JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 469–485
https://doi.org/10.1080/07350015.2021.1978469

Taylor & Francis
Taylor & Francis Group

Co-citation and Co-authorship Networks of Statisticians

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

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

ABSTRACT

We collected and cleaned a large dataset on publications in statistics. The dataset consists of the co-author relationships and citation relationships of 83,331 articles published in 36 representative journals in statistics, probability, and machine learning, spanning 41 years. The dataset allows us to construct many different networks, and motivates a number of research problems about the research patterns and trends, research impacts, and network topology of the statistics community. In this article we focus on (i) using the citation relationships to estimate the research interests of authors, and (ii) using the co-author relationships to study the network topology. Using co-citation networks we constructed, we discover a “statistics triangle,” reminiscent of the statistical philosophy triangle (Efron 1998). We propose new approaches to constructing the “research map” of statisticians, as well as the “research trajectory” for a given author to visualize his/her research interest evolvement. Using co-authorship networks we constructed, we discover a multi-layer community tree and produce a Sankey diagram to visualize the author migrations in different sub-areas. We also propose several new metrics for research diversity of individual authors. We find that “Bayes,” “Biostatistics,” and “Nonparametric” are three primary areas in statistics. We also identify 15 sub-areas, each of which can be viewed as a weighted average of the primary areas, and identify several underlying reasons for the formation of co-authorship communities. We also find that the research interests of statisticians have evolved significantly in the 41-year time window we studied: some areas (e.g., biostatistics, high-dimensional data analysis, etc.) have become increasingly more popular. The research diversity of statisticians may be lower than we might have expected. For example, for the personalized networks of most authors, the *p*-values of the proposed significance tests are relatively large.

ARTICLE HISTORY
Received April 2021
Accepted August 2021

KEYWORDS
Citation; Co-authorship; Community detection; Dynamic network; Hierarchical community tree; Mixed membership estimation; Personalized network

1. Introduction

In the past decades, the size of the scientific community has grown substantially. The rapid growth of the scientific community motivates many interesting Big Data projects, and one of them is how to use the vast volume of publications of a scientific field to delineate a complete picture of the research habits, trends, and impacts of this field. These studies are useful for examining national and global scientific publication-related activities, ranking universities, and making decisions of funding, promotions, and awards.

There are two main approaches to studying scientific publications, the subjective approach and the quantitative approach. The subjective approach is more traditional, but it is time-consuming and susceptible to bias. The quantitative approach (which uses statistical tools for analyzing such data) is comparably inexpensive, fast, objective, and transparent, and will play an increasingly more important role (Silverman 2016).

From a statistical standpoint, most existing quantitative approaches are overly simple, using preliminary metrics (e.g., counts of articles or citations) for analysis. The h-index and journal impact factor are examples of more sophisticated approaches, but they are still not principled statistical methods. Statistical modeling of publication data is a significantly underdeveloped area, where we have only a small number of interesting articles, sparsely scattered over the spectrum, and typically, each focusing on only a specific problem.

On the other hand, this can also be viewed as a golden opportunity for statisticians. The publication data provide a valuable data resource, important problems in science and social science, and interesting Big Data projects that demand sophisticated statistical tools. Having seen such an opportunity, Hall encouraged statisticians to take on a more active role in such research (Hall 2011). Hall’s viewpoint is shared by Donoho (2017), among others. In his illuminating article “50 Years of Data Science” (Donoho 2017), Donoho predicted that “science about data science” will become one of the major divisions of data science, and one task of this division is to evaluate scientific research outputs.

This article is a response to the call by Hall and others. We contribute a large-scale high-quality dataset on the publications of statisticians and use it to showcase how modern statistical tools can be employed for analysis of such kind of data.

1.1. A New Dataset About the Publications of Statisticians

We present a new dataset about the publications of statisticians, collected and cleaned by ourselves with enormous efforts. The

CONTACT Pengsheng Ji psji@uga.edu Department of Statistics, University of Georgia, 310 Herty Drive, Athens, 30602-0002.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

dataset consists of co-author relationships and citation relationships of 83K research articles published in 36 representative journals in statistics, probability, machine learning, and related fields, spanning 41 years. See the table below. More information of these journals is presented in Table B.1 of the supplement.

#Journals	Time span	#Authors	#Articles
36	1975–2015	47,311	83,331

One might think that the dataset is easy to obtain, as BibTeX and citation data seem to be easy to download. Unfortunately, when we need a large-volume, high-quality dataset, this is not the case. For example, the citation counts from Google Scholar are not always accurate, and many online resources do not allow for large volume downloads. Our data were downloaded from a handful of online resources by techniques including but not limited to web scraping. The dataset was also carefully cleaned by a combination of manual efforts and computer algorithms we developed. Both data collection and cleaning are sophisticated and time-consuming processes, during which we had to overcome a number of challenges. For a detailed discussion on data collection and cleaning, see Section B.2 of the supplement.

1.2. Results, Findings, and Challenges

First, we overview the results. Our dataset provides rich material for research and motivates many interesting problems for research trends, patterns, and impacts of the statistics community. In this article, we focus on two topics: (1) How to use the citation data to estimate the research interests of statisticians, and (2) How to use the co-authorship data to study the network topology of statisticians.

Section 2 studies the first topic. How to model the research interests of an author is an open problem in bibliometrics. Our idea is to first use the co-citation relationships to construct a *citee network* and then model the research interests of the author as the mixed-memberships he/she has over different network communities. This gives rise to the degree-corrected mixed-membership (DCMM) model (Jin, Ke, and Luo 2017). Such a framework allows us to use principled statistical tools to attack problems about research interests. Specifically, we develop new models, methods, and theory for (i) estimating the research interests of authors, (ii) clustering authors by research interests, (iii) studying how the research interests of an author evolve over time, and (iv) measuring the research interest diversity of individual authors. We discover a "Research Map" (a cloud of points in $\mathbb{R}^2$, each representing the research interests of an author), which consists of a "statistics triangle" and 15 sub-regions. The vertices of the triangle represent the three primary research areas in statistics: "Bayes," "Biostatistics," and "Nonparametric," and each sub-region represents an interpretable sub-area in statistics. The relative position of each author to the three vertices represents the weights of his/her research interests in the three primary areas. We also develop a new algorithm that allows us to plot the "research trajectory" on the "Research Map" for an author to visualize the evolvement of his/her research interests over time, and propose two new metrics to measure the citation diversity of individual authors.

Section 3 studies the second topic, where the focus is community detection. We develop new models and methods for (i) hierarchical clustering, (ii) dynamic clustering, and (iii) measuring the co-authorship diversity. For (i), we develop a new approach and build a 4-layer community tree with 26 leaves. Each leaf represents an interpretable co-authorship community where the authors may have some ties (e.g., colleagues, advisor-advisee) or share something (e.g., research interests or geological location) in common. For (ii), we use a Sankey plot to visualize the birth and growth of some communities and the migration of authors among different communities. For (iii), we propose a new idea to measure the research diversity of an author, by constructing the so-called "personalized networks."

Second, we discuss our findings. First, it is debatable what are *primary areas* and *representative sub-areas* in statistics. In Section 2, we suggest that "Bayes," "Biostatistics," and "Nonparametric" are the three primary areas in statistics, and identify 15 representative sub-areas. The "statistics triangle" is reminiscent of Efron's triangle of statistical philosophy (Efron 1998), where the three vertices are "Bayes," "Fisherian," and "Frequentist." Note that our triangle is based on data while Efron's triangle is more philosophical. Second, in the 41-year time span of our dataset, the research community of statistics has undergone significant changes: Some research areas (e.g., biostatistics) have become much more popular. Some research areas (e.g., nonparametric and semiparametric regressions) have significantly shifted the focus (e.g., with a significant surge of interest in high-dimensional data analysis after 2000). Last, the research of statisticians may be less diverse than expected: most researchers continue to collaborate with the same cluster of people over many years, with a large p-value for the significance test over his/her personalized network.

Last, we discuss some challenges we face. Getting meaningful results from a large dataset is never easy (let alone the time and efforts required for obtaining the dataset). We need new methods for computing trajectories in Section 2.2 and for constructing hierarchical community tree in Section 3.1. We also need new ideas to relate research interests to network mixed-memberships in Section 2.1 and to connect research diversity of an author to a network global testing problem on his/her personalized networks in Section 3.3.

Even with a handful of new approaches we develop, we still face great challenges: how to properly construct the network and choose the model, how to make inference, and how to interpret the results. To deal with such challenges, we need many new ideas. For example, in Section 2, we discover that ignoring some "old" citations makes the constructed citee network more useful. We also find that, to get meaningful results, it is critical to use a network model that allows for severe degree heterogeneity. Also, in our study for "research trajectory," we find that naively applying existing spectral approaches may face challenges, and to overcome the challenge, we propose *dynamic network embedding* as a new approach to dynamic network analysis. There are many such examples in Sections 2 and 3.

In summary, our findings are the combined results of (a) a large-scale high-quality dataset we collected, (b) many new approaches we developed, and (c) many new ideas and substantial efforts in data analysis. We will make our dataset and code available so researchers can conveniently use our study as a template to study other research communities.

1.3. Contributions, Broader Impacts, and Disclaimers

We have several major contributions. First, we contribute a high-quality, large-scale dataset, which provides material for research in bibliometrics, statistics, and data science. Second, we set an example for how quantitative analysis of large publication data can be executed. We create a template where we showcase how to use modern statistical tools to study a vast volume of publication data. We build large co-authorship and co-citation networks, propose new network models, and demonstrate how to use the output to label research areas, identify latent communities, and measure research diversities. While we use the statistics community as our object of study in this template, our approaches (data collection, research template, methods, and theory) are easily extendable to study other scientific communities (e.g., economics). Third, while our focus is on the new dataset, we also contribute in methods and theory. We introduce a handful of methods for network data analysis; some are new, and some are carefully adapted from the recent literature. Our approaches to computing research trajectory, building community tree, and measuring research diversity are especially novel. Last but not the least, as statisticians, we know partial ground truth of our community. For this reason, our dataset may provide a benchmark for comparing different methods in statistics, machine learning, and especially network analysis, and so largely help the development of methods and theory in these areas.

Our study has (potential) impacts in science, social science, and even real life. It provides an array of ready-to-use and easy-to-extend statistical tools which the administrators, award committee, and individuals can use to study the research profile of an individual, an area, or the whole statistics community. For example, suppose a committee wishes to learn the research profile of an individual researcher. Our study provides a long list of tools to help characterize and visualize the research profile of the researcher: his/her research interests and his/her position on the Research Map, his/her research interest trajectory, to which network community he/she belongs, his/her research diversity in terms of citation and in terms of co-authorship, his/her personalized networks, the importance of his/her research area, his/her research impact and ranking relative to his/her peers. Such information is not available from his/her curriculum vitae or profile on Google Scholar, and can be very useful for the award committee or administrators for decision making.

Our study also provides a useful guide for researchers (especially, junior researchers) in selecting research topics, looking for references, and building social networks. It also helps understand several important problems in social science and science: characterizing research evolvement, predicting emerging communities and significant advancement in each research area, checking whether the development of different areas is balanced, and identifying unknown biases in publications. We discuss these with more details in Section 4.

For disclaimers, note that we have to use real names as our data are about real-world publications, but we have not used any information that is not publicly available. It is not our intention to rank a researcher (or an article, or an area) over others. While we tried very hard to create a high-quality dataset, the time and effort one can invest is limited, so is the scope of our study; as a result, some of our results may have biases. Our article can be viewed as a starting point for an ambitious task, where we create a research template with which the researchers in other fields (e.g., economics) can use statisticians' expertise in data analysis to study their own fields. For this reason, the main contributions of our article are still valid. See Section A of the supplement for a longer version of the disclaimers.

1.4. Contents

Section 2 studies co-citation networks, where the focus is to study how to estimate the research interests of an author and how the research interests evolve over time. Section 3 focuses on co-authorship networks. It studies hierarchical and dynamic community detection, and proposes two new diversity measures. Section 4 is the conclusion.

2. Learning Research Interests by Co-Citation Networks

A good understanding of the research interests of statisticians helps understand the research trends, research impacts, and network topology of the statistics community, and also helps understand the research profile of individual statisticians. For example, suppose we are given an author with a total of 1000 citation counts. To decide whether he/she is highly cited, it is crucial to understand his/her major areas of interest, because the average citation count for a researcher in one area may be a few times higher than that of another.

The citation counts in our dataset provide a valuable resource to study the research interests. In this section, we consider four problems: (a) how to model the research interests of individual authors; (b) how to estimate his/her research interests and how to use the estimated research interests for author clustering; (c) how to study the dynamic evolvement of research interests of an author; (d) how to measure the diversity of research interests of an author. We propose new approaches to studying (a)–(d). Below is a sketch of our ideas.

Consider Problem (a) first. How to model research interests of individual authors is an open problem. We observe that two authors being frequently cited together in the same articles (i.e., co-cited) indicates that their works are scientifically related and that they share some common research interests. Motivated by this, we propose the following approach to tackling Problem (a). First, we use the co-citation relationship to construct an undirected network which we call the *citee network* (see Section 2.1). We assume that the citee network has K communities, each representing a primary research area in statistics (primary areas can be further divided into sub-areas). For author i, we model his/her research interest as a weight vector $\pi_i \in \mathbb{R}^K$, with $\pi_i(k)$ being the fraction of his/her interest in community k, $1 \leq k \leq K$. We further model the citee network with the recent *Degree Corrected Mixed-Membership (DCMM)* model, where π_i are the vectors of mixed-memberships.

In a network, communities are tight-knit groups of nodes that have more edges within than between (Goldenberg et al.

2010). For example, suppose $K = 3$ and we have three communities, each being a primary area in statistics: "Bayes," "Biostatistics," and "Nonparametric." Suppose for author i, $\pi_i = (0.5, 0.3, 0.2)'$. In this case, we think author i has 50%, 30%, and 20% of his research interest or impact in these primary areas, respectively.

The DCMM model is a recent network model (Jin, Ke, and Luo 2017; Zhang, Levina, and Zhu 2020). It models both severe degree heterogeneity and mixed-memberships and is reasonable for the current setting. Let $A \in \mathbb{R}^{n,n}$ be the adjacency matrix of the citee network, where $A(i,j) = 1$ if $i \neq j$ and there is an edge between nodes i and j and $A(i,j) = 0$ otherwise. As above, let π_i be the K-dimensional vector that models the research interests of author i, $1 \leq i \leq n$. For a nonnegative, unit-diagonal matrix $P \in \mathbb{R}^{K,K}$ that models the community structure and parameters $\theta_1, \theta_2, \ldots, \theta_n > 0$ that model the degree heterogeneity, we assume that the upper triangle of A contains independent Bernoulli variables, where for any $1 \leq i < j \leq n$,

$$\mathbb{P}\big(A(i,j) = 1\big) = \theta_i\theta_j \sum_{k,\ell=1}^{K} \pi_i(k)\pi_j(\ell)P(k,\ell) = \theta_i\theta_j \cdot \pi_i' P \pi_j. \tag{2.1}$$

This provides a reasonable model for the research interests of individual authors, and addresses an interesting problem in social science and bibliometrics.

Consider Problems (b) and (c). We first use the mixed-SCORE (Jin, Ke, and Luo 2017) to estimate the research interests of individual authors. We discover a *statistical triangle* and build the *Research Map* for statisticians. We then develop a new idea to compute the research trajectory of an author. To this end, we need a new clustering algorithm for building the research map, and a new algorithm to draw the trajectory. We now discuss them separately.

The clustering problem is well-studied (e.g., Zhao, Levina, and Zhu 2011; Amini et al. 2013, among others). Unfortunately, these algorithms have focused on the DCBM model (Karrer and Newman 2011). Compared to the DCMM model in (2.1), DCBM requires each π_i to be degenerate (one entry is 1, all other entries are 0), and is not appropriate for the citee network considered here. Our idea is to combine mixed-SCORE (Jin, Ke, and Luo 2017) with classical clustering algorithms. Suppose we have estimated the research interest vectors $\pi_1, \pi_2, \ldots, \pi_n$ by mixed-SCORE, and let $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n$ be the estimates. We view this step as a dimension reduction step, and propose an author clustering algorithm where we directly apply k-means to $\hat{\pi}_1, \ldots, \hat{\pi}_n$. Compared to existing clustering algorithms, our method works for the DCMM model where we allow mixed-memberships, and so is different.

The problem of estimating the trajectory is related to the problem of dynamic mixed-membership analysis. Consider a sequence of citee networks, each for a different time window. We extend the DCMM model for static networks in (2.1) to dynamic networks, where π_i may vary with time. In such a setting, how to estimate π_i is largely an open problem. Related works include Kim et al. (2018) and Liu et al. (2018), but these articles focus on settings where each static network satisfies the MMSB model (a special DCMM where we do not allow degree heterogeneity). For this reason, it is unclear how to extend their approaches to our setting. The approach of naively applying mixed-SCORE to each individual network in our setting does not work well either; see Section 2.2.

We propose the *dynamic network embedding* as a new approach to analyzing dynamic DCMM. For each author in our dataset, the approach produces a *research trajectory* which visualizes how his/her research interests evolve over time. Compared with the approach where we naively apply mixed-SCORE to each network in our setting separately, two approaches are the same for the first time window, but are significantly different for all other time windows; the new approach is more satisfactory both numerically and theoretically.

Consider Problem (d). How to measure the diversity of the research interests of individual authors is a problem of great interest. Using the *research trajectory* developed for Problem (c), we propose two diversity metrics: One measures the *significance* of research interest expansion of an author and the other measures his/her *persistence* of research interest expansion. Compared with other diversity metrics, our metrics are new, for they are based on our proposed new approach to estimating the research trajectories.

Sections 2.1, 2.2, and 2.3 discuss Problem (b), (c), and (d) respectively. Note that Problem (a) is already fully addressed.

