waves

Kobeissi Ziad[1,2]

1 - Inria Paris (Sierra team) (France)
2 - Institut Louis Bachelier (France)

Abstract: The goal of this talk is to investigate an instance of the tragedy of the commons in the optimal management of fisheries. The model we choose is that of a fishes' population that is governed by a bistable equation and that fishermen harvest. We assume that, when no fishermen are present, the fishes' population is invading (mathematically, there is an invading travelling front). Is it possible that fishermen, when acting selfishly, each in his or her own best interest, might lead to a reversal of the travelling wave and, consequently, to an extinction of the global population? To answer this question, we model the behaviour of individual fishermen using a Mean Field Game approach, and we show that the answer is yes. We then show that, at least in some cases, if the fishermen, instead of acting selfishly, coordinated, each of them could harvest a greater number of fishes, while still guaranteeing the survival of the population. Our study is illustrated by several numerical simulations. This is a joint work with Idriss Mazari-Fouquet and Domenec Ruiz-Balet.

Lauriere Mathieu: Deep Backward and Galerkin Methods for Learning Finite State Master Equations

Lauriere Mathieu[1]

1 - NYU Shanghai (China)

Abstract: We study two methods designed to efficiently solve high-dimensional PDEs in the setting of mean field master equations. Such master equations are independently interesting because they fully characterize the value of a game with a large number of players and have a myriad of applications in economics, finance, epidemiology, and more. The first method we explore is the deep backward dynamic programming (DBDP) method-we derive BSDEs for the master equation along some exploratory process and use this structure to train neural networks to approximate the master equation solution. Then we revisit the deep Galerkin method (DGM) as applied to the master equation and compare the efficacy of the two methods.

Yeung Lane Chun: A non-asymptotic perspective on mean field control via log-concavity

Yeung Lane Chun[1], Lacker Daniel[1], Mukherjee Sumit[2]

1 - Department of Industrial Engineering and Operations Research, Columbia University (United States)
2 - Department of Statistics, Columbia University (United States)

Abstract: We study a class of stochastic control problems in which a large number of players cooperatively choose their drifts to maximize an expected reward minus a quadratic running cost. For a broad class of potentially asymmetric rewards, we show that there exist approximately optimal controls which are decentralized, in the sense that each player's control depends only on its own state and not the states of the other players. Moreover, the optimal decentralized controls can be constructed non-asymptotically, without reference to any mean field limit. These results descend from a broader framework inspired by the theory of nonlinear large deviations of Chatterjee-Dembo, for which we offer an efficient non-asymptotic perspective in log-concave settings based on functional inequalities. Joint work with Daniel Lacker and Sumit Mukherjee.

Cohen Asaf: Markovian Equilibria In Ergodic Many-Player Games and Mean-Field Games

Cohen Asaf[1], Zell Ethan[1]

1 - Department of Mathematics - University of Michigan (United States)

Abstract: We consider a symmetric stochastic game with weak interactions between many players. Time is continuous, the number of states is finite, and costs are ergodic. We prove the existence of a unique Nash equilibrium in the game and show that its limiting behavior (as the number of players goes to infinity) is governed by the unique mean-field equilibrium of the corresponding mean-field game.

3.3.5 Invited session W3-5 (Olvera-Cravioto): Processes on graphs (Room 104)

Wednesday 28th, 15:30 - 17:00 - chaired by Mariana Olvera-Cravioto

Britton Tom: Semi-directed networks and their relation to epidemic models

Britton Tom[1]

1 - Stockholm University (Sweden)

Abstract: First we describe the configuration model for a semi-directed network consisting of both directed and undirected edges, and some large population properties of such networks. Then we show that the final outcome of a surprisingly large class of epidemic models can be characterized by such semi-directed networks.

Diallo Aoudi Mohamed Habib Aliou: Large graph limit of the matching coverage of local online algorithms on the Configuration model

Diallo Aoudi Mohamed Habib Aliou[1], Moyal Pascal[2], Robin Vincent[1]

1 - Laboratoire de Mathématiques Appliquées de Compiègne (France)
2 - Institut Élie Cartan de Lorraine (France)

Abstract: In this work, we study the hydrodynamic limits of various 'local' online matching algorithms on random graphs generated by the Configuration Model. By exploiting the Markov structure of a measure-valued process representing the joint construction of the graph together with an epxloration algorithm that constructs a matching on that graph, we obtain an approximation of the dynamics by a family of ordinary differential equations, for which a simple function of the solution provides the large-graph limits of the proportion of matched nodes. We then conduct a comparison of this performance metrics for various online matching algorithms, and study the sensibility on the degree distribution of the graph under consideration.

Ramanan Kavita: Large deviations for interacting diffusions on local tree-like graphs

Ramanan Kavita[1], Yasodharan Sarath[1], Chen I-Hsun[1], Salkeld William[1]

1 - Brown University (United States)

Abstract: We establish large deviation principles for a general class of interacting (reflected) diffusions on sparse graphs. A key ingredient of the proof, which is of possible independent interest, is a tractable characterization of the large deviation rate function for marked locally tree-like graphs, which can be viewed as a generalization of Sanov's theorem. The talk is based on joint works with I.-H. Chen, W. Salkeld and S. Yasodharan.

Olvera-Cravioto Mariana: Opinion dynamics on directed complex networks

Olvera-Cravioto Mariana[1], Lin Tzu-Chi[1], Fraiman Nicolas[1]

1 - University of North Carolina, Chapel Hill (United States)

Abstract: We study a general version of the classical DeGroot-Friedkin model on sparse random digraphs whose local weak limit is a marked Galton-Watson process. This model has been extensively used in the social sciences to study opinion dynamics on social networks. The type of graphs in our analysis include the directed versions of the Erdos-Renyi graph, the Chung-Lu model, the Norros-Reittu model, and the configuration model, among others. We show that the stationary distribution of a typical vertex in the graph converges in distribution to a random variable that can be constructed on the limiting tree. Furthermore, our approach allows us to provide exact formulas for the mean and variance of the limiting

stationary distribution, which can be used to characterize the conditions that lead to either "consensus" or "polarization". In particular, we explain how cognitive biases and the presence of stubborn agents (vertices that cannot be influenced) have on the evolution of opinions, in some cases promoting polarization and in others reducing it.

3.3.6 Invited session W3-6 (Xu): Experimentation and Learning with Stochastic Models (Room 105)

Wednesday 28th, 15:30 - 17:00 - chaired by Kuang Xu

Agarwal Anish: On Causal Inference with Temporal and Spatial Spillovers in Panel Data

Agarwal Anish[1]

1 - Columbia University (United States)

Abstract: Panel data is a ubiquitous setting where one collects multiple measurements over time of a collection of heterogeneous units (e.g., individuals, firms, geographic entities). Two pervasive challenges in doing causal inference in such settings are: (1) temporal spillovers - interventions in the past affect current outcomes; (2) spatial spillovers - interventions a unit receives affects the outcomes of its neighboring units. We develop a unified causal framework to tackle these two challenges. We do so by proposing a novel latent factor model that allows one to share information across units, measurements, and interventions, and incorporates both types of spillovers. The model subsumes linear time-varying dynamical systems and autoregressive processes as a special case, and previously studied models for spatial spillovers. We show how a variety of causal parameters are identified and estimated despite unobserved confounding.

Li Hannah: Scaling Up Healthcare Interventions: Insights from Experimentation

Li Hannah[1,2]

1 - Columbia University (United States)
2 - Massachusetts Institute of Technology (United States)

Abstract: A common concern about social interventions, often found in domains like public health, education, and vocational training, is that the interventions that seem promising in randomized controlled trials (RCTs) may not perform well when scaled to larger populations. Although many factors contribute to these difficulties in scaling up interventions, in this work we highlight and isolate the effects of one such factor: capacity constraints. We consider a case study of a mobile health platform designed to improve patients' adherence to treatment protocols. We model the dynamics of this platform as a queueing system and investigate patient adherence in such a system in (smaller) RCTs versus when the platform is scaled out to a larger population. We show that, due to capacity constraints on the system, the effect observed in an RCT may diminish when servicing a larger population. However, we show that the platform can utilize data from the RCT to inform how to scale to a larger population and still ensure performance close to that observed in the RCT.

Li Shuangning: Experimenting under Stochastic Congestion

Li Shuangning[1], Johari Ramesh[2], Wager Stefan[2], Xu Kuang[2]

1 - Harvard University (United States)
2 - Stanford University (United States)

Abstract: Stochastic congestion, a phenomenon in which a system becomes temporarily overwhelmed by random surges in demand, occurs frequently in service applications. While randomized experiments have been effective in gaining causal insights and prescribing policy improvements in many domains, using them to study stochastic congestion has proven challenging. This is because congestion can induce interference between customers in the service system and thus hinder subsequent statistical analysis. In this paper, we aim at getting tailor-made experimental designs and estimators for the interference induced by stochastic congestion. In particular, taking a standard queueing system as a benchmark model of congestion, we study how to conduct randomized experiments in a service system that has a single queue with an outside option. We study switchback experiments and a local perturbation experiment and propose estimators based on the experiments to estimate the effect of a system parameter on the average arrival rate. We establish that the estimator from the local perturbation experiment is asymptotically more accurate than the estimators from the switchback experiments because it takes advantage of the structure of the queueing system.

Xu Kuang: Service Slowdown Detection

Xu Kuang[1], Mendelson Gal[2]

1 - Stanford University (United States)
2 - Technion (Israel)

Abstract: Resource allocation control mechanisms and statistical analysis play key roles in the operation of dynamic service systems. Control is used to achieve desired performance and analysis is used for statistical tasks such as fault detection and parameter estimation. While control and analysis have been studied separately to a large extent, little is known on the interplay between the two. In this work, we derive principles for performing online detection of server slowdown in dynamic service systems. We show that the choice of control has a drastic impact on the subsequent statistical analysis. Specifically, we show that congestion based statistics can fail to detect server slowdown when applied to the data generated by systems which use popular adaptive control schemes. We propose using the controller's action data as a new statistic for analysis and prove its effectiveness.

3.3.7 Invited session W3-7 (Goulart): Random matrices and random tensors in data analysis and machine learning (Room 106)

Wednesday 28th, 15:30 - 17:00 - chaired by Henrique Goulart

Lebeau Hugo: Truncated HOSVD : An Analysis of a Fast Tensor Approximation

Lebeau Hugo[1]

1 - Université Grenoble Alpes (France)

Abstract: Tools from random matrix theory allow us to study a tensor spiked model describing the noisy observation of a low multilinear-rank signal. A precise characterization of the spectral behavior of the tensor unfoldings reveals a phase-transition phenomenon below which no information can be recovered. This analysis gives the exact performance of the approximation with truncated HOSVD. It paves the way towards a precise study of low-rank tensor approximation.

Mai Xiaoyi: A high dimensional asymptotic view on learning from large data sets

Mai Xiaoyi[1]

1 - Dept. Math-Info (France)

Abstract: In modern machine learning, we often encounter data sets with comparably large numbers of samples and features. It is argued that classical statistical learning theory is insufficient for explaining the generalization performance on such large data sets, due to the curse of dimensionality and the overfitting problem of overparametrized learning models. The approach of high dimensional asymptotic analyses is motivated by the need of modeling and understanding modern machine learning. To apply this increasingly popular approach, we develop a flexible framework based on the leave-one-out perturbation, notably capable of handling implicit optimization and iterative procedure involved in many learning algorithms. Our analyses allow for original understandings on fundamental questions such the challenge of high dimensional learning, the role of loss function and the effectiveness of learning with less labels. Practical improvements are proposed based on the insights gained from these analyses.

Seddik Mohamed El Amine: Asymptotics of Asymmetric Random Tensors

Seddik Mohamed El Amine[1]

1 - Technology Innovation Institute (United Arab Emirates)

Abstract: The purpose of this talk is to conduct an asymptotic analysis of large asymmetric spiked tensor models using random matrix theory. The presentation will be divided into two parts. In the first part, we will focus on analyzing a rank-one model and discussing its algorithmic implications, specifically in terms of signal recovery using a polynomial time algorithm. In the second part of the presentation, we will broaden the analysis to include a low-rank spiked model with non-independent components, through an analysis of tensor deflation procedures.

Zhu Yizhe: Overparameterized random feature regression with nearly orthogonal data

Wang Zhichao[1], Zhu Yizhe[2]

1 - University of California [San Diego] (United States)
2 - University of California [Irvine] (United States)

Abstract: We investigate the properties of random feature ridge regression (RFRR) given by a two-layer neural network with random Gaussian initialization. We study the non-asymptotic behaviors of the RFRR with nearly orthogonal deterministic unit-length input data vectors in the overparameterized regime, where the width of the first layer is much larger than the sample size. Our analysis shows high-probability non-asymptotic concentration results for the training errors, cross-validations, and generalization errors of RFRR centered around their respective values for a kernel ridge regression (KRR). This KRR is derived from an expected kernel generated by a nonlinear random feature map. We then approximate the performance of the KRR by a polynomial kernel matrix obtained from the Hermite polynomial expansion of the activation function, whose degree only depends on the orthogonality among different data points. This polynomial kernel determines the asymptotic behavior of the RFRR and the KRR. Our results hold for a wide variety of activation functions and input data sets that exhibit nearly orthogonal properties. Based on these approximations, we obtain a lower bound for the generalization error of the RFRR for a nonlinear student-teacher model.

3.3.8 Invited session W3-8 (Banerjee): Online decision making (Room 202)

Wednesday 28th, 15:30 - 17:00 - chaired by Sid Banerjee

Braun Alexander: Simplified Prophet Inequalities for Combinatorial Auctions

Braun Alexander[1]

1 - University of Bonn (Germany)

Abstract: In this talk, we consider online resource allocation problems where the input is drawn from known probability distributions. First, we focus on the special case of online bipartite max-weight matching: Nodes on one side (a.k.a. items) are available offline. Nodes on the other side (a.k.a. buyers) arrive one-by-one, incident edge-weights are sampled, and an irrevocable matching decision needs to be made. For this problem, we discuss a pricing-based algorithm: For every offline node, we set a price upfront based on the distributions. Then, any arriving online node is matched to maximize surplus given the prices. We give a simplified proof for the existence of static item prices which recover the tight competitive ratio from a paper of Feldman, Gravin and Lucier from 2015. We will then extend the techniques to more general settings where we can allocate multiple items per round and derive state-of-the-art competitive ratios for these. Based on joint work with Thomas Kesselheim.

Lomys Niccolò: A Mediator Approach to Mechanism Design with Limited Commitment

Lomys Niccolò[1], Yamashita Takuro[2]

1 - Center for Studies in Economics and Finance & Università di Napoli Federico II (Italy)
2 - Osaka School of International Public Policy, Osaka University (Japan)

Abstract: We study mechanism design with limited commitment. In each period, a principal offers a "spot" contract to a privately informed agent without committing to future contracts. In contrast to the classical model with a fixed information structure, we allow for all admissible information structures. We represent the information structure as a fictitious mediator and re-interpret the model as a mechanism design problem for the committed mediator. We construct examples to explain why new equilibrium outcomes can arise when considering general information structures. Next, we apply our approach to durable-good monopoly. In the seller-optimal mechanism, trade dynamics and welfare substantially differ from those in the classical model: the seller offers a discount to the high-valuation buyer in the initial period, followed by the high surplus-extracting price until an endogenous deadline, when the buyer's information is revealed without noise. The Coase conjecture fails. We also discuss unmediated implementation of the seller-optimal outcome.

Maggiori Andreas: Online Correlation Clustering

Maggiori Andreas[1]

1 - EPFL (Switzerland)

Abstract: We consider the classic correlation clustering objective in fully dynamic vertex streams: At each time a new vertex is added to or deleted from the stream and the goal is to maintain a solution to correlation clustering, i.e.: a partition of the vertex set. Each vertex is connected to each other vertex with either a + or - edge and the goal is to maintain a partition of the vertices that minimizes the number of + edges across clusters plus the number of - edges within clusters, at each time throughout the update sequence. We show that one can maintain an O(1)-approximation with O(polylog(n)) amortized update time in the sublinear regime (where we can make degree queries and sample neighbors in O(1) time). Prior to our work, the best known result for vertex streams was that of Behnezhad, Derakhshan, Hajiaghayi Stein and Sudan [FOCS'19], who showed how to maintain a 3-approximation in $O(log^4 n)$-update time in edge streams (which implies an $O(Dlog^4 n)$-update time 3-approximation for vertex streams of max degree D).

Lykouris Thodoris: Group fairness in dynamic refugee assignment

Lykouris Thodoris[1]

1 - Massachusetts Institute of Technology (United States)

Abstract: Ensuring that refugees and asylum seekers thrive (e.g., find employment) in their host countries is a profound humanitarian goal, and a primary driver of employment is the geographic location within a host country to which the refugee or asylum seeker is assigned. Recent research has proposed and implemented algorithms that assign refugees and asylum seekers to geographic locations in a manner that maximizes the average employment across all arriving refugees. While these algorithms can have substantial overall positive impact, using data from two industry collaborators we show that the impact of these algorithms can vary widely across key subgroups based on country of origin, age, or educational background. Thus motivated, we develop a simple and interpretable framework for incorporating group fairness into the dynamic refugee assignment problem. In particular, the framework can flexibly incorporate many existing and future definitions of group fairness from the literature (e.g., minmax, randomized, and proportionally-optimized within-group). Equipped with our framework, we propose two bid-price algorithms that maximize overall employment while simultaneously yielding provable group fairness guarantees. Through extensive numerical experiments using various definitions of group fairness and real-world data from the U.S. and the Netherlands, we show that our algorithms can yield substantial improvements in group fairness compared to state-of-the-art algorithms with only small relative decreases ($\approx$ 1%-2%) in global performance.

3.3.9 Invited session W3-9 (Daw & Yom-Tov): Hawkes Processes: New Models and Theory (Room 203)

Wednesday 28th, 15:30 - 17:00 - chaired by Andrew Daw & Galit Yom-Tov

Chen Xinyun: Diffusion limit and staffing of many-server queues with Hawkes input

Chen Xinyun[1], Gong Kang[1]

1 - The Chinese University of Hong Kong, Shenzhen (China)

Abstract: To investigate the impact of autocorrelation in the arrival process to the performance of queueing systems, we consider a single-class many-server system in which customers arrive according to a Hawkes process. We propose a new asymptotic regime so that the limiting arrival process still possesses a nontrivial autocorrelation structure. We derive the diffusion limit of the queue length process under this regime and obtain the heavy-traffic staffing level that is different from the classic square-root rule.

Daw Andrew: Conditional Uniformity and Hawkes Processes

Daw Andrew[1]

1 - University of Southern California (United States)

Abstract: Classic results show that the Hawkes self-exciting point process can be viewed as a collection of temporal clusters, in which exogenously generated initial events give rise to endogenously driven descendant events. This perspective provides the distribution of a cluster's size through a natural connection to branching processes, but this is irrespective of time. Insight into the chronology of a Hawkes process cluster has been much more elusive. Here, we employ this cluster perspective and a novel adaptation of the random time change theorem to establish an analog of the conditional uniformity property enjoyed by Poisson processes. Conditional on the number of epochs in a cluster, we show that the transformed times are jointly uniform within a particular convex polytope. Furthermore, we find that this polytope

leads to a surprising connection between these continuous state clusters and parking functions, discrete objects central in enumerative combinatorics and closely related to Dyck paths on the lattice. In particular, we show that uniformly random parking functions constitute hidden spines within Hawkes process clusters. This yields a decomposition that is valuable both methodologically and practically, which we demonstrate through application to the popular Markovian Hawkes model and proposal of a flexible and efficient simulation algorithm.

Mandjes Michel: Infinite-server queues with Hawkes input

Mandjes Michel[1]

1 - University of Amsterdam (Netherlands)

Abstract:

In this talk I'll discuss an analysis of the number of customers in infinite-server queues with a muktivariate self-exciting (Hawkes) arrival process. It is initially assumed that service requirements are exponentially distributed and that the Hawkes arrival process is of a Markovian nature. I obtain a system of differential equations that characterizes the joint distribution of the arrival intensity and the number of customers. Moreover, I provide a recursive procedure that explicitly identifies (transient and stationary) moments. Subsequently, I'll allow for non-Markovian Hawkes arrival processes and non-exponential service times. By viewing the Hawkes process as a branching process, I'll show that the probability generating function of the number of customers in the system can be expressed in terms of the solution of a fixed-point equation. I'll also discuss various asymptotic results: I derive the tail of the distribution of the number of customers for the case that the intensity jumps of the Hawkes process are heavy-tailed, and I also consider a heavy-traffic regime.

Weber Thomas: Estimation of Self-Exciting Point Processes from Time-Censored Data

Schneider Philipp[1], Weber Thomas[1]

1 - EPFL (Switzerland)

Abstract: Self-exciting point processes, widely used to model arrival phenomena in nature and society, are often difficult to identify. The estimation becomes even more challenging when arrivals are recorded only as bin counts on a finite partition of the observation interval. In this work, we propose the Recursive Identification with Sample Correction' (RISC) algorithm for the estimation of process parameters from time-censored data. In every iteration, a synthetic sample path is generated and corrected to match the observed bin counts. Then the process parameters update, and a new iteration is performed, to successively approximate the stochastic characteristics of the observed process. In terms of finite-sample approximation error, the proposed framework performs favorably over extant methods, including a naive locally uniform sample redistribution. The results of an extensive numerical experiment indicate that the reconstruction of an intra-bin history based on the conditional intensity of the process is crucial for obtaining the optimal performance in terms of estimation error.

3.3.10 Invited session W3-10 (Brown): Advances in Experimentation and Dynamic Models (Room 204)

Wednesday 28th, 15:30 - 17:00 - chaired by David Brown

Haugh Martin: Counterfactual Analysis in Dynamic Latent State Models

Haugh Martin[1]

1 - Imperial College London, Department of Analytics, Marketing and Operations (United Kingdom)

Abstract: We provide an optimization-based framework to perform counterfactual analysis in a dynamic model with hidden states. Our framework is grounded in the "abduction, action, and prediction" approach to answer counterfactual queries and handles two key challenges where (1) the states are hidden and (2) the model is dynamic. Recognizing the lack of knowledge on the underlying causal mechanism and the possibility of infinitely many such mechanisms, we optimize over this space and compute upper and lower bounds on the counterfactual quantity of interest. Our work brings together ideas from causality, state-space models, simulation, and optimization, and we apply it on a breast cancer case study. To the best of our knowledge, we are the first to compute lower and upper bounds on a counterfactual query in a dynamic latent-state model. (Joint work with Raghav Singal, Dartmouth)

Miao Sentao: Adaptive Lagrangian Policies for Multi-Warehouse Multi-Store Inventory System with Lost Sales

Chao Xiuli[1], Jasin Stefanus[1], Miao Sentao[2,3]

1 - University of Michigan (United States)
2 - McGill University (Canada)
3 - University of Colorado Boulder (United States)

Abstract: We consider the inventory control problem of a Multi-Warehouse Multi-Store (MWMS) system over a time horizon that the warehouses receive no external replenishment. This problem is prevalent in retail settings, and it is referred to in Jackson (1988) as "what to do until your (external) shipment comes in". The warehouses are stocked with initial inventories and the stores dynamically replenish inventory from the warehouses in each period of the planning horizon. Excess demand in each period at any store is lost. The optimal policy for this problem is complex and state-dependent, and due to the curse of dimensionality, computing the optimal policy using standard dynamic programming is numerically intractable. Static Lagrangian Base Stock (LaBS for short) policy has been developed for this problem (Miao et al. 2022) and shown to be asymptotically optimal. In this paper, we develop adaptive policies by dynamically adjusting the control parameters of a vanilla static LaBS policy using realized historical demands. We show, both theoretically and numerically, that the adaptive policies significantly improve the performance of LaBS policy, with the magnitude of improvement characterized by the number of times the policy can be adjusted. In particular, when the number of adjustment is a logarithm of the length of time horizon, the policy is rate-optimal in the sense that the rate of the loss (against the performance benchmark in terms of the length of the time horizon) matches that of the theoretical lower bound for any policy. Our analysis highlights the importance of a careful policy adjustment scheme to enjoy the most benefit of "pooling effect".

Zoumpoulis Spyros: A Sample Size Calculation for Training and Certifying Targeting Policies

Simester Duncan[1], Timoshenko Artem[2], Zoumpoulis Spyros[3]

1 - Sloan School of Management (United States)
2 - Northwestern University [Evanston] (United States)
3 - Institut Européen dádministration des Affaires (France)

Abstract: We propose an approach for determining the sample size required when using an experiment to train and certify a targeting policy. Calculating the rate at which the performance of a targeting model improves with additional training data is a complex problem. We approximate the problem by grouping customers into segments. This allows us to develop computationally efficient algorithms that calculate the required sample size using simulations or closed-form analytical approximations. We consider two problem formulations. The first formulation focuses on training a targeting policy that satisfies a predefined performance threshold. The second formulation involves out-of-sample certification of a trained targeting policy in a statistical test. We illustrate our analysis using data from a luxury fashion retailer.

Brown David: Sequential Search with Acquisition Uncertainty

Brown David[1], Uru Cagin

1 - Duke university [Durham] (Durham, NC 27708 United States)

Abstract: We study a variation of the classical Pandora's problem in which a decision-maker (DM) sequentially explores alternatives from a given set and learns their values while trying to acquire the best alternative. The variations in the model we study are (i) alternatives randomly become unavailable during exploration and (ii) the DM's ability to acquire a remaining alternative is uncertain and depends on a chosen offer price. Such acquisition uncertainties arise in many applications, including housing search, hiring problems, and e-commerce, but greatly complicate the search problem in that optimal policies retain all previously explored alternative values as part of the problem state, as opposed to only the highest explored value as in Pandora's rule. Our central insight is that despite the complexity that these acquisition uncertainties create, simple greedy policies based on static sequencing and a single threshold value enjoy strong performance guarantees. We develop such a class of policies and show how to compute them using a greedy algorithm whose worst-case run-time scales linearly (up to logarithmic factors) in the number of alternative types. We show that our policies (a) are asymptotically optimal in high multiplicity regimes with many alternatives and (b) obtain at least 63.2% of the optimal value under a broad set of conditions. Extensive numerical examples support this theory: we illustrate our policies on a variation of Pandora's problem with disappearing alternatives and housing search on models calibrated on data

from the online brokerage Redfin. In these examples, our policies significantly outperform policies based on Pandora's rule.

3.3.11 Contributed session W3-11 (Garcia): Stochastic analysis and estimation (Room 205)

Wednesday 28th, 15:30 - 17:00 - chaired by Celia Garcia

Gerencsér Balázs: Convergence and cutoff analysis for a class of random walks inspired by almost exchangeable data

Gerencsér Balázs[1], Ottolini Andrea[2]

1 - Alfréd Rényi Institute of Mathematics (Hungary)
2 - University of Washington (United States)

Abstract: We analyze the convergence rates for a family of auto-regressive Markov chains in a hypercube depending on a noise parameter c, where at each step a randomly chosen coordinate is replaced by a noisy damped weighted average of the others. The interest in the model comes from the connection with a certain Bayesian scheme introduced by de Finetti in the analysis of partially exchangeable data. We show that when c gets small (corresponding to a vanishing noise), a cutoff phenomenon occurs. However, without the damping, we get slower, diffusive mixing with no cutoff. Joint work with Andrea Ottolini.

Levering Nikki: Dynamic Routing in a Road Traffic Network

Levering Nikki, Boon Marko[1], Mandjes Michel, Nunez-Queija Rudesindo

1 - Eindhoven University of Technology (Eindhoven University of Technology Den Dolech 2 5612 AZ Eindhoven Netherlands)

Abstract: Despite measures to reduce congestion, occurrences of both recurrent and non-recurrent congestion are increasing in frequency, resulting in higher delays in road networks with important economic implications. Intelligent Transportation Systems (ITS) are able to provide vehicles with information on non-recurrent events and can therefore be used to reduce the delay in travel times. In this talk, we focus on a dynamic stochastic shortest path problem: our objective is to minimize the expected travel time of a vehicle, assuming the vehicle may adapt the chosen route while driving. To this end, we introduce a Markovian velocity model, based on an environmental background process that tracks both random and (semi-)predictable events affecting the vehicle speeds in a highway network. The resulting continuous routing model allows for correlation between velocities on the arcs and incorporates both recurrent and non-recurrent congestion. Routing in the proposed dynamic routing model can then be formulated as a semi-Markov decision process (SMDP). It can be shown that the optimal routing policy can be derived by the use of dynamic programming (DP), but the explosiveness of the size of the state space makes DP based algorithms computationally intractable. To overcome this, we present the EDSGER algorithm, a Dijkstra-like shortest path algorithm that can be used dynamically with real-time response. We quantify the performance of the algorithms and show how to operationalize this flexible data-driven model by providing numerical examples that use road network detector data provided by the National Data Warehouse for Traffic Information in The Netherlands.

Garcia Pareja Celia: Unbiased estimation of Wright-Fisher diffusions: an exact simulation approach

Garcia Pareja Celia[1], Nobile Fabio[1]

1 - Département de Mathématiques - EPFL (Switzerland)

Abstract: In this talk I will present an unbiased Monte Carlo maximum likelihood estimator for discretely observed Wright-Fisher diffusions. Our approach is based on exact simulation techniques that are of special interest for diffusion processes defined on a bounded domain, where numerical methods typically fail to remain within the required boundaries. We start by building unbiased maximum likelihood estimators for scalar diffusions and later present an extension to the multidimensional case. Consistency results of our proposed estimator are also presented and the performance of our method will be illustrated through numerical examples.

3.4 Thursday 29th, 10:30 - 12:00

3.4.1 Invited session T1-1 (Avrachenkov & Ayesta): Reinforcement learning (Auditorium)

Thursday 29th, 10:30 - 12:00 - chaired by Kostia Avrachenkov & Urtzi Ayesta

Borkar Vivek: Q-learning in unknown environments

Borkar Vivek[1]

1 - Indian Institute of Technology Bombay (India)

Abstract: The talk will describe some analytic results for Q-learning in an unknown but stationary environment, from a point of view of developing theoretically grounded approximation architectures.

Gaujal Bruno: The Sliding Regret or How to Control Bad Episodes in Reinforcement Learning

Gaujal Bruno[1]

1 - Inria Grenoble Alpes (France)

Abstract: In this presentation, we will examine the behavior of episodic reinforcement learning algorithms and question the cost of using sub-optimal policies for arbitrary long periods of time. This induces our first idea namely the introduction of a new performance measure of a RL algorithm that is more discriminating than the regret, that we call the sliding regret. Our second contribution is to present a new episodic reinforcement learning algorithm whose episodes are based on the performance of the current policy with respect to the best policy over the current confidence set. This is in contrast with all existing RL algorithms whose episode lengths are only based on the number of visits to the states. This new algorithm has $O(T^{1/2})$ regret as most RL algorithms but enjoys a unique additional property: We show that all current episodic RL algorithms have asymptotically linear sliding regret for all MDPs while the new one has a $O(\log(T)^{1/2})$ asymptotic sliding regret for generic deterministic MDPs.

Robledo Francisco: A Two-Timescale Approach for Learning Whittle Index in RMABPs: Tabular and Neural Network Implementations

Robledo Francisco[1], Ayesta Urtzi[2], Avrachenkov Konstantin[3], Borkar Vivek[4]

1 - University of the Basque Country/Euskal Herriko Unibertsitatea (Spain)
2 - Réseaux, Mobiles, Embarqués, Sans fil, Satellites (France)
3 - Inria Sophia Antipolis - Méditerranée (France)
4 - Indian Institute of Technology (India)

Abstract: In this work, we study the use of a two-time scale approach for learning the Whittle index in Restless Multi-Armed Bandit Problems (RMABPs). The Whittle index policy is a well-known heuristic that has demonstrated remarkable performance with guaranteed asymptotic optimality in RMABPs. The two-timescale approach used in this algorithm consists of a faster timescale for updating the state-action Q-values and a slower timescale for updating the Whittle indices. Our main theoretical result shows that our tabular implementation guarantees convergence to the theoretical Whittle indices, while our neural network implementation is able to extrapolate information from different state-action pairs and learn the Whittle index policy of larger state-space environments faster. Numerical results show that this approach outperforms standard Q-learning algorithms, approximate Q-learning based on neural networks, and other state-of-the-art algorithms.

Srikant R: Performance Bounds for Policy-Based Average Reward Reinforcement Learning Algorithms

Srikant R[1], Murthy Yashaswini, Moharrami Mehrdad

1 - Coordinated Science Libratory [University of Illinois] (United States)

Abstract: Many policy-based reinforcement learning (RL) algorithms can be viewed as instantiations of approximate policy iteration (PI), i.e., where policy improvement and policy evaluation are both performed approximately. In applications where the average reward objective is the meaningful performance metric, often discounted reward formulations are used with the discount factor being close to 1, which is equivalent to making the expected horizon very large. However, the corresponding theoretical bounds for error performance scale with the square of the horizon. Thus, even after dividing the total reward by the

length of the horizon, the corresponding performance bounds for average reward problems go to infinity. Therefore, an open problem has been to obtain meaningful performance bounds for approximate PI and RL algorithms for the average-reward setting. In this paper, we solve this open problem by obtaining the first non-trivial error bounds for average-reward MDPs which go to zero in the limit where when policy evaluation and policy improvement errors go to zero

3.4.2 Invited session T1-2 (Hasenbein): Strategic queueing IV (Room 101)

Thursday 29th, 10:30 - 12:00 - chaired by John Hasenbein

Snitkovsky Ran: Stochastic Approximation of Symmetric Nash Equilibria in Queueing Games

Snitkovsky Ran[1], Ravner Liron

1 - Coller School of Management, Tel Aviv University (Israel)

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 on the utility function and on the regenerative structure of the queueing 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.

Van Eekelen Wouter: Rational customers with limited information joining non-observable queues

Van Eekelen Wouter[1]

1 - Tilburg University, TiSEM, Department of Econometrics and Operations Research (Netherlands)

Abstract: Consider a firm that offers delay-prone services to a market of rational, delay-sensitive individuals. The customers in this market value the service, but dislike waiting, and will only use the service if the net benefit exceeds the waiting costs. As there is no real-time queue information available, they estimate their expected wait costs based on beliefs or information about the total arrival rate of all potential customers. The firm's service facility is modeled as an $M/M/s$ queue, with this total arrival rate as input. An important feature of this model is that the total arrival rate is a random variable of which only the mean, variance and range are known. Using semi-infinite linear programming and primal-dual techniques, we establish tight bounds for the expected waiting time in this setting, corresponding to an arrival rate that takes only two values. The proofs depend crucially on the fact that the expected waiting time, as a function of the arrival rate, has a convex derivative. We apply the novel tight bounds to the firm's pricing problem, in which both the firm and customers have partial knowledge about the total market size. Our findings provide valuable insights into the optimal maximin pricing strategy for the firm, which strikes the best balance between profit margin and market share.

Logothetis Dimitrios: Reneging behavior of strategic customers in a system with an unreliable server

Kanavetas Odysseas[1], Logothetis Dimitrios[1]

1 - Mathematical Institute, Leiden University (Netherlands)

Abstract: In this study, we analyze the reneging behavior of strategic customers in a Markovian queue with an unreliable server. In the event of a breakdown, the server is replaced with a secondary backup unit with a diminished service speed. The customers decide initially whether to join or balk and in case of joining, they continuously face the dilemma whether to stay or to renege. To focus on the effect of reneging, we also consider the case where reneging is prohibited and the customers only decide whether to join or to balk. For these models we derive the equilibrium strategies for the customers and we evaluate the performance of the system under both scenarios. Moreover, we discuss the effects of the server repairing procedure and we present numerical experiments that reveal key differences in the customer strategic behavior.

Dimitrakopoulos Yiannis: The impact of subcontracting practice on Markovian queues with strategic customers

Dimitrakopoulos Yiannis[1], Benioudakis Myron[2,3], Zissis Dimitris[3]

1 - National and Kapodistrian University of Athens (Greece)
2 - University of Thessaly [Volos] (Greece)
3 - Athens University of Economics and Business (Greece)

Abstract: We consider a service system consisting of a provider who offers a service in a make-to-order fashion to a market of potential customers, and a subcontractor who can be recruited to speed up the service process. The provider operates a Markovian queue at a standard service rate level, charging joining customers a fixed fee. At each busy cycle of the system, the provider has the option to recruit the sub-contractor and switch the implemented service rate to a higher level, offering him a salary. The sub-contractor in turn decides whether he will accept the provider's offer to increase the system's service rate level, and, also incurs a labor cost per time unit. Arriving customers, who are strategic and time-sensitive, are not aware of the number of customers in the system and the induced service policy upon arrival. They evaluate the received service with a fixed value and incur a cost per unit of time of delay. Based on their utility maximization, they make individual decisions to join the system or balk, considering the joining behavior of other customers and the strategic involvement of the subcontractor. We formulate a sequential non-cooperative game among the provider, the subcontractor, and the customers, where all entities act independently to maximize their individual utility, but they also interact since they take into account the decisions of the others. The game is solved backward to identify the customers' equilibrium joining strategies, and the subcontractor's and provider's optimal strategies, as well. Finally, we explore numerically the impact of subcontracting practice on the performance of the system in equilibrium.

3.4.3 Invited session T1-3 (Castellanos): Queueing Systems with Predictions (Room 102)

Thursday 29th, 10:30 - 12:00 - chaired by Antonio Castellanos

Hu Yue: Prediction-Driven Surge Planning with Application in the Emergency Department

Hu Yue[1], Chan Carri[2], Dong Jing[2]

1 - The University of Chicago (United States)
2 - Columbia University (United States)

Abstract: Determining emergency department (ED) nurse staffing decisions to balance the quality of service and staffing cost can be extremely challenging, especially when there is a high level of uncertainty in patient demand. Increasing data availability and continuing advancements in predictive analytics provide an opportunity to mitigate demand uncertainty by utilizing demand forecasts. In this work, we study a two-stage prediction-driven staffing framework where the prediction models are integrated with the base (made weeks in advance) and surge (made nearly real-time) staffing decisions in the ED. We quantify the benefit of having the ability to use the more expensive surge staffing and identify the importance of balancing demand uncertainty versus demand stochasticity. We also propose a near-optimal two-stage staffing policy that is straightforward to interpret and implement. Lastly, we develop a unified framework that combines parameter estimation, real-time demand forecasts, and capacity sizing in the ED. High-fidelity simulation experiments for the ED demonstrate that the proposed framework can reduce annual staffing costs by 11%-16% ($2M-3$ M) while guaranteeing timely access to care.

Singh Simrita: Feature-Based Priority Queuing

Singh Simrita[1], Gurvich Itai, Van Mieghem Jan

1 - Santa Clara University (United States)

Abstract: Traditional queuing theory assumes types are known or perfectly observed, and each type is typically put in its type-specific queue which is prioritized using some version of the celebrated c-mu rule; we call this type-based queueing. We study feature-based priority queuing where types are not perfectly observed but are inferred from observed features using a "classifier." A practically appealing approach combines an off-the-shelf classifier that predicts the type with type-based priority queueing. We propose a direct approach that optimizes the classifier to directly predict the priority queue from features.

The explicit modeling of the classifier in the queueing-system design is the novel contribution of this paper. We present an analytic model to study the optimal queue classification that minimizes queuing delay costs. We study how the optimal number of priority queues and the assignment of features to queues changes with the classifier accuracy. We present a numerical study on a real data set of medical images utilized in digital triage in radiology. We find that, relative to type classification, optimal feature-based priority queuing can improve delay costs by up to 54% using state-of-the-art image classifiers

Yom-Tov Galit: Optimising Hematology Patients LOS Using Mortality and Infection Risk Predictions

Yom-Tov Galit[1], Armony Mor

1 - Technion - Israel Institute of Technology (Israel)

Abstract: Previous research has shown that early discharge of patients may hurt their medical outcomes. However, in many cases the "optimal" length of stay (LOS) and the best location for treatment of the patient are not obvious. A case in point is hematology patients, for whom these are critical decisions. Patients with hematological malignancies are susceptible to life-threatening infections after chemotherapy. Sending these patients home early minimizes infection risk, while keeping them longer for hospital observation minimizes mortality risks if an infection occurs. We develop LOS optimization models for hematology patients that balance the risks of patient infection and mortality. We develop a Markov decision process formulation to explore the impact of the infection and mortality risk predictions on the optimal LOS from a single-patient perspective. We further consider the social optimization problem in which capacity constraints limit the ability of hospitals to keep patients for the entirety of their optimal LOS. Using fluid models, we find that the optimal solution takes the form of a two-threshold policy. This policy may block some patients and immediately route them to home care, or speed up some patients' LOS and send them to be home-cared early after an observation period at the hospital.

Physicians can use our model to determine a personalised optimal LOS for patients according to their infection and mortality risk predictions. Furthermore, they can adjust that decision according to the current hospital load. In a case study, we show that around 75% of the patient population need some observation period. If the hospital is overloaded, using a speedup only policy is optimal for 90% of the patient types; applying it to all patient types increases overall mortality risk by 0.5%.

Legros Benjamin: Why rational queueing of loss-averse users undermines the benefits of added resources

Legros Benjamin[1], Van Leeuwaarden Johan, Fransoo Jan

1 - Ecole de Management de Normandie (France)

Abstract: While policy makers regularly add capacity to reduce congestion on roads, hospitals, and other public services, empirical studies show that adding resources does not always reduce congestion and wait times. To clarify this counterintuitive phenomenon, we propose an explanatory model of rational loss-averse users who decide whether or not to access a service. Loss-averse users are more sensitive to time lost than to time gained. Using a rational queueing approach, we show that the equilibrium wait time increases monotonically with the number of resources. This insight might render investments in infrastructure such as roads and hospitals questionable, and creates the dilemma between facilitating more demand by adding resources and discouraging demand by reducing resources. We offer as explanation that adding resources reduces wait time variability, so that loss-averse users more readily join. In addition, we show that the number of resources does not only influence the volume of demand but also its nature: Adding resources attracts users who are less sensitive to expected wait and more sensitive to wait variability, which implies that large systems predominantly attract users who prefer reliability over efficiency.

3.4.4 Invited session T1-4 (Baccelli & Foss): Infinite-dimensional queueing/communication systems (Room 103)

Thursday 29th, 10:30 - 12:00 - chaired by François Baccelli & Sergey Foss

Curien Nicolas: Parking problems on random trees I

Curien Nicolas[1]

1 - Laboratoire de Mathématiques d'Orsay (France)

Abstract: Consider a rooted tree and let cars arrive sequentially, independently, and uniformly on its vertices. Each car tries to park on its arrival node, and if the spot is already occupied, it drives towards the root of the tree and parks as soon as possible. If it does not find an empty spot, it exists through the root. Recently, a phase transition has been established for this model and we shall survey the recent exciting developpments about it in connections with the theory of random graphs, coalescent processes and growth-fragmentation trees.

