The Annals of Applied Statistics
2016, Vol. 10, No. 4, 1846–1853
DOI: 10.1214/16-AOAS896G
Main article DOI: 10.1214/15-AOAS896

REJOINDER: "COAUTHORSHIP AND CITATION NETWORKS FOR STATISTICIANS"

BY PENGSHENG JI AND JIASHUN JIN

University of Georgia and Carnegie Mellon University

We would like to thank all discussants for very thoughtful and stimulating comments. We especially thank B. Silverman for a nice and timely introduction for our paper; it is well said and very much illuminating.

In the past decades, the scientific community has grown substantially: we have way more researchers and annual publications than we ever had before. For example, the statistics community has grown from a tight-knit community (where one statistician may know almost all other statisticians) to a much larger one, driven by the technology advancements in computing and data acquisition.

While undoubtedly we have numerous achievements in our time (of which we should be proud), we have also heard many critical criticisms, among which there are the paper by Ioannidis (2005), "Why most published research findings are false," and the paper by Geman and Geman (2016), "Opinion: Science in the age of selfies."

As Silverman points out, an interesting question is therefore how to scrutinize the vast volume of scientific research we have today. While we can always turn to the traditional *subjective approaches*, we must admit that such approaches may be biased or inadequate, and *quantitative approaches*, like it or not, will play an increasingly more important role.

Having overseen the need for statisticians to engage a more active role in quantitative evaluation of scientific impact and productivity, Peter Hall said the following in his Presidential address at the 2011 Institute of Mathematical Statistics Annual Meeting (Miami, FL) [Hall (2011)]:

"...As statisticians we should become more involved in these matters than we are... We should definitely take a greater interest in this area."

Hall's viewpoint is reminiscent of the recent proposal by Donoho (2015) in "50 years of data science," where he oversaw the need of a new research discipline called "Science for Science."

Our work is a response to Hall's calling, and we believe that our data set will provide a fertile ground for future research on network analysis and related fields. Our effort in data collection is continued, and we now have a data set much larger than the one presented in our paper, covering papers in 36 representative journals in statistics and related fields, spanning 40 years.

Received October 2016.

1. Scope of the data set. Our data set is based on research papers published in four journals (Annals of Statistics, Biometrika, JASA and JRSS-B) from 2003 to the first half of 2012. Several discussants, Crawford, Kolar and Taddy, and Regueiro, Sosa and Rodríguez, point out that we must not over-interpret the results presented in the paper, for the scope of the data set is limited.

This is certainly a legitimate concern. However, what is an appropriate data set depends on the scientific goal, and it is not always *"the bigger the data set, the better."* For example, it was suggested by Stigler, Stigler and Friedland (1995) in a related context that focusing on a subset of relatively homogeneous journals may lead to more meaningful results.

The focus of the paper is on the social network of statisticians who are primarily interested in statistical methods and theory and who have USA as their home base. For this purpose, using the four journals above for our study is appropriate, for these journals are representative journals in statistical methods and theory and form a homogeneous group.

Also, we note that a larger data set is usually harder to analyze. For example, the k-core networks contain less information than the original networks, but some discussants (e.g., Karwa and Petrović; Wang and Rohe) choose to use such networks for their analysis, for these networks are easier to analyze than the original ones.

On the other hand, one may hope that our study could cover a wider range of scientific problems, and, for that purpose, the current data set may be inadequate. While this is certainly a very interesting direction, we would like to mention that it merely takes a lot of time and effort to collect data of this kind and have them cleaned and prepared for study. Therefore, it is only feasible to divide our project into different phases, and to complete them one by one. In fact, we may call the research presented in the current paper as "Phase I" of our project.

For Phase II of our project, we have already made substantial progress. We now have a data set that consists of titles, authors and affiliations, abstracts, MSC numbers and keywords of about 70,000 papers published in 36 representative journals in statistics and related fields, spanning 40 years. The data set is expected to be ready for study some time soon.

2. Network modeling. For network community detection, we focus on the Degree Corrected Block Model (DCBM) [Karrer and Newman (2011)]. DCBM is a generalization of the classical Stochastic Block Model (SBM), and the major difference is that DCBM models degree heterogeneity while SBM does not.

In most real-world networks (including the Coauthorship networks and Citation networks in our paper), it is observed that the distribution of the degrees has approximately a power-law tail [Albert and Barabási (2002)]. Therefore, the modeling of the degree heterogeneity is indispensable.