2.1. Estimation of Research Interests, Author Clustering

We construct a citee network using the co-citations during 1991–2000. We limit the time to 1991–2000, for later we will use this network as a reference network to study the research trajectories of selected authors. For each year t, $1991 \leq t \leq 2000$, define a year-t weighted network where each node is an author, and for any two nodes i and j, the weight of the edge between them is the number of times that the articles by author i published between year $t - 9$ to t and the articles by author j published between year $t - 9$ and t have been cited *together* in an article by another author published in year t. This results in a weighted adjacency matrix for year t. Summing the adjacency matrices for $t = 1991, 1992, \ldots, 2000$ gives rise to a weighted network. Let the degree of node i be the sum of weights of edges between node i and the other nodes. We remove all nodes with a degree smaller than or equal to 60, and define a symmetric unweighted network using the remaining nodes, where two nodes have an edge if and only if the weight between them in the previous network is no less than 2. We call the giant component of this network the citee network for 1999 and 2000, which has 2831 nodes (these nodes form a subset of most active and most cited authors).

There are different ways to construct the citee network (we have studied many options and recommend the one above). We restricted to "fresh" citations only (a citation from one article to the other is considered "fresh" if the two publication times are no more than 10 years apart). We have removed low-degree nodes and low-weight edges in the intermediate weighted graph to reduce noise. In Section C.3 of the supplement, we have also studied the case where the threshold 60 is replaced by 50 and 70, and observed similar results (e.g., similar triangle and research map for statisticians). Thresholding the edge weights is a common practice. It may cause some information loss. But since the goal is to identify active communities, it is unclear how such a loss may affect the results. Also, just as in different fields of science, the average citations (per article or author) can vary dramatically in different areas. For this reason, we may threshold the edge weights adaptively with different thresholds for different areas. However, it is not immediately clear how to implement such an approach. We leave these studies to the future.

We wish to use this citee network to study the research interests of individual authors. We model this network with the aforementioned DCMM model (2.1). Under this model, each of the K communities can be interpreted as a research area, and the research interest of author i is modeled by the mixed-membership vector $\pi_i \in \mathbb{R}^K$. How to estimate π_i is known as the problem of mixed-membership estimation, where we use the method mixed-SCORE (Jin, Ke, and Luo 2017). The approach uses *SCORE embedding* which embeds all authors to a low dimensional space and provides a way to visualize the research interest of each author. Specifically, let $\hat{\xi}_1, \dots \hat{\xi}_K \in \mathbb{R}^n$ be the first K eigenvectors of the adjacency matrix. Each node i is embedded into a $(K-1)$-dimensional space by the vector

$$\hat{r}_i = \left[\hat{\xi}_2(i)/\hat{\xi}_1(i),\ \hat{\xi}_3(i)/\hat{\xi}_1(i),\ \dots, \hat{\xi}_K(i)/\hat{\xi}_1(i)\right], \qquad 1 \le i \le n. \tag{2.2}$$

Now, first, the embedded points are approximately contained in a *simplex with K vertices* in $\mathbb{R}^{K-1}$, where each vertex represents a community. Second, each embedded point $\hat{r}_i$ is approximately a convex combination of the vertices: $\hat{r}_i \approx \sum_{k=1}^K w_i(k) v_k$, where $v_1, v_2, \dots, v_K$ are the vertices of the simplex. The weight vector w_i is an order-preserving transformation of π_i, in the sense that $w_i \propto \pi_i \circ b$, where $\circ$ is the Hadamard product and $b \in \mathbb{R}^K$ is a positive vector (not depending on i). Therefore, if an embedded point $\hat{r}_i$ is close to one vertex, then w_i is nearly degenerate (with only one nonzero entry that is 1), and node i is a pure node (i.e., node i is called a pure node of community k if $\pi_i(k) = 1$ and $\pi_i(\ell) = 0$ for all $\ell \neq k$). If $\hat{r}_i$ is deeply in the interior of the simplex, then all entries of w_i are bounded away from 0 and node i is highly mixed; see Jin, Ke, and Luo (2017) for more discussions.

Why $K = 3$ is the Most Reasonable Choice. To use mixed-SCORE, we need to decide K, which is unknown. First, we use the scree plot of the adjacency matrix to determine the range of K as $[2, 6]$. Second, we implemented mixed-SCORE for each $K \in \{2, 3, \dots, 6\}$ and investigated the goodness of fit, by checking whether the rows of $\hat{R}$ fit the aforementioned $(K-1)$-dimensional simplex structure (it is hard to visualize the simplex when $K \ge 4$, so we plot two coordinates of $\hat{r}_i$'s at a time to visualize a projection of the simplex to $\mathbb{R}^2$). Last, for each K, we manually check the large-degree pure nodes in each community and see whether the results fit with our knowledge of the statistics community. The above analysis suggests $K = 3$ as the best choice. See Section C.2 of the supplement for details.

The Statistics Triangle. Since $K = 3$, the simplex in SCORE embedding is a triangle, each vertex representing (perceivably) a primary statistical research area. See Figure 1. To interpret these areas, we apply mixed-SCORE to the citee network with $K = 3$, and obtain an estimate for the membership vectors $\pi_1, \pi_2, \dots, \pi_n$ by $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_n$. We divide all the nodes into three groups: If the largest entry of $\hat{\pi}_i$ is the kth entry, then node i is assigned to group k, $1 \le k \le 3$. In Section C of the supplement, we investigate the research interests of authors in each group, using the topic weights estimated from abstracts of their articles. It suggests that the three vertices represent three primary research areas: "Bayes," "biostatistics," and "nonparametric statistics." This triangle is reminiscent of the *statistics philosophy triangle* by Efron (1998), where the three vertices are "Bayes," "Fisherian," and "frequentist." Efron argued that they are the three major philosophies in statistics, and most statistics methodologies (e.g., bootstrap) can be viewed as weighted averages of these three philosophies. Different from Efron's triangle, our statistics triangle is data-driven.

The Research Map. Perceivably, we can further split each primary area into sub-areas, and a convenient approach is to use SCORE embedding. For each author i in the citee network, $1 \le i \le n$, since $K = 3$, $\hat{r}_i$ can be viewed as a point in $\mathbb{R}^2$. The distance between authors in this space is a measure of closeness of their research areas. Therefore, it makes sense to further cluster the authors into sub-areas by applying the K-means algorithm to $\{\hat{r}_i\}_{i=1}^n$. We have tried the K-means algorithm with $L = 10, 11, \dots, 20$ clusters, and picked $L = 15$ due to that the result is most reasonable. We then apply the K-means with $L = 15$ and obtain 15 clusters, each of which can be interpreted as a sub-area in statistics after a careful investigation of the research works by representative authors in the cluster (while we try very hard to find a reasonable label for each cluster, we should not expect that a simple label is able to explain the research interests of all authors in the cluster).

Figure 1 shows the 15 clusters and their labels, which we call the *research map* of the citee network. In this map, each point represents $\hat{r}_i$ for some node i, $1 \le i \le n$, and the two axes are the two entries of $\hat{r}_i$, respectively. The statistics triangle is illustrated by the dashed green lines, where the three vertices are estimated by mixed-SCORE and represent the three primary areas "Bayes," "Biostatistics," and "Nonparametric." We also present the Voronoi diagram for the clusters (boundaries are illustrated by dashed blue lines), and the names for the 5 authors with the largest degrees in each cluster.

For each author, his/her position on the research map illustrates the weight his/her citation has in each of the three primary areas. For example, Raymond Carroll and Bradley Efron are located deeply in the interior of the triangle, suggesting that their citations between 1991 and 2000 have substantial weights in each of the three primary areas. Authors who are located around each corner of the triangle include Nicholas Polson ("Bayes"), Michael Proschan ("Biostatistics"), and Theo Gasser ("Nonparametric"), suggesting that their citations between 1991 and 2000 are mostly from one community. Note that, since the results are based on the citee network, the areas from which an author attracts citations may not be exactly the same as the areas he/she works on. For example, though Donald B. Rubin rarely works in *Longitudinal I (GEE)*, he is clustered to GEE for he is cited together with quite a few authors in GEE (e.g., Scott Zeger, Nan Laird, and Daniel F. Heitjan).

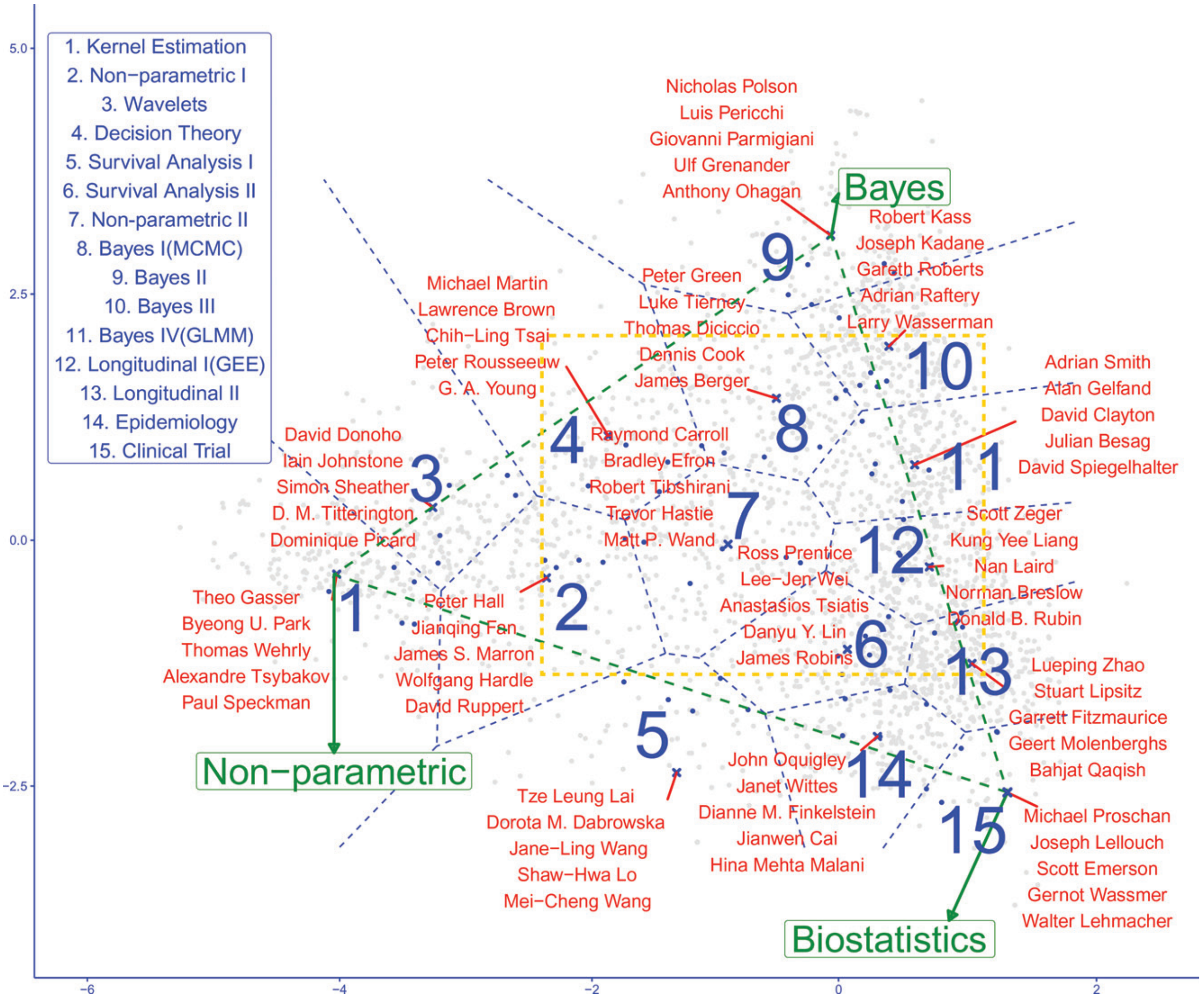

Figure 1. The research map. Each gray dot represents a 2-dimensional SCORE embedding vector $\hat{r}_i$, $1 \leq i \leq n$, and the 15 clusters and Voronoi diagram are obtained by applying the K-means algorithm to $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n$. The dashed green line represents the triangle, where the vertices represent the 3 primary areas. In each cluster, the cluster center is also presented (blue crosses), together with 5 authors with highest degrees (blue dots). The results are based on citations: it is possible that an author does not work in an area, but have many citations in that area.

2.2. Evolvement of Author Research Interests

The research map in Figure 1 was established using the co-citations during 1991–2000. We now study how individual authors' research interests evolve between 2001 and 2015, and propose *dynamic network embedding* as a new approach. For each author, the approach produces a trajectory on the research map to visualize his/her research interest evolvement.

We consider 21 time windows (see Table 1) and construct a citee network for each of them. As the numbers of articles published per year are steadily increasing, we use gradually smaller windows so the average node degrees of all 21 citee networks are roughly the same. We use the citee network for the first window (1991-2000) asthe reference network for our study below. This network is the same as the citee network that we use to study the statistics triangle and the research map in Figure 1. Recall that this network has 2831 nodes. We restrict each of the other 20 networks to the same set of nodes. We propose a *dynamic DCMM model* by extending the (static) DCMM model (2.1). Consider T citee networks for the same set of n nodes, and let $A_1, A_2, \ldots, A_T$ be the adjacency matrices. Let $P \in \mathbb{R}^{K,K}$ be the time-invariant community structure matrix, and let $\theta_i^{(t)} > 0$ and $\pi_i^{(t)} \in \mathbb{R}^K$ be the degree parameter and mixed membership vector of node i at time t, $1 \leq i \leq n$, $1 \leq t \leq T$. Write $\theta_t = \mathrm{diag}(\theta_{1t}, \ldots, \theta_{nt})$ and $\Pi_t = [\pi_{1t}, \ldots, \pi_{nt}]'$. Given $\{(\theta_t, \Pi_t\}_{t=1}^T$, we assume $A_1, A_2, \ldots, A_T$ are independently generated. Also, the upper triangle of A_t contains independent Bernoulli variables satisfying

Table 1. The 21 time windows we use to study the research trajectories.

Window	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Start	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11
End	00	01	01	02	03	04	04	05	06	07	07	08	09	10	10	11	12	13	13	14	15
Length	10	10	9	9	9	9	8	8	8	8	7	7	7	7	6	6	6	6	5	5	5

NOTE: For example, the first window is from 1991 to 2000, covering a 10-year time period.

$$\mathbb{P}\big(A_t(i,j) = 1\big) = \theta_i^{(t)}\theta_j^{(t)} \cdot (\pi_i^{(t)})' P(\pi_j^{(t)}), \qquad 1 \leq i < j \leq n. \tag{2.3}$$

Here, we assume $A_1, A_2, \ldots, A_T$ are independent given $\{(\theta_t, \Pi_t\}_{t=1}^T$, but this can be relaxed to allow for weak dependence. Also, to allow flexible temporal dependence in $\{(\theta_t, \Pi_t\}_{t=1}^T$, we do not impose any extra conditions on them.

How to estimate $\pi_i^{(t)}$ is known as the problem of dynamic mixed membership estimation. Existing works include Kim et al. (2018); Liu et al. (2018). However, these works focus on the dynamic MMSB model (a special dynamic DCMM) where it is required $\theta_i^{(t)} \equiv \alpha_t$ for all $1 \le i \le n$ at each time t. It is therefore unclear how to extend their ideas to our setting.

Alternatively, one may use naive mixed-SCORE (i.e., we apply mixed-SCORE to each network in the sequence separately). Unfortunately, the approach is also unsatisfactory. One challenge is that the estimates $\{\hat{\pi}_i^{(t)}\}_{1\le i\le n}$ for each time window t are up to an unknown permutation among the K communities. Since we have T different time windows, we have a large number of possible combinations of such permutations, and it is unclear how to pick the right one. The other challenge is that, each A_t is constructed for a relatively short time period, and can be very sparse. In such cases, spectral decomposition of A_t may be rather noisy, and the naive mixed-SCORE may perform unsatisfactorily.

We propose *dynamic network embedding* as a new approach to dynamic mixed membership estimation. Note that the network A_1 from the first window was used in Section 2.1 to build a "research map" for all the authors. This motivates us to treat A_1 as a reference network and project all the other networks onto this "research map." Let $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_K$ be the K largest eigenvalues (in magnitude) of A_1, and let $\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_K$ be the corresponding eigenvectors. For each $1 \le t \le T$ and each node $1 \le i \le n$, define a $(K-1)$-dimensional vector $\hat{r}_i^{(t)}$ by (e_i: the ith standard basis vector of $\mathbb{R}^n$)

$$\hat{r}_i^{(t)}(k) = [\hat{\lambda}_1(e_i' A_t \hat{\xi}_{k+1})]/[\hat{\lambda}_{k+1}(e_i' A_t \hat{\xi}_1)], \qquad 1 \le k \le K-1. \tag{2.4}$$

Now, for each time t, we obtain the low-dimensional embedding $\{\hat{r}_i^{(t)}\}_{1\le i\le n}$ of all n nodes, and for each node i, we obtain the embedded "trajectory" as $(\hat{r}_i^{(1)}, \hat{r}_i^{(2)}, \dots, \hat{r}_i^{(T)})$. For $t = 1$, $\hat{r}_i^{(1)}$ coincides with the SCORE embedding (2.2). It implies that the starting point of each embedded trajectory is always the position of this author in the "research map." For $t > 1$, the proposed embedding is different from the SCORE embedding (2.2) for A_t. Note that in (2.2), we use the eigenvectors of A_t to construct the embedding at t, while in Equation (2.4), we use the eigenvectors and eigenvalues of A_1 to construct the embeddings for all t.

We now explain how the approach overcomes the two challenges aforementioned. First, the new approach uses the same $(\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_K)$ to obtain the embeddings for all t, so that these networks are projected to the same low-dimensional space. Consequently, the projected points $\hat{r}_i^{(t)}$ are automatically aligned across time. Second, in spectral projection and its variants (e.g., SCORE), the data to project (rows of A_t) and the projection directions (eigenvectors of A_t) are *dependent* of each other. On the contrary, in Equation (2.4), the data to project, $A_t e_i$, and the projection direction, $\hat{\xi}_k$, are *independent* of each other, for any $t \ge 2$. Thus, the projected points are much less noisy. In the preliminary theoretical analysis, we find that $\hat{r}_i^{(t)}$ has a sharp large-deviation bound even when A_t is very sparse and when $\hat{\xi}_k$ is only a moderately good estimate of the population eigenvector of A_1.