Curien Nicolas: Parking problems on random trees II

Curien Nicolas[1]

1 - Laboratoire de Mathématiques d'Orsay (France)

Abstract: *See above.*

Sankararaman Abishek: Interference Queueing Networks on Grids

Sankararaman Abishek[1]

1 - Amazon Web Services [Seattle] (United States)

Abstract: Motivated by problems in wireless network dynamics, we study a countably infinite collection of interacting queues, with a queue located at each point of the d-dimensional integer grid, having independent Poisson arrivals, but dependent service rates. The service discipline is of the processor sharing type, with the service rate in each queue slowed down, when the neighboring queues have a larger workload. The interactions are translation invariant in space and is neither of the Jackson Networks type, nor of the mean-field type. Coupling and percolation techniques are first used to show that this dynamics has well-defined trajectories. Coupling from the past techniques are then proposed to build its minimal stationary regime. The rate conservation principle of Palm calculus is then used to identify the stability condition of this system, where the notion of stability is appropriately defined for an infinite dimensional process. We show that the identified condition is also necessary in certain special cases and conjecture it to be true in all cases. Remarkably, the rate conservation principle also provides a closed-form ex- pression for the mean queue size. When the stability condition holds, this minimal solution is the unique translation invariant stationary regime. In ad- dition, there exists a range of small initial conditions for which the dynamics is attracted to the minimal regime. Nevertheless, there exists another range of larger though finite initial conditions for which the dynamics diverges, even though stability criterion holds. We further discuss progress made on existence of exponential moments for this class of networks using algebraic rate conservation tools.

Wang Zhe: The Poisson HailModel, Stability, and Power Law Conditions

Mountford Thomas[1], Wang Zhe[1]

1 - Département de Mathématiques - EPFL (Switzerland)

Abstract: The Poisson hail model is a stochastic system of interacting queues in $\mathbb{Z}^d$. Points in $\mathbb{Z}^d$ represent servers, which receive jobs according to i.i.d. marked Poisson processes. Each job has random spatial and temporal sizes (R, T). Denote by $W(x, t)$ the workload of the system at a space-time point (x, t). The "stability" corresponds to the tightness of the family $\{W(0, t) : t > 0\}$. We will discuss power law conditions on the sizes (R, T) that guarantee "stability." In particular, we will deal with the case of infinite speed of propagation and describe a critical curve. This is joint work with Thomas Mountford.

3.4.5 Invited session T1-5 (Coron): Stochastic models for ecology and evolution (Room 104)

Thursday 29th, 10:30 - 12:00 - chaired by Camille Coron

Costa Manon: Eco-evolutionary cycles in a multi-type predator-prey system

Costa Manon[1], Czuppon Pete[2], Forien Raphaël[3]

1 - Institut de Mathématiques de Toulouse UMR5219 (France)
2 - University of Münster (Germany)
3 - INRAE, Avignon (France)

Abstract: In this talk, we present and study a model for two populations with a predator-prey interaction, where each population is composed of two types of individuals, denoted by 0 and 1, such that predators of a given type thrive in the presence of similar prey of type, while preys of a given type are more likely to survive in the presence of predators of the different type. We consider a limit in large population and rare mutations at an intermediate scale, namely the individual mutation rate vanishes while the total mutation rate tends to infinity. We prove that depending on the parameters of the model different scenario can occur: successive prey and predator invasion leading to the coexistence of four types, or successive invasion of prey in a population of resident predator leading either to prey extinction or to the coexistence of all types, ...

Marguet Aline: Spread of parasites affecting death and division rates in a cell population

Marguet Aline[1], Smadi Charline[2,3]

1 - Inria Grenoble (France)
2 - Laboratoire des EcoSystèmes et des Sociétés en Montagne (France)
3 - Institut Fourier (France)

Abstract: We introduce a general class of branching Markov processes for the modelling of a parasite infection in a cell population. Each cell contains a quantity of parasites which evolves as a diffusion with positive jumps. The drift, diffusive function and positive jump rate of this quantity of parasites depend on its current value. The division rate of the cells also depends on the quantity of parasites they contain. At division, a cell gives birth to two daughter cells and shares its parasites between them. Cells may also die, at a rate which may depend on the quantity of parasites they contain. We study the long-time behaviour of the parasite infection and exhibit conditions, on both the parasites and the cell population dynamics, for the survival of the cell population.

Pfaffelhuber Peter: Population genetic models for transposable elements

Pfaffelhuber Peter[1], Wakolbinger Anton[2]

1 - University of Freiburg (Germany)
2 - University of Frankfurt (Germany)

Abstract: Consider a population of size N , where each individual carries $k = 0, 1, 2, ...$ transposable elements (TEs). We follow the bi-parental Moran model for reproduction. This means that every individual dies at rate $N/2$, and is replaced by a child of two randomly chosen parents. If they have k and l TEs, each TE is inherited with probability $1/2$ to the child. Denoting by $X_k^N(t)$ the frequency of individuals carrying k TEs at time t, we show a weak limit for X^N as $N \to \infty$, and discuss shortcomings of the model with respect to the underlying dynamics of TEs. Joint work with Anton Wakolbinger (University of Frankfurt).

Tourniaire Julie: A branching particle system as a model of pushed fronts

Tourniaire Julie[1]

1 - Institute of Science and Technology [Klosterneuburg, Austria] (Austria)

Abstract: We consider a system of particles performing a one-dimensional dyadic branching Brownian motion with space-dependent branching rate $r(x)$, negative drift $-\mu$ and killed upon reaching 0. More precisely, the particles branch at rate $\rho/2$ in $[0, 1]$, for some $\rho \geq 1$, and at rate $1/2$ in $(1, +\infty)$. The drift is chosen in such a way that the system is critical. This system can be seen as an analytically tractable model for fluctuating fronts, describing the internal mechanisms driving the invasion of a habitat by a cooperating population. Recent studies from Birzu, Hallatschek and Korolev on the noisy FKPP equation with Allee effect suggest the existence of three classes of fluctuating fronts: pulled, semi pushed and fully pushed fronts. In this talk, we will show that the particle system exhibits the same phase transitions as the SPDE, suggesting the existence of a universality class.

3.4.6 Invited session T1-6 (Shi): Learning algorithms for managing service systems (Room 105)

Thursday 29th, 10:30 - 12:00 - chaired by Pengyi Shi

Bastani Hamsa: Decision-Aware Learning for Global Health Supply Chains

Chung Angel[1], Rostami Vahid[2], Bastani Hamsa[1], Bastani Osbert[3]

1 - The Wharton School (United States)
2 - Macro-Eyes (United States)
3 - School of Engineering and Applied Science [University of Pennsylvania] (United States)

Abstract: The combination of machine learning (for prediction) and optimization (for decision-making) is increasingly used in practice. However, a key challenge is the need to align the loss function used to train the machine learning model with the decision loss associated with the downstream optimization problem. Traditional solutions have limited flexibility in the model architecture and/or scale poorly to large datasets. We propose a principled decision-aware learning algorithm that uses a novel Taylor expansion of the optimal decision loss to derive the machine learning loss. Importantly, our approach only requires a simple re-weighting of the training data, allowing it to flexibly and scalably be incorporated

into complex modern data science pipelines, yet producing sizable efficiency gains. We apply our framework to optimize the distribution of essential medicines in collaboration with policymakers at the Sierra Leone National Medical Supplies Agency; highly uncertain demand and limited budgets currently result in excessive unmet demand. We leverage random forests with meta-learning to learn complex cross-correlations across facilities, and apply our decision-aware learning approach to align the prediction loss with the objective of minimizing unmet demand. Out-of-sample results demonstrate that our end-to-end approach significantly reduces unmet demand across 1000+ health facilities throughout Sierra Leone.

Che Ethan: Stochastic Gradient Descent with Adaptive Data: Applications to Online Learning

Che Ethan[1], Dong Jing[1], Tong Xin[2]

1 - Columbia Business School (United States)
2 - National University of Singapore (Singapore)

Abstract: We consider a stochastic optimization problem where the data is generated sequentially via Markov chains whose dynamics are determined by the decision variables and the loss function is defined with respect to the invariant measures of the Markov chains. Examples include finding the optimal price to maximize the long-run average revenue minus waiting cost in queueing systems; finding the optimal order-up-to level to minimize the long-run average holding cost plus backlog cost in inventory systems, etc. We study the convergence of the corresponding stochastic gradient descent algorithm with adaptive data. Our results establish an $O(1/\sqrt{T})$ convergence rate when the loss function is convex and an O(1/T) convergence rate when the loss function is strongly convex under reasonably general conditions on the ergodicity of the Markov chains and smoothness of the loss function. We demonstrate how to apply our results to several online-learning problems in operations research and a policy gradient algorithm in reinforcement learning.

Chen Xinyun: Data Pooling for Personalized Intervention via Contextual Bandits

Chen Xinyun[1], Shi Pengyi[2], Pu Shanwen[3]

1 - The Chinese University of Hong Kong, Shenzhen (China)
2 - Purdue University (United States)
3 - Shanghai University of Finance and Ecnomics (China)

Abstract: There is a growing interest in personalized preventative intervention across various fields, including healthcare and community-based support programs. However, personalization poses unique challenges in learning and optimization due to limited data samples. To overcome this challenge, we model the online optimal intervention problem as a contextual bandit with different target classes. To improve the sample efficiency of personalized planning and address the issue of small sample, we develop a novel data-pooling technique that is applicable to a broad class of contextual bandit algorithms. The effectiveness of our approach is demonstrated through theoretical regret analysis, simulation experiments, and empirical studies involving real datasets obtained from a partner hospital and a public probation dataset from the Department of Justice.

Kumar Rachitesh: Online Resource Allocation under Horizon Uncertainty

Balseiro Santiago[1], Kroer Christian[1], Kumar Rachitesh[1]

1 - Columbia University (United States)

Abstract: We study stochastic online resource allocation: a decision maker needs to allocate limited resources to stochastically-generated sequentially-arriving requests in order to maximize reward. At each time step, requests are drawn independently from a distribution that is unknown to the decision maker. Online resource allocation and its special cases have been studied extensively in the past, but prior results crucially and universally rely on the strong assumption that the total number of requests (the horizon) is known to the decision maker in advance. In many applications, such as revenue management and online advertising, the number of requests can vary widely because of fluctuations in demand or user traffic intensity. In this work, we develop online algorithms that are robust to horizon uncertainty. In sharp contrast to the known-horizon setting, no algorithm can achieve even a constant asymptotic competitive ratio that is independent of the horizon uncertainty. We introduce a novel generalization of dual mirror descent which allows the decision maker to specify a schedule of time-varying target consumption rates, and prove corresponding performance guarantees. We go on to give a fast algorithm for computing a schedule of target consumption rates that leads to near-optimal performance in the unknown-horizon setting. In particular, our competitive ratio attains the optimal rate of growth (up to logarithmic factors) as the horizon uncertainty grows large. Finally, we also provide a way to incorporate machine-learned

predictions about the horizon which interpolates between the known and unknown horizon settings.

3.4.7 Invited session T1-7 (Olvera-Cravioto): Graph dynamics (Room 106)

Thursday 29th, 10:30 - 12:00 - chaired by Mariana Olvera-Cravioto

Bhamidi Shankar: Dynamic networks: one has to do the "math"

Bhamidi Shankar[1]

1 - Department of Statistics and Operations Research, University of North Carolina (United States)

Abstract: Models for networks that evolve and change over time are ubiquitous in a host of domains in applied probability, including modeling social networks, understanding the evolution of systems in proteomics, the study of the growth and spread of epidemics etc. This talk will give a brief summary of three recent findings in this area: (a) Understanding the effect and detectability of change point in the evolution of the system dynamics. (b) Reconstructing the initial "seed" that gave rise to the current network, sometimes referred to as Network Archeology. (c) The disparity in the behavior of different centrality measures for measuring popularity in settings where there are vertices of different types. The main goal will to be convey unexpected findings in each of these three areas. Based largely on joint work with Sayan Banerjee, Iain Carmichael, Nelson Antunes and Vladas Pipiras.

Contat Alice: Critical core percolation on random graphs

Contat Alice[1]

1 - Laboratoire de Mathématiques dÃŞrsay (France)

Abstract: Motivated by the desire to construct large independent sets in random graphs, Karp and Sipser modified the usual greedy construction to yield an algorithm that outputs an independent set with a large cardinal called the Karp-Sipser core. When run on the Erdös-Rényi $G(n, c/n)$ random graph, this algorithm is optimal as long as $c <$ e. We will present the proof of a physics conjecture of Bauer and Golinelli (2002) stating that at criticality, the size of the Karp-Sipser core is of order $n^{3/5}$. Along the way we shall highlight the similarities and differences with the usual greedy algorithm and the k-core algorithm. Based on a joint work with Nicolas Curien and Thomas Budzinski.

Ray Rounak: Universality of the local limit of preferential attachment models

Ray Rounak[1], Van Der Hofstad Remco, Hazra Rajat Subhra, Garavaglia Alessandro

1 - Department of Mathematics and Computer Science, Eindhoven University of Technology (Netherlands)

Abstract: We study preferential attachment models where vertices enter the network with i.i.d. random numbers of edges that we call the out-degree. We identify local limit of such models, substantially extending the work of Berger et al. [2014]. The degree distribution of this limiting random graph, which we call the random Pólya point tree, has a surprising size-biasing phenomenon. Many of the existing preferential attachment models can be viewed as special cases of our preferential attachment model with i.i.d. out-degrees. Additionally, our models incorporates negative values of the preferential attachment fitness parameter, which allows us to consider preferential attachment models with infinite-variance degrees. Our proof of local convergence consists of two main steps: a Pólya urn description of our graphs, and an explicit identification of the neighbourhoods in them. We provide a novel and explicit proof to establish a coupling between the preferential attachment model and the Pólya urn graph. Our result proves a density convergence result, for fixed ages of vertices in the local limit.

Wang Tiandong: Random Networks with Heterogeneous Reciprocity

Wang Tiandong[1]

1 - Fudan University [Shanghai] (China)

Abstract: Users of social networks display diversified behavior and online habits. For instance, a user's tendency to reply to a post can depend on both the user and the person posting. For convenience, we group users into aggregated behavioral patterns, focusing here on the tendency to reply or reciprocate to a message. The reciprocity feature in social networks reflects the information exchange among users. We study properties of a preferential attachment model with heterogeneous reciprocity levels and give the growth rate of model edge counts as well as prove convergence of empirical degree frequencies to a limiting distribution. This limiting distribution is not only multivariate regularly varying, but also has the property of hidden regular variation.

3.4.8 Invited session T1-8 (Dorsman, Righter & Gardner): Load balancing II (Room 202)

Thursday 29th, 10:30 - 12:00 - chaired by Jan-Pieter Dorsman, Rhonda Righter & Kristen Gardner

Mukherjee Debankur: Load Balancing with Data Locality: Extending Mean-Field Framework to Constrained Heterogeneous Systems

Mukherjee Debankur[1], Rutten Daan[1], Zhao Zhisheng[1], Wu Ruoyu[2]

1 - School of Industrial and Systems Engineering [Georgia Tech] (United States)
2 - Iowa State University (United States)

Abstract: Over the past few years, the load balancing problem on bipartite graphs has gained popularity as it nicely models the data locality constraints in modern data centers. Due to such constraints, servers violate the exchangeability assumption that has been pivotal for the classical mean-field analysis. We will discuss several recent progress in this direction for broad classes of bipartite graphs including expanders, spatial graphs, and inhomogeneous graphs, each of which leads to a fundamentally distinct asymptotic behavior.

Jhunjhunwala Prakirt: Join-the-Shortest Queue with Abandonment: Critically Loaded and Heavily Overloaded Regimes

Jhunjhunwala Prakirt[1]

1 - Georiga Institute of Technology (United States)

Abstract: We consider a Load Balancing system operating under the well-known Join-the-Shortest Queue (JSQ) policy, where jobs/customers are impatient, and abandon if they do not receive service after some (random) time. In this setting, we study the effect of these abandonments on the steady-state queue length behavior. In particular, we characterize the 'distribution of appropriately centered and scaled steady-state queue length' (hereafter referred to as limiting distribution) as the abandonment rate goes to zero. Depending on the arrival, service, and abandonment rates, we consider three different regimes of operation, each yielding a different steady-state distribution. The first regime is when the system is underloaded, where we show that abandonments do not affect the limiting distribution. As a result, JSQ with Abandonment (JSQ-A) asymptotically behaves like JSQ without abandonments, for which the limiting distribution is known to be exponential. As the load increases, we move to the second regime, where abandonments become significant, and there is a critical balance between the arrivals/services and the abandonments. As such, the system undergoes a phase transition and the limiting distribution changes from exponential to a truncated normal. As the load increases even further, the system enters the heavily overloaded regime and the queue lengths become very large. In this case, we show that the (centered) steady state queue length distribution converges to a normal distribution. Our results use the Transform method to establish a one-dimensional differential equation in terms of the MGF (Moment Generating Function) of the steady-state queue lengths, which can be solved to get the limiting distribution.

Zubeldia Martin: Delayed Load Balancing

Zubeldia Martin[1]

1 - University of Minnesota [Twin Cities] (United States)

Abstract: We study a system comprimsed of a large number of homogeneous single-server queues, where a central dispatched is entrusted with dispatching an incoming stream of jobs based on information obtained through message exchanges with the servers. We assume that either these messages are delayed, or that the jobs are delayed when they are dispatched (that is, it takes a positive amount of time for jobs to arrive at a queue after they are dispatched). In this setting, we show that if the delays are small enough, it is still possible to achieve asymptotically zero queueing delay (as the number of servers goes to infinity). Alternatively, if the delays are above certain threshold, the expected queueing delays must be bounded away from zero.

Doroudi Sherwin: Queue-Length-Aware Dispatching in Large-Scale Heterogeneous Systems

Abdul Jaleel Jazeem[1], Doroudi Sherwin[2], Gardner Kristen[3]

1 - NXP Semiconductors [San Jose] (United States)
2 - University of Minnesota [Twin Cities] (United States)
3 - Amherst College (United States)

Abstract: One dominant approach for reducing response times in large-scale systems is Join-the-Shortest-Queue-d (JDQ-d): whenever a job arrives, the dispatcher queries d servers at random, and then assigns the job to the queried server with the shortest queue. While JSQ-d is known to perform quite well in systems where all servers run at the same speed, this is not the case in systems that exhibit heterogeneity with respect to server speeds. Unfortunately, there is no straightforward way to extend JSQ-d (or other so-called "power-of-d" policies) to heterogeneous systems. Should a job be assigned to the queried server with the shortest queue even if much faster servers were among those queried? Should a job be assigned to the queried server where it is expected to complete the soonest even if there is an idle, albeit slower, server available among those queried? And for that matter, should we query faster servers more often than their slower counterparts? Recent work has introduced a framework for developing strong dispatching policies by pairing suitably chosen querying and assignment rules. Within this framework, prior work has focused on finding strong-performing dispatching policies that use only the idle/busy statuses of the queried servers, rather than detailed queue length information. In this paper, we overcome the challenge of evaluating the performance of-and finding strong-performing-general scalable dispatching policies that make use of both server speed and detailed queue length information, through a combination of mean-field analysis and a sequence of modified optimization problems. We find that well-designed length-aware dispatching policies can significantly outperform their idleness-based counterparts in large-scale heterogeneous systems. While the best policies of this kind are often complicated to describe, we find that in the vast majority of cases the relatively simple Shortest Expected Wait policy performs nearly as well, without the need for an especially sophisticated and time-consuming optimization procedure.

3.4.9 Invited session T1-10 (Torrisi): Topics on point processes models : theory and applications (Room 204)

Thursday 29th, 10:30 - 12:00 - chaired by Nicolas Privault

Flint Ian: An Equilibrium Spatial Model of Crime and Surveillance

Flint Ian[1]

1 - Institut Mines-Télécom [Paris] (France)

Abstract: The spatial heterogeneity of crime incidence arises from the interplay between criminals, who decide where to commit crimes, and the police, who decide where to offer protection. In order to study this interdependence, we develop a game theoretical framework in which crime locations and police units are modeled as point processes. We derive (i) the equilibrium distribution of crime and police, (ii) the average probability of catching criminals, and (iii) the crime rates in terms of the economy fundamentals (given by, among others, the spatial preferences of criminals, police expenditures, inequality, and poverty). The theoretical results explain several empirical findings and our framework uncovers new processes driving the spatial distribution of crime.

Kirchner Matthias: Critical cluster cascades

Kirchner Matthias[1]

1 - University of Teacher Education NMS Bern (Switzerland)

Abstract: We consider a sequence of Poisson cluster point processes on R^d: At step $n \in \mathbb{N}_0$ of the construction, the cluster centers have intensity $c/(n+1)$ for some $c > 0$, and each cluster consists of the particles of a branching random walk up to generation n—generated by a point process with mean 1. We show that this critical cluster cascade' converges weakly, and that either the limit point process equals the a.s. void process (extinction), or it has the same intensity c as the critical cluster cascade (persistence). We obtain persistence, if and only if the Palm version of the outgrown critical branching random walk is locally a.s. finite. This result allows us to give numerous examples for persistent critical cluster cascades. In this talk, we first illustrate the construction and the main result. Then, we discuss properties and possible applications of the limit point process in the persistent case.

Privault Nicolas: Closed-form computation of the Hawkes cumulants with exponential kernels

Privault Nicolas[1]

1 - Division of Materials Science, Nanyang Technological University (Singapore)

Abstract: We present an algorithm for the computation of multivariate cumulants of Hawkes processes, based on a recursion for probability generating functionals and Bell polynomials. This approach yields

exact analytical expressions of any orders and is easier to apply to higher-order cumulants in comparison with other methods based on differential equations, tree enumeration or martingale arguments. Implementations in Maple, Mathematica and Sage are available, and provide a better alternative to Monte Carlo estimates via stochastic simulations. The closed-form modeling of neuronal membrane potentials driven by spike trains in a multivariate Hawkes process model with excitation and inhibition is considered as an application. N. Privault, Recursive computation of the Hawkes cumulants, Statistics & Probability Letters 177 (2021) Paper No. 109161. https://arxiv.org/abs/2012.07256 N. Privault, An algorithm for the computation of joint Hawkes moments with exponential kernel, Proceedings of the 53rd ISCIE International Symposium on Stochastic Systems Theory and Its Applications (SSS'21), 2022. https://doi.org/DOI:10.5687/sss.2022.72 N. Privault and M. Thieullen, Closed-form modeling of neuronal spike train statistics by multivariate Hawkes processes, Physical Review E: Biological physics 106 (2022) 054410. https://arxiv.org/abs/2210.15549

3.5 Thursday 29th, 13:30 - 15:00

3.5.1 Invited session T2-1 (Ramanan): Measure-valued process limits of stochastic networks (Auditorium)

Thursday 29th, 13:30 - 15:00 - chaired by Kavita Ramanan

Atar Rami: Many-server queues, invariant states, and entropy

Atar Rami[1], Kang Weining[2], Kaspi Haya[1], Ramanan Kavita[3]

1 - Technion (Israel)
2 - University of Maryland (United States)
3 - Brown University (United States)

Abstract: The fluid limit of the GI/G/N+G queue in the many-server ($N \to \infty$) regime is described by an evolution equation in the space of measures, where a measure represents the age profile of the jobs being processed. This equation can be classified as a transport equation with a nonlinear boundary condition. Whereas it is easy to calculate an invariant state, it is more difficult to prove that it describes the *only* possible long time behavior of this equation. Our result states that the only invariant distribution of the equation is the one given as an atom at the above "easy to guess" invariant state. One of the techniques used to achieve that, which is based on relative entropy, will be described. The result complements a line of work by Kang and Ramanan, and in particular allows us to state that the N-limit-t-limit diagram commutes. This is joint work with Weining Kang, Haya Kaspi, and Kavita Ramanan.

Maillard Pascal: Algorithmic hardness thresholds for the continuous random energy model

Maillard Pascal[1]

1 - Institut de Mathématiques de Toulouse (France)

Abstract: I will review recent results on algorithmic hardness thresholds for optimization and sampling in Derrida's continuous random energy model (CREM). The CREM is a toy model for spin glasses and other random non-convex functions on high-dimensional spaces. Our results may shed light on theoretical barriers of optimization and sampling algorithms in general settings.

Véber Amandine: A multitype growth-fragmentation process to model the growth of a filamentous fungus

Véber Amandine[1]

1 - MAP5, CNRS, Université Paris Cité (France)

Abstract: Filamentous fungi form a large family of species playing an important role in the functioning of many ecosystems. They develop in space thanks to the growth and multiplication of filaments (also called hyphae) which allow the absorption and sharing of nutrients and other molecules. In this talk, we shall present a toy model for the development of a hyphal network, whose main aim is to identify a small number of key parameters describing the growth of the fungus in homogeneous conditions (in particular, in lab conditions) and to understand and quantify the impact of different forms of stress on this growth.

The results presented are joint work with Vincent Bansaye (Ecole Polytechnique), Lena Kuwata (Univ. Paris Cité) and Milica Tomasevic (CNRS and Ecole Polytechnique) on the maths side, and Cécilia Bobée, Florence Chapeland-Leclerc, Thibault Chassereau, Pascal David, Eric Herbert, Christophe Lalanne, Clara Ledoux, Gwenaël Ruprich-Robert, all at LIED (Univ. Paris Cité) on the biology and physics side.

Yasodharan Sarath: Large deviations of interacting jump processes on locally tree-like graphs

Ramanan Kavita[1], Yasodharan Sarath[1]

1 - Brown University (United States)

Abstract: We study countable-state semi-Markov locally interacting particle systems on sparse random graphs in which the state at a vertex evolves depending on its age and the states and ages of its neighboring vertices. For certain sequences of random graphs, including Erdos-Renyi graphs G(n,c/n), we show that as the number of vertices becomes large, the family of component empirical measures satisfies a large deviation principle on the space of marked rooted graphs. The rate function is given in terms of certain relative entropies. This talk is based on joint work with Kavita Ramanan.

3.5.2 Invited session T2-2 (Kanavetas): Strategic queueing V (Room 101)

Thursday 29th, 13:30 - 15:00 - chaired by Odysseos Kanavetas

Kanavetas Odysseas: The effect of reneging in a fluid queue with alternating service process

Kanavetas Odysseas[1], Logothetis Dimitrios[1], Manou Athanasia[2]

1 - Mathematical Institute, Leiden University (Netherlands)
2 - Department of Mathematics, National and Kapodistrian University of Athens (Greece)

Abstract: We consider the fluid version of a queueing system that alternates between slow and fast service periods. We assume that customers (that are modeled as infinitesimal units at a fluid) decide whether to join or balk upon arrival and they have the option of abandoning the system at any time, without being served. In this system, we determine equilibrium strategies and evaluate the performance under equilibrium by computing the throughput and the social welfare. We also explore the effect of system parameters on these performance measures. Moreover, we explore the value of reneging option by comparing the performance measures of this system to the corresponding measures of a system with the reneging option closed.

Economou Antonis: The impact of informing strategic customers about other customers decisions

Economou Antonis[1]

1 - National and Kapodistrian University of Athens (Greece)

Abstract: The impact of information on strategic customer behavior in queueing systems is a classical topic that has attracted the interest of many researchers in the area of Rational Queueing. The information provided to the customers usually comprises the operation parameters of the system (as in the unobservable models), but it may also describe the state of the system, exactly or approximately (as in the fully and partially observable models). In the present talk, we will present some ongoing work that focuses on the value of providing some information about the decisions of other customers. More concretely, we will present a very simple stylized model where the customers are informed about other customers decisions. We will analyze the strategic behavior in this case and determine the equilbrium strategies of the customers. We will then compare the associated equlibrium throughput and social welfare of this model with the corresponding measures of the unobservable and observable models.

Baron Opher: Data Driven Causal Modeling of Queueing Systems

Baron Opher[1], Krass Dmitry[1], Senderovich Arik[2], Van Der Laan Mark[3], Xu Zhenghang[1]

1 - Rotman School of Management, University of Toronto (Canada)
2 - York University (Canada)
3 - University of California, Berkeley (United States)

Abstract: Simulation is a powerful tool for the prescriptive analysis of queueing models. With ample data and expert knowledge of the underlying system structure, a good model can be constructed and used to predict impact of various interventions. However, such manual construction is both time- and skill-demanding. Moreover it is somewhat subjective - if the expert failed to note an important feature of the system (e.g. different customer types receiving different service priorities), the model will not be accurate. As an alternative, we propose a data-driven representation of system building blocks, justified by the G-computation results from causal inference literature. We describe the queueing data generation process with structural equations and apply machine learning models to fit the equations to the data.

Through numerical experiments, we show that this approach can replace the explicit queueing dynamics of the simulator. Our model is shown to capture the intervention effect in M/G/c queues with independent hyper-exponential service time and first-come-first-serve queueing discipline. We discuss the limitation of this approach in the context of strategic customers.

Kerner Yoav: The Value of Service-Age Information in an Observable M/G/1 Queue

Kerner Yoav[1], Roet-Green Ricky[2], Zang Lin[2]

1 - Ben-Gurion University of the Negev (Israel)
2 - Simon School of business, University of Rochester (United States)

Abstract: We study the strategic join/balk behavior of customers in a fully observable M/G/1 queue. Upon arrival, customers are informed about the queue length and about the service age. We assumed a homogeneous reward from being served and waiting cost which is assumed to be linear. We obtain the Nash equilibrium across customers. We focus on service time distributions with monotone mean residual life and show that in such distributions the equilibrium strategy is threshold based, for any observed queue length. We show that when the potential arrival rate is relatively high (low, respectively), then providing the service age information yields better (worse, respectively) system's performance.

3.5.3 Invited session T2-3 (Buke): Many-Server Queues (Room 102)

Thursday 29th, 13:30 - 15:00 - chaired by Burak Buke

Bekker Rene: Multitasking Service Systems with Many Servers

Bekker Rene[1], Storm Jaap, Berkelmans Wouter

1 - Vrije Universiteit Amsterdam [Amsterdam] (Netherlands)

Abstract: We consider a many-server queue in which each server can serve multiple customers in parallel. Such multitasking phenomena occur in various applications areas (e.g. in hospitals and contact centers), although the impact of the number of customers that are simultaneously served on system efficiency may vary. During the presentation, we aim to discuss diffusion limits of the queueing process under the QED scaling and for different policies of assigning customers to servers depending on the number of customers they serve. For a broad class of routing policies, including routing to the least busy server, the same one-dimensional diffusion process is obtained in the heavy-traffic limit. In case of assignment to the most busy server, there is no state-space collapse and the diffusion limit involves a custom regulator mapping. Moreover, the aim is to discuss some health applications where models with multi-tasking servers may be useful.

Mukhopadhyay Arpan: Performance of Large Multi-Server Systems with Parallelisable Jobs

Mukhopadhyay Arpan[1]

1 - Department of Computer Science [Warwick] (United Kingdom)

Abstract: Modern data centres often handle computing jobs that are highly parallelisable. Such jobs can be processed at multiple servers in parallel. Although parallel processing reduces the latency of jobs, it comes at the cost of continuously maintaining synchronisation among parallel threads for each job. Therefore, a natural question to ask is whether it is possible to achieve most of the benefits of parallelism with only a small number of parallel threads per job. Motivated by this key question, in this talk, we shall consider a simple model in which each job can be served simultaneously by d distinct processor sharing (PS) servers. The job is considered complete when the total amount of work done on it by all the servers equals its size. Under standard memoryless assumptions, we shall show that when the number n of servers is large an exponential reduction in the average latency can be obtained using only $d \geq 2$ parallel servers per job compared to the case without parallelism (i.e., where $d = 1$). This is reminiscent of the classical result of [Mitzenmacher, 1996] and [Dobrushin and Vvedenskaya, 1996] on the Power-of-d choices load balancing algorithm and suggests that significant gains can be obtained with only small degree of parallelism. The system is however significantly more challenging to analyse using as standard mean-field techniques do not apply. We shall discuss how important properties such as stability, monotonicity, and uniform boundedness of the system can be established. We shall then discuss how these properties can be combined to obtain asymptotic results in the large system limit.

This talk will be based on a joint work with A. Ganesh.

Zhong Yueyang: Finite-Buffer Queues with Many Strategic Servers

Zhong Yueyang[1], Gopalakrishnan Raga[2], Ward Amy[3]

1 - The University of Chicago Booth School of Business (United States)
2 - Smith School of Business at Queen's University (Canada)
3 - The University of Chicago Booth School of Business (United States)

Abstract: Traditional queueing theory assumes that servers work at constant speeds. That is reasonable in computer science and manufacturing contexts. However, servers in service systems are people, and in contrast to machines, the incentives created by design decisions influence their work speeds. We study how server work speed is affected by managerial decisions concerning (i) how many servers to staff and how much to pay them and (ii) whether and when to turn away customers. We develop a game-theoretic many-server Markovian queueing model with a finite or infinite buffer in which the work speeds emerge as the solution to a non-cooperative game. In an asymptotic regime in which demand becomes large and the utility function becomes concave, we establish existence, uniqueness, and monotonicity properties of underloaded, critically loaded, and overloaded equilibria for various regions in the design space. The asymptotic regime we develop lays the foundation to study the interaction between customers that choose whether or not to join the queue based on wait time, and servers that choose their work speeds. We conclude by discussing some managerial consequences of ignoring strategic server behavior.

Buke Burak: Many-Server Systems with Heterogeneous and Strategic Servers

Buke Burak[1]

1 - The University of Edinburgh, School of Mathematics (James Clerk Maxwell Building The King's Buildings Peter Guthrie Tait Road Edinburgh EH9 3FD United Kingdom)

Abstract: In this talk, we focus on many server systems where servers are heterogeneous on an individual level. We present a framework to prove diffusion and fluid limits based on measure-valued processes that address individual differences in servers. We also show how this framework can be generalized to include parallel server systems with multiple pools of agents and customers. Finally, we show how this framework can be applied to characterize asymmetric equilibrium when servers act strategically to experience more idle times.

3.5.4 Invited session T2-4 (Nazarathy): Applied Probability for COVID-19 (Room 103)

Thursday 29th, 13:30 - 15:00 - chaired by Peter Taylor

Asanjarani Azam: Emulation of epidemics via Bluetooth-based virtual safe virus spread

Asanjarani Azam[1], Shausan Aminath[2], Chew Keng[2], Graham Thomas[2], Henderson Shane[3], Jansen Hermanus[4], Short Kirsty[2], Taylor Peter[5], Vuorinen Aapeli[6], Yadav Yuvraj[7], Ziedins Ilze[1], Nazarathy Yoni[2]

1 - The University of Auckland (New Zealand)
2 - The University of Queensland (Australia)
3 - Cornell University [New York] (United States)
4 - Delft University of Technology (Netherlands)
5 - University of Melbourne (Australia)
6 - Columbia University (United States)
7 - Indian Institute of Technology Delhi (India)

Abstract: We describe the Safe Blues Android app experiment at the University of Auckland City Campus. This experiment is designed to evaluate how physical interactions over time and between individuals affect the spread of epidemics. Our experiment involves the voluntary use of the Safe Blues Android app by participants. The app spreads multiple virtual safe virus strands via Bluetooth depending on the social and physical proximity of the subjects. The evolution of the virtual epidemics is recorded as they spread through the population. The data is presented as a real-time (and historical) dashboard. A simulation model is applied to calibrate strand parameters. Participants' locations are not recorded, but participants are rewarded based on the duration of participation within a geofenced area, and aggregate participation numbers serve as part of the data.

Britton Tom: Disentangling the most effective behavioural changes to reduce transmission during the Covid-19 pandemic

Britton Tom[1]

1 - Stockholm University (Sweden)

Abstract: Using various data sources, including Google trend (behavioural changes) temperature and circulating virus strains, and reported number of cases and hospitalizations, we estimate how much transmission that can be attributed to various behavioural patterns (work, public transportation, super-markets and shops/restaurants). The underlying infectious disease model also takes into account temperature and virus strain effects. The analysis is performed on data from Sweden and Norway. (Joint work with F Gunther, H Kjelgaard Brustad and A Frigessi).

Juneja Sandeep: Agent based simulators for epidemic modelling: Simulating larger models using smaller ones

Juneja Sandeep[1], Mittal Daksh, Agrawal Shubhada

1 - Tata Institute of Fundamental Research (India)

Abstract: Agent-based simulators (ABS) are a popular epidemiological modelling tool to study the impact of various non-pharmaceutical interventions in managing an epidemic in a city (or a region). They provide the flexibility to accurately model a heterogeneous population with time and location varying, person-specific interactions as well as detailed governmental mobility restrictions. Typically, for accuracy, each person is modelled separately. This however may make computational time prohibitive when the city population and the simulated time is large. In this paper, we dig deeper into the underlying probabilistic structure of a generic, locally detailed ABS for epidemiology to arrive at modifications that allow smaller models (models with less number of agents) to give accurate statistics for larger ones, thus substantially speeding up the simulation. We observe that simply considering a smaller aggregate model and scaling up the output leads to inaccuracies. We exploit the observation that in the initial disease spread phase, the starting infections create a family tree of infected individuals more-or-less independent of the other trees and are modelled well as a multi-type super-critical branching process. Further, although this branching process grows exponentially, the relative proportions amongst the population types stabilise quickly. Once enough people have been infected, the future evolution of the epidemic is closely approximated by its mean field limit with a random starting state. We build upon these insights to develop a shifted, scaled and restart-based algorithm that accurately evaluates the ABS's performance using a much smaller model while carefully reducing the bias that may otherwise arise. We apply our algorithm for Covid-19 epidemic in a city and theoretically support the proposed algorithm through an asymptotic analysis where the population size increases to infinity. We develop nuanced coupling based arguments to show that the epidemic process is close to the branching process early on in the simulation.

Litvak Nelly: The role of inter-regional mobility in forecasting COVID-19 transimission in the Netherlands

Litvak Nelly[1,2], Schoot Uiterkamp Martijn H. H.[3], Van Der Hofstad Remco[2], Gösgens Martijn[2], Heesterbeek Hans[4]

1 - University of Twente [Netherlands] (Netherlands)
2 - Eindhoven University of Technology [Eindhoven] (Netherlands)
3 - Tilburg University [Tilburg] (Netherlands)
4 - Utrecht University [Utrecht] (Netherlands)

Abstract: In The Netherlands, people comute from one municipality to another on daily basis, to the office, to schools, and just for one meeting or a party. At the very beginning of the COVID-19 pandemic, we teamed up with a company who had the inter-regional mobility data, with the goal to help local governments in predicting and controling pandemic on a regional level. This grew into an inter-disciplinary project, with consortium of mathematicians, epedimeologists, public health experts, policy scientists, and data science companies. In this talk I will present some of the results and lessons learned from this project, First, I will discuss the trade-off between mobility and epidemic transmission. I will present our heuristic to approximate optimal choices for mobility restrictions, using the mobility data and clustering algorithms from Network Science. Next, I will discuss under which epidemiological circumstances incorporating mobility into transmission models yields improvements in the accuracy of forecasting. If time permits, I will also touch upon other aspects of the project: interdisciplinarity, applications, and the organizational challenges around the use of the data.

3.5.5 Invited session T2-5 (Henry): Stochastic geometry (Room 104)

Thursday 29th, 13:30 - 15:00 - chaired by Benoît Henry

Blaszczyszyn Bartlomiej: Continuum Line-of-Sight Percolation on Poisson-Voronoi Tessellations

Blaszczyszyn Bartlomiej[1]

1 - Inria de Paris (France)

Abstract: In this work, we study a new model for continuum line-of-sight percolation in a random environment driven by the Poisson-Voronoi tessellation in the d-dimensional Euclidean space. The edges (one-dimensional facets, or simply 1-facets) of this tessellation are the support of a Cox point process, while the vertices (zero-dimensional facets or simply 0-facets) are the support of a Bernoulli point process. Taking the superposition Z of these two processes, two points of Z are linked by an edge if and only if they are sufficiently close and located on the same edge (1-facet) of the supporting tessellation. We study the percolation of the random graph arising from this construction and prove that a 0-1 law, a subcritical phase as well as a supercritical phase exist under general assumptions. Our proofs are based on a coarse-graining argument with some notion of stabilization and asymptotic essential connectedness to investigate continuum percolation for Cox point processes. We also give numerical estimates of the critical parameters of the model in the planar case, where our model is intended to represent telecommunications networks in a random environment with obstructive conditions for signal propagation. This is a joint work with Quentin Le Gall, Elie Cali, and Taoufik En-Najjary, published in Advances in Applied Probability in 2021.

Coupier David: Percolation on Delaunay triangulation to model long-range connection in 5G networks

Coupier David[1]

1 - Institut Mines Télécom Nord Europe (France)

Abstract: The 5G's Device-to-Device (D2D) technology consists in the possibility of short-range direct communications (according to microscopic rules) between two devices or users without the need for the signal to be routed through additional network infrastructure. Hence, a good connectivity in a D2D network, i.e. a long-range connection (macroscopic property), can be naturally interpreted as a percolation problem. Our goal is to model D2D networks using stochastic geometry tools, in particular Delaunay triangulation for the urban network and point processes for users, and then to study their percolation properties. Joint work with David Corlin Marchand and Benoît Henry (both from Institut Mines Télécom).

Heydenreich Markus: Phase transition of geometric inhomogeneous percolation models

Heydenreich Markus[1]

1 - Augsburg University (Germany)

Abstract: We investigate a geometric random graph model whose vertices are given as a marked Poisson process on R^d. Edges are inserted between any pair of points independently with probability depending on the Euclidean distance of the two endpoints and their marks. Upon variation of the Poisson density, a percolation phase transition occurs under mild conditions: for low density there are finite connected components only, while for large density there is an infinite component almost surely. Our interest is on the transition between the low- and high-density phase, where the system is critical. We establish that if dimension is high enough and the mark distribution satisfies certain conditions, then an infrared bound for the critical connection function is valid. As a consequence, we obtain that various critical exponents exist and take on their mean-field values. We finally present an asymptotic expansion of the critical density as a function of the dimension.

Jahnel Benedikt: Subcritical percolation phases for generalized weight-dependent random connection models

Jahnel Benedikt[1]

1 - Weierstrass Institute Berlin & TU Braunschweig (Germany)

Abstract: We present results on the existence of a subcritical percolation phase for a wide range of continuum percolation models where each vertex is embedded into Euclidean space and carries an independent weight. In contrast to many established models, the presence of an edge is not only allowed to depend on the distance and weights of its end vertices but can also depend on the surrounding vertex set. Our result can be applied in particular to models combining heavy-tailed degree distributions and long-range effects, which are typically well connected. Moreover, we establish bounds on the tail-distribution of the number of points and the diameter of the subcritical component of a typical point.

