Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

COAUTHORSHIP AND CITATION NETWORKS FOR STATISTICIANS

By Pengsheng Ji[†] and Jiashun Jin[‡]

University of Georgia[†] and Carnegie Mellon University[‡]

We have collected and cleaned two network data sets: Coauthorship and Citation networks for statisticians. The data sets are based on all research papers published in four of the top journals in statistics from 2003 to the first half of 2012. We analyze the data sets from many different perspectives, focusing on (a) productivity, patterns and trends, (b) centrality, and (c) community structures, and present an array of interesting findings.

For (a), we find that over the 10-year period, both the average number of papers per author and the fraction of self citations have been decreasing, but the proportion of distant citations has been increasing. These suggest that the statistics community has become increasingly more collaborative, competitive, and globalized, driven by the boom of online resources and Search Engines.

For (b), we identify the most prolific, the most collaborative, and the most highly cited authors. We also identify a handful of "hot" papers and suggest "Variable Selection" and several other areas as the "hot" areas in statistics.

For (c), we have investigated the data sets with 6 different methods, including D-SCORE (a new procedure we propose here). We have identified about 15 meaningful communities, including large-size communities such as "Biostatistics", "Large-Scale Multiple Testing", "Variable Selection" as well as small-size communities such as "Dimensional Reduction", "Objective Bayes", "Quantile Regression", and "Theoretical Machine Learning". We find that the community structures for the Coauthorship network and Citation network are connected and intertwined, but are also different in significant ways.

Our findings shed light on research habits, trends, and topological patterns of statisticians, and our data sets provide a fertile ground for future researches on or related to social networks of statisticians.

1. Introduction. It is frequently of interest to identify "hot" areas and key authors in a scientific community, and to understand the research habits, trends, and topological patterns of the researchers. A better understanding of such features is useful in many perspectives, ranging from that of adminis-

[‡]JJ was partially supported by NSF grant DMS-1208315.

MSC 2010 subject classifications: Primary 91C20, 62H30; secondary 62P25

Keywords and phrases: adjacent rand index, centrality, collaboration, community detection, Degree Corrected Block Model, productivity, social network, spectral clustering.

trations and funding agencies on priorities for support, to that of individual researchers on starting a new research topic or new research collaboration.

To study these features, one possible approach is to use statistical survey: to hand out questionnaires to people within or associated with a scientific community. However, such an approach is relatively expensive, and it is hard to persuade people to collaborate in the survey. Another approach is to use online resources, say, *Mathematical Genealogy*. This approach could be very useful to answer some of the questions, but the information available in such resources has been focused on some specific aspects (e.g., the focus of Mathematical Genealogy is student-advisor relationship) and is not very helpful in understanding the whole picture of the scientific community.

Coauthorship and Citation networks provide a convenient and yet appropriate approach to addressing many of these questions. On one hand, with the boom of online resources (e.g., MathSciNet) and Search Engines (e.g., Google Scholar), it is relatively convenient for us to collect the Coauthorship and Citation network data of a specific scientific community. On the other hand, these network data provide a wide variety of information (e.g., productivity, trends, impacts, and community structures) that can be extracted to understand many different aspects of the scientific community, and thus provide a more complete picture of the community.

Recent studies on such networks include but are not limited to the following: Grossman [17] studied the Coauthorship network of mathematicians; Newman [35, 37] studied the Coauthorship networks of biologists, physicists and computer scientists (see also Martin *et al.* [32], which studied networks of physicists using a much larger data set than that in [35, 37]); Ioannidis [23] used the Coauthorship network to help assess the scientific impacts.

Unfortunately, as far as we know, the Coauthorship and Citation networks for *statisticians* have not yet been studied. We recognize that

- The people who are most interested in social networks for statisticians are statisticians themselves or people with close ties to them. It is unlikely for researchers from other disciplines (e.g., physicists) to devote substantial time and efforts to pay *specific* attention to networks for statisticians: it is the statisticians' task to collect and analyze such network data about themselves and of interest to themselves.
- For many aspects of the networks, the "ground truth" is unavailable. However, as statisticians, we have the advantage of knowing (at least partially) many aspects (e.g., "hot" areas, community structures, trends and impacts of statistical researches) of our own community, and many times, such "partial ground truth" can be very helpful in analyzing the networks and interpreting the results.

Since the year of 2012, we have spent substantial efforts and time collecting two *new* network data sets: Coauthorship network and Citation network for statisticians. The data sets are based on all published papers from 2003 to the first half of 2012 in four of the top statistical journals: Annals of Statistics (AoS), Journal of American Statistical Association (JASA), Journal of Royal Statistical Society (Series B) (JRSS-B), and Biometrika.

The data sets provide a fertile ground for researches on social networks, especially to us statisticians, as we know the "partial ground truth" for many aspects of our community. For example, we can use the data sets to check and build network models, to develop new methods and theory, and to further understand the research habits, patterns, and topological structures of the networks of statisticians. Last but not least, we can use the data sets and the analysis in the paper as a starting point for a more ambitious project, where we collect network data sets of this kind but cover many more journals in or related to statistics and span a much longer time period.

1.1. *Our findings.* In this paper, we analyze the two network data sets, focusing on the following:

- *Productivity, patterns and trends.* We identify noticeable publication patterns of the statisticians, and how they evolve over time.
- *Centrality.* We identify "hot" areas as well as authors that are most collaborative or are most highly cited.
- *Community detection.* With possibly more sophisticated methods and analysis, we identify meaningful communities of statisticians.

For productivity, patterns and trends, we discuss the overall productivity, coauthor patterns and trends, and citation patterns and trends. We have many interesting findings, including but not limited to the following.

- In the 10-year period 2003-2012, the number of papers per author has been decreasing (Figure 1). Also, the proportion of self-citations has been decreasing while the proportion of distant citations has been increasing (Figure 4). These suggest that the statistics community has become increasingly more collaborative, competitive, and globalized.
- It seems that two authors who have a common coauthor are more likely to collaborate. However, the Coauthorship network data only marginally supports this, with a relatively small transitivity coefficient of .32 (usually, a coefficient in (.3, .6) is regarded as transitive [39]).
- The distribution of either the degrees of the author-paper bipartite network or the Coauthorship network has a power-law tail (Figures 2-3), a phenomenon frequently found in social networks [4, 36].

For centrality, we discuss several different centrality measures, and use them to identify "hot" authors and papers. Specifically,

- We identified Peter Hall, Jianqing Fan, and Raymond Carroll as the most prolific authors, Peter Hall, Raymond Carroll and Joseph Ibrahim as the most collaborative authors, Jianqing Fan, Hui Zou, and Peter Hall as the most highly cited authors. See Table 1.
- We identified 14 "hot" papers using several different measures of centrality. See Table 2.

Among these 14 papers, 10 are on "Variable Selection", suggesting "Variable Selection" as a "hot" area. Other "hot" areas may include "Covariance Estimation", "Empirical Bayes", and "Large-scale Multiple Testing".

For community detection, since the Coauthorship network is undirected and the Citation network is directed, we discuss them separately.

For the Coauthorship network, we discuss two versions of the network, (A) and (B). In Coauthorship network (A), in order for two authors to have an edge connecting them, we require that they have at least two collaborated papers. In Coauthorship network (B), however, we only require that they have at least one collaborated papers. While Coauthorship network (B) is defined in a more conventional way, Coauthorship network (A) is much more convenient to analyze, and presents many meaningful communities we can not find in Coauthorship network (B).

For each version of the network, we use the very recent *Degree Corrected Block Model (DCBM)* [27]. We investigate each network using four different community detection methods: Jin's SCORE [25], Newman's Spectral Clustering method (NSC) [38], Bickel and Chen's Profile Likelihood (BCPL) method [5, 47], and Armini *et al*'s Profile Likelihood (APL) method [1].

We find that Coauthorship network (A) is rather fragmented. It can be split into many disconnected components, many of which are groups with special characteristics. We present only the eight largest ones. The largest component is the "High Dimensional Data Analysis (Coauthorship (A))" (HDDA-Coau-A) community (Figure 6). This component has 236 nodes and seems to have sub-structures: namely, the "Carroll-Hall" community, the "North Carolina" community, and the "Fan" group. See Figures 7-8 where we present the community detection results for this component by four different methods (SCORE, APL, NSC, and BCPL). The next two largest components can be interpreted as communities of "Theoretical Machine Learning" (15 nodes) and "Dimension Reduction" (14 nodes) (presented side by side in Figure 9), respectively. The next 5 components can be, respectively, interpreted as communities of "Johns Hopkins", "Duke", "Stanford", "Quantile

Regression", and "Experimental Design" and are presented in Table 5. Note that all except the first component has a size ≤ 15, so there is no need for further study on sub-structures.

Similarly, for Coauthorship network (B), we restrict our attention to the giant component (2263 nodes), and apply all four community detection methods aforementioned. It seems that all these methods agree on that there are three meaningful communities as follows (arranged by size ascendingly): "Objective Bayes", "Biostatistics (Coauthorship (B))" (Biostat-Coau-B), "High Dimensional Data Analysis (Coauthorship (B))" (HDDA-Coau-B) (of course, the communities identified by different methods are also different in important ways; see Section 4.4). The three communities identified by SCORE are presented in Figures 10, 11, and 12, respectively.

Additionally, we carefully compare the results by different methods, measure how similar they are to each other, and address on their weakness and strengths. See Tables 3-4 for comparisons of these results with Coauthorship network (A), and Tables 6-7 for that with Coauthorship network (B). We also compare the communities identified in both Coauthorship network (A) and Coauthorship network (B) carefully, to see how they are connected, intertwined, and how they are different.

We now discuss the Citation network. Citation network is directed, and it remains largely unknown how to model such networks and how to do community detection. We extend the DCBM to directed networks, and propose Directed SCORE (D-SCORE) as a new community detection method.

We have applied D-SCORE to the Citation network and identified three meaningful communities: "Large-Scale Multiple Testing", "Biostatistics (Citation)" (Biostat-Cita), and "Variable Selection". These communities are presented in Figures 14, 15, and 16, respectively.

Similarly, we carefully compare these results with those for Coauthorship network (A) (e.g., Table 8) and Coauthorship network (B) (e.g., Table 9). We find that the community structures for the Citation network and Coauthorship network are connected, intertwined, but also very different: the study on one sheds additional light over that for the other, and combining two networks gives more insights over the community structures.

We have also applied Leicht and Newman's Spectral Clustering (LNSC) method [30] to the Citation network. However, the results are rather inconsistent to those by D-SCORE and are comparably less convincing.

1.2. *Data collection and cleaning.* We have faced substantial challenges in data collection and cleaning, and it has taken us more than 6 months to obtain high-quality data sets and prepare them in a ready-to-use format.

At first glance, it may be hard to understand why it is challenging to collect such data: the data seem to be everywhere, very accessible and free.

This is true to some extent. However, when it comes to high-volume high-quality downloads, the resources become surprisingly limited. For example, Google Scholar aggressively blocks any one (a person or a machine) who tries to download the data more than just a little; when you try to download little by little, you will see some portion of the data are made messy and incomplete intentionally. For other online resources, we face a similar problem.

We also face other challenges, both in data collection and in data cleaning: missing paper identifiers, ambiguous author names, and so on and so forth. In Section 7, we address all the challenges we have faced and how we have overcome them.

1.3. *Experimental design and scientific relevance.* We are primarily interested in the networks for statisticians home based in USA. For this reason, we have limited our attention to four journals (AoS, JASA, JRSS-B, Biometrika), which are regarded by many US-based statisticians the top statistical journals. We recognize that we may have different results when we include in our study either journals which are the main venues for statisticians from a different country or region, or journals which are the main venues for statisticians with a different focus (e.g., Bioinformatics).

We are also primarily interested in the time period when high dimensional data analysis emerged as a new statistical area. We may have different results if we extend the study to a much longer time period.

On the other hand, it seems that the data sets we have serve well for solving our targeted scientific problems: they provide many meaningful results in many aspects of our targeted community within the targeted time period. They also serve as a starting point for a more ambitious project in which we collect data from many more journals in a much longer time period.

