The Annals of Applied Statistics
2016, Vol. 10, No. 4, 1827–1834
DOI: 10.1214/16-AOAS978
Main article DOI: 10.1214/15-AOAS896

DISCUSSION OF "COAUTHORSHIP AND CITATION NETWORKS FOR STATISTICIANS"[1]

BY VISHESH KARWA AND SONJA PETROVIĆ

Harvard University and Illinois Institute of Technology

1. Introduction. Analyses of coauthorship and citation networks offer a fertile ground for studying research and collaboration patterns of scientific communities. Ji and Jin's efforts of collecting, cleaning and summarizing in various ways citation and coauthorship networks for statisticians is a great step forward to provide the community with a first such data set for self-study. They perform several descriptive analyses of the underlying networks to extract interesting patterns: they study trends of productivity over time, extract most prolific authors and research areas using various centrality measures, and find communities in these networks. We look forward to seeing this data set serving as a yardstick for fitting social network models to large data sets. Perhaps more interestingly, we see it as raising new research questions from the modeling, data representation and computational points of view and becoming a standard testbed for evaluating network models—both old and new—and testing scalability of inference procedures. In this regard, it is with great pleasure that we write this comment.

Here we take a model-based approach and consider the effects of various types of author interactions on the analysis and inference about the citation and coauthorship data sets. We are generally interested in three types of questions, two of which we discuss here: what are well-fitting models for the data? Is a simple network representation best for answering questions we ask, or should we be considering alternative representations? How can we scale existing network model fitting and goodness-of-fit testing procedures to networks of this size, as well as larger networks that the authors intend to collect? These forthcoming data sets should reduce sampling bias, but of course come at a price of a dramatic increase in network size and computational cost. We expect that availability of the data sets Ji and Jin have provided the community will encourage methodological research to push the limits of performing nonasymptotic inference in large and sparse networks.

We became aware of their data collection effort at a time when we were developing a basic exponential family model for hypergraphs, placing probabilities on

Received August 2016.

[1]Supported in part by the U.S. Air Force Office of Scientific Research Grant #FA9550-14-1-0141 to Illinois Institute of Technology and by the Singapore National Research Foundation under its International Research Centre Singapore Funding Initiative and administered by the IDM Programme Office through a grant for the joint Carnegie Mellon/Singapore Management University Living Analytics Research Centre.