Regueiro, Sosa and Rodríguez suggest using the SBM for network community detection, as SBM allows us to recover both assortative and dis-assortative communities. We wish to point out that DCBM also allows us to do so, as DCBM includes SBM as a special case.

Regueiro, Sosa and Rodríguez also argue that a more general definition of community might be "a group of nodes that interact similarly across the network." We agree, but such a definition is consistent with DCBM, and we see no contradictions. We wish to clarify that SBM does not require that the probability to have an edge between two nodes in the same community be larger than that between two nodes from different communities; the same is true for DCBM if we normalize each probability by the degree heterogeneity parameters; see our paper for details.

On the other hand, it is of great interest to study the networks with other models. Karwa and Petrović investigate the networks with the p_1 model [Holland and Leinhardt (1981)]. They test whether the p_1 model holds for the Citation network and Coauthorship network (A), and concluded that the p_1 model is not a good fit for the former, but may be a good fit for the latter. We find such results very interesting.

Based on the above results, the authors argue that the edges of the citation network may be "dyadic dependent" (meaning that the edges are not independent random variables). While it is very likely that the Citation network is "dyadic dependent" (and so are many real-world networks), we don't think such a conclusion can be drawn from the testing results by Karwa and Petrović. In fact, the p_1 model is only one of many "dyadic independent" models (SBM and DCBM are also "dyadic independent" models). While the p_1 model is not a good fit for the Citation network, it is still possible for other "dyadic independent" models to have a reasonably good fit.

Karwa and Petrović also use the p_1 model to test the *reciprocation effects* of citations[1] and the *triadic closure effects* in coauthorship.[2] Their results are very interesting, and are consistent with our findings presented in the paper. In fact, regarding the reciprocation effects, we find that the proportion of (either earlier or later) reciprocation among coauthor citations is 79%, while that among distant citations is 25% (much smaller); the high reciprocation of coauthor citations may due to that people tend to return favors or that coauthors tend to share similar research interests. Regarding the triadic closure effects, we find that the Coauthor (B) network has a transitivity coefficient of 0.32, where a value in $(0.3, 0.6)$ is often regarded as being transitive [Newman (2010)]; it was reported in Newman (2004) that the transitivity coefficients of the biology, mathematics and physics communities are 0.43, 0.15 and 0.43, respectively.

[1]That is, if author i cites a paper by author j, then author j is more likely to cite a paper by author i.

[2]That is, if authors i and j wrote a paper and authors j and k wrote a paper, it is more likely that authors i and k have also written a paper.

3. Community detection methods. A large part of our community detection results is based on the methods of SCORE and D-SCORE, but we also compare the two methods with several other methods, including NSC, BCPL, APL and LNSC (see Tables 4–5, 7–8, Figures 6–7 and Section 5.2.3). The discussants propose several different approaches and analyze the networks from many different perspectives. These include but are not limited to the SBM approach by Regueiro, Sosa and Rodríguez and the RSC approach by Wang and Rohe. All these are very interesting, and we invite all researchers to explore their ideas with our data set.

SCORE is attractive for (a) it is computationally fast and scalable, and so able to handle large networks, and (b) it is a simple (yet effective) modification of the classical PCA, and it is easily extendable to other settings; in fact, we find the idea of SCORE can be conveniently extended to mixed membership estimation [Jin, Ke and Luo (2016)], topic modeling [Ke (2016)] and nonnegative matrix factorization.

4. Combining several networks for community detection. Regueiro, Sosa and Rodríguez propose to combine the Coauthorship network and Citation network for community detection. In particular, they approach the problem by fitting a *latent space model* [Handcock, Raftery and Tantrum (2007)] for the adjacency matrix of the Coauthorship network and that of the Citation network. Their approach is very interesting.

SCORE and D-SCORE can also be extended to address such a situation. Let A_1 and A_2 be the adjacency matrices of the Coauthorship network and the Citation network, respectively. Let $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$ be the first K eigenvectors of A_1, $\hat{\eta}_1^{(L)}, \hat{\eta}_2^{(L)}, \ldots, \hat{\eta}_K^{(L)}$ be the first K left singular vectors of A_2, and $\hat{\eta}_1^{(R)}, \hat{\eta}_2^{(R)}, \ldots, \hat{\eta}_K^{(R)}$ be the first K right singular vectors of A_2. We construct three matrices of entry-wise ratios similarly as in the paper, each with a size of $n \times (K-1)$. We then combine the three matrices into an $n \times 3(K-1)$ matrix, and cluster with the classical k-means.

