The Hong Kong University of Science and Technology

Seminar on Business Data Science

Department of ISOM

Network modeling and goodness-of-Fit

by

Prof. Jiashun JIN
Department of Statistics & Data Science
Carnegie Mellon University, United States

Date: 4 December 2024 (Wednesday)
Time: 3:00pm – 4:00pm
Venue: Classroom 1007, LSK Business Building

Abstract

The block-model family has four popular network models: SBM, MMSBM, DCBM, and DCMM. A fundamental problem is, how well each of these models fits with real networks. We propose GoF-MSCORE as a new Goodness-of-Fit (GoF) metric for DCMM (the broadest one among the four), with two main ideas. The first is to use cycle count statistics asa general recipe for GoF. The second is a novel network fitting scheme. GoF-MSCORE isa flexible GoF approach. We adapt it to all four models in the block-model family.

We show that for each of the four models, if the assumed model is correct, then the corresponding GoF metric converges to N(0,1) as the network sizes diverge. We also analyze the powers and show that these metrics are optimal in many settings. For 12 real networks, we use the proposed GoF metrics to show that DCMM fits well with almost all of them. We also show that SBM, DCBM, and MMSBM do not fit well with many of these networks, especially when the networks are relatively large. Together with the mathematical tractability of the block-model family, these suggest that DCMM is a possible (or is close to the) sweet-spot for network modeling.

Bio

Jiashun Jin received his Ph.D in Statistics from Stanford University in 2003. He was trained in statistical inference for Big Data, specializing in dealing with the most challenging regime where the signals are both Rare and Weak. In such Rare/Weak settings, many conventional approaches fail, and it is desirable to find new methods and theory that are appropriate for such situations.

His earlier work was on large-scale multiple testing, focusing on (Tukey's) Higher Criticism and practical False Discovery Rate (FDR) controlling methods. He has developed the idea of Higher Criticism into a class of methods that are useful for solving problems in genetics and genomics and cosmology and astronomy, including cancer classification, cancer clustering, and nonGaussian signature detection in the Cosmic Microwave Background (CMB). He has proposed to use the so-called ``phase diagram" as a new optimality measure that is particularly appropriate for Big Data settings where the signals of interest are Rare/Weak, and worked out the phase diagrams for many seemingly unrelated settings.

His more recent interest is on complex graphs, social networks, and sparse PCA and Random Matrix Theory. He has developed a number of new methods, among which are the Graphlet Screening (GS) for high dimensional variable selection, IF-PCA for dimension reduction and high dimensional clustering, and SCORE for network community detection.

Jin and coauthors have collected and cleaned a data set for the coauthorship and citation networks for statisticians. The data set consists of titles, authors, keywords, abstracts, and citation counts of approximately 83,381 papers published in 36 journals in statistics and related fields, spanning 41 years. The data set provides a fertile ground for researches in social network of statisticians. It also opens doors for quantitative evaluation of the impacts of statistical research.

All interested are welcome!
Enquiries: Dept of ISOM