3.5.6 Invited session T2-6 (Sanders): Stochastic learning in structured systems (Room 105)

Thursday 29th, 13:30 - 15:00 - chaired by Jaron Sanders

Goldsztajn Diego: Utility maximizing load balancing policies

Goldsztajn Diego[1], Borst Sem[1], Van Leeuwaarden Johan[2]

1 - Eindhoven University of Technology [Eindhoven] (Netherlands)
2 - Tilburg University [Tilburg] (Netherlands)

Abstract: We consider a service system where incoming tasks are instantaneously assigned to one out of many heterogeneous server pools. All the tasks sharing a server pool are executed in parallel and the execution times do not depend on the class of the server pool or the number of tasks currently contending for service. However, associated with each server pool is a utility function that does depend on the class of the server pool and the number of tasks currently sharing it. These features are characteristic of streaming and online gaming services, where the duration of tasks is mainly determined by the application but congestion can have a strong impact on the quality-of-service (e.g., video resolution and smoothness). We provide an upper bound for the mean aggregate utility in steady state and we introduce a multi-threshold policy that achieves the upper bound in a large-scale regime. Since the optimal threshold values depend on the unknown offered load (i.e., the demand) of the system, our multi-threshold policy includes a learning scheme that adjusts the thresholds over time in order to find the optimal values.

Lee Junghyun: Nearly Optimal Latent State Decoding in Block MDPs

Jedra Yassir[1], Lee Junghyun[2], Proutiere Alexandre[1], Yun Se-Young[2]

1 - KTH School of Electrical Engineering and Computer Science (Sweden)
2 - Korea Advanced Institute of Science and Technology (South Korea)

Abstract: We consider the problem of model estimation in episodic Block MDPs. In these MDPs, the decision maker has access to rich observations or contexts generated from a small number of latent states. We are interested in estimating the latent state decoding function (the mapping from the observations to latent states) based on data generated under a fixed behavior policy. We derive an information-theoretical lower bound on the error rate for estimating this function and present an algorithm approaching this fundamental limit. In turn, our algorithm also provides estimates of all the components of the MDP. We apply our results to the problem of learning near-optimal policies in the reward-free setting. Based on our efficient model estimation algorithm, we show that we can infer a policy converging (as the number of collected samples grows large) to the optimal policy at the best possible asymptotic rate. Our analysis provides necessary and sufficient conditions under which exploiting the block structure yields improvements in the sample complexity for identifying near-optimal policies. When these conditions are met, the sample complexity in the minimax reward-free setting is improved by a multiplicative factor n, where n is the number of contexts.

Senen-Cerda Albert: Asymptotic convergence rate of Dropout in neural network models

Senen-Cerda Albert[1], Sanders Jaron[1]

1 - Eindhoven University of Technology [Eindhoven] (Netherlands)

Abstract: Dropout (Hinton et al., 2012) is a well-known method to prevent overfitting in neural networks that temporarily 'drops out' nodes of the network at random during each step of the training with stochastic gradient descent. In this talk we present general convergence guarantees for training neural networks with dropout or one of its variants, and investigate its convergence rate using stochastic approximation techniques on simplified models for deep and shallow neural networks. In particular, we examine theoretically and with simulations the dependence of the convergence rate on the width and depth of the neural network, and more importantly, on the dropout probability. For the considered models, we obtain explicit bounds for the convergence rate that depend on the dropout probability and discuss their consequences. This talk is based on results from two articles and is joint work with Jaron Sanders.

Van Werde Alexander: Detection and evaluation of clusters within sequential data

Van Werde Alexander[1]

1 - Eindhoven University of Technology (Netherlands)

Abstract: Motivated by theoretical advancements in dimensionality reduction techniques we use a recent model, called Block Markov Chains, to conduct a practical study of clustering in real-world sequential data. Clustering algorithms for Block Markov Chains possess theoretical optimality guarantees and can be deployed in sparse data regimes. Despite these favorable theoretical properties, a thorough evaluation of these algorithms in realistic settings has been lacking.

We address this issue and investigate the suitability of these clustering algorithms in exploratory data analysis of real-world sequential data. In particular, our sequential data is derived from human DNA, written text, animal movement data and financial markets. In order to evaluate the determined clusters, and the associated Block Markov Chain model, we further develop a set of evaluation tools. These tools include benchmarking, spectral noise analysis and statistical model selection tools. An efficient implementation of the clustering algorithm and the new evaluation tools is made available together with this paper.

Practical challenges associated to real-world data are encountered and discussed. It is ultimately found that the Block Markov Chain model assumption, together with the tools developed here, can indeed produce meaningful insights in exploratory data analyses despite the complexity and sparsity of real-world data. This talk is based on arXiv:2210.01679.

3.5.7 Invited session T2-7 (Avrachenkov): Network inference I (Room 106)

Thursday 29th, 13:30 - 15:00 - chaired by Kostia Avrachenkov

Gösgens Martijn: The Hyperspherical Geometry of Community Detection

Gösgens Martijn[1], Van Der Hofstad Remco[1], Litvak Nelly[1,2]

1 - Eindhoven University of Technology [Eindhoven] (Netherlands)
2 - University of Twente (Netherlands)

Abstract: We introduce a metric space of clusterings, where clusterings are described by a binary vector indexed by the vertex-pairs. We extend this geometry to a hypersphere and prove that maximizing modularity is equivalent to minimizing the angular distance to some modularity vector over the set of clustering vectors. In that sense, modularity-based community detection methods can be viewed as a subclass of a more general class of projection methods, which we define as the community detection methods that adhere to the following two-step procedure: first, mapping the network to a point on the hypersphere; second, projecting this point to the set of clustering vectors. We show that this class of projection methods contains many interesting community detection methods. Many of these new methods cannot be described in terms of null models and resolution parameters, as is customary for modularity-based methods. Instead, they are more naturally characterized by the latitude and the meridian of the hypersphere point that the network is mapped to. We discuss how this latitude and meridian affect the clustering that results from applying the projection method.

Leskelä Lasse: Community recovery from non-binary and higher-order network interactions

Leskelä Lasse[1]

1 - Aalto University (Finland)

Abstract: Community recovery is the task of learning a latent community structure from interactions in a population of N nodes. Efficient algorithms for sparse binary pairwise interaction data are well known, and so are their consistency properties with respect to data sampled from the stochastic block model (SBM), the canonical model for random graphs with a community structure. Instead of a binary variable indicating whether or not an interaction occurs, we often also observe a category, value, or shape of an interaction. This motivates the definition of a generalised SBM in which interactions can be of arbitrary type, including categorical, numeric, and vector-valued, and not excluding even more general objects such as Markov chains or Poisson processes. For this model, I will discuss information-theoretic bounds which characterise the existence of consistent estimators in terms of data sparsity, statistical similarity between intra- and inter-block interaction distributions, and the shape and size of the interaction space. Temporal networks with time-correlated interaction patterns of length T provide an important model instance, for which consistency can be analysed with respect to either N or T, or both, approaching infinity. Time permitting, I will also highlight recent findings and open problems related to data sets involving higher-order interactions which can be modelled using hypergraph stochastic block models.

Matias Catherine: Model-based clustering in simple hypergraphs through a stochastic blockmodel

Matias Catherine[1]

1 - Laboratoire de Probabilités, Statistique et Modélisation (France)

Abstract: We present a new hypergraph stochastic blockmodel and an associated inference procedure for model-based clustering of the nodes in simple hypergraphs. Simple hypergraphs, where a node may not appear several times in a same hyperedge, have been overlooked in the literature, though

they appropriately model some high-order interactions (such as co-authorship). The model assumes latent groups for the nodes and conditional independence of the hyperedges given the latent groups. We establish the first proof of generic identifiability of the parameters in such a model. We develop a variational approximation Expectation-Maximization algorithm for parameter inference and node clustering, and derive an integrated classification likelihood criterion for model selection. We illustrate the performance of our algorithm on synthetic data and analyse a real dataset of co-authorship. Our method called HyperSBM is implemented in C++ for efficiency and available as an R package at https://github.com/LB1304/HyperSBM.

Xu Jiaming: Random Graph Matching in Geometric Models

Xu Jiaming[1]

1 - Duke university [Durham] (United States)

Abstract: This talk focuses on the problem of matching two complete graphs with edge weights correlated through latent geometries, extending a recent line of research on random graph matching with independent edge weights to geometric models. We derive an approximate maximum likelihood estimator and characterize the sufficient conditions under which it achieves perfect recovery or almost perfect recovery. These conditions are further shown to be information-theoretically optimal even when the latent coordinates are directly observed. This work bridges several streams of literature including planted matching, feature matching, Procrustes matching, and graph matching. Based on joint work with Haoyu Wang, Yihong Wu, and Israel Yolou from Yale. Preprint available at https://arxiv.org/pdf/2202.10662.pdf

3.5.8 Invited session T2-8 (Dorsman, Righter & Gardner): Load balancing III (Room 202)

Thursday 29th, 13:30 - 15:00 - chaired by Jan-Pieter Dorsman, Rhonda Righter & Kristen Gardner

Wang Weina: The M/M/k with Deterministic Setup Times

Williams Jalani[1], Harchol-Balter Mor[1], Wang Weina[1]

1 - Computer Science Department - Carnegie Mellon University (United States)

Abstract: Capacity management, whether it involves servers in a data center, or human staff in a call center, or doctors in a hospital, is largely about balancing a resource-delay tradeoff. On the one hand, one would like to turn off servers when not in use (or send home staff that are idle) to save on resources. On the other hand, one wants to avoid the considerable setup time required to turn an "off" server back "on." However, surprisingly little is known about the effect of setup times on delay. While there has been some work on studying the M/M/k with Exponentially-distributed setup times, these works provide only iterative methods for computing mean delay, giving little insight as to how delay is affected by k, by load, and by the setup time. Furthermore, setup time in practice is much better modeled by a deterministic random variable. In this talk, we present the first analysis of the M/M/k with Deterministic setup times. We provide bounds on the effect of setup on delay, where our bounds are highly accurate for the common case where the setup time is much higher than the job service time. Our bounds are relatively simple algebraic formulas which provide insights into how delay scales with the input parameters. Our proofs use a combination of renewal theory, martingale arguments, and stopping-time constructions, providing strong intuition on the transient behavior of a system that turns servers on and off.

Walton Neil: Stability and Optimization of Speculative Queueing Networks

Walton Neil[1]

1 - Neil Walton, Durham (United Kingdom)

Abstract: We provide a queueing-theoretic framework for job replication schemes based on the principle "replicate a job as soon as the system detects it as a straggler". This is called job speculation. Recent works have analyzed replication on arrival, which we refer to as replication. Replication is motivated by its implementation in Google's BigTable. However, systems such as Apache Spark and Hadoop MapReduce implement speculative job execution. The performance and optimization of speculative job execution is not well understood. To this end, we propose a queueing network model for load balancing where each server can speculate on the execution time of a job. Specifically, each job is initially assigned to a single server by a frontend dispatcher. Then, when its execution begins, the server sets a timeout. If the job completes before the timeout, it leaves the network, otherwise the job is terminated and relaunched or

resumed at another server where it will complete. We provide a necessary and sufficient condition for the stability of speculative queueing networks with heterogeneous servers, general job sizes and scheduling disciplines. We find that speculation can increase the stability region of the network when compared with standard load balancing models and replication schemes. We provide general conditions under which timeouts increase the size of the stability region and derive a formula for the optimal speculation time, i.e., the timeout that minimizes the load induced through speculation. We compare speculation with redundant-d and redundant-to-idle-queue-d rules under an S&X model. For light loaded systems, redundancy schemes provide better response times. However, for moderate to heavy loadings, redundancy schemes can lose capacity and have markedly worse response times when compared with a speculative scheme. (Joint work with Jonathan Anselmi)

Perry Ohad: Stability of parallel server systems

Moyal Pascal[1], Perry Ohad[2]

1 - Institut Élie Cartan de Lorraine (France)
2 - SMU (United States)

Abstract: The fundamental problem in the study of parallel-server systems is that of finding and analyzing routing policies of arriving jobs to the servers that efficiently balance the load on the servers. The most natural, and well-studied policies are 'Join the shortest workload' (JSW), 'Join the shortest queue' (JSQ), and the 'Power-of-d' routing (PW(d)).

In this paper we study the stability problem of parallel-server systems, assuming that routing errors may occur, so that arrivals may be routed to the "wrong" queue (not the smallest among the relevant queues) with a positive probability. We treat this routing mechanism as a probabilistic routing policy, named a p-allocation policy, where p is an s-dimensional vector whose components are the routing probabilities. We study the (in)stability problem of the system under this routing mechanism, and under its "non-idling" version. We first show that the stability region is precisely the usual "rho less than one", whenever p is dominated by the uniform probability for a suitable ordering of measures. Otherwise, we show that the stability region is in general strictly smaller than the set "rho less than one". Our analyses build on Lyapunov techniques on an ordered space of s-dimensional real-valued vectors, and employing a generalized form of the Schur-convex order.

Hyytiä Esa: Are highly scalable sequential dispatching policies asymptotically optimal?

Hyytiä Esa[1], Righter Rhonda

1 - University of Iceland (Department of Computer Science, University of Iceland, Dunhagi 5, 107 Reykjavik, Iceland; Iceland)

Abstract: Job dispatching appears in many communication and computer systems. In this talk, we present a classification of dispatching policies according to their scalability, and introduce the class of sequential dispatching policies, which are a special case of index policies. Sequential policies are highly scalable. We focus on a sequential policy where each server has a single control parameter that can be determined in adaptive manner. In principle, the optimal (size-aware) policies can be determined numerically via value iteration. This approach, however, becomes computationally demanding as the size of the system increases. In the numerical examples, we compare sequential policies to the optimal policy and other well-known heuristic dispatching policies. Our numerical examples suggest that efficient and scalable load balancing can be realized by simple primitive dispatching elements. This approach seems also suitable for machine learning.

3.5.9 Invited session T2-10 (Decreusefond): Stein method I (Room 204)

Thursday 29th, 13:30 - 15:00 - chaired by Laurent Decreusefond

Coutin Laure: Diffusive limits of Lipschitz functionals of Poisson measures

Coutin Laure[1]

1 - Mathematical Institute of Toulouse (France)

Abstract: Continuous Time Markov Chains, Hawkes processes and many other interesting processes can be described as solution of stochastic differential equations driven by Poisson measures. First, using the Stein's method, we give the convergence rate of a sequence of renormalized Poisson measures towards the Brownian motion in several distances, constructed on the model of the Kantorovitch-Rubinstein (or Wasserstein-1) distance. We show that many operations (like time change, convolution) on continuous functions are Lipschitz continuous to extend these quantified convergences to diffusive limits of Markov processes and long-time behavior of Hawkes processes.

Nourdin Ivan: Total variation bound for Hadwiger's functional using Stein's method

Nourdin Ivan[1]

1 - University of Luxembourg [Luxembourg] (Luxembourg)

Abstract: Let K be a convex body in $\mathbb{R}^d$. Let X be a d-dimensional random vector distributed according to the Hadwiger-Wills density associated with K. Finally, let the information content H be defined as $H = dist^2(X, K)$. In this talk, we will study the fluctuations of H around its expectation as the dimension d go to infinity. Stein's method plays a crucial role in our analysis. This is joint work with Valentin Garino.

Reveillac Anthony: The Malliavin-Stein method for Hawkes functionals

Reveillac Anthony[1,2]

1 - Institut de Mathématiques de Toulouse UMR5219 (France)
2 - Institut National des Sciences Appliquées - Toulouse (France)

Abstract: In this talk we combine the Malliavin calculus and Stein's method to provide general bounds on the Wasserstein distance between functionals of a compound Hawkes process and a given Gaussian density. To achieve this, we rely on the Poisson embedding representation of an Hawkes process to provide a Malliavin calculus for the Hawkes processes, and more generally for compound Hawkes processes. As an application, we close a gap in the literature by providing the first Berry-Esséen bounds associated to Central Limit Theorems for the compound Hawkes process. This talk is based on joint works with Caroline Hillairet, Lorick Huang, Mahmoud Khabou and Nicolas Privault.

Vuong Christophe: Malliavin-Stein method for conditionally independent random variables

Vuong Christophe[1], Decreusefond Laurent[1]

1 - Télécom Paris (France)

Abstract: On any denumerable product of probability spaces, we extend the discrete Malliavin structure for conditionally independent random variables. As a consequence, we obtain the chaos decomposition for functionals of conditionally independent random variables. The Malliavin-Stein method yields Berry-Esseen bounds for U-Statistics of such random variables. It leads to quantitative statements of conditional limit theorems: Lyapunov's central limit theorem, partial fourth moment theorem for multilinear forms. As an application, we obtain rates of normal approximation for subhypergraph counts in random exchangeable hypergraphs.

3.6 Thursday 29th, 15:30 - 17:00

3.6.1 Invited session T3-1 (Blanchet): Advances in Theory and Algorithms for Decision Making Under Uncertainty (Auditorium)

Thursday 29th, 15:30 - 17:00 - chaired by Jose Blanchet

Dong Jing: A primal-dual approach to constrained Markov decision processes with applications to queue scheduling and inventory management

Dong Jing[1], Chen Yi[2], Wang Zhaoran[3]

1 - Columbia University (United States)
2 - Hong Kong University of Science and Technology (Hong Kong SAR China)
3 - Northwestern University (United States)

Abstract: In many sequential decision-making problems, a single objective might not suffice to describe the real considerations faced by decision-makers. The constrained Markov decision process (CMDP) has become an important modeling tool in this context. In this talk, I will introduce a data-driven primal-dual algorithm to solve CMDPs. Our approach alternatively applies regularized policy iteration to improve the policy and subgradient ascent to maintain the constraints. Under mild regularity conditions, we show that the algorithm converges at rate $O(1/\sqrt{T})$. Our algorithm can be easily combined with advanced value function approximation techniques to deal with large-scale problems, with the added benefit of straightforward convergence analysis. We apply the algorithm to solve multi-class queue scheduling and multi-product inventory control problems, and show that it generates policies that achieve superior performance compared to state-of-art policies.

Qu Yanlin: Contractive Drift Conditions for Rates of Convergence of Markov Chains in Wasserstein Distance

Qu Yanlin[1], Blanchet Jose[1], Glynn Peter[1]

1 - Stanford University (United States)

Abstract: We introduce a condition called "contractive drift" to bound the convergence of Markov chains in Wasserstein distance. Under this condition, we derive computable convergence bounds with different rates, from polynomial to exponential. These bounds are computable in the sense that they only contain fully explicit constants and one-step transition expectations. To enhance practical application, we develop various techniques to verify the contractive drift condition so that computable convergence bounds can be conveniently obtained for queueing models as well as optimization algorithms. Moreover, our bounds have the ability to capture optimal rates of convergence in terms of model parameters such as the traffic intensity in queueing models and the step size in optimization algorithms.

Ghosh Soumyadip: Distributionally Robust Optimization of Boosting for Efficient Ensemble Learning

Ghosh Soumyadip[1], Squillante Mark[1]

1 - IBM [Yorktown] (United States)

Abstract: Boosting is a learning approach that combines weak learners into an ensemble, where each learner is picked adaptively by training on data that are re-weighed to emphasize those data points on which the current ensemble performs poorly. The ensemble's output is obtained as a weighed combination of the individual learner's outputs. Various algorithms have been proposed to determine the data and weak learner weights, with the popular AdaBoost multiplicative weights scheme being the most prominent and having rigorously established theoretical properties. We propose and analyze a sequence of distributionally robust formulations for both re-weighing problems with the objective of picking a compact ensemble that optimizes generalization performance. We devise computationally efficient estimation algorithms to implement these formulations, which provably balance a fundamental tradeoff between computation time and statistical variance. Numerical experiments demonstrate the significant benefits of our approach over previous related work.

Huang Zhiyuan: Conformal Data-Driven Chance Constraints

Lam Henry[1], Huang Zhiyuan[1]

1 - Columbia University (United States)

Abstract: Chance-constrained optimization is a widely popular approach to obtain decisions that balance objective performance and risk, by requiring the solution to satisfy safety constraints with a high probability. In data-driven contexts, existing approaches are designed to attain such a high-probability feasibility at a confidence level with respect to past data. This confidence guarantee, while intuitive, requires extra parameter specification that adds sensitivity and conservativeness. We propose an alternative statistical framework for data-driven chance-constrained optimization that gives conformal guarantees on the obtained decision. Our framework naturally integrates the probability level in the chance constraint and the data confidence level, thus removing the confidence parameter, and subsequently leading to less conservative sample size requirements and better objective performances than existing methods. Moreover, our framework applies to "conformalize" a range of more advanced methods including sampling-and-discarding, FAST and learning-based robust optimization.

3.6.2 Invited session T3-2 (Koole): Call centers (Room 101)

Thursday 29th, 15:30 - 17:00 - chaired by Ger Koole

Garlati Elena: Fractional staffing levels in Erlang models

Garlati Elena[1,2]

1 - Vrije Universiteit, Amsterdam (Netherlands)
2 - Internship at CCmath BV (Netherlands)

Abstract: Even though Erlang calculators give powerful results on safety staffing decicions, the Markov chains theory behind allows us to only work with integer staffing levels. In practice, when scheduling agents, an additional percentage is added to this staffing level. Since we are scheduling an integer number of agents, we are effectively rounding twice. In order to avoid this we should get a fractional staffing

level from the Erlang calculation. With this research, we give an overview on which are the different approaches to extend the Erlang models to a continuous setting in terms of staffing levels with the purpose of making thousands of computations in fractions of seconds. First, we derive the re-formulation of the performance measures as continuous functions in terms of number of agents for Erlang C systems from a theoretical point of view. Since this is not applicable to all the Erlang models, we subsequently introduce new approaches for the research of fractional staffing levels in Erlang C in order to extend this problems to other more complicated systems, such as Erlang A.

Li Siqiao: Data-driven prediction of service level for call centers

Li Siqiao[1], Akidas Hatim, Koole Ger

1 - Department of Mathematics, Vrije Universiteit Amsterdam (Netherlands)

Abstract: Predicting service levels accurately can help reduce over- /understaffing in workforce planning in call centers so as to save cost. The commonly used Erlang models in practice cannot capture agent behaviors such as different average handling times, taking break patterns, and so on. Moreover, the Erlang models fail in multi-skill situations. Whereas using simulation can capture more details in the call center, building a simulation model is challenging and expensive. In this work, we propose a data-driven method: using one-year call center data to train a machine-learning model to predict the service level. Our numerical results show that the proposed machine learning method is able to have a more accurate prediction of the service level than Erlang C and Erlang A with negligible computing time. Our method also requires less effort than Erlang models and simulation because dumping a historical dataset into the method is all it needs.

Özümerzifon Ömer: Analysis of time-dependent queues with retrials: Impact of retrial distribution

Özümerzifon Ömer[1], Legros Benjamin[2], Stolletz Raik[1]

1 - University of Mannheim (Germany)
2 - École de Management de Normandie (France)

Abstract: Queueing systems are used in various service systems, such as call centers, health care, emergency services, and repair facilities. In many of these service systems, especially in call centers, customers leave the queue before being served due to a lack of patience. However, impatient users may join the system at a later time again after some time (retrials). This research project focuses on the time-dependent performance evaluation of multi-server queueing systems with retrials, and the impact of retrial time distribution. In order to approximate performance measures in a time-dependent setting, we develop a stationary backlog-carryover (SBC) approach. We present the basic ideas of this approach and findings of a numerical study.

Ger Koole: The validation of call center workforce staffing models

Ger Koole[1]

1 - Vrije Universiteit, Amsterdam (Netherlands)

Abstract: Forecasting and mathematical models are used on a large scale to predict call center performance. Validating these models is difficult because we have non-identical dependent realizations, defying standard statistical validation methods. We discuss ways to decompose the errors into its components such as forecasting error, noise and model error. This is ongoing work currently undertaken with Siqiao Li.

3.6.3 Invited session T3-3 (Ayhan): Control of Queueing Systems (Room 102)

Thursday 29th, 15:30 - 17:00 - chaired by Hayriye Ayhan

Lewis Mark: A stochastic decision-making scenario in Telemedicine

Lewis Mark[1,2], Pender Jamol[2], Lu Shuwen[1]

1 - Cornell University (United States)
2 - School of Operations Research and Information Engineering (United States)

Abstract: We consider the question of when a nurse practitioner should call for a telemedicine consultation when low acuity patients are receiving service. A "minute clinic" is modeled using a stochastic network. Where in the past nurse practitioners made these decisions without explcitly considering upstream effects, incorporating the impact on other patients complicates decision-making. After modeling the clinic with a clearing system, we show when the structure of an optimal control is monotone and then develop heuristic strategies for implementation. Our results also provide guidance as to when a minute clinic should NOT be equpped with a telemedicine capability.

Rebuffi Louis-Sébastien: Reinforcement Learning in a Birth and Death Process: Breaking the Dependence on the State Space

Rebuffi Louis-Sébastien[1,2], Anselmi Jonatha[2], Gaujal Bruno[2]

1 - Université Grenoble Alpes (France)

2 - Inria (France)

Abstract: In this talk, we revisit the regret of undiscounted reinforcement learning in MDPs with a birth and death structure. Specifically, we consider a controlled queue with impatient jobs and the main objective is to optimize a trade-off between energy consumption and user-perceived performance. Within this setting, the diameter D of the MDP is $\Omega(S^S)$, where S is the number of states. Therefore, the existing lower and upper bounds on the regret at time T , of order $O(\sqrt{DSAT})$ for MDPs with S states and A actions, may suggest that reinforcement learning is inefficient here. In our main result however, we exploit the structure of our MDPs to show that the regret of a slightly-tweaked version of the classical learning algorithm UCRL2 is in fact upper bounded by $\tilde{O}(\sqrt{E_2AT})$ where E_2 is related to the weighted second moment of the stationary measure of a reference policy. Importantly, E_2 is bounded independently of S. Thus, our bound is asymptotically independent of the number of states and of the diameter. This result is based on a careful study of the number of visits performed by the learning algorithm to the states of the MDP, which is highly non-uniform.

Righter Rhonda: Policy-Space Collapse in Matching Models

Righter Rhonda[1], Xie Runhan, Gardner Kristen

1 - IEOR, UC Berkeley (IEOR UC Berkeley, CA 94720 United States)

Abstract: We consider one- and two-sided multi-class matching models and give sufficient conditions for policy-space collapse: the distribution of the number of each items of each class at any time is the same for any queue-position-based matching order, such as first-come-first-matched and last-come-first-matched. An important special case is redundancy(d) models, both under cancel-on-complete and cancel-on-start protocols.

Ayhan Hayriye: Optimal Pricing and Information Sharing in Queueing Systems

Ayhan Hayriye[1], Andradottir Sigrun[1], Xinchang Wang[2]

1 - School of Industrial and Systems Engineering [Georgia Tech] (United States)

2 - Department of Finance and Management Science, Carson College of Business, Washington State University (United States)

Abstract: We study optimal pricing in a single-server queueing system that can be observable or unobservable, depending on how customers receive information to estimate sojourn time. Our primary objective is to determine whether the service provider is better off making the system observable or unobservable under optimal pricing. We formulate the optimal pricing problem using Markov decision process (MDP) models for both observable and unobservable systems. For unobservable systems, the problem is studied using an MDP with a fixed-point equation as equilibrium constraints. We show that the MDPs for both observable and unobservable queues are special cases of a generalized arrivals-based MDP model, in which the optimal arrival rate (rather than price) is set in each state. Then, we show that the optimal policy that solves the generalized MDP exhibits a monotone structure in that the optimal arrival rate is non-decreasing in the queue length, which allows for developing efficient algorithms to determine optimal pricing policies. Next, we show that if no customers overestimate sojourn time in the observable system, it is in the interest of the service provider to make the system observable. We also show that if all customers overestimate sojourn time, the service provider is better off making the system unobservable. Lastly, we learn from numerical results that when customers are heterogeneous in estimating their sojourn time, the service provider is expected to receive a higher gain by making the system observable if on average customers do not significantly overestimate sojourn time.

3.6.4 Invited session T3-4 (Tran): Contact-tracing (Room 103)

Thursday 29th, 15:30 - 17:00 - chaired by Viet Chi Tran

Vo Thi Phuong Thuy: Exploration of SBM graphs by the Respondent Driven Sampling methods.

Vo Thuy[1]

1 - Laboratoire de Mathématiques et Modélisation d'Evry (France)

Abstract: The study of Respondent Driven Sampling (RDS) is invested for the discovery of a social network of hidden populations. It leads to the study of a Markov chain on a random graph whose vertices represent individuals and whose edges describe the relationships between the people connected. The RDS survey searches for hidden nodes in the population by randomly following the edges of the underlying social network, which allows us to trace the sampled individuals. We are interested in the problem of recovering statistical information on a SBM from the subgraph discovered by an exploring random walk (RDS with 1 coupon per interviewee). We consider here the dense case where the random network can be approximated by a graphon. First, we write the probability of the subgraph discovered by the random walk: biases emerge because the hubs and the majority types are more likely to be sampled. Even for the case where the types are observed, the maximum likelihood estimator is not explicit any more. When the types of the vertices are unobserved, we use an SAEM (Stochastic approximation of Expectation-Maximization) algorithm to maximize the likelihood. Second, we propose a different estimation strategy using new results by Athreya and Röllin. It consists in de-biasing the variational EM estimator proposed in Daudin et al. and that ignores the biases.

Zhang Dongni: Epidemic models with manual and digital contact tracing

Zhang Dongni[1], Britton Tom[1]

1 - Stockholm University (Sweden)

Abstract: We start with a Markovian SIR epidemic model in a homogeneous mixing community with a constant diagnosis rate. First, we introduce manual contact tracing by assuming that once an infective is diagnosed (tested positive and isolated), each of her/his contacts is immediately traced and tested independently with some fixed probability. Using large population approximations, we analyzed the early stage of the outbreak when the process of "to-be-traced components" behaves like a branching process. The component and individual reproduction numbers are derived. Then we focus on the more recent digital contact tracing via a tracing app (only app-users can trigger and be traced by digital tracing). And we assume that there is a fixed app-using fraction and that digital tracing occurs instantaneously and recursively. The model with digital tracing is analysed by a two-type branching process relying on a large community, where one type is "app-using components," and another is non-app-users. Further, we investigate the combined preventive effect of manual and digital tracing. This combined model is analysed by a different two-type branching process with both types being the "to-be-traced components" but starting with different "roots". The corresponding reproduction numbers are derived. We conclude that to control the epidemic, we always need to have a large fraction of app-users compared to the fraction of individuals being successfully reached by manual tracing. Another important conclusion is that the combined effect is bigger than the product of two separate preventive effects. Finally, our ongoing work is based on an SEIR epidemic model on a configuration model, where transmission could happen from random contacts (e.g., on a bus), which are usually more easily identified by using the tracing app. Each infective remains infectious for a constant period. Meanwhile, we assume that such an infective is diagnosed and asked for an interview with some probability otherwise, we say that the infective recovers naturally. Once diagnosed, each of her/his infectee neighbours (manual tracing only on the network) is reported with some probability independently. If such reported neighbours are infectious after some delay, they are isolated and said to be traced. If the diagnosed person is also an app-user, all of her/his app-using contacts are traced immediately (digital tracing both on the network and among global contacts). And in this case, we assume that only the diagnosed individuals can trigger contact tracing.

Stegehuis Clara: Contact tracing and network structures

Stegehuis Clara[1]

1 - University of Twente [Netherlands] (Netherlands)

Abstract: Quarantining and contact tracing are popular practices to mitigate epidemic outbreaks. However, the effectiveness of these practices depends on the network on which the epidemic spreads. We therefore study how the final size of an epidemic is influenced by contact tracing and quarantining on a network null model: the configuration model. In this process, infected vertices may self-quarantine and trace their infector with a given success probability. A traced infector is less likely to infect others. We show that the effectiveness of such tracing processes strongly depends on the network structure. However, in contrast to previous findings, the tracing procedure is not necessarily more effective on networks with

heterogeneous degrees. We also show that network clustering influences the effectiveness of the tracing process in a non-trivial way.

Doroudi Sherwin: Evaluating Disease Transmission Risks in Small-Scale Settings Under the Possibility of Multiple Infectors

Doroudi Sherwin[1], Kang Kang

1 - University of Minnesota [Twin Cities] (United States)

Abstract: Mathematical models of disease transmission in small scale setting typically focus on static environments with a limited number of individuals. Recent work inspired by the ongoing COVID-19 pandemic has considered disease transmission in environments that behave like queueing models (i.e., where individuals arrive and depart over time). One such framework builds upon the exponential dose response model in order to extend the epidemiological concept of the basic reproduction rate to the "system-specific basic reproduction rate." This system-specific metric is defined as the expected number susceptible individuals that a single infectious individual will infect during a single visit to a queueing system, under the single-infector assumption that all other individuals in the system are susceptible. Our work expands this framework by consider multiple-infector models, which relax the assumption that at most one individual is infectious, and account for the fact that susceptible individuals are more at risk when in the presence of a greater number of infectious individuals. While the analysis of single-infector models is often simplified by employing the linearity of expectation, multiple-infector models require nuanced analysis that takes into account a variety of correlations. Using a variety of queueing-theoretic techniques to overcome these challenges, we derive expressions for the probability that a susceptible individual becomes infected in a variety of multiple-infector models, including those built upon the M/M/1 (under both FCFS and PLCFS scheduling orders), M/M/1/k, and M/M/infinity, queueing systems; we show that these results are consistent with existing results for the system-specific base reproduction rate in limiting regime where the probability that an individual is infectious tends to zero. Our findings enable to us to assess interventions and identify settings where it is necessary to consider multiple-infector models.

3.6.5 Invited session T3-5 (Penington): Branching processes and their applications (Room 104)

Thursday 29th, 15:30 - 17:00 - chaired by Sarah Penington

Gonzalez Casanova Adrian: The speed of (a variant of) Muller's ratchet can be calculated using multi-type logistic branching processes

Gonzalez Casanova Adrian[1], Smadi Charline[2], Wakolbinger Anton[3]

1 - UC Berkeley and UNAM (Mexico)
2 - Grenoble (France)
3 - University of Frankfurt (Germany)

Abstract: Consider a population of N individuals, each of them carrying a type in $\mathbb{N}_0$. The population evolves according to a Moran dynamics with selection and mutation, where an individual of type k has the same selective advantage over all individuals with type $k' > k$, and type k mutates to type $k+1$ at a constant rate. This model is thus a variation of the classical Muller's ratchet: there the selective advantage is proportional to $k' - k$. For a regime of selection strength and mutation rates which is between the regimes of weak and strong selection/mutation, we obtain the asymptotic rate of the click times of the ratchet (i.e. the times at which the hitherto minimal best' type in the population is lost), and reveal the quasi-stationary type frequency profile between clicks. The large population limit of this profile is characterized as the normalized attractor of a "dual" hierarchical multitype logistic system, and also via the distribution of the final minimal displacement in a branching random walk with one-sided steps. An important role in the proofs is played by a graphical representation of the model, both forward and backward in time, and a central tool is the ancestral selection graph decorated by mutations.

Hartung Lisa: Systems of F-KPP equations

Hartung Lisa[1]

1 - JGU Mainz (Germany)

Abstract: In this talk I will give an overview on the conncetion between (systems of) F-KPP equations and spatial branching processes. Then I explain how probabbilistic tools can be used to analyse the solutions to these equations. The talk is based on joint work with A. Bovier

Pain Michel: The height of weighted recursive trees

Pain Michel[1], Sénizergues Delphin[2]

1 - Université de Toulouse (France)

2 - Université Paris Nanterre - Département de Mathématiques et Informatique (France)

Abstract: Weighted recursive trees are built by adding successively vertices with predetermined weights to a tree: each new vertex is attached to a parent chosen at random with probability proportional to its weight. I will present results obtained with Delphin Sénizergues concerning the height of these trees. When the total weight of the tree grows polynomially, we show that the height is in the same universality class as the maximum of supercritical branching random walks. When the total weight grows sub-polynomially, universality is broken and we exhibits several new regimes.

Villemonais Denis: Some asymptotic results for discrete time weighted branching particle systems

Villemonais Denis[1]

1 - Institut Élie Cartan de Lorraine (France)

Abstract: Discrete time weighted branching particle systems are branching processes where each offspring is attributed a weight and a type, and where each generation can be composed of countably infinitely many individuals with a finite total weight. During this talk, I will present time asymptotic results for these processes.

3.6.6 Invited session T3-6 (Wang): Queues, bandits, and reinforcement learning (Room 105)

Thursday 29th, 15:30 - 17:00 - chaired by Weina Wang

Agrawal Shubhada: CRIMED: Lower and Upper Bounds on Regret for Bandits with Unbounded Stochastic Corruption

Agrawal Shubhada[1], Mathieu Timothee[2], Basu Debabrota[2], Maillard Odalric-Ambrym[2]

1 - School of Industrial and Systems Engineering [Georgia Tech] (United States)

2 - Scool (France)

Abstract: We study the regret-minimization problem in the multi-armed bandit setting with *unbounded stochastic corruption*, where the agent might not always observe the reward generated from the selected arm but instead sees a sample generated from an arbitrary corruption distribution with potentially *unbounded support*. The amount of corruption is controlled by a corruption proportion $\varepsilon \in (0, \frac{1}{2})$. In this talk, we will first develop a generic problem-dependent lower bound on the regret that holds for a given family of reward distributions with arbitrary corruption distributions. For the setting with Gaussian arm-distributions with known variance, we will then see an asymptotically-optimal algorithm (`CRIMED`) matching the lower bound exactly. Notably, `CRIMED` is the first algorithm that can handle corruption levels up to 1/2. We also provide numerical results showing that `CRIMED` significantly outperforms the existing algorithms for this setting.

You Wei: Information-Directed Selection for Top-Two Algorithms

You Wei[1], Qin Chao[2], Wang Zihao[1], Yang Shuoguang[1]

1 - the Hong Kong University of Science and Technology (Hong Kong SAR China)

2 - Columbia University (United States)

Abstract: We consider the best-k-arm identification problem for multi-armed bandits, where the objective is to select the exact set of k arms with the highest mean rewards by sequentially allocating measurement effort. We characterize the necessary and sufficient conditions for the optimal allocation using dual variables. Remarkably these optimality conditions lead to the extension of top-two algorithm design principle (Russo, 2020), initially proposed for best-arm identification. Furthermore, our optimality conditions induce a simple and effective selection rule dubbed information-directed selection (IDS) that selects one of the top-two candidates based on a measure of information gain. As a theoretical guarantee, we prove that integrated with IDS, top-two Thompson sampling is (asymptotically) optimal for Gaussian best-arm identification, solving a glaring open problem in the pure exploration literature (Russo, 2020). As a by-product, we show that for $k > 1$, top-two algorithms cannot achieve optimality even with an oracle tuning parameter. Numerical experiments show the superior performance of the proposed top-two

algorithms with IDS and considerable improvement compared with algorithms without adaptive selection.

3.6.7 Invited session T3-7 (Backhausz): Random graphs and matrices (Room 106)

Thursday 29th, 15:30 - 17:00 - chaired by Agnes Backhausz

Holmgren Cecilia: The asymptotic distribution of cluster sizes for supercritical percolation on random split trees

Holmgren Cecilia[1], Berzunza Gabriel[2]

1 - Uppsala Universitet [Uppsala] (Sweden)
2 - University of Liverpool (United Kingdom)

Abstract: We consider the model of random trees introduced by Devroye (1998), the so-called random split trees. The model encompasses many important randomized algorithms and data structures. We then perform supercritical Bernoulli bond-percolation on those trees and obtain the asymptotic distribution for the sizes of the largest clusters. Gabriel Berzunza, Cecilia Holmgren, Random Structures and Algorithms, 2022

Mathien Joffrey: Diameter of random surfaces built from random graphs.

Mathien Joffrey[1]

1 - Aix-Marseille Université - Faculté des Sciences (France)

Abstract: The study of random surfaces, especially in the asymptotic of large genus, is of increasing interest in recent years. Many geometrical questions have analogous formulations in the theory random graphs with a large number of vertices, and results obtained in one domain can inspire the other. In this way, we get interested in the diameter of random surfaces, which is a basic measure of the geometry of the surface. In 2019, Budzinski, Curien and Petri studied this measure for a special model of surfaces, built from random graphs. We extend it to obtain a richer class of models of random surfaces and we compute the asymptotic of the diameter of these surfaces. The strategy of the proof relies on a detailed study of an exploration process which is the analog to the breadth-first search exploration of a random graph. Its analysis is notably based on subadditive and concentration techniques.

Müller Noela: The rank of sparse symmetric matrices over arbitrary fields

Müller Noela[1], Van Der Hofstad Remco[1], Zhu Haodong[1]

1 - Eindhoven University of Technology (Netherlands)

Abstract: The talk discusses a combinatorial approach to study the asymptotic normalised rank of sparse random matrices with arbitrary edge weights. Here, the edge weights can be assigned from an arbitrary field, e.g. a finite field, and thus also the ensuing rank is considered with respect to that field. The presented result shows that for a mildly general class of sparse symmetric matrices, the asymptotics of the normalised rank are independent of potential edge weights in the graph and the field, in the sense that the limiting constant of the normalised rank coincides with previously established results for symmetric matrices over the reals with 0/1-entries. The underlying proof is based on an intricate extension of the novel perturbation approach from Coja-Oghlan, Ergür, Gao, Hetterich and Rolvien to the symmetric setting.

Shneer Seva: First passage percolation on Erdos-Renyi graphs with general weights

Daly Fraser[1], Schulte Matthias[2], Shneer Seva[1]

1 - Heriot-Watt University [Edinburgh] (United Kingdom)
2 - Technical University Hamburg Harburg [Hamburg] (Germany)

Abstract: We consider an Erdos-Renyi random graph on n nodes where the probability of an edge being present between any two nodes is equal to λ/n with $\lambda > 1$. Every edge is assigned a (non-negative) weight independently at random from a general distribution. For every path between two typical vertices we introduce its hop-count (which counts the number of edges on the path) and its total weight (which adds up the weights of all edges on the path). We prove a limit theorem for the joint distribution of the appropriately scaled hop-count and general weights. This theorem, in particular, provides a limiting result for hop-count and the total weight of the shortest path between two nodes. This is a joint work

with Fraser Daly and Matthias Schulte.