1.4. *Disclaimers.* Our primary goal in the paper is to present the data sets we collect, and to report what we find with them. We wish to clarify:

- It is not our intention to rank one author or a paper over the others. We wish to clarify that "highly cited" is not exactly the same as "important" or "influential", and "not highly cited" is not exactly the same as "unimportant" or "not influential".
- It is not our intention to rank one area over the other. A "hot" area is not exactly the same as an "important" area or an area that needs the most of our time and efforts. It is also not exactly the area that is

so-much over-studied or exhausted that we should not dive in either.
- It is not our intention to label an author/paper/topic with a certain community/group/area. A community or a research group may contain many authors, and can be hard to interpret. For presentation, we need to assign names to such communities/groups/areas, but the names do not always accurately reflect all the authors/papers in them.

Also, social networks are about "real people", and this time, "us". In order to obtain meaningful and interpretable results, we have to use real names. We have not used any data beyond those which are publicly available. The interest of the paper is on the statistics community *as a whole*, not on any individual statisticians.

1.5. *Contents.* The paper is organized as follows. Section 2 discusses the productivity, patterns and trends, and Section 3 discusses the centrality. In Sections 4-5, we discuss community detection for the Coauthorship network and Citation network, respectively. Section 6 contains a brief summary and discusses the limitations of the paper and suggests some future directions. Section 7 addresses the challenges in data collection and cleaning.

2. Productivity, patterns and trends. In this section, we report our findings, focusing on three interconnected aspects: productivity, coauthor patterns and trends, citation patterns and trends.

2.1. *Productivity.* Overall, there are 3248 papers and 3607 authors in the data set, suggesting an average of 0.90 paper per author. It is of interest to investigate how the productivity evolves over the years. In Figure 1, we present the total number of papers published in each year (left panel) and the average number of papers per author in each year (right panel), i.e., the ratio of the total number of papers published that year over the total number of authors who published at least once that year (it seems the result is inconsistent to that of an overall mean of .90, but this is due to that authors in different years largely overlap with each other). It is interesting to note that over the 10-year period, the number of papers published each year has been increasing, but the average number of papers per author has been decreasing (drop about 18% in ten years). Possible explanations include:

- *More collaborative.* Collaboration between authors has been increasing.
- *More competitive.* Statistics has become a more competitive area, and there are more people who enter the area than who leave the area. Also, it becomes increasingly more difficult to publish in these 4 journals (which are viewed by many as top journals in statistics).

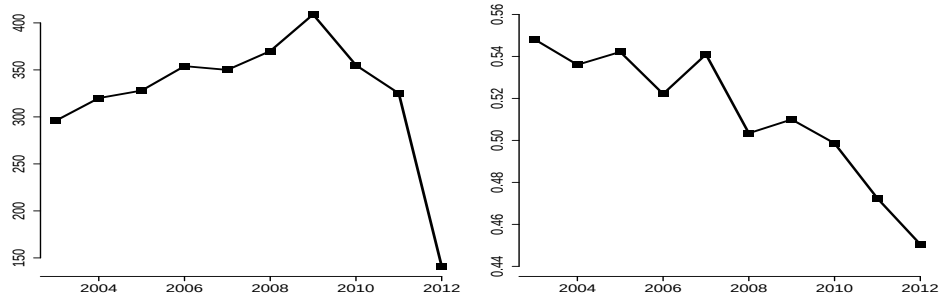

FIG 1. *Left: total number of papers published each year from* 2002 *to* 2012 *(for the year 2012, we have only data for the first half). Right: the ratios between the number of papers published in each year and the number of authors who has published in the same year.*

We also present the distribution of the numbers of papers per author. For any K-author paper, $K \geq 1$, we have two different ways to count each coauthor's contribution to this particular paper, either divided or non-divided.

- *Non-divided.* We count every coauthor as has published one paper.
- *Divided.* We count every coauthor as has published $1/K$ paper.

Both approaches have their virtues and disadvantages. The first way may cause substantial "inflation" in counting, and the second way may be insufficient, especially since for many papers, there are one or more "leading authors" who contribute most of the work.

Following the first approach, we have the left panel of Figure 2, where the x-axis is the number of papers, and the y-axis is the proportion of authors who have written more than a certain number of papers. Approximately, the curve looks like a straight line, especially to the right tail. This suggests that the distribution of the number of papers has a power law tail.

Following the second approach, we present the Lorenz curve [39] of the number of papers by each author (where for a K-author paper, each author is counted as having $1/K$ paper) in the right panel of Figure 2, which suggests the distribution does not have a power law tail but is still very skewed. The figure shows that the top 10% most prolific authors contribute 41% of the papers, and the top 20% most prolific authors contribute 58% of the papers. Our findings are similar to that in [32] for the physics community.

The Gini coefficient [15] is a well-known measure of dispersion for a distribution. For our data set, the Gini coefficient for the distribution of the number of papers by different authors is 0.51, which is much smaller than the Gini coefficient of 0.70 for that associated with the physics community [32]. This seems to suggest that the published papers are more evenly distributed

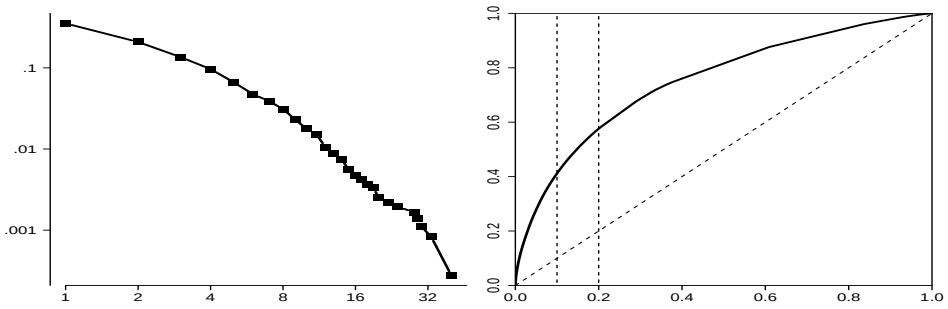

FIG 2. *Left: The proportion of authors who have written more than a certain number of papers (for a better view, both axes are evenly spaced on the logarithmic scale). Right: The Lorenz curve for the number of papers each author with divided contributions.*

among authors in the statistics community than the physics community. Another possible explanation is that the data set in [32] is based on all published papers in physics spanning more than 100 years, while our data set is based on four journals in statistics for a 10-year period. It is expected that in the latter, the distribution of the number of papers by different authors (with divided contributions) is less dispersed. It is interesting to note that the Gini coefficient of the income inequality for the USA in the year of 2011 is 0.48, which is slightly smaller than 0.51.

2.2. *Coauthor patterns and trends.* In the coauthorship network, the degree of a node is also the number of coauthors for the node. The degrees range from 0 to 65, where Peter Hall (65), Raymond Caroll (55), Joseph Ibrahim (41), Jianqing Fan (38) and David Dunson (32) are the ones with the highest degrees (and so they are the most collaborative authors). Also, 154 authors have degree 0, and 913 authors have degree 1. The degree distribution is shown in Figure 3 (left panel), suggesting a power law tail.

It is of interest to investigate how the number of coauthors changes over time. In Figure 3 (right panel), we present the average number of coauthors in each of the 10 years (for each year, we consider only the authors who published in these journals). It is seen that overall the average number of coauthors is steadily increasing. Again, this suggests that the statistics community has become increasingly more collaborative.

Many social network are transitive (e.g., a friend of a friend is likely to be a friend) [45]. For the coauthorship network based our data sets, the transitivity is 0.32, compared to 0.066 for the biology community, 0.15 for the mathematics community, and 0.43 for the physics community [37]. For real-world social networks, the usual range of transitivity is between 0.3 and

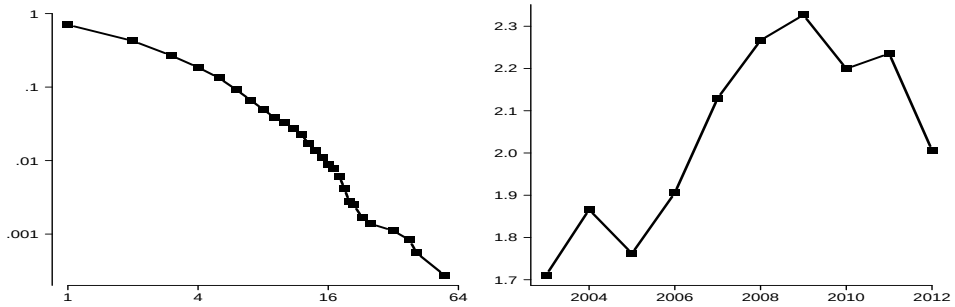

FIG 3. *Left: The proportion of authors with more than a given number of coauthors (for a better view, both axes are evenly spaced on the logarithmic scale). Right: The average number of coauthors for all authors who has published in these journals that year.*

0.6 [39], suggesting that the Coauthorship network is moderately transitive.

2.3. *Citation patterns and trends.* For the 3248 papers (3607 authors) in our data sets, the average citation per paper is 1.76, which is significantly lower than the Impact Factor (IF) of these journals. Based on ISI 2010, the IFs for AoS, JRSS-B, JASA, and Biometrika are 3.84, 3.73, 3.22, and 1.94, respectively. This is largely due to that we count only the citations between papers in these 4 journals in a 10-year period. Among these papers, (a) 1693 (52%) are not cited by any other paper in the data set, (b) 1450 (45%) do not cite any other paper in the data set, and (c) 778 (24%) neither cite nor are cited by any other papers in the data sets.

The distribution of the in-degree (the number of citations received by each paper) is highly skewed. The top 10% highly cited papers receive about 60% of all citation counts, while the top 20% receive about 80% of all citation counts. The Gini coefficient is 0.77 [15] suggesting that the in-degree is highly dispersed. The Lorenz curve [39] is shown in Figure 4 (left panel), confirming that the distribution of the in-degrees is highly skewed.

We also observe some very interesting patterns. First, the authors return a favor of citation, especially if it is from a coauthor. The proportion of (either earlier or later) reciprocation among coauthor citations is 79%, while that among distant citations is 25%.

In Figure 4 (right panel), we show that over the 10-year period, (a) the proportion of self-citations has been slowly decreasing, (b) the proportion of citations from a coauthor remains roughly the same, and (c) the proportion of distant citations (citations that are not from oneself or a coauthor) has been slowly increasing. The last item is a little unexpected, but it probably makes sense in that over the years, the publications have become increas-

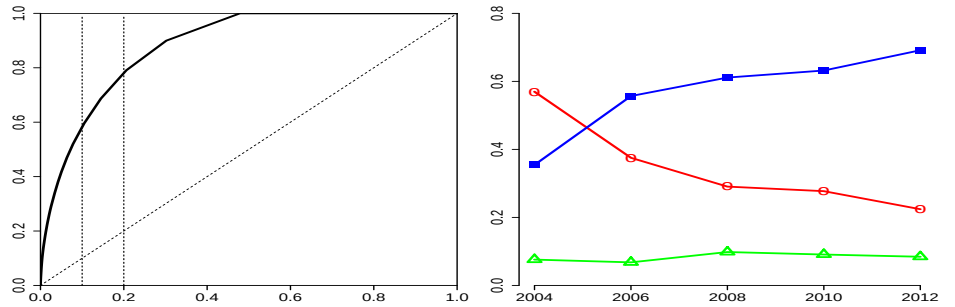

FIG 4. *Left: The Lorenz curve for the number of citation received by each paper. Right: The proportions of self-citations (red circles), coauthor citations (green triangles) and distant citations (blue rectangles) for each two-year block.*

ingly more accessible online and communications have become increasingly easier and more efficient. That the blue curve and the red cross crossover with each other on the left is probably due to the "boundary effect": for papers published in 2003 (say), most the papers they have cited are probably published earlier than 2002, which are not included in our data sets. Below, we show that the mean delay of citation is about 3 years. For this reason, the "boundary effect" is probably negligible in the later half of the time period. Note that the overall proportions for self-citations, coauthor citations and distant citations are 27%, 9%, and 64%, respectively.

The data set also confirms a reasonable delay in citations, despite the fact that most papers appear online (such as personal website, arXiv, department archives) much earlier than the time when the paper is published. The overall mean delay (e.g., the average difference between the years of the publication of a new paper and the papers it cites) is 3.30 years, and the mean delay for self-citations, coauthor citations, and distant citations, are 2.81, 3.36 and 3.51 years, respectively, suggesting the authors cite their own or their coauthors' work more quickly than that of others.

