

STATISTICS COLLOQUIUM

TUESDAY, FEBRUARY 22, 2011

TALK: 4:00 PM — SCIENCE CENTER RM. 705

RECEPTION: 5:15 PM — SCIENCE CENTER, 7TH FLOOR

“Online Model-Based Clustering for Crisis Identification in Distributed Computing”

Dawn Woodard

**Operations Research and Information Engineering
Cornell University**

ABSTRACT

Large-scale distributed computing systems can suffer from occasional severe violation of performance goals; due to the complexity of these systems, manual diagnosis of the cause of the crisis is too slow to inform interventions taken during the crisis. Rapid automatic recognition of the recurrence of a problem can lead to cause diagnosis and informed intervention. We frame this as an online clustering problem, where the labels (causes) of some of the previous crises may be known. We give a fast and accurate solution using model-based clustering based on a Dirichlet process mixture; the evolution of each crisis is modeled as a multivariate time series.

In the periods between crises we perform full Bayesian inference for the past crises, and as a new crisis occurs we apply fast approximate Bayesian updating. These inferences allow real-time expected-cost-minimizing decision making that fully accounts for uncertainty in the crisis labels and other parameters. We apply and validate our methods using simulated data and data from a production computing center with hundreds of servers running a 24/7 email-related application.