We explain why the approach is reasonable. Define a population counterpart of Equation (2.4). In model (2.3), let $\Theta^{(t)} = \mathrm{diag}(\theta_1^{(t)}, \dots, \theta_n^{(t)})$, $\Pi^{(t)} = [\pi_1^{(t)}, \dots, \pi_n^{(t)}]'$, and $\Omega_t = \Theta^{(t)}\Pi^{(t)}P(\Pi^{(t)})'\Theta^{(t)}$, $1 \le t \le T$. Let $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_K)$ and $\Xi = [\xi_1, \xi_2, \dots, \xi_K]$, where λ_k is the kth largest (in magnitude) eigenvalue of Ω_1 and ξ_k is the corresponding eigenvector. For $1 \le t \le T$ and $1 \le i \le n$, define $r_i^{(t)} \in \mathbb{R}^{K-1}$ by

$$r_i^{(t)}(k) = [\lambda_1(e_i'\Omega_t\xi_{k+1})]/[\lambda_{k+1}(e_i'\Omega_t\xi_1)], \qquad 1 \le k \le K-1. \tag{2.5}$$

Theorem 2.1. Consider the dynamic DCMM model (2.3). For each $1 \le t \le T$, letting $M_t = P(\Pi^{(t)})'\Theta^{(t)}\Xi\Lambda^{-1} \in \mathbb{R}^{K,K}$, we suppose $\mathrm{rank}(M_t) = K$ and $\min_{1\le k\le K}\{M_t(1,k)\} > 0$. Let $v_k^{(t)} = \frac{1}{M_t(k,1)}[M_t(k,2), M_t(k,3), \cdots, M_t(k,K)]'$, $1 \le k \le K$, and let $\mathcal{S}_t \subset \mathbb{R}^{K-1}$ be the simplex with K vertices $v_1^{(t)}, \dots, v_K^{(t)}$. For all $1 \le t \le T$, first, each $r_i^{(t)}$ is contained in the simplex $\mathcal{S}_t$. If i is a pure node of community k ($\pi_i^{(t)} = e_k$), then $r_i^{(t)}$ is located on the vertex $v_k^{(t)}$. If i is not a pure node of any community, then $r_i^{(t)}$ is in the interior of $\mathcal{S}_t$ (including the edges and faces, but not any of the vertices). Second, each $r_i^{(t)}$ is a convex combination of $v_1^{(t)}, v_2^{(t)}, \dots, v_K^{(t)}$, denoted by $r_i^{(t)} = \sum_{k=1}^K w_i^{(t)}(k)v_k^{(t)}$. The coefficient vector $w_i^{(t)} \in \mathbb{R}^K$ satisfies that $w_i^{(t)} = (\pi_i^{(t)} \circ h_t)/\|(\pi_i^{(t)} \circ h_t)\|_1$, where $\circ$ is the Hadamard product and $h_t \in \mathbb{R}^K$ is a positive vector that does not depend on i.

Theorem 2.1 is proved in the supplement. By Theorem 2.1, in the noiseless case, the embedded data cloud $\{r_i^{(t)}\}_{1\le i\le n}$ at every t form a low-dimensional simplex, similar to that in Jin, Ke, and Luo (2017). We can then borrow the idea there and estimate $\pi_i^{(t)}$ from the embedded data cloud via a simplex vertex hunting algorithm. This explains the rationale of our procedure. To focus on real data analysis, we relegate more detailed analysis of the approach to a forthcoming article. We now apply the procedure to our dataset.

Research Trajectories for Individual Authors. Recall that we have constructed a 2831-node citee network for each of the 21 time windows in Table 1. Applying Equation (2.3), we get an embedding $\hat{r}_i^{(t)}$ for each author i at each time t. Viewing $\hat{r}_i^{(t)}$ as a point on the research map, we have 21 points for author i, each corresponding to a time window. Connecting these time-ordered points gives rise to the research trajectory of author i, which visualizes how the research interests of author i evolve over time. The starting point of his/her research trajectory is the same as his/her position in the research map in Figure 1.

In Figure 2, we present the research trajectories of a handful of representative authors in statistics. For better visualization, note that the whole region covered by Figure 2 is the zoom-in of the rectangular region bounded by dashed yellow lines in Figure 1. Since all of these authors happen to be in the reference citee network, the starting point of each author's trajectory is the same as his/her position on the research map in

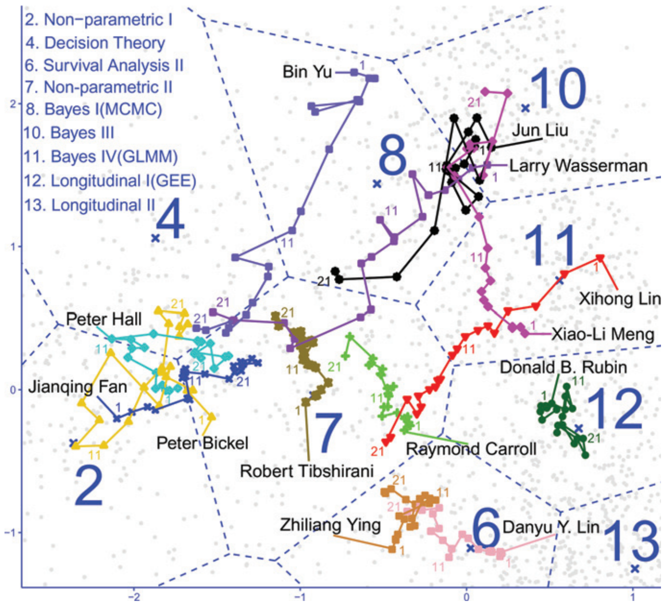

Figure 2. Research trajectories of representative authors (this is a zoomed-in view of the region in Figure 1 within the dashed yellow square, with the same Voronoi diagram). Each trajectory has 21 knots, corresponding to the 21 time windows in Table 1 (knots 1, 11, and 21 are marked with 1, 11, and 21, respectively). The starting point (marked with 1) is the same as the author's position in Figure 1. For interpretation, we selected some authors we are familiar with, but we can plot the trajectory for any author with a reasonably long publication history in our dataset. The results are based on citations: it may happen that an author (e.g., D. Rubin) does not work in an area, but have many citations in that area.

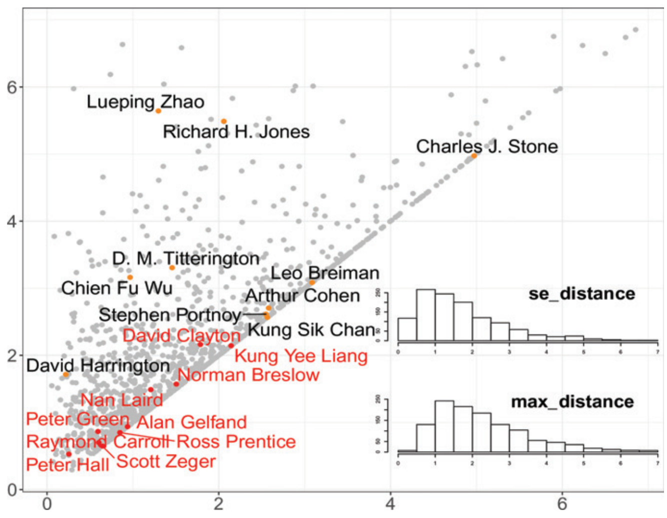

Figure 3. The two diversity metrics of 1,202 authors (x-axis: se_distance; y-axis: max_distance). The red dots represent the 10 highest-degree authors. The orange dots represent (among the top 200 highest-degree nodes) the 5 authors with the largest se_distance and the 5 authors with the largest differences between max_distance and se_distance.

Figure 1. We have the following observations: (a) A few authors (e.g., Xihong Lin, Jun Liu, Xiao-Li Meng, Larry Wassermann, and Bin Yu) exhibit a significant change of research interest from 2000 to 2015, suggesting that they persistently tried to broaden their research horizon and scope of interest. (b) The research trajectories of Peter Bickel, Raymond Carroll, Jianqing Fan, Peter Hall, and Robert Tibshirani stayed in the regions of *Decision Theory* and *Non-parametric I and II*, and the research trajectories of Danyu Lin, Donald Rubin, and Zhiliang Ying stayed in the regions of *Survival Analysis II* and *Longitudinal I (GEE)*. A possible reason is that the research areas of these authors in 1991–2000 continued to be "hot areas" for the time period 2000–2015. (c) The two subregions, *Non-parametric I and II*, are among the most "popular" research areas between 1991 and 2015. Research leaders (e.g., Peter Bickel, Jianqing Fan, Peter Hall, and Robert Tibshirani) who worked in these areas in 1990s continued to work in these research areas in 2000–2015. At the same time, research leaders who used to work on some seemingly distant areas or in distant regions (e.g., Xihong Lin, Jun Liu, Larry Wasserman, and Bin Yu) gradually migrate to the center of these two regions. These two sub-areas highly overlap with the research area of *high-dimensional data analysis*, which was one of the most rapidly growing areas in statistics between 2000 and 2015. The claim is confirmed by investigating more authors in these two subregions.

2.3. Diversity of Author Research Interests

The research trajectories in Section 2.2 suggest that research interests of some authors may vary more significantly than those of others. This motivates us to propose some metrics for research diversity of individual authors. Recall that the 21 knots for the trajectory of author i are $\hat{r}_i^{(1)}, \dots, \hat{r}_i^{(21)}$. We introduce two diversity metrics: $E_i = ||\hat{r}_i^{(21)} - \hat{r}_i^{(1)}||$ and $M_i = \max_{2 \le k \le 21} ||\hat{r}_i^{(t)} - \hat{r}_i^{(1)}||$,[1] where E_i is called *se_distance* (distance between the starting point and the ending point) and M_i is called *max_distance* (maximum distance between a point and the starting point). A large E_i suggests that the research areas for author i in 2011–2015 (the last time window) are significantly different from his/her research areas in 1991–2000, and a large M_i suggests that the research areas for author i in at least some of the time windows are significantly different from his/her research areas in 1991–2000.

Figure 3 presents the two metrics for a total of 1202 authors. The reference network has 2,831 nodes in total, but in the 21 citee networks (each for a different time window) only 1202 authors are always in the giant component, so we present only the E_i and M_i for these 1202 authors. In this figure, the 10 highest-degree nodes are marked with red dots, where their names are also presented in red. Also, among the 200 authors who have the largest degrees, the five authors who have the largest E_i values (Charles J. Stone, Leo Breiman, Arthur Cohen, Kun Sik Chan, Stephen Portnoy) are marked with orange dots, and the 5 authors who have the largest $(M_i - E_i)$ values (Luoping Zhao, Richard H. Jones, Chien Fu Wu, D.M. Titterington, David Harrington) are also marked with orange dots.

For author i, if both M_i and E_i are large, we call the changes of the research areas of author i *significant and persistent (SP)*, and for short, author i is an SP type. If M_i is large but E_i is relatively small, we call the changes of the research areas of author i *significant but not persistent (SnP)*, and for short, author i is an SnP type. For the 20 authors whose names are showed in

[1] Here, $\hat{r}_i^{(t)}$ are defined by (2.5) through the leading eigenvalues and eigenvectors $(\hat{\lambda}_k, \hat{\xi}_k)$ of A_{t_0} with $t_0 = 1$. Since we use the first one in the 21 networks as the reference, $t_0 = 1$ is the most natural choice. For robustness check, we have also studied the case of $t_0 \in \{2, 5, 10\}$; see Section C.4 of the supplement. The results are largely similar to those in this section.

the figure, Charles J. Stone has the largest E_i value and is seen to be an SP type, and Lueping Zhao has the largest M_i value and is seen to be an SnP type.

3. Learning Communities from Co-authorship Networks

The study of co-authorship patterns and community structures in an academic society is an interesting topic (Newman 2004). The co-author relationship in our dataset provides a valuable resource to study the community structure, which is the focus of this section. Compared to the co-citation relationship (focus of Section 2), the co-author relationship is quite different in nature: Citations are primarily driven by scientific relevance, but collaborations may be driven by many factors (e.g., geographical proximity, academic genealogy, and cultural ties). Therefore, the study below may shed new insight which we do not see in Section 2. We focus on the following problems: (a) hierarchical community detection (and especially interpretation of different communities), (b) evolvement of communities, and (c) diversity measure of individual authors. We discuss these in Sections 3.1-3.3 separately.

3.1. Estimation of the Hierarchical Community Structure

Compared to the citee networks, the effect of mixed-memberships in co-authorship networks is notably less significant; see Section D.5 of the supplement for detailed discussion. So instead of focusing on the mixed-memberships as in Section 2, we focus on the problem of *recursive community detection*: We think that the co-authorship network has many communities (each is a research sub-area in statistics), and the sub-areas may have a tree structure. The goal is to (possibly recursively) cluster the authors into these sub-areas.

A popular strategy to recursive community detection is as follows: First, we partition the network into K_0 groups, for a small integer $K_0 < K$, where K is the total number of communities. This gives rise to K_0 subnetworks restricted to each group. Next, for each subnetwork, we test whether it has only one community (null hypothesis) or multiple communities (alternative hypothesis). If the null hypothesis is rejected, then this subnetwork is further split. The algorithm stops when the null hypothesis is accepted in every subnetwork. The output is a hierarchical tree, with each leaf being an estimated community.

As the mixed-membership effect here is less significant than that in citee networks, it is reasonable to use the DCBM model (Karrer and Newman 2011). Compared with the DCMM model in (2.1), DCBM is a special case where we require all vectors π_i to be degenerate (i.e., one entry is 1, all other entries are 0), and so the nodes partition to non-overlapping communities $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$. Let $A \in \{0,1\}^{n\times n}$ be the symmetrical adjacency matrix of a co-authorship network, where $A(i,j) = 1$ if and only if authors i and j have co-authored articles in the range of interest. In DCBM, we assume

$$\mathbb{P}(A(i,j)=1) = \theta_i\theta_j P_{k\ell}, \qquad \text{if } i \in \mathcal{C}_k,\ j \in \mathcal{C}_\ell, \quad \text{for all } 1 \le k,\ell \le K. \tag{3.6}$$

where $(P, \theta_1, \theta_2, \ldots, \theta_n)$ are the same as those in Equation (2.1). In this subsection, we assume both the whole network and subnetworks satisfy the DCBM. A more careful modeling for the hierarchical structure is possible (e.g., Li et al. 2020). But since our primary focus here is to analyze a valuable new dataset, we leave this to the future.

There are many interesting works on recursive community detection (e.g., Li et al. 2020), but they focused on the stochastic block models, a special case of the DCBM model in Equation (3.6) that does not allow degree heterogeneity. It is unclear how to extend their methods to our settings. We propose a new algorithm for recursive community detection, consisting of a community detection module and a hypothesis testing module. Both modules are able to properly deal with severe degree heterogeneity. We now discuss them separately.

The community detection module clusters the nodes in a network into K_0 communities, for a given $K_0 \ge 2$. We use the following algorithm. For a tuning parameter $c_0 > 0$, let I_n be the identity matrix, let $\hat\mu_k$ be the kth largest eigenvalue (in magnitude) of $A + c_0 I_n$, and let $\hat\xi_k$ be the corresponding eigenvector, $1 \le k \le K_0$. Define a matrix $\hat R \in \mathbb{R}^{n,K_0-1}$ by $\hat R(i,k) = \hat\xi_{k+1}(i)/\hat\xi_1(i)$. For a threshold $t > 0$, we apply element-wise truncation on $\hat R$ and obtain a matrix $\hat R^* \in \mathbb{R}^{n,K_0-1}$ by $\hat R^*(i,k) = \mathrm{sgn}(\hat R(i,k)) \cdot \min\{|\hat R(i,k)|, t\}$, $1 \le i \le n, 1 \le k \le K_0 - 1$. We then apply the k-means algorithm to the rows of $\hat R^*$, assuming there are $\le K_0$ clusters. There are two tuning parameters (c_0, t). We set $c_0 = 1$ and $t = \log(n)$.

The approach extends SCORE (Jin 2015), where $c_0 = 0$. Recall that we call $\hat\xi_k$ the kth largest eigenvector of A if it corresponds to the kth largest (in magnitude) eigenvalue of A. SCORE uses the first K eigenvectors of A for clustering, but unfortunately, the estimated network is dis-assortative (a network is assortative if for any pair of communities, they have more edges within than between (Lu and Szymanski 2019). For co-authorship networks, such a result is hard to interpret. Note that for an assortative network, a negative eigenvalue is more likely to be spurious than a positive one. This motivates the above approach, where we replace A by $A + c_0 I_n$: the term $c_0 I$ penalizes the rankings of negative eigenvalues, so the set of first K eigenvectors of $A + c_0 I_n$ is different from those of A. How to choose c_0 is an interesting problem. We find all estimated networks for $c_0 \ge 1$ are assortative, so we choose c_0 as 1 for convenience. The asymptotic consistency of the proposed approach is similar to that of the original SCORE.

