Co-citation and Co-authorship Networks of Statisticians

Pengsheng Ji*, Jiashun Jin**, Zheng Tracy Ke†, Wanshan Li**
University of Georgia*, Carnegie Mellon University** and Harvard University†

Abstract

We collected and cleaned a large data set on publications in statistics. The data set consists of the coauthor relationships and citation relationships of 83,331 papers published in 36 representative journals in statistics, probability, and machine learning, spanning 41 years. The data set 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 paper we focus on (i) using the citation relationships to estimate the research interests of authors, and (ii) using the coauthor 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.

Keywords. Citation, coauthorship, community detection, dynamic network, mixed membership estimation, personalized network, hierarchical community tree.
AMS 2010 subject classification. 62P25, 62-07, 62G05, 62G10.

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

A new data set about the publications of statisticians. We present a new data set about the publications of statisticians, collected and cleaned by ourselves with enormous

efforts. The data set consists of coauthor relationships and citation relationships of $83K$ research papers 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	#papers
36	1975-2015	47,311	83,331

One might think that the data set is easy to obtain, as BibTeX and citation data seem to be easy to download. Unfortunately, when we need a large-volume, high-quality data set, 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 data set 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.

Results, findings, and challenges. First, we overview the results. Our data set provides rich material for research and motivates many interesting problems for research trends, patterns, and impacts of the statistics community. In this paper, 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 coauthorship 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 et al., 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 coauthorship 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 Sections 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 data set, 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 data set is never easy (let alone the time and efforts required for obtaining the data set). 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 2-3.

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

Contributions, broader impacts, and disclaimers. We have several major contributions. First, we contribute a high-quality, large-scale data set, 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 data set, 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 data set 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 a paper, or an area) over others. While we tried very hard to create a high-quality data set, 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 paper 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 paper are still valid. See Section A of the supplement for a longer version of the disclaimers.

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 coauthorship 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 data set 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 papers (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 "Non-parametric". 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 et al., 2017; Zhang et al., 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)-(c). We first use the mixed-SCORE (Jin et al., 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 et al. (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 et al., 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 papers 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 data set, 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.

Below, 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 papers by author i published between year $t-9$ to t and the papers by author j published between year $t-9$ and t have been

cited *together* in a paper 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 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 1999-2000, which has 2,831 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 paper 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 paper 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 et al., 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 $\hat{\xi}_1, \ldots \hat{\xi}_K \in \mathbb{R}^n$ be the first K eigenvectors of the adjacency matrix. Each node i is embedded into a $(K-1)$-dimensional space by the vector

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

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 $v_1, v_2, \ldots, v_K$ 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 et al. (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 $K \in \{2, 3, \ldots, 6\}$ 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 $K \geq 4$, 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 $\pi_1, \pi_2, \ldots, \pi_n$ by $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n$. 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, $1 \leq k \leq 3$. 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 papers. 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.

