Journal of Business & Economic Statistics

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

Rejoinder: "Co-citation and Co-authorship Networks of Statisticians"

Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke & Wanshan Li

To cite this article: Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke & Wanshan Li (2022) Rejoinder: "Co-citation and Co-authorship Networks of Statisticians", Journal of Business & Economic Statistics, 40:2, 499-504, DOI: 10.1080/07350015.2022.2055358

To link to this article: https://doi.org/10.1080/07350015.2022.2055358

Published online: 21 Apr 2022.

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 499–504
https://doi.org/10.1080/07350015.2022.2055358

Rejoinder: "Co-citation and Co-authorship Networks of Statisticians"

Pengsheng Ji[a], Jiashun Jin[b], Zheng Tracy Ke[c], and Wanshan Li[b]

[a]University of Georgia, Athens, GA; [b]Carnegie Mellon University, Pittsburgh, PA; [c]Harvard University, Cambridge, MA

We would like to thank all discussants for their thoughtful and stimulating comments. We are especially glad to hear that our dataset is a valuable contribution to modern large-scale datasets and that our approaches and findings will likely inspire many other research projects. Below are our responses.

1. The Two Traditions: Data-First and Model-First

We thank David Donoho for very encouraging comments. As always, his penetrating vision and deep thoughts are extremely stimulating. We are glad that he summarizes a major philosophical difference between statistics in earlier years (e.g., the time of Francis Galton) and statistics in our time by just a few words: data-first versus model-first. We completely agree with his comment that "each effort by a statistics researcher to understand a newly available type of data enlarges our field; it should be a primary part of the career of statisticians to cultivate an interest in cultivating new types of datasets, so that new methodology can be discovered and developed"; these are exactly the motivations underlying our (several-year) efforts in collecting, cleaning, and analyzing a large-scale high-quality dataset.

We would like to add that both traditions have strengths, and combining the strengths of two sides may greatly help statisticians deal with the so-called *crisis of the 21st century in statistics* we face today.

Let us explain the crisis above first. In the model-first tradition, with a particular application problem in mind, we propose a model, develop a method and justify its optimality by some hard-to-prove theorems, and find a dataset to support the approach. In this tradition, we put a lot of faith on our model and our theory: we hope the model is adequate, and we hope our optimality theory warrants the superiority of our method over others. Modern machine learning literature (especially the recent development of deep learning) provides a different approach to justifying the "superiority" of an approach; we compare the proposed approach with existing approaches by the real data results over a dozen of benchmark datasets. To choose an algorithm for their dataset, a practitioner does not necessarily need warranties from a theorem; a superior performance over many benchmark datasets says it all. To some theoretical statisticians, this is rather disappointing, as they come from a long *model-first* tradition where they believe that numerical study alone is inadequate for justifying the optimality of a method, and the best way to construct a superior method is by careful modeling and careful analysis. What is even more disappointing to them is that, frequently, over these benchmark datasets, the methods with support of optimality theorems underperform those without. This is what some statisticians call *the crisis of statistics in the 21st century*: Statistical models and methods—bread and butters to statisticians—face unprecedented challenges in finding their relevance and significance in modern scientific research, and fears that statistics will be crushed by some other fields spread on social media such as Facebook and WeChat, day after day, in recent years.

There are no easy ways to deal with such a major challenge, but many statisticians are trying. In doing so, we must combine the strengths of both traditions, and especially, put a lot more efforts in generating large-scale modern datasets. Our article is a combined effort of both traditions: On one hand, we collected and cleaned a large-scale high quality dataset, which motivates a long list of interesting problems and generates several research areas. On the other hand, to solve these problems, we need to use our training in statistical modeling and theory to develop new methods. Especially, since we emphasize on methods that are truly effective in analyzing our dataset instead of methods with strong theoretical support, our methods are more competitive in real applications. Our results will be much less satisfying if we only do one of the two. By combining the strengths of the two traditions, we believe that we can firmly keep the statistical models and theories in the central stage of modern scientific research.