Given a cluster (subnetwork), the hypothesis testing module determines whether the cluster should be further split. To abuse the notation a little bit, let A be the adjacency matrix of the network formed by restricting nodes and edges to the set of nodes in the current cluster. As before, we assume A follows a DCBM model with K_0 communities and test the null hypothesis $K_0 = 1$. We use the Signed-Quadrilateral (SgnQ) test by Jin, Ke, and Luo (2021). Define $\hat\eta = \frac{1}{\sqrt{\mathbf{1}_n' A \mathbf{1}_n}} A\mathbf{1}_n \in \mathbb{R}^n$ and $A^* = A - \hat\eta\hat\eta' \in \mathbb{R}^{n,n}$. The SgnQ test statistic is

$$\psi_n = \frac{1}{\sqrt{2}}\left(\frac{\sum_{i_1,i_2,i_3,i_4 (\text{distinct})} A^*_{i_1 i_2} A^*_{i_2 i_3} A^*_{i_3 i_4} A^*_{i_4 i_1}}{2(||\hat\eta||^2 - 1)^2} - 1\right). \tag{3.7}$$

It was showed in Jin, Ke, and Luo (2021) that under mild conditions, $\psi_n \to N(0,1)$ in the null hypothesis. This asymptotic normality holds even when the network has severe degree heterogeneity. Then, we can compute the p-value conveniently

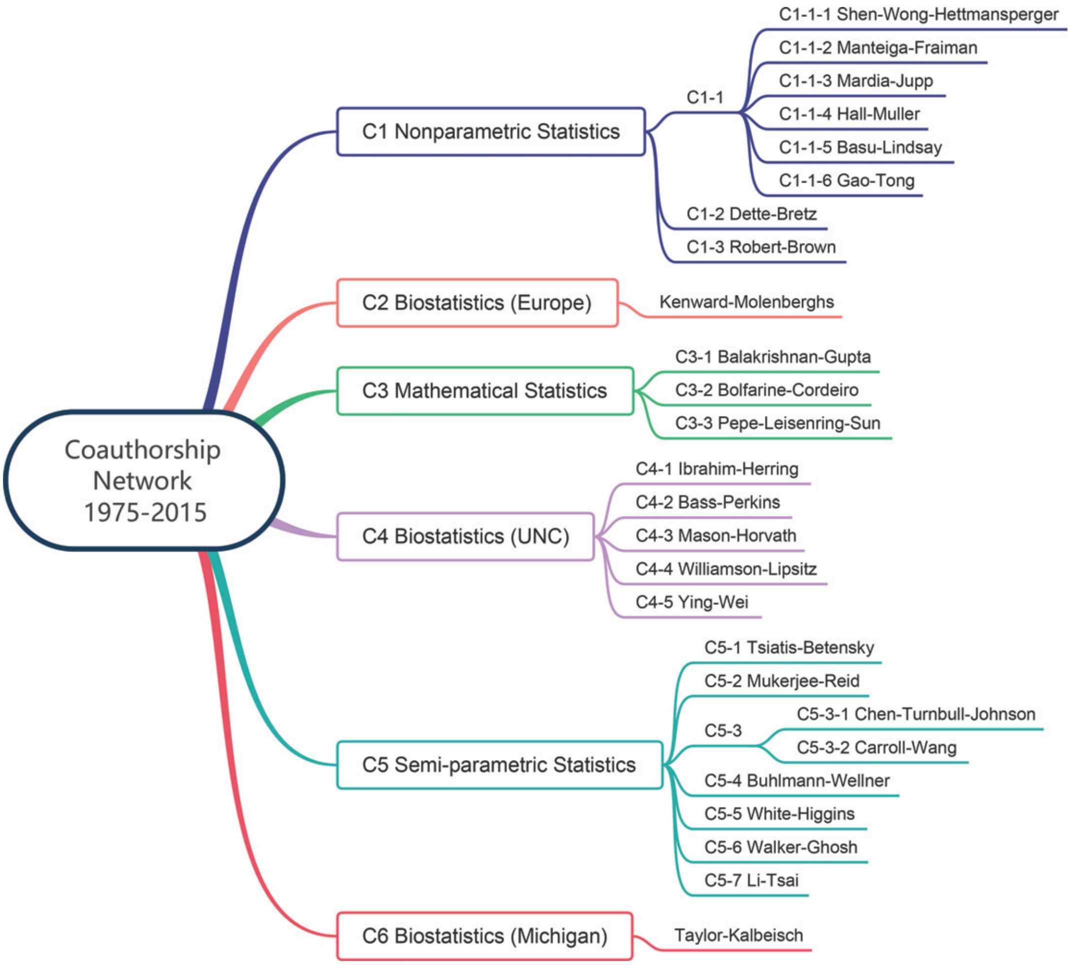

Figure 4. The community tree for co-authorship network. Each rightmost leaf community is labeled with the last names of 2 or 3 authors, selected by node betweenness and closeness. For each leaf, the representative nodes are shown in Table 3 (and Tables D.4–D.6 in the supplement).

and use it to set the stopping rule of the recursive algorithm (e.g., when p-value is ≥ 0.05, a cluster will not be split).

The Co-authorshp Network (36 Journals). We build a co-authorship network using all the data in 36 journals during 1975-2015 as follows: Each node is an author; there is an edge between two nodes if they have co-authored at least m_0 articles in the data range. As we wish to focus on (a) the subset of long-term active researchers, and (b) solid collaborations, choosing $m_0 = 1$ would be too low (see Ji and Jin 2016): we may include too many edges between active researchers and nonactives ones (e.g., a Ph.D. advisee who joined industry and stopped publishing in academic journals). We take $m_0 = 3$ and focus on the giant component, which has 4383 nodes. Taking $m_0 = 2$ may also be a reasonable choice, but the network is comparably denser and larger (10,741 nodes), and so requires more time and efforts to interpret the results (as we need to check each identified community one by one manually). Below, we present the result for $m_0 = 3$, and leave the results for $m_0 = 2$ to Section D.6 of the supplement, where we see the results of two cases are largely consistent.

We now apply our proposed algorithm. Note that the community detection module still requires an input of K_0. Similar to that in Section 2.1, we choose K_0 by combining the scree plot, goodness-of-fit, and evaluation of output communities (details are in Section D.4 of the supplement). Since we use the eigenvectors of $(A + I_n)$ for community detection, the scree plot contains the absolute eigenvalues of $(A + I_n)$ instead of those of A. The stopping rule of the recursive algorithm is set as follows: Either the SgnQ p-value is > 0.001 or the community has ≤ 250 nodes. The output is a hierarchical community tree in Figure 4.

The Hierarchical Community Tree. First, we investigate the 6 communities in the first layer. To help for interpretation, we apply topic modeling on article abstracts (see Section D of the supplement, especially Figure D.6). Combining the topic modeling results with a careful read of the large-degree nodes in each community, we propose to label these communities as in Table 2, where we also list some comments on each community.[2]

Next, we look at the other layers of the tree. The stopping rule of recursive partition is that either the SgnQ p-value is > 0.001 or the community size is ≤ 250, but there are a few exceptions in Figure 4: (a) C6 has 264 nodes, but its giant component has no more than 250 nodes. We thus keep C6 unchanged. (b) The second largest component of C4 contains 60 nodes which

[2] In Section 2.1, "Bayes" is one of the three vertices of the statistics triangle. Here, Bayes continues to play an important role, but it splits into multiple communities and so the word "Bayes" does not appear in the community labels.

Table 3. The leaf communities and the representative authors (ordered by degree within leaf community).

ID	Name	#Authors	p-value	Representative Authors
C1-1-1	Shen-Wong–Hettmansperger	144	0	Hannu Oja, Harvard Rue, Friedrich Gotze, Wei Pan, *Thomas P. Hettmansperger*, Jun Liu, *Xiaotong Shen*, Douglas A. Wolfe, Ishwar Basawa, Leonhard Held
C1-1-2	Manteiga-Fraiman	118	0.04	*Wenceslao Gonzalez-manteiga*, Graciela Boente, Juan Antonio Cuesta, Daniel Pena, Antonio Cuevas, *Ricardo Fraiman*, Richard Johnson, Michael Akritas
C1-1-3	Mardia-Jupp	102	0	Christian Genest, Ian Dryden, *Kanti V. Mardia*, Rainer Von Sachs, Wensheng Guo
C1-1-4	Hall-Müller	331	0.34	*Peter Hall*, James S. Marron, Jianqing Fan, Liang Peng, Byeong U. Park, *Hans-Georg Müller,* M. C. Jones, Laurens De Haan, Theo Gasser, Wolfgang Hardle
C1-1-5	Basu-Lindsay	68	0.012	*Bruce Lindsay*, Dankmar Bohning, Domingo Morales, Leandro Pardo, Dongwan Shin, *Ayanendranath Basu*, Maria Luisa Menendez, Konstantinos Zografos
C1-1-6	Gao-Tong	189	0	Marc Hallin, Wai Keung Li, David Nualart, David Nott, *Howell Tong*, Vo Anh
C1-2	Dette-Bretz	104	0.0049	*Holger Dette*, *Frank Bretz*, Axel Munk, Tony Hayter, Wei Liu, Henry Wynn
C1-3	Robert-Brown	249	0	William Strawderman, George Casella, Kerrie Mengersen, *Christian Robert*, *Lawrence Brown*, Tony Cai, Eric Moulines, Murad Taqqu, Anthony Pettitt
C2	Kenward-Molenberghs	202	0	*Geert Molenberghs*, Emmanuel Lesaffre, Marc Aerts, Christophe Croux, Helena Geys, *Mike Kenward*, Paddy Farrington, Byron J. T. Morgan, Ariel Alonso
C3-1	Balakrishnan-Gupta	311	0	*Narayanaswamy Balakrishnan*, *Arjun Gupta*, Manlai Tang, Yasunori Fujikoshi
C3-2	Bolfarine-Cordeiro	58	0.0003	*Gauss M. Cordeiro*, *Heleno Bolfarine*, Victor H. Lachos, Reinaldo B. Arellano-valle
C3-3	Pepe-Leisenring-Sun	86	0.0002	*Jianguo Sun*, Govind S. Mudholkar, *Margaret Pepe*, Liuquan Sun, *Wendy Leisenring*, Yudi Pawitan, Xinyuan Song, Xingwei Tong, Xian Zhou, Ziding Feng
C4-1	Ibrahim-Herring	142	0.003	*Joseph Ibrahim*, David Dunson, Hongtu Zhu, Andy Lee, Ming-hui Chen, Keith E. Muller, Kelvin K. W. Yau, Haitao Chu, Wing Fung
C4-2	Bass-Perkins	104	0	Yuval Peres, *Richard Bass*, Zhen Qing Chen, Frank Den Hollander, Davar Khoshnevisan, Donald Dawson, Klaus Fleischmann, *Edwin Perkins*, Jay Rosen
C4-3	Mason-Horvath	109	0	*Lajos Horvath*, Josef Steinebach, Miklos Csorgo, Luc Devroye, Piotr Kokoszka, Evarist Gine, Armelle Guillou, Marie Huskova, *David Mason*, Ricardas Zitikis
C4-4	Williamson-Lipsitz	120	0.0003	*Stuart Lipsitz*, Robert H. Lyles, Enrique Schisterman, Brian Reich, *John Williamson*, Peter Diggle, Nan Laird, Huiman X. Barnhart, Amita Manatunga
C4-5	Ying-Wei	60	0.008	*Lee-jen Wei*, *Zhiliang Ying*, Tze Leung Lai, Danyu Y. Lin, David Siegmund, Daniel Krewski, Lu Tian, Tianxi Cai, Louis Gordon, Sin-ho Jung
C5-1	Tsiatis-Betensky	185	0.009	Paul Yip, Xiaohua Zhou, *Rebecca Betensky*, John Crowley, Adrian Raftery, *Anastasios Tsiatis*, Ji Zhu, Richard Huggins, George Michailidis, John Oquigley
C5-2	Mukerjee-Reid	193	0	*Rahul Mukerjee*, Zhidong Bai, Christos Koukouvinos, Kashinath Chatterjee
C5-3-1	Chen-Turnbull–Johnson	201	0.31	*Wesley Johnson*, Brian Caffo, Dongchu Sun, Weichung J. Shih, *Bruce Turnbull*, Richard Lockhart, Richard Simon, *Gemai Chen*, Mathias Drton, Galin L. Jones
C5-3-2	Carroll-Wang	231	0	*Raymond Carroll*, Mitchell Gail, Xihong Lin, Laurence Freedman, Hua Liang, Jianhua Huang, David Ruppert, Suojin Wang, Kevin W. Dodd, Dean Follmann
C5-4	Buhlmann-Wellner	166	0.0013	Mark Van Der Laan, Aad Van Der Vaart, *Peter Buhlmann*, Subhashis Ghosal, Ram Tiwari, Larry Wasserman, Bin Yu, Joseph Kadane, Thomas Kneib
C5-5	Whilte-Higgins	71	0.016	Martin Schumacher, Simon Thompson, John Whitehead, Nicky Best, *Ian White*, *Julian P. T. Higgins*, Jon Wakefield, Dan Jackson, Sylvia Richardson
C5-6	Walker-Ghosh	197	0	*Stephen Walker*, *Malay Ghosh*, Alan Gelfand, Pranab Kumar Sen, Robert Kohn
C5-7	Li-Tsai	159	0.034	Lixing Zhu, Robert Tibshirani, Dennis Cook, *Chih-ling Tsai*, *Runze Li*, Jun Shao, Trevor Hastie, Shein-chung Chow, Riquan Zhang, Andreas Buja
C6	Taylor-Kalbfleisch	264	0	*Jeremy Taylor*, Xin Tu, Daniel Commenges, Donald R. Hoover, Thomas Ten Have

NOTE: To label a community, two or three authors are selected by node betweenness and closeness; if any of them is also a representative author, we present his/her full name in italics. More details are in Tables D.4–D.6 of the supplement.

Table 2. The communities C1, C2, . . ., C6 and a brief description for each community.

Community	Description
C1. Nonparametric Statistics	Decision theory, nonparametric methods, high-dimensional statistics
C2. Biostatistics (Europe)	Biostatisticians from Europe, and their close collaborators
C3. Mathematical Statistics	Testing, computational statistics, probability, and other classical topics in probability and statistical theory
C4. Biostatistics (UNC)	Survival analysis, longitudinal data analysis, Biostatisticians from University of North Carolina (UNC) and collaborators
C5. Semi-parametric Statistics	Semiparametric methods, machine learning, variable selection, biostatistics
C6. Biostatistics (UM)	Biostatisticians from University of Michigan (UM) and close collaborators

form a tight-knit group. While these nodes are not in the giant component, we keep them as a separate community C4-5. (c) C3-1 has 311 nodes and its p-value $\approx$ 0. However, after we further split it into 2 sub-communities by SCORE, one sub-community contains only 8 nodes, and the other has a p-value 0.1. We thus keep C3-1 unchanged.

For each leaf community (i.e., the community corresponding to a leaf in the tree), we provide a manual label using two commonly used centrality measures, the *betweenness* (Freeman 1977) and the *closeness* (Bavelas 1950). For a node in a community, its *betweenness* is defined as the number of pairs of nodes in the same community that are connected through this node via the shortest path (therefore, a node with a large betweenness plays an important role in bridging other nodes), and the *closeness* of the node is defined as the reciprocal of the sum of distances from all other nodes in the same community to this node. Given a leaf community, we use the last names of the two nodes with largest betweenness and the one node with largest closeness to label the community (of course, if the latter happens to be one of the former, we will not use the same name twice). As a result, each leaf community is labeled with the last names of either two or three authors (not necessarily in alphabetical order). Table 3 presents a few representative nodes for each leaf community. More information of each leaf community is in Tables D4–D.6 of the supplement.

The results confirm that there are multiple factors for the formation of a tightly knit cluster of co-authorship: similar research interest, academic genealogy, friendship, colleague relationship, geological proximity, or close cultural ties. Below are some examples.

Example 1. Similar research interest. A number of leaf communities can be interpreted as groups of researchers sharing similar research interest. For example: *C1-3: Robert-Brown (Decision theory)*, *C1-1-4: Hall-Müller (Nonparametric statistics)*, *C4-2: Bass-Perkins (Probability)*, *C4-5: Ying-Wei (Sequential data analysis)*, *C5-4: Bühlmann-Wellner (Theoretical machine learning)*, *C5-3-2: Carroll-Wang (Semi-parametric statistics)*, *C5-7: Li-Tsai (Variable selection and dimension reduction)*.

Example 2. Geological and cultural factors. It is more likely for people who are geologically or culturally close to each other (e.g., colleagues, researchers in neighboring institutes or in the same region or country) to form tightly knit clusters. For example: *C2: Kenward-Molenberghs (Biostatisticians in Belgium)*, *C4-1: Ibrahim-Herring (Statisticians in the North Carolina research triangle)*, and *C5-5: White-Higgins (Biostatisticians in the U.K.)*. Additionally, C4-1 also contains a group of statisticians in Hong Kong, China. This group is brought together with the North Carolina group largely due to the collaboration between Joseph Ibrahim (faculty at University of North Carolina (UNC)) and Qi-Man Shao (faculty at the Chinese University of Hong Kong). Our analysis also suggests that the geological and cultural effect plays a more important role in forming clusters among biostatisticians than (say) among theoretical statisticians, and a possible reason is that collaborated research in biostatistics depends more on manpower and data sharing. For example, to comply with the data-sharing policies, it is simply easier for one to collaborate with someone in the same institute/country than with others.

Example 3. Academic genealogy. The academic advisor-advisee relationship is also a common source of collaboration. For example, the leaf community *C1-1-1 Shen-Wong-Hettmansperger* has a component of 29 nodes, which is largely formed by students of three authors, Wing H Wong, Jun Liu, and Xiaotong Shen; Liu and Shen are former students of Wong. We also note that this leaf community has sub-communities. For example, the network has a component of 24 nodes containing Thomas P. Hettmansperger. We did not further split C1-1-1 simply because its size falls below 250.

Recall that we name the first-layer communities, C1, C2, …, C6, using the results of topic learning (see Figure D.6 and Table 2). In most cases, the interpretations of umbrellaed leaf communities match with the name of the first-layer community. One exception is "C3-3 Pepe-Leisenring-Sun." It is under "C3 Mathematical Statistics" but consists of a group of biostatisticians. After some investigation, we find that this group is brought together with other groups in C3 largely by the author Xingqiu Zhao. She collaborated with both Narayanaswamy Balakrishnan, a hub node of C3, and Jianguo Sun, a hub node of C3-3.

The community tree is constructed by SCORE. To compare with other clustering methods, we apply Newman-Girvan's modularity approach (Newman's spectral approximation) (Newman 2006) to the same co-authorship network, and obtain six communities. We then check the numbers of nodes in the intersection between each of these communities and each of 26 leaves in our tree. The results are in Table D.7 of the supplement. We find that for most of the 26 leaf communities identified by SCORE, the majority of nodes in the community are contained in one of the six communities identified by Newman's approach. Therefore, at least to some extent, two clustering results are consistent with each other.

