

INVITED SESSIONS FOR VALUETOOLS

Title: *Simulation and Learning*

Organizers: *Ioannis Kontoyiannis (Brown Univ. and Athens Univ. of Econ. & Business) and Sean Meyn (University of Illinois at Urbana-Champaign)*

Sessions: **6 and 13**

These eight talks cover a range of topics in the rapidly evolving field of simulation and learning. Four of these talks concern importance sampling. This technique is commonly used to reduce the variance of simulations when estimating rare events. More recently it has found application to combinatorial problems found in computer science. These talks will touch on all of these applications. Two other lectures concern improved estimation techniques in very different contexts. Finally, there are two talks on the MCMC method

The talk by *P. Glynn* and *J. Blanchet* offers a fresh look at the foundations of the importance sampling method. Based on a more exact large deviations limit theorem they pinpoint a source of bias in the standard algorithm, and show how to correct for this using a state-dependent method.

J. Blanchet will show how state-dependent importance sampling methods can be used to circumvent the curse of dimensionality in counting problems found in computer science, and will describe techniques to establish the efficiency of this approach.

Pierre L'Ecuyer introduces a novel combination of importance sampling and state-dependent "splitting" to improve the variance of the importance sampling method.

V. Nicola and *T. Zaburnenko* will describe methods for estimating rare events in simulating networks using a state-dependent approach to importance sampling. This represents the first asymptotically efficient method for networks.

Casella and *Roberts* apply a very recent technique known as the "exact algorithm" to improve estimates obtained from estimating a diffusion sampled at a first exit time. Applications include models in finance and networks.

I. Kontoyiannis, *L.A. Lastras-Montano*, and *S. Meyn* introduce an alternative to the control variate method for variance reduction in simulation. A carefully constructed sampling method gives rise to an unbiased estimator that admits exponentially small error bounds in cases where no such bound is possible in a standard estimator.

C. Andrieu and *Y. Atchade* consider the asymptotic properties of adaptive MCMC algorithms, and proposes a new class of method called quasi-perfect MCMC.

G. Fort, E. Moulines, S. Meyn and P. Priouret show how the deterministic fluid model now commonly applied in network analysis can be used to analyse the MCMC algorithm. Sufficient conditions for stability of the algorithm are obtained, as well as new insight regarding the dynamics of this popular algorithm.

Title: *Web system-oriented performance*

Organizer: *Michele Colajanni (University of Modena and Reggio Emilia)*

Session: 8

Web-based services are nowadays at the basis of several interactive and complex applications. The system infrastructures underlying these environments have to satisfy scalability and availability requirements, and have to avoid performance degradation and system overload in critical conditions where Web-based applications may be reached by unpredictable or hard to predict loads.

A similar scenario requires adequate models and tools in the design phase of the system infrastructures and, at run-time, the utilization of many mechanisms for adapting the system behavior to different conditions. Indeed, there is a large set of algorithms and mechanisms that are integrated into these infrastructures, especially for load balancing, load sharing, overload control, admission control, and job dispatching purposes. To take appropriate decisions, all these mechanisms should be able to continuously monitor the system components, automatically infer the right information about the internal state, and possibly predict future load.

This invited session consists of three papers, that share a common keyword (*prediction*) in the context of design and run-time management of Web-based application environments.

The first paper “Capacity Planning Tools for Web and Grid Environments” (authors: *Sugato Bagchi, Eugene Hung, Arun Iyengar, Norbert Vogl, and Noshir Wadia*, from IBM T.J. Watson Research Center, Yorktown Heights, and IBM Software Group, San Jose) presents an overview of two sets of capacity planning tools for e-commerce and GRID-based applications. The first set of tools, known as the On Demand Performance Advisor (OPERA), uses analytical models to predict performance. It can predict solutions quickly and can be used for both offline and online capacity planning due to its fast response times. The second set of tools, known as Grid Value at Work, uses discrete event simulation. The two approaches are complementary, and OPERA and Grid Value at Work have been integrated into a single tool. This paper is completed by a results section where the authors present test cases for real customer workloads.