2. The Dataset We Collected and Cleaned (MADStat)

While small-size datasets on scientific publications are easily accessible nowadays (e.g., by queries with Google Scholar), they are no substitute for large-scale high-quality datasets which require many online resources and web scraping techniques and demand substantial efforts in cleaning and wrangling the data.

Recent literature discusses a few well-known datasets on scientific publications (based on CiteSeer, Cora, PubMed, WebKB, and ArXiv; see *https://linqs.soe.ucsc.edu/data* and *https://getoor.soe.ucsc.edu/bio*). Compared with those sources, our dataset

CONTACT Jiashun Jin *jiashun@cmu.edu* Carnegie Mellon University, Pittsburgh, PA; Zheng Tracy Ke *zke@fas.harvard.edu*, Harvard University, Cambridge, MA.

is the first high-quality large-scale *paper-level* dataset on the publications of statisticians; it not only has many more entries (each entry being one paper) but also has more features. Our dataset offers 83,331 entries, while most earlier datasets provide no more than 4K entries (the ArXiv dataset is larger, with about 30K entries). For each entry, our dataset contains many attributes or features including title, authors, abstract, keywords, MSC subject classification, references, and citation counts.

We call our dataset *MADStat* (which stands for *Multi-Attribute Dataset on Statisticians*). Note that the dataset reported in Ji and Jin (2016) is a subset of MADStat.

In comparison, each entry of the ArXiv dataset only contains a binarized word count vector and a list of keywords. Other datasets are similarly short on features, and only one of them (CiteSeer for Entity Resolution) contains author information. Using these features in MADStat, we can tackle many problems that cannot even be properly stated based on other datasets. For example, we can use MADStat to study the citation patterns and personalized co-authorship networks of individual authors, dynamic evolvement of citations and co-authorships (for an individual or for a group of authors), and journal ranking; such studies are out of reach for alternative datasets, lacking, as they do, author attributes, publication year, or journal information. We can also apply Natural Language Processing (NLP) tools, since MADStat contains the original text of abstract of each article. Competing datasets may only contain word counts (insufficient for advanced NLP). Our forthcoming article Ke et al. (2022) uses MADStat for text learning, journal ranking, topic ranking, and citation prediction.

3. Incorporating Edge Weights in the Citee Networks

The dynamic citee network in Section 2 in our article is a collection of 21 unweighted citee networks, each for a different time window. These unweighted networks are constructed from the original weighted networks by hard thresholding the edge weights. As a result, the adjacency matrix of each unweighted network is binary, so DCMM model is natural. The DCMM model is well-studied; see Jin and Ke (2021) for a survey of recent literature.

Weng and Feng pointed out that using the DCMM model may lose some information hidden in edge weights, and proposed to study the 21 original weighted citee networks instead, modeling each of them by a Poisson-DCMM model (a variant of DCMM which assumes that the upper triangle of A contains independent Poisson variables). They made a great point by arguing that one can continue to use mixed-SCORE for analysis of Poisson-DCMM, as mixed-SCORE is a nonparametric method that is robust to parametric model specification and is expected to work well as long as the model is first-order correct. Weng and Feng also reported that the memberships inferred from a Poisson-DCMM model differ notably from a DCMM model (e.g., some nodes have purer memberships).

Weng and Feng's study is very interesting and opens door for a new line of research. We also agree that the membership matrix Π under the DCMM and the membership matrix under the Poisson-DCMM (denoted by $\widetilde{\Pi}$) can be quite divergent. For explanation, let $\widetilde{A}$ be the adjacency matrix of an original weighted network, and let A be the binary adjacency matrix by hard thresholding the entries of $\widetilde{A}$ at a threshold $t > 0$. We model A with DCMM, where $\mathbb{E}[A] = \Omega - \mathrm{diag}(\Omega)$ and $\Omega = \Theta\Pi P\Pi'\Theta$. We model $\widetilde{A}$ with Poisson-DCMM, where we similarly have $\mathbb{E}[\widetilde{A}] = \widetilde{\Omega} - \mathrm{diag}(\widetilde{\Omega})$ and $\widetilde{\Omega} = \widetilde{\Theta}\widetilde{\Pi}\widetilde{P}\widetilde{\Pi}'\widetilde{\Theta}$. By definitions, for $1 \le i \ne j \le n$,

$$\widetilde{\Omega}(i,j) = \mathbb{E}[\widetilde{A}(i,j)], \qquad \text{and} \qquad \Omega(i,j) = \mathbb{P}(\widetilde{A}(i,j) \ge t).$$