3. Centrality. It is frequently of interest to identify the most "important" authors or papers, and one possible approach is to use centrality. There are many different measures of centrality. In this section, we use the degree centrality, the closeness centrality, and the betweenness centrality.

The degree centrality is conceptually simple, but the definition varies with the types of networks. For the author-paper bipartite network, the centrality of an author is the number of papers he/she publishes. For Coauthorship network, the centrality of an author is the number of his/her coauthors. For Citation network of *authors*, we are primarily interested in the in-degree,

and the centrality of an author is the number of citers (i.e., authors who cite his or her papers). For Citation network of *papers*, the centrality is the in-degree (i.e., the number of papers which cite this paper).

The closeness centrality is defined as the reciprocal of the total distance to all others (Sabidussi [41]), $C_{clo}(v) = 1/(\Sigma_{u\in V}\text{dist}(v,u))$, where V is the set of all nodes and $\text{dist}(v,u)$ is the distance between nodes v and u.

The betweenness centrality measures the extent to which a node is located "between" other pairs of nodes. The most commonly used definition (e.g., Freeman *et al.* [14]) is $C_{bet}(v) = \Sigma_{s\neq t\neq v\in V}\frac{\sigma(s,t|v)}{\Sigma_v\sigma(s,t|v)}$, where $\sigma(s,t|v)$ is the total number of shortest paths between s and t that pass through v.

Table 1 presents the key authors identified by different measures of centrality. The results suggest that different measures of centrality are largely consistent with each other, which identify Raymond Carroll, Jianqing Fan, and Peter Hall (alphabetically) as the "top 3" authors.

TABLE 1

Top 3 authors identified by the degree centrality (Columns 1-3; corresponding networks are the author-paper bipartite network, Coauthorship network, and Citation network for authors), the closeness centrality and the betweenness centrality.

# of papers	# of coauthors	# of citers	Closeness	Betweenness
Peter Hall	Peter Hall	Jianqing Fan	Raymond Carroll	Raymond Carroll
Jianqing Fan	Raymond Carroll	Hui Zou	Peter Hall	Peter Hall
Raymond Carroll	Joseph Ibrahim	Peter Hall	Jianqing Fan	Jianqing Fan

Table 2 presents the "hot" papers identified by 3 different measures of centrality. For all these measures, the "hottest" papers seem to be in the area of variable selection. In particular, the top 3 most cited paper are Zou [48] (75 citations; adaptive lasso), Meinshansen and Buhlmann [34] (64 citations; graphical lasso), and Candès and Tao [8] (49 citations; Dantzig Selector). The three papers are all in a specific sub-area in high dimensional variable selection, where the theme is to extend the well-known penalization methods of the lasso [9, 43] in various directions (these fit well with the impression of many statisticians: in the past 10-20 years, there is a noticeable wave of research papers devoted to the penalization methods).

These results suggest "Variable Selection" as one of the "hot" areas. Other "hot" areas may include "Covariance Estimation", "Empirical Bayes", and "Large-Scale Multiple Testing"; see Table 2 for details.

For more information, note that at `www.stat.uga.edu/~psji/`, we have listed the 30 most cited papers in the file top-cited.xlsx. These 30 papers account for 16% of the total number of citation counts. The list furthers shows that the most highly cited papers are on the regularization methods (e.g., adaptive lasso, group lasso, etc.).

On the other hand, we must note that some important and innovative works in the particular area of variable selection have significantly fewer citations. This includes but is not limited to the phenomenal paper by Efron *et al.* (2004) [10] on least angle regression, which has received a lot of attention from a broader scientific community (4365 citations on Google Scholar), but was cited only 11 times by papers in our data set (in comparison, the adaptive lasso paper [48] has received 75 citations). A similar claim can be drawn on other areas or topics.

TABLE 2

Fourteen "hot" papers (alphabetically) identified by degree centrality (Column 2; for citation networks of papers), closeness centrality, and betweenness centrality. Numbers in Column 2-4 *are the ranks (only shown when the rank is smaller than* 5*).*

Paper (Area)	Citations	Closeness	Betweenness
Bickel & Levina (2008) [6] (Covariance Estimation)			4
Candes & Tao (2007) [8] (Variable Selection)	3		
Fan & Li (2004) [11] (Variable Selection)		2	
Fan & Lv (2008) [12] (Variable Selection)			1
Fan & Peng (2004) [13] (Variable Selection)	4	1	
Huang et al (2006) [20] (Covariance Estimation)			3
Huang et al (2008) [19] (Variable Selection)			5
Hunter & Li (2005) [22] (Variable Selection)		4	
Johnstone & Silverman (2005) [26] (Empirical Bayes)		5	
Meinshausen & Buhlmann (2006) [34] (Variable Selection)	2		
Storey (2003) [42] (Multiple Testing)		3	
Zou (2006) [48] (Variable Selection)	1		
Zou & Hastie (2005) [49] (Variable Selection)	5		
Zou & Li (2008) [50] (Variable Selection)			2

That statisticians have been very much focused on a very specific research topic and a very specific approach is an interesting phenomenon that deserves more explanation by itself.

The centrality measures we use here are either natural choices or existing measures. We are merely reporting what the data sets tell us, with no intention to rank one author or an area over the others; see Section 1.4.

4. Community detection for Coauthorship networks. In this section, we discuss community detection for Coauthorship networks. We investigate two Coauthorship networks, (A) and (B), to be introduced shortly. We first discuss models and methods in Sections 4.1-4.2, and then apply the methods to Coauthorship networks (A) and (B) and report the results in Sections 4.3-4.4. In Section 4.5, we briefly comment on community extraction (a problem that is closely related but is also very different).

There are many different ways to define the Coauthorship network, and in this paper, we use the following definition. Let n be the number of authors.

The Coauthorship network is the undirected network $\mathcal{N} = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of nodes and E is the set of edges. Fixing an integer $t \geq 1$, we assume

(4.1) nodes i and j have an edge $\Longleftrightarrow$ they have coauthored $\geq t$ papers.

In this paper, we focus our study on the following two networks.

- *Coauthorship network (A).* In this network, we choose $t = 2$.
- *Coauthorship network (B).* In this network, we choose $t = 1$.

While Coauthorship network (B) is the most natural choice, Coauthorship network (A) is also of interest: it is not only easier to analyze but also has some very different structures. Our study on Coauthorship (A) identifies an array of meaningful communities that are hard to find by using Coauthorship (B). See Section 4.3 for details.

In principle, networks with $t \geq 3$ could also be of interest. However, such networks are very much fragmented, and provide limited additional insight to those of $t = 1, 2$. For reasons of space, we skip discussions along this line.

Community detection is one of the problems that of major interest in studies on social networks [16, 29]. Consider an *undirected* and *connected* network $\mathcal{N} = (V, E)$. We think V as the union of a few (disjoint) subsets which we call the "communities":

$$V = V^{(1)} \cup V^{(2)} \ldots \cup V^{(K)},$$

where "$\cup$" stands for the union of sets and has nothing to do with networks (same below). A community can be thought of as a subset of nodes where there are more edges 'within' than 'across' different communities; see for example [7]. The goal of community detection is for each node $i \in V$, to decide to which community it belongs.

Note that for simplicity, we assume each pair of communities are disjoint with each other in this paper. Also, in practice, a given network might not always be connected. For the purpose of community detection, we can always first split the network into different disconnected components, and then apply community detection to each component separately.

4.1. *Degree Corrected Block Model for undirected networks.* For an undirected network $\mathcal{N} = (V, E)$ where $V = \{1, 2, \ldots, n\}$ as before, let A be the associated adjacency matrix (symmetric since $\mathcal{N}$ is undirected):

$$A(i, j) = \begin{cases} 1, & \text{if there is an edge between nodes } i \text{ and } j, \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

We view the upper triangular of A as independent (but possibly not-identically distributed) Bernoulli random variables, and decompose A as the sum of a 'signal' component and a 'noise' component:

$$A = E[A] + W, \qquad W \equiv (A - E[A]);$$

$E[A]$ is the matrix satisfying $(E[A])(i,j) = P(A(i,j) = 1)$, $1 \le i \le j \le n$.

By default, we think that there is no edge between a node and itself, so the diagonals of A are all 0s. In light of this, for a symmetric matrix Ω to be determined, we can further write

$$A = \Omega - \mathrm{diag}(\Omega) + (A - E[A]). \tag{4.3}$$

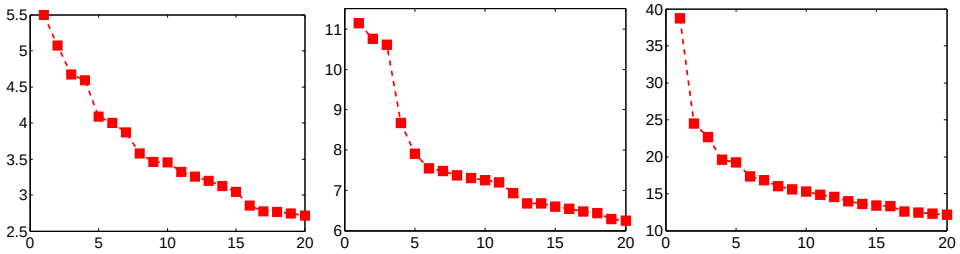

FIG 5. *Scree plots. From left to right: the giant component of Coauthorship network(A), Coauthorship network(B), Citation network (in the last one, we display singular values instead of eigenvalues).*

Degree Corrected Block Model (DCBM) is a model proposed by Karrer and Newman [27] for undirected networks, and has become popular recently. In this model, we have n positive parameters for degree heterogeneities, $\theta(1), \theta(2), \ldots, \theta(n)$, and a $K \times K$ matrix P such that

$$\Omega = \Theta L \Theta, \qquad \text{where} \qquad L = \sum_{k=1}^{K}\sum_{\ell=1}^{K} P(k,\ell)\mathbf{1}_k\mathbf{1}_\ell'. \tag{4.4}$$

Here, Θ is the $n \times n$ diagonal matrix with $\theta(i)$ being the i-th diagonal entry, and $\mathbf{1}_k$ is the $n \times 1$ indicator vector of k-th community satisfying $\mathbf{1}_k(i) = 1$ if $i \in V^{(k)}$ and $\mathbf{1}_k(i) = 0$ otherwise, $1 \le k \le K$. For analysis, we usually need some mild regularity conditions on P; see [25] and also sections below.

4.2. *Community detection methods for undirected networks.* There are many approaches to community detection for undirected networks, including but not limited to Jin's SCORE [25], Newman and Girvan's Modularity

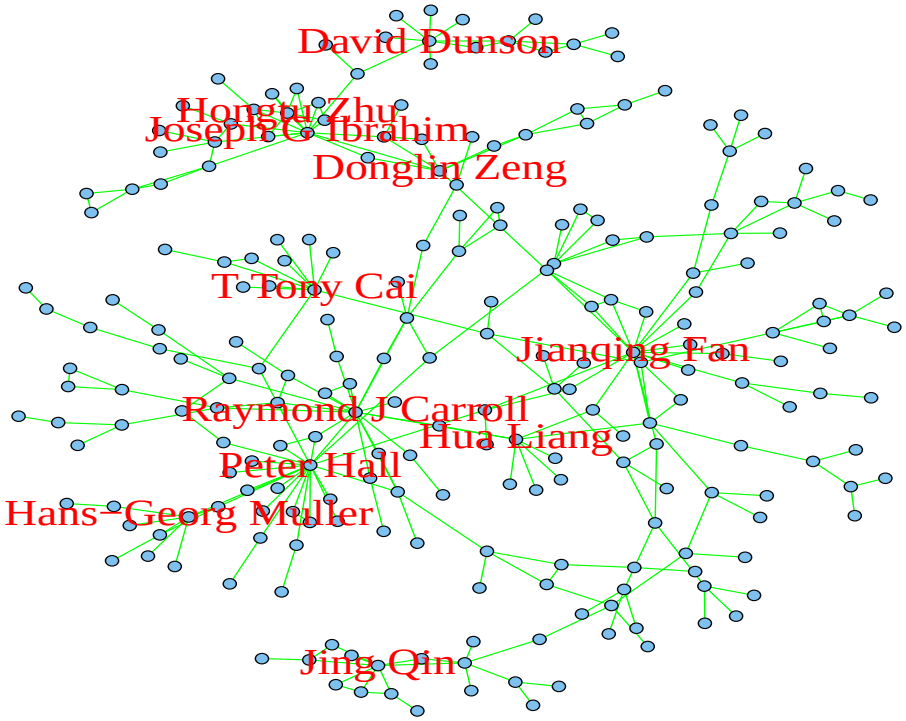

FIG 6. *The giant component of Coauthorship network (A). It could be interpreted as the "High Dimensional Data Analysis (Coauthorship (A))" (HDDA-Coau-A) community. Names are only shown for* 11 *nodes with a degree of* 8 *or larger.*

approach (NGM) [40], Newman's Spectral Clustering approach (NSC) [38], Bickel and Chen's Profile Likelihood approach (BCPL) [7, 47], and Armini *et al.*'s Pseudo Likelihood approach (APL) [1].