3.2. Evolvement of Co-authorship Clusters

Our dataset spans a relatively long time period (1975–2015), and it is interesting to study and visualize how the network communities evolve over time. The Sankey plot is a popular visualization tool for dynamic networks. However, to have a nice plot with interpretable results, we face many challenges: (a) the co-authorship network constructed using all data has too many communities (so it is hard to interpret all of them, and the resultant Sankey plot will also be too crowded); (b) it is unclear how to determine the number of communities; (c) it is also unclear how to interpret each community.

For (a), we decided to focus on the co-authorship network constructed with only articles from four representative journals, *AoS*, *Bka*, *JASA*, and *JRSSB* (the full journal names are in Table B.1). Compared to the co-authorship network constructed with the articles in all 36 journals, research interests of the authors in the current network are more homogeneous. As a result, the network has many fewer communities and is comparably easier to analyze. We have also spent a lot of efforts in dealing with challenges (b)–(c); see details below.

The Dynamic Co-authorship Networks (4 Journals). We consider three time windows in our study: (i) 1975–1997, (ii) 1995–2007, and (iii) 2005–2015. As in many works on dynamic network analysis (Kim et al. 2018), we let the adjacent time windows be slightly overlapping, so the results on community detection will be much more stable. For each time period, we construct a co-authorship network where each author who has ever published in any of the four aforementioned journals during this time period is a node, and two nodes have an edge if and only if they have co-authored one or more articles. For each network, there are relatively few nodes outside the giant component, so we remove them and consider the giant component only. Denote the resultant co-authorship networks for the three time periods by G_1, G_2, and G_3, respectively.

The Sankey Diagram. By careful investigation, we found that the three networks have 3, 4, and 3 communities, respectively. Once these numbers are determined, we first perform a community detection for each network by applying the modified SCORE described in Section 3.1, and then use the estimated community labels to generate a Sankey diagram; see Figure 5. Since the sets of nodes of three networks are different, we focus on the set $V = (G_1 \cap G_2) \cup (G_2 \cap G_3)$, which has 1687 nodes, for the Sankey diagram.

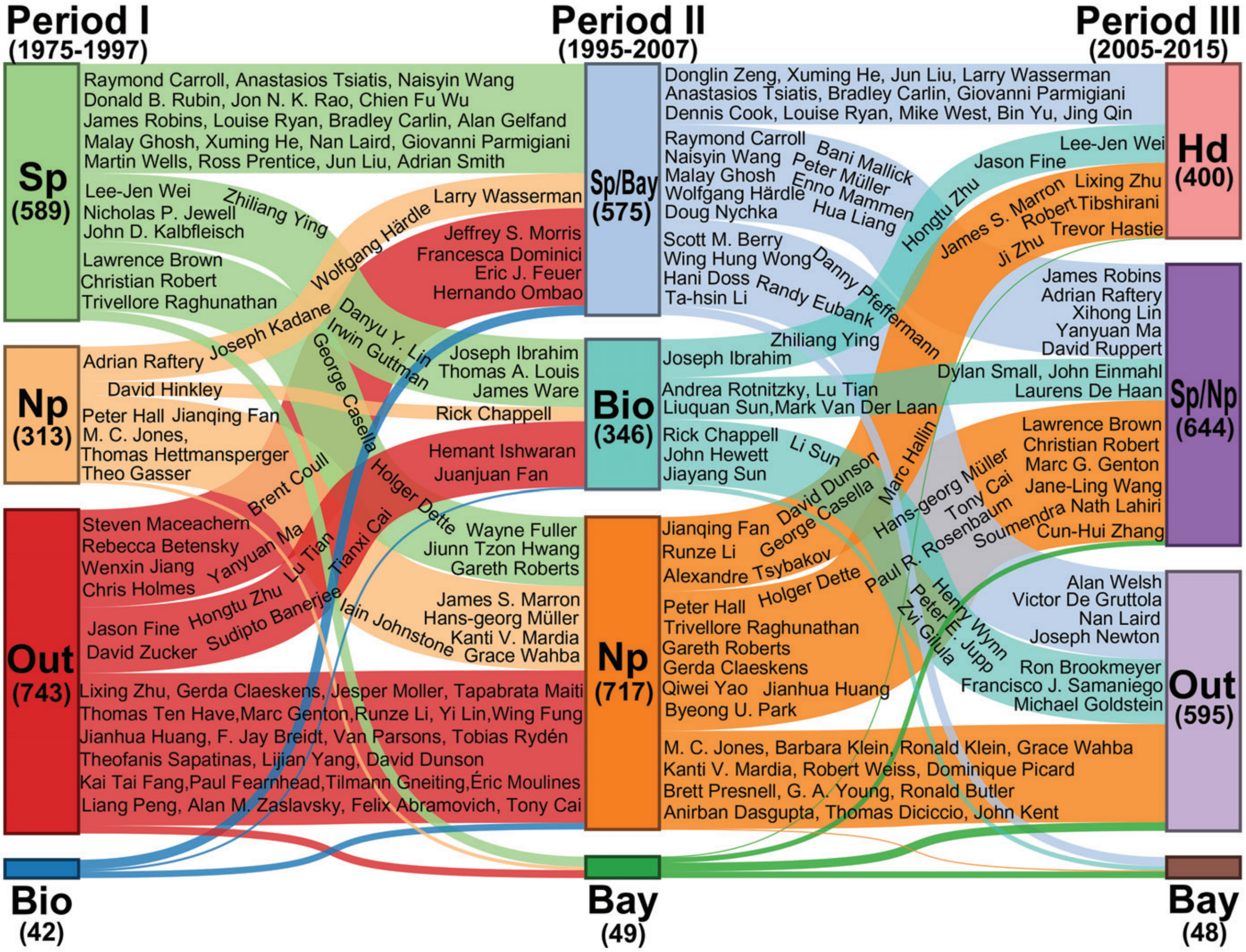

Figure 5. Evolution of communities in the dynamic co-authorship network (based on articles in 4 journals). The representative authors are selected by average degree in two adjacent networks.

We explain some notations in Figure 5. Consider the network for the time Period 1 first. By similar analysis as before, we propose to label the three communities obtained from applying modified SCORE to the network by *semiparametric statistics (SP)*, *nonparametric statistics (NP)*, and *Bayes (Bay)*. We do not have a separate community for biostatisticians, but a significant number of biostatisticians (e.g., Jason Fine, Lu Tian, Hongtu Zhu) are outside V, and another significant number of them (e.g., Lee-jen Wei, Zhiliang Ying, Joseph Ibrahim, Nicholas P. Jewell) are in SP. Let SP1, NP1, and Bio1 be the intersection of V and each community, respectively. We have $V = SP1 \cup NP1 \cup Bio1 \cup O_1$, where $O_1 = V \setminus G_1$.

The discussion of the third network is similar, except that the estimated communities are interpreted as *high-dimensional data analysis (HD)*, *nonparametric and semiparametric (NP/SP)*, and *Bayes (Bay)*. Similarly, $V = HD \cup (NP/SP) \cup Bay3 \cup O_3$, where $O_3 = V \setminus G_3$.

Last, consider the second network. The four communities obtained by applying SCORE can be similarly interpreted as *seimparametric statistics and Bayes (SP/Bay)*, *nonparametric (NP)*, *Bayes (Bay)*, and *biostatistics (Bio)*. We have $V = (SP/Bay) \cup NP2 \cup Bay2 \cup Bio2$, where NP2 is the intersection of NP with G_2; similar for Bay2 and Bio2. Note here that V is a subset of G_2 (but not a subset of G_1 or G_3), and so $O_2 = V \setminus G_2$ is an empty set. See Figure 5 for details.

The Sankey diagram suggests several noteworthy observations. First, in time Period 1, our algorithm suggests that there is no "Bio" community, although many biostatisticians (e.g., Jason Fine, Hongtou Zhu, Lu Tian) are outside the set V (recall that $V = (G_1 \cap G_2) \cup (G_2 \cap G_3)$). In time Period 2, our algorithm suggests that there is a "Bio" community, where a significant fraction of the members come from the outside of V, and another significant fraction (e.g., Lee-jen Wei, Zhiliang Ying, Joseph Ibrahim, Nicholas P. Jewell) come from SP in time Period 1. Second, from time Period 2 to time Period 3, a noticeable point is the rise of the community of *high dimensional data analysis (HD)*, which attracts authors from nonparametric statistics (e.g., Jianqing Fan, David Dunson, James S. Marron, Lixing Zhu), semiparametric statistics and Bayes (e.g., Dongling Zeng, Xuming He, Jun Liu, Larry Wassermann), and biostatistics (e.g., Joseph Ibrahim, Zhiliang Ying, Hongtu Zhu, Jason Fine). Last, in all three time periods, there are significant migrations between semiparametric statistics and nonparametric statistics.

Also, as examples, we note that (a) Raymond Carroll, Malay Ghosh, Bruce Lindsay, Ross Prentice, Jon N. K. Rao, James Robins, and Naisyin Wang remain in "SP" all the time; (b)

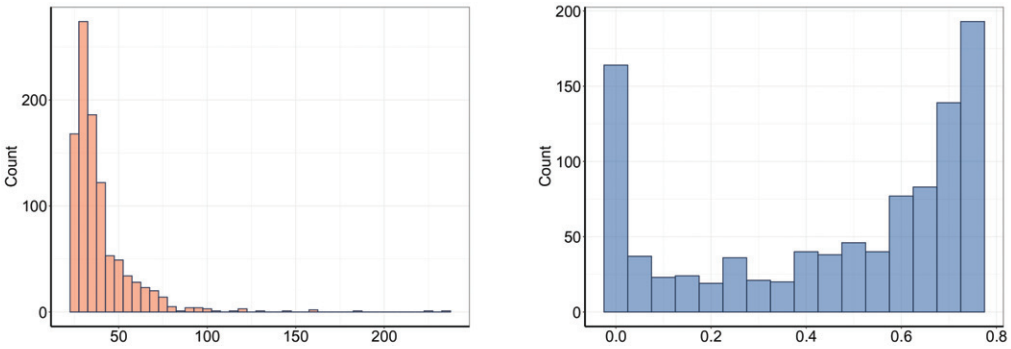

Figure 6. Left: histogram for the numbers of co-authors of 1000 authors who have the largest number of co-authors in our dataset. Right: histogram for the SgnQ *p*-values for the 1000 personalized co-authorship networks. A smaller *p*-value suggests that the personalized network is more likely to have multiple tight-knit groups (so the author is more diverse in terms of co-authorship).

Peter Hall, Hans-Georg Müller remain in "NP" all the time; (c) Jianqing Fan, Trevor Hastie, James S. Marron, Robert Tibshirani stay in "NP" in time Period 1, 2, and migrate to "HD" in Period 3; (d) Bradley Carlin, Xuming He, Jun Liu, Rahul Mukerjee, Lous Ryan, Anastasios Tsiatis, and Martin Wells, stay in "SP" in time Period 1, 2 and migrate to "HD" in Period 3. (e) Danyu Y. Lin, Lee-jen Wei, Zhiliang Ying start from "SP" in time Period 1, migrate to "Bio" in Period 2, and migrate to "HD" in Period 3.

3.3. A New Approach to Measuring an Author's Research Diversity

In Section 2.3, we have proposed two diversity metrics for the research interests of individual authors, using the trajectory. In this section, we propose a new approach to measuring research diversity by using the personalized networks and a recent tool in network global testing. The approach is quite different from that in Section 2.3 (and also those in the literature), and provides new insight on the research diversity of statisticians.

Fixing a node in a symmetrical network, the *personalized network* (also called the ego network) is the subnetwork consisting of the node itself and all of its adjacent nodes. We construct a co-authorship network similar to that in Section 3.1 but with $m_0 = 1$: Every author who ever published an article in any of the 36 journals between 1975 and 2015 is a node, and two nodes have an edge if and only if they co-authored one or more articles. Once this large network is constructed, for every author, we can obtain a personalized co-authorship network accordingly.

We model each personalized co-authorship network with a DCBM model (2.1) with K communities. We consider the global testing problem (Yuan, Feng, and Shang 2018) where we test H_0: $K = 1$ versus H_1: $K > 1$. Viewing each community as a tight-knit group, this is testing whether the given personalized co-authorship network has only one or multiple tight-knit groups. We approach the testing problem by the SgnQ test (Jin, Ke, and Luo 2021) which was already described in Section 3.1. Let Q_i be the test score ψ_n in Equation (3.7) for the personalized co-authorship network of author i. According to Jin, Ke, and Luo (2021), when the null hypothesis is true, $Q_i \to N(0, 1)$ as the size of the personalized network grows to ∞. We thus calculate the p-value by $p_i = \mathbb{P}(N(0, 1) \geq Q_i)$ and assign p_i to author i. We propose to use p_i to measure the co-authorship diversity of author i: a large p-value suggests that his/her co-authors form a tightly knit group, and a small p-value suggests that his/her co-authors are from two or more groups and so he/she is more diverse in co-authorship.

Figure 6 presents the results for the personalized co-authorship networks of 1000 authors who have the largest numbers of co-authors in our dataset. The left panel presents the histogram for the numbers of co-authors of these 1000 authors, and the right panel presents the histogram for the p-values of their personalized co-authorship networks. The p-values spread between 0 and 0.8, and 190 of them are smaller than 5%. Therefore, for about 80% of these 1000 authors, their co-authors form a tight-knit group.

Moreover, Table 4 presents the p-values from the SgnQ test for the personalized networks of 15 authors who have the largest numbers of co-authors. Take the first two authors, for example. They both have a large number of co-authors, but the p-value for Raymond Carroll is 0.02 while the p-value for Peter Hall is 0.23. This suggests that Hall's co-authors are likely to form a tight-knit group, while Carroll's co-authors may come from multiple groups. To identify such groups, we perform a community detection on Carroll's personalized co-authorship network (excluding Carroll[3]) by SCORE (see Section 3.1 and Jin 2015) and find that the research areas of a group of co-authors (e.g., Laurence Freedman, Victor Kipnis, Douglas Midthune, etc.—they work or used to work for National Cancer Institute (NCI)) are quite different from those of the other co-authors of Carroll. This explains why Carroll's network has a relatively small p-value. See Figure 7 (left panel) for the personalized co-authorship network of Carroll, where the p-value of any node presented there is the p-value for his/her own personalized co-authorship network.

Extension to Measuring the Diversity of Citers and Citees. We extend the study to personalized citer/citee networks. In a citer network, two authors have an edge if they have both cited some other authors. In a citee network, two authors have an edge if they have been both cited by some other nodes. Similarly as above, we construct a personalized citer network and a personalized citee network for each author i. We apply the SgnQ test and denote the two test scores by T_i^{citer} and T_i^{citee}, respectively.

[3] We exclude Carroll here for the edges between him and all other nodes contain little information of the community structure, but have a significant effect in the spectral domain, which makes the estimated communities by SCORE (a spectral method) less clear.

Table 4. Numbers of co-authors and p-values of the personalized co-authorship networks for the 15 authors who have the largest numbers of co-authors in our dataset (zero p-value means $< 10^{-6}$).

Name	#Coau	p-value	Name	#Coau	p-value	Name	#Coau	p-value
Raymond Carroll	234	0.02	Geert Molenberghs	146	0	Pranab Kumar Sen	112	0.71
Peter Hall	222	0.23	James S. Marron	130	0.007	Lixing Zhu	103	0.65
Naray. Balakrishnan	186	0.70	Malay Ghosh	119	0.51	David Dunson	101	0.64
Jeremy Taylor	159	0	Emmanuel Lesaffre	119	0	Jianqing Fan	101	0.38
Joseph Ibrahim	158	0.01	Xiaohua Zhou	119	0.31	Stuart Lipsitz	98	0.11

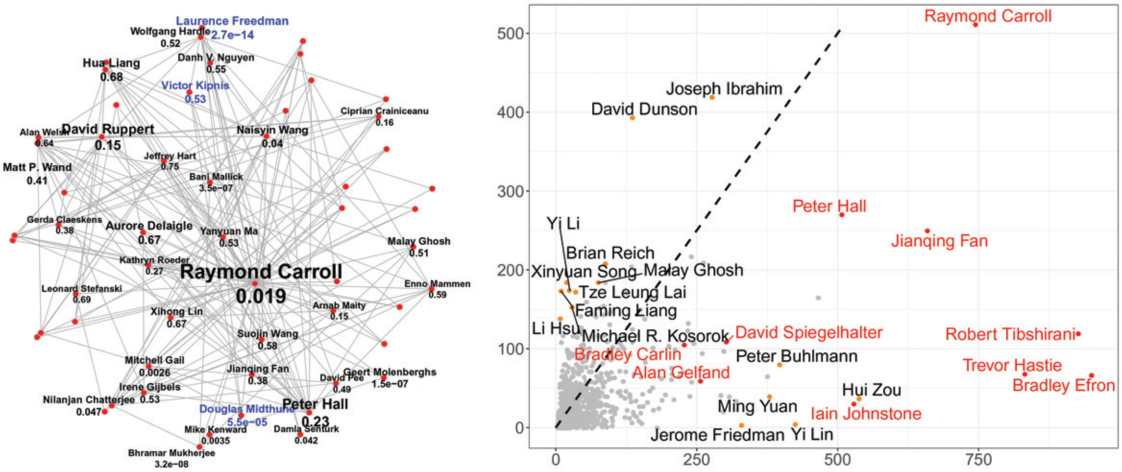

Figure 7. Left: The personalized co-authorship network of Raymond Carroll (the most collaborative author; see Table 4). Only nodes with 40 or more co-authors are shown. Different colors of names indicate two communities identified by SCORE. Similar plot can be generated for any author whose personalized network is reasonably large ($\geq$ 50 nodes, say). Right: The pair SgnQ test statistics (T_i^{citer}, T_i^{citee}) on personalized citer and citee networks of 1000 authors with highest degrees. The red dots correspond to high-degree authors. The yellow dots correspond to authors with either the largest or the smallest values of ($T_i^{\text{citer}} - T_i^{\text{citee}}$)

Figure 7 shows the two test scores for 1000 authors with the largest numbers of co-authors. First, for most authors (705 out of 1000) the personalized citer network is more diverse than the personalized citee network. This is because each author typically focuses on only a few research areas, but his/her work may be cited by researchers from various areas. Second, there is a group of authors whose T_i^{citee} is much smaller than T_i^{citer}, most of whom are theoretical statisticians (e.g., Bradley Efron, Iain Johnstone). This is probably because theoretical articles mainly cite theoretical articles but can be cited by many methodology and applied articles. Third, there is a group of authors in biostatistics (e.g., Michael Kosorok, Tze Leung Lai), whose test score for the citee network is much larger than that for the citer network. This is probably because biostatistics articles cite a variety of methodology articles; another reason is that many citations to articles in biostatistics are from other disciplines not covered by our dataset. Last, for Raymond Carroll, Jianqing Fan, Peter Hall, and Joseph Ibrahim, both test scores are relatively large, suggesting that they are diverse both in citer and citee.