Therefore, while perhaps for some parameter range both models turn out to be reasonable, in general the two triplets, $(\widetilde{\Theta}, \widetilde{\Pi}, \widetilde{P})$ and (Θ, Π, P), can be quite different, and we should not be surprised by divergences in membership estimation. Also, the two matrices Π and $\widetilde{\Pi}$ should be interpreted differently: the former is the membership matrix where we use the co-citation counts in a conservative way (by only considering whether the count exceeds t), and the latter corresponds to a more aggressive use of co-citation counts.

We chose to use the unweighted networks for two main reasons. First, on one hand, the co-citation counts have severe heterogeneity: they may range from 1 to a few thousands for different nodes; on the other hand, co-citation counts should be largely ancillary to the membership vectors π_i: for example, an adviser and his/her advisee may have very different co-citation counts but similar research interests. We believe DCMM is more robust than Poisson-DCMM to severe heterogeneity in co-citation counts (this was also noted by Weng and Feng in Section 1.1 of their discussion). Second, from a theoretical perspective, membership estimation under DCMM has been carefully analyzed (Jin, Ke, and Luo 2017; Zhang, Levina, and Zhu 2020; Ke and Wang 2022), while Poisson-DCMM lacks such results.

Chen and Loyal also noted the possible information loss by using unweighted networks, and proposed to tackle the problem by a Latent Space Model (LSM). For $1 \le t \le 21$, let A_t be the adjacency matrix for the tth weighted citee network. They proposed to model A_t with a generalized mixed effect model:

$$g(\mathbb{E}[A_t(i,j)]) = \beta^T X_{ijt} + h_\psi(u_{it}, u_{jt}),$$

where g and h_ψ are prespecified functions, X_{ijt} are covariates, and u_{it} are latent variables similar to π_{it} in our dynamic DCMM model. Chen and Loyal further proposed to model u_{it} with a Markov process prior and obtain the posterior of u_{it} with a Markov chain Monte Carlo (MCMC) algorithm. See Sewell and Chen (2015, 2016) for details. In comparison, LSM is more flexible to incorporate edge weights and dyadic covariates than DCMM, but the MCMC algorithm for model fitting can be harder to analyze and computationally more challenging than mixed-SCORE (mixed-SCORE is a spectral method, which is computationally fast and minimax optimal (Ke and Wang 2022). It remains unclear which of the two approaches perform better in analyzing the citee networks. For limit of space, we leave the study to future work.

4. Dynamic Network Modeling

As pointed out by MacDonald, Levina and Zhu, there are two common approaches to modeling the citation counts. The first one is the *event approach*, where we treat citation counts as a stream of time-stamped events. For example, Zhu and Kolaczyk used this approach in their discussion and constructed a

dynamic citation network with directed and time stamped edges (see Section 7 for more discussions). The second one is the *aggregation approach*: we divide time into a number of windows, treat data points in each window as a *snapshot*, and aggregate the data of each snapshot to obtain a static network. We took the second approach in modeling the citee network. This approach is popular in dynamic network analysis and has some advantages. First, aggregating many time-stamped citation counts together is an important step to ensure the stability of downstream analysis. Second, aggregating the data into 21 (slightly overlapping) static networks allows us to conveniently adapt the well-studied tools for static networks (e.g., Jin, Ke, and Luo 2017; Zhang, Levina, and Zhu 2020) to analyze dynamic networks.

