

STATISTICS COLLOQUIUM

MONDAY, NOVEMBER 15, 2010

TALK: 4:00 PM — SCIENCE CENTER RM. 309

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

“Community Structure in Large Social and Information Networks”

Michael W. Mahoney
Mathematics Department
Stanford University

ABSTRACT

The concept of a community is central to social network analysis, and thus a large body of work has been devoted to identifying community structure. For example, a community may be thought of as a set of WebPages on related topics, a set of advertisers in a similar economic market, or more generally as a set of nodes in a network more similar amongst themselves than with the remainder of the network. Motivated by difficulties we experienced at actually finding meaningful communities in very large real-world networks, we have performed a large scale analysis of a wide range of social and information networks. Our main methodology used local spectral methods and involved computing isoperimetric properties of the networks at various size scales---a novel application of ideas from statistics and scientific computation to internet data analysis, which required the development of new algorithmic tools, as well as the reinterpretation of the statistical basis underlying traditional spectral approximation algorithms.

Our empirical results suggest a significantly more refined picture of community structure than has been appreciated previously. Our most striking finding is that in nearly every network dataset we examined, we observe tight but almost trivial communities at very small size scales, and at larger size scales, the best possible communities gradually “blend in” with the rest of the network and thus become less “community-like.” This behavior is not explained, even at a qualitative level, by any of the commonly-used network generation models. Moreover, this behavior is exactly the opposite of what one would expect based on experience with and intuition from expander graphs, from graphs that are well-embeddable in a low-dimensional structure, and from small social networks that have served as test beds of community detection algorithms. Possible mechanisms for reproducing our empirical observations will be discussed, as will implications of these findings for clustering, classification, and more general data analysis in modern large social and information networks.