The second paper “Dynamic Estimation of CPU Demand of Web Traffic” (authors: *Giovanni Pacifici, Wolfgang Segmuller, Mike Spreitzer, and Asser Tantawi*, from IBM T.J. Watson Research Center, Yorktown Heights) is oriented to the run-time estimation of dynamic resource needs. Their proposal rely only on external and high-level measurements, such as overall resource utilization and request rates, instead of

application and/or kernel profilers. The authors formulate the problem as a linear regression problem and obtain its basic solution, but they also present techniques to deal with typical issues, such as data aging, flow rejection, flow combining, noise reduction, and smoothing. The experimental results obtained in realistic scenarios demonstrate that the estimations are close to the right values for request with high CPU requirements.

The third paper “Load Prediction Models in Web-based Systems” (authors: *Mauro Andreolini and Sara Casolari*, from the University of Modena, Italy) considers a problem close to that of the previous paper. The authors aim to predict at run-time the future load conditions of a resource by considering measures obtained from the load monitors of servers. These raw data are extremely variable even at different time scales, and tend to become obsolete rather quickly, hence it is quite difficult to forecast the behavior of future resource measures, and to deduce a clear trend about the load behavior of a resource, for example to find out whether a resource is offloading, overloading or stabilizing. The authors propose a two-step approach that first aims to get a representative view of the load trend from measured raw data, and then applies a load prediction algorithm to the load trends. They demonstrate in a multi-tier Web-based system, that the proposed approach is suitable to support different decision systems even for highly variable contexts and it is characterized by a computational complexity that is fully compatible to run-time decisions.

Title: *Control and Analysis of Communication Networks*

Organizers: *Maury Bramson (University of Minnesota) and Ruth Williams (University of California, San Diego)*

Session: 17

This invited session features four papers illustrating a variety of techniques used in the control and analysis of Internet-type communication network models. The first two papers related to congestion control and the last two relate to performance analysis for communication networks.

The proposed speakers are R. Srikant (University of Illinois), Tom Voice (Cambridge University), Thomas Bonald (France Telecom) and Francois Baccelli (Ecole Normale Supérieure, Paris).

R. Srikant will describe joint work with *S. Liu* and *T. Basar* on the design and analysis of a loss-based congestion control algorithm for high speed networks. Under this algorithm, the window increment/decrement adapts to the measured delay in the network. The design is justified using simple queueing models and its stability is analyzed using properties of stochastic matrices. The algorithm scales well with large link capacities and has the same fairness properties as TCP-Reno.

Tom Voice considers networks with multi-path routing capabilities, in which there is more than one available route between some source-destination pairs. A congestion avoidance mechanism dynamically regulates the flow along these multiple paths, based on information about the congestion level on the routes. Voice shows that a fluid model for the class of natural multi-path dual algorithms is globally stable in the absence of propagation delays. However, a counterexample shows that these algorithms may have undesirable delay stability properties. An alternative class, called controlled splitting multi-path dual algorithms, is proposed. For this class, fluid model global stability in the absence of delays is shown, and for one particular scheme, decentralized, scalable conditions for fluid model stability with delay are found.

Thomas Bonald will describe necessary and sufficient conditions for queueing networks, with state-dependent arrival and service rates, to be insensitive in the sense that the steady-state distribution depends on the service time distribution at each queue through the mean only. This insensitivity property is important for the development of simple engineering rules for communication networks that do not require the knowledge of fine traffic statistics.

Francois Baccelli will report on work with *D. McDonald* that analyzes a simplified stochastic model for HTTP flows. The model has a succession of idle and download

periods, with the file downloads being subject to a fixed packet loss probability. The same TCP connection might be used for the download of a random number of files, for which the effect of a slow start is taken into account. A closed form formula for the stationary distribution of the throughput obtained by a flow is given, as well as a closed form expression for the mean time to transfer a file. Several laws for file sizes and idle times are considered, including heavy tailed distributions. A brief discussion is given of how the formulae can be applied to predict bandwidth sharing among competing HTTP flows.