While MacDonald, Levina and Zhu largely agreed that the aggregation approach is a reasonable choice for dynamic network modeling, they pointed out some practical issues: (a) the window size needs to be chosen carefully, (b) there may be an identifiability issue and an alignment issue across different snapshots, (c) there may be a smoothness issue across different snapshots, and (d) the node set may not remain constant across different snapshots. Some of these issues are faced by a general dynamic network modeling strategy, not necessarily tied to the approach in our article.

For (a), we completely agree. In fact, as the statistical community has been steadily growing, in our dataset, we see far more authors per year in 2010s than in 1990s. Therefore, we allow the window sizes to vary, so that the networks corresponding to different time windows have similar numbers of nodes.

For (b)–(c), our approach was designed to tackle such issues. In the proposed *dynamic network embedding* algorithm, we create a universal embedding that embeds all nodes at all time t to the *same* low-dimensional space (i.e., the Statistics Triangle defined by the reference network). This offers an alignment for networks corresponding to different snapshots that is naturally smooth; for a detailed explanation, see the paragraphs above Theorem 2.1 of our article. McDonald, Levina and Zhu agreed that this is a solution to the alignment issue and raised a great question—how much the approach "relies on the assumption of homogeneity of the community structure matrix over time." We indeed need some temporal smoothness conditions on parameters of the dynamic DCMM model, to guarantee that the embedding, which is defined by the eigenvalues and eigenvectors of the first snapshot, maintains high signal-to-noise ratios for all snapshots. Such conditions are given explicitly in our forthcoming article (Cammarata et al. 2022). McDonald, Levina and Zhu also pointed out other approaches to network alignment in a dynamic setting, such as Procrustes analysis (Sanna Passino et al. 2021) and the omnibus embedding (Levin et al. 2017). We note that, first, these approaches still need temporal smoothness conditions to maintain high signal-to-noise ratios for all snapshots; second, they, at least in their current form, do not allow for degree heterogeneity. In comparison, our dynamic network embedding approach always accommodates degree heterogeneity. We believe our approach provides a reasonably good solution to the alignment issue and the smoothness issue. It is of great interest to study other alignment approaches and adapt them to the dynamic DCMM model, which we leave to future work.

For (d), this is an issue faced by all approaches that use the snapshot data. Fortunately, in the citee networks, most of the "leading nodes" (i.e., authors with large degrees) are also "active nodes," who remain active across the whole range of time. For the dynamic network embedding approach in our article, the effect of high-degree nodes is considerably larger than of small-degree nodes, so at least for some tasks (e.g., following the trajectory of a representative author), this issue does not have a major effect in our analysis. Furthermore, in our forthcoming article (Cammarata et al. 2022), we propose a slightly different embedding approach where instead of using the first citee network as the reference network, we use the pooled network (the network constructed by using all data points in the whole time range) as the reference network. This can largely alleviate the issue.

Loyal and Chen proposed an alternative aggregation approach, where they used the same way to construct the 21 citee networks. However, instead of modeling each of these citee networks with a DCMM model, they proposed to model it with a latent space model (LSM). This gives rise to the dynamic LSM. They proposed to analyze dynamic LSM with a Bayesian nonparametric approach, and use the results to infer changes of communities and to measure "research attraction." Loyal and Chen argued, by studying a concept called edge attraction in dynamic LSM, one can visualize co-movements of research interests of multiple authors, and also illustrate how individuals influence the research trajectories of each other; see Sewell and Chen (2015) for details. These comments suggested new research topics and pointed out new uses of the MADStat dataset, worthy of careful investigations in the future.

5. The Spectral Embedding and Visualization of the Estimated Memberships

At the heart of our citee network analysis is the SCORE embedding (Jin 2015), which produces the low-dimensional vectors $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n$. Weng and Feng raised several questions about this embedding: (a) In Figure 1, which is the better way to visualize the research triangle, the plot of $\hat{r}_1, \ldots, \hat{r}_n$ or the plot of $\hat{\pi}_1, \ldots, \hat{\pi}_n$? (b) How to derive the limiting distribution of $\hat{r}_i$, and (c) how to utilize such limiting distribution to improve community detection, diversity metric and other inference tasks? (d) What is an appropriate distance metric for r_i or π_i that can faithfully reflect the closeness of author research interests?

