Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Discussion of "Cocitation and Coauthorship Networks of Statisticians"

Haolei Weng & Yang Feng

To cite this article: Haolei Weng & Yang Feng (2022) Discussion of "Cocitation and Coauthorship Networks of Statisticians", Journal of Business & Economic Statistics, 40:2, 486-490, DOI: 10.1080/07350015.2022.2037432

To link to this article: https://doi.org/10.1080/07350015.2022.2037432

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 486–490
https://doi.org/10.1080/07350015.2022.2037432

Discussion of "Cocitation and Coauthorship Networks of Statisticians"

Haolei Weng[a] and Yang Feng[b]

[a]Department of Statistics and Probability, Michigan State University, East Lansing, MI; [b]Department of Biostatistics, New York University, New York, NY

ABSTRACT

We congratulate the authors for their stimulating and thought-provoking work on network data analysis. In the article, the authors not only introduce a new large-scale and high-quality publication dataset that will surely become an important benchmark for further network research, but also present novel statistical methods and modeling which lead to very interesting findings about the statistics community. There is much material for thought and exploration. In this discussion, we will focus on the cocitation networks, and discuss a few points for the coauthorship networks toward the end.

ARTICLE HISTORY
Received January 2022
Accepted January 2022

KEYWORDS
Covariates; Spectral clustering; Weighted network

1. Statistical Analysis of Cocitation Networks

As pointed out in the article, cociting two authors in an article provides evidence that they tend to share some common research interests. Based on this key observation, the authors used the cocitation relationship to construct citee networks for different time windows, and have performed solid statistical citee network analysis to study various aspects of research interests of statisticians including clustering, dynamic evolvement and diversity. In the rest of the section, we provide three sets of comments pertaining to weighted networks, model diagnostics and spectral embedding.

1.1. Weighted Citee Network

The citee networks analyzed in the article are undirected binary networks where an edge between two nodes is present if the edge weight is above a certain threshold. While thresholding the edge weights can reduce noise, it may result in loss of information. It would be interesting to directly investigate the weighted citee networks and check the impact of the weight information on the conclusions about research interests. To this end, we show how the authors' methods can be easily adapted to handle weighted networks, and provide some numerical results to discuss the impacts of weights. Our analysis is intended to stimulate more discussions and hence, illustrative rather than exhaustive.

We focus on the important citee network constructed using the cocitations during 1991–2000. This is the network employed to produce the research map in Figure 1 of the article. Following the notation used in the article, let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix of the citee network. The authors model the citee network with the *Degree Corrected Mixed-Membership* (DCMM) model which specifies that $\mathbb{P}(A) = \prod_{i<j}(\theta_i\theta_j\pi_i'P\pi_j)^{A(i,j)}(1-\theta_i\theta_j\pi_i'P\pi_j)^{1-A(i,j)}$, where $\theta_i > 0$ is the degree heterogeneity parameter of author i, and the weight vector $\pi_i \in \mathbb{R}^K$ models the research interests of author i. A spectral method called mixed-SCORE (Jin, Ke, and Luo 2017) is then performed for mixed-membership estimation. Now, let $\tilde{A}$ be the weighted adjacency matrix of the weighted citee network, where $\tilde{A}(i,j) \in \{0,1,2,\ldots\}$ denotes the edge weight between node i and node j. We could model $\{\tilde{A}(i,j)\}_{i<j}$ with some independent parametric distributions supported on the integers such as Poisson and negative binomial. However, after a careful inspection of the mixed-SCORE approach, we found that as with many other spectral methods (e.g., Rohe, Chatterjee, and Yu 2011; Jin 2015; Lei and Rinaldo 2015; Zhang, Levina, and Zhu 2020), mix-SCORE can be considered as a nonparametric method that is robust to parametric model specification. In particular, the method is expected to work well (e.g., achieving estimation consistency) as long as the model is first-order correct, that is, $\mathbb{E}[\tilde{A}(i,j)] = \theta_i\theta_j\pi_i'P\pi_j$, for $1 \le i < j \le n$, along with certain regularity conditions on the tail distribution decay. Thusly motivated, we will apply the same mixed-SCORE method to the weighed adjacency matrix $\tilde{A}$, and perform the same downstream analysis to produce the research map. The procedure can be easily implemented via a minor modification of the code provided by the authors.

