# Discussion of "Co-authorship and citation networks for statisticians" by Jiashun Jin and Pengsheng Ji

Pedro Regueiro * Juan Sosa* Abel Rodríguez*

## 1. General comments

We would like to start by congratulating the authors for taking the initiative in gathering this very interesting and novel dataset. Given the substantial amount of work involved in scrapping the data, their focus on four periodicals (Journal of the American Statistical Association, Journal of the Royal Statistical Society Series B, Biometrika and Annals of Statistics) is understandable. However, this relatively narrow choice raises some concerns. The most obvious one relates to the robustness of the results to the choice of periodicals, particularly for authors/papers concentrating on areas for which specialized high-quality alternative publications exist. Two examples are biomedical applications and Bayesian methods. Furthermore, although the four journals selected are mainly methodological, the inclusion of the Applications and Case Studies section of JASA was unfortunate. Manuscripts published there can be expected to have more in common with papers published in Annals of Applied Statistics or the Journal of the Royal Statistical Society, Series C than with manuscripts in the Theory and Methods section of JASA itself.

The analysis in the paper feels a little bit like a "fishing expedition". The paper lacks a clear question that motivates and shapes the data collection. The use of multiple alternative methods (both for constructing the networks and for analyzing them) yielding different results also detracts from a sense of purpose. This is a pity, because there are a number of interesting questions that could be explored if the data collection exercise had been slightly expanded with a clear objective in mind. Some examples include:

1. What are the main drivers of collaboration in statistics?
2. How have the collaboration networks evolved over time?
3. How likely are researchers to publish with their PhD mentors as time goes by?
4. Are there regional biases in citation and/or publication patterns?
5. How prevalent is "self-referencing" (both at the author and journal level)?

The feeling of lack of focus is reinforced by the fact that the clusters generated by the community identification methods in the paper are puzzling. For example, the fact that only three clusters are identified in the connected component of the author citation network is quite surprising. This small number could be driven by the fact that the two-mode relational data has been projected into one-mode networks, by the fact that the resulting one-mode network is converted into a binary network instead of treated as weighted, or by the lack of formality in the choice of the number of networks. This observation also suggests that co-authorship and citation data separately are not enough to create a taxonomy of the statistical literature; multiple sources of information are necessary to produce a more fine-grained partition that better reflects most people's understanding of the community structure. The remainder of the discussion explores some of these issues.

*Department of Applied Mathematics and Statistics, University of California, Santa Cruz

| Coauthorship (A) | Coauthorship (B) |
|---|---|
| Peter Hall | Joseph G Ibrahim |
| Raymond J Carroll | Hongtu Zhu |
| Yanyuan Ma | Weili Lin |
| Aurore Delaigle | Yimei Li |
| Hans-Georg Muller | Xiaoyan Shi |
| Enno Mammen | Bradley S Peterson |
| Hua Liang | Daniel B Rowe |
| Alexander Meister | Hongyu An |
| Fang Yao | Wei Gao |
| Naisyin Wang | Yashen Chen |

Table 1: Top-ten authors based on eigenvalue centrality for the coauthorship (A) and coauthorship (B) networks.

## 2. Eigenvalue centrality

We complemented the centrality measures presented in the paper with the eigenvalue centrality (see Table 1). Interestingly, note that while the top-ten list contains two of the three highly highly central authors identified in Section 3 of the manuscript (Peter Hall and Raymond Carroll), it does not contain the third (Jinquiang Fan). This suggests that, even though all these three authors are highly collaborative themselves, the coauthors of Jinquiang Fan tend to be less collaborative than those of Peter Hall and Raymond Carroll. Furthermore, note that whereas the top-ten lists generated by other centrality measures substantially overlap for the two networks, the lists for eigenvalue centrality are completely different, suggesting that this metric is much more sensitive to the procedure used to dichotomize the weighted network.

## 3. How many communities? Reanalyses using stochastic blockmodels

In this section we explore using an stochastic blockmodel to fit the coauthorship (A) and coauthorship (B) networks, and compare the results to those presented in Section *4* of the

manuscript. The model assumes that the edges in the network ($y_{i,i'}$) are conditionally independent given the set of interaction probabilities $\Theta$, and that the probability of observing an edge between two vertices $i$ and $i'$ depends exclusively on the community membership of $i$ and $i'$,

$$y_{i,i'} \overset{\text{ind}}{\sim} \text{Ber}(\theta_{\xi_i,\xi_{i'}}), \tag{1}$$

where $\boldsymbol{\xi}$ is a vector of *community indicators* taking values in $\{1, 2, \ldots, K\}$ and $K$ is the maximum number of communities. In a Bayesian setting the model is completed with priors for the parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\xi}$. For simplicity, the interaction probabilities are assigned independent uniform priors, $\theta_{k,l} \overset{\text{ind}}{\sim} \text{Uni}[0, 1]$ and the prior on the community indicators are constructed by assuming the entries of $\boldsymbol{\xi}$ are exchangeable and follow a Categorical distribution in $\{1, 2, \ldots, K\}$

$$Pr(\xi_i = k \mid w_k) = w_k; \qquad i = 1, 2, \ldots, I, \tag{2}$$

with weights vector $\boldsymbol{w} \sim \text{Dir}\left(\frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$, such that the marginal likelihood from this model converges to the marginal likelihood of a mixture model with a Chinese restaurant process prior. The parameter $\alpha$, which controls the effective number of components $K^\star \leq K$, is assigned a Gamma prior.