For (a), we think both visualization approaches are interesting, but to save space, we chose the first approach, and the main reason is that $\hat{r}_1, \ldots, \hat{r}_n$ contain more information from the raw data. To see the point, recall that $\hat{\pi}_1, \ldots, \hat{\pi}_n$ are obtained as follows. First, we use $\hat{r}_1, \ldots, \hat{r}_n$ to estimate the vertices of the Research Triangle, and use the leading eigenvalues and eigenvectors of A to obtain an estimate of b by $\hat{b}$ (see Jin, Ke, and Luo 2017 for details). We then express each $\hat{r}_i$ as a convex combination of the estimated vertices, with $\hat{w}_i$ being the resulting combination coefficient vector. Finally, letting $\tilde{\pi}_i$ be the vector where $\tilde{\pi}_i(k) = \hat{w}_i(k)/\hat{b}(k)$, $1 \le k \le K$, we obtain $\hat{\pi}_i$ by first replacing each negative entry of $\tilde{\pi}_i$ by 0 and then rescaling the resultant vector so all of its entries sum up to 1. Due to regularization in the last step, it is relatively easy to find $\hat{\pi}_i$ by $\hat{r}_i$, but harder to find $\hat{r}_i$ by $\hat{\pi}_i$. Moreover, $\hat{\pi}_i$ depends on the algorithm of estimating the vertices but $\hat{r}_i$ does not. Vertex

hunting can bring additional errors. For the above reasons, the plot of $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n$ is more informative and less prone to noise. For (b), we echo that this is an important problem, as knowing the limiting distribution of $\hat{r}_i$ and $\hat{\pi}_i$ can help many inference problems (e.g., confidence band for π_i, ranking of π_i in a multiple testing setting, and membership pairwise comparison; see, e.g., Huang, Weng, and Feng 2020). This problem is closely related to the literature of entry-wise eigenvector analysis (Tang and Priebe 2018; Abbe et al. 2020; Fan et al. 2022). In the case of severe degree heterogeneity, Ke and Wang (2022) derived sharp large-deviation bounds for every $\hat{r}_i$ and characterized precisely how these bounds vary with the individual degree parameters. This proved Weng and Feng's conjecture that the "the asymptotic covariance matrix of each $\hat{r}_i$ may vary considerably." For (c), we completely agree that it is beneficial to account for the asymptotic behavior of $\hat{r}_i$ in estimation and inference. For example, we may draw a confidence ball for each $\hat{r}_i$; since these confidence balls have different diameters, we may use them to have a better assessment of author closeness in the Research Triangle or develop a better test for the null hypothesis of $\pi_i = \pi_j$. For (d), Weng and Feng suggested that a good distance metric should satisfy some faithfulness properties such that $d(\pi_i, \pi_j) < d(\pi_i, \pi_k)$ always implies $\pi_i' P \pi_j > \pi_i' P \pi_k$. This is an interesting point. In fact, for those real data where K is small and P is strongly diagonal dominating, the Euclidean distance metrics (ℓ^2-norm or ℓ^1-norm) seem to work reasonably well for visualization and interpretation of memberships, but we agree with Weng and Feng that designing a more appropriate distance metric is practically valuable.

6. Joint Modeling of Different Data Sources

The MADStat dataset provides several different data sources, including but not limited to (a) co-authorships, (b) citation relationships, and (c) title, keywords, and abstracts (which can be used as text documents). In Section 2 of our article, we focus on a dynamic citee network constructed from (b); in Section 3, we focus on a dynamic co-authorship network constructed from (c). Seemingly, our study only covers a very small proportion of research one can do with the dataset. The discussants have suggested a few ideas for future research. Among them, joint modeling and analysis of different data sources is especially interesting, so we discuss it below.