3.6.8 Invited session T3-8 (Comte): Online stochastic matching: a diversity of viewpoints (Room 202)

Thursday 29th, 15:30 - 17:00 - chaired by Céline Comte

Jonckheere Matthieu: Online matching for the multitype stochastic block model

Jonckheere Matthieu[1], Moyal Pascal, Soprano Loto Nahuel

1 - CNRS (France)

Abstract: We consider the problem of sequential matching in a stochastic block model (SBM) with several types of nodes and generic compatibility constraints. When the probabilities of connections do not scale with the size of the graph, we show that under the NCOND condition, a simple max-weight policy allows to attain an asymptotically perfect matching while no sequential algorithm attain perfect matching otherwise. The proof relies on a specific Markovian representation of the dynamics associated with Lyapunov techniques.

Kerimov Süleyman: Near-Optimal Policies for Dynamic Matching

Kerimov Süleyman[1], Ashlagi Itai[2], Gurvich Itai[3]

1 - Rice University (United States)
2 - Stanford University (United States)
3 - Northwestern University (United States)

Abstract: We study centralized dynamic matching markets with finitely many agent types and heterogeneous match values. An inherent trade-off arises between short- and long-term value. A social planner may delay match decisions to thicken the market and increase match opportunities to generate high value. This inevitably compromises short-term objectives, and the planner may match greedily to maximize short-term value. A matching policy is hindsight optimal if the policy can (nearly) maximize the total value simultaneously at all times. We first establish that in multi-way networks, where a match can include more than two agent types, a simple periodic clearing policy with a carefully chosen period length is hindsight optimal. Interestingly, in two-way networks, where any match includes two agent types, suitably designed greedy policies also achieve hindsight optimality.

Reiffenhäuser Rebecca: Online Max-Weight Matchings with Limited Priors

Reiffenhäuser Rebecca[1], Caramanis Constantine[2], Lazos Philip[3], Fusco Federico[4], Dütting Paul[5], Faw Matthew[2], Leonardi Stefano[6], Papadigenopoulos Orestis[7], Pountourakis Emmanouil[8]

1 - Universiteit van Amsterdam (Netherlands)
2 - University of Texas at Austin [Austin] (United States)
3 - Input Output (United Kingdom)
4 - Sapienza - University of Rome (Italy)
5 - google research (Switzerland)
6 - Sapienza University of Rome (Italy)
7 - Columbia University (United States)
8 - Drexel University (United States)

Abstract: When matchings are computed online, access to distributional information has a huge impact on the achieved competitive guarantees. This talk explores a setting with a minimal amount of such knowledge, i.e. single-sample prophet inequalities, and gives small constant-factor approximations for different settings, including that of matching strategic agents to goods, as well as discussing extensions to combinatorial auctions.

Zubeldia Martin: Matching Queues with Abandonments in Quantum Switches: Stability and Throughput Analysis

Zubeldia Martin[1], Jhunjhunwala Prakirt[2], Maguluri Siva Theja[2]

1 - University of Minnesota [Twin Cities] (United States)
2 - Georgia Institute of Technology [Atlanta] (United States)

Abstract: Inspired by quantum switches, we consider a discrete-time multi-way matching system with two classes of arrivals: requests for entangled pair of qubits between two nodes, and qubits from each node that can be used to serve the requests. An important feature of this model is that qubits decohere and so

abandon over time. In contrast to classical server-based queueing models, the combination of queueing, server-less multi-way matching, and abandonment make the analysis a challenging problem. The primary focus of this paper is to study a simple system consisting of two types of requests and three types of qubits operating under a Max-Weight policy. In this setting, we characterize the stability region under the Max-Weight policy by adopting a two-time scale fluid limit to get a handle on the abandonments. In particular, we show that Max-Weight is throughput optimal and that it can achieve throughputs larger than the ones that can be achieved by non-idling policies when the requests are infinitely backlogged. Moreover, despite the use of the Max-Weight policy, we show that there can be a counter-intuitive behavior in the system: the longest requests queue can have a positive drift for some time even if the overall system is stable.

3.6.9 Invited session T3-10 (Decreusefond): Stein method II (Room 204)

Thursday 29th, 15:30 - 17:00 - chaired by Laurent Decreusefond

Halconruy Hélène: Stochastic analysis for marked binomial processes and Poisson approximations

Halconruy Hélène[1]

1 - ESILV (France)

Abstract: Clumps can be observed by studying the number of occurrences of a rare word in a DNA sequence. The Chen-Stein method provides tools to bound the error by approximating the distribution of the number of clusters by a Poisson law (possibly compound). In the work that supports this presentation, we propose to revisit Chen-Stein method by reducing the initial problem to that of a compound Poisson approximation for functionals of binomial marked processes (MBPs). Those are discrete analogues of compound Poisson processes for which we develop stochastic analysis tools as well as a Malliavin calculus based on a family of operators (gradient, divergence, etc.). Under this new formalism, we obtain a general criterion - for the total variation distance - of the poissonian approximation for MBP functionals in terms of Malliavin operators. In this talk, I will present elements of the Malliavin calculus for MBPs, before stating the general approximation result and illustrating it to study the distribution of a rare word in a DNA sequence.

Peccati Giovanni: Stein's method and quantitative two-scale stabilization

Peccati Giovanni[1]

1 - University of Luxembourg [Luxembourg] (Luxembourg)

Abstract: I will discuss some bounds on the normal approximation of (possibly multidimensional) functionals of Poisson random measures, implicitly using the geometric notion of "two-scale stabilization". These estimates can be used to deduce explicit estimates on the rate of convergence of geometric functionals of random point configurations, in situations (related e.g. to problems of combinatorial optimization) where iterated add-one cost operators (and, consequently, second-order Poincaré inequalities) are not amenable to analysis.

Swan Yvik: Normal approximation for the posterior in exponential families

Swan Yvik[1]

1 - Université Libre de Bruxelles (Belgium)

Abstract: We obtain quantitative Bernstein-von Mises type bounds on the normal approximation of the posterior distribution in exponential family models when centering either around the posterior mode or around the maximum likelihood estimator. Our bounds, obtained through a version of Stein's method, are non-asymptotic, and data dependent; they are of the correct order both in the total variation and Wasserstein distances, as well as for approximations for expectations of smooth functions of the posterior. All our results are valid for univariate and multivariate posteriors alike, and do not require a conjugate prior setting. We illustrate our findings on a variety of exponential family distributions, including Poisson, multinomial and normal distribution with unknown mean and variance. The resulting bounds have an explicit dependence on the prior distribution and on sufficient statistics of the data from the sample, and thus provide insight into how these factors may affect the quality of the normal approximation. The performance of the bounds is also assessed with simulations.

Decreusefond Laurent: Stein's method for discrete alpha stable point processes

Decreusefond Laurent[1], Vasseur Aurélien[1]

1 - Télécom Paris (France)

Abstract: Many probability distributions satisfy some stability properties (alpha-stable distribution and among them, the Gaussian distribution, Poisson distribution, max-stable, etc.) We are here interested in a class of point processes introduced by Davydov et al. which obeys a stability with respect to thinning operations. We define a Ornstein-Uhlenbeck like semigroup thanks to this property and then show how to construct a Malliavin derivative. We can then the usual machinery of Stein-Malliavin calculus to derive some new convergence rates for stable-CLTs.

3.7 Friday 30th, 10:30 - 12:00

3.7.1 Invited session F1-1 (Robert): Stochastic Models of Biology (Auditorium)

Friday 30th, 10:30 - 12:00 - chaired by Philippe Robert

Zaherddine Jana: Stochastic Models of Regulation of Gene Expression in Biological Cells

Zaherddine Jana[1]

1 - Inria (France)

Abstract: Gene expression is a fundamental process in cellular life. It is a two step process: Transcription (production of RNAs) and Translation (production of proteins). We study an important regulation mechanism of transcription of bacterial cells using specific macro-molecules, 6S RNAs. The functional property of these 6S RNAs is of blocking the transcription of RNAs when the cell environment is not favorable to its growth. In this talk, after a brief introduction, we will investigate the efficiency of this mechanism with a stochastic model in different growth regimes of the cell. The results are mainly based on the proof of averaging principles for the associated Markov processes. The consequences of these results are discussed.

Vignoud Gaetan: Spike-timing dependent plasticity in stochastic neuronal networks

Vignoud Gaetan[1,2,3]

1 - Dioxycle (France)
2 - Inria de Paris (France)
3 - Centre interdisciplinaire de recherche en biologie (France)

Abstract: In neuroscience, synaptic plasticity refers to the set of mechanisms driving the dynamics of the connections between neurons. These synapses and their evolution have a great influence on learning and memory, and can be represented by a scalar value, the synaptic weight. We will consider a stochastic system with two connected neurons, with a variable synaptic weight that depends on point processes associated to each neuron. The input neuron spikes are generated by an homogeneous Poisson process, while the output neuron has a spiking intensity that depends on the jumps of the input neuron and the synaptic weight. A scaling regime where the rate of both point processes is large compared to the dynamics of the synaptic weight corresponds to a classical assumption in computational neuroscience, and offers interesting properties. Using this stochastic averaging principle, we will present different regimes of the synaptic weight dynamics for various plasticity rules, with both analytical and numerical results.

Véber Amandine: Growth properties of the infinite-parent spatial Lambda-Fleming-Viot process

Véber Amandine[1]

1 - MAP5, CNRS, Université Paris Cité (France)

Abstract: The infinite-parent spatial Lambda-Fleming-Viot process is a model for spatially expanding populations in a two dimensional continuum, in which empty areas are filled with ghost individuals. This model can be seen as a continuous-space version of the Eden growth model, and it comes with a dual process that allows us to trace back the origins of a sample of individuals taken from the current

population. In this talk, we shall focus on the growth properties of the area covered by real individuals. With the help of a simple toy model, we shall also investigate how the fluctuations at the front edge lead to a much larger speed of growth of the occupied region than that predicted by simple first-moment estimates. Joint work with Apolline Louvet (University of Bath).

Campillo Fabien: Nonlinear filtering in neurosciences

Campillo Fabien[1]

1 - Equipe MODEMIC INRA/INRA (UMR MISTEA, 2 Place Pierre Viala, Bat 29, Montpellier, France France)

Abstract: Nonlinear filtering is a set of tools that allow to identify, in a dynamic and recurrent way, a hidden component from observations in a Markov context. Does this take a particular form in neuroscience?

3.7.2 Invited session F1-2 (Stolletz): Stochastic modeling of manufacturing and service operations (Room 101)

Friday 30th, 10:30 - 12:00 - chaired by Raik Stolletz

Karabag Oktay: Production and Energy Mode Control of a Production-Inventory System

Karabag Oktay[1,2], Tan Barıs[3], Khayyati Siamak[4]

1 - Yaşar University (Turkey)
2 - Erasmus University Rotterdam (Netherlands)
3 - Koç University (Turkey)
4 - University of Mannheim (Germany)

Abstract: Energy efficiency in manufacturing can be improved by controlling energy modes and production dynamically. We examine a production-inventory system that can operate in Working, Idle, and Off energy modes with mode-dependent energy costs. There can be a warm-up delay to switch from one mode to another. With random inter-arrival, production, and warm-up times, we formulate the problem of determining in which mode the production resource should operate at a given time depending on the state of the system as a stochastic control problem under the long-run average profit criterion considering the sales revenue together with energy, inventory holding and backlog costs. The optimal solution to the problem for the exponential inter-arrival, production, and warm-up times is determined by solving the Markov Decision Process with a linear programming approach. The structure of the optimal policy for the exponential case uses two thresholds to switch between the Working and Idle or Working and Off modes. We use the two-threshold policy as an approximate policy to control a system with correlated inter-event times with general distributions. This system is modeled as a Quasi Birth and Death Process and analyzed by using a matrix-geometric method. Our numerical experiments show that the joint production and energy control policy performs better compared to the pure production and energy control policies depending on the system parameters. In summary, we propose a joint energy and production control policy that improves energy efficiency by controlling the energy modes depending on the state of the system.

Khayyati Siamak: A Hybrid Approximation Method for Time-Dependent Queues in Service Systems

Khayyati Siamak[1], Foroughi Amir, Stolletz Raik[1]

1 - University of Mannheim (Germany)

Abstract: Many service systems have time-dependent parameters e.g. queues in healthcare facilities and call centers. For this reason, analysis of time-dependent queues is of great interest. Among the various available approximation methods for time-dependent queues, the fluid/diffusion-based methods, and the stationary-backlog-carryover (SBC) method are efficient and accurate. However, the fluid/diffusion based models do not perform well for under-loaded periods and the SBC method's performance is negatively affected by the over-loaded periods. To remedy these drawbacks, we develop a hybrid method that alternates between the two methods aiming to apply each one when they perform best. Finally using numerical experiments, we show the accuracy and the efficiency of the hybrid method.

Schwarz Justus Arne: A production planning model for non-stationary and stochastic yield in the semiconductor industry

Schwarz Justus Arne[1], Diefenbach Johannes[2], Karaesmen Fikri[3], Stolletz Raik[2]

1 - University of Regensburg (Germany)
2 - Universität Mannheim [Mannheim] (Germany)
3 - Koç University (Turkey)

Abstract: In the semiconductor industry, the yield of a production process tends to start low but increases with cumulative production volume when a new product or machine is introduced. Due to long production lead times, the company has to choose the production quantity before observing the realized yield. We formalize the firm's problem as a newsvendor problem with stochastic and non-stationary yield and deterministic demand. We derive analytical and numerical insights into the optimal quantity to produce and the expected profit. For stationary yield, we characterize the optimal production by a critical-fractile and discuss the sensitivity of the results to non-stationary yield. A numerical study shows that the optimal production quantity using the proposed model is close to the ex-post optimal production quantity for real yield data.

Rochlin Igor: The Impact of Outdated Information and the Optimal Up- date Policy

Rochlin Igor[1], Kerner Yoav[1], Roet-Green Ricky[2]

1 - Ben-Gurion University of the Negev (Israel)
2 - University of Rochester [USA] (United States)

Abstract: Many services, such as emergency rooms, border-crossing agencies and airports provide wait time information to their customers. However, the information is not updated continuously, due to limitations of the wait-time tracking technology. We propose a model in which the information regarding the system's state is not updated continuously. We assume that the provider shares accurate information with their customers. The updating process follows an endogenous Poisson process. Customers make joining decisions knowing that the information might be outdated. They use the posted information to infer their expected waiting time in the system. Our model bridges between the classical observable and unobservable queueing models: if the time between updates goes to zero, the queue becomes observable; if the time between updates goes to infinity, the queue becomes unobservable. In our research, we calcualted the expected waiting given the outdated queue-length information. It followed by a proof of the uniquenss of equlibirum strategy of the customers as a function of the update rate and system's parameters. We then show some numerical examples of system's performence measures: throughput and social welfare. In most cases, as expected, either the fully observable or the unobservable model, is optimal. However, there are cases in which the throughput and/or the social welfare are unimodal functions with respect to the update rate. We conclude with the question whether there is value of sharing the update age with the customers

3.7.3 Invited session F1-3 (Henderson): Healthcare (Room 102)

Friday 30th, 10:30 - 12:00 - chaired by Shane Henderson

Chick Stephen: Adaptive Sequential Clinical Trials for Precision Medicine with Predictive and Prognostic Covariates

Chick Stephen[1], Alban Andres[2], Zoumpoulis Spyros[1]

1 - INSEAD (France)
2 - Massachussetts General Hospital (United States)

Abstract: Motivated by work with collaborators at a university medical center, we sought to develop techniques improve the rate of learning in clinical trials for precision medicine. In such applications, there may be structural knowledge as to which covariates might react with various treatments in the trial, and other covariates that may affect outcomes independent of treatment. We extend the expected value of information/knowledge gradient approach to account for such covariate structures, differences from LASSO-type approaches, and illustrate computational results for a realistic example.

Van Der Mei Rob: Reducing waiting times by coordination in complex acute healthcare systems

Van Der Mei Rob[1,2]

1 - Centrum Wiskunde & Informatica (P.O. Box 94079 1090 GB Amsterdam Netherlands Netherlands)
2 - VU University Amsterdam (Faculty of Sciences De Boelelaan 1081 1081 HV Amsterdam The Netherlands Netherlands)

Abstract: In many countries, the current system of acute care (e.g., elderly care, youth care, mental healthcare) is not functioning optimally. As a consequence, patients-in-need often experience major difficulties in receiving prompt, adequate and appropriate care when really needed. Temporary care is for example not always available, patients who cannot go straight-home after a hospital admission are often difficult to place, and the hospital's emergency department is structurally overloaded. These problems typically arise from a lack of coordination in today's complex system of healthcare at the macro level, and from an interest bottleneck' between the different stakeholders involved. In this talk, I will show how quantitative models and methods can make the difference.

Henderson Shane: Community First Responders: Recruitment Guidelines and Dispatch

Henderson Shane[1], Van Den Berg Pieter[2], Fourmentraux Oceane[3], Jagtenberg Caroline[3], Li Hemeng (maggie)[4]

1 - Cornell University [New York] (United States)
2 - Erasmus University Rotterdam (Netherlands)
3 - Vrije University (Netherlands)
4 - Cornell University (United States)

Abstract: Patient survival from out-of-hospital cardiac arrest (OHCA) can be improved by augmenting traditional ambulance response with the dispatch of community first responders (volunteers) who are alerted via an app. How many volunteers are needed, from where should volunteers be recruited, and how should they be dispatched? We use a combination of Poisson point process modeling and convex optimization to address the first two questions; the right areas from which to recruit are not always obvious, because volunteers recruited from one area may spend time in various areas across a city. We use a combination of dynamic programming and decision trees to answer the last question, balancing the goal of a fast response to the current patient with the need to avoid disengagement of volunteers that arises when multiple volunteers respond. A case study for Auckland, New Zealand demonstrates the ideas.

3.7.4 Contributed session F1-4 (Wang): Networks and optimization (Room 103)

Friday 30th, 10:30 - 12:00 - chaired by Jinting Wang

Lim Andrew: Mechanisms for Coordinating Systems of Decentralized Agents

Lim Andrew[1]

1 - National University of Singapore (Singapore)

Abstract: We consider a decentralized stochastic service system where arrival rates and service rates are controlled by different agents. Not surprisingly, the aggregate system is not efficient if agents optimize individual costs in isolation, so the challenge is to come up with incentives under which the collection of decentralized decisions optimizes the system. We characterize transfer payments between agents under which the collection of decentralized decisions coincides with the centralized optimal, and propose a decentralized algorithm for computing the optimal transfers. The transfer payments are related to the VCG mechanism from mechanism design, and the algorithm for computing them is incentive compatible and does not require the services of a mechanism designer to be implemented. Convergence and robustness properties of both mechanisms will be discussed.

Avramidis Thanos: Dynamic Pricing with Finite Price Sets via Upper Confidence Bounds

Avramidis Athanasios (thanos)[1]

1 - School of Mathematics, University of Southampton (United Kingdom)

Abstract: We study price optimization of perishable inventory over multiple, consecutive selling seasons in the presence of demand uncertainty. Each selling season consists of a finite number of discrete time periods, and demand per time period is a Bernoulli random variable with price-dependent parameter. The set of feasible prices is finite, and the expected demand (Bernoulli parameter) corresponding to each price is unknown to the seller, whose objective is to maximize cumulative expected revenue. Under full information (that is, upfront knowledge of the Bernoulli parameters), an optimal pricing policy is determined by a dynamic program with finite horizon and finite state and action sets. In the absence of such knowledge, we propose a pricing algorithm such that in each season the pricing policy is determined as the solution to the revised dynamic program where the Bernoulli parameters are replaced by respective (current) upper confidence bounds. Revenue performance is measured by the regret: the expected revenue

loss relative to the optimal attainable revenue under full information. With the number of seasons n known upfront, our main result is an $O(log(nlog(n)))$ upper bound on the regret. The analysis hinges on concentration results (McDiarmid 1998, Cappe et al 2013). Numerical comparisons to alternatives from the literature indicate the competitiveness of our approach. References McDiarmid, C. (1998), Concentration, in Probabilistic methods for algorithmic discrete mathematics', Springer, pp. 195-248. CappÂťe, O., Garivier, A., Maillard, O.-A., Munos, R. and Stoltz, G. (2013), Kullback-leibler upper confidence bounds for optimal sequential allocation', The Annals of Statistics 41(3), 1516-1541.

Jiang Jiashuo: Degeneracy is OK: Logarithmic Regret for Network Revenue Management with Indiscrete Distributions

Jiang Jiashuo[1], Ma Will[2], Zhang Jiawei[3]

1 - Hong Kong University of Science and Technology (Hong Kong SAR China)
2 - Columbia University School of Business (United States)
3 - NYU-Stern School of Business (United States)

Abstract: We study the classical Network Revenue Management (NRM) problem with accept/reject decisions and T IID arrivals. We consider a distributional form where each arrival must fall under a finite number of possible categories, each with a deterministic resource consumption vector, but a random value distributed continuously over an interval. We develop an online algorithm that achieves $O(\log^2 T)$ regret under this model, with no further assumptions. We develop another online algorithm that achieves an improved $O(\log T)$ regret, with only a second-order growth assumption. To our knowledge, these are the first results achieving logarithmic-level regret in a continuous-distribution NRM model without further "non-degeneracy" assumptions. Our results are achieved via new techniques including: a new method of bounding myopic regret, a "semi-fluid" relaxation of the offline allocation, and an improved bound on the "dual convergence".

Wang Jinting: Value-at-Risk Based Queueing

Wang Jinting[1], Zhu Sheng[2]

1 - Central University of Finance and Economics (China)
2 - Henan Polytechnic University (China)

Abstract: The value-at-risk (VaR) based queueing economics in observable/unobservable queues with strategic but non-risk-neutral customers is studied. Confidence level and risk tolerance value constitute the VaR measure that is used to characterize customers' risk preference in which homogeneous and heterogeneous cases are both studied. We find that the optimal strategies of non-risk-neutral customers are still of threshold-type in the observable model, albeit in a quite different way compared with risk-neutral customers. The individually optimal threshold is non-increasing in the confidence level and nondecreasing in the tolerance value. In the unobservable model, we obtain the one-dimensional (multi-dimensional) equilibrium joining probability when customers have the homogeneous risk preference (heterogeneous risk preferences). From the viewpoint of social welfare maximization, in the case of homogeneous risk preference, the socially optimal risk preference belongs to a "ribbon" region in the observable model, while it forms a curve in the unobservable case. When customers have heterogeneous risk preferences, the socially optimal confidence level is a point if the real-time queue length is unobservable, and it belongs to a region if the real-time queue length is observable. We reveal that from the perspective of throughput maximization, it is better not to disclose the real-time queue length information when the risk confidence level is sufficiently low. From the perspective of social welfare, disclosing the real-time information is more conducive to the social welfare in the case of homogeneous risk preference, while hiding the information can gain more social welfare if customers are heterogeneous.

3.7.5 Invited session F1-5 (De Saporta): Learning while controling: crossroads between stochastic control and statistical learning (Room 104)

Friday 30th, 10:30 - 12:00 - chaired by Benoîte De Saporta

Le Quellennec Orlane: An example of medical treatment optimization under model uncertainty

Le Quellennec Orlane[1], De Saporta Benoîte[1], Cleynen Alice[1], Sabbadin Régis[2]

1 - IMAG, Univ Montpellier, CNRS, Montpellier, France (France)
2 - Univ Toulouse, INRAE-MIAT, Toulouse, France (France)

Abstract: Human diseases such as cancer involve long-term follow-up. A patient alternates phases of remission with relapses. A biomarker is monitored throughout the follow-up. Its dynamic is modeled by a controlled piecewise deterministic Markov process (PDMP). The PDMP evolves in continuous time and space, the process is partially observed through noise and some of its parameters are unknown, making the control problem especially difficult. To our knowledge, there is no method to control such a PDMP, i.e. to maximize the life of the patient while minimizing the treatment cost and side effects. First, we consider discrete dates only for the decisions, thus turning the controlled PDMP into a partially observed Markov decision process (POMDP), with a continuous state space. Then, through simulations, we compare Bayesian and non-Bayesian reinforcement learning methods to solve this POMDP.

Maillard Odalric-Ambrym: Regret minimization in MDPs: Lower bound-guided strategies.

Maillard Odalric-Ambrym[1]

1 - Inria Lille - Nord Europe (France)

Abstract: When considering Markov Decision Problem (MDP) from a learning perspective, the transitions and reward models are unknown from the agent that must interact with the environment in order to receive observations (samples). We are interested in the problem of learning while controling an MDP, that is when the learner interacts in a single-trajectory of state-action-rewards and the learner competes with an omnisicent oracle perfeclty knowing the MDP and how to optimally control it. More precisely, we consider the problem of regret minimization in a discrete, undiscounted, infinite-horizon Markov Decision Problem (MDP) under the average reward criterion, and focus on the minimization of the regret with respect to an optimal policy, when the learner does not know the rewards nor the transitions of the MDP. Characterizing the optimal regret achievable in such MDPs is not trivial as it involves a complextity terms written as an intricate optimization problem. In certain cases, this problem simplifies and can exploited to guide the construction of a sound, efficient learning strategy. In this talk, we will shed a novel light on regret minimization strategies, by extending to reinforcement learning the computationally appealing Indexed Minimum Empirical Divergence (IMED) bandit algorithm. We will then discuss how this fruitful idea can be modified to make computations of the index fully iterative, hence reducing computational complexity while retaining sound guarantees. We will illustrate this promising strategy and how discuss how they may permeate to revisit gradient descent approaches for learning in larger MDPs.

Pham Huyên: Policy gradient learning methods for stochastic control with exit time and applications to share repurchase pricing

Hamdouche Mohamed[1], Henry-Labordere Pierre[2], Pham Huyên[1]

1 - Université Paris Cité (France)

2 - Qube RT (France)

Abstract: We develop policy gradients methods for stochastic control with exit time in a model-free setting. We propose two types of algorithms for learning either directly the optimal policy or by learning alternately the value function (critic) and the optimal control (actor). The use of randomized policies is crucial for overcoming notably the issue related to the exit time in the gradient computation. We demonstrate the effectiveness of our approach by implementing our numerical schemes in the application to the problem of share repurchase pricing. Our results show that the proposed policy gradient methods outperform PDE or other neural networks techniques in a model-based setting. Furthermore, our algorithms are flexible enough to incorporate realistic market conditions like e.g. price impact or transaction costs.

Trottner Lukas: Learning to reflect - Data-driven solutions to singular control problems

Trottner Lukas[1]

1 - Aarhus University (Denmark)

Abstract: Even though theoretical solutions to stochastic optimal control problems are well understood in many scenarios, their practicability suffers from the assumption of known dynamics of the underlying stochastic process, raising the statistical challenge of developing purely data-driven strategies. For the classes of continuous diffusion processes and Lévy processes, we show that developing efficient strategies for related singular stochastic control problems can essentially be reduced to finding rate-optimal estimators with respect to the sup-norm risk of objects associated to the invariant distribution of ergodic processes which determine the theoretical solution of the control problem. From a statistical perspective, we exploit exponential mixing properties as the common factor of both scenarios to drive the convergence analysis, indicating that relying on general stability properties of Markov processes is a sufficiently powerful and flexible approach to treat complex applications requiring statistical methods.

3.7.6 Invited session F1-6 (Jonckheere): Probabilistic tools for learning (Room 105)

Friday 30th, 10:30 - 12:00 - chaired by Matthieu Jonckheere

Ferragut Andres: Spatial estimation of EV energy demand based on aggregated measurements

Ferragut Andres[1], Espindola Emiliano[2]

1 - Universidad ORT Uruguay (Uruguay)
2 - Universidad ORT Uruguay (Uruguay)

Abstract: Electrical vehicle (EV) adoption is currently growing exponentially, and may even speed up once adoption prices become competitive against carbon based alternatives. However, EV charging is a new energy and power intensive application for the electrical grid. Preparing and deploying the necessary charging infrastructure is thus a challenging task, particularly in the grid last mile. In our work, we develop techniques to spatially estimate the amount of "miles" or "kilometers" that are currently required from vehicles based on the total amount of gas sales in a given region. Our goal is to provide a suitable geographic estimate of the energy density that needs to be fulfilled, in order to inform future investment decisions for grid operators. Mathematically, the main challenge is to estimate a distributed density function when measurements are only taken at certain points, e.g. gas sales at given gas stations. These measurements correspond to integrals of the underlying density over a space partition. Given this input, we propose two parametric models based on Gaussian Radial Basis Functions: one based on minimizing mean square error and a further approach using maximum likelihood estimation, and discuss how to build the estimators based on a gradient descent approach. We test the results on real data from California.

Jonckheere Matthieu: Metric Learning : Choosing the parameter of the Fermat distance

Ferraris Laure[1], Chazal Frédéric[1], Groisman Pablo[2], Jonckheere Matthieu[3], Pascal Frederic[4], Sapienza Facundo[5]

1 - Inria (France)
2 - Universidad de Buenos Aires (Argentina)
3 - CNRS (France)
4 - CentraleSupélec (France)
5 - UC Berkeley (United States)

Abstract: In this talk, I will present a "measure-dependent metric learning" problem. Inspired from optics, the Fermat distances have been recently established as a powerful tool for machine learning tasks. Given a non-vanishing density function f on a domain or a submanifold in R^D the Fermat distances define the length of a shortest path between two points. The optimal path remains in the support of f and travels faster into high density regions, where this latter measure-dependent effect is calibrated by a parameter α. Hence, the family of Fermat distances is able to capture both the geometry of the support and the underlying density. Considering the estimator of the Fermat distance inferred from a dataset, I will focus on the choice of the parameter α. This problem turns out to be crucial for the performance of the subsequent tasks. In the context of applications to clustering, i will show by a combination of theoretical and heuristic arguments that there is a window of useful α, while a well chosen parameter scales with the dimension of the data. This work is a collaboration with Frédéric Chazal, Pablo Groisman, Matthieu Jonckheere, Frédéric Pascal and Facundo Sapienza.

Mastropietro Daniel: Fleming-Viot particle systems as a mechanism to improve learning from rare events

Mastropietro Daniel[1,2], Ayesta Urtzi[2,3,4,5], Jonckheere Matthieu[5,6], Majewski Szymon[7]

1 - Institut National Polytechnique (Toulouse) (France)
2 - Institut de recherche en informatique de Toulouse (France)
3 - Ikerbasque - Basque Foundation for Science (Spain)
4 - University of the Basque Country/Euskal Herriko Unibertsitatea (Spain)
5 - Laboratoire dánalyse et dárchitecture des systèmes (France)
6 - Institut de recherche en informatique de Toulouse (France)
7 - Département de Mathématiques de l'École polytechnique (France)

Abstract: In this talk we introduce Fleming-Viot particle systems as a mechanism to improve exploration and learning from rare events in contexts where importance sampling is not an option. Queueing systems serve as illustration, where the objective is to learn the optimum acceptance policy of incoming jobs of different classes in terms of lost opportunity costs due to rejection. Policy gradient methods are used

to learn the optimum integer-valued thresholds that are a function of the queue occupancy. Learning happens by observing the cost of rejecting incoming jobs of each of all possible classes which, depending on the thresholds, can occur very rarely. Fleming-Viot particle systems leverage information about the structure of the queueing system in order to boost the observation of these rejection events. This gives rise to the FVRL algorithm that combines Fleming-Viot with Reinforcement Learning to efficiently learn the optimum thresholds. A Monte-Carlo-based reinforcement learning method is used as benchmark for comparison.

Keppo Jussi: Diversified Learning: Bayesian Control with Multiple Biased Information Sources

Keppo Jussi[1,2]

1 - Xinyuan Zhang (Canada)
2 - Michael Jong Kim (Canada)

Abstract: We consider a decision-maker (DM) who can sample from multiple information sources to learn the underlying state before making an earning decision. The DM optimizes her sampling and earning decisions to maximize her payoffs. The problem is motivated by financial and healthcare applications with multiple information sources. We develop a Bayesian control framework for such problem and an efficient algorithm to solve it. Furthermore, we solve the problem explicitly in estimation and testing settings.

3.7.7 Invited session F1-7 (Hachem): Large random matrices and statistical learning (Room 106)

Friday 30th, 10:30 - 12:00 - chaired by Walid Hachem

Fan Zhou: Random linear estimation with rotationally-invariant designs: Asymptotics at high temperature

Fan Zhou[1],Yufan Li, Subhabrata Sen, Yihong Wu

1 - Yale University (United States)

Abstract: We study estimation in the linear model, in a Bayesian setting where the signal has an entrywise i.i.d. prior and the design is rotationally-invariant in law. In the large system limit as dimension and sample size increase proportionally, a set of related conjectures have been postulated for the asymptotic mutual information, Bayes-optimal mean squared error, and TAP mean-field equations that characterize the Bayes posterior mean of the true signal. In this work, we prove these conjectures for a general class of signal priors and for arbitrary rotationally-invariant designs, under a "high-temperature" condition that restricts the range of eigenvalues of the design. Our proof uses a conditional second-moment method argument, where we condition on the iterates of a version of the Vector AMP algorithm for solving the TAP mean-field equations.

This is joint work with Yufan Li, Subhabrata Sen, and Yihong Wu

Loubaton Philippe: Testing that a large number of time series are independent using linear spectral statistics of the estimated spectral coherence matrix

Loubaton Philippe[1]

1 - Université Paris-Est-Marne La Vallée (France)

Abstract: We address the problem of testing that M jointly stationary complex gaussian times series $(y_{1,n})_{n\in\mathbb{Z}},\ldots,(y_{M,n})_{n\in\mathbb{Z}}$ are independent when the number of time series M is large and that the number of available observations N is not unlimited. We model this general context by the asymptotic regime where $M=M(N)$ verifies $C_1N^\alpha\leq M\leq C_2N^\alpha$ where $C_1>0$ and $C_2>0$ are two constants and where $\alpha<1$. In order to propose relevant test statistics, we consider the frequency smoothed estimated spectral coherence matrix $\hat{\mathbf{C}}_N(\nu)$ defined at each frequency ν by $\hat{\mathbf{C}}_N(\nu)=\left(\mathrm{Diag}(\hat{\mathbf{S}}_N(\nu))\right)^{-1/2}\hat{\mathbf{S}}_N(\nu)\left(\mathrm{Diag}(\hat{\mathbf{S}}_N(\nu))\right)^{-1/2}$ where $\hat{\mathbf{S}}_N(\nu)$ is the frequency smoothed estimated spectral density matrix of the M-variate time series $(\mathbf{y}_n=(y_{1,n},\ldots,y_{M,n})^T)_{n\in\mathbb{Z}}$ defined by $\hat{\mathbf{S}}_N(\nu)=\frac{1}{B+1}\sum_{b=-B/2}^{B/2}\boldsymbol{\xi}_{\mathbf{y}}(\nu+b/N)\boldsymbol{\xi}_{\mathbf{y}}(\nu+b/N)^*$ with $\boldsymbol{\xi}_{\mathbf{y}}(\nu)=\frac{1}{\sqrt{N}}\sum_{n=1}^N\mathbf{y}_ne^{-2i\pi(n-1)\nu}$. Assuming that the smoothing span $B=B(N)$ verifies $c_N=\frac{M}{B}\to c$, $0<c<1$ when $N\to+\infty$, the properties of $\hat{\mathbf{C}}_N(\nu)$ under the hypothesis $\mathcal{H}_0$ that the M time series are independent were studied in [1]. It was in particular shown that the empirical eigenvalue

distribution of $\hat{\mathbf{C}}_N(\nu)$ converges almost surely towards the Marcenko-Pastur distribution μ_c with parameter c. Therefore, if ϕ is a smooth function, the linear spectral statistics $\hat{\phi}_N(\nu)$ of $\hat{\mathbf{C}}_N(\nu)$ defined by $\hat{\phi}_N(\nu) = \frac{1}{M}\sum_{m=1}^{M}\lambda_m(\hat{\mathbf{C}}_N(\nu)) = \frac{1}{M}\mathrm{Tr}\,\phi(\hat{\mathbf{C}}_N(\nu))$ verifies $\hat{\phi}_N(\nu) - \int \phi(\lambda)d\mu_{c_N}(\lambda) \to 0$. [1] also studied the rate of convergence of $\hat{\phi}_N(\nu) - \int \phi(\lambda)d\mu_{c_N}(\lambda)$. In this presentation, we take benefit of the results of [1] to establish that for a well chosen frequency grid $\mathcal{E}_N$ included into $\{\frac{k}{N}, k = 0, \ldots, N-1\}$, then, under the hypothesis $\mathcal{H}_0$, the integrated statistics $\zeta_N = \sum_{\nu\in\mathcal{E}_N}\left(\hat{\phi}_N(\nu) - \int \phi(\lambda)d\mu_{c_N}(\lambda)\right)$ converges, after convenient recentering and renormalization, towards a standard gaussian random variable. This allows to evaluate the asymptotic type I error of the corresponding statistical test. [1] P. Loubaton, A. Rosuel, "Properties of linear spectral statistics of frequency-smoothed estimated spectral coherence matrix of high-dimensional Gaussian time series", Electronic J. of Statistics, vol. 15, no. 2, pp. 5380-5454, 2021. Joint work with Alexis Rosuel and Pascal Vallet.

Maillard Antoine: Some advances on extensive-rank matrix denoising and factorization

Maillard Antoine[1]

1 - ETH Zurich (Switzerland)

Abstract: Factorization of matrices where the rank of the two factors diverges linearly with their sizes has many applications in diverse areas such as unsupervised representation learning, dictionary learning or sparse coding. We consider a setting where the two factors are generated from known component-wise independent prior distributions, and the statistician observes a (possibly noisy) component-wise function of their matrix product. In the limit where the dimensions of the matrices tend to infinity, but their ratios remain fixed, we expect to be able to derive closed form expressions for the optimal mean squared error on the estimation of the two factors. However, this remains a very involved mathematical and algorithmic problem. A related, but simpler, problem is extensive-rank matrix denoising, where one aims to reconstruct a matrix with extensive but usually small rank from noisy measurements. In this talk we will start by approaching the latter problem, and relate previously-known rotationally-invariant estimators to Bayes-optimal estimation, and show how exact asymptotic calculations of the minimal error can be performed using extensive-rank Harish-Chandra-Itzykson-Zuber integrals. We will then highlight the additional difficulties posed by the matrix factorization setting, and will layout a systematic way to derive the corrections to existing approximations, using high-temperature expansions at fixed order parameters which originated in statistical physics.

Stephan Ludovic: Sparse random hypergraphs: Non-backtracking spectra and community detection

Stephan Ludovic[1], Zhu Yizhe[2]

1 - Ecole Polytechnique Fédérale de Lausanne (Switzerland)

2 - University of California [Irvine] (United States)

Abstract: We consider the community detection problem in a sparse q-uniform hypergraph G, assuming that G is generated according to the Hypergraph Stochastic Block Model (HSBM). We prove that a spectral method based on the non-backtracking operator for hypergraphs works with high probability down to the generalized Kesten-Stigum detection threshold conjectured by Angelini et al. (2015). We characterize the spectrum of the non-backtracking operator for the sparse HSBM and provide an efficient dimension reduction procedure using the Ihara-Bass formula for hypergraphs. As a result, community detection for the sparse HSBM on n vertices can be reduced to an eigenvector problem of a $2n \times 2n$ non-normal matrix constructed from the adjacency matrix and the degree matrix of the hypergraph. To the best of our knowledge, this is the first provable and efficient spectral algorithm that achieves the conjectured threshold for HSBMs with r blocks generated according to a general symmetric probability tensor.

3.7.8 Invited session F1-8 (Perchet): Online and dynamic matchings (Room 202)

Friday 30th, 10:30 - 12:00 - chaired by Vianney Perchet

Aouad Ali: A Nonparametric Framework for Online Stochastic Matching with Correlated Arrivals

Aouad Ali[1], Ma Will[2]

1 - London Business School (United Kingdom)

2 - Columbia Business School (United States)

Abstract: The design of online policies for stochastic matching and revenue management settings is usually bound by the Bayesian prior that the demand process is formed by a fixed-length sequence of queries with unknown types, each drawn independently. This assumption of serial independence implies that the demand of each type, i.e., the number of queries of a given type, has low variance and is approximately Poisson-distributed. Thus, matching policies are often based on "fluid" linear programs that only use the expectations of these distributions. This paper explores alternative stochastic models for online matching that allow for nonparametric, higher variance demand distributions. We propose two new models, INDEP and CORREL, that relax the serial independence assumption in different ways by combining a nonparametric distribution for the demand with standard assumptions on the arrival patterns – adversarial or random-order. In our INDEP model, the demand for each type follows an arbitrary distribution, while being mutually independent across different types. In our CORREL model, the total demand follows an arbitrary distribution, and conditional on the sequence length, the type of each query is drawn independently. In both settings, we show that the fluid LP relaxation based on only expected demands can be an arbitrarily bad benchmark for algorithm design. We develop tighter LP relaxations for the INDEP and CORREL models that leverage the exact distribution of the demand, leading to matching algorithms that achieve constant-factor performance guarantees under adversarial and random-order arrivals. More broadly, our paper provides a data-driven framework for expressing demand uncertainty (i.e., variance and correlations) in online stochastic matching models.

Lykouris Thodoris: Efficient Decentralized Multi-Agent Learning in Asymmetric Bipartite Queueing Systems

Lykouris Thodoris[1]

1 - Massachusetts Institute of Technology (United States)

Abstract: Learning in multi-agent systems often poses significant challenges due to interference between agents. In particular, unlike classical stochastic systems, the performance of an agent's action is not drawn i.i.d. from some distribution but is directly affected by the (unobserved) actions of the other agents. This is the reason why most collaborative multi-agent learning approaches aim to globally coordinate all agents' actions to evade this interference. In this talk, we focus on agents in a decentralized bipartite queueing system, where N agents request service from K servers. Prior decentralized approaches aim to decentrally identify a globally coordinated schedule, which leads to various shortcomings: they are restricted to symmetric systems, have performance that degrades exponentially in the number of servers, require communication through shared randomness, two-way message passing or unique identifiers, or are computationally demanding. In contrast, we provide a low-complexity algorithm that is run decentrally by each agent, avoids the shortcomings of "global coordination", and leads the queueing system to have efficient performance in asymmetric bipartite queuing systems while also having additional robustness properties. Paper information: The paper on which this talk is based is joint work with Daniel Freund and Wentao Weng and can be found here: https://arxiv.org/abs/2206.03324. The paper is currently under major revision in Operations Research, a preliminary version appeared at COLT'22, and Wentao was selected as a finalist in the Applied Probability Society student paper competition for this work.