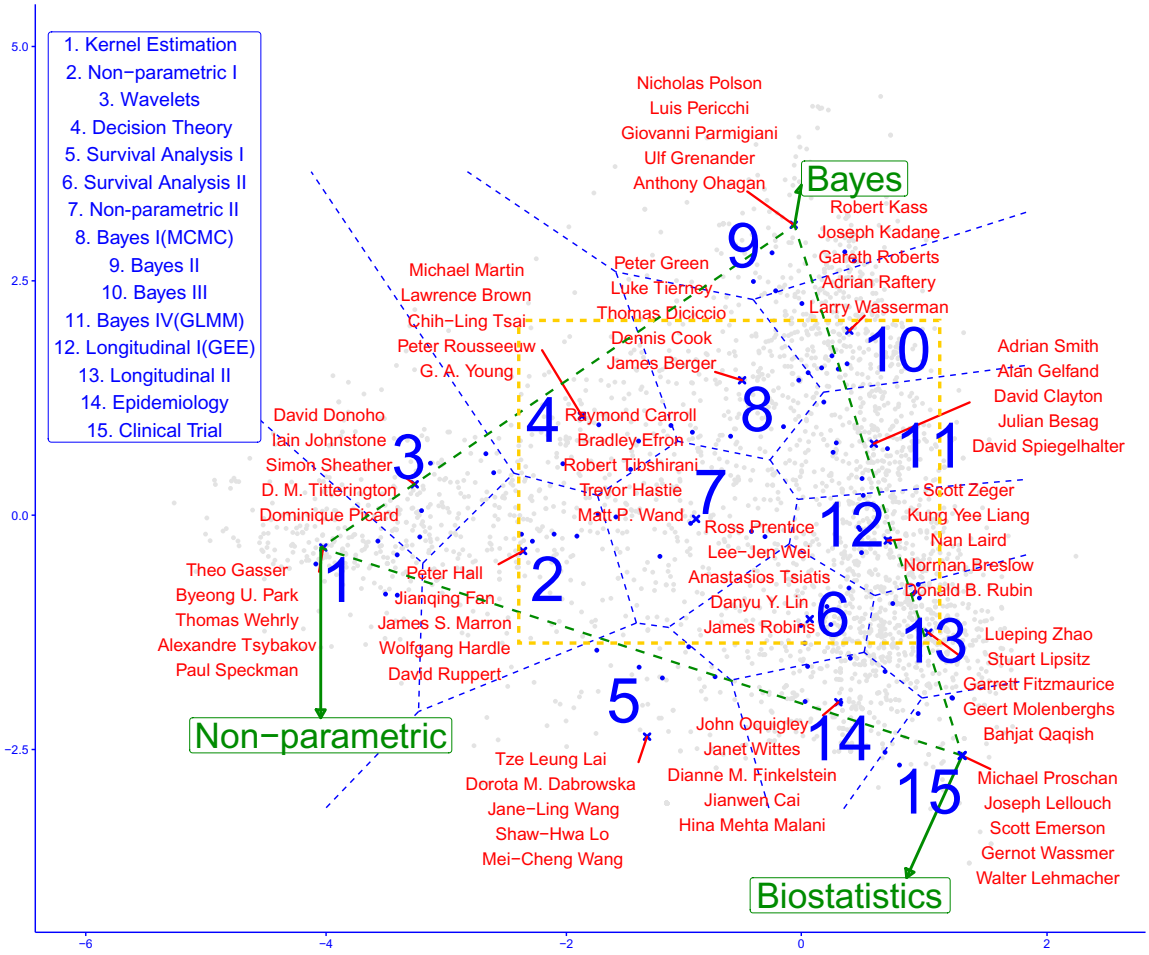

Figure 1: The research map. Each grey dot represents a 2-dimensional SCORE embedding vector $\hat{r}_i$, $1 \le i \le n$, and the 15 clusters and Voronoi diagram are obtained by applying the K-means algorithm to $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n$. The dashed green line represents the triangle, 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.

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 $L = 10, 11, \ldots, 20$ 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, $1 \leq i \leq n$, 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).

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.

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

Table 1: The 21 time windows we use to study the research trajectories. For example, the first window is from 1991 to 2000, covering a 10-year time period.

We consider 21 time windows (see Table 1) and construct a citee network for each of them. As the numbers of papers 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) as the 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 2,831 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

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

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 $1 \leq i \leq n$ 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 $\{\hat{\pi}_i^{(t)}\}_{1\leq i\leq n}$ 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 $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K$ be the K largest eigenvalues (in magnitude) of A_1, and let $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$ be the corresponding eigenvectors. For each $1 \leq t \leq T$ and each node $1 \leq i \leq n$, define a $(K-1)$-dimensional vector $\hat{r}_i^{(t)}$ by (e_i: the ith standard basis vector of $\mathbb{R}^n$)