First, several discussants (Loyal and Chen, Weng and Feng) suggested a combined analysis of the co-authorship network and the co-citation network. This is a very interesting problem. To approach it, one possibility is modeling these two networks with two different DCMM models, with some constraints on parameters (e.g., the two models share the same membership matrix). The spectral method, mixed-SCORE, in our article can be extended to this setting. Let $\hat{r}_i^{\text{coau}}$ and $\hat{r}_i^{\text{cite}}$ be the embeddings of node i in the co-authorship network and the co-citation network, respectively. We concatenate them to get an embedding

$$\hat{r}_i = \begin{bmatrix} \hat{r}_i^{\text{coau}} \\ \hat{r}_i^{\text{cite}} \end{bmatrix}, \qquad \text{where } \hat{r}_i \text{ is of dimension } 2(K-1).$$

It is not hard to see that $\hat{r}_i$ inherits the simplex geometry, as long as the two models share the same membership matrix. Therefore, we can similarly develop a spectral method for estimating the common membership matrix Π. Another possibility is suggested by Loyal and Chen, where they proposed to model the two networks with two different latent space models (LSMs) sharing the same latent space. Let $A^{(1)}$ and $A^{(2)}$ be the adjacency matrices of the co-authorship network and the co-citation network, respectively. In their discussion, they suggested the following models:

$$\text{logit}(\mathbb{E}[A^{(m)}(i,j)]) = \theta_i^{(m)} + \theta_j^{(m)} + u_{it}^T \Lambda^{(m)} u_{jt}, \qquad m \in \{1,2\}.$$

Here, the latent variables u_{it} are shared by two models. Similar to the DCMM approach, the LSM approach also pools information of two networks.

Moreover, Weng and Feng suggested a combined analysis of the networks with the text documents (title, abstract and keywords) in our dataset. This is a great idea, and in fact, in our forthcoming article (Ke et al. 2022), we have done two lines of research. In the first one, we combine ideas on journal ranking and text learning and propose the Hoffman–Stigler model as a new model for jointly modeling citation counts and article abstracts. We then analyze it by the topic-SCORE algorithm (Ke and Wang 2017) and use the results to identify representative topics in statistics, study how topic weights of a given author evolve over time, identify the friendliest journal for a given topic, and perform topic ranking and journal ranking. In the second line, we extract 22 features by combining the text learning results above with manual efforts and use them to predict whether a given article will be highly cited in the near future.

Finally, Weng and Feng also suggested us to combine the MADStat dataset with other data resources, such as the mathematical genealogy, for analysis. This is a very interesting suggestion, as the adviser-advisee relationship is one of the most important co-authorship patterns; see Section 3 of our article. If we have the mathematical genealogy data, we can have a more careful study on how the relationship of adviser-advisee affects the long-term co-author relationships and evolvement of research interest. To incorporate such additional features to our network analysis, we may use the LSM approach. In the discussion of Loyal and Chen, they mentioned that the LSM framework can admit dyadic attributes such as the advisor-advisee dummy and geographical proximity between nodes. They suggested to use this model-based approach to study those factors that affect the formation of collaboration. These are all great suggestions, which we leave to future work.

7. Counting Motifs, Graphlets, and Cycles

Zhu and Kolacyzk raised an excellent point that we may gain interesting insights of the networks by counting the numbers of small-size subgraphs (e.g., motifs, cycles, graphlets). Especially, by treating the citation counts in MADStat as a time-stamped stream of events, they closely investigated the frequencies of 36 motifs in four different settings, and discovered some interesting patterns of these motifs. For example, they found that the reciprocal citations across time occur relatively rare in the statistical community. Their study points out a new use of the MADStat dataset and opens doors for new research.