We have proposed five metrics for measuring the research interests and diversity: two (denoted by A1 and A2) in Section 2.3 where we measure the diversity using the research trajectory computed from the co-citation networks, and three (denoted by B1-B3) in this section, for the co-authorship, citer, and citee networks, respectively. These metrics measure diversity from different angles using different types of networks. Also, the networks are based on data in different ranges. For these reasons, our results on diversity may have some inconsistencies, and we must interpret them with caution. For example, it is not rare that an article on one research topic may impact several other research topics, so an author who is not diverse in co-authorship can be significantly diverse in research impacts. For example, most articles by Donald Rubin are in Bayesian statistics and causal inferences, but he has impacts over many other areas (e.g., GEE); see Figures 1 and 2. Xihong Lin is regarded as highly diverse in research impact, but not regarded as diverse in co-authorship (based on results in our data range); see Figures 2 and 7. Also, while Approaches A1–A2 and B3 are both for citee networks, A1–A2 are for a dynamic DCMM setting and measures how the membership vector, π_{it}, evolve over time, and B3 considers a (static) DCMM setting and measures whether the personalized network has only one or multiple communities.

For reasons of space, we focus on the network approach in this article where we model the co-author relationships by networks. As an extension, we may model the co-author relationships by the more sophisticated hypergraph model (e.g., Ke, Shi, and Xia 2019; Jin, Ke, and Liang 2021; Yuan et al. 2021). In comparison, the literature on the hypergraph approach is much less developed than that of the network approach, so we leave the study on the hypergraph approach to the future.

4. Conclusion

We have several contributions. First, we produce a large-scale high-quality dataset. Second, we set an example for how to conduct a data science project that is highly demanding (in data resource, tools, computing, and time and efforts). We showcase this by creating a research template where we (a) collect and clean a valuable large-scale dataset, (b) identify a list of interesting problems in social science and science, (c) attack these problems by developing new tools and by adapting exiting tools, (d) deal with a long array of challenges in real data analysis so as to get meaningful results, and (e) use multiple resources to interpret the results, from perspectives in science and social science. We have also made significant contributions in methods and theory by developing an array of ready-to-use tools (for analysis and for visualization).

Our study has (potential) impact in social science, science, and real life. For example, suppose an administrator (in an university or a funding agency) wants to learn the research profile of a researcher. Our study provides a long list of tools to characterize and visualize the research profile of the researcher. Such information can be very useful for decision making. Our study also provides a useful guide for researchers (especially junior researchers) in selecting research topics, looking for references, and building social networks.

In social science, an important problem is to study the evolvement of a scientific community (Rosvall and Bergstrom 2010). We attack the problem by providing several tools (e.g., research map, research trajectory, Sankey plot) for characterizing and visualizing the evolvement of the statistical community. Another important problem is to check whether the development of a research field is balanced (e.g., if some areas are over-studied or under-studied) and whether there are unknown biases (e.g., whether scientists have biases when publishing articles related to COVID-19) (Foster, Rzhetsky, and Evans 2015). Our study can tell which areas have far more researchers, articles, or citations than others, and so helps check the balance of the field. Our study is also potentially useful for checking unknown biases.

In science, an important problem is how to identify patterns and so to predict new discoveries ahead of time. For example, in material science, one can use the abstracts of published articles to recommend materials for functional applications several years ahead of time (Tshitoyan et al. 2019). We can do similar things with our dataset to predict emerging new areas and significant advancements. For example, in Ji et al. (2021), we combine our citation data with the article abstracts (treated as text data) to rank different research topics and identify the most active research topics. We find that in the past decade, machine learning has been rising to one of the active research topics in statistics.

Though our dataset is high quality, we still need some necessary data preprocessing, and focus on networks with sizes much smaller than $47K$. The bottleneck for studying much larger networks is the time and efforts required to manually label each research area and to interpret the results in each case. For better use of such a valuable dataset, our hope is that, the dataset (which will be publicly available soon) will motivate many lines of researches, so over the years, researchers may continue to use different parts of the dataset for new projects and new discoveries.

For future work, note that our dataset provides at least two data resources: co-author relationships and citation relationships. It is noteworthy that most existing works in bibliometrics have been focused on one data source and one specific problem. Our results suggest the following: (a) The two data resources provide different information for the same group of researchers, and analysis of different data resources may have different results. The data resources and the results complement with each other. (b) Analysis focusing on only one aspect may have limited insight. Combining analysis of different aspects helps paint a more complete picture. (c) Therefore, it is highly preferable to combine the data resources for our study, with a multi-dimensional framework and multi-way analysis. In our real data analysis, we have combined the two data resources. For example, in Section 3.3, we use different metrics to measure the diversity of an author, where some metrics are based on the co-citation data and others are based on the co-authorship data. How to combine different data resources more efficiently is an interesting problem. We leave this to the future work.

Supplemental Material

Supplemental material contains a disclaimer, details of the dataset, supplemental data analysis results, and proof of Theorem 2.1.

Acknowledgments

The authors thank to the associate editor and referees for very helpful comments. They thank to Yoav Benjamini, Raymond Carroll, David Donoho, Yi Li, Jun S. Liu, Xiao-Li Meng, Neil Shephard, Bill Shi and Peter Song for many helpful comments and encouragements.

Funding

The research of J. Jin is supported in part by NSF Grant DMS-2015469, and the research of Z. T. Ke is supported in part by NSF Grant DMS-1943902.

References

Amini, A. A., Chen, A., Bickel, P. J., Levina, E. (2013), "Pseudo-Likelihood Methods for Community Detection in Large Sparse Networks," *Annals of Statistics*, 41, 2097–2122. [472]

Bavelas, A. (1950), "Communication Patterns in Task-Oriented Groups," *Journal of the Acoustical Society of America*, 22, 725–730. [479]

Donoho, D. L. (2017), "50 Years of Data Science," *Journal of Computational and Graphical Statistics*, 26, 745–766. [469]

Efron, B. (1998), "Fisher in the 21st Century," *Statistical Science*, 13, 95–114. [469,470,473]

Foster, J. G., Rzhetsky, A., and Evans, J. A. (2015), "Tradition and Innovation in Scientists' Research Strategies," *American Sociological Review*, 80, 875–908. [484]

Freeman, L. C. (1977), "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, 40, 35–41. [479]

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010), "A Survey of Statistical Network Models," *Foundations and Trends in Machine Learning*, 2, 129–233. [472]

Hall, P. G. (2011), " 'Ranking Our Excellence' or 'Assessing Our Quality' or Whatever," *Institute of Mathematical Statistics Bulletin*, 40, 12–14. [469]

Ji, P., Jin, J., Ke, Z. T., and Li, W. (2021), *Journal Ranking, Topic Modeling, and Citation Prediction for Statistical Publications*. Manuscript. [484]

Jin, J. (2015), "Fast Community Detection by SCORE," *Annals of Statistics*, 43, 57–89. [477,482]

Jin, J., Ke, Z. T., and Liang, J. (2021), "Sharp Impossibility Results for Hypergraph Testing," manuscript. [483]

Jin, J., Ke, Z. T., and Luo, S. (2017), "Estimating Network Memberships by Simplex Vertex Hunting," arXiv:1708.07852. [470,472,473,475]

——— (2021), "Optimal Adaptivity of Signed-Polygon Statistics for Network Testing," *Annals of Statistics* (to appear). [477,482]

Karrer, B., and Newman, M. E. J. (2011), "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E 83*, 016107. [472,477]

Ke, Z. T., Shi, F., and Xia, D. (2019), "Community Detection for Hypergraph Networks Via Regularized Tensor Power Iteration," arXiv:1909.06503. [483]

Kim, B., Lee, K. H., Xue, L., and Niu, X. (2018), "A Review of Dynamic Network Models With Latent Variables," *Statistics Surveys*, 12, 105. [472,475,480]

Li, T., Lei, L., Bhattacharyya, S., Van den Berge, K., Sarkar, P., Bickel, P. J., and Levina, E. (2020), "Hierarchical Community Detection by Recursive Partitioning," *Journal of the American Statistical Association*, 1–18. [477]

Liu, F., Choi, D., Xie, L., and Roeder, K. (2018), "Global Spectral Clustering in Dynamic Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 115, 927–932. [472,475]

Lu, X., and Szymanski, B. K. (2019), "A Regularized Stochastic Block Model for the Robust Community Detection in Complex Networks," *Science Reports*, 9, 1–9. [477]

Newman, M. E. J. (2004), "Coauthorship Networks and Patterns of Scientific Collaboration," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 5200–5205. [477]

——— (2006), "Modularity and Community Structure in Networks," *Proceedings of the National Academy of Sciences*, 103, 8577–8582. [480]

Rosvall, M., and Bergstrom, C. T. (2010), "Mapping Change in Large Networks," *PLOS ONE 5*(1), e8694. [484]

Silverman, B. W. (2016), Introduction to Discussion of "Coauthorship and Citation Networks for Statisticians," *Annals of Applied Statistics*, 10, 1777–1778. [469]

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. (2019), "Unsupervised Word Embeddings Capture Latent Knowledge From Materials Science Literature," *Nature*, 571, 95–98. [484]

Yuan, M., Feng, Y., and Shang, Z. (2018), "A Likelihood-Ratio Type Test for Stochastic Block Models With Bounded Degrees," arXiv:1807.04426. [482]

Yuan, M., Liu, R., Feng, Y., and Shang, Z. (2021), "Testing Community Structures for Hypergraphs," *Annals of Statistics*. [483]

Zhang, Y., Levina, E., Zhu, J. (2020), "Detecting Overlapping Communities in Networks Using Spectral Methods," *SIAM Journal of Mathematics of Data Science*, 2, 265–283. [472]

Zhao, Y., Levina, E., and Zhu, J. (2011), "Community Extraction for Social Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 108, 7321–7326. [472]

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

Taylor & Francis
Taylor & Francis Group

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.

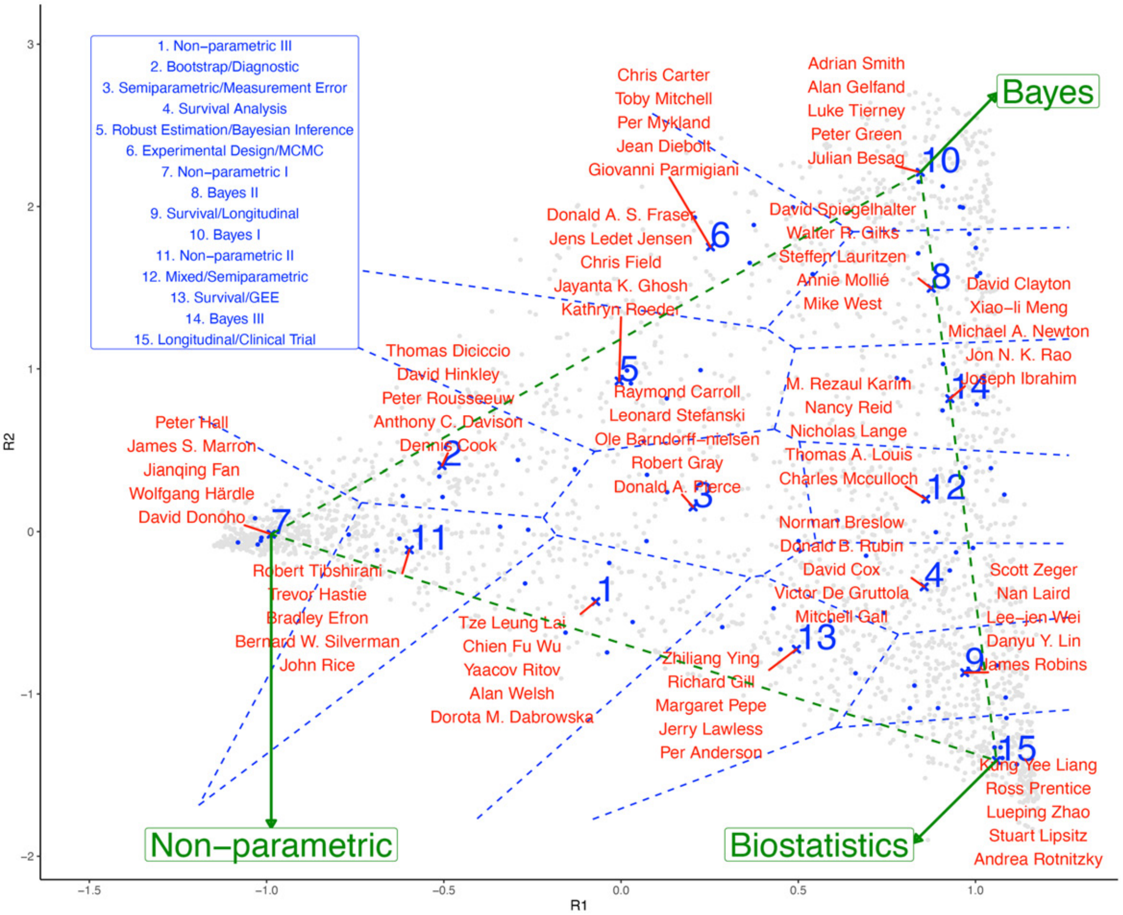

Figure D1. The research map obtained from the weighted citee network.

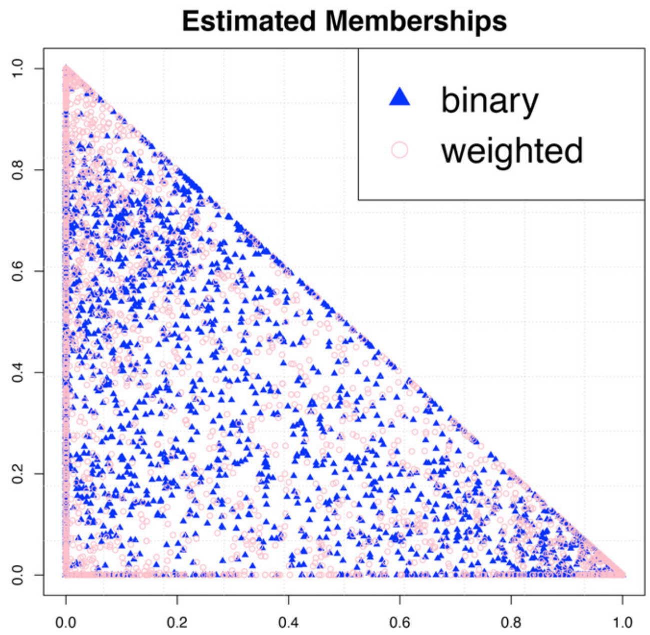

Figure D2. The estimated memberships $\{\hat{\pi}\}_{i=1}^{n}$ based on the binary and weighted citee networks. The X and Y coordinates correspond to the "nonparametric statistics" and "biostatistics" research areas.

- 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 blockmodels 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 $\Theta = \mathrm{diag}(\theta_1, \dots, \theta_n), \Pi = (\pi_1, \pi_2, \dots, \pi_n)'$, 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 $B = (b_1, \dots, b_K) \in \mathbb{R}^{K\times K}$ 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 $\hat{r}_i = \left[\frac{\hat{\xi}_2(i)}{\hat{\xi}_1(i)}, \frac{\hat{\xi}_3(i)}{\hat{\xi}_1(i)}, \dots, \frac{\hat{\xi}_K(i)}{\hat{\xi}_1(i)}\right]$, where $\hat{\xi}_1, \dots, \hat{\xi}_K$ are the first K eigenvectors of the adjacency matrix. The embedding is further generalized to handle dynamic citee networks.

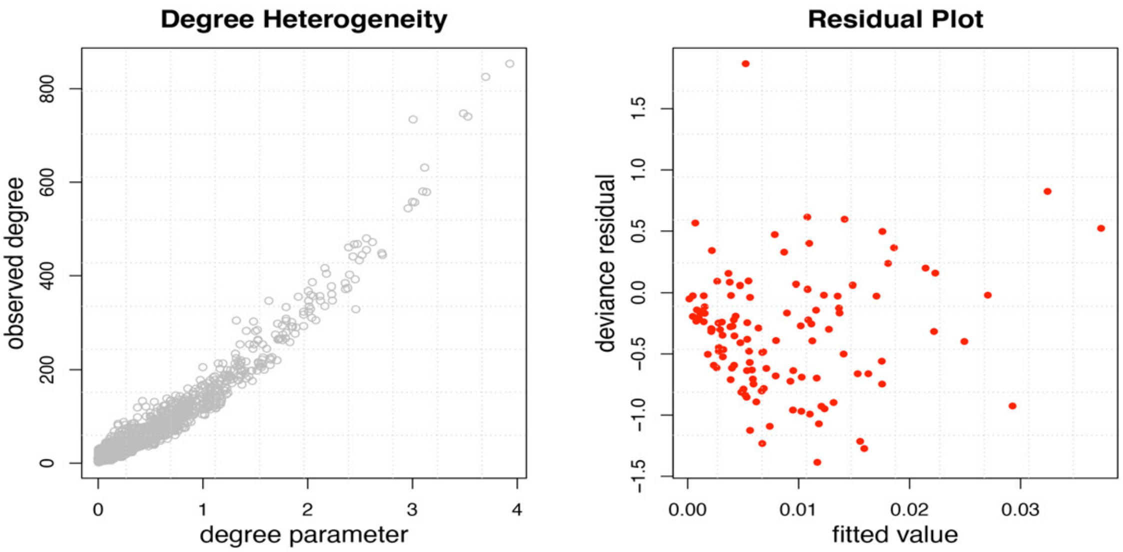

Figure D3. Left plot shows the relationship between observed degrees and estimated degree parameters based on mixed-SCORE method. Right plot is the deviance residual plot, where X-axis and Y-axis are the average of fitted values and average of deviance residuals (with normalization) within each bin,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

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 491
https://doi.org/10.1080/07350015.2022.2055356

Taylor & Francis
Taylor & Francis Group

Data Come First: Discussion of “Co-citation and Co-authorship Networks of Statisticians”

David Donoho

Stanford University, Stanford, CA

I salute the authors for their gift to the world of this new dataset! They have clearly invested plenty of time, effort, and IQ points in the study of the statistics literature as a bibliometric laboratory, and our field will grow and develop because of this dataset, as well as methodology the authors developed and/or fine-tuned with those data.