In this paper, we only investigate SCORE, NSC, BCPL, and APL, and do not include NGM for comparisons: on one hand, NSC is closely related to the NGM and relies on an approximation of the Newman and Girvan's modularity for inference; on the other hand, NSC is computationally more efficient than NGM, especially when the size of the network is large.

NSC is a spectral method, where the key observation that Newman and Girvan's modularity matrix can be approximated by the leading eigenvectors of the matrix [38]. Newman introduced NSC as a general idea for spectral clustering, and there are several different ways for implementations. Following [38], we cluster by using the signs of the first leading eigenvectors when $K = 2$, and by using the recursive bisections approach when $K = 3$.

TABLE 3

The Adjusted Random Index (ARI) and Variation of Information (VI) for the vectors of predicted community labels by four different methods for the giant component of Coauthorship (A), assuming $K = 2$. A large ARI/small VI suggests that the two predicted label vectors are similar to each other.

	SCORE	NSC	BCPL	APL
SCORE	1.00/.00	-.04/.95	.09/1.05	.72/.33
NSC		1.00/.00	.21/1.06	-.06/.91
BCPL			1.00/.00	.09/.87
APL				1.00/.00

TABLE 4

Comparison of community sizes by different methods assuming $K = 2$ for the giant component of Coauthorship network (A).

	North Carolina	Carroll-Hall
SCORE	45	191
NSC	155	81
APL	31	205
SCORE ∩ NSC	45	81
SCORE ∩ APL	31	191
NSC ∩ APL	31	81
SCORE ∩ NSC ∩ APL	31	81

BCPL is a penalization method proposed by Bickel and Chen [7] which uses greedy search to maximize the profile likelihood and works well for networks with thousands of nodes. When the network size is large, BCPL may be computationally slow. In light of this, Amini *et al.* [1] propose a different Profile Likelihood approach which aims to improve the speed of BCPL. By doing so, the price it pays is to ignore some dependence structures of the data so as to simplify the likelihood and make it more tractable.

SCORE, or **S**pectral **C**lustering **O**n **R**atios of **E**igenvectors, is a recent spectral method proposed by Jin [25]. Assuming K (number of communities) as known, SCORE consists of the following simple steps.

- Obtain the K leading (unit-norm) eigenvectors of A, say, $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$.
- Obtain the $n \times (K-1)$ matrix $\hat{R}$ of *entry-wise ratios* by $\hat{R}(i,k) = \hat{\xi}_{k+1}(i)/\hat{\xi}_1(i)$, $1 \le i \le n$, $1 \le k \le K-1$.
- Cluster by applying the classical k-means to $\hat{R}$, assuming there are $\le K$ communities.

At the heart of SCORE is the observation that under DCBM, the degree heterogeneity parameters $\theta(i)$'s are nearly ancillary, and can be conveniently removed by taking entry-wise ratios between $\hat{\xi}_k$ and $\hat{\xi}_1$, $k = 2, \ldots, K$. In detail, let $\xi_1, \xi_2, \ldots, \xi_K$ be the K leading (unit-norm) eigenvectors of Ω.

Denote $\hat{\Xi}$ and Ξ by the two $n \times K$ matrices

$$\hat{\Xi} = [\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K], \qquad \text{and} \qquad \Xi = [\xi_1, \xi_2, \ldots, \xi_K],$$

and let R be the non-stochastic counterpart of $\hat{R}$ given by

$$R(i,k) = \xi_{k+1}(i)/\xi_1(i), \qquad 1 \le k \le K-1, 1 \le i \le n.$$

Recall that $\mathbf{1}_k$ denotes the indicator vector for community k, $1 \le k \le K$. Write $\theta = \sum_{k=1}^{K}(\Theta \mathbf{1}_k)$ and let D be the $K \times K$ diagonal matrix such that $D(k,k) = \|\Theta \mathbf{1}_k\|/\|\theta\|$, $1 \le k \le K$, where $\|\cdot\|$ denotes the ℓ^2-norm. Let T be the $n \times K$ matrix given by $\Xi = \Theta T$. We have the following lemma, whose proof is elementary so we omit it.

LEMMA 4.1. *Assume that (4.4) holds, that P is symmetric, non-singular, non-negative and irreducible, and that all K eigenvalues of DPD are distinct. Consider either the $n \times (K-1)$ matrix R or the $n \times K$ matrix T. The matrix has exactly K distinct rows, according to which all n rows of the matrix partition into K different groups, where the partition coincides with the partition of all nodes into K different communities.*

As a result, (a) the heterogeneity parameters $\theta(i)$ are nearly ancillary and can be removed by taking entry-wise ratios between the leading eigenvectors of Ω, (b) under DCBM and some regularity conditions, we expect that $A = \Omega - \mathrm{diag}(\Omega) + W \approx \Omega$, and so (up to a ± 1 sign for each column of Ξ or R) $\hat{\Xi} \approx \Xi$, and $\hat{R} \approx R$. Applying k-means to $\hat{R}$ is then a reasonable approach to community detection. This is carefully justified in [25]; see details therein.

While this is for undirected networks, the above idea continues to be valid for directed networks, with some careful adaptions. In Section 5, we extend SCORE to directed-SCORE (D-SCORE) as an approach to community detection for directed networks, and use it to analyze the Citation network.

In Sections 4.3-4.4, we apply SCORE, NSC, BCPL, and APL to Coauthorship network (A) and (B), and report the results.

Remark 1. Note that the vectors of predicted labels by different methods could be very different. For a pair of the predicted label vectors, we measure the similarity by the Adjusted Rand Index (ARI) [21] and the Variation of Information (VI) [33]; a large ARI or a small VI suggests that two predicted label vectors are similar to each other.

4.3. *Coauthorship network (A).* Coauthorship network (A) has a total of 3607 nodes. Partially due to that we choose $t = 2$ in the definition of the network (i.e., (4.1)), the network is very much fragmented: it consists of

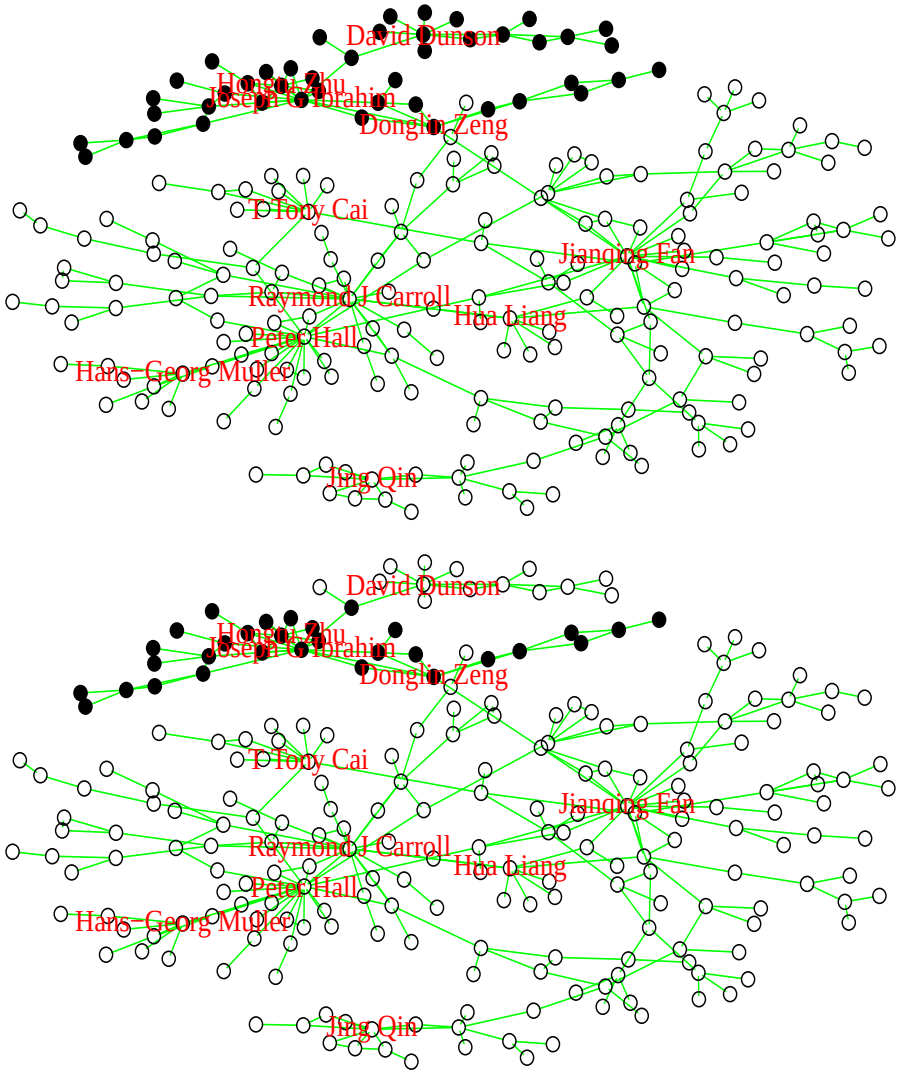

FIG 7. *Community detection results by SCORE (top) and APL (bottom) for the giant component of Coauthorship network (A), assuming $K = 2$. Nodes in black (solid) dots and white circles represent two different communities.*

2985 different components, 2805 (94%) of them are singletons, 105 (3.5%) of them are pairs, and the average component size is 1.2.

The giant component has 236 nodes, which can be roughly interpreted as the "High Dimensional Data Analysis (Coauthorship (A))" group (HDDA-Coau-A). We present this group in Figure 6 where we only show the name

of an author if the degree is 8 or larger.

Given that the size of the giant component is relatively large, it is of interest to see if it consists of sub-structures (i.e., communities). In the left panel of Figure 5, we plot the scree-plot of this group. The elbow point of the scree-plot maybe at the $3rd$, $5th$, or 8th largest eigenvalue, suggesting that there may be 2, 4, or 7 communities. In light of this, for each K with $2 \leq K \leq 7$, we run SCORE, NSC, BCPL and APL and record the corresponding vectors of predicted labels. We find that (see Remark 1 in Section 4.2 for discussions on ARI and VI):

- For $K \geq 3$, the results by different methods are largely inconsistent with each other: the maximum of ARI and the minimum VI across different pairs of methods are 0.15 and 1.19, respectively.
- For $K = 2$, we present the ARI and VI for each pair of the methods in Table 3. The table suggests that: the 4 methods split into two groups where SCORE and APL are in the same group with an ARI of 0.72, and NSC and BCPL are in the other group with an ARI of 0.21.

Note that results for methods in each groups are moderately consistent to each other, but those for methods in different groups are rather inconsistent. See Table 4 for more comparisons.

The case of $K = 2$ is worthy of further investigation. In Figures 7-8, we present the community detection results by each of the four methods. In each panel, nodes are marked with either black dots or white circles, representing two different communities. It seems that all four methods agree that there are two communities as follows.

- "North Carolina" community. This includes a group of researchers from Duke Univ., Univ. of North Carolina, North Carolina State Univ..
- "Carroll-Hall" community. This includes a group of researchers in non-parametric and semi-parametric statistics, functional estimation, and high dimensional data analysis.

It seems that the four methods split into two groups according to clustering results: SCORE and APL in one, and NSC and BCPL in the other. Methods in two groups have very different results, especially with the Fan's group and Dunson's group (we think the latter as a branch of the North Carolina community): SCORE and APL cluster the Fan's group into the "Carroll-Hall" community, and NSC and BCPL cluster both the "Fan" group and the "Dunson" group into the"North Carolina" community. See Figures 7-8.

