NYU Stern School of Business
Department of Information, Operations & Management Sciences
STATISTICS RESEARCH SEMINAR

TOPIC: Network-based auto-probit modeling with application to protein function prediction.
SPEAKER: Eric Kolaczyk, Boston University
DATE: Friday, March 12th, 2010
TIME: 11:30 AM- 12:30PM
PLACE: KMC 5-90

ABSTRACT

Consider the problem of predicting a binary process indexed on nodes in a network, where network edges indicate relationshipss among nodes, which are presumed to be predictive of the state of the binary process. A canonical example of this problem is that of predicting the functional roles of proteins based on various genome-wide data, such as protein-protein association networks. In this particular context, there is the additional difficulty of a high percentage of false negative labels in training data, a serious short-coming endemic to many types of biological databases. I will describe an approach for modeling data of this type and, in particular, for producing predictions in a manner that accounts for such labeling problems. Approaching the task as a binary classification problem, we develop a hierarchical Bayesian probit-based framework for modeling binary network-indexed processes, subject to artificially inflated false negative training labels, with a latent multivariate conditional autoregressive (CAR) Gaussian process. Our method performance is evaluated and compared with standard algorithms on weighted yeast protein-protein association networks, extracted from a recently developed integrative database called STRING. Results show that our basic method is competitive with or better than these other methods, and that incorporating the uncertainty in negative labels among the training data can yield substantial improvements in predictive accuracy.