We can also use text mining techniques for community detection. For example, we can run a text mining algorithm on the titles and abstracts, and treat each identified keyword as a feature. This gives us a bipartite network between papers and features (or between authors and features). We can then assess the similarity between two papers (or two authors) with some similarity measure. The similarity metrics can then be combined with the networks for community detection, interpretation and validation.

5. About the number of communities. Most existing community detection methods require the knowledge of K (i.e., the number of communities). However, how to estimate K is a challenging problem.

In fact, the problem is challenging even in much simpler settings. Consider a case where we have i.i.d. samples $X_1, X_2, \ldots, X_n$ from a (one-dimensional) K-

component Gaussian location mixture:

$$X_i \overset{\text{i.i.d.}}{\sim} \delta_0 N(0,1) + \sum_{k=1}^{K-1} \delta_k N(\mu_k, 1), \qquad \sum_{k=0}^{K-1} \delta_k = 1,$$

where $\delta_0 \geq \delta_1 \geq \cdots \geq \delta_{K-1} > 0$ calibrate the sizes of the mixing components. Even in such a simple setting, it is impossible to estimate K if some of the (δ_k, μ_k) fall very close to $(0, 0)$; one can only hope to provide a confidence lower bound for K, but not to estimate K consistently.

For network data, the problem is even more challenging because (a) it is still unclear what would be realistic yet tractable mathematical models for real-world networks, and (b) even when such a model exists, it is too complicated, at least for estimating K; it is also likely that some small-size communities are undetectable, and so it is impossible to estimate K consistently. There are some interesting recent works addressing this problem [e.g., Bickel and Sarkar (2016), Daudin, Picard and Robin (2008), Le and Levina (2015), Saldana, Yu and Feng (2016)], but how these methods perform for our networks remains unclear, especially because we don't know the true K in our networks.

Our strategy is different from that in these works. Our point is that it is hard to have an approach that works well and that only depends on the networks, and so it is preferable to choose K by combining such approaches with the "partial ground truth" (which is fortunately available to us).

In our paper, we first use the scree plot to suggest a possible range for K. We then try SCORE with all K in the range, and use the "partial ground truth" to help us pick the K that we think gives the most reasonable community partition. This is of course only a heuristic approach, but it reveals many meaningful and interpretable community structures.

Another possible approach is to apply SCORE iteratively. Recall that in our paper, the third community in the Citation network identified by SCORE (i.e., "Spatial and Semiparametric/Nonparametric Statistics") is hard to interpret. We tackle the problem by applying SCORE to the network formed by the nodes in this community only, and produce three communities that are much easier to interpret; see Figure 15. We plan to develop such an idea into an easy-to-use "iterative SCORE" algorithm in the future.

Wang and Rohe study the paper-paper citation network. They suggest that choosing K to the right of the elbow point in the scree plot may reveal new interpretable clusters. We find such analysis very interesting.

Regueiro, Sosa and Rodríguez also have a very interesting study on this topic. They investigate the networks with a different method and different choices of K; their model is SBM, which is also different from the DCBM model we use. Take the giant component of Coauthorship network (A), for example. They find $K = 3$ to be a good choice, and we find both $K = 2$ and $K = 3$ are reasonable choices, but neither provides very convincing results.

In fact, for this particular network, we notice that many members in the Fan group may have mixed memberships (many of them have strong ties to both the Carroll–Hall community and the North Carolina community), and so neither SBM nor DCBM is appropriate for the network, for they do not accommodate mixed memberships. In a recent manuscript [Jin, Ke and Luo (2016)], we find that a degree-corrected mixed membership model is more appropriate for this particular network. We propose mixed-SCORE as a new version of SCORE and obtain more meaningful results on this network; see details therein.

6. Data analysis with meta information. Several discussants (Crawford; Kolar and Taddy) suggest that we should collect and use the meta information of the published papers, such as keywords, abstracts and author characteristics (institution, thesis advisor, etc.) for our study. This is a great suggestion. We wish to point out that our Phase I data set has already included some meta information (e.g., DOIs, years of publication, titles, abstracts). Also, our Phase II data set includes more meta information (e.g., keywords, author institutions, funding agencies).

As some discussants illustrate, meta information can be very useful. Kolar and Taddy apply topic analysis to the abstracts and study how the topic usage (i.e., proportion of documents devoted to each topic) changes over time. Wang and Rohe apply text mining over the paper abstracts, and then use the identified key words to interpret the clustering results. We find these results very interesting and illuminating.