Sentenac Flore: Online Matching in Bipartite Random Graphs

Sentenac Flore[1]

1 - CREST (France)

Abstract: Motivated by sequential budgeted allocation problems, we investigate online matching problems on classes of random graphs. In particular, we look at the so-called configuration model and geometric random graphs. We estimate the competitive ratio of the simplest algorithm, GREEDY, by approximating some relevant stochastic discrete processes by their continuous counterparts, that are solutions of an explicit system of partial differential equations. This technique gives precise bounds on the estimation errors, with arbitrarily high probability as the problem size increases. In particular, it allows the formal comparison between different configuration models. We also prove that, quite surprisingly, GREEDY can have better performance guarantees than RANKING, another celebrated algorithm for online matching that usually outperforms the former.

Tercieux Olivier: Unpaired Kidney Exchange: Overcoming Double Coincidence of Wants without Money

Tercieux Olivier[1], Akbarpour Mohammad, Combe Julien, He Yinghua, Hiller Victor, Shimer Rob

1 - PARIS SCHOOL OF ECONOMICS (France)

Abstract: For an incompatible patient-donor pair, kidney exchanges often forbid receipt-before-donation (the patient receives a kidney before the donor donates) and donation-before-receipt, causing a double-coincidence-of-wants problem. We study an algorithm, the Unpaired kidney exchange algorithm, which eliminates this problem. In a dynamic matching model, we show that waiting time of patients under the Unpaired is close to optimal and substantially shorter than widely used algorithms. Using a rich administrative dataset from France, we show that Unpaired achieves a match rate of 63 percent and an average waiting time of 176 days for transplanted patients. The (infeasible) optimal algorithm is only slightly better (64 percent and 144 days); widely used algorithms deliver less than 40 percent and at least 232 days. We discuss a range of solutions that can address the potential practical incentive challenges of the Unpaired. In particular, we extend our analysis to an environment where a deceased donor waitlist can be integrated to improve the performance of algorithms. We show that our theoretical and empirical comparisons continue to hold. Finally, based on these analyses, we propose a practical version of the Unpaired algorithm.

3.7.9 Invited session F1-9 (Kazi-Tani): Recent advances in actuarial science (Room 203)

Friday 30th, 10:30 - 12:00 - chaired by Nabil Kazi-Tani

Ecoto Geoffrey: Forecasting the cost of drought events in France by super learning

Ecoto Geoffrey[1,2]

1 - Mathématiques Appliquées Paris 5 (France)
2 - Caisse Centrale de Réassurance (France)

Abstract: Recent years in France have been marked by the occurrence of intense drought events, in line with the conclusions of numerous studies analyzing the impact of climate change on natural disasters in terms of frequency and intensity. Confronted with the increased exposure of the insurance and reinsurance market, the actuarial community needs to assess as acurately as possible the cost of these events. In this talk, we present a new methodology to estimate the damage due to a geotechnical drought event in France. We develop a dedicated sequential super-learning algorithm, which is adapted to the spatio-temporal dependency structure at play in the data. A theoretical analysis of the procedure provides interesting performance guarantees. The exploitation of the algorithm on the data available to CCR yields very relevant results.

Goffard Pierre-Olivier: Stochastic models for blockchain analysis

Goffard Pierre-Olivier[1]

1 - Institut de Recherche en Mathématiques Avancées (France)

Abstract: This talk focuses on the consensus protocol at the heart of blockchain systems. Simple stochastic models are introduced to measure the efficiency, decentralization and security of a blockchain systems. The model for efficiency is a single server queue with bulk service, the model for centralization is a simple Polya's urn and the model for security is analysed using standard insurance risk theory tools.

Weber Stefan: Modeling and Pricing Cyber Insurance

Weber Stefan[1]

1 - House of Insurance, Leibniz Universität Hannover (Germany)

Abstract: The talk will discuss recent developments and current challenges in cyber risk and insurance.

Kazi-Tani Nabil: Stochastic differential Colonel Blotto games on graphs - Applications to cyber security

Hazzami Mohamed Amine[1], Kazi-Tani Nabil[1], Varma Vineeth Satheeskumar[2]

1 - Institut Élie Cartan de Lorraine (France)
2 - Centre de Recherche en Automatique de Nancy (France)

Abstract: We extend standard Colonel Blotto games to a dynamic setting, using the framework of continuous-time two-player zero-sum stochastic differential games. The game is played on a graph, where each edge represents a battlefield, and where players control the drift of their output process on each battlefield, subject to a budget constraint. Using a dynamic programming approach, we explicitly

characterize the Nash equilibrium strategies and the value function of the game, in terms of a Hamilton-Jacobi-Bellman PDE that admits a smooth solution. Our formulation is general enough to include various rewards, allowing different interpretations of the game. In particular, we allow reward functionals with interactions betwen the battlefields: for example, player 1, interpreted as a defender, maximizes the expected algebraic connectivity associated to the subgraph given by the edges that she won, while player 2, the attacker, minimizes the same criteria. This is a joint work with Mohamed Amine Hazzami (IECL, Univ. Lorraine) and Vineeth S Varma (CRAN, CNRS).

3.7.10 Invited session F1-10 (Murthy): Advances in stochastic modelling for decision-making under uncertainty (Room 204)

Friday 30th, 10:30 - 12:00 - chaired by Karthyek Murthy

Feng Yifan: Nested Elimination: A Simple Algorithm for Best-Item Identification from Choice-Based Feedback

Yang Junwen[1], Feng Yifan[1]

1 - National University of Singapore (Singapore)

Abstract: We study the problem of best-item identification from choice-based feedback. In this problem, a company sequentially and adaptively shows display sets to a population of customers and collects their choices. The objective is to identify the most preferred item with the least number of samples and at a high confidence level. We propose an elimination-based algorithm, namely Nested Elimination (NE), which is inspired by the nested structure implied by the information-theoretic lower bound. NE is simple in structure, easy to implement, and has a strong theoretical guarantee for sample complexity. Specifically, NE utilizes an innovative elimination criterion and circumvents the need to solve complex combinatorial optimization problems. We provide an instance-specific and non-asymptotic bound on the expected sample complexity of NE. We also show NE achieves high-order worst-case asymptotic optimality. Finally, numerical experiments from both synthetic and real data corroborate our theoretical findings.

Deo Anand: Risk Averse Decision Making Using Tail Self-Similarity

Deo Anand[1], Murthy Karthyek[2]

1 - Indian Institute of Management Bangalore (India)
2 - Singapore University of Technology and Design (Singapore)

Abstract: In this talk, we attempt to address impractically large sample requirements in optimisation formulations involving tail risk measures such as Conditional Value at Risk. Based on importance sampling, we propose a novel scheme for risk averse optimisation assuming access to the evaluations of the objective loss, and the distribution of the underlying covariates. We demonstrate that our algorithm requires exponentially fewer samples that plain sample averages as the problem is made harder by increasing the level of risk aversion to meet a particular accuracy requirement in the optimal objective value. The algorithm itself is derived using notion of self-similarity in distribution tails - rare events happen in structurally similar (and most likely) ways. We leverage this observation to infer the distribution of samples in the target rare region by appropriately extrapolating representative but less rare samples observed more frequently on data. Our algorithm is an improvement over the state of the art methods, most of which either Fail to give any guarantees on the amount of sample complexity reduction obtained or Have to be carefully tuned to the underlying loss and covariate distribution. Numerical experiments validate our theoretical claims and showcase practical utility of the proposed methods.

Das Bikramjit: Aggregating heavy-tails: tracking large jumps from random sums to Lévy processes

Das Bikramjit[1], Fasen-Hartmann Vicky

1 - Singapore University of Technology and Design (Singapore)

Abstract: The tail behavior of aggregates of heavy-tailed random vectors is known to be determined by the so-called principle of "one large jump". We establish that, in fact, a more general principle is at play. Assuming that the random vectors are multivariate regularly varying (heavy-tailed) on various subcones of the positive quadrant, first we show that their aggregates are also regularly varying (heavy-tailed) on these subcones. This allows us to approximate certain tail probabilities which were rendered asymptotically negligible under classical regular variation, despite the "one large jump" asymptotics. We

also discover that depending on the structure of the tail event of concern, the tail behavior of the aggregates may be characterized by more than a single large jump. Insurance risk processes have often been represented using Levy processes. We establish a relationship between multivariate regularly varying Levy processes and multivariate regular variation of its Levy measure on different subcones; this which simplifies computing certain ruin probabilities. We illustrate our results using simulated data. The talk is based on a joint work with Vicky Fasen-Hartmann.

3.8 Friday 30th, 13:30 - 15:00

3.8.1 Invited session F2-2 (Ravner): Statistical inference for queueing systems (Room 101)

Friday 30th, 13:30 - 15:00 - chaired by Liron Ravner

Bodas Shreehari: Parametric inference of the periodic arrival rate to a queue with unobserved balking

Bodas Shreehari[1], Mandjes Michel[1], Ravner Liron[2]

1 - University of Amsterdam [Amsterdam] (Netherlands)

2 - University of Haifa [Haifa] (Israel)

Abstract: Consider a multi-server queueing system in which arriving customers are provided information about the delay they will experience. Based on this information, they decide to wait for service or leave the system. While previous work considered the case of a constant arrival rate of potential customers, here we assume that they arrive according to a time-dependent periodic arrival rate. Each customer has a patience level sampled from a distribution. The objective is to estimate the arrival rate and patience distribution parameters using only knowledge of the actual queue-length process over time, i.e., customers who decide not to join are not observed. We assume that the structure of the arrival rate and the patience distribution are known beforehand, making this a parametric inference problem. We devise an MLE procedure to estimate the parameters. For certain parametric assumptions, consistency and asymptotic normality of the estimator is proved by breaking down the observed queue-length process into distinct regeneration cycles to invoke i.i.d structure.

Bodas Tejas: Some recent results on the convert queueing problem

Bodas Tejas[1], Yardi Arti[1]

1 - IIIT Hyderabad (India)

Abstract: In the covert queueing problem, we consider a setup where server Bob is in agreement with customer Willie to serve his jobs exclusively and yet would like to admit non-Willie jobs covertly. Seeing that the server is occasionally idle, Bob is tempted to allow a stream of non-Willie jobs to generate some extra revenue. However, Bob must do so covertly without Willie being able to detect the presence of outside traffic. As a part of the contract, Bob is typically obliged to provide Willie with some information or statistic on the utilization of the server. Examples of such statistic could be the arrival and departure times of Willie jobs, length of each busy session, the number of successive Willie jobs served in each session etc. Given the information provided by Bob, Willie can perform optimal hypothesis testing to infer whether Bob is serving non-Willie stream of jobs or not. Thus Bob aims to insert stream of non-Willie jobs covertly such that Willie is not able to detect the presence of non-Willie traffic with high probability, leading to the term covert queueing. In this talk, we will look at some recent results on the covert queueing problem that consider both i.i.d and non-iid statistic and discuss some interesting future directions.

Wichelhaus Cornelia: Nonparametric approaches for unreliable queueing systems

Wichelhaus Cornelia[1]

1 - Technical University of Darmstadt (Germany)

Abstract: In this talk we will present statistical approaches for unreliable queueing systems with servers which are unreliable, i.e. the servers interrupt the service process from time to time and take a random vacation period. Based only on observations of the arrival and departure epochs of the customers (without the possibility of their correct matching) we obtain information about the unknown repair time

distributions and the breakdown rates of the servers. Explicit constructions of the estimators are presented as well as their analytic performance. Decompounding techniques for random sums play a major role in the proofs. By simulation results we illustrate the practicability of our analysis.

Ravner Liron: Nonparametric estimation of the M/G/1 job-size distribution from discrete workload observations

Ravner Liron[1]

1 - University of Haifa [Haifa] (Israel)

Abstract: Consider an M/G/1 queue with an unknown job-size distribution. The workload process is observed according to an independent Poisson sampling process. A non-parametric estimator for the cdf of the job-size distribution is constructed by first estimating the characteristic function and then applying an inversion formula. The convergence rate of the asymptotic risk is shown to depend on the smoothness class of the underlying job-size distribution.

3.8.2 Invited session F2-3 (Hu): Queueing theory and its applications (Room 102)

Friday 30th, 13:30 - 15:00 - chaired by Yue Hu

Ata Baris: Dynamic Control of MultiClass Queue in Halfin-Whitt Regime: A Computational Approach for High-Dimensional problems

Ata Baris[1], Harrison J. Michael[2], Kasikaralar Ebru[3]

1 - University of Chicago Booth School of Business (United States)
2 - Stanford University, Graduate School of Business (United States)
3 - Universiity of Chicago, Booth School of Business (United States)

Abstract: We consider a dynamic control problem motivated by call centers and study it in the Halfin-Whitt regime. We propose an effective method for solving the resulting diffusion control problem in high dimensions. We illustrate the effectiveness of the solution in several test problems, including one designed using a call center data set.

Castellanos Antonio: An Infinite Server Queueing Model for Incarceration Diversion Decisions

Castellanos Antonio[1], Shi Pengyi[2], Ward Amy R.[1]

1 - University of Chicago (United States)
2 - Purdue University (United States)

Abstract: Community-based incarceration diversion programs have been shown to be an effective strategy to reduce recidivism (when formerly incarcerated individuals commit an offence and are re-incarcerated). These programs have garnered much interest due to the high rate of recidivism in the United States, where 58% of individuals released on parole were rearrested for a new offense within a two-year period. In collaboration with a government agency in Illinois we develop a data driven framework for modeling incarceration diversion decisions, with an eye towards using our model to provide decision support. Specifically, we model the system with a multi-class infinite server queue model. In our model arrivals correspond to program referrals, service times to length of stay in the programs, and in-service reneging to when a person recidivates. We address important questions, such as: How should the population be split into distinct subgroups of interest? Is a Poisson arrival process consistent with the data, and, if so, should the rate be time-stationary or time varying? What distribution best models service times? We end by discussing how to use our framework to explore the tradeoffs among admission policy, program utilization, and recidivism.

Hu Yue: Optimal Scheduling of Proactive Service with Customer Deterioration and Improvement

Hu Yue[1], Chan Carri[2], Dong Jing[2]

1 - The University of Chicago (United States)
2 - Columbia University (United States)

Abstract: Service systems are typically limited resource environments where scarce capacity is reserved for the most urgent customers. However, there has been a growing interest in the use of proactive service when a less urgent customer may become urgent while waiting. On one hand, providing service for customers when they are less urgent could mean that fewer resources are needed to fulfill their service requirement. On the other hand, utilizing limited capacity for customers who may never need the service in the future takes the capacity away from other more urgent customers who need it now. To understand this tension, we propose a multi-server queueing model with two customer classes: moderate and urgent. We allow customers to transition classes while waiting. In this setting, we characterize how moderate and urgent customers should be prioritized for service when proactive service for moderate customers is an option. We identify an index, the modified $c\mu/\theta$-index, which plays an important role in determining the optimal scheduling policy. This index lends itself to an intuitive interpretation of how to balance holding costs, service times, abandonments, and transitions between customer classes.

Daw Andrew: Asymmetries of Service: Interdependence and Synchronicity

Daw Andrew[1], Yom-Tov Galit[2]

1 - University of Southern California (United States)

2 - Technion - Israel Institute of Technology [Haifa] (Israel)

Abstract: We propose and analyze a Hawkes cluster model of service interactions that captures two (a)symmetries between the customer and agent: co-production vs self-production, synchrony vs asynchrony. We analyze the stochastic process through a novel decomposition in terms of objects from probabilistic combinatorics. This model reveals connection to the behavioral operations literature, such as non-monotonic system performance from monotonic agent-load slowdown, which we prove will exist under customer-agent asynchrony. Given this existence, we can prescribe an optimal concurrency, or level of multi-tasking, for the service agents. However, the efficiency of such operational improvements is limited by symmetry in interdependence: the more co-productive the service, the less opportunity there is to improve.

3.8.3 Invited session F2-5 (Bhulai): Markov decision problems (Room 104)

Friday 30th, 13:30 - 15:00 - chaired by Sandjai Bhulai

Koopmans Camiel: Optimal service control and observation restoration decisions in Markovian queues with fallible sensors.

Kanavetas Odysseas[1], Koopmans Camiel[1], Spieksma Floske[1]

1 - Mathematical Institute, Leiden University (Netherlands)

Abstract: Markov decision processes (MDPs) have been a source of many techniques and models frequently used nowadays. This talk aims to show how to apply techniques stemming from MDPs to Markovian single-queue models where service rate control is performed with fallible sensors. More specifically, "fallible"means that after a sensor failure the queue cannot be observed until it is repaired. This differs from noisy partially-observable MDPs. In practice, such a model is applicable to situations where sensors break down and their repairs have to be optimized. We search for policies that minimize costs by specifying the service rate to use and by scheduling the repair of the sensor. One of the approaches is based on the fact that cer- tain mathematical properties (e.g., monotonicity and convexity) are preserved under Value Iteration, which allows us to establish certain (threshold) structures in optimal policies.

Singh Vikas Vikram: Joint chance-constrained Markov decision Processes

Singh Vikas Vikram[1], Venktesan Varagapriya[1], Lisser Abdel[2]

1 - Indian Institute of Technogy Delhi (India)

2 - Université Paris-Saclay (France)

Abstract: We consider a finite state-action uncertain constrained Markov decision process under discounted and average cost criteria. The running costs are defined by random variables and the transition probabilities are known. The uncertainties present in the objective function and the constraints are modelled using chance constraints. We assume that the random cost vectors follow multivariate elliptically symmetric distributions and dependence among the random constraints is driven by a Gumbel-Hougaard copula. We propose two second order cone programming problems whose optimal values give lower and upper bounds of the optimal value of the uncertain constrained Markov decision process. As an application, we study a stochastic version of a service and admission control problem in a queueing system. The proposed approximation methods are illustrated on randomly generated instances of queueing control problem.

Spieksma Floske: The deviation matrix of a denumerable, multichain Markov process in continuous time

Spieksma Floske[1], Avrachenkov Kostya[2]

1 - Universiteit Leiden (Netherlands)
2 - Inria Sophia Antipolis (France)

Abstract: The deviation matrix of a denumerable, irreducible non-explosive Markov process is well-understood. However, to our knowledge no results so far exist, when the irreducibility requirement is dropped. In this talk, we will present necessary and sufficient conditions for the deviation matrix to exist in the case of a potentially infinite number of closed classes, and a non-closed subset of transient states. Under mild conditions, the deviation matrix has an alternative representation based on rank one perturbations of the generator of the Markov process. An analog of this representation has been studied earlier by Kartashov [1] under stronger assumptions for an irreducible, discrete time queueing system. We will illustrate the use of this formula by an application to a restart process. [1] N. V. Kartashov. Strong stable Markov chains. VSP, 1996.

Bhulai Sandjai: Optimal stopping policies when assessing students' mastery

Sapountzi Anni, Storm Jaap, Bhulai Sandjai[1], Meeter Martijn

1 - Vrije Universiteit, Amsterdam (Netherlands)

Abstract: An important challenge in assessments and testing of learning outcomes is estimating whether an individual has attained mastery of a single skill on the basis of sequences of responses. It is not trivial to obtain a reliable diagnosis of mastery with as few students' responses as possible and provide an adaptive length assessment in an online and personalized manner. We propose an approach to solve that problem. Our solution consists of three components: (1) a Markov reward model in which the value function serves as a proxy for a student's mastery level; (2) a reinforcement learning and machine learning technique to evaluate the value function, and (3) stopping policies, based on the value function, that balance the accuracy and the efficiency of the mastery assessment. These policies are intuitive and are easy to use in practice by, e.g., teachers. Finally, we demonstrate the efficacy of these stopping policies using synthetic and historical student response data. The policies enable the efficient assessment of mastery across broad students' mastery levels.

3.8.4 Contributed session F2-6 (Hautphenne): Advanced stochastic models (Room 105)

Friday 30th, 13:30 - 15:00 - chaired by Sophie Hautphenne

Khezeli Ali: An Improved Lower Bound on the Largest Common Subtree of Random Leaf-Labeled Binary Trees

Khezeli Ali[1]

1 - Inria de Paris (France)

Abstract: In some models of random combinatorial structures, an object of interest is the largest common substructures of two samples. A well-studied example is the size of the largest increasing subsequence in a random permutation, which can be regarded as the largest common subset in which the permutation induces the same total order as the identity permutation (or another random permutation). In the context of trees, the largest common subtree has been proposed in biology as a measure of similarity between two phylogenetic trees (also known as the maximum agreement subtree). With this motivations, some works have studied the expected size of the largest common subtree of two independent random trees. In this paper, we consider two random binary trees built on n common labeled leaves. It is known that the expected size of the largest common subtree is of order between $n^{0.366}$ and $n^{1/2}$. We improve the lower bound to order $n^{0.4464}$ by constructing a common subtree recursively and by proving a lower bound for its asymptotic growth. The construction is a modification of an algorithm proposed by D. Aldous by splitting the tree at the centroid and by proceeding recursively.

Ganguly Ankan: Hydrodynamic Limits of Interacting Particle Systems on Sparse Graphs

Ganguly Ankan[1], Ramanan Kavita[2]

1 - University of Pennsylvania (United States)
2 - Brown University (United States)

Abstract: Consider an interacting particle system indexed by the vertices of a (possibly random) locally finite graph. Each particle takes values in a countable state space and evolves according to a pure jump process whose jump intensities depend only on its own state and states of particles in its neighborhood. Under mild conditions on the jump intensities, it is shown that if the sequence of interaction graphs converges locally in probability to a limit graph that satisfies a certain finite dissociability property, then the corresponding sequence of empirical measures of the particle trajectories converges weakly to the law of the marginal dynamics at the root vertex of the limit graph. The proof of this hydrodynamic limit relies on several auxiliary results of potentially independent interest. First, such interacting particle systems are shown to be well-posed on (almost surely) finitely dissociable graphs, which include graphs of maximal bounded degree and any Galton-Watson tree whose offspring distribution has a finite first moment. A counterexample is also provided to show that well-posedness can fail for dynamics on graphs outside this class. Next, the dynamics on a locally convergent sequence of graphs are shown to converge in the local weak sense to the dynamics on the limit graph when the latter is finitely dissociable. Finally, the dynamics are also shown to exhibit an (annealed) asymptotic correlation decay property. This talk is based on joint work with Kavita Ramanan.

Note to the committee: the author believes this talk would pair well with the talk by Juniper Cocomello.

Mertz Laurent: A Piecewise Deterministic Markov Process approach modeling a dry friction problem with noise

Mertz Laurent[1]

1 - City University of Hong Kong [Hong Kong] (Hong Kong SAR China)

Abstract: Understanding and predicting the dynamical properties of systems involving dry friction is a major concern in physics and engineering. It abounds in many mechanical processes, from the sound produced by a violin to the screeching of chalk on a blackboard to human infant crawling dynamics and friction-based locomotion of a multitude of living organisms (snakes, bacteria, scallops..) to the displacement of mechanical structures (building, bridges, nuclear plants, massive industrial infrastructures) under earthquakes and beyond. Surprisingly, even for low-dimensional systems, the modeling of dry friction in the presence of random forcing has not been elucidated. In this work, we propose a piecewise deterministic Markov process approach modeling a system with dry friction including different coefficients for the static and dynamic forces. In this mathematical framework, we derive the corresponding Kolmogorov equations to compute statistical quantities of interest related to the distributions of the static (sticked) and dynamic phases. We show ergodicity and provide a representation formula of the stationary measure using independent identically distributed portions of the trajectory (excursions). We also obtain deterministic characterizations of the Laplace transforms of the probability density functions of the durations of the static and dynamic phases. In particular, the analysis of the power spectral density of the velocity reveals a critical value of the noise correlation time below which the correlations of the dynamic behaviors coincide with those of the white noise limit. The existence of such a critical value was already mentioned in the physics literature [Geffert and Just, Phys. Rev. E. 95 062111 (2017)]. This is joint work with J. Garnier and Z. Lu.

Hautphenne Sophie: Consistent estimation for population-size-dependent branching processes

Braunsteins Peter[1], Hautphenne Sophie[2], Minuesa Carmen[3]

1 - University of New South Wales (Sydney) (Australia)
2 - The University of Melbourne (Australia)
3 - Universidad de Extremadura - University of Extremadura (Spain)

Abstract: In this talk we consider population-size-dependent branching processes, which are stochastic models designed to capture the dynamics of populations living in restricted habitats. We derive several estimators for parameters in these processes and demonstrate that some of these estimators do not satisfy the classical consistency property called C-consistency. This leads us to define a new type of consistency called Q-consistency. We then investigate the circumstances when C-consistent estimators are preferable to Q-consistent estimators. We illustrate our results on the Chatham Island black robin population.

3.8.5 Invited session F2-7 (Leskela): Network inference II (Room 106)

Friday 30th, 13:30 - 15:00 - chaired by Lasse Leskela

Deka Prabhanka: PageRank Nibble on the sparse directed stochastic block model

Banerjee Sayan[1], Deka Prabhanka[1], Olvera-Cravioto Mariana[1]

1 - University of North Carolina, Chapel Hill (United States)

Abstract: We present results on community recovery in a sparse, directed Stochastic Block Model (dSBM) where the identities of a positive proportion of seeds in one community is known. Our method is based on the PageRank Nibble algorithm, which uses personalized PageRank and a simple threshold rule to label vertices in the same community as the seeds. This method leverages information about the community label of some vertices, a situation often encountered in practice. Moreover, since PageRank is a local algorithm that can be computed in a distributed and asynchronous fashion, it is highly efficient even in large graphs. Our estimates on the probability of misclassification are derived from a tractable expression for the limiting PageRank distribution. This distribution is in turn obtained by using a new local weak limit for a very general class of inhomogeneous random digraphs, which allows us to characterize it via a system of distributional fixed point equations.

Lelarge Marc: Graph matching: from fundamental limits to algorithms

Lelarge Marc[1]

1 - inria (France)

Abstract: Alignment of graphs consists in finding a mapping between the nodes of two graphs which preserves most of the edges. For two correlated Erdös-Rényi, we present information-theoretic results in the sparse regime. We then propose an algorithm based on local comparisons and prove theoretical guarantees. Finally, we show the empirical success of learning algorithms.

Sanders Jaron: Spectral norm bounds for, and singular value distributions of, block Markov chains

Sanders Jaron[1], Senen-Cerda Albert[1], Van Werde Alexander[1]

1 - Eindhoven University of Technology [Eindhoven] (Netherlands)

Abstract: A Block Markov Chain (BMC) is a Markov chain whose state space can be partitioned into a finite number of clusters such that the transition probabilities only depend on the clusters. BMCs thus serve as a model for Markov chains with communities. In this talk, I will share with you some of our latest results on the spectrum of random matrices built from the sample path of a BMC. Specifically, for the centered, empirical frequency matrix of a BMC, we have quantified the asymptotic order of the largest singular value [1] and established limiting laws for the singular value distributions [2]. The former result holds even in sparse regimes when the sample path is relatively short. [1] Jaron Sanders, Albert Senen-Cerda (2022). Spectral norm bounds for block Markov chain random matrices. Stochastic Processes and their Applications. https://doi.org/10.1016/j.spa.2022.12.004 [2] Jaron Sanders, Alexander Van Werde (2023). Singular value distribution of dense random matrices with block Markovian dependence. Stochastic Processes and their Applications. https://doi.org/10.1016/j.spa.2023.01.001

Zhang Anderson Ye: Fundamental Limits of Spectral Clustering in Stochastic Block Models

Zhang Anderson Ye[1]

1 - University of Pennsylvania [Philadelphia] (United States)

Abstract: We give a precise characterization of the performance of spectral clustering for community detection under Stochastic Block Models by carrying out sharp statistical analysis. We show spectral clustering has an exponentially small error with matching upper and lower bounds that have the same exponent, including the sharp leading constant. The fundamental limits established for the spectral clustering hold for networks with multiple and imbalanced communities and sparse networks with degrees far smaller than $\log n$. The key to our results is a novel truncated ℓ_2 perturbation analysis for eigenvectors and a new analysis idea of eigenvectors truncation.

3.8.6 Invited session F2-8 (Busic): Stochastic matching and applications (Room 202)

Friday 30th, 13:30 - 15:00 - chaired by Ana Busic

Banerjee Siddhartha: Priorities and Quotas in Online Matching

Banerjee Siddhartha[1], Eichhorn Matthew[1], Kempe David[2]

1 - Cornell University (United States)
2 - USC (United States)

Abstract: In many applications such as rationing medical care and supplies, university admissions, and the assignment of public housing, the decision of who receives an allocation can be justified by various normative criteria (ethical, financial, legal, etc.). Such settings motivate a natural priority-respecting allocation problem: several categories, each with a quota of interchangeable items, wish to allocate the items among a set of agents. Each category has a list of eligible agents and a priority ordering over these agents; agents may be eligible in multiple categories. The goal is to select a valid allocation: one that respects quotas, eligibility, and priorities, and ensures Pareto efficiency. I will describe some recent work of ours, which provides a complete algorithmic characterization of all valid allocations. I will then show how this characterization allows us to understand the strange complexity landscape of such problems, as well as characterize optimal tradeoffs in dynamic priority-respecting allocation. In particular, in settings where allocations must be made to agents arriving online via some stochastic process, we show that while insisting on zero priority violations leads to an $\Omega(T)$ loss in efficiency, one can design allocation policies that ensure that the ex post sum of the efficiency loss and priority violations is $O(1)$.

Comte Céline: Stochastic Dynamic Matching: A Mixed Graph-Theory and Linear-Algebra Approach

Comte Céline[1], Mathieu Fabien[2], Bušić Ana[3,4]

1 - LAAS-CNRS (France)
2 - Laboratory of Information, Network and Communication Sciences (France)
3 - Inria (France)
4 - Département d'informatique de l'École normale supérieure (France)

Abstract: Stochastic dynamic matching problems have recently drawn attention in the stochastic-modeling community due to their numerous applications, ranging from supply-chain management to kidney exchange programs. In this paper, we consider a matching problem in which items of different classes arrive according to independent Poisson processes. Unmatched items are stored in a queue, and compatibility constraints are described by a simple graph on the classes, so that two items can be matched if their classes are neighbors in the graph. We analyze the efficiency of matching policies, not only in terms of system stability, but also in terms of matching rates between different classes. Our results rely on the observation that, under any stable policy, the matching rates satisfy a conservation equation that equates the arrival and departure rates of each item class. This presentation is based on a joint work with Fabien Mathieu (LINCS) and Ana Bušić (Inria and PSL University).

Fourneau Jean-Michel: Performance Paradox of Dynamic Matching under Greedy Policies

Fourneau Jean-Michel[1], Busic Ana, Cadas Arnaud, Doncel Josu

1 - Université de Versailles St Quentin (France)

Abstract: In our previous work [CDFB21], we focus on the First Come First Match policy and we study the performance paradox that arises when we increase flexibility in the matching model. More precisely, we analyze whether adding an edge to a matching graph leads to an improvement of the expectation of the number of items in the system. The main result of [CDFB21] provides sufficient conditions for the existence of the performance paradox: despite a new edge in the matching graph, the expectation of the total number of items increase. It also shows some numerical examples of such a paradox. These conditions are related to the heavy-traffic assumptions in queueing systems. The intuition behind the performance paradox existence result is that the performance paradox occurs when the added edge in the compatibility graph disrupts the draining of a bottleneck. The main contribution of this work consists of generalizing the performance paradox existence analysis of [CDFB21] to a wide range of matching policies and explore the type of matching graph where such paradoxes occur. First we formally define a large set of policies based on the arriving letter and the state of the system which is assumed to be a word. Then we define greedy matching policies. For such a policy, one may have several possible deletions of matched letters associated with a distribution of probability. But only one letter will be deleted. FCMF, Match the longest, Random selection, Priority are examples of such disciplines. We prove several results about the existence of performance paradoxes for greedy disciplines for some family of graphs. More precisely, we prove several results about the transfert of the paradox from one graph to another one, or for one discipline to another one. Most of these results are based on Strong Aggregation of Markov chains and some graph theoretical properties. We also consider several matching models that lie outside of the

conditions of our main theorem. We first consider a matching model with several saturated independent set. Then, we study the performance paradox when the saturated independent set is not that of the maximum cardinality and for a different family of arrival distribution than that of required in our main theorem.

Moyal Pascal: Stability regions of systems with compatibilities, and ubiquitous measures on graphs

Begeot Jocelyn[1], Marcovici Irène[1], Moyal Pascal[1]

1 - Institut Élie Cartan de Lorraine (France)

Abstract: In many service systems, it is necessary to take into account the compatibility constraints between users, or between supplies and demands, and so on. The stability region of such systems can then be seen as a set of measures (representing the arrival flows of the various classes of users or supplies, demands, etc.) on a graph (representing the compatibilities between those classes). In this paper, we show that these 'stabilizing' measures can always be easily constructed as a simple function of a family of weights on the edges of the graph. Second, we show that the latter measures always coincide with invariant measures of random walks on the graph under consideration. Some arguments in the proofs rely on the matching rates of specific stochastic matching models. As a by-product of these arguments we show that, in several cases, the matching rates are independent of the matching policy, that is, the rule for choosing a match between various compatible elements.

3.8.7 Invited session F2-9 (Carassus): Robust finance (Room 203)

Friday 30th, 13:30 - 15:00 - chaired by Laurence Carassus

Denis Laurent: Robust maximization of recursive utility with penalization on the model

Denis Laurent[1]

1 - Laboratoire Manceau de Mathématiques[Le Mans] (France)

Abstract: We show that robust control under volatility uncertainty in a continuous time setup allows for a dynamic description in terms of second order backward differential equations (2BSDE). The relevant penalizing mechanism is modeled by the distance of any two stochastic volatility models. Apart from unique existence of related utility processes, we establish the related Bellman-Equation.

Pergamenchtchikov Serguei: Robust portfolio optimisation for jump-diffusion market models with transaction costs

Pergamenchtchikov Serguei[1], Egorov Sergei[1]

1 - Laboratoire de Mathématiques Raphaël Salem (France)

Abstract: We consider optimal consumption and investment problems under logarithmic utilities for financial markets described by semi-martingales with jumps defined through Lévy processes. For this problem, we show the corresponding verification theorem and construct the optimal strategies in the explicit form. Then, we apply these strategies to markets with transaction costs. Through the Leland - Lepinette approach we construct the strategies based on the finite number of portfolio revisions for which the objective function tends to its optimal value uniformly over the market distributions when the number of portfolio revisions tends to infinity, i.e. the proposed strategies are robust and optimal in the asymptotic setting.

Wiesel Johannes: The Uniform Diversification Strategy Is Optimal for Expected Utility Maximization under High Model Ambiguity

Wiesel Johannes[1], Carassus Laurence

1 - Columbia University (United States)

Abstract: In a one-period model with d risky assets, we investigate an expected utility maximization problem when there is ambiguity on the probability P modeled by a Wasserstein ball of radius k around P. We show that when k goes to infinity, the optimal solutions tend to some (rescaled) uniform diversification strategy.

Carassus Laurence: Quasi-sure essential supremum and applications to finance

Carassus Laurence[1,2]

1 - Resarch Center Pôle Léonard de Vinci (France)

2 - Laboratoire de Mathématiques de Reims (France)

Abstract: We place ourselves in the so-called quasi-sure setting where the unique probability measure prevailing on the state space is replaced by a set of measures which is neither supposed to be dominated, nor supposed to be compact. A notion of essential supremum for a family of real-valued random variables is developed in terms of upper semi-analytic functions. We show that this notion is indeed well-suited for financial problems as super-replication or Absence of Instantaneous Profit (AIP).

3.8.8 Invited session F2-10 (Rhee): Rare Events: Large Deviations, Simulation, and AI (Room 204)

Friday 30th, 13:30 - 15:00 - chaired by Chang-Han Rhee

Murthy Karthyek: Letting the samples speak: A new approach towards efficient Importance Sampling for tail events

Deo Anand[1], Murthy Karthyek[2]

1 - Indian Institute of Management, Bangalore (India)

2 - Singapore University of Technology and Design (Singapore)

Abstract: The ability to estimate and control tail risks, besides being an integral part of quantitative risk management, is central to running operations requiring high service levels and cyber-physical systems with high-reliability specifications. Despite this significance, scalable algorithmic approaches have remained elusive: This is due to the rarity with which relevant risky samples get observed, and the critical role experts play in devising variance reduction techniques based on instance-specific large deviations studies. Our goal in this talk is to examine if such tailored variance reduction benefits can be instead achieved by instance-agnostic algorithms capable of scaling well across multitude of tail estimation and optimisation tasks. To this end, we identify an elementary transformation whose push-forward automatically induces efficient importance sampling distributions across a variety of models by replicating the concentration properties observed in less rare samples. This obviates the need to explicitly identify a good change of measure, thereby overcoming the primary bottleneck in the use of importance sampling beyond highly stylized models. Our novel approach is guided by a large deviations principle which brings out the phenomenon of self-similarity of zero variance distributions. Being a nonparametric phenomenon, this self-similarity phenomenon is manifest in a rich set of objectives modeled with tools such as linear programs, piecewise linear/quadratic objectives, feature maps modeled with neural networks, etc. together with a spectrum of light and heavy-tailed multivariate distributions.

Wang Xingyu: Eliminating Sharp Minima from SGD with Truncated Heavy-tailed Noise

Wang Xingyu[1], Oh Sewoong[2], Rhee Chang-Han[1]

1 - Northwestern University (United States)

2 - University of Washington (United States)

Abstract: The empirical success of deep learning is often attributed to the stochastic gradient descent (SGD) algorithm's mysterious ability to avoid sharp local minima in the loss landscape, as sharp minima are known to lead to poor generalization. Recently, evidence of heavy-tailed gradient noise was reported in many deep learning tasks, and Simsekli et al. (2019) argued that SGDs are capable of escaping sharp local minima under the presence of such heavy-tailed gradient noise. This suggests a new promising perspective to the mystery, but when heavy-tailed gradient noises are present, the pure form of SGD often fails to train DNNs altogether, and one needs to deploy a variant of SGD, where gradients are truncated above a fixed threshold. In this talk, we analyze the global dynamics of such a variation (often called gradient clipping) and show that they are fundamentally different from those of the pure form of SGD. In fact, this variant attains a strong notion of avoiding sharp minima: it can effectively eliminate sharp local minima entirely from its training trajectory. We prove this by rigorously establishing an Eyring-Kramers-type first exit time formula and metastability analysis for heavy-tailed SGDs: We reveal the large deviations in heavy-tailed SGDs and obtain a sharp characterization of the relationship between the first exit time from a local basin and the ratio between the truncation threshold and the width of the basin. Moreover, when the objective function satisfies appropriate structural conditions, we prove that the heavy-tailed truncated SGD converges to a continuous-time Markov chain that never visits any sharp minima as the step size decreases. Real data experiments on deep learning confirm our theoretical prediction that heavy-tailed SGDs with gradient clipping find flatter local minima and achieve better generalization.

Zhe Su: Sample path large deviations for Levy processes and random walks with log-normal increments

Zhe Su[1], Rhee Chang-Han[1], Zwart Bert[2]

1 - Northwestern University [Evanston] (United States)
2 - Centrum Wiskunde & Informatica (Netherlands)

Abstract: The past two decades have seen breakthroughs in understanding the general principle that dictates the asymptotics of the rare events that arise in heavy-tailed systems. Hult, Lindskog, Mikosch, and Samorodnitsky (2005) was the first systematic confirmation of long-standing folklore in the heavy-tail literature called the principle of a single big jump: that is, the systemwide rare events are often driven by the catastrophic failure of a single component of the system when the underlying uncertainties are heavy-tailed, in particular, regularly varying. More recently, Rhee, Blanchet, and Zwart (2019) characterized the 'catastrophe principle' in the form of sample path large deviations, which addresses a more general class of heavy-tailed rare events far beyond the principle of a single big jump-i.e., the rare events that are driven by an arbitrary number of unusually large shocks. Bazhba, Blanchet, Rhee, and Zwart(2020) further extended the large deviations approach to the processes with (heavy-tailed) Weibul increment distributions, showing that a version of the catastrophe principle still holds in such systems, but the exact characterization is more subtle. Among the three most popular classes of distributions-regularly-varying, Weibull, and log-normal-for modeling heavy-tailed phenomena, this leaves only the log-normal case open. In this talk, we establish the sample path large deviations for random walks and Lévy processes and characterize the catastrophe principle for the log-normal case. We also discuss its implications in the tail asymptotics of queueing systems with log-normal job sizes.

Rhee Chang-Han: From heavy-tailed large deviations to local stability analysis to global dynamics

Rhee Chang-Han[1], Zwart Bert[2], Chen Bohan, Bazhba Mihail[3], Blanchet Jose[4], Seo Insuk[5], Ryu Jeeho[5]

1 - Northwestern University [Evanston] (United States)
2 - Centrum Wiskunde & Informatica (Netherlands)
3 - University of Amsterdam [Amsterdam] (Netherlands)
4 - Stanford University (United States)
5 - Seoul National University [Seoul] (South Korea)

Abstract: Stochastic dynamical systems often exhibit some form of local stability (metastability) and phase transition from a domain of local stability (a metastable set) to another domain of local stability. For example, under the presence of multiple modes in the training landscape, stochastic gradient descent (SGD) can be attracted to and "stabilizes" at a suboptimal local minimum for a long time, even if its true stationary distribution has little mass at the local minimum and is predominantly concentrated elsewhere. Such phenomena has a close connection to the large deviations of small-noise dynamical systems: If one considers a finite time horizon, escaping from a local minimum is a large deviation event. On the other hand, if one waits for a long enough time, the SGD is guaranteed to escape any local minima and eventually find the global minimum.

In this talk, we propose a new local uniform version of the heavy-tailed large deviations formulation and introduce the notion of asymptotic atom. We then show that one can transform such large deviations formulation to a sharp local stability analysis (i.e., convergence of exit time and location distributions). These machineries provide a streamlined framework for the heavy-tailed counterpart of the classical Freidlin-Wentzel theory and Eyring-Kramers formula which are limited to the light-tailed contexts. Moreover, we show that, due to the generality and precise nature of the asymptotics we develop, the analysis can be further elevated to a crisp characterization of the global dynamics of the stochastic systems. We illustrate this in the context of heavy-tailed SGDs and show that our framework successfully captures intricate mathematical structures that is unique to the heavy-tailed SGDs.

3.9 Friday 30th, 15:30 - 17:00

3.9.1 Invited session F3-1 (Lelarge): Graph matching and planted models in random graphs (Auditorium)

Friday 30th, 15:30 - 17:00 - chaired by Marc Lelarge

Araya Valdivia Ernesto: Seeded graph matching under the Correlated Gaussian Wigner Model

Braun Guillaume, Tyagi Hemant, Araya Valdivia Ernesto[1]

1 - Inria Lille - Nord Europe (France)

Abstract: Graphs are a convenient way to represent relational datasets. It has been widely used in various domains such as computer vision, biology, and sociology. In many applications, one needs to find a correspondence between these graphs. This problem is known as the graph matching problem. For example, in computer vision, identifying key points of an object in pictures of the same object in different positions is an important problem. In this work, we focus on the case where some partial and noisy knowledge about the correspondence between two graphs is available. We propose a refinement algorithm, based on the generalized power method, that iteratively refines the initial correspondence between the two graphs. We show that the algorithm, when properly initialized, provably converges after $O(\log n)$ iterations, under the Correlated Gaussian Wigner Model, a classical generative model for dense and weighted graphs. We complement our theoretical results with numerical experiments on synthetic and real datasets showing that using this refinement algorithm can considerably improve the accuracy of the obtained matching.