Strikingly, the article also conveys a great deal of enthusiasm for the data! This seems such a departure from the pattern of many articles in statistics today.

The enthusiastic spirit reminds me of some classic work by great figures in the history of statistics, who often were fascinated by new kinds of data which were just becoming available in their day, and who were inspired by the new data to invent fundamental new statistical tools and mathematical machinery. Francis Galton was interested in the relationships between father’s height and son’s height, himself compiling an extensive bivariate dataset of such heights, leading to the invention of the bivariate normal distribution and the correlation coefficient.

Time and time again, new types of data came first, new types of models and methodology later. Indeed, this seems almost inevitable. As new technologies come onstream, new kinds of measurements become available, and new settings for data analysis and statistical inference emerge. This is plain to see in recent decades, where computational biology produced gene expression data, DNA sequence data, SNP data, and RNA-Seq data, each new data type leading to interesting methodological challenges and scientific progress.

For me, 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.

However, many Ph.D. statisticians, particularly those who are early in their careers, might not agree; they might even have difficulty “getting” the point I’m trying to make. The data-first approach I cited earlier, giving the example of Galton, has not been academically dominant in recent decades. This approach might be considered something like footprints left in the forest by our ancestors: and in this case, those footprints became overgrown and hidden over the last century. It’s a wonder they were not lost forever.

Since the 1930s really, the literature of statistics—the very topic of this article (!)—has focused on models and methodology first; in some articles, occasionally data are provided as a kind of afterthought, simply to illustrate the article’s methodology concretely—the same way we might “tack on” a bibliography at the end of a article, we “tack on” a data example.

Indeed, this article uncovers a statistics triangle, revealing that our field’s literature is clustered around specific types of models and methodology. If the field were instead data-first, its literature might instead be organized around datasets; in which case the authors might have uncovered a very different type of clustering.

Remarkably, the authors discovered this model-first structure of the literature because they were willing to depart from the models-first tradition and work instead within the data-first tradition, developing a fresh high-quality dataset which could support scrutiny and discovery.

The authors could, amazingly, transcend their own modern, theory-driven upbringing and reinvent or relearn the largely forgotten data-first attitudes of an era prior to today’s. But doing so has paid off! Kudos to the authors for showing us the way, I hope that others can be inspired and that the literature of statistics will grow because of this shining example of the fruits of a data-first orientation.

CONTACT David Donoho donoho@stanford.edu Stanford University, Stanford, CA.

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 492–493
https://doi.org/10.1080/07350015.2022.2041423

Taylor & Francis
Taylor & Francis Group

Discussion of "Co-citation and Co-authorship Networks of Statisticians" by Pengsheng Ji, Jiashun Jin, Zheng Tracy Ke, and Wanshan Li

Peter W. MacDonald, Elizaveta Levina, and Ji Zhu

Department of Statistics, University of Michigan, Ann Arbor, MI

1. Introduction

We congratulate the authors on an interesting paper and on making an important contribution to the network analysis community through compiling a large new dataset which will spur further work on multilayer, dynamic and other complex network settings. This discussion focuses on the paper's particular methods and applications in dynamic network analysis. Complexity of dynamic network data leads to many necessary analyst choices in both data processing and network modeling. Where possible, we will compare the choices made in this paper with other possibilities from recent literature on dynamic network analysis.

One of the important points of the paper is that much of our network data has always been dynamic. For instance, communication networks consisting of sent and received E-mails come with time stamps, whether we choose to incorporate them or not. Developing statistical methods that take advantage of this time varying structure will lead to greater efficiency, novel insights, and generally allow us to take full advantage of rich modern datasets like the one featured in this paper.

2. Choices in Data Processing

Dynamic network data typically arrive in one of two general settings, which we term *snapshots* and *events*. Snapshot data, consisting of time stamped networks collected at prespecified times, generally arise from the active querying of a complex system at those times. Conversely, event data, consisting of time stamped dyadic events between nodes, is a natural outcome from passive observation of a complex system. It is common, however, as this paper does, to aggregate event data into fixed time windows, which results in a format similar to snapshot data. Typically, this makes it easier to extend single network methods to the aggregated data, but there is a growing literature on directly modeling event data (Crane and Dempsey 2018; Matias, Rebafka, and Villers 2018; Kreiß, Mammen, and Polonik 2019, among others), including first steps toward models for nondyadic (hyperedge) events (Mulder and Hoff 2021).

Windowing, the choice of how to aggregate the observed events, is an important stage in the processing of dynamic network data. This paper chooses to summarize 30 years of data into 21 overlapping time windows of varying length from 5 to 10 years in Section 2, and then summarizes the same data into three overlapping time windows in Section 3. For spectral methods like those applied here, varying window length gives the analyst direct control of the edge density and ultimately the eigengap, especially important for methods which operate on snapshots independently, and thus, require sufficient signal strength for each snapshot. Overlapping and nonoverlapping windows will influence downstream modeling assumptions, with overlapping windows suggesting serial dependence among snapshot entries. As noted by Kim et al. (2018), overlapping windows can also aid interpretation by stabilizing results over time. However, stability of results to window choice is also important, and would help increase confidence in the final results.

The node set will not necessarily remain constant across dynamic network snapshots. In Section 2, the authors restrict the node set to nodes present in the first snapshot; for citations that makes sense, since a paper can continue to be cited indefinitely. However, the node set of the coauthorship network in Section 3 changes across the three snapshots. Since the proposed estimation method operates on each snapshot independently, nodes have no effect on the analysis of snapshots where they do not appear. An alternative approach might be to explicitly model the arrival and departure of nodes from the system, as done, for example, in the dynamic stochastic block model (SBM) proposed by Matias and Miele (2017). This can potentially allow for more efficient information sharing across snapshots.

3. Choices in Network Modeling

The citation network analysis in Section 2 works with 21 network snapshots constructed by aggregating overlapping time windows. The dynamic degree-corrected mixed-membership (DCMM) model assumes independence of the network snapshots after accounting for the latent community structure. Dependence across networks is a challenging new area of research, including recent work on multiple community detection with dependence (Yuan and Qu 2021), and applications of more classical time series models to individual node-pair series (Jiang, Li, and Yao 2021).

CONTACT Peter W. MacDonald pwmacdon@umich.edu Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1382.

The dynamic DCMM model presented here, along with many other dynamic extensions of the SBM, and more general network latent space models can be categorized according to which parameters are constant over time and which are allowed to vary. In contrast to the dynamic DCMM model, Pensky and Zhang (2019) and Matias and Miele (2017) propose dynamic extensions of the SBM in which both the community memberships and community structure matrix are allowed to vary. In order to still share information across time, Pensky and Zhang (2019) assume smoothness of the parameters, while Matias and Miele (2017) community membership is governed by a Markov chain with nontime varying parameters. These models allow for greater flexibility and potentially better fit to the data, but they are typically challenging to fit, and can present issues in model identifiability and interpretability, as we discuss below.

An often cited property of network latent space models, including the SBM, is their intrinsic nonidentifiability to a class of transformations, which could be permutations (SBM), rotations (RDPG, the random dot product graph model), or indefinite orthogonal transformations (generalized RDPG (Rubin-Delanchy et al. 2020)). In the dynamic setting, this can lead to time varying nonidentifiability, in which the parameters at each snapshot are unknown up to a snapshot-specific unknown transformation. In the case of the dynamic DCMM model, parameter sharing (a constant $K \times K$ community structure matrix P) makes the model well identified, limiting the nonidentifiability to a single unknown transformation.

Although the dynamic DCMM model is well identified, the authors still need to account for time varying nonidentifiability in their estimation procedure. An intermediate stage of their estimation procedure embeds each network snapshot, and a naive approach could produce embeddings $\{\hat{r}_i^{(t)}\}_{i=1}^n$ for each snapshot with arbitrarily rotated (misaligned) columns. While the theoretically aligned embeddings may exhibit smoothness over time, this smoothness can be lost due to misalignment. The interpretability of a dynamic network model via plots of embedded trajectories like Figure 2, or those presented in Sewell and Chen (2015), also relies on correct alignment.

The reference network projection approach taken in Section 2.2 can be viewed as a particular method of aligning a collection of $(K-1)$-dimensional network embeddings (which may be the end goal of analysis or an intermediate step followed by clustering) by representing them all in the same basis. The authors note that it also serves as a denoising step. We also note that it relies on the assumption of homogeneity of the community structure matrix over time, so that projection onto the reference network eigenvectors does not interfere with the signal strength in the other snapshots.

Other approaches could align embeddings at consecutive times by solving a (possibly indefinite) Procrustes problem (Sanna Passino et al. 2021). This approach is a simple post-processing step which is logical under the assumption of network smoothness over time, but it does not necessarily fix issues caused by misalignment in the modeling and estimation stages. Alignment can be ensured during the estimation stage by jointly embedding a concatenated object, such as an omnibus adjacency matrix (Levin et al. 2017). However, the omnibus adjacency matrix is an $nT \times nT$ object that can be computationally unwieldy for operations like singular value decomposition. Estimation efficiency of the omnibus embedding depends on the homogeneity of the network over time (Draves and Sussman 2021).

4. Conclusion

In this discussion, we have used the methodology and analysis presented in the accompanying paper to consider new opportunities presented by dynamic network data, but also new issues which require careful consideration in data processing, modeling and interpretation. At the data processing stage, we have highlighted how control over windowing and aggregation can influence signal strength, smoothness across snapshots, and heterogeneity of the (possibly time varying) node set. At the modeling stage we discuss dependence, the key issue of alignment, and argue that nonidentifiability in network models requires careful consideration in the dynamic setting. The authors of this paper have revealed many insights about their new dataset by taking advantage of its dynamic structure, but as with any statistical analysis, these insights depend on choices, and the number of choices grows quickly with the complexity of the data. Recognizing these choices and their potential alternatives is a necessary step toward a complete and principled framework for dynamic network analysis.

Funding

Levina's research is partially supported by NSF DMS grants 1916222 and 2052918. Zhu's research is partially supported by NSF DMS grant 1821243.

References

Crane, H., and Dempsey, W. (2018), "Edge Exchangeable Models for Interaction Networks," *Journal of the American Statistical Association*, 113, 1311–1326. [492]

Draves, B., and Sussman, D. L. (2021), "Bias-Variance Tradeoffs in Joint Spectral Embeddings," *arXiv:2005.02511 [math, stat]*. [493]

Jiang, B., Li, J., and Yao, Q. (2021), "Autoregressive Networks," *arXiv:2010.04492 [stat]*. [492]

Kim, B., Lee, K. H., Xue, L., and Niu, X. (2018), "A Review of Dynamic Network Models with Latent Variables," *Statistics Surveys*, 12, 105–135. [492]

Kreiß, A., Mammen, E., and Polonik, W. (2019), "Nonparametric Inference for Continuous-Time Event Counting and Link-Based Dynamic Network Models," *Electronic Journal of Statistics*, 13, 2764–2829. [492]

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. [493]

Matias, C., and Miele, V. (2017), "Statistical Clustering of Temporal Networks Through a Dynamic Stochastic Block Model," *Journal of the Royal Statistical Society*, 79, 1119–1141. [492,493]

Matias, C., Rebafka, T., and Villers, F. (2018), "A Semiparametric Extension of the Stochastic Block Model for Longitudinal Networks," *Biometrika*, 105, 665–680. [492]

Mulder, J., and Hoff, P. D. (2021), "A Latent Variable Model for Relational Events with Multiple Receivers," arXiv:2101.05135 [stat]. [492]

Pensky, M., and Zhang, T. (2019), "Spectral Clustering in the Dynamic Stochastic Block Model," *Electronic Journal of Statistics*, 13, 678–709. [493]

Rubin-Delanchy, P., Cape, J., Tang, M., and Priebe, C. E. (2020), "A Statistical Interpretation of Spectral Embedding: The Generalised Random Dot Product Graph," arXiv:1709.05506 [cs, stat]. [493]

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. [493]

Sewell, D. K., and Chen, Y. (2015), "Latent Space Models for Dynamic Networks," *Journal of the American Statistical Association*, 110, 1646–1657. [493]

Yuan, Y., and Qu, A. (2021), "Community Detection with Dependent Connectivity," *The Annals of Statistics*, 49, 2378–2428. [492]

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 494–496
https://doi.org/10.1080/07350015.2022.2044335

Discussion of “Co-citation and Co-authorship Networks of Statisticians”

Xiaojing Zhu and Eric D. Kolaczyk

Department of Mathematics and Statistics, Boston University, Boston, MA

1. Introduction

We thank the authors for their new contribution to a high quality dataset and interesting findings from the modeling and analysis of the co-citation and co-authorship networks of statisticians. Leveraging this dataset, there are lots of additional questions that might be answered, and analyses done. Network motif analysis is one such, with roots in the triad census of traditional social network analysis (Wasserman and Faust 1994, chap. 14.2.1) and first introduced in its modern form by Milo et al. (2002) in systems biology. It has since been applied to various scientific domains, for example, social science, neuroscience, to study network structures and the underlying complex systems (see Stone, Simberloff, and Artzy-Randrup (2019) for a survey article).

While the notion of network motif was originally defined for static networks as small subgraph patterns occurring frequently in a given network, several ways have been proposed to extend it to dynamic networks consisting of a set of vertices and a collection of timestamped edges. One widely used one is from Paranjape, Benson, and Leskovec (2017), where temporal motifs are defined as an ordered sequence of timestamped edges among a subset of nodes conforming to a specified pattern as well as a specified duration of time δ in which the edges must occur. In contrast to their static counterparts, such temporal motifs take into account not only subgraph isomorphism but also edge ordering and duration, which can be regarded as the simple building blocks for temporal structures of dynamic networks.

There are a few works in the literature on motif analysis for journal citation networks (Wu, Han, and Li 2008; Zeng and Rong 2021) and author collaboration networks (Chakraborty, Ganguly, and Mukherjee 2015), but none of them seem to be from the perspective of temporal motifs. In this discussion, we construct temporal citation networks among statisticians using the publication data provided in the article, and focus on analyzing the frequency and distribution of temporal motifs in such dynamic networks. This analysis provides initial insights into the temporal patterns of citing behaviors among authors of various statistics journals from 1975 to 2015.

2. Definition of Temporal Author Citation Networks

The co-citation and co-authorship networks studied in the article are in the form of matrices or time series of matrices. Here, we consider networks in a different form consisting of a set of vertices and a collection of timestamped directed edges.

The dataset provided in the article contains two data resources: one is the citation records consisting of pairs of citing and cited article; the other is the article related information including authors, year and journal name of publication for each article. Using both sources of information, we construct a temporal author citation network as a series of timestamped directed edges, where nodes represent authors and a directed timestamped edge pointing from a citer to a citee represents that an author cited at least one article from another author within a given year. We exclude all self-loops. The resulting network includes 1,768,050 citing interactions among 43,521 authors from year 1975 to year 2015.

We also construct temporal networks in the same manner for the following three different subcategories of journals: (a) the four flagship journals consisting of AoS, Bka, JASA and JRSSB; (b) the IMS family, consisting of AoAS, AoP, AoS, StSci, EJS and JCGS; and (c) the RSS family, consisting of JRSSA, JRSSB and JRSSC (see Table B.1 in the article for full journal names), where only citing interactions among authors and their articles published within each journal subcategory are considered for each of the three temporal networks constructed. The resulting networks for the three journal subcategories contain 230,821, 93,180 and 25,940 citing interactions among 8998, 7363 and 4001 authors, respectively, from year 1975 to year 2015.

With all these networks, we are interested in understanding the temporal patterns of citing behavior among authors and its change over time.

3. Temporal Motif Analysis Results

For each of the constructed temporal networks, we use the `snap`[1] package to count the number of occurrences of several

[1] *https://snap.stanford.edu/temporal-motifs/code.html*.

CONTACT Eric D. Kolaczyk kolaczyk@bu.edu Department of Mathematics and Statistics, Boston University, Boston, MA.