In connection with their study, we proposed to apply the SgnQ test on personalized networks to measure the coauthor-

ship diversity and citation diversity of individual authors; see Section 3.3 of our article. The SgnQ statistic is a member of the class of Signed-Polygon test statistics (Jin, Ke, and Luo 2021) constructed from cycle counts. Such test statistics can be viewed as some (properly centered and normalized) motif counts in a symmetrical network, with the appealing property of a limiting distribution of $N(0, 1)$ under a null DCMM model with $K = 1$. This poses an interesting question: Is it possible to borrow the idea of SgnQ to develop a statistic from the temporal motif counts such that it has a tractable distribution? We believe this is possible. Assuming a dynamic DCMM model with $K = 1$, we can estimate the mean and standard deviation of the temporal motif counts and come up with a properly standardized test statistic. At least for those two-node and three-node temporal motifs discussed by Zhu and Kolacyzk, it is feasible to derive the asymptotic distributions of such test statistics. We leave the study to future work. It is worth mentioning that Zhu and Kolaczyk (2022) and Chang, Kolaczyk, and Yao (2022) have studied the distributions of temporal motif counts, have studied the distributions of temporal motif counts in some related but different settings.

We further point out some other applications of temporal motif counts in the MADStat dataset. First, we can use the *personalized motif counts* (i.e., count of motifs in a properly defined ego dynamic citation network of a given author) to measure the citation diversity of this author. Second, the personalized motif counts can be used for citation prediction. Given an author, the problem of citation prediction is to use his/her past citation patterns to predict his/her total citation counts in the next 5 years (say). In our forthcoming article (Ke et al. 2022), we use the MADStat dataset to extract 22 features and show that these features are relatively powerful in predicting future citations. Zhu and Kolacyzk mentioned that the motifs M34-36 reflect the broad impact of some seminal works and that if an individual frequently serves as the top left node in their motifs M34-36 (see Figure 1 of their discussion), then he/she is likely to receive high citations. These findings suggest that the counts of some particular motifs may be predictive for future citations.

8. Goodness of Fit (GoF) and Model Diagnostics

The DCMM model allows for severe degree heterogeneity and mixed-memberships, and achieves a good balance between practical feasibility and mathematical tractability. An interesting question is whether DCMM is adequate for most real networks. Weng and Feng proposed a deviance residual plot for model diagnostics, and their results suggest that, at least for the reference citee network, the DCMM model is adequate.

Weng and Feng's approach is very interesting, but they did not provide a goodness-of-fit (GoF) test that can output an explicit p-value. From a practical perspective, it is desirable to have a GoF metric with an explicit limiting null distribution. We now borrow the ideas of model fitting and cycle counting (Jin et al. 2022) to propose such a GoF metric. Given a symmetric network with K communities, we test whether it satisfies a DCMM model with K communities (i.e., goodness of fit). We prefer not to specify the alternative hypothesis, leaving it flexible to incorporate various cases where the assumed model does not hold (e.g., misspecified K, outlier nodes, edge dependency, etc.).

Our approach is a 4-step recipe. In step 1, we estimate Π by a spectral method (e.g., mixed-SCORE). In Step 2, we estimate Θ and P by refitting the adjacency matrix A using the estimated Π. This gives rise to an estimate of Ω, denoted by $\widehat{\Omega}$. In step 3, we apply a cycle count statistic (see Section 7 and Jin, Ke, and Luo 2021) to the matrix $\widehat{A} = A - \widehat{\Omega}$. In Step 4, we standardize the statistic by its estimated mean and standard deviation. Details are in the forthcoming article (Jin and Ke 2022). In this recipe, Steps 1–2 share a similar spirit as the approach of Weng and Feng by creating a residual matrix $A - \widehat{\Omega}$ (Weng and Feng also used mixed-SCORE to estimate Π first, but their refitting procedure to obtain $\widehat{\Omega}$ is different), and Steps 3–4 serve to create a GoF metric with a known limiting null distribution.

The above approach has been justified in the simpler DCBM setting (i.e., the network satisfies a DCBM model with K communities in the null hypothesis, where DCBM is a special case of DCMM with no mixed-memberships). In this case, we use SCORE (Jin 2015) as the spectral method in Step 1, and our recipe coincides with one step of the StGoF algorithm (Jin et al. 2022) at $m = K$ (StGoF is a stepwise algorithm where we run a GoF test successively for $m \geq 1$). By Theorem 3.1 of Jin et al. (2022), under the null hypothesis, the test statistic converges to $N(0, 1)$ in law as n diverges to ∞, and so we can use it as a GoF metric. For the DCMM setting of interest here, we follow the same recipe but use mixed-SCORE as the spectral method in Step 1 and modify Steps 2–4 to accommodate mixed memberships; the study of the asymptotic null distribution of the GoF metric is technically more demanding, and details are in Jin and Ke (2022).