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

Now, for each time t, we obtain the low-dimensional embedding $\{\hat{r}_i^{(t)}\}_{1\leq i\leq n}$ of all n nodes, and for each node i, we obtain the embedded "trajectory" as $(\hat{r}_i^{(1)}, \hat{r}_i^{(2)}, \ldots, \hat{r}_i^{(T)})$. 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 (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 utilizes the same $(\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K)$ 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 (2.4), the data to project, $A_t e_i$, and the projection direction, $\hat{\xi}_k$, are *independent* of each other, for any $t \geq 2$. 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 (2.4). In model (2.3), let $\Theta^{(t)} = \mathrm{diag}(\theta_1^{(t)}, \ldots, \theta_n^{(t)})$, $\Pi^{(t)} = [\pi_1^{(t)}, \ldots, \pi_n^{(t)}]'$, and $\Omega_t = \Theta^{(t)}\Pi^{(t)}P(\Pi^{(t)})'\Theta^{(t)}$, $1 \leq t \leq T$. Let $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_K)$ and $\Xi = [\xi_1, \xi_2, \ldots, \xi_K]$, where λ_k is the k-th largest (in magnitude) eigenvalue of Ω_1 and ξ_k is the corresponding eigenvector. For $1 \leq t \leq T$ and $1 \leq i \leq n$, define $r_i^{(t)} \in \mathbb{R}^{K-1}$ by

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

Theorem 2.1. *Consider the dynamic DCMM model (2.3). For each $1 \leq t \leq T$, letting $M_t = P(\Pi^{(t)})'\Theta^{(t)}\Xi\Lambda^{-1} \in \mathbb{R}^{K,K}$, we suppose $\mathrm{rank}(M_t) = K$ and $\min_{1\leq k\leq K}\{M_t(1,k)\} > 0$.*

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* $v_1^{(t)}, \ldots, v_K^{(t)}$. *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* $v_1^{(t)}, v_2^{(t)}, \ldots, v_K^{(t)}$, *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 et al. (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 manuscript We now apply the procedure to our data set.

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 (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 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-*

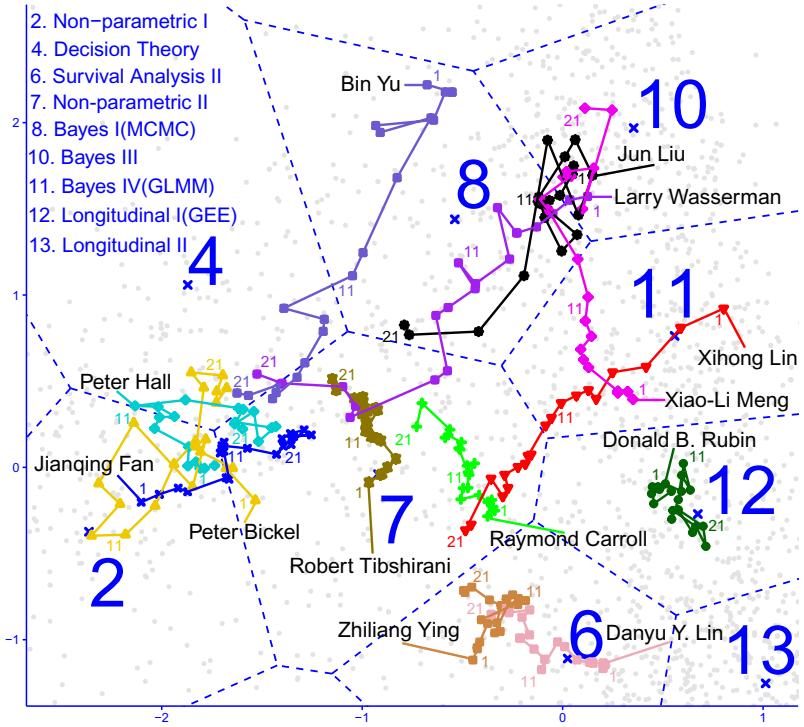

Figure 2: Research trajectories of representative authors (this is a zoomed-in view of the region in Figure 1 within the dashed yellow square, with the same Voronoi diagram). Each trajectory has 21 knots, corresponding to the 21 time windows in Table 1 (knots 1, 11, and 21 are marked with 1, 11, and 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 data set. 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.

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.