Ganassali Luca: Aligning graphs via detecting correlation in trees

Ganassali Luca[1]

1 - EPFL (Switzerland)

Abstract: Graph alignment refers to recovering the underlying vertex correspondence between two random graphs with correlated edges. This problem can be viewed as an average-case and noisy version of the well-known graph isomorphism problem. For correlated Erdös-Rényi random graphs, we will first give insights on the fundamental limits for the planted formulation of this problem, establishing statistical thresholds for partial recovery. Then, motivated by designing an efficient (polynomial-time) algorithm to recover the underlying alignment in a sparse regime, a message-passing algorithm based on testing correlation in trees is proposed. We study this related correlation detection problem in trees and identify a phase transition in the limit of large degrees for the existence of suitable tests, hence giving insights on the performance of the above method for our initial problem on graphs. Based on joint works with Laurent Massoulié, Marc Lelarge and Guilhem Semerjian.

Mao Cheng: Detection-Recovery Gap for Planted Dense Cycles

Mao Cheng[1]

1 - Georgia Institute of Technology (United States)

Abstract: Planted dense cycles are a type of latent structure that appears in many applications, such as small-world networks in social sciences and sequence assembly in computational biology. We consider a model where a dense cycle with expected bandwidth $n\tau$ and edge density p is planted in an Erdős–Rényi graph $G(n, q)$. We characterize the computational thresholds for the associated detection and recovery problems for the class of low-degree polynomial algorithms. In particular, a gap exists between the two thresholds in a certain regime of parameters. For example, if $n^{-3/4} \ll \tau \ll n^{-1/2}$ and $p = Cq = \Theta(1)$ for a constant $C > 1$, the detection problem is computationally easy while the recovery problem is hard for low-degree algorithms.

Yu Sophie: Random graph matching at Otter's tree-counting threshold via counting chandeliers

Mao Cheng[1], Wu Yihong[2], Xu Jiaming[3], Yu Sophie[4]

1 - Georgia Institute of Technology (United States)
2 - Yale University (United States)
3 - Duke University (United States)
4 - Duke university [Durham] (United States)

Abstract: Given a pair of graphs, the problem of graph matching refers to finding the underlying vertex correspondence that maximally aligns the edges. This is a ubiquitous problem arising in a variety of applications across diverse fields, such as network privacy, computational biology, computer vision, and natural language processing. Graph matching is an instance of the notoriously difficult quadratic assignment problem (QAP), which is NP-hard to solve or approximate.

Despite the worst-case computational hardness of QAP, I will present a computationally efficient graph matching algorithm based on counting a special family of trees. When the two graphs are Erdös-Rényi random graphs with correlated edges through the hidden vertex correspondence, we show that our algorithm correctly matches all but a vanishing fraction of vertices with high probability as soon as the edge correlation exceeds the square root of Otter's constant. Moreover, we further upgrade the almost exact recovery to exact recovery whenever it is information-theoretically possible. This is the first polynomial-time algorithm that achieves exact and almost exact matching with an explicit constant correlation for both dense and sparse graphs.

3.9.2 Invited session F3-2 (Huang & Xie): Stochastic models and their application (Room 101)

Friday 30th, 15:30 - 17:00 - chaired by Junfei Huang & Jingui Xie

Gao Xuefeng: Asymptotically Optimal Control of Make-to-Stock Systems

Gao Xuefeng[1], Huang Junfei

1 - The Chinese University of Hong Kong [Hong Kong] (China)

Abstract: We consider multi-class make-to-stock production/inventory systems in which the manager makes three decisions, including pricing, outsourcing, and scheduling, to maximize the long-run average profit. For a sequence of systems in the heavy-traffic regime, with linear or strictly convex holding/waiting cost functions, we propose a sequence of policies and establish its asymptotic optimality. Our proof combines the lower bound approach and a thorough steady-state analysis of the systems. We also establish general results on the existence and tightness of the stationary distributions of the state processes under a more general family of policies.

Dong Jing: Scheduling queues with different and multiple resource requirements

Dong Jing[1], Chan Carri[1], Xi Weiye[2]

1 - Columbia University (United States)
2 - Tsinghua University (China)

Abstract: Motivated by the hospital inpatient setting, we study the scheduling of a new class of multi-class queueing systems with multiple types of resources where different classes of customers have heterogeneous - in terms of the type and amount - resource requirements. In particular, a customer may require different numbers of "servers" from different server types to be allocated simultaneously in order to be served. Some service capacity constraints cannot be violated while others can be violated with a penalty. We characterize a near-optimal scheduling policy. Our results provide insights into how the capacity violation penalty affects the scheduling decision.

Shi Pengyi: Inpatient Overflow Management With Proximal Policy Optimization

Shi Pengyi[1], Sun Jingjing[2], Dai Jim[3]

1 - Purdue University [West Lafayette] (United States)
2 - CUHK-Shenzhen [Shenzhen] (China)
3 - Cornell University [Ithaca] (United States)

Abstract: One effective strategy for reducing congestion in hospitals is to overflow patients to non-primary wards. However, this approach can lead to other issues such as mismatches in skillsets and compromised quality of care. Therefore, it is necessary to balance the trade-off between congestion and overflow costs. We model this overflow management problem as a discrete-time long-run average cost Markov Decision Process, which have extremely large state and action spaces for systems with practical relevance. To overcome the curse-of-dimensionality, we develop a scalable actor-critic algorithm based on Proximal Policy Optimization (PPO). Our approach incorporates two key innovations: (i) the integration of atomic action and randomization to reduce the action space, and (ii) the use of a tailored time-dependent approximation structure to estimate the advantage function. We show that our algorithm can handle realistic problem sizes with over 20 server pools, overcoming the limitations of prior approaches. Moreover, we show that our algorithm consistently outperforms state-of-the-art policies, and innovative features in the algorithm design tailored to this domain application are essential for achieving superior performance.

Wang Zhengli: A Stochastic Control Framework for New Venture Creation

Wang Zhengli[1], Zenios Stefanos[2]

1 - The University of Hong Kong (Hong Kong SAR China)
2 - Stanford University (United States)

Abstract: We model the creation of a new venture with a novel stochastic control framework in which the state of the venture is captured by a diffusion process. The entrepreneur creating the venture chooses costly controls, which determine both the drift and the variance of the process. When the process reaches an upper boundary, the venture succeeds and the entrepreneur receives a reward. When the process reaches a lower boundary, the venture fails. The entrepreneur can choose between different controls and wishes to determine the policy that maximizes the expected total reward minus total cost. We derive closed-form expressions under which the optimal policy will be dynamic versus static and characterize the structure of the optimal policy.

3.9.3 Invited session F3-3 (Stolyar): Large-scale and heavy-traffic models (Room 102)

Friday 30th, 15:30 - 17:00 - chaired by Alexander Stolyar

Jhunjhunwala Prakirt: Rate of Convergence of Tail Probabilities to Heavy Traffic

Jhunjhunwala Prakirt[1], Hurtado-Lange Daniela[2], Maguluri Siva Theja[3]

1 - Georgia Institute of Technology (United States)
2 - William & Mary (United States)
3 - School of Industrial and Systems Engineering [Georgia Tech] (United States)

Abstract: A popular approach to computing performance measures of queueing systems (such as delay and queue length) is studying the system in an asymptotic regime. However, these results are only valid in the limit and often provide bounds for the pre-limit systems that are not optimized and, hence, give loose bounds for the tail probabilities. In this work, we provide optimized bounds for the tail probabilities of the scaled total queue length in a load-balancing system under Join the Shortest Queue (JSQ). Our bounds characterize the rate of convergence of the tail probabilities to the corresponding heavy traffic values. For the tail probability of the JSQ system, our bounds yield a multiplicative error that arises from three factors: pre-limit errors, pre-exponent errors, and State-Space Collapse (SSC).

Mukherjee Debankur: A Stochastic Approximation Approach to Large-Scale Distributed Rate Scaling

Mukherjee Debankur[1], Rutten Daan[1], Zubeldia Martin[2]

1 - School of Industrial and Systems Engineering [Georgia Tech] (United States)
2 - University of Minnesota (ISyE) (United States)

Abstract: In this talk, we will discuss a distributed rate-scaling mechanism for large-scale parallel-server systems. The goal is to strike a delicate balance between the cost of maintaining the respective server-processing speeds and the average sojourn time of tasks in the entire system. The key challenge arises from that servers need to dynamically scale their processing speeds in a fully distributed way without any centralized control or exchange of information among the servers and global quantities like the arrival rate, are unknown to the servers. Using insights from the stochastic approximation, we develop a novel algorithm where the processing speed of each server converges to the globally optimum value in the steady state as the system size becomes large. En route, we also analyze the performance of a fully heterogeneous parallel-server system (i.e, each server has a different processing speed), which might be of independent interest.

Zhong Yuan: Beating the curse of dimensionality in high-dimensional reflected diffusion control problems

Zhong Yuan[1]

1 - University of Chicago (United States)

Abstract: Inspired by recent advances in efficient numerical schemes for high-dimensional partial differential equations, we develop a multi-level Monte Carlo Picard approximation scheme for a class of reflected diffusion control problems. These problems arise as approximations to heavy-traffic diffusion limits of parallel server systems, which have broad applications in service and technological systems. We

show that the proposed numerical approximation scheme is provably efficient, in that the computational effort grows at most polynomially in both the dimension of the state space and the inverse of the required accuracy.

Stolyar Alexander: Data Flow Dissemination in a Network

Gopalan Aditya[1], Stolyar Alexander[1]

1 - University of Illinois at Urbana-Champaign (United States)

Abstract: We consider a network where multiple packet flows, originating in different nodes, need to be disseminated to all other nodes. (This model is motivated, in particular, by blockchains and peer-to-peer live streaming.) We show that Send-Random-Useful-Packet (RU) discipline is throughput optimal (has maximum stability region). Send-Oldest-Useful-Packet (OU) discipline is a natural candidate for reducing the network Age-of-Information (AoI), i.e. the oldest packet age. We show that, surprisingly, OU discipline is not throughput optimal. We prove that OU is throughput optimal in the important special case of a complete-graph symmetric model. We also study the cumulative delays experienced by a packet as it propagates through the network, specifically their limiting behavior as the network size becomes large.

3.9.4 Contributed session F3-4 (Remiche): Stochastic processes and Markov modeling (Room 103)

Friday 30th, 15:30 - 17:00 - chaired by Marie-Ange Remiche

Yang Yunfang: Load Balancing in Geographically Separated Service Systems

He Shuangchi[1], Yang Yunfang[1]

1 - National University of Singapore (Singapore)

Abstract: Nowadays, smartphone apps allow customers to access real-time information (such as waiting times or queue lengths) about service facilities (e.g., clinics and bank branches). A customer may use this information to decide where to go, if there are multiple facilities he can choose. If a facility is distant from the customer, the information may become obsolete when the customer arrives. It is known that load balancing policies that rely on current information, such as the join-the-shortest-queue (JSQ) policy, may induce queue length oscillations in the presence of information delays and seriously hurt the system performance. In this work, we study a randomized variant of the JSQ policy for a service system with geographically separated facilities in a region where customers randomly appear from different locations. We prove that the proposed randomized policy is asymptotically optimal for minimizing a customer's time to service (i.e., the sum of the traveling delay and the waiting time). We also establish a lower bound for the time to service and demonstrate how to minimize the performance gap by fine-tuning the randomized policy.

Dendievel Sarah: Analysis of Singular Perturbation for a Markovian Fluid Model

Dendievel Sarah[1,2]

1 - Haute Ecole ICHEC - ECAM - ISFSC, Département ICHEC Brussels Management School, QUARESMI Laboratory (Belgium)

2 - Ghent University, TELIN Department, SMACS Research Group (Belgium)

Abstract: In this talk, we present a Markovian fluid model where the states are nearly completely decomposable. Phases are thus organised in aggregates. The purpose of this research is to study the stationary density of the level of the complete model in the case of positive recurrent aggregates but also to develop further analysis for the case when there are two aggregates with one being positive recurrent and the other being transient. We provide some numerical example.

Ali Mohammadi Yeganeh: The Power of a Few Local Samples in Predicting Epidemics

Ali Mohammadi Yeganeh[1], Saberi Amin[1], Borgs Christian[2], Van Der Hofstad Remco

1 - Stanford (United States)

2 - UC Berkeley (United States)

Abstract: The spread of infectious diseases is heavily influenced by the interactions between individuals in a population. Accurately mapping these interactions can be expensive and sometimes impossible, making it difficult to predict the likelihood and outcome of an outbreak. In this talk, I will show that it is

not necessary to map the entire network to make these predictions. By using a non-parametric estimator based on local graph features, it is possible to accurately estimate the likelihood and size of an outbreak and the time evolution of epidemics using only a few samples from the population. This result applies to a large class of networks and can be extended to general random processes on graphs.

Remiche Marie-Ange: Sojourn time distribution in a fluid queue with server vacations

Deiana Eleonora[1], Latouche Guy[2], Remiche Marie-Ange[1]

1 - Université de Namur, Faculté d'Informatique (Belgium)
2 - Universite Libre de Bruxelles (Belgium)

Abstract: We use a Markov driven fluid queue with infinite buffer to model a system with server vacations. In our setting, servers enter into vacation as soon as two conditions are fulfilled. Not only must the buffer be empty but also the input rate of arriving fluid has to be null. The stationary distribution of the buffer content is obtained using a regenerative approach. We determine the sojourn time distribution using its Laplace-Stieltjes transform. To that end, we define a new clock that increases at the same rate as new fluid enters the buffer. In this talk, we highlight the machinery behind the resulting equations, keeping our argument on an intuitive level as much as possible.

3.9.5 Invited session F3-5 (Agrawal): Learning, Optimization and Applied Probability (Room 104)

Friday 30th, 15:30 - 17:00 - chaired by Shubhada Agrawal

Sundaresan Rajesh: Learning to detect an anomalous Markov process

Sundaresan Rajesh[1,2]

1 - Robert Bosch Centre for Cyber-Physical Systems (Indian Institute of Science Bangalore - 560012, INDIA India)
2 - Department of Electrical Communication Engineering (Indian Institute of Science Bangalore - 560012, INDIA India)

Abstract: The talk will discuss the problem of finding an anomalous arm in a multi-armed bandit when (a) each arm is a finite-state Markov process, and (b) the arms are restless. Here, anomaly means that the transition probability matrix (TPM) of one of the arms (the odd arm) is different from the common TPM of each of the non-odd arms. The TPMs are unknown to a decision entity that wishes to find the index of the odd arm as quickly as possible, subject to an upper bound on the error probability. We will discuss a problem instance-specific asymptotic lower bound on the expected time required to find the odd arm index, where the asymptotics is as the error probability vanishes. Further, we will highlight a policy based on the principle of certainty equivalence that, under a continuous selection assumption and a certain regularity assumption on the TPMs, achieves the lower bound arbitrarily closely. The achievability analysis is based on resolving the identifiability problem in the context of a certain lifted countable-state controlled Markov process. This is joint work with P. N. Karthik.

Juneja Sandeep: Guided top-2 approach to optimally solve the best arm identification problem

Agrawal Shubhada[1], Juneja Sandeep[2,3], Shanmugam Karthikeyan[3], Suggala Arun[3]

1 - Georgia Tech (United States)
2 - Tata Institute of Fundamental Research (India)
3 - Google Research (India)

Abstract: We consider the popular best arm identification problem where given finitely many probability distributions or arms that can be sampled from, the aim is to sequentially generate a minimum expected number of samples and arrive at the correct arm with a high pre-specified degree of confidence 1- delta. Elegant lower bound max-min formulations on the sample complexity exist for this BAI problem, and algorithms that plug-in empirical data based estimates into the lower bound and track the resulting prescriptive allocations have been shown to match the lower bound as delta reduces to zero. These are however computationally inefficient as they rely on repeatedly solving the max-min optimization problem. Lately, top two methods have been devised that are much simpler and exploit some properties of the optimal lower bound solution. They rely on sequentially sampling either the empirical best arm with probability beta or the empirical best challenger with probability 1-beta. The drawback of these methods is that the optimal beta is not known. In this talk, we introduce the guided top 2 algorithm that samples the empirically best versus the challenger so that the resulting sample complexity matches the lower bound as delta reduces to zero.

Maguluri Siva Theja: Concentration of Contractive Stochastic Approximation: Additive and Multiplicative Noise

Chen Zaiwei[1], Zubeldia Martin[2], Maguluri Siva Theja[3]

1 - California Institute of Technology (United States)
2 - University of Minnesota (United States)
3 - School of Industrial and Systems Engineering [Georgia Tech] (United States)

Abstract: We study the concentration behavior of a stochastic approximation algorithm under a contractive operator with respect to an arbitrary norm. We consider two settings where the iterates are potentially unbounded: (1) bounded multiplicative noise, and (2) additive sub-Gaussian noise. We obtain concentration bounds on the convergence errors, and show that these errors have sub-Gaussian tails in the additive noise setting, and super-polynomial tails (faster than polynomial decay) in the multiplicative noise setting. Moreover, our bounds hold anytime in the sense that the entire sample path lies within a tube of decaying radius with high probability. To establish these results, we develop a novel bootstrapping argument that involves bounding the moment generating function of the generalized Moreau envelope of the error, and the construction of an exponential supermartingale to enable the use Ville's maximal inequality to obtain anytime bound. These results are motivated by application in Reinfrocement Learning, and can be used to obtain high probablity convergence bounds and sample complexity for a large class of RL algorithms including TD learning and its variants, as well as Q learning.

Agrawal Shubhada: Bandits, Empirical Likelihood Confidence Intervals, and Heavy Tailed Distributions

Agrawal Shubhada[1], Juneja Sandeep[2], Koolen Wouter[3], Glynn Peter[4]

1 - School of Industrial and Systems Engineering [Georgia Tech] (United States)
2 - Tata Institute of Fundamental Research (India)
3 - Centrum Wiskunde & Informatica (Netherlands)
4 - Stanford University (United States)

Abstract: Multi-armed bandit (MAB) is a popular framework for sequential decision-making in an uncertain environment. In the classical setup of MAB, the algorithm has access to a fixed and finite set of K unknown, independent probability distributions or arms. At each time step, having observed the outcomes of all the previous actions, the algorithm chooses one of the K arms and receives an independent sample drawn from the underlying distribution, which may be considered a reward. The algorithm's goal is either to maximize the accumulated rewards or to identify the arm with the maximum mean in as few samples as possible. These problems are well-studied in literature, and tight lower bounds and optimal algorithms exist when the arm distributions are known to belong to simple classes of distributions such as single-parameter exponential family or distributions that have bounded support, etc. However, in practice, the distributions may not satisfy these assumptions and may even be heavy-tailed. In this talk, we will look at techniques and algorithms for optimally solving these problems with minimal assumptions on the arm distributions. A key component of designing an optimal algorithm for MAB is constructing tight, anytime valid confidence intervals for mean. We will look at new concentration inequalities for heavy-tailed distributions, which may be of independent interest.

3.9.6 Invited session F3-6 (Chen): Learning and queueing theory (Room 105)

Friday 30th, 15:30 - 17:00 - chaired by Xinyun Chen

Hong Guiyu: An online learning method for the staffing problem of Hawkes queues

Hong Guiyu[1], Chen Xinyun[1]

1 - The Chinese University of Hong Kong, Shenzhen (China)

Abstract: We investigate the capacity sizing problem for a single-sever queue with Hawkes arrivals and general service time distributions. Due to the complexity of system dynamics, there is no analytic solution to the problem. We develop an online learning algorithm to numerically compute the optimal service capacity with theoretic guarantee. In particular, we prove that with properly chosen hyper-parameters, our algorithm obtains regret of logarithmic order. To obtain the regret bound, we establish geometric ergodicity for single-server queues with Hawkes input using coupling technique, which we believe is of independent interest. In the end, we apply the online learning algorithm to solve capacity sizing problem for a variety of single-server queues with Hawkes input. The numerical results indicate a sharp difference between GI/GI/1 and Hawkes/GI/1 queues, especially in some asymptotic regime, indicating the importance of taking into account autocorrelation in arrival processes in queueing controls.

Si Nian: Drift control of high-dimensional RBM: A computational method based on neural networks

Si Nian[1], Ata Baris[1], Harrison Michael[2]

1 - The University of Chicago Booth School of Business (United States)
2 - Stanford Graduate School of Business (United States)

Abstract: We consider a stochastic control problem whose state space is the d-dimensional non-negative orthant. The controlled process Z evolves as a Brownian motion whose covariance matrix is exogenously specified, as are its directions of reflection from the orthant's boundary surfaces. A controlling agent chooses a drift vector q(t) at each time based on the history of Z, and the cost rate at time depends on both Z(t) and q(t). In our initial problem formulation, the objective is to minimize expected discounted cost over an infinite planning horizon, but we also treat the corresponding ergodic control problem. Extending earlier work by Han et al. (Proceedings of the National Academy of Sciences, 2018, 8505-8510), we develop and illustrate a simulation-based computational method that relies heavily on deep neural network technology. Our method is accurate to within a fraction of one percent, and is computationally feasible in dimensions up to at least d = 20. (Based on joint work with Baris Ata and Mike Harrision.)

Zhang Hailun: The Generalized c/μ Rule for Queues with Heterogeneous Server Pools

Long Zhenghua[1], Zhang Hailun[2], Zhang Jiheng[3], Zhang Zhe George[4]

1 - Nanjing University (China)
2 - The Chinese University of Hong Kong, Shenzhen (China)
3 - Hong Kong University of Science and Technology (Hong Kong SAR China)
4 - Western Washington University (United States)

Abstract: We study the optimal control of a queueing model with a single customer class and heterogeneous server pools under two kinds of objectives. The first is to strike a balance between the holding cost of the queue and the operating costs of the server pools. We introduce a target-allocation policy, which assigns higher priority to the queue or pools without enough customers, for general cost functions. Although we can prove its asymptotic optimality implementation requires solving a nonlinear optimization problem. When the cost functions are convex, we propose a dynamic priority policy referred to as the Gc/μ rule, which is much easier to implement. When the cost functions are concave, it turns out that a fixed priority policy is optimal. The second objective is to minimize the operating cost of the server pools while satisfying a service-level target for customers waiting in the queue. We develop hybrid routing policies, combining a threshold policy for the queue and the aforementioned policies for the server pools, for different types of operating cost functions. Moreover, the hybrid routing policies coincide with several classical policies in the literature in special cases. Extensive simulation experiments demonstrate the efficacy of our proposed policies.

Cheung Wang Chi: Non-Stationary Bandits with Knapsack Problems with Advice

Lyu Lixing[1], Cheung Wang Chi[1]

1 - National University of Singapore (Singapore)

Abstract: We consider a non-stationary Bandits with Knapsack problem. The outcome distribution at each time is scaled by a non-stationary quantity that signifies changing demand volumes. Instead of studying settings with limited non-stationarity, we investigate how online predictions on the total demand volume Q allows us to improve our performance guarantees. We show that, without any prediction, any online algorithm incurs a linear-in-T regret. In contrast, with online predictions on Q, we propose an online algorithm that judiciously incorporates the predictions, and achieve regret bounds that depends on the accuracy of the predictions. These bounds are shown to be tight in settings when prediction accuracy improves across time. Our theoretical results are corroborated by our numerical findings.

3.9.7 Invited session F3-7 (Shneer & Zocca): Mathematics of energy (Room 106)

Friday 30th, 15:30 - 17:00 - chaired by Seva Shneer & Alessandro Zocca

Christianen Mark: Comparison of stability regions for a line distribution network with stochastic load demands

Christianen Mark[1], Cruise James[2], Janssen Guido[1], Shneer Seva[2], Vlasiou Maria[1,3], Zwart Bert[1,4]

1 - Eindhoven University of Technology (Netherlands)

2 - Heriot-Watt University [Edinburgh] (United Kingdom)
3 - University of Twente (Netherlands)
4 - Centrum voor Wiskunde en Informatica (Netherlands)

Abstract: We compare stability regions for different power flow models in the process of charging electric vehicles (EVs) by considering their random arrivals, their stochastic demand for energy at charging stations, and the characteristics of the electricity distribution network. We assume the distribution network is a line with charging stations located on it. We consider the Distflow and the Linearized Distflow models and we assume that EVs have an exponential charging requirement, that voltage drops on the distribution network stay under control and that the number of charging stations N goes to infinity. We investigate the stability of utility-optimizing power allocations in large distribution networks for both power flow models by controlling the arrival rate of EVs to charging stations. For both power flow models, we show that, to obtain stability, the maximum feasible arrival rate, i.e. stability region of vehicles is decaying as $1/N^2$, and the difference between those arrival rates is up to constants, which we compare explicitly.

Moriarty John: The structure of optimal stopping problems under risk

Moriarty John[1], Kosmala Tomasz[1]

1 - Queen Mary University of London (United Kingdom)

Abstract: We present recent work on optimally stopping a diffusion process under a risk mapping. The problem is solved by extending an approach which was originally developed by S. Christensen to deal with bounded ambiguity over the drift term. Our method reveals the general structure of optimal stopping problems under a broad class of Markovian risk mappings, and we discuss motivation from energy problems and potential applications.

Van Der Sar Erica: Optimizing Power Grid Control with Reinforcement Learning

Van Der Sar Erica[1], Bhulai Sandjai[1], Zocca Alessandro[1]

1 - Vrije Universiteit Amsterdam [Amsterdam] (Netherlands)

Abstract: In the past few decades, power transmission networks have changed tremendously, mostly due to a dramatic increase in energy demand and ambitious sustainability goals. In particular, the increasing penetration of renewable energy sources, such as wind and solar, introduces a substantial amount of stochasticity in these already complex networks. As a result, it is becoming more and more difficult for network operators to control power grids and keep them safe and reliable at all times. The only way to achieve this is by using the network flexibility to its fullest extent and thus being able to identify in real-time optimal control actions in a combinatorially very large action space. Multiple reinforcement learning frameworks have been proposed in recent years to assist network operators in the decision-making for real-time operations. In such settings, the RL agent learns how to optimally control the power network using various actions that can change the network topology and thus, indirectly, the power flows. In this talk, we illustrate how existing RL approaches can be improved using Graph Neural Networks (GNN) and additional information, such as renewable and load forecasts.

Zocca Alessandro: Generating synthetic power grids using Exponential Random Graphs

Zocca Alessandro[1], Giacomarra Francesco[2], Bet Gianmarco[3]

1 - Vrije Universiteit, Amsterdam (Netherlands)
2 - University of Trieste (Italy)
3 - University of Florence (Italy)

Abstract: Power grids are fundamental infrastructures of our modern societies and economies. However, only limited information on transmission power grids is released by network operators due to security concerns. Thus, there is a substantial lack of high-quality real data. To address this problem, in the last two decades, synthetic grid generation approaches have been extensively investigated by the research community. The main challenge has been developing models flexible enough to replicate the very heterogeneous nature and peculiar properties of real power grids.

In this talk, I will describe a novel method for the generation of synthetic transmission power grids using Exponential Random Graph (ERG) models. From a modeling perspective, we identify edge count and k-triangles as the crucial topological characteristics for synthetic power grid generation. From a technical perspective, we develop a rigorous methodology to estimate the parameters of an ERG constrained to the space of connected graphs. The proposed ERG model to generate the topological structure of power grids is flexible, easy to implement, and successfully captures the desired topological properties of

power grids.

3.9.8 Invited session F3-8 (Varma & Zubeldia): Stochastic Matching Networks and their Applications (Room 202)

Friday 30th, 15:30 - 17:00 - chaired by Sushil Varma & Martin Zubeldia

Hillas Lisa: Designing Service Menus for Bipartite Qeueing Systems

Caldentey Rene[1], Gupta Varun[1], Hillas Lisa[1]

1 - Booth School of Business [Chicago] (United States)

Abstract: We consider a multi-class multi-server queueing system, in which customers of different types have heterogeneous preferences over the many servers available. The goal of the service provider is to design a menu of service classes that balances two competing objectives: (1) maximize customers' average matching reward and (2) minimize customers' average waiting time. A service class corresponds to a single queue served by a subset of servers under a FCFS-ALIS service discipline. Customers act as rational self-interested utility maximizing agents when choosing which service class to join. In particular, they join the class that maximizes their expected ex-ante net utility, which is given by the difference between the server-dependent service reward they receive minus a disutility based on the mean steady-state waiting time of the service class they join. We study the problem under (conventional) heavy traffic conditions, that is, in the limit as the traffic intensity of the system approaches one from below. For an arbitrary number of servers, we prove that if the service provider only cares about minimizing average delay or maximizing total matching reward then very simple menus are optimal. We also provide Mixed Integer Linear Programming (MILP) formulations for optimizing the delay-reward trade-off within fairly rich and practically relevant families of menus.

Wang Qiong: A Stochastic Tracking Model: Asymptotic Analysis and Applications

Wang Qiong[1], Martin Reiman[2], Haohua Wan[3]

1 - Department of ISE, University of Illinois at Urbana-Champaign (United States)
2 -Columbia University (United States)
3 - University of Illinois at Urbana Champaign (United States)

Abstract: In inventory systems, control of some key processes, such as an inventory position or a backlog level, can only be executed from one direction. For instance, an inventory position can be raised instantly to a target level for minimizing a cost objective according to the current state. However, when the change of the state requires the position to be reduced to a lower target, no action can be taken except to wait for some external events, such as a new demand arrival, to bring about the reduction. The situation motivates us to develop and analyze a generic stochastic tracking model that is defined by three processes: a compound Poisson demand process with appropriate moment conditions on the jump size; a target process that evolves over time; and a tracking process that involves negative jumps trigged by demand arrivals and subjects to controls that can raise its level to meet a higher target. We consider an asymptotic regime constructed by scaling up the demand rate. We show that if the target process is Lipschitz continuous in increments of the demand process over some periods of fixed lengths, then the expected difference between the target and tracking processes converges uniformly to zero on the diffusion scale over an infinite time horizon. We discuss implications of this result for inventory management, for instance, the assemble-to-order systems, and other possible applications.

Yang Xiaotang: Human in The Loop Automation: Ride-Hailing with Remote (Tele-) Drivers

Benjaafar Saif[1], Wang Zicheng[1], Yang Xiaotang[1]

1 - University of Minnesota [Twin Cities] (United States)

Abstract: Tele-driving refers to a novel concept where drivers can remotely operate vehicles (without being physically in the vehicle). By putting the human back "in the loop," tele-driving has emerged recently as a more viable alternative to fully automated vehicles, with ride-hailing (and other on-demand transportation-enabled services) being an important application. Because remote drivers can be operated as a shared resource (any driver can be assigned to any customer regardless of trip origin or destination), tele-driving has the potential to reduce the severity of the spatial mismatch between vehicle supply and customer demand that is often experienced in these services and the number of drivers needed. In this paper, we quantify the potential gains that could be realized by switching from in-vehicle drivers to remote drivers. We compare a traditional ride-hailing system with one with tele-drivers in three regimes defined

by vehicle capacity. We find that: (1) if customers are impatient, a system with appropriately selected driver capacity may significantly improve the service level (the fraction of demand that is served), or significantly reduce the number of drivers while maintaining a similar service level; and (2) if customers are patient, a system with remote drivers may stabilize an otherwise unstable system with in-vehicle drivers or significantly reduce the number of drivers while maintaining a similar service level (as measured by the expected delay experienced by customers).

Varma Sushil: Electric Vehicle Fleet and Charging Infrastructure Planning

Varma Sushil Mahavir[1], Castro Francisco[2], Maguluri Siva Theja[1]

1 - Georgia Institute of Technology [Atlanta] (United States)
2 - UCLA Anderson School of Management (United States)

Abstract: We analyze an optimal electric vehicle (EV) fleet and charging infrastructure capacity planning problem in a spatial setting. As customer requests arrive at rate λ, the system operator must determine the minimum number of vehicles and chargers along with a matching and charging policy that maximizes the service level. We provide a sharp characterization of the fleet size and the charging infrastructure requirements as the demand grows. While a system in which charging times are negligible needs extra $\theta(\lambda^{2/3})$ vehicles on top of the nominal capacity, we show that an EV system has a fundamentally different scaling. Due to charging times, the nominal capacity of the system is increased, but this extra capacity allows for an optimal EV dispatching policy to result in a fleet requirement of only $\theta(\lambda^{\nu})$ for $\nu \in (1/2, 2/3]$, depending on the number of charging stations and the size of the EV battery packs. We propose the Power-of-d dispatching policy, which achieves this performance by selecting the d closest vehicles to a trip request and choosing the one with the highest battery level, thus optimizing the trade-off between the pickup distance and balancing the state of charge across the fleet. Our study provides valuable guidelines for determining the optimal fleet and charging infrastructure capacity for an EV-based on-demand transportation system.

3.9.9 Invited session F3-9 (Zähle): Stability and statistical inference of stochastic optimization problems in quantitative risk management (Room 203)

Friday 30th, 15:30 - 17:00 - chaired by Henryk Zähle

Krätschmer Volker: Nonasymptotic error rates of the sample average approximation method to solve risk averse stochastic progams

Krätschmer Volker[1]

1 - University Duisburg-Essen (Germany)

Abstract: We study statistical properties of the optimal value of the Sample Average Approximation. The focus is on rates dependent on the sample sizes for the tail function of the absolute error induced by the Sample Average Approximation. They allow to conclude immediately convergence rates for the optimal value of the Sample Average Approximation. As a crucial point the investigations are based on a new type of conditions from the theory of empirical processes which do not rely on pathwise analytical properties of the goal functions. In particular, continuity in the parameter is not imposed in advance as usual in the literature on the Sample Average Approximation method. The main results are shown for classical risk neutral stochastic programs, but we also demonstrate how to apply them to the sample average approximation of risk averse stochastic programs. In this respect we consider stochastic programs expressed in terms of mean upper semideviations and divergence risk measures.

Wiesel Johannes: Sensitivity analysis of Wasserstein distributionally robust optimization problems

Bartl Daniel[1], Drapeau Samuel[2], Obloj Jan[3], Wiesel Johannes[4]

1 - University of Vienna (Austria)
2 - Shanghai Jiao Tong University (China)
3 - Oxford University (United Kingdom)
4 - Columbia University (United States)

Abstract: In this talk I consider sensitivity of a generic stochastic optimization problem to model uncertainty, where I take a non-parametric approach and capture model uncertainty using Wasserstein balls around the postulated model. I provide explicit formulae for the first order correction to both the value function and the optimizer and further extend our results to optimization under linear constraints. Then I

present applications to statistics, machine learning, mathematical finance and uncertainty quantification. In particular, I prove that LASSO leads to parameter shrinkage, propose measures to quantify robustness of neural networks to adversarial examples and compute sensitivities of optimised certainty equivalents in finance. I also propose extensions of this framework to a multiperiod setting.

Xu Huifu: Robust Spectral Risk Optimization When Information on Risk Spectrum Is Incomplete

Wang Wei[1], Xu Huifu[1]

1 - The Chinese University of Hong Kong (Hong Kong SAR China)

Abstract: A spectral risk measure (SRM) is a weighted average of value at risk where theweighting function (also known as risk spectrum or distortion function) characterizes a decision maker's risk attitude. In this work, we consider the case where the decision maker's risk spectrum isambiguous and introduce a robust SRM model based on the worst risk spectrum from a ball of riskspectra centered at a nominal risk spectrum. When the ball consists of step-like risk spectra, we showthat the robust SRM can be computed by solving a linear programming problem. For the general case, we propose a step-like approximation scheme and derive an error bound for the approximation.As an application, we apply the proposed robust SRM to one-stage stochastic optimization with theobjective of minimizing the robust SRM and propose an alternating iterative algorithm for solvingthe resulting minimax optimization problem. Moreover, to examine stability of the robust spectralrisk optimization model with respect to perturbation of observed data from the underlying exogenous uncertainty in data-driven environments, we investigate statistical robustness of the model and derive sufficient conditions for the required stability.

Zähle Henryk: A concept of copula robustness and its applications

Zähle Henryk[1]

1 - Saarland University (Germany)

Abstract: In financial and actuarial applications, marginal risks and their dependence structure are often modelled separately. While it is sometimes reasonable to assume that the marginal distributions are "known", it is usually quite involved to obtain information on the copula. Therefore copula models used in practice are quite often only rough guesses. For many purposes, it is thus relevant to know whether certain characteristics derived from d-variate risks are robust with respect to (at least small) deviations in the copula. In this talk, a general concept of copula robustness is introduced and criteria for copula robustness are presented and illustrated by examples.

3.9.10 Contributed session F3-10 (Minabutdinov): Contagion and Systemic risk (Room 204)

Friday 30th, 15:30 - 17:00 - chaired by Aleksei Minabutdinov

Hakim Arief: Conditional quantile and expectile-based systemic risk measures and systemic networks for dirty and clean energy markets

Hakim Arief[1], Syuhada Khreshna[1]

1 - Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung (Indonesia)

Abstract: Conditional quantiles are crucial for many statistical problems, including systemic risk quantification. Specifically, we can determine conditional quantile-based Value-at-Risk (hereinafter, CoQVaR), which measures a financial entity's risk conditional on another entity's risk and thus quantifies the systemic impact of the latter on the former. However, CoQVaR depends only on the probability of risk occurrence and exhibits incoherence. This leads us to propose conditional expectile-based Value-at-Risk (henceforth, CoEVaR), which accounts for risk magnitude and satisfies risk measure coherence. In this study, we aim to formulate CoQVaR and CoEVaR. Since their bases (i.e., conditional quantiles and expectiles) are determined by optimizing expected conditional loss functions, we adopt an efficient method for the computation of conditional expectations. We compare the performances of the resulting CoQVaR and CoEVaR forecasts by evaluating the estimates of their respective expected conditional loss functions. As an application, we perform an empirical study using the daily return series of dirty energy commodities and clean energy indices. In addition, we compute the so-called Delta-CoQVaR and Delta-CoEVaR to quantify the systemic risk contribution of one energy market to the other. Using these measures, we then construct systemic networks to describe the mechanism of systemic risk transmission across these energy markets. This approach helps us investigate the systemic relevance and fragility of each energy market and maintain the stability of global energy markets.

Lagos Guido: An asymptotic systemic failure time result for mixed coherent systems

Lagos Guido[1], Barrera Javiera[1,2], Romero Pablo[3,4], Valencia Juan[5]

1 - Universidad Adolfo Ibáñez [Viña del Mar] (Chile)
2 - Northwestern University [Evanston] (United States)
3 - Universidad de la República [Montevideo] (Uruguay)
4 - Universidad de Buenos Aires [Buenos Aires] (Argentina)
5 - Universidad de Santiago de Chile [Santiago] (Chile)

Abstract: In this work we show a limit result for the probability of a system being in an operational (working) state as the size of the system grows to infinity. More specifically, we consider a sequence of mixed coherent systems whose components are homogeneous and non-repairable, with failure times that allow simultaneous failures of the components. We show that under dominated convergence conditions the system failure time converges in distribution to a first-passage time of a Levy subordinator process. To illustrate, we give a parametric family of systems where the time of systemic failure converges to an exponential distribution. To the best of our knowledge, this is the first result to tackle the asymptotic behavior of the system failure time as the number of components of the system grows.

Amanihamedani Alireza: Spatial Matching under Resource Competition

Amanihamedani Alireza[1], Freund Daniel[2], Aouad Ali[1]

1 - London Business School (United Kingdom)
2 - MIT (United States)

Abstract: Motivated by ride-hailing marketplaces, we present a model of two-sided matching platforms in a duopoly market with spatial frictions. Both platforms face disjoint streams of rider requests, but they share the same resources on the supply side, i.e., drivers multi-home. Due to spatial frictions, supply efficiency requires an admission control to maintain a buffer of available drivers to balance matching distances with other supply costs. We consider two types of platform admissions controls, "protection" and "distance thresholds", under a large-market stochastic model. For both controls, we find that a duopoly equilibrium entails at least one platform "undercutting", by accepting all requests, to gain market share. When both platforms undercut, the supply buffer is not maintained and inefficiencies ensue. In contrast, when just one platform undercuts, then the duopoly is asymptotically as efficient as a monopolist. For the protection thresholds, we provide a classifier of market characteristics that delineates the two regimes and allows us to characterize the prices of anarchy and stability in these settings.

Minabutdinov Aleksei: Misfortunes Never Come Singly: Managing the Risk of Chain Disasters

Minabutdinov Aleksei[1], Bretschger Lucas, Renoir Clément, Brausmann Alexandra

1 - Department of Management, Technology, and Economics [ETH Zürich] (Switzerland)

Abstract: The recent COVID-19 pandemic has confirmed two facts: First, a large initial disaster - the pandemic - is capable of triggering a whole chain of secondary economy-wide disasters, like a domino effect. Second, governments around the world implemented disaster-mitigation efforts only after the gravity of the pandemic became obvious, as opposed to taking precautionary measures ahead of time. In this paper, we question the optimality of such a "reactive" approach to disaster management by studying disaster-prevention and growth policies in an environment where a primary shock to the economy may provoke future calamities through contagion effects. We develop a novel dynamic stochastic framework, where disaster arrivals are modeled via the Hawkes process. We derive analytical solutions and show that the optimal policy consists of devoting a stochastic fraction of output to disaster-mitigation. The mitigation propensity is an increasing function of the Hawkes intensity and essentially tracks disaster arrivals. The latter implies that the policy is indeed reactive. This result is in contrast with the existing literature, which does not take into account the possibility of contagion and therefore finds a constant mitigation propensity to be optimal.

Chapter 4

Abstracts of posters

4.1 Poster session I

4.1.1 Bharti Bharti: Spatial appointment scheduling in a random environment

Bharti Bharti[1], Mandjes Michel[1], Bekker René[2]

1 - Korteweg-de Vries Institute for Mathematics (Netherlands)
2 - Vrije Universiteit, Amsterdam (Netherlands)

Abstract: In this work, we investigate single server delivery problems using a stochastic approach. In our model we distinguish two factors; namely given the number of delivery locations, we wish to decide both on an appropriate route, as well as on appropriate delivery times. Here both the travel times between the locations and the service times at the locations are random variables. The aim is to optimize the objective function determined (which contains the nodes' idle- and waiting times) by a given route and inter-appointment times. The overall process is challenging since it is difficult to obtain a closed-form expression for the objective function. In this work, therefore, we follow three approaches: (1) an accurate (but slow) approximation approach using phase-type distributions that are mixtures of Erlang distributions and hyperexponential distributions, (2) a simplification of the objective function based on a heavy-traffic approximations, and (3) a 'lag-based' approach in which some 'memory' is ignored so as to avoid the multiplication of large matrices. For the routing component, which in itself is an NP-hard problem, classical heuristic methods which destroy and repair edges have been tested and implemented. Empirical results illustrate that the combination of these techniques lead to accurate and efficient results.