The new research map is shown in Figure D1. Comparing it to the original map in Figure 1 of the article, we have the following observations and comments:

- We come up with the 15 cluster labels by carefully checking research works of representative authors (with large degrees) in each cluster. These labels are similar to the ones in the original map and provide a reasonable representation of subareas of the three primary research areas.
- A notable change is that the new map has more pure nodes especially in the "nonparametric statistics" and "biostatistics" research areas. Such a difference can be further confirmed from the estimated mixed-membership vectors $\{\hat{\pi}\}_{i=1}^n$ as shown in Figure D2.

CONTACT Yang Feng yang.feng@nyu.edu Department of Biostatistics, New York University, New York, NY.

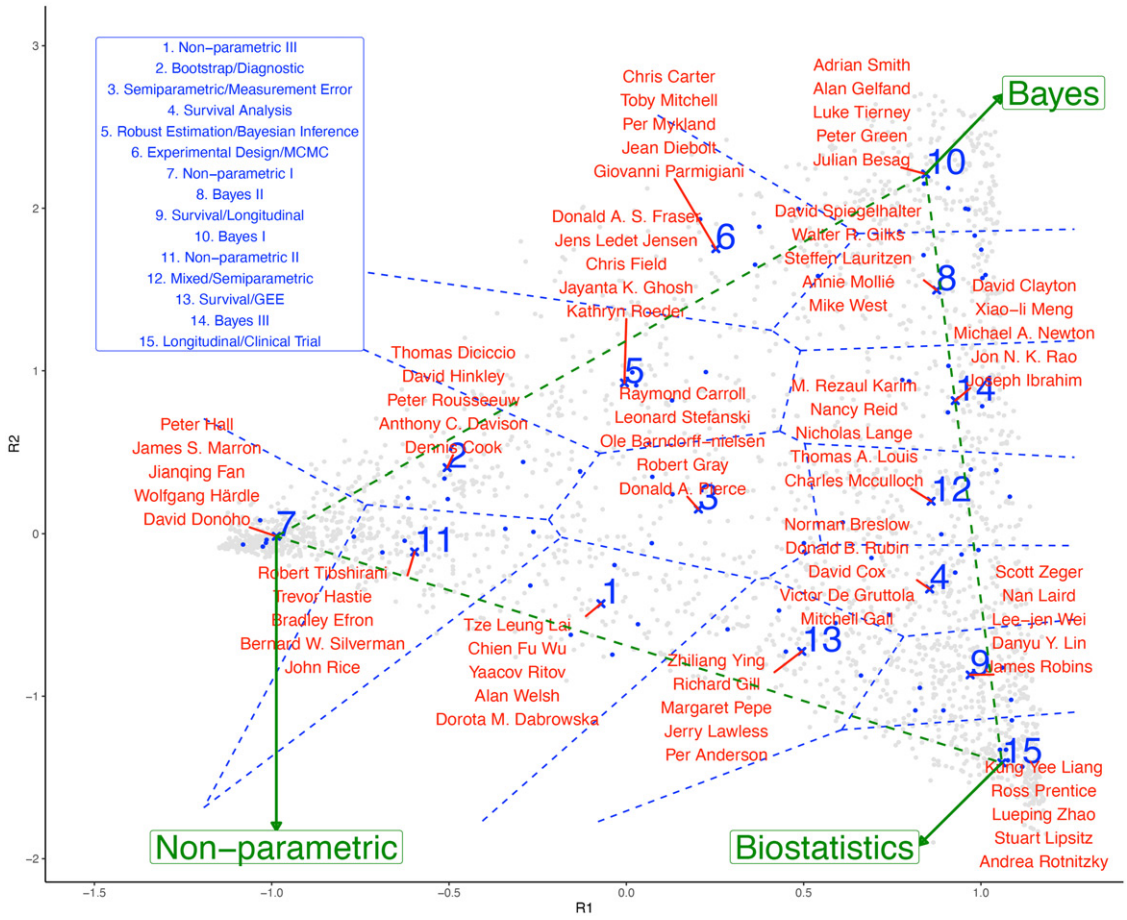

Figure D1. The research map obtained from the weighted citee network.