Why methods in two groups do not agree with each on the Fan's group? A possible reason is that Fan's group has strong ties to both the "North Carolina" community and the "Carroll-Hall" community. This may also suggest

there are 3 communities (instead of 2) in this component. However, as mentioned before, when we assume $K = 3$, the results by all four methods are rather inconsistent with each other. How to obtain a more convincing explanation is an interesting but challenging problem. We omit further discussions along this line for reasons of space.

At the same time, for methods in the same group, despite their similarity, there are also some major differences. In detail,

- SCORE and APL differ on the "Dunson" branch. SCORE includes the "Dunson" branch in the "North Carolina" group, but APL excludes it from the group, and clusters them into the "Carroll-Hall" group to which they are not directly connected. In this regard, it seems that results by SCORE are more meaningful.
- NSC and BCPL differ on several small branches, including the aforementioned "Dunson" branch and two small branches connecting to the hub node marked as Jianqing Fan. In comparison, the results by NSC seem more meaningful.

For the second point, the possible reason is that BCPL uses a random start as originally proposed [7, 47]. One could use BCPL with a different start, say, the predicted labels by SCORE. However, this leads to very similar results to that of using SCORE alone. The results of BCPL highly depend on the starting label vectors it uses, and how to find the best starting vector remains an open problem; we leave the discussions to the future.

We now move away from the giant component. The next two largest components seem to the "Theoretical Machine Learning" community (15 nodes) and the "Dimension Reduction" community (14 nodes), presented in Figure 9. The first community is a small group of researchers (including Peter Buhlmann, Alexandre Tsybakov, Jon Wellner, Bin Yu) who work on Machine Learning topics using sophisticated statistical theory. The second community is a tight research group working on dimension reduction, including Francesca Chiaromonet, Denis Cook, Bing Li and their collaborators.

A conversation with Qunhua Li [31] helps to illuminate why these communities are meaningful and how they evolve over time. In the first community, Marloes H. Maathuis obtained her Ph.D from University of Washington (jointly supervised by Jon Weller and Piet Groeneboom) in 2006 and then went on to work in ETH, Switzerland, and she is possibly the "bridge" connecting the Seattle group and the ETH group (Markus Kalische, Peter Buhlmann, Markus Kalische, Sara van de Geer). Nocolai Meinshausen could be one of the "bridge" nodes between ETH and Berkeley: he was a Ph.D student of Peter Buhlmann and then a post-doctor at Berkeley. In the

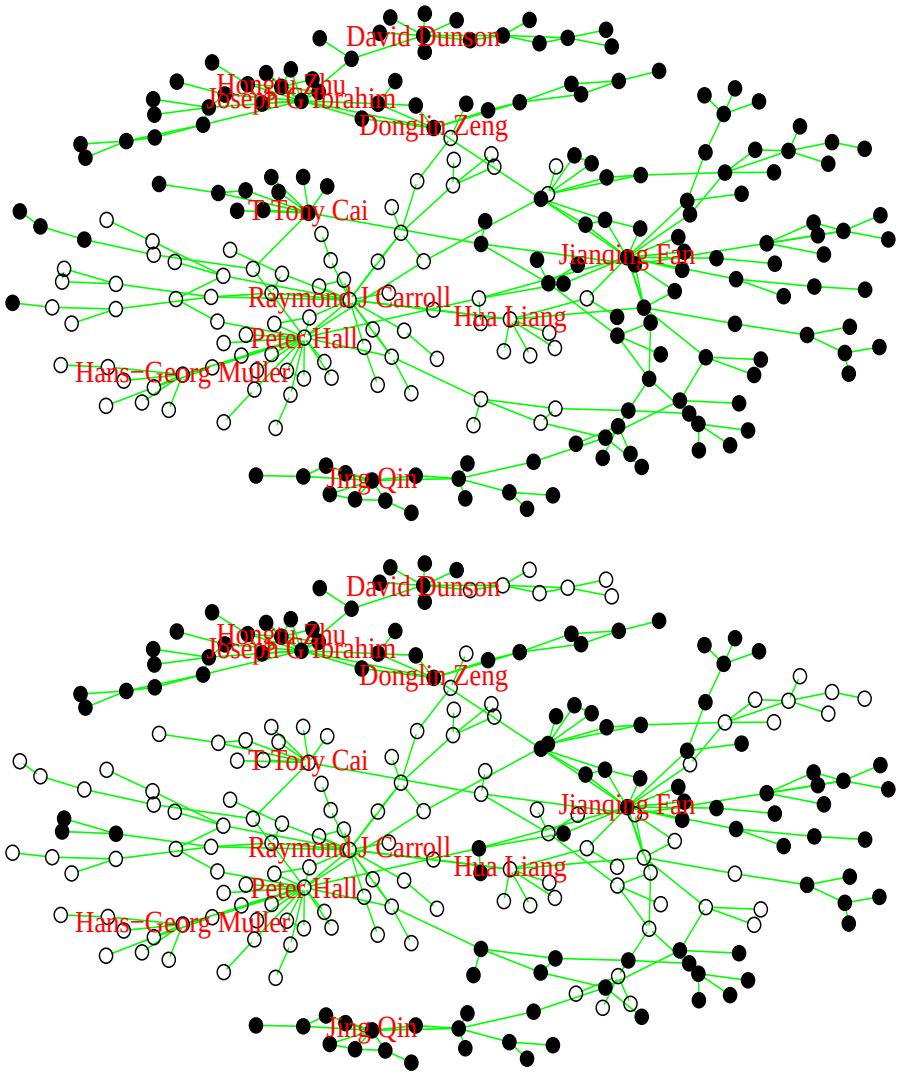

FIG 8. *Community detection results by NSC (top) and BCPL (bottom) for the giant component of Coauthorship network (A), assuming $K = 2$. Nodes in black (solid) dots and white circles represent two different communities.*

second community, Ms. Chiaromonet obtained her Ph.D from University of Minnesota, where Denis Cook served as the supervisor. She then went on to work in Statistics at Pennsylvania State University, and started to collaborate with Bing Li there for researches on dimension reduction [31].

In the Coauthorship network (A), further down the list are the "Johns

FIG 9. *The second largest (left) and third largest (right) components of Coauthorship network (A). They can be possibly interpreted as the "Theoretical Machine Learning" and "Dimension Reduction" communities, respectively.*

Hopkins research group" (13 nodes; including faculty at Johns Hopkins University and their collaborators; similar below), "Duke research group" (10 nodes; including Mike West, Jonathan Stroud, Carlos Caravlaho, etc.), "Stanford research group" (9 nodes including David Siegmund, John Storey, Ryan Tibshirani, and Nancy Zhang, etc.), "Quantile Regression group" (9 nodes; including Xuming He and his collaborators), and "Experimental Design group" (8 nodes). These communities are presented in Table 5.

4.4. *Coauthorship network (B).* In Coauthorship network (B), there is an edge between nodes i and j if and only if they have coauthored 1 or more papers. Compared to Coauthorship network (A), this definition is more conventional, but it also makes the network harder to analyze.

Coauthorship network (B) has a total of 3607 nodes, where the giant component consists of 2263 (63% of all nodes). For analysis in this section, we focus on the giant component. Also, for simplicity, we call the giant component the Coauthorship network (B) whenever there is no confusion.

We are primarily interested in community detection. Figure 5 (middle panel) presents the scree plot associated with Coauthorship network (B), suggesting 3 or more communities. We apply all four methods: SCORE, NSC, BCPL, and APL assuming $K = 3$ and below are the findings.

First, in Table 6, we compare all 4 methods pair-wise and tabulate the corresponding ARI and VI (see Remark 1). Somewhat surprisingly, the results

TABLE 5

Top: the 4-th, 5-th, and 6-th largest components of Coauthorship network (A) which can be interpreted as the groups of "Johns Hopkins", "Duke", and "Stanford"). Bottom: the 7-th and 8-th largest components of Coauthorship network (A) which can be interpreted as the groups of "Quantile Regression" and "Experimental Design".

Barry Rowlingson	Carlos M Carvalho	Armin Schwartzman
Brian S Caffo	Gary L Rosner	Benjamin Yakir
Chong-Zhi Di	Gerard Letac	David Siegmund
Ciprian M Crainiceanu	Helene Massam	F Gosselin
David Ruppert	James G Scott	John D Storey
Dobrin Marchev	Jonathan R Stroud	Jonathan E Taylor
Galin L Jones	Maria De Iorio	Keith J Worsley
James P Hobert	Mike West	Nancy Ruonan Zhang
John P Buonaccorsi	Nicholas G Polson	Ryan J Tibshirani
John Staudenmayer	Peter Muller	
Naresh M Punjabi		
Peter J Diggle		
Sheng Luo		

Hengjian Cui	Andrey Pepelyshev
Huixia Judy Wang	Frank Bretz
Jianhua Hu	Holger Dette
Jianhui Zhou	Natalie Neumeyer
Valen E Johnson	Stanislav Volgushev
Wing K Fung	Stefanie Biedermann
Xuming He	Tim Holland-Letz
Yijun Zuo	Viatcheslav B Melas
Zhongyi Zhu	

of BCPL are inconsistent with those by all other methods. For example, the maximum ARI between BCPL and each of the other three methods is .00, and the smallest VI between BCPL and each of the other three methods is 1.29, showing a substantial disagreement. This is possibly due to that BCPL highly depends on the starting vector it uses, and may not always converge to a meaningful community partition. This could be improved by starting the algorithm with the vector of predicted labels by (say) SCORE, but the resultant partition is usually close to that of SCORE. For this reason, we omit BCPL for comparison in the analysis below.

At the same time, the results by SCORE, NSC, and APL are reasonably consistent with each other: the ARI between the vector of predicted labels by SCORE and that by NSC is 0.55 and the ARI between the vector of predicted labels by NSC and that by APL is 0.41; see Table 6 for details. In particular, the three methods agree on that, the three communities each of them identifies can be interpreted as follows (arranged in sizes ascendingly).

- "Objective Bayes" community. This community includes a small group of researchers (group sizes are different for different methods, ranging from 20 to 69) including James Berger and his collaborators. Figure 10 presents the "Objective Bayes" community identified by SCORE, where the names for a handful of high-degree nodes are presented.

TABLE 6
The Ajusted Rand Index (ARI) and Variation of Information (VI) for the vectors of predicted community labels by four different methods in Coauthorship network (B), assuming $K = 3$. A large ARI/small VI suggests that the two predicted label vectors are similar to each other.

	SCORE	NSC	BCPL	APL
SCORE	1.00/.00	.55/.51	.00/1.65	.19/.59
NSC		1.00/.00	.00/1.46	.41/.36
BCPL			1.00/.00	.00/1.21
APL				1.00/.00

- "Biostatistics (Coauthorship (B))" (Biostat-Coau-B) community. The sizes of this community by three different methods have quite a bit variability and range from 50 to 388. While it is probably not exactly right to call this community "Biostatistics", the community consists of a number of statisticians and biostatisticians in North Carolina Research Triangle (University of North Carolina (UNC), Duke University (Duke), and North Carolina State University (NCSU)). It also includes many statisticians and biostatisticians from Harvard University, University of Michigan at Ann Arbor, University of Wisconsin at Madison. Figure 11 presents the "Biostatistics" community identified by SCORE, where we similarly show the names of a handful of high-degree nodes.
- "High Dimensional Data Analysis (Coauthorship (B))" (HDDA-Coau-B) community. The sizes of this community by three different methods range from 1811 to 2193. The community include researchers from a wide variety of research areas in or related to high dimensional data analysis (e.g., Bioinformatics, Machine Learning, non-parametric regression, Quantile Regression). Figure 12 presents HDDA-Coau-B community identified by SCORE, with the names of a handful of high-degree nodes shown.

While it seems that three methods agree on that there are three communities as described above, they also present substantial differences. In Table 7, we compare the sizes of the three communities identified by each of the three methods. There are two points worth noting.

First, while SCORE and NSC are quite similar to each other, there is a major difference: NSC clusters about 200 authors, mostly biostatisticians from Harvard University, University of Michigan at Ann Arbor, and University of Wisconsin at Madison, into the HDDA-Coau-B community, but SCORE clusters them into the Biostat-Coau-B community. It seems that the results by SCORE are more meaningful.