4.1.2 Brodu Virgile: An individual-based model for food webs: how can simple biological constraints enforce precise allometric coefficients.

Billiard Sylvain[1], Brodu Virgile[2], Champagnat Nicolas[2,3], Fritsch Coralie[2,3]

1 - Évolution, Écologie et Paléontologie (Evo-Eco-Paleo) - UMR 8198 (France)
2 - Institut Élie Cartan de Lorraine (France)
3 - Inria Nancy - Grand Est (France)

Abstract: For a few decades, a lot of papers in the field of evolutionary ecology have stood for a very particular relationship between body mass or length and metabolism, across several orders of magnitude of species size. It is now well-known as an *allometric* relationship, or simply *allometry*, presented in the form $B \propto M^\alpha$, where B is the metabolism, M the mass and α the so-called allometric coefficient. Even if these allometries seem to be a key ingredient for modelling the behavior of species, their mass and ecological features, it is still mainly justified by experimental approaches.

We design a simple individual-based model, structured by a trait called *energy*. We study species characterized by a typical energy, consuming a resource maintained at a fixed level. The whole system is ruled by two sorts of mechanisms: random jumps conforming to births and deaths in the population; and a continuous and deterministic evolution of individual energies over time. Importantly, we will allow ourselves to consider the broad range of possible characteristic energy, either really small or high, and thus consider very different species and not one in particular. It means that we design a model focusing on interspecific allometries, rather than intraspecific dynamics with a fixed typical energy.

First, we enforce with basic probabilistic tools sharp values for the allometric coefficients involved in our model, showing that this allometric setting is very restrictive. This idea is also supported by a more complex property of asymptotic pseudotrajectory verified by our underlying stochastic process, leading to precise asymptotic behaviors.

4.1.3 Cocomello Juniper: Tractable Approximations of non-Markovian S(E)IR dynamics on locally tree-like graphs

Cocomello Juniper[1], Ramanan Kavita[2]

1 - Brown University (United States)
2 - Brown University (United States)

Abstract: The Susceptible-Infected-Recovered (SIR) process is a widely used mathematical model of infectious diseases. We develop asymptotically exact and computationally tractable approximations of the ensemble behavior of the SIR process on a large class of sparse graphs that converge locally to a tree (which includes uniform d-regular graphs, sparse Erdos-Rényi and the Configuration Model). The model we consider allows for age-dependent recovery rates, and is therefore non-Markovian. We also generalize our result to the SEIR model.

4.1.4 Deep Vikas: Optimal inference of ATE in A/B testing

Deep Vikas[1], Juneja Sandeep[2], Bassamboo Achal[1]

1 - Northwestern University (United States)
2 - Tata Institute of Fundamental Research (India)

Abstract: A/B Testing is a popular statistical test used in practice on various online platforms for website optimization, product offerings etc. For example, if a firm wants to asses some performance metric of the existing version of their product, named A, against a new version of product B, they conduct A/B testing. Most importantly, firms would like to do valid statistical inference about the effect size, i.e., average treatment effect (ATE) of a variation from the existing version of the product to a new version of the product. In this paper, Our objective is to arrive at a confidence interval (CI) of a pre-specified width ε, and with a coverage guarantee that the ATE lies in the CI with probability at least 1δ for pre-specified $\delta \in (0, 1)$. The novelty of our analysis is the derivation of the asymptotic lower bound on the number of samples required by any algorithm to construct a CI of $\varepsilon-$width with the coverage guarantee for fixed $\varepsilon > 0$ in the regime when $\delta \to 0$, and the construction of an algorithm that, under mild assumptions, matches the lower bound. We utilize bandit algorithms (adaptive algorithms) for reliable inference methods based on the resulting adaptively-collected data. Surprisingly, we get that there are two optimal adaptive experimentation policies and there is value to use adaptive experimental policies over randomized policies.

4.1.5 Deka Prabhanka: PageRank Nibble on the sparse directed stochastic block model

Banerjee Sayan[1], Deka Prabhanka[1], Olvera-Cravioto Mariana[1]

1 - University of North Carolina, Chapel Hill (United States)

Abstract: We present results on community recovery in a sparse, directed Stochastic Block Model (dSBM) where the identities of a positive proportion of seeds in one community is known. Our method is based on the PageRank Nibble algorithm, which uses personalized PageRank and a simple threshold rule to label vertices in the same community as the seeds. This method leverages information about the community label of some vertices, a situation often encountered in practice. Moreover, since PageRank is a local algorithm that can be computed in a distributed and asynchronous fashion, it is highly efficient even in large graphs. Our estimates on the probability of misclassification are derived from a tractable expression for the limiting PageRank distribution. This distribution is in turn obtained by using a new local weak limit for a very general class of inhomogeneous random digraphs, which allows us to characterize it via a system of distributional fixed point equations.

4.1.6 Gupta Hritika: Expected Number of Call Abandonments in a Call Centre

Gupta Hritika[1,2]

1 - University of Melbourne (Australia)
2 - ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA) (Australia)

Abstract: Most call centres are a mix of dedicated servers and multi-skilled servers. The dedicated servers are trained to solve one type of customer query and the multi-skilled servers are trained to solve more than one type of customer query. In such call centres, there are generally multiple ways of allocating

an incoming call to a server. Some allocation methods might perform better than others. One of the measures that can be used to compare the performance of these methods is the total number of calls that were abandoned in a given time interval when using that method to allocate the calls. Assessing call abandonments is crucial for a call centre as they reflect an unsatisfactory customer experience and loss of revenue in some cases. Our aim is to calculate the expected number of call abandonments during a future time interval given the number of callers currently present in the system for any given allocation method. We can then use it to find an optimal call allocation policy that minimises the expected number of call abandonments in a call centre. We will make use of Laplace transforms in order to achieve this. This method can also be used to calculate other performance measures related to call centres.

4.1.7 Hanukov Gabi: Postponed Complementary Services

Hanukov Gabi[1]

1 - Ariel University (Israel)

Abstract: Studies on service systems and queueing theory classically assume that a customer who obtains service must be present in the system during the whole time the service is provided. However, as time passed, many service systems realized that part of the service can be performed while the customer is absent. The customer can use this time efficiently outside the system. Adopting this potentially favorable approach, we study a Markovian service system in which the service can be split into two phases: an opening service (OS), which can be provided only when the customer is present, and a complementary service (CS), which can be carried out in the absence of the customer. To improve the system's overall performance, after providing the OS, the server can postpone executing the CS to time intervals in which the system is empty of customers. The prepared OSs are stored in a designated storage facility. Once the system becomes empty, the server executes the CS for the stored OSs, one at a time. We consider two unbounded dimensions system involving a multivariate probability generating function and derive closed-form expressions for the system's steady-state probabilities and its various performance measures. An economic analysis compares the described model with a corresponding limited OS capacity model.

4.1.8 Hillas Lisa: Designing Service Menus for Bipartite Qeueing Systems

Caldentey Rene[1], Gupta Varun[1], Hillas Lisa[1]

1 - Booth School of Business [Chicago] (United States)

Abstract: We consider a multi-class multi-server queueing system, in which customers of different types have heterogeneous preferences over the many servers available. The goal of the service provider is to design a menu of service classes that balances two competing objectives: (1) maximize customers' average matching reward and (2) minimize customers' average waiting time. A service class corresponds to a single queue served by a subset of servers under a FCFS-ALIS service discipline. Customers act as rational self-interested utility maximizing agents when choosing which service class to join. In particular, they join the class that maximizes their expected ex-ante net utility, which is given by the difference between the server-dependent service reward they receive minus a disutility based on the mean steady-state waiting time of the service class they join. We study the problem under (conventional) heavy traffic conditions, that is, in the limit as the traffic intensity of the system approaches one from below. For an arbitrary number of servers, we prove that if the service provider only cares about minimizing average delay or maximizing total matching reward then very simple menus are optimal. We also provide Mixed Integer Linear Programming (MILP) formulations for optimizing the delay-reward trade-off within fairly rich and practically relevant families of menus.

4.1.9 Huang Junfei: Capacity Allocation and Scheduling in Two-Stage Service Systems with Multi-Class Customers

Zhong Zhiheng, Cao Ping, Huang Junfei[1], Zhou Sean

1 - The Chinese University of Hong Kong (Hong Kong SAR China)

Abstract: This paper considers a tandem service system and the objective is to design a joint capacity allocation and scheduling policy to minimize the long-run average cost. We develop an iterative procedure to solve the fluid problem, and translate the solution into an easy-to-implement policy for the stochastic system. For two special structures, we derive explicit index-based policies.

4.1.10 Johlke Anika: Control Charts for Statistical Process Control - A Literature Overview

Johlke Anika[1]

1 - Friedrich-Schiller-Universität = Friedrich Schiller University Jena [Jena, Germany] (Germany)

Abstract: Control charts as a tool for statistical process control have been regularly researched and applied for decades. If the underlying assumptions are guaranteed, traditional charts such as the Shewhart, EWMA or CUSUM chart are a reasonable choice for monitoring production or service processes. However, the charts themselves have evolved in many aspects, partly due to the increasing availability of data or the need for more complex monitoring processes, so that this field of research has become extremely broad and diverse. This literature review aims to categorize the different characteristics of control charts, namely general and specific types of control charts, different types of data, multivariate and distribution-free charts, or the economic approach for designing a control chart. Especially, a categorization scheme for different performance metrics is provided. All categories are briefly summarized and representative articles are overviewed, the most influential authors in the field are introduced, and several suggestions for future research are concluded. Unlike other recent literature reviews, this review does not focus on a specific property of control charts, but rather provides a broad overview. Research on this topic has not ceased, in fact it has increased in the past year, making control charts as important as ever.

4.1.11 Khodadadian Sajad: Federated reinforcement learning: Linear speedup under Markovian sampling

Khodadadian Sajad[1], Sharma Pranay, Joshi Gauri, Maguluri Siva Theja

1 - School of Industrial and Systems Engineering [Georgia Tech] (United States)

Abstract: Reinforcement learning (RL) is a sequential decision making paradigm where an agent learns to accomplish certain tasks by interacting with the environment. It is known that RL algorithms are data-intensive, and require a large set of data to train. One way to boost the learning of RL algorithms is to employ multiple agents to collect data. Furthermore, in certain applications such as medical settings, data of the local agents might be sensitive and needs to be kept private. In this talk, we consider a federated RL framework where multiple agents collaboratively learn a global model, without sharing their sensitive individual data and policies. Although having N agents enables the sampling of N times more data, it is not clear if it leads to proportional convergence speedup. We consider federated versions of on-policy TD, off-policy TD and Q-learning, and establish that there is a speedup in learning that is linear in the number of agents. In particular, we show this even in the presence of Markovian noise and multiple local updates. We do this by developing a federated stochastic approximation algorithm with Markovian noise (FedSAM) and establishing linear speedup under a very general framework.

4.1.12 Koopmans Camiel: Optimal service control and observation restoration decisions in Markovian queues with fallible sensors.

Kanavetas Odysseas[1], Koopmans Camiel[1], Spieksma Floske[1]

1 - Mathematical Institute, Leiden University (Netherlands)

Abstract: Markov decision processes (MDPs) have been a source of many techniques and models frequently used nowadays. This talk aims to show how to apply techniques stemming from MDPs to Markovian single-queue models where service rate control is performed with fallible sensors. More specifically, “fallible”means that after a sensor failure the queue cannot be observed until it is repaired. This differs from noisy partially-observable MDPs. In practice, such a model is applicable to situations where sensors break down and their repairs have to be optimized. We search for policies that minimize costs by specifying the service rate to use and by scheduling the repair of the sensor. One of the approaches is based on the fact that cer- tain mathematical properties (e.g., monotonicity and convexity) are preserved under Value Iteration, which allows us to establish certain (threshold) structures in optimal policies.

4.1.13 Kraakman Yanna: Configuration models of random directed hypergraphs

Kraakman Yanna[1], Litvak Nelly[1,2], Stegehuis Clara[1]

1 - University of Twente (Netherlands)

2 - Eindhoven University of Technology [Eindhoven] (Netherlands)

Abstract: Many real-world networks exhibit interactions beyond pairwise connections, which are called hyperedges. What is often overlooked, is that hyperedges can also be directed. For chemical reaction networks for example, the reaction A+B -> C+D is a hyperedge connecting two sets of interacting species. Such networks can be represented by directed hypergraphs, which can be categorized into different classes. The central question we investigate is: how can we generate directed hypergraphs in such a way that they are sampled uniformly? We therefore introduce configuration models that can generate random directed hypergraphs for the different classes, and investigate when they generate these uniformly. Subsequently, we propose a Markov chain Monte Carlo method to sample them. Interestingly, we show that the proposed Markov chain Monte Carlo method generates uniformly distributed samples from the configuration models of certain classes of directed hypergraphs, and not from others.

4.1.14 Lee Junghyun: Communication-Efficient Collaborative Heterogeneous Bandits in Networks

Lee Junghyun[1], Schmid Laura[1], Yun Se-Young[1]

1 - Korea Advanced Institute of Science and Technology (South Korea)

Abstract: The multi-agent multi-armed bandit problem has been studied extensively due to its ubiquity in many real-life applications, such as online recommendation systems and wireless networking. We consider the setting where agents should minimize their group regret while collaborating over a given graph via some communication protocol and where each agent is given a different set of arms. Previous literature on this problem only considered one of the two desired features: agents communicate over a general graph while having the same arm set, or the graph is fully connected while each agent has a different arm set. We first introduce a more general problem setting that encompasses all the desired features. For this novel setting, we first provide theoretical analyses of the standard flooding protocol combined with the UCB algorithm. Then, to mitigate the issue of huge communication costs incurred by flooding, we propose a new protocol called Flooding with Absorption (FWA). We provide theoretical analyses and informal intuitions on the advantages of using FWA over flooding. Lastly, we verify empirically that using FWA leads to significantly lower communication costs despite minimal regret performance loss compared to flooding.

4.1.15 Nagy Oliver: Mixing of fast random walks on dynamic random permutations

Nagy Oliver[1]

1 - Universiteit Leiden (Netherlands)

Abstract: We analyse the mixing profile of a random walk on a dynamic random permutation, focusing on the regime where the walk evolves much faster than the permutation. Two types of dynamics generated by random transpositions are considered: one allows for coagulation of permutation cycles only, the other allows for both coagulation and fragmentation. We show that for both types, in the limit as the length of the permutation tends to infinity and after an appropriate scaling of time, the total variation distance between the current distribution and the uniform distribution drops down in a single jump. This jump is similar to a one-sided cut-off, occurs after a random time, and goes from the value 1 to a value that is a strictly decreasing and deterministic function of the time of the jump. After the jump, the total variation distance follows the aforementioned function down to 0. Joint project with Luca Avena (Leiden U.), Remco van der Hofstad (TU Eindhoven) and Frank den Hollander (Leiden U.)

4.1.16 Nieto Benoît: Estimation of threshold diffusions

Mazzonetto Sara[1], Nieto Benoît[2]

1 - Institut Élie Cartan de Lorraine (France)

2 - Institut Camille Jordan [Villeurbanne] (France)

Abstract: We refer by threshold diffusions to a class of continuous-time Markov process which admits a change of dynamics on a fixed level. We study the (quasi)-maximum likelihood estimation of the drift parameters, for continuous and discrete time observations. Two consistent estimators for the volatility parameters are also proposed. We discuss about the consistency and the speed convergence of these estimators in long time and high frequency for the ergodic case. Based of these results, we propose an application to the Threshold Cox-Ingersoll-Ross (CIR) model.

4.1.17 Qu Yanlin: Contractive Drift Conditions for Rates of Convergence of Markov Chains in Wasserstein Distance

Qu Yanlin[1], Blanchet Jose[1], Glynn Peter[1]

1 - Stanford University (United States)

Abstract: We introduce a condition called "contractive drift" to bound the convergence of Markov chains in Wasserstein distance. Under this condition, we derive computable convergence bounds with different rates, from polynomial to exponential. These bounds are computable in the sense that they only contain fully explicit constants and one-step transition expectations. To enhance practical application, we develop various techniques to verify the contractive drift condition so that computable convergence bounds can be conveniently obtained for queueing models as well as optimization algorithms. Moreover, our bounds have the ability to capture optimal rates of convergence in terms of model parameters such as the traffic intensity in queueing models and the step size in optimization algorithms.

4.1.18 Reid Mirabel: Convergence of the k-cap Process on Graphs with Weight Reciprocity

Kailasa Sameer[1], Moondra Jai[2], Reid Mirabel[2]

1 - University of Michigan (United States)
2 - Georgia Institute of Technology (United States)

Abstract: Artificial neural networks utilizing a k-winners-take-all activation mechanism, or k-cap, have been proposed as a model of excitatory-inhibitory biological networks dominated by strong lateral inhibition. In conjunction with Hebbian plasticity of weights, such networks form the backbone of the assembly calculus, an emerging mathematically tractable theory for formation and manipulation of memories.

In this work, we study the dynamics of random recurrent neural networks employing the k-cap operation, with underlying network connectivity drawn from a chosen graph model. The process leads to a sequence of k-subsets of vertices, called 'active sets'; at each discrete time step, the active set consists of the k vertices with the highest weighted degree sum from the previous active set. We establish the convergence of this process on unweighted undirected graphs and some weighted random graphs. Complementing recent work which analyzes the same process on soft geometric random graphs, our results do not require plasticity of weights, indicating two distinct mechanisms by which activity in biological neural networks may become localized.

4.1.19 Schmocker Philipp: Universal Approximation Property of Random Neural Networks

Neufeld Ariel[1], Schmocker Philipp[1]

1 - Nanyang Technological University (Singapore)

Abstract: In this talk, we study single-hidden-layer feedforward neural networks with randomly initialized weights and biases, which are inspired by the works on extreme learning machines, random feature models, and reservoir computing. After the random initialization, only the linear readout needs to be trained, which can be performed, e.g., by a linear regression. Despite the popularity of this approach in empirical tasks, only little is known about the approximation properties. By considering these "random neural networks" as Banach space-valued random variables, we prove some universal approximation theorems within Bochner spaces. Hereby, the Banach space can be a more general function spaces such as L^p-spaces or Sobolev spaces, where the latter includes the approximation of the derivatives. In addition, we provide some approximation rates for random neural networks, which shows in particular that they overcome the curse of dimensionality on a dimension-scaled neighborhood of the training data. This talk is based on joint work with Ariel Neufeld.

4.1.20 Vande Velde Sylvie: Branching Processes and Approximate Bayesian Computation Applied to the Initiation of Chronic Myeloid Leukaemia

Vande Velde Sylvie[1], Engelbeen Céline[2], Hautphenne Sophie[3]

1 - Université libre de Bruxelles (Belgium)
2 - Brussels Management School [Bruxelles] (Belgium)
3 - University of Melbourne (Australia)

Abstract: Chronic Myeloid Leukemia (CML) is a type of cancer that occurs when there is an exchange of genetic material between chromosomes 9 and 22 in certain blood-forming cells of the bone marrow. This results in the growth and division of leukemia cells that accumulate in the bone marrow and differentiate into more mature cells, which then spread throughout the blood. Previous studies have demonstrated that the formation of normal blood cells can be modelled using a 32-compartment model, where each compartment represents a stage of cell differentiation from stem cells to mature blood cells [1, 2]. Building upon this assumption, we use a special class of branching processes known as Markovian Binary Tree (MBT) [3] to model the development of the number of normal and leukemia cells in the 32 compartments. Our project aims to determine the stage of cell differentiation where the first chromosomal exchange occurred with the highest probability. To achieve this goal, we use real patient data [5] and apply the Approximate Bayesian Computation (ABC) method [4]. This method provides a probability distribution over the compartment where the initial leukemia cell appeared, which enables us to test the correlation between the origin compartment and the treatment response for groups of patients with similar characteristics. References [1] D. Dingli, A. Traulsen, and J. M. Pacheco, Compartmental architecture and dynamics of hematopoiesis, PLOS ONE, vol. 2, pp. 14, 2007. [2] B. Werner, D. Dingli, T. Lenaerts, J. M. Pacheco, and A. Traulsen, Dynamics of mutant cells in hierarchical organized tissues, PLOS Computational Biology, vol. 7, pp. 19, 2011. [3] S. Hautphenne, M. Massaro, and K. Turner, Fitting Markovian binary trees using global and individual demographic data, Theoretical Population Biology, vol. 128, pp.39-50, 2019. [4] M. SunnËŽaker, A.G. Busetto, E. Numminen, J. Corander, M. Foll, C. Dessimoz, Approximate Bayesian Computation, PLoS Comput Biol, vol. 9(1), 2013. [5] I. S. Pagani et al, BCR-ABL1 genomic DNA PCR response kinetics during first-line imatinib treatment of chronic myeloid leukemia, Haematologica, vol. 103(12), pp. 2026-2032, 2018.

4.1.21 Van Der Sar Erica: Optimizing Power Grid Control with Reinforcement Learning

Van Der Sar Erica[1], Bhulai Sandjai[1], Zocca Alessandro[1]

1 - Vrije Universiteit Amsterdam [Amsterdam] (Netherlands)

Abstract: In the past few decades, power transmission networks have changed tremendously, mostly due to a dramatic increase in energy demand and ambitious sustainability goals. In particular, the increasing penetration of renewable energy sources, such as wind and solar, introduces a substantial amount of stochasticity in these already complex networks. As a result, it is becoming more and more difficult for network operators to control power grids and keep them safe and reliable at all times. The only way to achieve this is by using the network flexibility to its fullest extent and thus being able to identify in real-time optimal control actions in a combinatorially very large action space. Multiple reinforcement learning frameworks have been proposed in recent years to assist network operators in the decision-making for real-time operations. In such settings, the RL agent learns how to optimally control the power network using various actions that can change the network topology and thus, indirectly, the power flows. In this talk, we illustrate how existing RL approaches can be improved using Graph Neural Networks (GNN) and additional information, such as renewable and load forecasts.

4.1.22 Wang Xingyu: Eliminating Sharp Minima from SGD with Truncated Heavy-tailed Noise

Wang Xingyu[1], Oh Sewoong[2], Rhee Chang-Han[1]

1 - Northwestern University (United States)

2 - University of Washington (United States)

Abstract: The empirical success of deep learning is often attributed to the stochastic gradient descent (SGD) algorithm's mysterious ability to avoid sharp local minima in the loss landscape, as sharp minima are known to lead to poor generalization. Recently, evidence of heavy-tailed gradient noise was reported in many deep learning tasks, and Simsekli et al. (2019) argued that SGDs are capable of escaping sharp local minima under the presence of such heavy-tailed gradient noise. This suggests a new promising perspective to the mystery, but when heavy-tailed gradient noises are present, the pure form of SGD often fails to train DNNs altogether, and one needs to deploy a variant of SGD, where gradients are truncated above a fixed threshold. In this talk, we analyze the global dynamics of such a variation (often called gradient clipping) and show that they are fundamentally different from those of the pure form of SGD. In fact, this variant attains a strong notion of avoiding sharp minima: it can effectively eliminate sharp local minima entirely from its training trajectory. We prove this by rigorously establishing an Eyring-Kramers-type first exit time formula and metastability analysis for heavy-tailed SGDs: We reveal the large deviations in heavy-tailed SGDs and obtain a sharp characterization of the relationship between the first exit time from a local basin and the ratio between the truncation threshold and the width of

the basin. Moreover, when the objective function satisfies appropriate structural conditions, we prove that the heavy-tailed truncated SGD converges to a continuous-time Markov chain that never visits any sharp minima as the step size decreases. Real data experiments on deep learning confirm our theoretical prediction that heavy-tailed SGDs with gradient clipping find flatter local minima and achieve better generalization.

4.1.23 Zhang Dongni: Epidemic models with manual and digital contact tracing

Zhang Dongni[1], Britton Tom[1]

1 - Stockholm University (Sweden)

Abstract: We start with a Markovian SIR epidemic model in a homogeneous mixing community with a constant diagnosis rate. First, we introduce manual contact tracing by assuming that once an infective is diagnosed (tested positive and isolated), each of her/his contacts is immediately traced and tested independently with some fixed probability. Using large population approximations, we analyzed the early stage of the outbreak when the process of "to-be-traced components" behaves like a branching process. The component and individual reproduction numbers are derived. Then we focus on the more recent digital contact tracing via a tracing app (only app-users can trigger and be traced by digital tracing). And we assume that there is a fixed app-using fraction and that digital tracing occurs instantaneously and recursively. The model with digital tracing is analysed by a two-type branching process relying on a large community, where one type is "app-using components," and another is non-app-users. Further, we investigate the combined preventive effect of manual and digital tracing. This combined model is analysed by a different two-type branching process with both types being the "to-be-traced components" but starting with different "roots". The corresponding reproduction numbers are derived. We conclude that to control the epidemic, we always need to have a large fraction of app-users compared to the fraction of individuals being successfully reached by manual tracing. Another important conclusion is that the combined effect is bigger than the product of two separate preventive effects. Finally, our ongoing work is based on an SEIR epidemic model on a configuration model, where transmission could happen from random contacts (e.g., on a bus), which are usually more easily identified by using the tracing app. Each infective remains infectious for a constant period. Meanwhile, we assume that such an infective is diagnosed and asked for an interview with some probability otherwise, we say that the infective recovers naturally. Once diagnosed, each of her/his infectee neighbours (manual tracing only on the network) is reported with some probability independently. If such reported neighbours are infectious after some delay, they are isolated and said to be traced. If the diagnosed person is also an app-user, all of her/his app-using contacts are traced immediately (digital tracing both on the network and among global contacts). And in this case, we assume that only the diagnosed individuals can trigger contact tracing.

4.2 Poster session II

4.2.1 Abdelhakmi Anas: A multi-period Black-Litterman model with desynchronized expert views

E.b. Lim Andrew[1], Abdelhakmi Anas[2]

1 - Department of Analytics and Operations, Department of Finance, and Institute of Operations Research and Analytics, National University of Singapore (Singapore)

2 - Institute for Operations Research and Analytics, National University of Singapore (Singapore)

Abstract: The Black-Litterman model is a framework for incorporating forward-looking expert views in a portfolio optimization problem. Existing work focuses almost exclusively on single-period problems where the horizon of expert forecasts matches that of the investor. We consider a multi-period generalization of the Black-Litterman model where the investor trades dynamically, and the horizon of expert views may differ from that of the investor. By exploiting an underlying graphical structure relating the asset price process and the forward-looking expert views, we derive the conditional distribution of asset returns when the price process is assumed to be geometric Brownian motion. The new price process is an affine factor model where an adjustment of the unconditional log-price process is a factor, and is closely related to the smoothed estimates of future asset prices. We derive an explicit expression for the optimal dynamic investment policy and analyze the hedging demand associated with the new covariate. More generally, this application shows that smoothed estimates of underlying random processes appear naturally as state variables when there are forward looking forecasts.

4.2.2 Allmeier Sebastian: Bias of Multiscale Mean Field Models and Averaging Methods

Allmeier Sebastian[1], Gast Nicolas[1]

1 - INRIA Grenoble-Rhône Alpes (France)

Abstract: The mean field approximation is a powerful technique which has been used in many settings to study large-scale stochastic systems. In the case of two-timescale systems, the approximation is obtained by a combination of scaling arguments and the use of the averaging principle. In this talk, I will present how to analyze the approximation error of this average' mean field approach for a two-timescale model (X, Y), where the slow component X describes a population of interacting particles which is fully coupled with a rapidly changing environment Y.

We show that under relatively mild conditions the 'average' mean-field approximation has a bias of order $O(1/N)$, where N is a scaling factor. We also derive a steady state bias correction term that reduces the error to order $O(1/N^2)$. This refined 'average' mean-field approximation allows the computation of an accurate approximation even for small scaling factors, i.e., $N \approx 10$–50. We illustrate the developed framework and accuracy results through an application to a random access CSMA model.

4.2.3 Amanihamedani Alireza: Spatial Matching under Resource Competition

Amanihamedani Alireza[1], Freund Daniel[2], Aouad Ali[1]

1 - London Business School (United Kingdom)
2 - MIT (United States)

Abstract: Motivated by ride-hailing marketplaces, we present a model of two-sided matching platforms in a duopoly market with spatial frictions. Both platforms face disjoint streams of rider requests, but they share the same resources on the supply side, i.e., drivers multi-home. Due to spatial frictions, supply efficiency requires an admission control to maintain a buffer of available drivers to balance matching distances with other supply costs. We consider two types of platform admissions controls, "protection" and "distance thresholds", under a large-market stochastic model. For both controls, we find that a duopoly equilibrium entails at least one platform "undercutting", by accepting all requests, to gain market share. When both platforms undercut, the supply buffer is not maintained and inefficiencies ensue. In contrast, when just one platform undercuts, then the duopoly is asymptotically as efficient as a monopolist. For the protection thresholds, we provide a classifier of market characteristics that delineates the two regimes and allows us to characterize the prices of anarchy and stability in these settings.

4.2.4 Bhambay Sanidhay: Asymptotic Optimality of Speed-Aware JSQ for Heterogeneous Service Systems

Bhambay Sanidhay[1], Mukhopadhyay Arpan

1 - University of Warwick (United Kingdom)

Abstract: The Join-the-Shortest-Queue (JSQ) load-balancing scheme is known to minimise the average delay of jobs in homogeneous systems consisting of identical servers. However, it performs poorly in heterogeneous systems where servers have different processing rates. Finding a delay optimal scheme remains an open problem for heterogeneous systems. In this work, we consider a speed-aware version of the JSQ scheme for heterogeneous systems and show that it achieves delay optimality in the fluid limit. One of the key issues in establishing this optimality result for heterogeneous systems is to show that the sequence of steady-state distributions indexed by the system size is tight in an appropriately defined space. The usual technique for showing tightness by coupling with a suitably defined dominant system does not work for heterogeneous systems. To prove tightness, we devise a new technique that uses the drift of exponential Lyapunov functions. Using the non-negativity of the drift, we show that the stationary queue length distribution has an exponentially decaying tail - a fact we use to prove tightness. Another technical difficulty arises due to the complexity of the underlying state-space and the separation of two time-scales in the fluid limit. Due to these factors, the fluidlimit turns out to be a function of the invariant distribution of a multi-dimensional Markov chain which is hard to characterise. By using some properties of this invariant distribution and using the monotonicity of the system, we show that the fluid limit is has a unique and globally attractive fixed point.

4.2.5 Blomerus Wessel: Structured learning of the optimal replacement rule beyond Markov settings

Blomerus Wessel[1], Kapodistria Stella[1], Borst Sem[1]

1 - Eindhoven University of Technology [Eindhoven] (Netherlands)

Abstract: In Markov decision processes, the archetypical point of departure is the Bellman optimality equations. For such problems, these equations rely on the underlying stochastic model being a Markov chain. However, the analysis and modelling of real data rarely produces a simple Markov chain model. Instead, it tends to result in history-dependent models (e.g., autoregressive models) or models with partial information. This necessitates two extensions of the current solution framework. Firstly, that the theoretical foundation of the Bellman optimality equations is extended non-Markovian settings. Secondly, that the learning of the underlying model is integrated with the decision analysis. In this presentation we demonstrate how one can overcome these two challenges. Specifically, we focus on the replacement problem of an asset such as a technological device or component, and observe its condition, with the objective to determine when to optimally replace the asset so as to minimize cost. The key assumption is that the underlying condition (stochastically) renews when the asset is replaced by a good-as-new asset. In this case, the optimal replacement rule is much more complicated than in the Markovian case. Despite this complication, we present an algorithm which requires a relatively small number of iterations to converge to the optimal replacement rule. Furthermore, we show that this framework can be extended to the case where the underlying model is learned in real time whilst also learning the optimal replacement rule. Lastly, the proposed algorithm is also applicable for optimal stopping problems with a non-Markovian structure.

4.2.6 Briend Simon: Finding adam under polynomial connectivity assumption

Briend Simon[1,2]

1 - Universitat Pompeu Fabra [Barcelona] (Spain)
2 - Université Paris-Saclay (France)

Abstract: We study the problem of root finding in growing random graphs under general hypothesis on connection probabilities. A simple algorithm is introduced, that runs in polynomial time and that outputs a set of vertices containing the original vertex with high probability. We present its performance in the particular case of a "general graph" version of the uniform attachment tree. Unlike in the tree, each new vertex connects to several ancestors (and not only one) chosen uniformly at random. We then explain why this method can be extended to a wide range of growing graphs models and give other examples of such models.

4.2.7 Cheikhi David: On the Statistical Benefits of Temporal Difference Learning

Cheikhi David[1], Daniel Russo

1 - Columbia University [New York] (United States)

Abstract: Given a dataset on actions and resulting long-term rewards, a direct estimation approach fits value functions that minimize prediction error on the training data. Temporal difference learning (TD) methods instead fit value functions by minimizing the degree of temporal inconsistency between estimates made at successive time-steps. Focusing on finite state Markov chains, we provide a crisp asymptotic theory of the statistical advantages of this approach. First, we show that an intuitive inverse trajectory pooling coefficient completely characterizes the percent reduction in mean-squared error of value estimates. Depending on problem structure, the reduction could be enormous or nonexistent. Next, we prove that there can be dramatic improvements in estimates of the difference in value-to-go for two states: TD's errors are bounded in terms of a novel measure - the problem's trajectory crossing time - which can be much smaller than the problem's time horizon.

4.2.8 Corujo Rodriguez Josué: The standard augmented multiplicative coalescent revisited

Corujo Rodriguez Josué[1]

1 - Institut de Recherche Mathématique Avancée (France)

Abstract: The Erd ́os-Rényi random graph is a classic random graph model where vertices are connected with a certain probability independently among them. We consider a continuous-time version of this model, where all vertices have the same mass and any edge appears with a rate proportional to the product of the sizes of the vertices. In addition, we do not restrict the graph to be simple, i.e. we allow the existence of multi-edges and self-loops. It is well-known that the size of the connected component of this graph follows the standard multiplicative coalescent dynamics. Besides, when accounting for the additional information consisting in the number of surplus edges, the resulting process follows the so-called standard augmented multiplicative coalescent dynamic. In this talk, we will explain how using a breadth-first exploration process we can prove in a simple way the scaling limit of the standard augmented multiplicative coalescent. Even if this result is already known, our methods are simpler and more direct than those encounters in the bibliography. Furthermore, this framework could potentially be extended to the study of general non-standard augmented multiplicative coalescent as scaling limits of inhomogeneous random graphs.

This is a work in collaboration with Vlada Limic (CNRS, UniversitÃĺ de Strasbourg).

4.2.9 Ertel Sebastian: Ensemble Kalman-Bucy filters for finite and infinite dimensional signals

Ertel Sebastian[1], Stannat Wilhelm[1]

1 - Technische Universität Berlin (Germany)

Abstract: In this talk we first derive a (deterministic) mean-field Ensemble Kalman–Bucy filter (EnKBF) for finite dimensional, nonlinear, continuous time filtering problems as a constant gain approximation to the Feedback Particle Filter. Other well known versions of the EnKBF are then obtained by replacing Brownian motions with deterministic inflation terms and vice versa. We then prove the well posedness of this class of mean-field equations via a combination of a fixed point and a partial stopping argument. Next we investigate the mean-field EnKBF when the signal is given by a nonlinear SPDE, i.e. for infinite dimensional signals. In this case proving the well posedness requires an adaptation of the fixed point argument, which makes use of the robustness of the law of total variance with respect to the observation function. Finally, if time permits, we discuss the analysis of the approximating particle system and a propagation of chaos result.

4.2.10 Jhunjhunwala Prakirt: Join-the-Shortest Queue with Abandonment: Critically Loaded and Heavily Overloaded Regimes

Jhunjhunwala Prakirt[1]

1 - Georiga Institute of Technology (United States)

Abstract: We consider a Load Balancing system operating under the well-known Join-the-Shortest Queue (JSQ) policy, where jobs/customers are impatient, and abandon if they do not receive service after some (random) time. In this setting, we study the effect of these abandonments on the steady-state queue length behavior. In particular, we characterize the 'distribution of appropriately centered and scaled steady-state queue length' (hereafter referred to as limiting distribution) as the abandonment rate goes to zero. Depending on the arrival, service, and abandonment rates, we consider three different regimes of operation, each yielding a different steady-state distribution. The first regime is when the system is underloaded, where we show that abandonments do not affect the limiting distribution. As a result, JSQ with Abandonment (JSQ-A) asymptotically behaves like JSQ without abandonments, for which the limiting distribution is known to be exponential. As the load increases, we move to the second regime, where abandonments become significant, and there is a critical balance between the arrivals/services and the abandonments. As such, the system undergoes a phase transition and the limiting distribution changes from exponential to a truncated normal. As the load increases even further, the system enters the heavily overloaded regime and the queue lengths become very large. In this case, we show that the (centered) steady state queue length distribution converges to a normal distribution. Our results use the Transform method to establish a one-dimensional differential equation in terms of the MGF (Moment Generating Function) of the steady-state queue lengths, which can be solved to get the limiting distribution.

4.2.11 Joshi Purva: Performance Evaluation of Intersection Access Mechanisms for Autonomous Vehicles

Joshi Purva[1], Boon Marko[1], Borst Sem[1]

1 - Department of Mathematics and Computer Science, Eindhoven University of Technology (Netherlands)

Abstract: The anticipated launch of fully autonomous vehicles presents an opportunity to develop and implement novel traffic management systems, such as for urban intersections. Platoon-forming algorithms, in which vehicles are grouped together with short inter-vehicular distances just before arriving at an intersection at high speed, seem promising from a capacity-improving standpoint. In this talk, we present a performance evaluation framework which not only captures the intersection access dynamics via a queueing model (or, more specifically, a polling model) with multiple customer types, but also explicitly accounts for the joint optimisation of vehicle trajectories. We focus on deriving computationally fast and interpretable closed-form expressions for safe and efficient vehicle trajectories during the process of platoon formation, and show that these closed-form trajectories are equivalent to those obtained via the joint optimisation process. Additionally, we conduct a numerical study to obtain approximations for the capacity of an intersection under the platoon-forming framework.

4.2.12 Kalvit Anand: Diffusion limits of multi-armed bandit experiments under optimism-based policies

Kalvit Anand[1], Zeevi Assaf[1]

1 - Columbia University (United States)

Abstract: Our work provides new results on the arm-sampling behavior of the classical UCB family of optimism-based multi-armed bandit algorithms, leading to several important insights. Among these, it is shown that arm-sampling rates under UCB are asymptotically deterministic, regardless of the problem complexity. This discovery facilitates new sharp asymptotic characterizations that among other things, reveal profound distinctions between UCB and Thompson Sampling such as an “incomplete learning" phenomenon characteristic of the latter.

4.2.13 Karunarathne Wathsala: Scheduling customers in the presence of random customers into stochastic server systems

Karunarathne Wathsala[1], Taylor Peter[1], Fackrell Mark[1]

1 - University of Melbourne (Australia)

Abstract: We consider an appointment scheduling problem in the presence of random arrivals into a queueing system with stochastic service times. While there have been discussions about scheduling appointments in stochastic server systems in the literature, little is known about managing appointments in the presence of random arrivals. As different queueing systems have disparate objectives to be achieved, we first consider a multi-objective system where the objectives are to maximise expected revenue, minimise the total expected customer waiting cost, and minimise the total expected server idle cost. We formulate a mathematical model to derive the optimum schedule for a fixed number of customers. Later, we study the behavior of optimum inter-arrival times and optimum objective function values for individual objectives and different combinations of objectives. We assume that the service times are exponentially distributed, and the random customers arrive according to a Poisson process. Further, we assume that the scheduled customers get non-preemptive priority over random customers, in which the expected waiting times of scheduled customers can be affected by random customers. The problem is computationally expensive and was solved using a meta-heuristic search algorithm named covariance matrix adaptation evolutionary strategy on the University of Melbourne high-performance computing cluster. We show that the system's efficiency can be improved by considering random customer arrivals in scheduling. We believe that the presented scheduling model and analysis lay a foundation for the problem of static scheduling for scheduling customers in the presence of random customers into a stochastic server system and provide better insights for relevant enterprises.

4.2.14 Lee Seunghyun: Fluctuations in random field Ising models

Lee Seunghyun[1], Deb Nabarun[2], Mukherjee Sumit[1]

1 - Department of Statistics, Columbia University (United States)
2 - University of British Columbia (Canada)

Abstract: The Ising model was originally proposed to explain sharp changes in magnetization for heated magnets, making it essential to examine fluctuations in the average magnetization across different temperatures. However, previous studies on this topic often assume a specific underlying graph, and the fluctuations have only been examined on complete graphs for random field Ising models. In this work, we answer this question for a broader range of models. We present both quenched and annealed central limit theorems for random field Ising models on approximately regular graphs, where the average degree is larger than $\sqrt{n}$. Notably, our findings suggest that fluctuations at criticality differ from those

in the Ising model without a random field, both in terms of the magnitude and the resulting limiting distribution. Our results can be applied to various examples in the litterature, including Erdös-Rényi graphs and block-spin models.

4.2.15 Loubaton Rodolphe: Modelisation of dynamic gene regulatory networks and prediction of biological experiments modifying genes in cancerous cells.