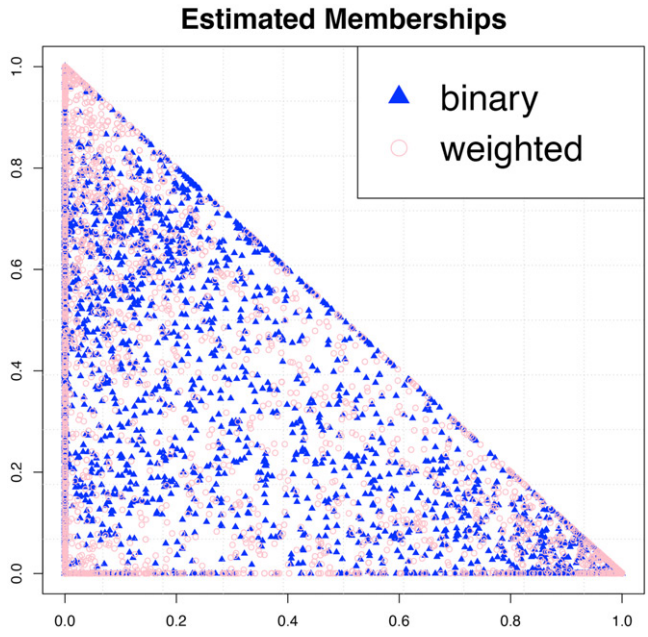

Figure D2. The estimated memberships $\{\hat{\pi}\}_{i=1}^{n}$ based on the binary and weighted citee networks. The X and Y coordinates correspond to the "nonparametric statistics" and "biostatistics" research areas.

- To examine the meaning of the aforementioned change, we select representative authors who are not pure nodes in the original map but become (nearly) pure in the new map. This is summarized in Table D1. Let's take David Donoho as an example. According to his articles published in the 36 journals during 1982–2000 (there is a 10-year time lag in citation), Donoho's research works concentrated on nonparametric estimation, wavelets and robust statistics, in our opinion, of which little fraction is related to "Bayes" or "Biostatistics." Hence, it is perhaps more accurate to classify him as a nearly pure node in the "nonparametric" vertex (note that "nonparametric" is a broad notion in the statistics triangle). Similar results apply to several other authors that we checked in Table D1 including Jianqing Fan, Peter Hall and Kung Yee Liang. Why is there such a change? A plausible explanation is that in the citation network two authors working on different areas can be cocited in the same article, though usually with a smaller chance compared with those working on the same area. This type of edges can pull a pure node away from the triangle corners. However, if we incorporate the edge weight information, it is very likely that those noisy edges will be down-weighted so that the location of (nearly) pure nodes can be more accurately estimated. We did not go over all the authors in Table D1 (and other authors with significant changes) and thus, do not claim that all the changes are positive.
- A more complete comparative study of binary and weighted citee network is desirable. However, since there is no ground truth, it is not immediately clear how to evaluate various

Table D1. Selective authors whose estimated membership vector has a large change.