Notice that the communities in the stochastic blockmodel have an interpretation that is slightly different from the the communities obtained from the algorithms considered by the authors (NSC, BCPL, APL and SCORE). Specifically, rather than groups of vertices with relatively large number of edges *within* and small number of edges *across*, communities in the stochastic blockmodel are formed by vertices that interact similarly across the network and, thus, these clusters can be thought of as functional structures in the network. In the setting of coauthorship networks this distinction turns out to be relevant as, a priori, one would expect to observe disassortative communities that arise from multiple students collaborating almost exclusively with their advisors and, at the same time, assortative communities that represent close-knit research groups with few outside collaborators. Therefore, the stochastic blockmodel seems as a natural modeling choice, as it is capable of simultaneously recovering assortative and disassortative mixing in a network.

### 3.1 Coauthorship network (A)

In this section we examine *coauthorship network (A)*. Following the manuscript, we focus on the largest connected component of this network. A first difference that can be appreciated in Figure 1 is the fact that the stochastic blockmodel supports the existence of three –rather than two– communities. As seen in this plot, Peter Hall, Raymond Carroll, Jianqing Fan and Tony Cai are clustered into a single community that can be interpreted as being composed by the network's "hubs". Although a direction for further investigation would be the use of degree-corrected blockmodels Karrer and Newman (2011), the fact that Joseph Ibrahim is not included in this community, despite having the fourth largest number of publications in the network, is evidence that the partition obtained by the stochastic blockmodel is not exclusively driven by vertex degree.

Table 2 compares the partition from the stochastic blockmodel to those obtained from NSC, BCPL, APL and SCORE using the Adjusted Random Index (ARI) and the Variation of Information (VI). Here, it can be seen that the communities from the stochastic blockmodel are closest to those from APL. In particular, Community 2 in the SBM corresponds almost perfectly to the Carroll-Hall community identified in the main paper, and Community 3 corresponds to the North Carolina community (Community 1 in the SBM is made of the four high-degree authors identified above, which APL assigns to the Caroll-Hall community).

| ARI/VI | SCORE | NSC | BCPL | APL |
|---|---|---|---|---|
| SBM | 0.64/0.43 | -0.05/0.99 | 0.08/0.95 | 0.90/0.13 |

Table 2: Adjusted Random Index and Variation of Information comparing the communities from the stochastic blockmodel to the communities obtained by the different methods presented in Ji and Jin (in press), using the giant component of Coauthorship (A)

### 3.2 Coauthorship network (B)

We also examined the *coauthorship network (B)* where two researchers are connected with an edge if they share one or more publications, focusing again on the largest connected component. In this case the stochastic blockmodel suggests six communities in the data, although two of them containing only a very small fraction of the vertices in the network (6 and 2 observations, respectively).

To compare the partitions obtained from the stochastic blockmodel with those derived from NSC, BCPL, APL and SCORE, Table 3 presents ARI and VI measures. These indexes suggest that, unlike the case of Coauthorship (A), the communities identified by the stochastic blockmodel have little overlap with any of those identified by other metrics. An inspection of the estimated interaction probabilities $\boldsymbol{\Theta}$ suggests that these differences might be driven by the fact that the stochastic blockmodel identifies a couple of disassortive communities.

| ARI/VI | SCORE | NSC | BCPL | APL |
|---|---|---|---|---|
| SBM | 0.04/1.57 | 0.03/1.38 | 0.00/2.09 | 0.04/1.11 |

Table 3: Adjusted Random Index and Variation of Information comparing the communities from the stochastic blockmodel to the communities obtained by the different methods presented in Ji and Jin (in press), using the giant component of Coauthorship (B).

To investigate this relationship further, table 4 shows the intersection of the communities from the stochastic blockmodels with those from APL. The stochastic blockmodel suggests that the "HDDA" community can be further partitioned into smaller blocks.

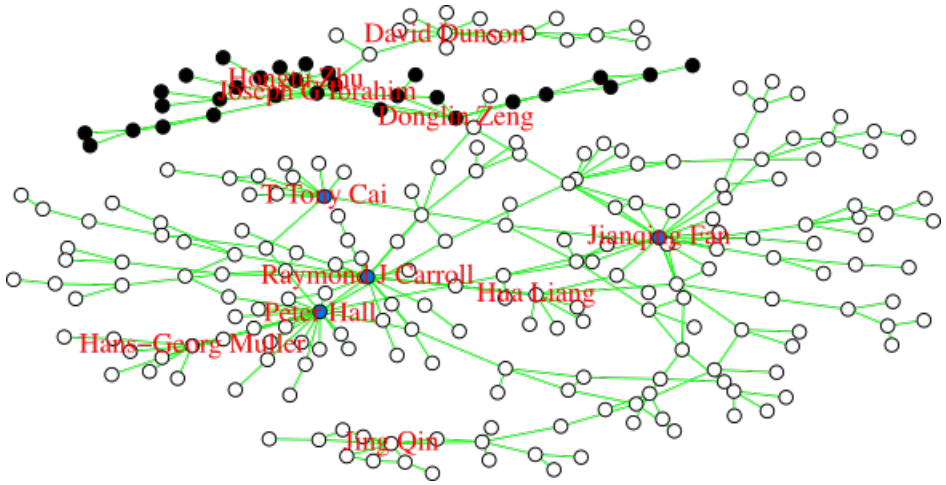


Figure 1: Communities resulting from fitting a stochastic blockmodel to Coauthorship Network (A)