Loubaton Rodolphe[1], Champagnat Nicolas[1,2], Vallois Pierre[3], Vallat Laurent[4]

1 - Institut Élie Cartan de Lorraine (France)
2 - Inria Nancy - Grand Est (France)
3 - Institut Élie Cartan de Lorraine (France)
4 - Interface de Recherche Fondamentale et Appliquée en Cancérologie (France)

Abstract: The cells of the human organism are in constant contact with their environment and changes in the environment induce adapted responses from the cells. These cellular responses take place through the expression of thousands of genes that are activated in a concerted manner and these interactions are represented as a gene regulatory network. In the case of cancer cells, there are abnormalities in the expression of certain genes that alter the normal functioning of these gene and protein networks and thus of the gene programmes. Gene programmes correspond to the switching on, control and switching off of the activity of sets of genes as part of the orderly performance of a biological operation that requires many molecular players. These abnormalities, which are responsible for the malfunctioning of the gene programs that control cell division, lead to uncontrolled cell proliferation. The general project of this thesis, which is in line with the work of Laurent Vallat et al. [2, 3], con- sists of modelling the functioning of these gene programmes using temporal data on the expression of genes and proteins. Such a model will make it possible to identify the genes whose expression modulation will make it possible to improve cellular behaviour for therapeutic purposes. Laurent Vallat and his team collected healthy and cancerous cells from control subjects and pa- tients. They then measured, at different times, the expression of genes by RNA-seq (measurement of the number of RNA copies of more than 20 000 genes) and the expression of proteins by mass spectrometry (quantity of about 5 000 proteins) after modulation of the cellular environment. This modulation is achieved by stimulation of a membrane receptor. First of all, I worked on a code (R software) that allows a pre-analysis of an RNA-seq data set. This pre-analysis, based on the R package DESeq2 [1], allows the selection of a subset of genes "of interest" (about 1000 genes) among all the genes (about 20 000). I then built a parametric Gaussian statistical model that allows to account for the dynamics of gene expressions using parameters describing, among others, the interactions between these "genes of interest". The statistical inference of the model parameters combines steps of penalized linear regressions (lasso) and regressions with positivity constraints and constraints on the sum of the coefficients. I then studied the modelling of biological experiments modifying genes. In practice, this mainly involves the inhibition of gene expression (silencing). We have developed a so-called "mechanistic" method and a second so-called "conditional" method that make it possible to construct two new models associated with a biological intervention. Finally, I developed a code for inference and the two prediction models

4.2.16 Marek Oheim: Risk Class Management in Life Insurance Markets

Marek Oheim[1]

1 - RPTU Kaiserslautern-Landau (Germany)

Abstract: Grouping customers of different risk types into contracts is important for the stability and the robustness of an insurance market. In order to analyze the effects of these risk coalitions, we set up a market model which is inspired by a paper of Sass and Seifried (2014). Afterwards, we provide studies on the equilibrium premiums of classic life insurance products in monopolistic and competitive insurance markets. In our model we can show that the market specifications have a high impact on the optimal choice of different contracts in the insurance market. Regarding markets with capacity constraints serve as a window for a deeper understanding of the model. Further extensions of the model and its applications to other insurance types complete the talk.

References:

J. Sass, F.T. Seifried (2014). Insurance markets and unisex tariffs: Is the European Court of Justice improving or destroying welfare? Scandinavian Actuarial Journal. 2014, 228-254.

4.2.17 Miguelez Fernando: Price of Anarchy in an Age-aware multi-source system

Miguelez Fernando[1], Doncel Josu[2], Ayesta Urtzi[2,3,4,5]

1 - Public University of Navarre, UPNA (Spain)
2 - University of the Basque Country (Spain)
3 - Ikerbasque - Basque Foundation for Science (Spain)
4 - Institut de recherche en informatique de Toulouse (France)
5 - Institut National Polytechnique (Toulouse) (France)

Abstract: In this article we define a system in which K sources generate updates following a Poisson process that are routed to a monitor through a single channel with a hard-preemption policy and exponential service times. By means of the Price of Anarchy we compare the performance of a competing setting versus a cooperative one. We assume an evaluation function that is the combination of an economic expense for the server utilisation and the Average Age of Information, a metric that accounts for the freshness of the last information available at the monitor. In our main result we show that the inneficiency due to the competition between sources is strongly dependent on the disparity among service rates, and some numerical experiments support our conclusions.

4.2.18 Pakyari Reza: Spacings-Based Goodness-of-Fit Testing

Pakyari Reza[1]

1 - Qatar University (Qatar)

Abstract: One of the important steps before performing any parametric statistical analysis is to conduct a formal goodness-of-fit (GOF) testing. In this talk, we introduce somehow less known GOF testing procedure based on spacings which has been shown to have larger power in some cases in compare to the conventional Empirical Distribution Function (EDF)-based GOF methods. Indeed, GOF testing based on spacings cover a large class of tests including the Greenwood's statistic, the Moran's statistic and the Quesenberry and Miller's statistic. GOF tests based on higher order spacings and some suggestions for new testing methodologies for the case of censored data will also be discussed.

4.2.19 Paraggio Paola: A perturbed Bertalanffy-Richards growth: deterministic and stochastic models

Paraggio Paola[1], Di Crescenzo Antonio[1], Torres-Ruiz Francisco[2]

1 - Department of Mathematics, University of Salerno (Italy)
2 - Department of Statistics and Operational Research, University of Granada (Spain)

Abstract: In [6] Richards introduced a curve suitable to describe asymmetrical growth patterns. Such curve represents a generalization of the classical logistic function which is characterized by a symmetric behavior with respect to the inflection point (e.g. [1]) Specifically, in the logistic case, the ratio between the carrying capacity and the value of the curve at the inflection point is equal to $1/2$ whereas in the Richards case it is equal to $(1+1/q)^q$. Hence, the Richards curve seems to be more reasonable to describe phenomena in which the carrying capacity is not necessarily twice the value at the inflection point. In this work, we propose a modified Richards-type curve by introducing a perturbation in the growth rate starting at a critical time instant t^*. The definition of such generalized model is stimulated by several real situations in which the growth rate may be modified by external factors as soon as the curve describing the specific evolution crosses a critical threshold. Think about, for example, the case of oil production of a country: when the quantity of produced oil is lower than a fixed threshold, the government may decide to support new explorations to increase the amount of the production. The modified Richards-type curve is analyzed both from a deterministic and a stochastic point of view. The modification of the growth rate is chosen so that (i) the properties of the classical Richards curve are preserved and (ii) the new growth rate is greater than the previous one. We provide explicit expressions of the quantities of interest, such as the inflection point, the carrying capacity, the lag time and the maximum specific growth rate. Furthermore, two stochastic counterparts of the model are considered. More in detail, we define two special birth-death processes (similar to the one proposed in [4]) and a lognormal diffusion process (following the strategy of [2], [3] and [8]). In both cases, the mean of the stochastic process is a modified Richards curve. The following problems are also addressed: - finding the maximum likelihood estimates of the parameters of the diffusion process, - investigating the critical time t^* when it is regarded as a first-passage-time. In particular, the maximization of the likelihood function is performed by means of two different heuristic algorithms: Simulated Annealing (as done in [3]) and Ant Lion Optimizer (cf. [5]). Whereas, the R package fptdApprox (see [7] and [9]) is used to find an approximation of the critical time t^*. Finally, a simulation study and an application to real data are also presented. This contribution is based on joint works with A. Di Crescenzo, F. Torres-Ruiz and P. Román-Román. Bibliography [1] Di Crescenzo A., Paraggio P. (2019) Logistic growth described by birth-death and diffusion processes. Mathematics 7 (6) 489. doi: https://doi.org/10.3390/math7060489 [2] Di Crescenzo A., Paraggio P., Román-Román P., Torres-Ruiz F. (2021) Applications of the multi-sigmoidal deterministic and stochastic logistic models for plant

dynamics. Applied Mathematical Modelling, 92, 884–904. https://doi.org/10.1016/j.apm.2020.11.046 [3] Di Crescenzo A., Paraggio P., Román-Román P., Torres-Ruiz F. (2022) Statistical analysis and first-passage-time applications of a lognormal diffusion process with multi-sigmoidal logistic mean. Stat Papers (online first) https://doi.org/10.1007/s00362-022-01349-1 [4] Majee S., Jana S., Khatua A., Kar T. K. (2022) Growth of single species population: a novel approach. In: Nonlinear Dynamics and Applications. Proceedings of the ICNDA 2022. pp. 907-915. Springer Proceedings in Complexity. Springer, Cham. doi: https://doi.org/10.1007/978-3-030-99792-276 [5] Mirjalili S. (2015) The ant lion optimizer. Advances in Engineering Software, 83,80–98. doi: https://doi.org/10.1016/j.advengsoft.2015.01.010 [6] Richards F. J. (1959) A flexible growth function for empirical use. J Exp Bot 10 (2) 290-300. doi:10.1093/jxb/10.2.290. [7] Román-Román P., Serrano-Pérez J.J., Torres-Ruiz F. (2008) First-passage-time location function: application to determine first-passage.time densities in diffusion processes. Comput. Stat. Data Anal., 52, 4132–4146. doi: https://doi.org/10.1016/j.csda.2008.01.017 [8] Román-Román P., Serrano-Pérez J. J., Torres-Ruiz F. (2018) Some notes about inference for the lognormal diffusion process with exogenous factors. Mathematics 6 85 doi: https://doi.org/10.3390/math6050085 [9] Román-Román P., Serrano-Pérez J.J.M., Torres-Ruiz F. fptdApprox: Approximation of First-Passage-Time Densities for Diffusion Processes, version 2.2. https://cran.r-project.org/web/packages/fptdApprox/

4.2.20 Sarkovic Andjela: Phase transition for random walks on graphs with added weighted random matching

Baran Zsuzsanna[2], Hermon Jonathan[1], Šarković Anđela[2], Sousi Perla[2]

1 - University of British Columbia (Canada)

2 - University of Cambridge (United Kingdom)

Abstract: For a finite graph $G = (V, E)$, we let G^* be the random graph obtained by placing edges of weight ε between pairs of vertices of random perfect matching. We establish a phase transition in the occurrence of cutoff of a weighted random walk on G^* in terms of the weight ε for graphs of polynomial volume growth and for expanders. Joint work with Zsuzsanna Baran, Jonathan Hermon and Perla Sousi.

4.2.21 Shadmi Yonatan: Importance sampling in high dimensions

Shadmi Yonatan[1], Simatos Florian[2], Beh Jason

1 - Department of Electrical Engineering - Technion [Haïfa] (Israel)

2 - Institut Supérieur de l'Aéronautique et de l'Espace (France)

Abstract: In this study, we challenge the common belief that the number of samples required for high-dimensional importance sampling is exponential in the dimension. We find conditions that allow one to use importance sampling with number of samples that scales polynomially with the dimension. Our analysis leverages results by Chatterjee and Diaconis and covers both cases where the importance density is known in advance and where it must be adaptively estimated from random samples. In the latter case, the total number of samples required for both estimations of the density and the final integral is polynomial in the dimension. We also provide a theoretical foundation for understanding the performance of adaptive importance sampling schemes that rely on projection-based estimations and demonstrate the scalability of the cross entropy algorithm using the insights and techniques developed in this works.

4.2.22 Sun Maotong: Towards Safe Extubation Decision Making for Mechanically Ventilated Patients in ICUs: A Deep Constrained Offline Reinforcement Learning Approach

Sun Maotong[1], Xie Jingui[1]

1 - Technical University of Munich (Germany)

Abstract: Weaning patients from invasive mechanical ventilation and the regulation of sedation and analgesia during ventilation are critical decisions in intensive care unit (ICU) management, significantly affecting patient outcomes and the throughput of ICU, especially during the COVID-19 pandemic. Both prolonged dependences on mechanical ventilation and premature extubation are associated with an increased risk of complications and higher hospital costs. However, the clinical opinion on the best extubation protocol varies. In this study, we aim to develop a decision support tool that can utilize pre-collected patient information to predict time-to-extubation readiness and to have a more personalized extubation strategy for patients in the ICU. To this end, we propose a deep offline constrained reinforcement learning-based decision support tool to help physicians determine the best action at a given patient state from sub-optimal historical ICU data. We evaluate our model using the state-of-art off-policy evaluation (OPE) and show that our model enables substantial improvements in expected health outcomes and in consistency with relevant practice and safety guidelines.

4.2.23 Varma Sushil: A Heavy Traffic Theory of Matching Queues

Varma Sushil Mahavir[1], Maguluri Siva Theja[1]

1 - Georgia Institute of Technology [Atlanta] (United States)

Abstract: Motivated by emerging applications in online matching platforms and marketplaces, we study a matching queue. Customers and servers that arrive into a matching queue depart as soon as they are matched. It is known that a state-dependent control is needed to ensure the stability of a matching queue. However, analytically studying the steady-state behaviour of a matching queue, in general, is challenging. Therefore, inspired by the heavy-traffic regime in classical queueing theory, we study a matching queue in an asymptotic regime where the state-dependent control decreases to zero. It turns out that there are two different ways the control can be sent to zero, and we model these using two parameters viz., ε is the magnitude scaling parameter that goes to zero and τ is the time scaling parameter that goes to infinity. Intuitively, ε scales the magnitude of the state-dependent control, and τ is the threshold after which we modulate the control. We show that depending on the relative rates of ε and τ, there is a phase transition in the limiting regime. We christen the regime when $\varepsilon\tau \to 0$, the time dominates magnitude (TDM) regime, and the limiting behaviour is a Laplace distribution. The phase transition starts in the regime when, $\varepsilon\tau$ goes to a nonzero constant when the limiting distribution is a Gibbs distribution, and so we call it the hybrid regime. When $\varepsilon\tau \to \infty$, the limiting distribution converges to a uniform distribution and we call this the magnitude dominates time (MDT) regime. These results are established using two related proof techniques. Both the proof techniques generalizes the characteristic function method. The first one exploits the underlying structure by engineering complex exponential Lyapunov functions, and the second is a novel inverse Fourier transform method.

4.2.24 Wei Zhaoxuan: Online Inventory Control with Partial Backorder

Lim Andrew[1], Wei Zhaoxuan[2], Zhang Hanqin[3]

1 - Institute of Operations Research and Analytics, National University Singapore (Singapore)
2 - Institute of Operations Research and Analytics, National University of Singapore (Singapore)
3 - Department of Analytics and Operations, NUS Business School , National University of Singapore (Singapore)

Abstract: We consider a single-product inventory model with patient demand, where the unmet demands may be either patient and choose to wait (backorder) or lose their patience and choose to leave (lost sale), which can be taken as a hybrid of the classic lost-sales and backorder model and is referred to as *partially backorder model*. With perfect information about the distribution of the demand and customers' patience, we identify the (asymptotically) optimal control policy. We then consider a more practical online inventory control problem where the system manager doesn't have perfect information about the distribution and the patience of the demand and only observes the sales. Compared to the literature, our problem has unique challenges due to the complex nature of the inventory system because (a) the state of the system is only partially observable as the backorder is unobservable and (b) the sales (the data observed by the agent) could be the exogenously arrived demand and the patient customers. An UCB-type algorithm is then developed to obtain the near-optimal policy. We prove the regrets of our learning algorithm are (nearly) tight in the planning horizon T, and it can also recover the state-of-art regret bound in the lost-sales model developed in the literature.

4.2.25 Yu Sophie: Constant Regret Primal-Dual Policy for Multi-way Dynamic Matching

Mao Cheng[1], Wu Yihong[2], Xu Jiaming[3], Yu Sophie[4]

1 - Georgia Institute of Technology (United States)
2 - Yale University (United States)
3 - Duke University (United States)
4 - Duke university [Durham] (United States)

Abstract: We study a discrete-time dynamic multi-way matching model. There are finitely many agent types that arrive stochastically and wait to be matched. State-of-the-art dynamic matching policies in the literature require the knowledge of all system parameters to determine an optimal basis of the fluid relaxation, and focus on controlling the number of waiting agents using only matches in the optimal basis (Kerimov et al., 2021a,b; Gupta, 2021). In this paper, we propose a primal-dual policy that schedule matches for future arrivals based on an estimator for the dual solution. Our policy does not require the knowledge of optimal bases, and is the first to achieve constant regret at all times under unknown arrival rates. In addition, we show that when the arrival rates are known, the primal-dual policy achieves the optimal scaling as the lower-bound described in Kerimov et al. (2021a,b).. Furthermore, we find that when the arrival rates are known, the primal-dual policy can significantly outperform alternative dynamic matching policies in numerical simulations.

4.2.26 Zhe Su: Sample path large deviations for Levy processes and random walks with log-normal increments

Zhe Su[1], Rhee Chang-Han[1], Zwart Bert[2]

1 - Northwestern University [Evanston] (United States)
2 - Centrum Wiskunde & Informatica (Netherlands)

Abstract: The past two decades have seen breakthroughs in understanding the general principle that dictates the asymptotics of the rare events that arise in heavy-tailed systems. Hult, Lindskog, Mikosch, and Samorodnitsky (2005) was the first systematic confirmation of long-standing folklore in the heavy-tail literature called the principle of a single big jump: that is, the systemwide rare events are often driven by the catastrophic failure of a single component of the system when the underlying uncertainties are heavy-tailed, in particular, regularly varying. More recently, Rhee, Blanchet, and Zwart (2019) characterized the 'catastrophe principle' in the form of sample path large deviations, which addresses a more general class of heavy-tailed rare events far beyond the principle of a single big jump-i.e., the rare events that are driven by an arbitrary number of unusually large shocks. Bazhba, Blanchet, Rhee, and Zwart(2020) further extended the large deviations approach to the processes with (heavy-tailed) Weibul increment distributions, showing that a version of the catastrophe principle still holds in such systems, but the exact characterization is more subtle. Among the three most popular classes of distributions-regularly-varying, Weibull, and log-normal-for modeling heavy-tailed phenomena, this leaves only the log-normal case open. In this talk, we establish the sample path large deviations for random walks and Lévy processes and characterize the catastrophe principle for the log-normal case. We also discuss its implications in the tail asymptotics of queueing systems with log-normal job sizes.

4.2.27 Zhong Yueyang: Learning the Scheduling Policy in Time-Varying Multiclass Many Server Queues with Abandonment

Zhong Yueyang[1], Birge John[1], Ward Amy[1]

1 - The University of Chicago Booth School of Business (United States)

Abstract: We consider a learning variant of a canonical scheduling problem in a multiclass many server queue with abandonment (specifically, the $M_t/M/N+M$ and the $GI/GI/N+GI$ queues). The objective is to minimize the long-run average class-dependent expected linear holding and abandonment costs when the class-dependent model parameters (arrival rates, service rates and abandonment rates) are a priori unknown. The difficulty is that even when parameters are known, characterizing an optimal scheduling policy appears intractable. Fortunately, the simple $c\mu/\theta$ rule, that prioritizes classes in accordance with a static ranking that depends on the costs, the service rates, and the abandonment rates, is asymptotically optimal as the arrival rates and number of servers become large under certain conditions. Then, our task is to learn the service and abandonment rates well enough to determine an optimal static priority ranking for the classes, and we can benchmark our performance by defining the regret relative to the $c\mu/\theta$ rule. We propose a Learn-Then-Schedule algorithm, which is composed of a learning phase during which point estimates of the mean service and patience times are formed, and an exploitation phase during which the $c\mu/\theta$ rule with empirical mean estimates as a surrogate for actual parameters is followed. It is shown that the smallest achievable regret for static priority scheduling policies in T periods is $\omega(\log T)$, and we prove that our proposed algorithm achieves a regret upper bound of $O(\log T)$ which matches the lower bound.

4.2.28 Zur (Schwarz) Elisheva: Priority queues with arrival rate uncertainty

Zur (Schwarz) Elisheva[1], Oz Binyamin[1], Skir Gregory[1]

1 - The Hebrew University of Jerusalem (Israel)

Abstract: We study an unobservable single-server queue with arrival-rate uncertainty. Customers arrive according to a Poisson process with a random arrival rate and decide whether to pay for priority. We characterize the equilibrium customer behavior under different pricing and information disclosure policies. The model is analyzed under two different assumptions regarding the customers' degree of rationality. The first assumes that customers are fully rational and take into account the RASTA (rate-biased ASTA) phenomenon. The second assumes that customers are naive and base their decision on the PASTA phenomenon even though the arrival process is not Poisson.

4.3 Poster session III

4.3.1 Baars Justin: Delayed Hawkes infinite server queue: excitation delayed by sojourn time

Baars Justin[1], Mandjes M. R. H., Laeven Roger

1 - Universiteit van Amsterdam (Netherlands)

Abstract: The Hawkes process is a self-exciting point process where each arrival increases the conditional intensity instantenously. A variant of this model is introduced that allows for departures, in which the conditional intensity does not increase at arrivals, but at departures from the system. In other words, arrivals increase the intensity after a delay equal to the service requirements; therefore this model is called a delayed Hawkes infinite server queue. In the Markovian case, an explicit recursive procedure is presented to calculate the dth order moments analytically. Next, this is generalized to a multivariate setting. Furthermore, in a non-Markovian setting a family of models having sojourn time dependent excitation is introduced, containing the Hawkes, delayed Hawkes and ephemeral self-exciting processes as special cases, for which the transform is characterized by a fixed-point of a certain operator, that can be approximated by an iterative procedure. Next, for the delayed Hawkes process it is shown that if marks are heavy-tailed — in the sense of being regularly varying — then these heavy tails propagate across the system. Finally, I compare the delayed Hawkes process to the regular Hawkes process in the stochastic ordering, which will enable us to describe heavy-traffic behaviour.

4.3.2 Chenchen Xing: Estimating customer valuation in a service system with unobserved balking

Chenchen Xing[1], Taylor Peter[1], Jiesen Wang[2]

1 - The University of Melbourne (Australia)
2 - Tel Aviv University (Israel)

Abstract: We consider a service system selling goods to strategic customers. Customers are inhomogeneous on the valuation of goods. At their arrival times, customers are informed about the information of the current system. Customers probably arrive and then immediately balk if the waiting cost is above the valuation. Those balking customers are not observed by the system manager. The aim of the system manager is to estimate customers' valuation distribution as well as the arrival rate of customers including balking customers. The modelling framework that we will be working with is a double sided queue with Poisson streams of goods and customers on both sides. We find the maximum likelihood estimators for the customer's valuation and the arrival rate. Consistency of MLEs holds based on simulation results.

4.3.3 Dieleman Nanne: A Pseudo-Gradient Approach for Parameter-free Optimization of Cost Functions over the Stationary Distribution of a Discrete Time Markov Chain

Dieleman Nanne A.[1], Heidergott Bernd[1], Berkhout Joost[2]

1 - Department of Operations Analytics, Vrije Universiteit Amsterdam [Amsterdam] (Netherlands)
2 - Department of Mathematics, Vrije Universiteit Amsterdam [Amsterdam] (Netherlands)

Abstract: Discrete Time Markov Chains (DTMCs) are used to model a wide variety of networks, such as social, road congestion and hyperlink networks. By optimising these DTMCs, insights and guidelines are obtained on how the performance of these networks can be improved in practice. In this talk, we consider problems in which (a function of) the stationary distribution of a DTMC is optimised by adjusting the entries of the transition matrix of the DTMC, i.e., we consider a parameter-free setting. Algorithms exist that can handle linear and quadratic cost functions over the stationary distribution, but these algorithms do not scale well and cannot handle general costs functions. To optimise a general cost function, gradient search techniques are required. Unfortunately, gradient search techniques are not straightforward for this problem, as a violation of the stochasticity of the DTMC transition matrix during a search step makes it impossible to calculate the stationary distribution and, therefore, the cost function. In principle, a projected gradient could guarantee the feasibility of the problem throughout the search, but this is computationally challenging and does not scale well. Therefore, we will present a pseudo-gradient method that scales well in this talk. More specifically, we develop an algorithm based on the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm. We give insights into why solving this optimization problem is notoriously challenging and show the necessary adaptations to the SPSA algorithm. We compare the developed algorithm to a benchmark algorithm in the case of linear and quadratic cost functions and show how our algorithm can also be used in the case of general cost

functions. We conclude by providing examples of how this algorithm can be used in practice, for instance, in the ranking of web pages.

4.3.4 Janicka Agnieszka: Buffer content analysis of a two-class tandem fluid model

Janicka Agnieszka[1], Zwart Bert[1]

1 - Department of Mathematics and Computer Science [Eindhoven] (Netherlands)

Abstract: We study a tandem fluid model with an On-Off heavy-tailed Source A and a constant Source B. We assume that Fluid A leaves the system once processed by the first server, while Fluid B continues to the second buffer. We develop large deviation results for the buffer contents. In particular, we provide analytic arguments supporting that $\lim_{n\to\infty} n^{-\gamma}\cdot\mathbb{P}(Q_i(nT) \geq na) = C$, $i = 1, 2$. For Buffer 1, the primary cause of congestion is an On period of order n, as expected based on the Single Big Jump Principle. For Buffer 2, an additional Off period of order n might be required to cause congestion, depending on model parameters. The value of γ is determined by the number and type of big' periods of order n. The analytical results are complemented by a stochastic simulation of the model, with a rare-event sampling component that improves simulation efficiency.

4.3.5 Le Corre Thomas: Feature Projection for Optimal Transport

Le Corre Thomas[1,2], Bušić Ana[1,2], Meyn Sean[3]

1 - Département d'informatique - ENS Paris (France)
2 - Inria de Paris (France)
3 - University of Florida [Gainesville] (United States)

Abstract: Optimal transport is now a standard tool for solving many problems in statistics and machine learning. The optimal "transport of probability measures" is also a recurring theme in stochastic control and distributed control, where in the latter application the probability measure corresponds to an empirical distribution associated with a large collection of distributed agents, subject to local and global control. The goal of this work is to make precise these connections, which inspires new relaxations of optimal transport for application in new and traditional domains. The proposed relaxation replaces a target measure with a "moment class": a set of probability measures defined by generalized moment constraints. This is motivated by applications to control, outlier detection, and to address computational complexity. The main conclusions are (i) A characterization of the solution is obtained, similar to Kantorovich duality, in which one of the dual functions in the classical theory is replaced by a linear combination of the features defining the generalized moments. Hence the dimension of the optimization problem coincides with the number of constraints, even with an uncountable state space; (ii) By introducing regularization in the form of relative entropy, the solution can be interpreted as replacing a maximum with a soft-max in the dual; (iii) In applications such as control for which it is not known a-priori if the moment class is non-empty, a relaxation is proposed whose solution admits a similar characterization; (iv) The gradient of the dual function can be expressed in terms of the expectation of the features under a tilted probability measure, which motivates Monte-Carlo techniques for computation. We illustrate the approach on two applications: i) Optimal charging of EV arriving at a car park ii) Signal tracking by controlling the consumption of a set of consumers in an electricity network.

4.3.6 Levering Nikki: Externalities in the M/G/1 queue: LCFS-PR versus FCFS

Jacobovic Royi[1], Levering Nikki[2], Boxma Onno[3]

1 - University of Amsterdam (Netherlands)
2 - University of Amsterdam (Netherlands)
3 - Eindhoven University of Technology (Netherlands)

Abstract: Externalities are the costs that a user of a common resource imposes on others. In a stable M/G/1 system, the externalities created by an arriving customer with known service requirement equal the total waiting time that others will save if she would reduce her service requirement to zero. Naturally, these externalities are stochastic, and the corresponding analysis relies heavily on the underlying service discipline. In this talk, we compare the externalities under the last-come first-served with preemption (LCFS-PR) and first-come first-served (FCFS) service distributions. Specifically, we establish a joint decomposition for the externalities under LCFS-PR and FCFS in terms of a bivariate compound Poisson process. This decomposition can be used to derive several other results regarding the externalities: moments, asymptotic approximations as the service requirement goes to infinity, asymptotics of the tail distribution, and a functional central limit theorem.

4.3.7 Loeser Eva: Fluid Limit for a Stochastic Model of Enzymatic Processing with General Distributions

Loeser Eva[1], Williams Ruth[1]

1 - University of California [San Diego] (United States)

Abstract: We consider a stochastic chemical reaction system arising as a model for enzymatic processing in a cell. This can also be thought of as a multiserver multiclass queue with reneging operating under the random order of service discipline. Stochastic primitives for the model such as production/interarrival times, processing/service times, and lifetimes are assumed to be generally distributed. We establish a fluid limit for a measure valued process that keeps track of the remaining lifetime for each entity in the system. We prove uniqueness for fluid model solutions under mild conditions and study the asymptotic behavior of fluid model solutions as time goes to infinity. This poster is based on joint work with Ruth Williams.

4.3.8 Luo Yuwei: A Geometric Approach to Improve the Worst Case Performance of Thompson Sampling

Luo Yuwei[1], Bayati Mohsen[1]

1 - Stanford Graduate School of Business [Stanford] (United States)

Abstract: In this paper, we study Thompson Sampling (TS) for the stochastic linear bandit problem. TS is a popular policy to use in practice partly due to its strong empirical performance. In addition, it is known that TS achieves minimax Bayesian regret. However, its worst-case regret suffers from an extra multiplicative factor of order $\sqrt{d}$, which is recently shown to be unavoidable. Motivated by this discrepancy between the worst-case versus typical performance of TS, we derive an instance-dependent regret bound for TS that can be calculated efficiently in a "data-driven" manner through a geometric approach. This result enables us to detect and "course-correct" if TS performs poorly in any specific problem instance. We show that a variant of TS based on our approach enjoys the minimax optimal worst-case regret of $\tilde{O}(d\sqrt{T})$ while keeping most of the properties of TS. Simulation results are presented to validate our insights and to compare our algorithm with the baselines.

4.3.9 Mcarthur Harry: Preserving Privacy while Publishing Information

Mcarthur Harry[1], Smith-Miles Kate[1], Taylor Peter[1]

1 - University of Melbourne (Australia)

Abstract: The central goal of data-dissemination agencies around the world is to collect and publish information about individuals in a population to inform important policy decisions and research objectives. Can we ensure the privacy of the individuals is maintained while retaining the utility of the data? The current privacy preserving techniques used in practice rely on injecting noise into the statistics prior to publishing. In order to quantify the level of leakage a given technique has, we can consider what an attacker can learn from the released information, or equivalently what they can reconstruct about the underlying database. Anything that can be learned about an individual is a potential violation of privacy. We discuss the design of different reconstruction methods for the purpose of informing the decision on how best to inject this noise, or implement other privacy preserving techniques — optimising the utility of the data subject to certain privacy constraints.

4.3.10 Mohammadi Narges: Efficient Discovery of Cost-effective Policies in Sequential, Medical Decision-Making Problems

Mohammadi Narges[1], Skandari Reza[2], Shah Anand[3]

1 - PhD Student, Imperial College Business School (United Kingdom)
2 - Assistant Professor, Imperial College Business School (United Kingdom)
3 - School of Public Health, Imperial College (United Kingdom)

Abstract: Cost-effectiveness analysis is widely used by policymakers to prioritize interventions that improve a population's health. Net monetary benefit (NMB) is a metric used for the comparison of medical care strategies, which converts an intervention's health-benefits to monetary value using the willingness to pay (WTP) as the exchange rate. There is no universally accepted value for WTP, and its value affects the optimal choice of strategies. In this paper, we propose constructing a comprehensive menu of cost-effective strategies for the policymaker to choose from and develop a novel algorithm that efficiently generates the cost-effectiveness frontier. The algorithm is based on successive application of the

incremental cost-effectiveness ratio (ICER) minimization problem. We apply our modelling framework to design cost-effective hearing loss screening strategies for patients with cystic fibrosis (CF) disease. We build a stochastic optimization problem based on a hidden Markov reward model to optimally decide the timing and modality of audiometry. We solve the associated NMB and ICER optimization problems using a partially observable Markov decision process (POMDP) and a constrained, non-linear POMDP model, respectively. We build a data-driven model using evidence from the literature and use a grid-based approximation method to solve the problem numerically. We develop multiple heuristic and approximate policies and evaluate their performance against the optimal policy based on several metrics using a simulation model. The policymaker can use the simulation results to decide the trade-off between performance and ease of implementation of the various policies. We prove several theoretical properties of the proposed solution methods.

4.3.11 Pijnappel Tom: Optimal Deployment of Drone Base Stations in Wireless Networks

Pijnappel Tom[1], Van Den Berg Hans[2], Borst Sem[1], Litjens Remco[3,4]

1 - Eindhoven University of Technology (Netherlands)
2 - University of Twente (Netherlands)
3 - TNO (Netherlands)
4 - Technische Universiteit Delft (Netherlands)

Abstract: Reliable mobile communications is of critical importance, and should be maintained even in case of extremely crowded events or emergency scenarios. In such scenarios the deployment of drone-mounted base stations offers an agile and cost-efficient way to sustain coverage and/or provide capacity relief. In this talk we present a stochastic framework based on multi-class loss systems to estimate the blocking and coverage probabilities of drone-assisted wireless networks using information that is readily available to network operators. We demonstrate how this framework can be used to determine the minimum required number of drones and their corresponding locations for a given target performance level.

4.3.12 Ray Rounak: Percolation on Preferential Attachment Models

Ray Rounak[1], Van Der Hofstad Remco, Hazra Rajat Subhra, Garavaglia Alessandro

1 - Department of Mathematics and Computer Science, Eindhoven University of Technology (Netherlands)

Abstract: We study the bond percolation on the preferential attachment models. We have explicitly identified the critical percolation threshold for preferential attachment models with $m(\geq 2)$ out-edges and any affinity parameter $\delta > -m$. We have shown that the preferential attachment models are large-set expander and the survival probability is continuous at the critical percolation threshold. The work of Alimohammadi, Borga and Saberi (2021) suggests that for such cases, the critical percolation threshold for the graphs are that of the local limit. In our previous work (2022), we have shown that Pólya point tree is the local limit of a large class of preferential attachment models. For the case of negative δ, we have shown that the percolation threshold for Pólya point tree is 0. For positive values of δ, the critical percolation threshold turns out to be the spectral radius of the offspring generator of the Pólya point tree.

References:

- Yeganeh Alimohammadi and Christian Borgs and Amin Saberi (2021), Locality of Random Digraphs on Expanders
- Alessandro Garavaglia, Rajat Hazra, Remco van der Hofstad, Rounak Ray (2022), Universality of the local limit of preferential attachment models

4.3.13 Shah Abhin: On counterfactual inference with unobserved confounding

Shah Abhin[1,2,3], Dwivedi Raaz[1,3,4], Shah Devavrat[1,3], Wornell Gregory[2,3,5]

1 - Laboratory for Information and Decision Systems - Massachusetts Institute of Technology (United States)
2 - Research Laboratory - Massachusetts Institute of Technology (United States)
3 - MIT Institute for Data, Systems, and Society [Cambridge, MA] (United States)
4 - Harvard University Statistics Department (United States)
5 - MIT Computer Science & Artificial Intelligence Lab (United States)

Abstract: Given an observational study with n independent but heterogeneous units, our goal is to learn the counterfactual distribution for each unit using only one p-dimensional sample per unit containing covariates, interventions, and outcomes. Specifically, we allow for unobserved confounding that introduces statistical biases between interventions and outcomes as well as exacerbates the heterogeneity across units. Modeling the underlying joint distribution as an exponential family, we reduce learning the unit-level counterfactual distributions to learning n exponential family distributions with heterogeneous parameters and only one sample per distribution. We introduce a convex objective that pools all n samples to jointly learn all nparameter vectors, and provide a unit-wise mean squared error bound that scales linearly with the metric entropy of the parameter space. For example, when the parameters are s-sparse linear combination of k known vectors, the error is O(s log k/p). En route, we derive sufficient conditions for compactly supported distributions to satisfy the logarithmic Sobolev inequality. As an application of the framework, our results enable consistent imputation of sparsely missing covariates.

4.3.14 Van Santvoort Mike: Mathematically mapping the network of cells in the tumor microenvironment

Van Santvoort Mike[1], Lapuente-Santana ÃŞscar[1], Finotello Francesca[2], Van Der Hoorn Pim[1], Eduati Federica[1]

1 - Eindhoven University of Technology [Eindhoven] (Netherlands)
2 - Universität Innsbruck [Innsbruck] (Austria)

Abstract: In a malignant cancer, tumor cells exploit communication with immune cells to build resistance to one's own self-defense mechanisms. They break down and rebuild the communication network of interacting cells to misuse it for their own benefit. Treatment strategies like immunotherapy aim to interfere with this pro-tumor communication network, but for such treatments to be effective, it is important to map out what the cell-cell communication network in a patient's tumor looks like. To this end a lot of data about interacting proteins can be extracted from a patient's body, but it is often unclear how this data can be mapped consistently to a digraph of cells (nodes) and protein interactions (arcs). To intuitively bridge the gap between prior knowledge and a patient's cell-cell communication network, we propose a random graph model and analyze its properties, such that they can be used as predictive features of a patient's cancer. In our model nodes have types and arcs have colors. A type and color distribution is given as input, which encodes the probability that a uniformly chosen node/arc has a certain type/color. Additionally, a ruleset is given that fixes which arc colors can possibly connect to a given node type. Graphs of a fixed number of nodes and arcs are generated by first giving each node/arc a type/color independently of the others. Thereafter, each arc is placed uniformly at random between two nodes that can accept the arc according to the ruleset. We mathematically analyze the model by coupling probabilities of monotone events to probabilities in the inhomogeneous random digraph model. Then, we use recent literature by Cao et al. to translate properties of the inhomogeneous random digraph to our model. Examples of properties we investigate include the degree distribution, wedge/triangle counts, and the emergence of a giant strongly connected component.

4.3.15 Van Werde Alexander: Universality-based concentration for sums of dependent random matrices

Van Werde Alexander[1]

1 - Eindhoven University of Technology (Netherlands)

Abstract: We establish concentration inequalities for sums of dependent random matrices. Our results allow for two types of dependencies: First, we consider a model where the summands are generated by a Markov chain. Second, we consider a model where the summands are deterministic matrices scaled by random coefficients. The leading-order term in our results is asymptotically sharp and is identified by a quantity from free probability theory. We discuss applications related to community detection in block Markov chains.

4.3.16 Wellalage Achini: A Monte Carlo simulation-based simulated annealing algorithm for predicting the minimum staffing requirement at a blood donor centre

Wellalage Achini[1], Fackrell Mark[1], Zhang Lele (joyce)[1]

1 - The University of Melbourne (Australia)

Abstract: Australian Red Cross Lifeblood collects blood from non-remunerated voluntary donors. Thus, it is important to ensure that donors experience good service so they will return to donate blood again. Donor experience is adversely influenced by prolonged waiting times, but they may be reduced by determining the staffing demand over the day. In this talk, I propose a Monte-Carlo simulation-based simulated annealing algorithm that seeks the minimum number of employees to meet demand over a single day while ensuring the system's predicted average waiting time does not exceed a specified threshold. To enhance the efficiency of our simulated annealing algorithm, we develop a novel neighbourhood search method based on the staff occupancy levels. We use data from four different Australian Red Cross Lifeblood donor centres, demonstrating that our methodology can be adapted to any donor centre to determine the minimum staffing demand. Since these staffing demands ensure the donor waiting time target is met for each donor centre, they have the potential to improve both donor and staff satisfaction as well as streamline the donor flow.

4.3.17 Yeung Lane Chun: A non-asymptotic perspective on mean field control via log-concavity

Yeung Lane Chun[1], Lacker Daniel[1], Mukherjee Sumit[2]

1 - Department of Industrial Engineering and Operations Research, Columbia University (United States)
2 - Department of Statistics, Columbia University (United States)

Abstract: We study a class of stochastic control problems in which a large number of players cooperatively choose their drifts to maximize an expected reward minus a quadratic running cost. For a broad class of potentially asymmetric rewards, we show that there exist approximately optimal controls which are decentralized, in the sense that each player's control depends only on its own state and not the states of the other players. Moreover, the optimal decentralized controls can be constructed non-asymptotically, without reference to any mean field limit. These results descend from a broader framework inspired by the theory of nonlinear large deviations of Chatterjee-Dembo, for which we offer an efficient non-asymptotic perspective in log-concave settings based on functional inequalities. Joint work with Daniel Lacker and Sumit Mukherjee.

4.3.18 Zhang Kelly: Inference after Using RL to Learn Across Users for Digital Interventions

Zhang Kelly[1]

1 - Harvard University [Cambridge] (United States)

Abstract: Online reinforcement learning (RL) algorithms are a key tool for personalizing decision-making for digital interventions, as these algorithms can use previously collected data from users to learn and improve future treatment decisions. We provide methods to perform a variety of statistical analyses using data collected by RL and other adaptive sampling algorithms for digital interventions. In this work, we focus on data collected by online RL algorithms that can learn across users, i.e., use the data of multiple users to learn and inform treatment decisions. This data type is important since highly noisy outcomes (as is the case for digital interventions) mean RL algorithms that learn using the data of multiple users can significantly reduce noise and learn faster. At the same time, this data type is challenging to develop inferential theory for because online RL algorithms that combine data across induce dependence between the collected user data trajectories. We develop a general inferential approach for this non-i.i.d. data type that allows one to, for example, construct confidence intervals for time-varying treatment effects and perform off-policy analyses.

4.3.19 Zhu Feng: Stochastic Multi-armed Bandits: Trade-off among Optimality, Consistency, and Tail Risk

Simchi-Levi David[1], Zheng Zeyu[2], Zhu Feng[1]

1 - Massachusetts Institute of Technology (United States)
2 - University of California, Berkeley (United States)

Abstract: We study the stochastic K-armed bandit and fully characterize the interplays among three desired properties for policy design: worst-case optimality, instance-dependent consistency, and light-tailed risk. We show how the order of expected regret exactly affects the decaying rate of the regret tail probability for both the worst-case and instance-dependent scenario. A novel policy is proposed such that for any given $\alpha \in [1/2, 1)$ and $\beta \in [0, 1)$, our policy achieves a worst-case expected regret of $\tilde{O}(T^{\alpha})$ and instance-dependent expected regret of $\tilde{O}(T^{\beta})$, while enjoys a probability of incurring an $\Omega(T^{\delta})$ regret

that decays exponentially with a polynomial T term. Such decaying rate is proved to be best achievable. Our proposed policy achieves a delicate balance between doing more exploration at the beginning of the time horizon and doing more exploitation when approaching the end, compared to standard policies such as Successive Elimination and Upper Confidence Bound. A brief account of numerical experiments is conducted to illustrate the theoretical findings. Finally, we extend our proposed policy design to (1) a stochastic multi-armed bandit setting with adversarial baseline rewards (2) a stochastic linear bandit setting. Our results reveal insights on the trade-off between expected regret and tail risk for both worst-case and instance-dependent scenario, indicating that more sub-optimality and inconsistency leaves space for more light-tailed risk of incurring large regret.

4.3.20 Zhou Han: Optimal starting for Renal Replacement Therapy in Intensive Care Unit

Zhou Han[1], Xie Jingui[1]

1 - Technische Universität München = Technical University of Munich (Germany)

Abstract: Whether and when to initiate renal replacement therapy (RRT) is a controversial question in critical care. Benefits of delaying therapy include enabling a substantial number of patients to recover from acute kidney injury without undergoing RRT, and avoiding exposure to an unnecessary therapy with its negative side effects and costs. On the other hand, early therapy could improve electrolyte and fluid balance, enhance the removal of toxins, and thereby prevent potential life-threatening harm. Thus, there is a trade-off between short-term benefits and long-term recovery risks. Aiming to identify a treatment policy to minimize the mortality of a patient, we develop Markov decision processes. We prove conditions that establish the structural properties of the optimal solution. To estimate model parameters, we utilize instrumental variables and prediction methods.