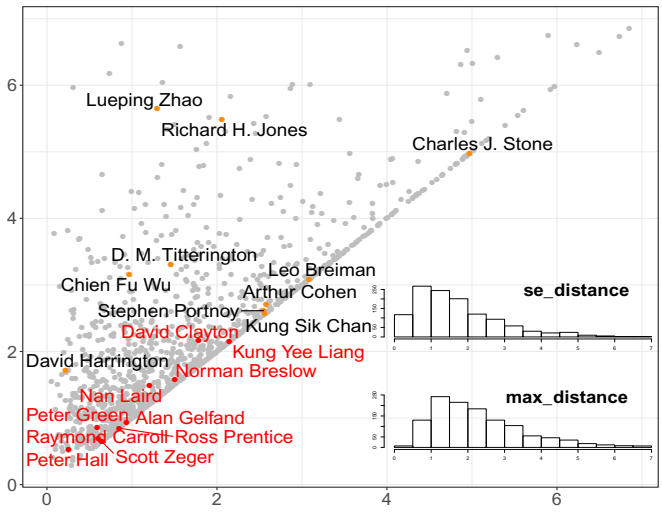

Figure 3: The two diversity metrics of 1,202 authors (x-axis: se_distance; y-axis: max_distance). The red dots represent the 10 highest-degree authors. The orange dots represent (among the top 200 highest-degree nodes) the 5 authors with the largest se_distance and the 5 authors with the largest differences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

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 $\hat{r}_i^{(1)}, \ldots, \hat{r}_i^{(21)}$. 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 1,202 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 1,202 authors. In this figure, the 10 highest-degree nodes are marked with red

[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.

dots, where their names are also presented in red. Also, among the 200 authors who have the largest degrees, the 5 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 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 coauthorship networks

The study of coauthorship patterns and community structures in an academic society is an interesting topic (Newman, 2004). The co-author relationship in our data set 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, 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, 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 coauthorship network, where $A(i,j) = 1$ if and only if authors i and j have co-authored papers in the range of interest. In DCBM, we assume

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

where $(P, \theta_1, \theta_2, \ldots, \theta_n)$ are the same as those in (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 data set, 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 (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 k-th 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 k-th largest eigenvector of A if it corresponds to the k-th 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 et al. (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 et al. (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 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 coauthorshp network (36 journals). We build a coauthorship 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 coauthored at least m_0 papers 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

Community	Description
C1. Non-parametric Statistics	Decision theory, non-parametric 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

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

between active researchers and non-actives 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 4,383 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 paper 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

[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.

