The Annals of Applied Statistics
2016, Vol. 10, No. 4, 1835–1841
DOI: 10.1214/16-AOAS896D
Main article DOI: 10.1214/15-AOAS896

DISCUSSION OF "COAUTHORSHIP AND CITATION NETWORKS FOR STATISTICIANS"

BY MLADEN KOLAR[*] AND MATT TADDY[*,†]

University of Chicago[*] and Microsoft Research[†]

This article by Ji and Jin (JJ throughout) provides a network analysis by two experts in the field in the connections between statisticians and statistics research papers. This is not just an exercise in navel-gazing, but also an opportunity to compare results obtained by different methods in an area we know well: our profession. We think that the article leads to lessons for how we use network models and what data we choose to analyze. First, one can gain insight into a network by considering meta-information for the nodes—in this case, the research paper abstracts. Second, since summary statistics like closeness and betweenness centrality are sensitive to partial network observation, one needs to take care in defining the universe of nodes.

Topic analysis. In our first study, we consider decomposing the abstracts of the articles into latent "topics," for example, as in the LDA of Blei, Ng and Jordan (2003). The properties of the citation network can then be considered in light of the topical *content* of the articles. We use the `maptpx` R package [Taddy (2012)] to obtain posterior maximizing point estimates for LDA topics. The `maptpx` package applies the Bayes factors of Taddy (2012) in model selection, and for this data we find that a 15 topic decomposition is optimal.

We focus on topics that have seen their usage change over time—their mean proportions within documents during the first and last five years differ by more than 0.01. Most topics were stable over time so that only three meet this threshold. These three topics are shown in the left panel of Figure 1. The list of words given for each topic are those with the highest *lift*: within-topic probability over the average corpus-wide occurrence rate.

Topic 1 seems to contain traditional mathematical statistics content, especially for nonparametric and semiparametric analysis. The three articles most representative of this topic (i.e., have the highest estimated usage) are as follows:

Backfitting and smooth backfitting for additive quantile models [Lee, Mammen and Park (2010)],

Depth weighted scatter estimators [Zuo and Cui (2005)] and

Estimating invariant laws of linear processes by U-statistics [Schick and Wefelmeyer (2004)].

Received August 2016.