Discussion on “Statistical Modelling of Citation Exchange between Statistics Journals” by Varin, Cattelan and Firth

Pengsheng Ji (*University of Georgia, Athens*), **Jiashun Jin** (*Carnegie Mellon University, Pittsburgh*) **and Zheng Tracy Ke** (*University of Chicago*)

We congratulate Varin, Cattelan and Firth for a very stimulating paper. They use the Stigler model on cross-citation data and provide a model-based method to rank statistical journals. Their approach allows for evaluation of uncertainty of rankings, and sheds light on how to avoid overinterpretation of the insignificant difference among journal ratings.

In a related context, we study social networks for authors (instead of journals) with a data set we collect (based on all papers in the *Annals of Statistics*, *Biometrika*, the *Journal of the American Statistical Association* and the *Journal of the Royal Statistical Society*, Series **B**, 2003–2012). The data set will be publicly available soon.

The data set provides a fertile ground for studying networks for statisticians. In Ji and Jin (2014), we have presented results including

(a) “hot” authors and papers,
(b) many meaningful communities and
(c) research trends.

Here, we report results only on community detection of the citation network (for authors). Intuitively, network communities are groups of nodes that have more edges within than across (Jin, 2015). The goal of community detection is to identify such groups (i.e., clustering).

We have analyzed the citation network with the method of D-SCORE (Ji and Jin, 2014; Jin, 2015) and identified three meaningful communities; see Fig 16. The first community is “Large-Scale Multiple Testing”, including

(a) a Bayesian group: James Berger, Peter Müller,

(b) a Carnegie Mellon group: Christopher Genovese, Jiashun Jin, Isabella Verdinelli, Larry Wasserman,

(c) a causal inference group: Donald Rubin, Paul Rosenbaum,

(d) Three Berkeley-Stanford groups,

  (i) Bradley Efron, David Siegmund and John Storey,

  (ii) David Donoho, Iain Johnstone, Mark Low (University of Pennsylvania) and John Rice,

  (iii) Eric Lehmann and Joseph Romano, and

(e) a Tel Aviv group: Felix Abramovich, Yoav Benjamini, Abba Krieger (University of Pennsylvania), Daniel Yekutieli.

The second community is "Spatial Statistics" and can be further split into three groups

(a) a non-parametric spatial statistics group, including David Blei, Alan Gelfand, Yi Li, Trivellore Raghunathan,

(b) a parametric spatial statistics group, including Tilmann Gneiting, Douglas Nychka, Anthony O'Hagan, Adrian Raftery, Nancy Reid, Michael Stein,

(c) a semiparametric/non-parametric statistics group, including Raymond Carroll, Ciprian Crainiceanu, David Ruppert, Naisyin Wang.

The third community is "Variable Selection", including researchers on dimension reduction (Dennis Cook), quantile regression (Xuming He), variable selection (Peter Bickel, Peter Bühlmann, Emmanuel Candés, Jianqing Fan, Peter Hall, Trevor Hastie, Runze Li, Terrence Tao, Robert Tibshirani, Alexandre Tsybakov, Ming Yuan, Cun-Hui Zhang, Ji Zhu, Hui Zou).

Our results must be interpreted with caution, for the scope of the data set is limited. Also, it is not our intention to rank authors or papers.

References

[1] Ji, P. and Jin, J. (2014). Coauthorship and citation networks for statisticians. *arXiv:1410.2840*.

[2] Jin, J. (2015). Fast community detection by SCORE. *Ann. Statist* **43**(1), 57–89.

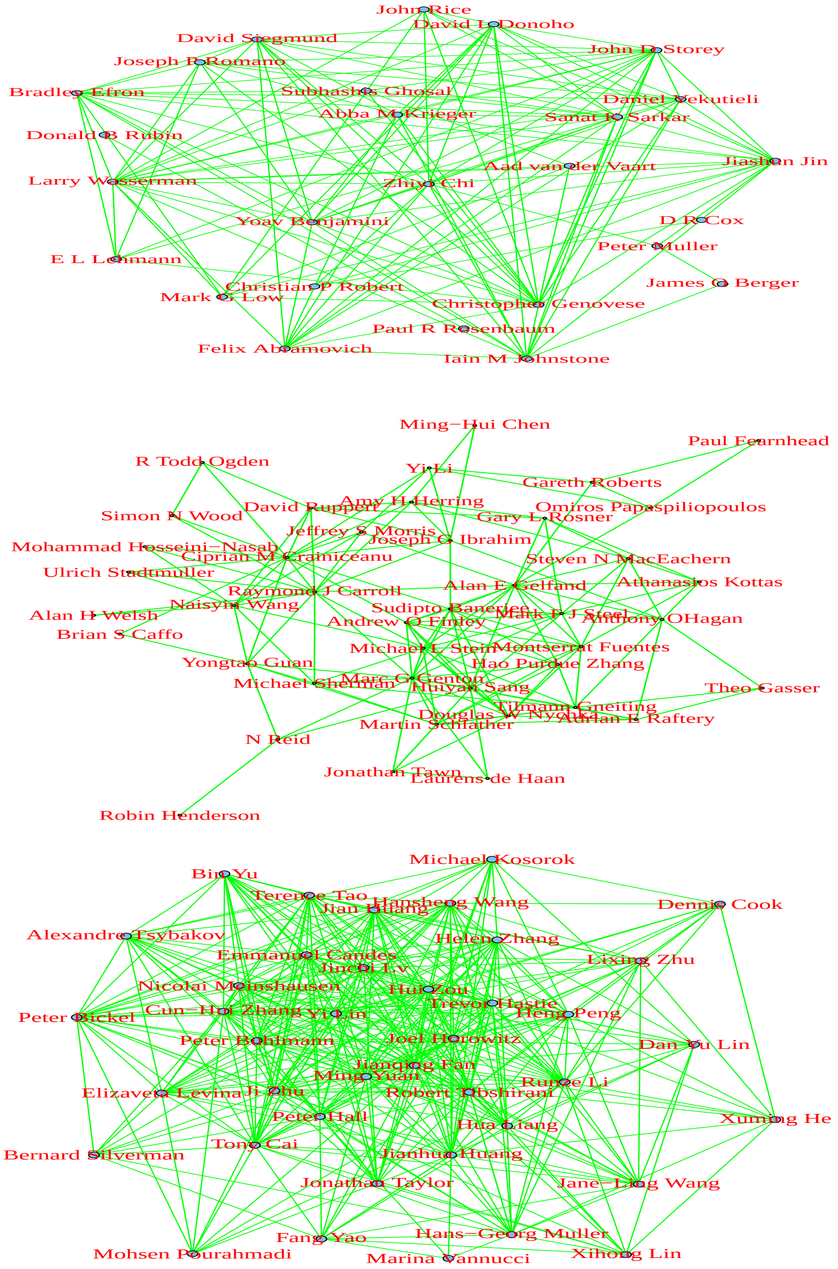

Figure 16: Communities found in the Citation network. Top: "Large-Scale Multiple Testing" (359 nodes; only 26 nodes with 24 or more citers are shown). Middle: "Spatial Statistics" (1010 nodes; only 42 nodes with 24 or more citers are shown. Bottom: "Variable Selection" community (1285 nodes; only 40 nodes with 54 or more citers are shown).