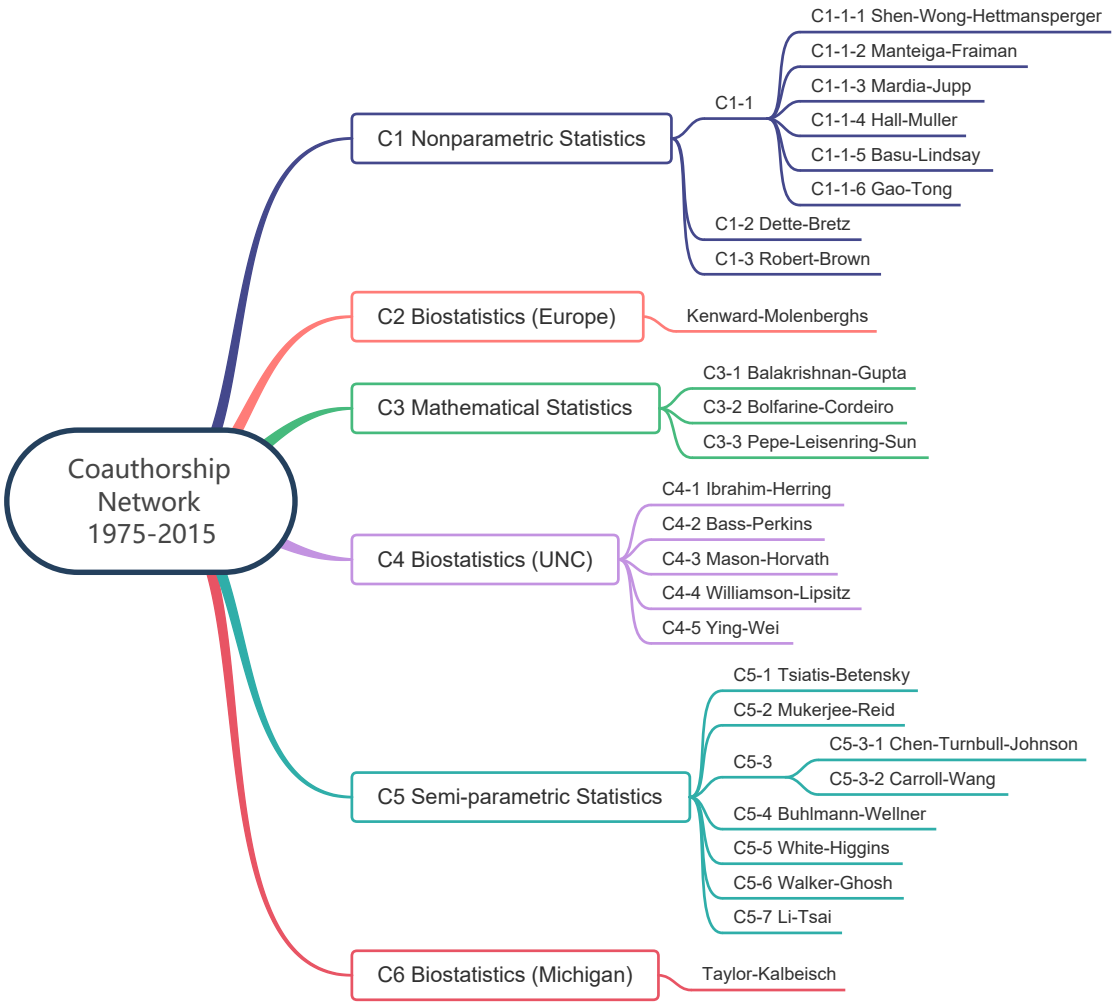

Figure 4: The community tree for coauthorship network. Each rightmost leaf community is labeled with the last names of 2 or 3 authors, selected by node betweenness and closeness. For each leaf, the representative nodes are shown in Table 3 (and Tables D.4-D.6 in the supplement).

250 nodes. We thus keep C6 unchanged. (b) The second largest component of C4 contains 60 nodes which 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 ≈ 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 coauthorship: 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

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	.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	.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	.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	.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	.0003	*Gauss M. Cordeiro*, *Heleno Bolfarine*, Victor H. Lachos, Reinaldo B. Arellano-valle
C3-3	Pepe-Leisenring-Sun	86	.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	.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	.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	.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	.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	.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	.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	.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	.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

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

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 6 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 6 communities identified by Newman's approach. Therefore, at least to some extent, two clustering results are consistent with each other.

3.2 Evolvement of coauthorship clusters

Our data set 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 coauthorship 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 coauthorship network constructed with only papers from 4 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 papers 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 coauthorship 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 coauthorship network where each author who has ever published in any of the 4 aforementioned journals during this time period is a node, and two nodes have an edge if and only if they have coauthored one or more papers. 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 coauthorship 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 1,687 nodes, for the Sankey diagram.