7. Centrality measures. Several discussants investigate the networks with different centrality measures. For example, Karwa and Petrović apply a centrality measure using the k-core network decomposition, Kolar and Taddy investigate the sensitivity of centrality measures to journal choice, and Regueiro, Sosa and Rodríguez use *eigenvector centrality* [Bonacich (1972)] as the centrality measure.

All these approaches are very interesting. And since these measures are different from those we use in the paper, they may lead to different results and thus shed additional insights on the networks. These measures complement each other, but it is hard to say one is "better" than the other.

In their Section 2, Reguerio, Sosa and Rodríguez seem to misunderstand what networks we refer to in Table 2 of our paper. In this table, the first column corresponds to the author-paper bipartite network, the second column corresponds to Coauthorship network (B), and all other columns correspond to the Citation network. We wish to clarify that the table does not claim Jianqing Fan as the third most collaborative author.

8. Summary. We thank all discussants for very stimulating comments. Silverman suggests that our data set can be used for quantitative evaluation of the quality and impact of scientific research. Along these lines, some problems that

are of interest include journal ranking [Stigler (1994), Varin, Cattelan and Firth (2016)], studying long-term scientific impact [Wang, Song and Barabási (2013)] and "metaknowledge" investigation for studying innovations [Evans and Foster (2011)]. Also, several discussants propose to use meta information for network analysis, Crawford suggests studying the network evolution over time, Karwa and Petrović point out the need of new models and representations of the networks, Kolar and Taddy make a very interesting connection between network analysis and topic modeling, and Regueiro, Sosa, and Rodríguez and Wang and Rohe have very stimulating discussions on selecting K. All these are very interesting topics for future research.

The Phase I of our data set can be downloaded either from http://faculty.franklin.uga.edu/psji/scc/ or from http://www.stat.cmu.edu/~jiashun/StatNetwork/PhaseOne; see details therein. We will continue our effort in collecting new data sets, and we are close to finishing the Phase II of our data collection project. We hope our data set will provide a fertile ground for research on networks and related areas, and we welcome all researchers to investigate their ideas with our data sets.

REFERENCES

ALBERT, R. and BARABÁSI, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys*. **74** 47–97. MR1895096

BICKEL, P. J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **78** 253–273. MR3453655

BONACICH, P. (1972). Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol*. **2** 113–120.

DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Stat. Comput*. **18** 173–183. MR2390817

DONOHO, D. (2015). 50 years of data science. *Unpublished manuscript*.

EVANS, J. A. and FOSTER, J. G. (2011). Metaknowledge. *Science* **331** 721–725. MR2798026

GEMAN, D. and GEMAN, S. (2016). Opinion: Science in the age of selfies. *Proc. Natl. Acad. Sci. USA* **113** 9384–9387.

HALL, P. G. (2011). "Ranking our excellence" or "assessing our quality," or whatever. *Inst. Math. Statist. Bull*. **September** 12–14.

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc*. **76** 33–65. MR0608176

IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *PLoS Med*. **2** e124.

JIN, J., KE, Z. T. and LUO, S. (2016). Estimating network memberships by simplex vertices hunting. *Manuscript*.

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206

KE, Z. T. (2016). A geometrical approach to topic model estimation. Available at arXiv:1608.04478.

LE, C. M. and LEVINA, E. (2015). Estimating the number of communities in networks by spectral methods. Available at arXiv:1507.00827.

NEWMAN, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci. USA* **101** 5200–5205.

NEWMAN, M. E. J. (2010). *Networks*: *An Introduction*. Oxford Univ. Press, Oxford. MR2676073

SALDANA, D. F., YU, Y. and FENG, Y. (2016). How many communities are there? *J. Comput. Graph. Statist*. To appear.

STIGLER, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statist. Sci*. **9** 94–108.

STIGLER, G. J., STIGLER, S. M. and FRIEDLAND, C. (1995). The journals of economics. *J. Polit. Econ*. **103** 331–359.

VARIN, C., CATTELAN, M. and FIRTH, D. (2016). Statistical modelling of citation exchange between statistics journals. *J. Roy. Statist. Soc. Ser. A* **179** 1–63.

WANG, D., SONG, C. and BARABÁSI, A.-L. (2013). Quantifying long-term scientific impact. *Science* **342** 127–132.

DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602
USA
E-MAIL: psji@uga.edu

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jiashun@stat.cmu.edu