Nonparametric	Bayes	Biostatistics
Matt P. Wand (0.55, 0.98)	Adrian Smith (0.55, 0.95)	Ross Prentice (0.59, 0.97)
Peter Hall (0.58, 1.00)	Alan Gelfand (0.57, 0.95)	Alan Agresti (0.58, 0.95)
Joseph Romano (0.54, 0.94)	Peter Green (0.55, 0.91)	Kung Yee Liang (0.55, 0.91)
Jianqing Fan (0.67, 1.00)	Julian Besag (0.65, 0.98)	Steve Self (0.65, 0.98)
David Ruppert (0.63, 0.96)	Amy Racinepoon (0.67, 1.00)	Andrea Rotnitzky (0.64, 0.97)
M. C. Jones (0.69, 1.00)	Walter R. Gilks (0.60, 0.91)	David Harrington (0.67, 1.00)
Wolfgang Härdle (0.70, 1.00)	Bradley Carlin (0.62, 0.93)	Garrett Fitzmaurice (0.67, 1.00)
James S. Marron (0.72, 1.00)	Wing Hung Wong (0.63, 0.93)	Lueping Zhao (0.67, 1.00)
Hans-georg Müller (0.75, 1.00)	Martin Tanner (0.63, 0.93)	John Neuhaus (0.60, 0.92)
David Donoho (0.73, 0.97)	Luke Tierney (0.69, 0.99)	Geert Molenberghs (0.68, 1.00)

NOTE: For each column, the pair (a, b) denotes the estimates for the corresponding membership coordinate based on the binary and weighted citee networks, respectively.

comparison results. For instance, the adjusted rand index between the clustering results based on binary and weighted citee networks is 0.226, indicating a significant change on the partitions of the subareas. But it is hard to argue which one is better. One possible direction is to use article abstracts. We noticed that the authors have applied topic modeling techniques on abstracts to label the three vertices in the triangle. Similar modeling strategies may be used to create a research interest profile for each author. We believe that these profiles can serve as very informative nodal features to help assess the results obtained from citee networks.

1.2. Model Diagnostics

The study of research interests in the article relies on the interpretation and estimation of the membership vectors $\{\pi_i\}_{i=1}^n$ in the Degree Corrected Mixed-Membership (DCMM) model or its dynamic version. While the authors have obtained various interesting results that are interpretable and sensible via these models, it remains important to perform statistically sound model diagnostics. This is crucial for drawing valid conclusions from real data analysis. Model checking and diagnostics has not yet been well studied in the community detection and block-models literature. Existing related works focus on determining the number of communities and model selection for stochastic block model and degree corrected variants Yan et al. (2014), Li, Levina, and Zhu (2016), Lei (2016), Saldana, Yu, and Feng (2017), Wang and Bickel (2017), and Chen and Lei (2018). While developing novel diagnostic tools for DCMM is beyond the scope of this discussion, nevertheless, we would like to point out several relevant points as follows:

- It is fairly challenging to derive a goodness-of-fit test for DCMM. First of all, given that DCMM is already a very general blockmodel, what is the appropriate full model to test against? Second, the large-sample analysis of a given test statistic is a nonstandard asymptotic problem, and the asymptotic distribution will critically depend on the sparsity level of the network.
- Can we have useful residual diagnostic plots to assess the goodness of fit of DCMM? We describe a procedure directly built on top of the mixed-SCORE method (Jin, Ke, and Luo 2017). Using the notation in Section 1.1 and further defining $\Theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$, $\Pi = (\pi_1, \pi_2, \ldots, \pi_n)'$, DCMM assumes that $\mathbb{E}(A) = \Theta\Pi P\Pi'\Theta$. The following has been proved in Jin, Ke, and Luo (2017): (i) There exists a unique nonsingular matrix $B = (b_1, \ldots, b_K) \in \mathbb{R}^{K\times K}$ such that $\Xi = \Theta\Pi B$, where the columns of Ξ are eigenvectors corresponding to the K nonzero eigenvalues of $\mathbb{E}(A)$. (ii) $\{b_i\}_{i=1}^K$ can be explicitly expressed using the K nonzero eigenvalues of $\mathbb{E}(A)$ and the K vertices in the simplex formed by the embedding of $\mathbb{E}(A)$. We then proceed with the following steps:

  1. Based on (ii), use the vertices found by mixed-SCORE and the K largest eigenvalues (in magnitude) of A to obtain $\hat{B}$.
  2. Based on (i), solve
  $$\hat{\Theta} = \arg\min_{\theta_i>0, 1\le i\le n} \|\hat{\Xi} - \Theta\hat{\Pi}\hat{B}\|_F^2,$$
  where $\hat{\Pi}$ is the estimated memberships from mixed-SCORE and $\hat{\Xi}$ is the eigenvector matrix of A.
  3. Solve the least-squares problem with linear constraints to obtain:
  $\hat{P} = \arg\min_{P:\hat{\Theta}\hat{\Pi}P\hat{\Pi}'\hat{\Theta}\in[0,1]^{n\times n}} \|A - \hat{\Theta}\hat{\Pi}P\hat{\Pi}'\hat{\Theta}\|_F^2$.
  4. Compute the residual matrix $E \triangleq A - \hat{\Theta}\hat{\Pi}\hat{P}\hat{\Pi}'\hat{\Theta}$.

  Having the raw residuals, as in logistic regression diagnostics, we can construct a deviance residual plot by grouping the deviance residuals into bins. In the current case, the bins are chosen to correspond to the inter-cluster and intra-cluster edges where the clusters are the 15 subareas in the research map.
- Some preliminary results obtained from the above method are shown in Figure D3. It is clear from the left plot that the estimated degree parameters based on mixed-SCORE capture well the degree heterogeneity in the data. The correlation between the observed degrees and estimated degree parameters is 0.965. For the deviance residual plot on the right, we see a rather even variation as the fitted value varies, thus, revealing no significant inadequacy in the model. There is one suspicious point that may deserve further investigation. This point represents the bin formed by the intracluster data from the 15th cluster whose cluster size is 88, the second smallest among the 15 clusters.

1.3. Spectral Embedding and Mixed-Membership Vectors

At the heart of the citee network analysis is the *SCORE embedding*, which embeds each node i into a $(K-1)$-dimensional space with $\hat{r}_i = \left[\frac{\hat{\xi}_2(i)}{\hat{\xi}_1(i)}, \frac{\hat{\xi}_3(i)}{\hat{\xi}_1(i)}, \ldots, \frac{\hat{\xi}_K(i)}{\hat{\xi}_1(i)}\right]$, where $\hat{\xi}_1, \ldots, \hat{\xi}_K$ are the first K eigenvectors of the adjacency matrix. The embedding is further generalized to handle dynamic citee networks.

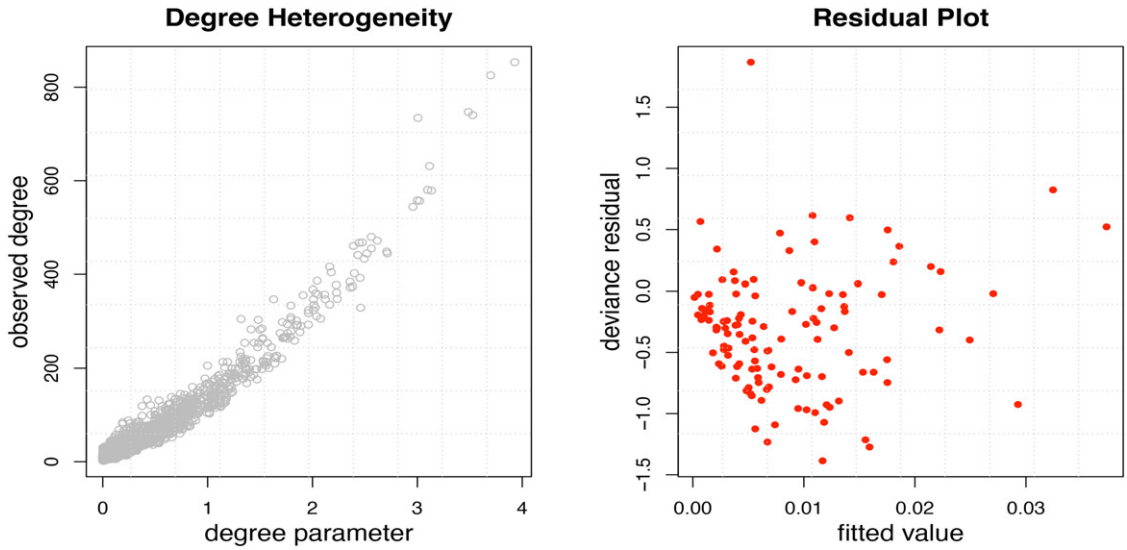

Figure D3. Left plot shows the relationship between observed degrees and estimated degree parameters based on mixed-SCORE method. Right plot is the deviance residual plot, where X-axis and Y-axis are the average of fitted values and average of deviance residuals (with normalization) within each bin, respectively.