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)*, *nonparamet-*

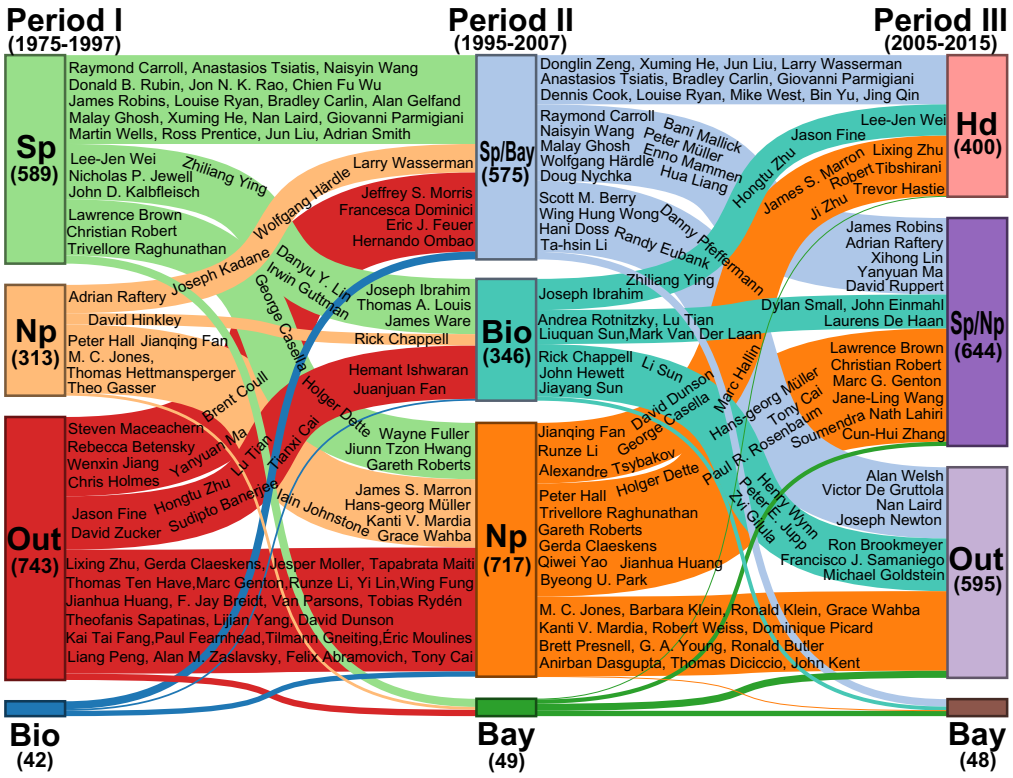

Figure 5: Evolution of communities in the dynamic coauthorship network (based on papers in 4 journals). The representative authors are selected by average degree in two adjacent networks.

ric (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) 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 coauthorship network similar to that in Section 3.1 but with $m_0 = 1$: Every author who ever published a paper in any of the 36 journals between 1975 and 2015 is a node, and two nodes have an edge if and only if they coauthored one or more papers. Once this large network is constructed, for every author, we can obtain a personalized coauthorship network accordingly.

We model each personalized coauthorship network with a DCBM model (2.1) with K communities. We consider the global testing problem (Yuan et al., 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 coauthorship network has only one or multiple tight-knit groups. We approach the testing problem by the SgnQ test (Jin et al., 2021) which was already described in Section 3.1. Let Q_i be the test score ψ_n in (3.7) for the personalized

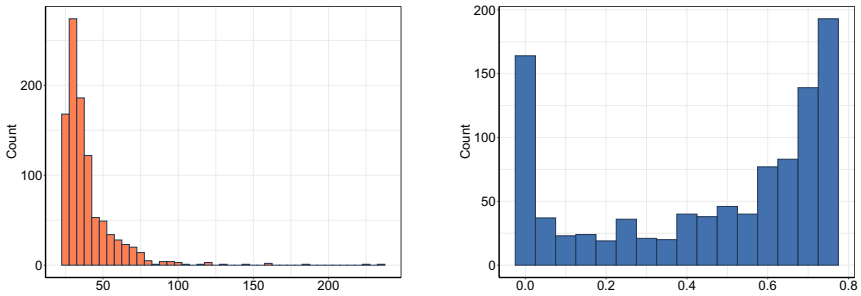

Figure 6: Left: histogram for the numbers of coauthors of 1,000 authors who have the largest number of coauthors in our data set. Right: histogram for the SgnQ p-values for the 1,000 personalized coauthorship networks. A smaller p-value suggests that the personalized network is more likely to have multiple tight-knit groups (so the author is more diverse in terms of coauthorship).

Name	#Coau	p-value	Name	#Coau	p-value	Name	#Coau	p-value
Raymond Carroll	234	.02	Geert Molenberghs	146	0	Pranab Kumar Sen	112	.71
Peter Hall	222	.23	James S. Marron	130	.007	Lixing Zhu	103	.65
Naray. Balakrishnan	186	.70	Malay Ghosh	119	.51	David Dunson	101	.64
Jeremy Taylor	159	0	Emmanuel Lesaffre	119	0	Jianqing Fan	101	.38
Joseph Ibrahim	158	0.01	Xiaohua Zhou	119	.31	Stuart Lipsitz	98	.11

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

coauthorship network of author i. According to Jin et al. (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 coauthorship diversity of author i: a large p-value suggests that his/her coauthors form a tightly knit group, and a small p-value suggests that his/her coauthors are from two or more groups and so he/she is more diverse in coauthorship.