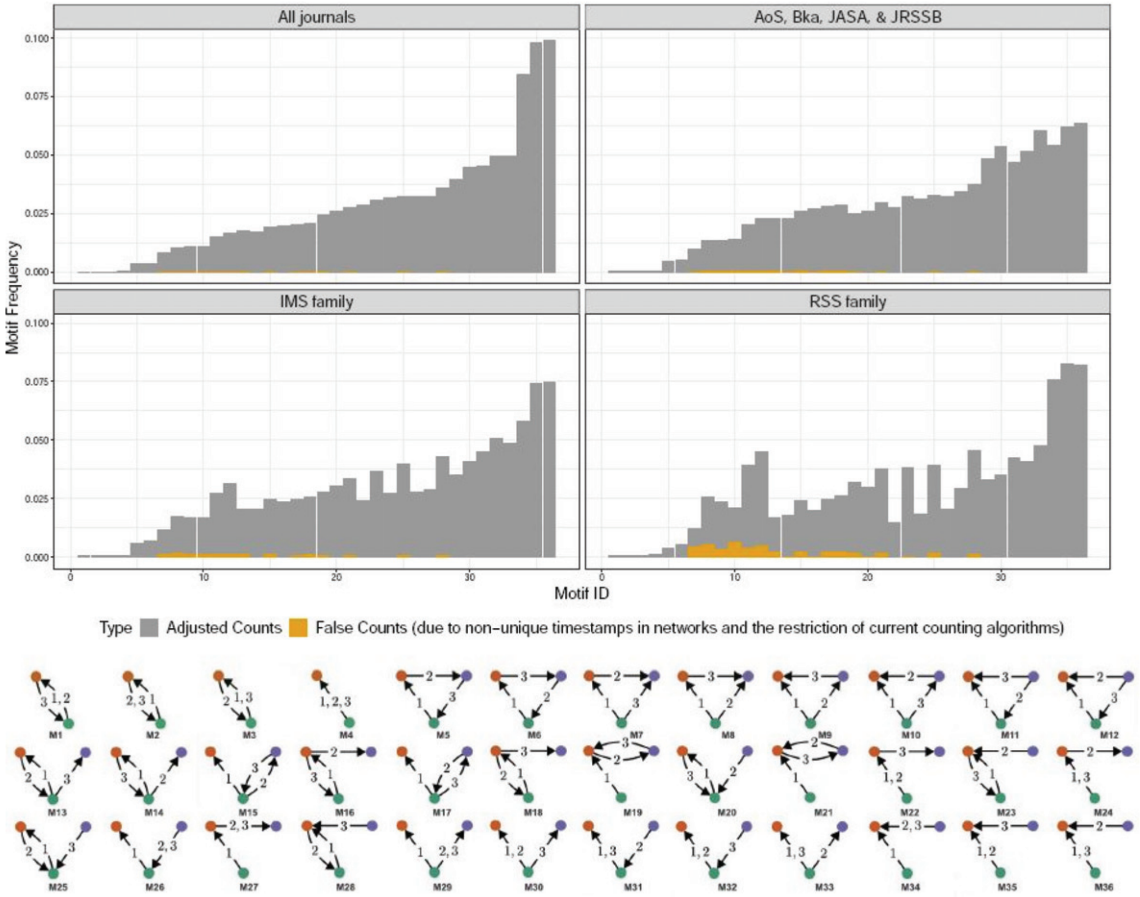

Figure 1. *Top*: frequency distributions of the 36 two/three-node, three-edge temporal motifs in four temporal citation networks constructed within different categories of journals. *Bottom*: the 36 temporal motifs with IDs corresponding to the *x*-axis of the bar plots, ordered by relative frequency for all journals.

temporal motifs in the network within a sliding observation window of $\delta = 5$ years, with results[2] shown in Figure 1. However, this is not an exact but an overestimated motif count in our networks since all temporal motif counting methods are currently built upon the assumption that each event in the given network has a unique timestamp, and it is not the case for our networks with edge resolution of one year. Consequently, some citing interactions with nonunique timestamps are mistakenly counted as motif instances by the algorithm. We adjust for one source of such error where all three citing interactions in the motif instance are of the same timestamps by subtracting such false counts (as illustrated in yellow in Figure 1) from the total counts, while the other source of false motif counts where only two citing interactions are of the same timestamps cannot be easily adjusted for given the restrictions in current motif counting methods. To fully address this issue, either finer time resolution of timestamped edges or a motif counting method adapted for nonunique timestamps is needed. For now, we use the partially adjusted counts in the following temporal motif analysis, which are the best estimates we can come up with so far for the true motif counts.

Figure 1 shows frequency distribution of all the two/three-node, three-edge temporal motifs (36 motifs in total shown on the bottom of Figure 1 with motif ID from 1 to 36), reflecting the behavior of citing patterns among two or three authors publishing in different categories of journals. The plots are rich and space here precludes a comprehensive analysis, but here are a few examples of findings that can be obtained from the plots. In the full journal analysis, we can see that the most frequently occurring citing patterns (M34, M35, M36) are two authors successively citing another author in 5 years which reflects the broad impact of some seminal works, while the least frequent ones (M1, M2, M3, M4) are two authors citing each other, or one author citing another author multiple times in 5 years, which demonstrates that reciprocal citations across time occur relatively more rarely in the community of statistics. The motif frequency distributions for the three subsets of journals look different from each other and also different from the one for all journals. The most eye-catching differences are: (a) the frequency of motif M34, M35 and M36 decreases in the AoS, Bka, JASA and JRSSB category compared with other journal categories although they still occurred quite frequently; (b) while all the triadic patterns seem to occur less frequently in each journal category, the frequency of two triadic patterns (M11 and M12) seems to be more prominent in the RSS family.

Figure 2 shows the frequency distribution of the 36 temporal motifs across four decades, from which we can see a change in the behavior of citing patterns from 1975 to 2015. One interesting observation is that the two lines for 1996–2005 and

[2] Analysis code is available at *https://github.com/KolaczykResearch/TempMotifOnStatCitationNets*.

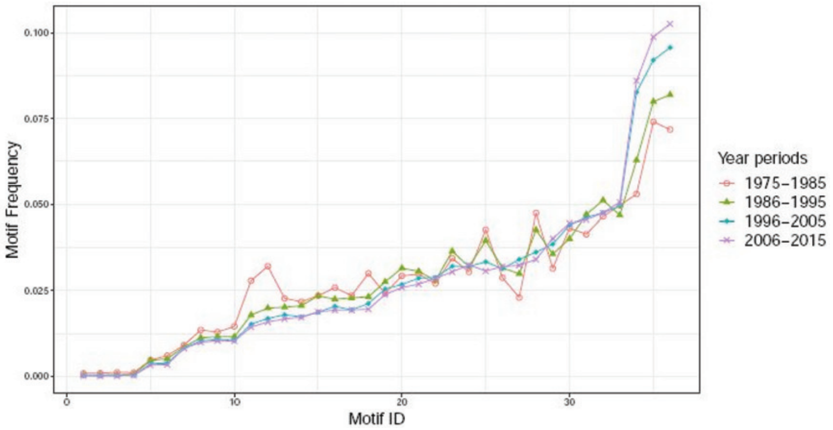

Figure 2. The comparison of motif frequency distributions across four decades: 1975–1985, 1986–1995, 1996–2005, and 2006–2015. Motif IDs in the *x*-axis correspond to the temporal motifs shown on the bottom of Figure 1.

2006–2015 align quite well with each, indicating that the citation pattern during 2006–2015 didn't change much from that during 1996–2005. However, the citation patterns from 1975 to 2005 seem to evolve every 10 years. For example, the frequency of motif M27, M34, M35 and M36 increases while that of motif M11, M12, M25 and M28 decreases from 1975 to 2005.

Funding

This work was supported by NSF award SES-2120115.

References

Chakraborty, T., Ganguly, N., and Mukherjee, A. (2015), "An Author is Known by the Context She Keeps: Significance of Network Motifs in Scientific Collaborations," *Social Network Analysis and Mining*, 5, 16. [494]

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002), "Network Motifs: Simple Building Blocks of Complex Networks," *Science*, 298, 824–827. [494]

Paranjape, A., Benson, A. R., and Leskovec, J. (2017), "Motifs in Temporal Networks," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 601–610. [494]

Stone, L., Simberloff, D., and Artzy-Randrup, Y. (2019), "Network Motifs and their Origins," *PLoS Computational Biology*, 15, e1006749. [494]

Wasserman, S., and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press. [494]

Wu, W., Han, Y., and Li, D. (2008), "The Topology and Motif Analysis of Journal Citation Networks," in *2008 International Conference on Computer Science and Software Engineering* (Vol. 1), pp. 287–293. IEEE. [494]

Zeng, X., and Rong, Z. (2021), "Evolution of the Physics Citation Network with Motifs," in *2021 40th Chinese Control Conference (CCC)*, pp. 776–780. IEEE. [494]

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 2, 497–498
https://doi.org/10.1080/07350015.2022.2044828

Taylor & Francis
Taylor & Francis Group

Discussion of "Co-citation and Co-authorship Networks of Statisticians"

Joshua Daniel Loyal and Yuguo Chen

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL

We want to congratulate the authors on a fascinating article containing an insightful analysis and their hard work curating the high-quality co-citation and co-authorship networks. These datasets alone are a valuable contribution to the statistics profession, which will undoubtedly inspire future data science projects and advances in methodology. In fact, we are eager to use these networks in our own classrooms and research. Furthermore, the authors use these networks to tackling exciting questions in network science that go beyond the familiar problems of edge imputation and predicting node labels. In doing so, the authors perform a terrific analysis accompanied by exciting new methodology. This analysis serves as a great first step in understanding these networks, and the ideas initiated in this article will certainly stimulate many further research questions. For example, how do individuals influence the research trajectory of others? Or, how do the components of the proposed "research map" change over time? As statisticians, we have a first-hand understanding of the complex system these networks describe, which can help us contextualize these problems and validate our inferences. As such, we look forward to this dataset becoming a standard benchmark to test new models and scalable inference procedures.

A central challenge of the work is rigorously quantifying the time-varying research patterns and trends of the statistics community, which naturally leads to the statistical modeling of dynamic networks. The authors skillfully use various dynamic block models to uncover statisticians' community structure. In the remainder of this discussion, we focus on an alternative statistical network model known as latent space models. Specifically, we briefly describe the latent space modeling approach, highlight five further research questions, and demonstrate how latent space models may be used to answer them. Although other models, such as block models, may be appropriate to tackle these questions as well, we hope that this discussion gives future researchers an expanded toolset to investigate this rich data source.

Latent space models (LSMs) are a popular approach to modeling networks first proposed by Hoff, Raftery, and Handcock (2002) for static networks and later generalized to dynamic networks by Sarkar and Moore (2006) and Sewell and Chen (2015). These models embed the nodes of a network into a low-dimensional latent space, which can provide meaningful visualizations and insights into the evolution of a network. In particular, consider T binary undirected networks on a common set of n nodes, and let $A_1, \ldots, A_T$ be their adjacency matrices with entries $A_t(i,j)$. Also, let $\mathbf{u}_{it} \in \mathbb{R}^d$ be the latent position of the ith node at time t. Dynamic LSMs posit that

$$\mathbb{P}(A_t(i,j) = 1) = f(h_{\boldsymbol{\psi}}(\mathbf{u}_{it}, \mathbf{u}_{jt}), \boldsymbol{\theta}), \qquad (1)$$

where $h_{\boldsymbol{\psi}}$ is a similarity function that depends on parameters $\boldsymbol{\psi}$, f is an inverse-link function, and $\boldsymbol{\theta}$ are additional parameters. To capture temporal correlations, the latent positions evolve over time as Markov processes:

$$\mathbf{u}_{i1} \overset{\text{iid}}{\sim} N(0, \tau^2 I_d), \quad \mathbf{u}_{it} \sim N(\mathbf{u}_{i(t-1)}, \sigma^2 I_d), \quad t = 1, \ldots, T.$$

Furthermore, one assumes $A_1, \ldots, A_T$ are independent given the latent positions. As defined, LSMs are flexible models that can capture various properties of dynamic networks.

1. *The role of node and dyad attributes.* Incorporating additional features such as author characteristics (e.g., institution, department, academic rank, etc.) could yield interesting insights into the statistics community's co-citation and co-authorship patterns. As the authors observe in the text: "collaborations may be driven by many factors (e.g., geographical proximity, academic genealogy, cultural ties)." In Section 3, the authors answer this question by associating clusters inferred with a degree-corrected block model with attributes in a post hoc manner. Another approach is to incorporate the features directly into the network model. The LSM framework can formally quantify the effect of covariates on edge formation by using the following likelihood in Equation (1):

$$\text{logit}\{\mathbb{P}(A_t(i,j) = 1)\} = \beta_t^{\mathrm{T}} \mathbf{X}_{ijt} + \mathbf{u}_{it}^{\mathrm{T}} \mathbf{u}_{jt},$$

where $\mathbf{X}_{ijt}$ is a vector of dyad-specific covariates and β_t is a time-varying vector of coefficients. This approach can be understood as a generalized bilinear mixed-effect model (Hoff 2005, 2021). The latent positions are mean-zero random-effects ($\mathbb{E}[\mathbf{u}_{it}^{\mathrm{T}} \mathbf{u}_{jt}] = 0$) that capture residual network correlations such as transitivity. For example, we can use this model to investigate whether geographical proximity has had a decreasing effect over time on co-authorship as virtual communication platforms became popular.

CONTACT Yuguo Chen yuguo@illinois.edu Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL.

2. *Inferring an evolving research map.* Just as the co-authorship community structure changes over time, it is reasonable to assume that the research areas of the research map do not remain static from 1991 to 2015. In fact, statistical network analysis has emerged as a popular research topic during this time. An alternative to the mixed-membership model for community detection involves clustering the nodes according to their positions in latent space (Handcock, Raftery, and Tantrum 2007; Sewell and Chen 2017). To infer an evolving community structure, Loyal and Chen (in press) focused on the following LSM likelihood

$$\text{logit}\{\mathbb{P}(A_t(i,j) = 1)\} = \beta_0 - \left\|\mathbf{u}_{it} - \mathbf{u}_{jt}\right\|_2,$$

and proposed a Bayesian nonparametric approach that can infer additions, deletions, splits, and mergers of communities. This model could elicit changes in statistics research areas when applied to the co-citation networks.

3. *Measuring research attraction.* We can use dynamic LSMs to answer our previous question on how individuals influence the research trajectory of others through a concept called edge attraction (Sewell and Chen 2015). The edge attraction between nodes *i* and *j* measures the tendency of node *i* to move through the latent space in the direction of another node *j*. Sewell and Chen (2015) developed a test for the presence of edge attraction between two nodes. It would be exciting to develop a similar concept for the research trajectories estimated by the dynamic DCMM model to study the co-movement of statisticians' research interests.

4. *Accounting for co-citation and co-authorship counts.* When constructing the co-citation and co-authorship networks, the authors convert the weighted networks of counts into unweighted networks by applying a threshold to the edge weights. This procedure may affect the detected communities since it equates edges with low and high counts. It would be interesting to compare how the research map and co-authorship communities change (or not) when accounting for an edge's strength. In the context of LSMs, the model accounts for weighted edges by assuming the dyads in the networks, $A_t(i,j)$, arise from a generalized mixed model

$$g(\mathbb{E}[A_t(i,j)]) = \beta^{\mathrm{T}}\mathbf{X}_{ijt} + h_{\boldsymbol{\psi}}(\mathbf{u}_{it}, \mathbf{u}_{jt}),$$

where *g* is a link function. Sewell and Chen (2016) introduced likelihoods for various weighted networks, including networks with count-valued edges. As before, a clustering model can be applied to the latent positions to detect communities in the networks.

5. *Pooling information across co-citation and co-authorship networks.* The analysis in Section 3 indicates that both co-citation and co-authorship relations contain information about statistics research areas with many communities corresponding to statistics sub-fields. It would be interesting to combine these two relations by viewing the co-authorship and co-citation networks as components of a dynamic multilayer network, a collection of dynamic networks defined on a common set of nodes. Specifically, let A_t^k indicate the adjacency matrix for relation *k* (i.e., co-citation or co-authorship) measured at time *t* with entries $A_t^k(i,j)$. To infer structure shared across the two relations, Loyal and Chen (2021) proposed modeling these adjacency matrices with a shared dynamic latent space as follows:

$$\text{logit}\{\mathbb{P}(A_t^k(i,j) = 1)\} = \theta_{it}^k + \theta_{jt}^k + \mathbf{u}_{it}^{\mathrm{T}}\Lambda_k\mathbf{u}_{jt},$$

where $\theta_{it}^k \in \mathbb{R}$ models degree heterogeneity across time and relation, and Λ_k is a diagonal matrix that allows the relations to apply different weights to the shared latent features. One can infer communities shared by the co-citation and co-authorship relations by clustering the latent positions.

Again, we want to congratulate the authors for a fine contribution. The authors do a tremendous job developing methods and theory to answer complex questions in network science. In particular, it is exciting to see the power of modern statistical network analysis in uncovering information about our academic community. We look forward to the ideas presented in this article and the co-citation and co-authorship networks stimulating more exciting research in the future.

References

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007), "Model-Based Clustering of Social Networks," *Journal of the Royal Statistical Society*, Series A, 170, 301–354. [498]

Hoff, P. D. (2005), "Bilinear Mixed-Effects Models for Dyadic Data," *Journal of the American Statistical Association*, 100, 286–295. [497]

——— (2021), "Additive and Multiplicative Effects Network Models," *Statistical Science*, 36, 34–50. [497]

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098. [497]

Loyal, J. D. and Chen, Y. (2021), "An Eigenmodel for Dynamic Multilayer Networks," arXiv preprint arxiv:2103.12831. [498]

——— (in press), "A Bayesian Nonparametric Latent Space Approach to Modeling Evolving Communities in Dynamic Networks," *Bayesian Analysis*. [498]

Sarkar, P., and Moore, A. W. (2006), "Dynamic Social Network Analysis using Latent Space Models," in *Advances in Neural Information Processing Systems*, pp. 1145–1152. [497]

Sewell, D. K. and Chen, Y. (2015), "Latent Space Models for Dynamic Networks," *Journal of the American Statistical Association*, 110, 1646–1657. [497,498]

——— (2016), "Latent Space Models for Dynamic Networks with Weighted Edges," *Social Networks*, 44, 105–116. [498]

——— (2017), "Latent Space Approaches to Community Detection in Dynamic Networks," *Bayesian Analysis*, 12, 351–377. [498]

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 $1 \le i \neq j \le n$,

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

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 Kolacyzk 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 $\hat{r}_1, \hat{r}_2, \dots, \hat{r}_n$. 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 $\hat{r}_1, \dots, \hat{r}_n$ or the plot of $\hat{\pi}_1, \dots, \hat{\pi}_n$? (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 $\hat{r}_1, \dots, \hat{r}_n$ contain more information from the raw data. To see the point, recall that $\hat{\pi}_1, \dots, \hat{\pi}_n$ are obtained as follows. First, we use $\hat{r}_1, \dots, \hat{r}_n$ 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 $\tilde{\pi}_i(k) = \hat{w}_i(k)/\hat{b}(k)$, $1 \leq k \leq K$, 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 Kolaczyk 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 coauthorship 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]