These embedded points are the key statistics that the authors use to produce research map, generate research trajectory for individual author, and quantify diversity of author research interests. Given the importance of the embedding, we would like to highlight a few points that might be helpful toward a better understanding of the embedding.

- *Research interest representation*. The embedded points $\{\hat{r}_i\}$ are used to represent the research interests of authors. As explained in the article, each embedded point $\hat{r}_i$ is approximately contained in a simplex with $\hat{r}_i \approx \sum_{k=1}^{3} w_i(k)v_k$ where $\{v_k\}_{k=1}^{3}$ are the vertices representing three primary research areas, and $w_i \propto \pi_i \circ b$ characterizes the position of $\hat{r}_i$ within the simplex. However, when the elements of the positive vector b take substantially different values, the weight vector w_i and the membership vector π_i can differ significantly. As a result, the research map that consists of $\{\hat{r}_i\}$ may not be an authoritative description of research interests, because the positions of embedded points do not truly reflect the research interests which are modeled by $\{\pi_i\}$ in DCMM. We have checked the estimated b from the mixed-SCORE method for the citee network constructed during 1991–2000. It is $(0.037, 0.026, 0.035)$ for the binary network and $(0.014, 0.009, 0.011)$ for the weighted network. Hence, this may not be of concern in the current article, but nonetheless deserves attention when applying the embedding to other citation data. The general question here can be formulated as: is it better to use $\{\hat{\pi}_i\}$ instead of $\{\hat{r}_i\}$ to study research interests?
- *Statistical variations of the embedding*. The theoretical analysis in Jin, Ke, and Luo (2017) shows that $\{\hat{r}_i\}$ and $\{\hat{\pi}_i\}$ obtained from mixed-SCORE enjoy nice convergence properties (along with some minimax optimality). An interesting and important future research topic is to characterize the asymptotic distributions. Such asymptotic results can be used to perform more sophisticated data analysis. For instance, depending on the parameter configurations in the model, the asymptotic covariance matrix of each $\hat{r}_i$ may vary considerably. Incorporating the heterogeneity could lead to a better clustering method for the partition of sub-areas. Similar ideas have been explored in other network models Athreya et al. (2016), Tang (2018), and Huang, Weng, and Feng (2020). Also, by taking into account the statistical variations of embedding, it is possible to design more accurate metrics for the diversity of author research interests (e.g., using a weighted distance). Finally, results on asymptotic distributions enable us to quantify uncertainty and answer various questions via statistical inferential techniques such as hypothesis tests and confidence intervals.
- *Geometry of membership vectors*. In DCMM, the membership vectors $\{\pi_i\}$ model research interests of authors and lie in a simplex. Is there a more appropriate distance function than the common distances such as Euclidean distance for the space? Addressing this question may help us better interpret embedding results from estimated vectors $\{\hat{\pi}_i\}$, and better quantify the difference of membership vectors in problems such as evolvement of author research interests. Let $d(\cdot,\cdot)$ be a distance function. One potentially desirable property for $d(\cdot,\cdot)$ to satisfy is that $\pi_i' P \pi_j > \pi_i' P \pi_k$ whenever $d(\pi_i, \pi_j) < d(\pi_i, \pi_k)$. This property implies that if author i's research interests are more similar (in terms of the metric d) to author j's than to author k's, then author i has a higher probability to connect with author j than author k in the network, modulo degree heterogeneity of authors j and k. Other properties that account for the community structure matrix P are possible. We do not aim to advocate a specific property here, but rather bring up the issue that common distances may fall short of characterizing the complete role of membership vectors in the network.