Second, APL behaves very differently from either SCORE or NSC. Its estimate of the "Objective Bayes" community is (almost) a subset of its counterpart by either SCORE or NSC, and is much smaller in size (sizes are 20, 64, and 69 for that by APL, SCORE, and NSC). A similar claim applies to the Biostat-Coau-B community identified by each of the methods (sizes are 50, 388, and 169 for that by APL, SCORE, and NSC). This suggests that APL may have underestimated these two communities but overestimated the HDDA-Coau-B community.

It is also interesting to compare these results with those we obtain in Section 4.3 for Coauthorship network (A). Below are three noteworthy points.

First, recall that in Figure 9 and Table 5, we have identified a total of 7 different components of Coauthorship network (A). Among these components, the Duke component (middle panel on top row in Table 5) splits into three parts, each belongs to the three of the communities of Coauthorship network (B) identified by SCORE. The other 6 components fall into the HDDA-Coau-B community identified by SCORE almost completely.

Second, for the giant component of Coauthorship (A), there is a close draw on whether we should cluster the Carroll-Hall's group and Fan's group into two communities: SCORE and APL think that two groups belong to one community, but NSC and BCPL do not agree with this. In Coauthorship (B), both groups are in the HDDA-Coau-B community. Also, in previous studies on this giant component, BCPL and APL separate the nodes in Dunson's branch from the North Carolina group, and cluster them into the Carroll-Hall group. In the current study, however, the whole North Carolina group (including Dunson's branch) are in the Biostat-Coau-B community.

Third, in Coauthorship (A), Gelfand's group is included in this 236-node giant component, where James Berger is not a member. In Coauthorship network (B), Gelfand's group now becomes a subset of "Objective Baye" community where James Berger is a hub node.

4.5. *Community extraction.* The goals of community detection and community extraction are related but also subtly different. The former attempts to assign a class label to each node. The latter, however, attempts to extract one or more meaningful communities, without assigning labels to nodes outside the extracted communities. In principle, methods for community detections can be adapted to methods for community extraction, and vice versa.

A noteworthy approach to community extraction is the approach by Zhao *et al.* [46], which is related to APL [47] in a high level. We have applied this procedure for community detection with the Coauthorship network (B), and have extracted three communities with sizes 493, 214, and 1556. The 493-

TABLE 7
Comparison of sizes of the three communities identified by each of the three methods in Coauthorship network (B), assuming $K = 3$. BCPL is not included for comparisons for its results are inconsistent with those by the other three methods.

	Objective Bayes	Biostat-Coau-B	HDDA-Coau-B
SCORE	64	388	1811
NSC	69	163	2031
APL	20	50	2193
SCORE ∩ NSC	55	162	1807
SCORE ∩ APL	20	50	1811
NSC ∩ APL	20	50	2032
SCORE ∩ NSC ∩ APL	20	50	1807

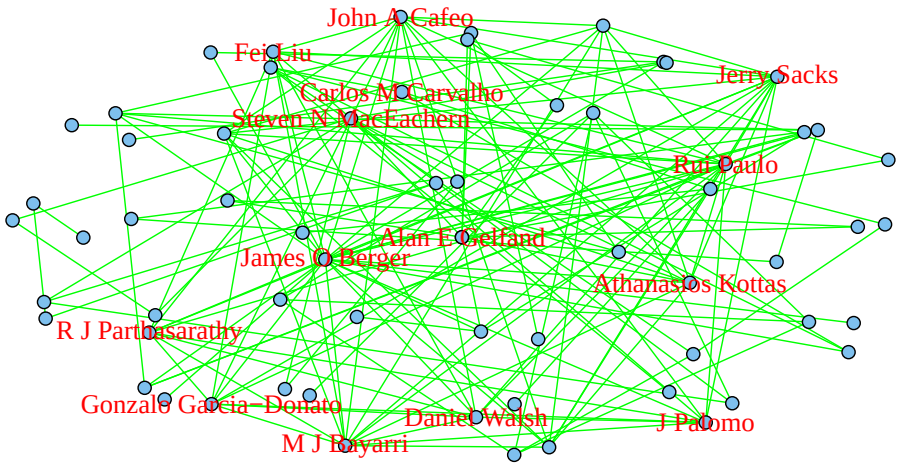

FIG 10. *The "Objective Bayes" community in Coauthorship network (B) identified by SCORE (64 nodes). Only names for 14 nodes with a degree of 9 or larger are shown.*

node one can be interpreted as the HDDA-Coau-B community, including many hub nodes in the HDDA-Coau-B community identified by SCORE. The other two are unfortunately hard to interpret.

5. Community detection for Citation network. The Citation network is a directed network. As a result, the corresponding discussion is different in important ways to that in Section 4, and provides additional insight into the structures of the networks of statisticians. In Section 5.1, we extend DCBM to directed networks, and in Section 5.2, we discuss methods for community detection. In Section 5.3, we analyze the Citation network, and compare the results with those in Section 4.

Denote the Citation network by $\tilde{\mathcal{N}} = (\tilde{V}, \tilde{E})$, where $\tilde{V} = \{1, 2, \ldots, n\}$ is

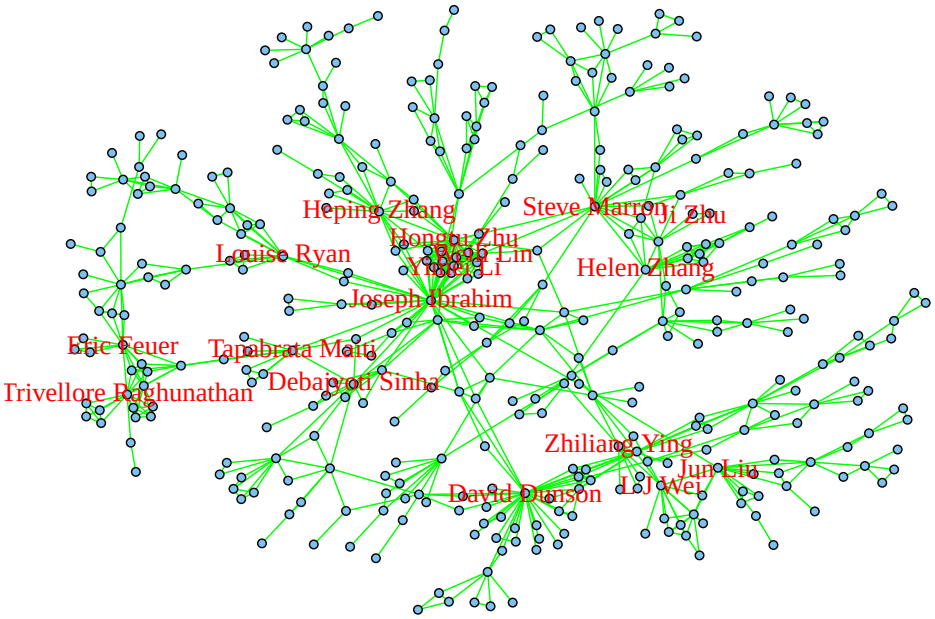

FIG 11. *The "Biostatistics" community (Biostat-Coau-B) in Coauthorship network (B) identified by SCORE (*388 *nodes). Only names for* 17 *nodes with a degree of* 13 *or larger are shown. A "branch" in the figure is usually a research group in an institution or a state.*

the set of nodes (i.e., authors), and $\tilde{E}$ is the set of edges, where for any two distinct nodes $i, j \in \tilde{V}$ (self-citations are not counted by default),

(5.1) there is a directed edge from i to $j \iff i$ has cited j at least once.

To analyze the Citation network, one usually focuses on the *weakly connected giant component* [3]. This is the giant component of the so-called *weakly connected network* associated with the Citation network, where each $i \in \tilde{V}$ is a node and for any two distinct nodes $i, j \in \tilde{V}$, there is an (undirected) edge between them if either i has cited j at least once or j has cited i at least once. Restricting all nodes in $\tilde{\mathcal{N}}$ to the weakly connected giant component gives a (directed) sub-network which we denote by $\mathcal{N}$.

From now on, we restrict our attention to $\mathcal{N}$, and still call it the Citation network for notational simplicity. Note that

(5.2) The weakly connected network associated with $\mathcal{N}$ is connected.

For any citation network $\mathcal{N} = (V, E)$, we can define two associated (undirected) networks as follows.

FIG 12. *The "High Dimensional Data Analysis" community (HDDA-Coau-B) in Coauthorship network (B) identified by SCORE (*1181 *nodes). Only names for* 22 *nodes with degree of* 18 *or larger are shown.*

- *Citer network*. In this network, each $i \in V$ is a node, and there is an (undirected) edge between two distinct nodes i and j if and only if both of them have cited a node k at least once, for some $k \in (V \setminus \{i, j\})$ (i.e., they have a common citee).
- *Citee network*. In this network, each $i \in V$ is a node, and there is an (undirected) edge between two distinct node i and j if and only if each of them has been cited at least once by the same node $k \notin (V \setminus \{i, j\})$ (i.e., they have a common citer).

We shall use these terminologies for the descriptions of both our models and methods. For general directed networks, some other terminologies may be more appropriate. However, for simplicity, we stick to the terminology in this paper, even though the network is not for citations.

Writing $\mathcal{N} = (V, E)$, we are interested in community detection. Similarly as before, we think the nodes in V splits into K different (disjoint) communities

$$V = V^{(1)} \cup V^{(2)} \ldots \cup V^{(K)}.$$

The goal is to assign a community label for each node i, $1 \leq i \leq n$.

Remark 2. In (5.1), we may change the right hand side to that of "i has cited j at least t times" for some $t \geq 2$, but the incentive for doing so lies in the hope that the network automatically splits into many components by choosing an appropriately small t; this is the case for Coauthorship network (A). Unfortunately, this does not work well with the Citation network, where the degree density is much larger than that of Coauthorship network.

5.1. *DCBM for directed network.* We now extend DCBM to directed networks. Let A be the adjacency matrix of a directed network $\mathcal{N}$, where

$$A(i,j) = \begin{cases} 1, & \text{there is a directed edge from } i \text{ to } j, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the diagonals of A are all 0's since we don't consider self citations. Similar as in Section 4.1, we view the off-diagonals of A as Bernoulli random variables and write

$$A = \Omega - \mathrm{diag}(\Omega) + W, \qquad W = (A - E[A]).$$

Introduce two vectors with all positive entries

$$\theta = (\theta(1), \theta(2), \ldots, \theta(n))', \qquad \delta = (\delta(1), \delta(2), \ldots, \delta(n))', \tag{5.3}$$

where $\theta(i)$ models the *degree heterogeneity parameter* for node i as a *citer*, and $\delta(i)$ models the *degree heterogeneity parameter* for node i as a *citee*. Let Θ and Δ be two diagonal matrices such that the i-th diagonals of Θ and Δ are $\theta(i)$ and $\delta(i)$, respectively, and let P be a $K \times K$ symmetric irreducible, and non-negative matrix as before. We model

$$\Omega = \Theta L \Delta, \tag{5.4}$$

where similarly as before, L is the $n \times n$ matrix satisfying

$$L(i,j) = \sum_{k=1}^{K} \sum_{\ell=1}^{K} P(k,\ell) \mathbf{1}_k(i) \mathbf{1}_\ell(j), \qquad 1 \leq i, j \leq n. \tag{5.5}$$

Compared to DCBM for undirected network (4.3)-(4.4), the main difference is that we do not require $\Theta = \Delta$, so Ω is not necessarily a symmetric matrix.

5.2. *Community detection methods for directed networks.* For community detection for directed networks, there are relatively few approaches. In this section, we consider two methods: LNSC and D-SCORE.

LNSC stands for Leicht and Newman's Spectral Clustering approach proposed in [30], where the authors extended the spectral modularity methods aforementioned [38] to directed networks, using the so-called generalized modularity [2]. However, it is pointed out in [28] that LNSC can not properly distinguish the directions of the edges and can not detect communities representing directionality patterns among the nodes. See details therein.

D-SCORE is an extension of SCORE discussed in Section 4.2. SCORE was originally proposed by [25] as a community detection method for undirected network. Below, we carefully adapt it to directed networks, and call the resultant method Directed-SCORE (D-SCORE).

In detail, recall that the key observation underlying SCORE is that the degree heterogeneity parameters $\theta(i)$ are nearly ancillary, and can be largely removed by taking entry-wise ratios between the k-th leading eigenvector $\hat{\xi}_k$ and the first leading eigenvector $\hat{\xi}_1$, $2 \le k \le K$. In the current setting, such an observation is still valid, provided that we replace eigenvalues/eigenvectors by singular values/singular vectors.