Figure 6 presents the results for the personalized coauthorship networks of 1,000 authors who have the largest numbers of coauthors in our data set. The left panel presents the histogram for the numbers of coauthors of these 1,000 authors, and the right panel presents the histogram for the p-values of their personalized coauthorship networks. The p-values spread between 0 and 0.8, and 190 of them are smaller than 5%. Therefore, for about 80% of these 1,000 authors, their coauthors 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 coauthors. Take the first two authors, for example. They both have a large number of coauthors, 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 coauthors

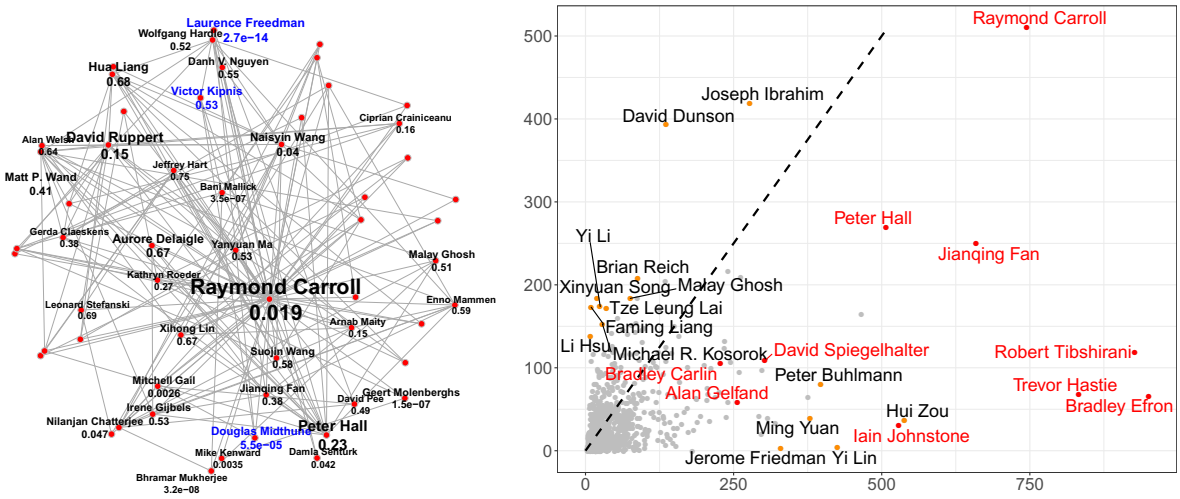

Figure 7: Left: The personalized coauthorship network of Raymond Carroll (the most collaborative author; see Table 4). Only nodes with 40 or more coauthors are shown. Different colors of names indicate two communities identified by SCORE. Similar plot can be generated for any author whose personalized network is reasonably large (≥ 50 nodes, say). Right: The pair SgnQ test statistics $(T_i^{citer}, T_i^{citee})$ on personalized citer and citee networks of 1,000 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 $(T_i^{citer} - T_i^{citee})$

are likely to form a tight-knit group, while Carroll's coauthors may come from multiple groups. To identify such groups, we perform a community detection on Carroll's personalized coauthorship network (excluding Carroll [3]) by SCORE (see Section 3.1 and Jin (2015)) and find that the research areas of a group of coauthors (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 coauthors of Carroll. This explains why Carroll's network has a relatively small p-value. See Figure 7 (left panel) for the personalized coauthorship network of Carroll, where the p-value of any node presented there is the p-value for his/her own personalized coauthorship 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

[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.

and denote the two test scores by T_i^{citer} and T_i^{citee}, respectively. Figure 7 shows the two test scores for 1,000 authors with the largest numbers of coauthors. 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 papers mainly cite theoretical papers but can be cited by many methodology and applied papers. 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 papers cite a variety of methodology papers; another reason is that many citations to papers in biostatistics are from other disciplines not covered by our data set. 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 5 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 a paper 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 papers by Donald Rubin's are in Bayesian statistics and causal inferences, but he has impacts over many other areas (e.g., GEE); see Figures 1-2. 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 paper 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., Jin et al. (2021); Yuan et al. (2021); Ke et al. (2019)). 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 data set. 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 data set, (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 papers related to COVID-19) (Foster et al., 2015). Our study can tell

which areas have far more researchers, papers, 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 papers to recommend materials for functional applications several years ahead of time (Tshitoyan et al., 2019). We can do similar things with our data set to predict emerging new areas and significant advancements. For example, in Ji et al. (2021), we combine our citation data with the paper 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 data set 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 data set, our hope is that, the data set (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 data set for new projects and new discoveries.