2. Statistical Analysis of Coauthorship Networks

In the analysis of coauthorship networks, the article focused on using only the coauthorship information. The authors provided a very intuitive view on the hierarchical community structure. It

occurs to us that the articles contain many other useful information, including the title, abstract, keywords, author affiliations, all of which could be used in the network model. There are two categories of auxiliary information.

- *Incorporate the information for each article in the model.* The available information could include the title, keywords, abstract, the journal published, as well as the number of citations. Of course, some of them are unstructured information, which may require further modeling. For example, topic modeling could be applied to abstract to extract compact information. There has been a great deal of research on using the edge information in the stochastic block model (e.g., Wu, Levina, and Zhu 2017; Huang and Feng 2018).
- *Incorporate the information of each author in the model.* This could include the authors' affiliation, the year of getting Ph.D., and advisor name(s). There exist many works on incorporating these nodal information into the stochastic block model (e.g., Yan et al. 2019; Weng and Feng 2021).

Funding

Partially supported by NSF CAREER grant DMS-2013789 and NIH grant 1R21AG074205-01.

References

Athreya, A. Priebe, C. E., Tang, M., Lyzinski, V., Marchette, D. J., and Sussman, D. L. (2016), "A Limit Theorem for Scaled Eigenvectors of Random Dot Product Graphs," *Sankhya A*, 78, 1–18. [489]

Chen, K., and Lei, J. (2018), "Network Cross-validation for Determining the Number of Communities in Network Data," *Journal of the American Statistical Association*, 113, 241–251. [488]

Huang, S., and Feng, Y. (2018), "Pairwise Covariates-Adjusted Block Model for Community Detection," arXiv preprint arXiv:1807.03469. [490]

Huang, S., Weng, H., and Feng, Y. (2020), "Spectral Clustering via Adaptive Layer Aggregation for Multi-Layer Networks," arXiv preprint arXiv:2012.04646. [489]

Jin, J. (2015), "Fast Community Detection by Score," *The Annals of Statistics*, 43, 57–89. [486]

Jin, J., Ke, Z. T., and Luo, S. (2017), "Estimating Network Memberships by Simplex Vertex Hunting," arXiv preprint arXiv:1708.07852, 2017. [486,488,489]

Lei, J. (2016), "A Goodness-of-Fit Test for Stochastic Block Models," *The Annals of Statistics*, 44, 401–424. [488]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [486]

Li, T., Levina, E., and Zhu, J. (2016), "Network Cross-Validation by Edge Sampling," arXiv preprint arXiv:1612.04717. [488]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [486]

Saldana, D. F., Yu, Y., and Feng, Y. (2017), "How Many Communities are There?" *Journal of Computational and Graphical Statistics*, 26, 171–181. [488]

Tang, M., and Priebe, C. E. (2018), "Limit Theorems for Eigenvectors of the Normalized Laplacian for Random Graphs," *The Annals of Statistics*, 46, 2360–2415. [489]

Wang, Y. X. R., and Bickel, P. J. (2017), "Likelihood-Based Model Selection for Stochastic Block Models," *The Annals of Statistics*, 45, 500–528. [488]

Weng, H., and Feng, Y. (2021), "Community Detection with Nodal Information: Likelihood and its Variational Approximation," *Stat*, 11, e428. [490]

Wu, Y.-J., Levina, E., and Zhu, J. (2017), "Generalized Linear Models with Low Rank Effects for Network Data," arXiv preprint arXiv:1705.06772. [490]

Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2019), "Statistical Inference in a Directed Network Model with Covariates," *Journal of the American Statistical Association*, 114, 857–868. [490]

Yan, X., Shalizi, C., Jensen, J. E., Krzakala, F., Moore, C., Zdeborová, L., Zhang, P., and Zhu, Y. (2014), "Model Selection for Degree-Corrected Block Models," *Journal of Statistical Mechanics: Theory and Experiment*, 2014, P05007. [488]

Zhang, Y., Levina, E., and Zhu, J. (2020), "Detecting Overlapping Communities in Networks Using Spectral Methods," *SIAM Journal on Mathematics of Data Science*, 2, 265–283. [486]