In detail, recall that the rank of the $n \times n$ matrices Ω is K. Let

$$\Omega = U\Lambda V', \qquad U, V \in R^{n,K}, \qquad \Lambda \in R^{K,K} \tag{5.6}$$

be the Singular Value Decomposition (SVD) where the diagonals of Λ are sorted descendingly so that the SVD is unique. Define two $n \times K$ matrix $T^{(l)}$ and $T^{(r)}$ and two $n \times (K-1)$ matrices $R^{(l)}$ and $R^{(r)}$ (where the superscripts (l) and (r) stand for left and right, respectively) by

$$U = \Theta T^{(l)}, \quad R^{(l)}(i,k) = \frac{U(i,k+1)}{U(i,1)}, \ \ 1 \le i \le n, \ 1 \le k \le K-1, \tag{5.7}$$

and

$$V = \Delta T^{(r)}, \quad R^{(r)}(i,k) = \frac{V(i,k+1)}{V(i,1)}, \ \ 1 \le i \le n, \ 1 \le k \le K-1. \tag{5.8}$$

Let (θ, δ) be as in (5.3) and recall that $\mathbf{1}_k$ denotes the indicator vector for community k, $1 \le k \le K$. Write $\theta = \sum_{k=1}^{K}(\Theta \mathbf{1}_k)$ and $\delta = \sum_{k=1}^{K}(\Delta \mathbf{1}_k)$. Let D and F be the $K \times K$ diagonal matrices such that $D(k,k) = \|\Theta \mathbf{1}_k\|/\|\theta\|$, $F(k,k) = \|\Delta \mathbf{1}_k\|/\|\delta\|$, $1 \le k \le K$. We have the following lemma, the proof of which is elementary so is omitted.

LEMMA 5.1. *Suppose (5.4)-(5.5) and (5.7)-(5.8) hold where P is non-negative and non-singular, and both PP' and $P'P$ are irreducible, and that all singular values of DPF are simple. Let R be any of the four matrices $R^{(l)}$, $R^{(r)}$, $T^{(l)}$ and $T^{(r)}$. Then R has K district rows, according to which all n rows of R partition into K different groups. Moreover, the partition coincides with the partition of n nodes into K different communities.*

Note also that by Perron's theorem [18, Page 500], all entries of the first column of U and V are strictly positive so both $R^{(l)}$ and $R^{(r)}$ are well-defined.

Lemma 5.1 is similar to Lemma 4.1 but also provides something quite useful which we don't have before: in Lemma 4.1, we have only one matrix R to help us for inferences, but here we have two matrices $R^{(l)}$ and $R^{(r)}$ (which are generally unequal) for inferences. Note that for authors whose in-degrees and out-degrees are both large, it does not make much difference whether we use both matrices or one of them, but for authors for whom either the in-degree or the out-degree is low, it is important to use both.

Lemma 5.1 motivates the following procedure, which we call D-SCORE. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the giant components of the Citer and Citee networks associated with $\mathcal{N}$, respectively. Suppose the first K leading singular values of A are all simple (multiplicity is 1) [18], let $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_K$ be the first K leading left singular vectors of A, and let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$ be the first K leading right singular vectors of A. Define two $n \times (K-1)$ matrices $\hat{R}^{(l)}$ and $\hat{R}^{(r)}$ where for $1 \le k \le K-1$,

$$\hat{R}^{(l)}(i,k) = \begin{cases} \operatorname{sgn}(\hat{u}_{k+1}(i)/\hat{u}_1(i)) \cdot \min\{|\frac{\hat{u}_{k+1}(i)}{\hat{u}_1(i)}|, \log(n)\}, & i \in \mathcal{N}_1, \\ 0, & i \notin \mathcal{N}_1, \end{cases} \tag{5.9}$$

and

$$\hat{R}^{(r)}(i,k) = \begin{cases} \operatorname{sgn}(\hat{v}_{k+1}(i)/\hat{v}_1(i)) \cdot \min\{|\frac{\hat{v}_{k+1}(i)}{\hat{v}_1(i)}|, \log(n)\}, & i \in \mathcal{N}_2, \\ 0, & i \notin \mathcal{N}_2. \end{cases} \tag{5.10}$$

Note that we have thresholded both $|\hat{u}_{k+1}(i)/\hat{u}_1(i)|$ and $|\hat{v}_{k+1}(i)/\hat{v}_1(i)|$ with a threshold $\log(n)$; this is recommended by Jin [25] and applies to both SCORE and D-SCORE. We show that, under some mild conditions, $\hat{u}_1(i) \neq 0$ for all $i \in \mathcal{N}_1$ and $\hat{v}_1(i) \neq 0$ for all $i \in \mathcal{N}_2$, so both matrices are well-defined. In detail, for any $S \subset \{1, 2, \ldots, n\}$ and any $n \times n$ matrix A, let $A^{S,S}$ be the sub-matrix of A formed by restricting the rows and columns of A to S. By definitions of the Citer and Citee networks, both $(AA')^{\mathcal{N}_1,\mathcal{N}_1^c}$ and $(A'A)^{\mathcal{N}_2,\mathcal{N}_2^c}$ are matrices of 0's, where for $m = 1, 2$, $\mathcal{N}_m^c = \mathcal{N} \setminus \mathcal{N}_m$. We assume

$$\|(AA')^{\mathcal{N}_1,\mathcal{N}_1}\| > \|(AA')^{\mathcal{N}_1^c,\mathcal{N}_1^c}\|, \qquad \|(A'A)^{\mathcal{N}_2,\mathcal{N}_2}\| > \|(A'A)^{\mathcal{N}_2^c,\mathcal{N}_2^c}\|. \tag{5.11}$$

Note that by Random Matrix Theory, (5.11) holds with overwhelming probabilities, assuming DCBM and some mild regularity conditions.

LEMMA 5.2. *Consider a directed network $\mathcal{N} = (V, E)$ where (5.11) holds. If the multiplicity of the first leading singular value of A is 1, then $\hat{u}_1(i) \neq 0$ if and only if $i \in \mathcal{N}_1$ and $\hat{v}_1(i) \neq 0$ if and only if $i \in \mathcal{N}_2$.*

Lemma 5.2 is the direct result of Perron's theorem [18, Page 500], so we omit the proof. In a way, $\hat{R}^{(l)}$ and $\hat{R}^{(r)}$ can be viewed as the stochastic counterparts of $R^{(l)}$ and $R^{(r)}$, respectively. They display similar but non-identical patterns, and each of them contains valuable information for the community structures that can be combined for clustering.

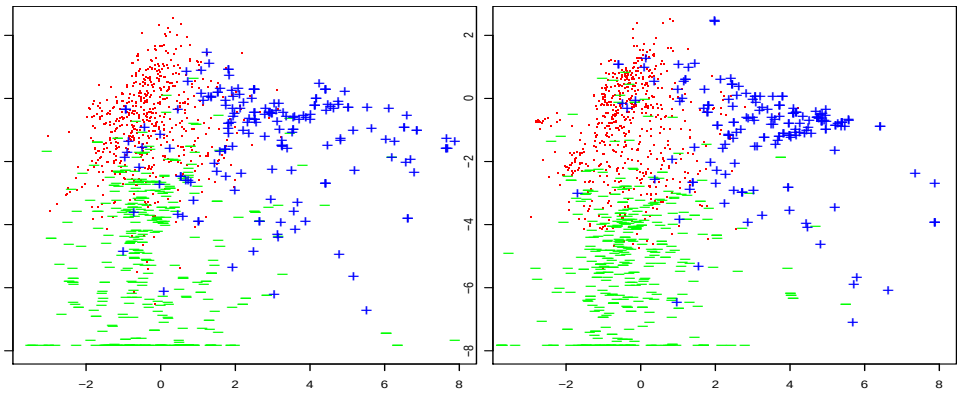

FIG 13. *Left: each point represents a row of the matrix $\hat{R}^{(l)}$ associated with the statistical Citation network ($K = 3$; x-axis: first column, y-axis: second column). Only rows with indices in $\mathcal{N}_1$ are shown. Blue pluses, green bars, and red dots represent 3 different communities identified by SCORE, which can be interpreted as "Large-Scale Multiple testing", "Biostatistics (Citation)", and "Variable Selection", Right: similar but with $(\hat{R}^{(l)}, \mathcal{N}_1)$ replaced by $(\hat{R}^{(r)}, \mathcal{N}_2)$.*

We now introduce D-SCORE. Given a directed network $\mathcal{N} = (V, E)$ where (5.2) holds and the number of communities K. As before, let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the giant components of the Citer network and Citee network, respectively. Obtain two $n \times (K-1)$ matrices $\hat{R}^{(l)}$ and $\hat{R}^{(r)}$ as in (5.9)-(5.10). Note that all nodes split into four disjoint subsets:

$$\mathcal{N} = (\mathcal{N}_1 \cap \mathcal{N}_2) \cup (\mathcal{N}_1 \setminus \mathcal{N}_2) \cup (\mathcal{N}_2 \setminus \mathcal{N}_1) \cup (\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2)).$$

D-SCORE clusters nodes in each subset separately.

1. $(\mathcal{N}_1 \cap \mathcal{N}_2)$. We first restrict the rows of $\hat{R}^{(l)}$ and $\hat{R}^{(r)}$ to the set $\mathcal{N}_1 \cap \mathcal{N}_2$ and obtain two matrices $\tilde{R}^{(l)}$ and $\tilde{R}^{(r)}$. We cluster all nodes in $\mathcal{N}_1 \cap \mathcal{N}_2$

by applying the k-means to the matrix $[\tilde{R}^{(l)}, \tilde{R}^{(r)}]$ assuming there are $\leq K$ communities.

2. $(\mathcal{N}_1 \setminus \mathcal{N}_2)$. Note that according to the communities we identified above, the rows of $\tilde{R}^{(l)}$ partition into $\leq K$ groups. For each group, we call the mean of the row vectors the *community center*. For a node i in $\mathcal{N}_1 \setminus \mathcal{N}_2$, if the i-th row of $\hat{R}^{(l)}$ is closest to the center of the k-th community for some $1 \leq k \leq K$, then we assign it to this community.
3. $(\mathcal{N}_2 \setminus \mathcal{N}_1)$. We cluster in a similar fashion to that in the last step, but we use $(\tilde{R}^{(r)}, \hat{R}^{(r)})$ instead of $(\tilde{R}^{(l)}, \hat{R}^{(l)})$.
4. $(\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2))$. We say there is a weak-edge between i and j if there is an edge between i and j in the weakly connected citation network. By 1-2, all nodes in $\mathcal{N}_1 \cup \mathcal{N}_2$ partition into $\leq K$ communities. For each node in $\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2)$, we assign it to the community to which it has the largest number of weak-edges.

For 4, our assumption is that $|\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2)|$ is small, so we don't have to have a sophisticated clustering method. For the statistical citation network data set we study in this paper, this is true with $|\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2)| = 14$.

In Figure 13, we illustrate how D-SCORE works by using the statistical citation network data set with $K = 3$ (left: rows of $\hat{R}^{(l)}$; right: rows of $\hat{R}^{(r)}$). Two panels show similar clustering patterns, confirming the main message of Lemma 5.1. D-SCORE combines the information in both $\hat{R}^{(l)}$ and $\hat{R}^{(r)}$ for clustering, and suggests that there are three communities in the network, which can be interpreted as "Large-Scale Multiple Testing", "Biostatistics (Citation)", and "Variable Selection". See details in Section 5.3.

5.3. *Citation network.* The original Citation network $\tilde{N} = (\tilde{V}, \tilde{E})$ consists of 3607 nodes (i.e., authors). The associated weakly connected network has 927 components. The giant component has 2654 authors, accounting 74% of all nodes. All other components have no more than 5 nodes.

We now restrict our attention to the sub-network $\mathcal{N} = (V, E)$ of $\tilde{N}$, where V consists of all nodes in the weakly connected giant component. As before, let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the giant components of the Citer and Citee networks associated with $\mathcal{N}$, respectively. We have $|\mathcal{N}_1| = 2126$, $|\mathcal{N}_2| = 1790$, $|\mathcal{N}_1 \cap \mathcal{N}_2| = 1276$, and $|\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2))| = 14$.

We are primarily interested in community detection. In Figure 5 (right panel), we present the scree plot of A. Note that since A is non-symmetric, we use the singular values instead of the eigenvalues in the plot. The plot suggests that there are $K = 3$ communities in $\mathcal{N}$.