| | | APL | | |
|---|---|---|---|---|
| | | Bayes | Biostat | HDDA |
| SBM | Community 1 | 14 | 12 | 211 |
| | Community 2 | 2 | 1 | 284 |
| | Community 3 | 2 | 5 | 202 |
| | Community 4 | 8 | 9 | 1505 |
| | Community 5 | 0 | 0 | 6 |
| | Community 6 | 0 | 0 | 2 |

Table 4: Comparison of communities obtained for the stochastic blockmodel and the APL algorithm.

| | | APL | |
|---|---|---|---|
| | | North Carolina | Carroll-Hall |
| LS | 1 | 23 | 159 |
| | 2 | 8 | 46 |

Table 5: Comparison of communities obtained for the latent space modeling (LS) and the APL algorithm.

## 4. Embeddings and combining information from citation and co-authorship networks

An alternative approach to community identification involves first embedding the probabilities in a Euclidean latent "social" space, and then clustering the nodes according to their position in the latent space (e.g., see Handcock et al., 2007). For example, for an undirected network we could proceed with a two-step approach where

$$y_{i,i'} \overset{\text{ind}}{\sim} \text{Ber}\left(\Phi(\beta + \mathbf{u}_i^T \mathbf{u}_{i'})\right), \qquad \mathbf{u}_i \overset{\text{ind}}{\sim} \text{N}_L(\mathbf{0}, \sigma^2 \mathbf{I}).$$

with further hyperpriors for $\beta$ and $\sigma^2$. The dimension $L$ of the latent space is selected using the *Deviance Information Criterion* (DIC) (Gelman et al., 2014, Ch. 6). Once point estimates $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_I$ are obtained (e.g., the posterior means after the enforcement of an appropriate identifiability constraint), communities can be determined using a finite mixture model for clustering, such as that implemented in the `R` package `mclust` (Fraley and Raftery, 2002; Fraley et al., 2012).

We use this procedure on the giant component of the co-authorship (A) network. DIC selects a three-dimensional social space (i.e., $L = 3$), and `mclust` identifies $K = 2$ communities. Table 5 compares the communities obtained using this procedure with those identified by APL; note that the results vary substantially.

The approach we just described can be extended to two or more adjacency matrices $\mathbf{Y}_1, \ldots, \mathbf{Y}_J$ defined over a common set of $I$ actors by letting

$$y_{i,i',j} \mid \beta_j, \boldsymbol{u}_{i,j}, \boldsymbol{u}_{i',j} \overset{\text{ind}}{\sim} \text{Ber}\left(\Phi\left(\beta_j + \boldsymbol{u}_{i,j}^T \boldsymbol{u}_{i',j}\right)\right),$$

with

$$\beta_j \overset{\text{ind}}{\sim} \text{N}(\mu, \tau^2), \qquad \mathbf{u}_{i,j} \mid \boldsymbol{\eta}_i, \sigma^2 \overset{\text{ind}}{\sim} N(\boldsymbol{\eta}_i, \sigma^2 \mathbf{I})$$

and $\mu \sim \text{N}(0, b_\mu^2)$, $\boldsymbol{\eta}_i \overset{\text{ind}}{\sim} \text{N}(\mathbf{0}, b_\eta^2 \mathbf{I})$, $\tau^2 \sim \text{IG}(a_\tau, b_\tau)$ and $\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma)$. Community identification proceeds then by clustering the "average" random position $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_I$.

We used the extended model to obtain a set of communities of authors that combines co-authorship and citation information. To facilitate comparisons, we focus again only on those authors included in the giant component of the co-authorship (A) network. The joint model identifies $K = 5$ communities, again with $L = 3$. Table 6 compares these 5 communities to those identified by the model based only on co-authorship data. Note that while the second original community remains largely unaffected by the inclusion of citation information (roughly corresponding to our new community 5), but the first one is split into four subgroups.

## References

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). *mclust Version 4 for R: Normal Mixture Model-*

| | | Joint | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Co A | 1 | 48 | 37 | 40 | 34 | 23 |
| | 2 | 0 | 1 | 2 | 4 | 47 |

Table 6: Comparison of communities obtained for the latent space modeling based only on co-authorship data (Co A) and both co-authorship and citation information (Joint).

*ing for Model-Based Clustering, Classification, and Density Estimation*.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

Ji, P. and Jin, J. (in press). Coauthorship and citation networks for statisticians. *Annals of Applied Statistics*.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.