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020), "Entrywise Eigenvector Analysis of Random Matrices with Low Expected Rank," *The Annals of Statistics*, 48, 1452–1474. [502]

Cammarata, L., Jiang, K., Jin, J., and Ke, Z. T. (2022), "Estimating Dynamic Mixed Memberships by Trajectory Embedding," (manuscript) [501]

Chang, J., Kolaczyk, E. D., and Yao, Q. (2022), "Estimation of Subgraph Density in Noisy Networks," *Journal of the American Statistical Association*, 117, 361–374. [503]

Fan, J., Fan, Y., Han, X., and Lv, J. (2022), "SIMPLE: Statistical Inference on Membership Profiles in Large Networks," *Journal of the Royal Statistical Society*, Series B (to appear). [502]

Huang, S., Weng, H., and Feng, Y. (2020), "Spectral Clustering via Adaptive Layer Aggregation for Multi-layer Networks," arXiv:2012.04646. [502]

Ji, P., and Jin, J. (2016), "Coauthorship and Citation Networks for Statisticians," (with discussions), *The Annals of Applied Statistics*, 10, 1779–1812. [500]

Jin, J. (2015), "Fast Community Detection by SCORE," *The Annals of Statistics*, 43, 57–89. [501,503]

Jin, J., and Ke, Z. T. (2021), "The SCORE Normalization, Especially for Heterogeneous Network and Text Data," (manuscript). [500]

——— (2022), "A Goodness-of-Fit Test for Social Networks," (manuscript). [503]

Jin, J., Ke, Z. T., and Luo, S. (2017), "Estimating Network Memberships by Simplex Vertex Hunting," arXiv:1708.07852. [500,501]

——— (2021), "Optimal Adaptivity of Signed-Polygon Statistics for Network Testing," *The Annals of Statistics*, 49, 3408–3433. [503]

Jin, J., Ke, Z. T., Luo, S., and Wang, M. (2022), "Optimal Estimation of the Number of Network Communities," *Journal of the American Statistical Association* (to appear) DOI: 10.1080/01621459.2022.2035736. [503]

Ke, Z.T., Ji, P. Jin, J. and Li, W. (2022), "Recent Advances in Text Learning and a Case Study,"" (manuscript). [500,502,503]

Ke, Z. T., and Wang, J. (2022), "The Minimax Rates of Network Membership Estimation Under Severe Degree Heterogeneity," (manuscript). [500,502]

Ke, Z. T., and Wang, M. (2017), "A New SVD Approach to Optimal Topic Estimation," arXiv:1704.07016. [502]

Levin, K., Athreya, A., Tang, M., Lyzinski, V., and Priebe, C. E. (2017), "A Central Limit Theorem for an Omnibus Embedding of Multiple Random Dot Product Graphs," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 964–967. IEEE. [501]

Sanna Passino, F., Bertiger, A. S., Neil, J. C., and Heard, N. A. (2021), "Link Prediction in Dynamic Networks Using Random Dot Product Graphs," *Data Mining and Knowledge Discovery*, 35, 2168–2199. [501]

Sewell, D. K., and Chen, Y. (2015), "Latent Space Models for Dynamic Networks," *Journal of the American Statistical Association*, 110, 1646–1657. [500,501]

——— (2016), "Latent Space Models for Dynamic Networks with Weighted Edges," *Social Networks*, 44, 105–116. [500]

Tang, M., and Priebe, C. (2018), "Limit Theorems for Eigenvectors of the Normalized Laplacian for Random Graphs," *The Annals of Statistics*, 46, 2360–2415. [502]

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. [500,501]

Zhu, X., and Kolaczyk, E. D. (2022), "Quantifying Uncertainty for Temporal Motif Estimation in Graph Streams under Sampling," arXiv:2202.10513. [503]