For future work, note that our data set 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 coauthorship data. How to combine different data resources more efficiently is an interesting problem. We leave this to the future work.

Acknowledgments. The authors thank the Associate Editor and referees for very

helpful comments. They thank 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. 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.

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

References

Amini, A. A., A. Chen, P. J. Bickel, and E. Levina (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist. 41*(4), 2097–2122.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *J. Acoust. Soc. Am. 22*, 725–730.

Donoho, D. L. (2017). 50 years of data science. *J. Comput. Graph. Stat. 26*(4), 745–766.

Efron, B. (1998). Fisher in the 21st century. *Statist. Sci. 13*(2), 95–114.

Foster, J. G., A. Rzhetsky, and J. A. Evans (2015). Tradition and innovation in scientists' research strategies. *Am. Sociol. Rev. 80*(5), 875–908.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry 40*(1), 35–41.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2010). A survey of statistical network models. *Found. Trends Mach. Learn. 2*(2), 129–233.

Hall, P. G. (2011). "Ranking our excellence" or "assessing our quality" or whatever. *Inst. Math. Statist. Bull. 40*, 12–14.

Ji, P., J. Jin, Z. T. Ke, and W. Li (2021). Journal ranking, topic modeling, and citation prediction for statistical publications. *Manuscript*.

Jin, J. (2015). Fast community detection by SCORE. *Ann. Statist. 43*(1), 57–89.

Jin, J., Z. T. Ke, and J. Liang (2021). Sharp impossibility results for hypergraph testing. *Manuscript*.

Jin, J., Z. T. Ke, and S. Luo (2017). Estimating network memberships by simplex vertex hunting. *arXiv:1708.07852*.

Jin, J., Z. T. Ke, and S. Luo (2021). Optimal adaptivity of signed-polygon statistics for network testing. *Ann. Statist. (to appear)*.

Karrer, B. and M. E. J. Newman (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E 83*, 016107.

Ke, Z. T., F. Shi, and D. Xia (2019). Community detection for hypergraph networks via regularized tensor power iteration. *arXiv:1909.06503*.

Kim, B., K. H. Lee, L. Xue, and X. Niu (2018). A review of dynamic network models with latent variables. *Stat. Surv. 12*, 105.

Li, T., L. Lei, S. Bhattacharyya, and et al. (2020). Hierarchical community detection by recursive partitioning. *J. Amer. Statist. Assoc.*, 1–18.

Liu, F., D. Choi, L. Xie, and K. Roeder (2018). Global spectral clustering in dynamic networks. *Proc. Natl. Acad. Sci. 115*(5), 927–932.

Lu, X. and B. K. Szymanski (2019). A regularized stochastic block model for the robust community detection in complex networks. *Sci. Rep. 9*(1), 1–9.

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proc. Nat. Acad. Sci. 101*(suppl 1), 5200–5205.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. 103*(23), 8577–8582.

Rosvall, M. and C. T. Bergstrom (2010). Mapping change in large networks. *PLOS ONE 5*(1), e8694.

Silverman, B. W. (2016). Introduction to discussion of "Coauthorship and citation networks for statisticians". *Ann. Appl. Stat. 10*(4), 1777–1778.

Tshitoyan, V., J. Dagdelen, L. Weston, et al. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature 571*(7763), 95–98.

Yuan, M., Y. Feng, and Z. Shang (2018). A likelihood-ratio type test for stochastic block models with bounded degrees. *arXiv:1807.04426*.

Yuan, M., R. Liu, Y. Feng, and Z. Shang (2021). Testing community structures for hypergraphs. *Ann. Statist. (to appear)*.

Zhang, Y., E. Levina, and J. Zhu (2020). Detecting overlapping communities in networks using spectral methods. *SIAM J. Math. Data Sci. 2*(2), 265–283.

Zhao, Y., E. Levina, and J. Zhu (2011). Community extraction for social networks. *Proc. Natl. Acad. Sci. 108*(18), 7321–7326.