# DISSERTATION DEFENSE

**Sherwin Doroudi**

Monday, April 11, 2016
12:30 pm
388 Posner Hall

## STOCHASTIC ANALYSIS OF MAINTENANCE AND ROUTING POLICIES IN QUEUEING SYSTEMS

This dissertation focuses on reexamining traditional management problems that emerge in service systems where customers or jobs queue for service. In particular we examine maintenance and routing policies with an emphasis on novel methodological contributions and modeling approaches that seek to capture nuances in server or job behavior that have been unaddressed by prior work.

In the first chapter, we study a large class of quasi-birth-and-death Markov Chains that can model a variety of problems in computing, service, and manufacturing systems. These Markov chains consist of an infinite number of repeating levels and a finite number of phases, with transitions that are *skip-free* in level and *unidirectional* in phase. We present a procedure, which we call Clearing Analysis on Phases (CAP), for determining the limiting probabilities of such Markov chains exactly. The CAP method yields the limiting probability of each state in the repeating portion of the chain as a linear combination of *scalar bases* raised to a higher power based on the level of the state. This solution form is convenient for efficiently computing performance metrics (e.g., average queue lengths) of interest.

In the second chapter, we focus on applying the CAP method to evaluate maintenance policies in a setting where an online customer-facing service is vulnerable to persistent malware infections. These infections can cause performance degradation (i.e., drops in service rate) and facilitate data theft, both of which have monetary repercussions. Infections can go undetected and can only be removed by a time-consuming cleanup procedure, which takes the service offline and causes all existing jobs to be dropped without service, both of which lead to revenue loss. From a security perspective, cleanup should be undertaken as frequently as possible. On the other hand, from a performance-oriented perspective, frequent clean-ups are *desirable* because they maintain faster service, but they are simultaneously *undesirable* because they lead to greater downtime and loss of revenue. We ask how often and in response to what observable events cleanups should happen. In order to quantify the efficiency of various clean-up (maintenance) policies, we propose a revenue model which incorporates delay-based pricing and data theft costs. We model malware infections as a stochastic process that can evolve in stages. Unlike traditional maintenance and repair problems, our model necessitates analyzing a Markov chain that simultaneously tracks queue lengths and the (possibly unobservable) evolution of the malware infection process. The CAP method enables the analysis of such chains and the comparison of various clean-up policies. Previous work on service system maintenance under cyber-attacks has focused only on heuristic approaches, making this the first analytic investigation of the subject.

The third and fourth chapters focus on two different routing problems that emerge in service systems. Traditionally, routing in service systems attempts to minimize customer response times under the

assumption that all servers work at a fixed rate. In reality, many service systems, such as call centers, are staffed by people who respond to workload incentives. In the third chapter, we use a game-theoretic framework to model servers as agents who can choose how fast they work in response to the decisions of their coworkers (i.e., the other servers) and their manager (who sets the routing policy). We introduce a utility model where these strategic servers choose their service rate in order to maximize a tradeoff between an "effort cost" and a "value of idleness." Under this behavior, we find that some traditional policies such as routing to the fastest (or the slowest) free server do not admit symmetric equilibria, while random routing or routing to the server that was most (or least) recently idle *do* admit such equilibria. We also identify novel policies, which both admit symmetric equilibria and improve upon the traditional policies with respect to customer delay.

The fourth chapter addresses routing (or dispatching) jobs in web server farms. Unlike human servers, computer servers are not strategic, but routing in web server farms requires sending jobs to a server *immediately upon arrival*. Once a job is sent to a server, the job is served according to the processor-sharing service discipline, and it *cannot* be moved to another server in the future. However, the common assumption in the literature has been that all jobs are equally important or valuable, in that they are equally sensitive to delay. Our work departs from this assumption: we model each arrival as having a randomly distributed value parameter (i.e., a waiting cost rate), independent of the arrival's service requirement. Given such value heterogeneity, the correct metric is no longer the minimization of response time, but rather, the minimization of value-weighted response time (i.e., long-run waiting costs). In this context, we ask "what is a good routing policy for minimizing the value-weighted response time metric?" We propose a number of new routing policies that are motivated by the goal of minimizing this metric. Via a combination of exact analysis, asymptotic analysis, and simulation, we are able to deduce many unexpected results regarding routing.