We have applied D-SCORE and LNSC to $\mathcal{N}$. The results by SCORE are reported with details below in this section. We find that the results of LNSC

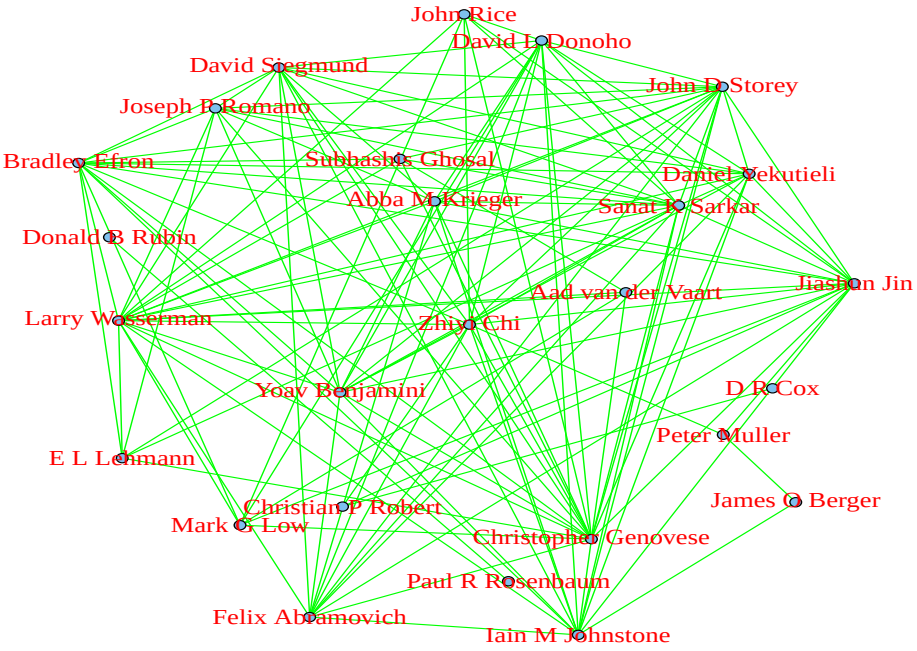

FIG 14. *The "Large-Scale Multiple Testing" community identified by D-SCORE ($K = 3$) in the Citation network (359 nodes). Only 26 nodes with 24 or more citers are shown here.*

are rather inconsistent with those of SCORE, so we only discuss them briefly in the end of this section (Section 5.3.3).

We now present the results by D-SCORE. The method identifies there communities which can be interpreted as follows.

- "Large-Scale Multiple Testing" (Multiple Tests) community (359 nodes). This consists of researchers in multiple testing and control of False Discovery Rate. It includes the Tel-Aviv group (e.g., Felix Abramovich, Yoav Benjamini, Daniel Yekutieli), Stanford group (e.g., David Donoho, Bradley Efron, Iain Johnstone, Joseph Romano, David Siegmund), Carnegie Mellon group (e.g., Christopher Genovese, Jiashun Jin, Isbella Verdini, Larry Wasserman), etc.
- "Biostatistics (Citation)" (Biostat-Cita) community (1010 nodes). This includes most authors from the "Biostatistics (Coauthorship (B))" community identified by SCORE in Section 4.4 (388 nodes). The high-

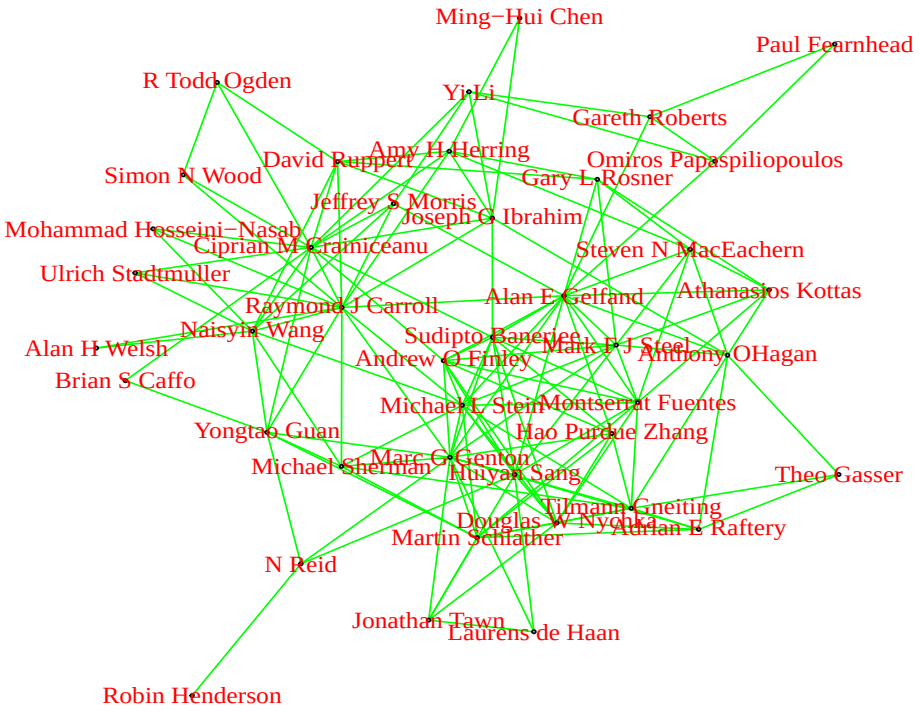

FIG 15. *The "Biostatistics (Citation)" community identified by D-SCORE ($K = 3$) in the Citation network (*1010 *nodes). Only* 42 *nodes with* 24 *or more citers are shown here.*

degree nodes include (sorted descendingly by the number of citers) Raymond Carroll, Gareth Roberts, Joseph Ibrahim, Naisyin Wang, Adrian Raftery, Omiros Papaspiliopoulos, David Ruppert, Alan Gelfand, Tilmann Gneiting, Jeffrey Morris, Michael Stein, Ciprian Crainiceanu, Marc Genton, Hao Zhang, Fernando Quintana, Nicolas Chopin, Alan Welsh, Anthony OHagan, Fadoua Balabdaoui, Sudipto Banerjee, Nancy Reid, Paul Fearnhead, Steven MacEachern, Douglas Nychka, Gary Rosner.

- "Variable Selection" (Var. Selection) community (1285 nodes). This includes the high-degree nodes such as (sorted descendingly by the number of citers) Jianqing Fan, Hui Zou, Peter Hall, Nicolai Meinshausen, Peter Buhlmann, Ming Yuan, Yi Lin, Runze Li, Peter Bickel, Trevor Hastie, Hans-Georg Muller, Emmanuel Candes, Cun-Hui Zhang, Heng Peng, Jian Huang, Tony Cai, Terence Tao, Jianhua Huang, Alexandre Tsybakov, Jonathan Taylor, Xihong Lin, Jane-Ling Wang, Dan Yu

Lin, Fang Yao, Jinchi Lv.

The three communities are presented in Figures 14-16, respectively.

It is interesting to compare these results with those of Coauthorship network. In Sections 5.3.1 and 5.3.2, we compare the results presented above with those of Coauthorship networks (A) and (B), respectively.

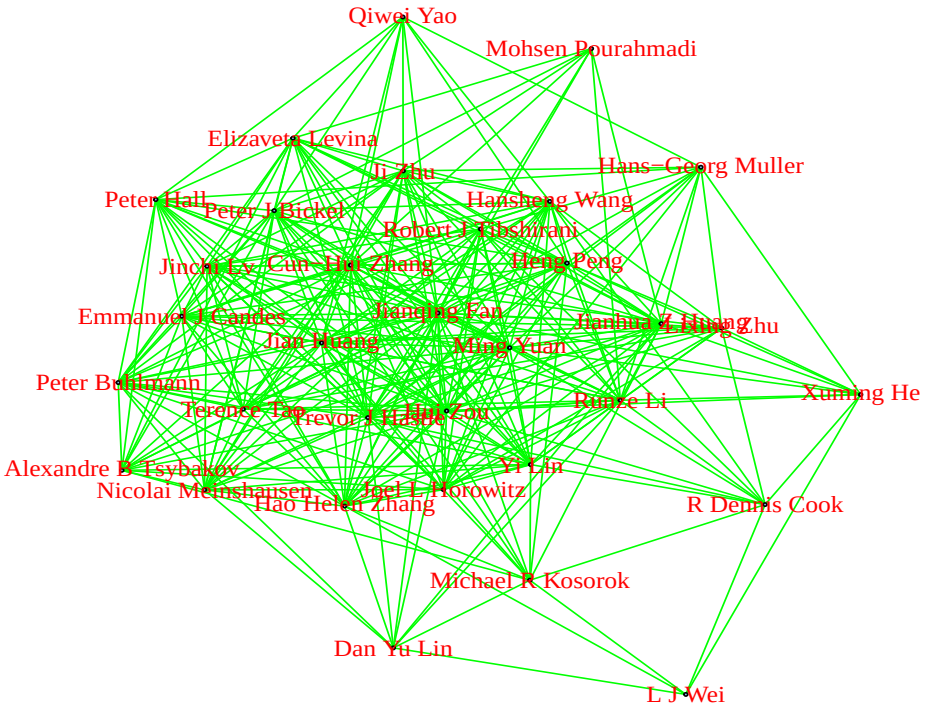

FIG 16. *The "Variable Selection" community identified by D-SCORE ($K = 3$) in the Citation network (*1285 *nodes). Only* 40 *nodes with* 54 *or more citers are shown here.*

5.3.1. *Comparison with Coauthorship network (A).* In Section 4.3, we present 8 different components of Coauthorship network (A). In Table 8, we reinvestigate all these components in order to understand their relationship with the 3 communities identified by D-SCORE in the Citation network.

Among these 8 components, the first one is the giant component, consisting of 236 nodes. All except 3 of these nodes fall in the 3 communities identified by D-SCORE in the Citation network, with 60 nodes in "Biostatistics (Citation)" including (sorted descendingly by the number of citers; same

TABLE 8

Sizes of the intersections of the communities identified by D-SCORE ($K = 3$) in the Citation network (rows; "other" stands for nodes outside the weakly connected giant component) and the 8 largest components of Coauthorship network (A) as presented in Figures 6 and 9 and Tables 5 (columns).

	giant	Mach. Learn.	Dim. Reduc.	John Hopkins	Duke	Stanford	Quant. Reg.	Exp. Design
Biostatistics	60	1		12	1			3
Var. Selection	166	15	14	1	7	2	8	2
Multiple Tests	7	2			2	7	1	3
Other	3							
	236	18	14	13	10	9	9	8

below) Raymond Carroll, Joseph Ibrahim, Naisyin Wang, Alan Gelfand, Jeffrey Morris, Marc Genton, Sudipto Banerjee, Hongtu Zhu, Jeng-Min Chiou, Ju-Hyun Park, Ulrich Stadtmuller, Ming-Hui Chen, Yi Li, Nilanjan Chatterjee, Andrew Finley, 166 nodes in "Variable Selection" including Jianqing Fan, Hui Zou, Peter Hall, Ming Yuan, Yi Lin, Runze Li, Trevor Hastie, Hans-Georg Muller, Emmanuel Candes, Cun-Hui Zhang, Heng Peng, Jian Huang, Tony Cai, Jianhua Huang, Xihong Lin, and 7 nodes in "Large-Scale Multiple Testing" including David Donoho, Jiashun Jin, Mark Low, Wenguang Sun, Ery Arias-Castro, Michael Akritas, Jessie Jeng.

This is consistent with our previous claim that this 236-node giant component contains a "Carroll-Hall" community and a "North Carolina" community: The "Carroll-Hall" community has strong ties to the area of variable selection, and the "North Carolina" community has strong ties to Biostatistics. Raymond Carroll has close ties to both of these two communities, and it is not surprising that SCORE assigns him to the "Carroll-Hall" community in Section 4.3 in Coauthorship network (A) but D-SCORE assigns him to the "Biostatisticis (Citation)" community in the Citation network.

For the remaining 7 components of Coauthorship network (A), "Theoretical Machine Learning", "Dimension Reduction", "Duke", "Quantile Regression" are (almost) subsets of "Variable Selection", "Stanford" (including John Storey, Johathan Taylor, Ryan Tibshirani) is (almost) a subset of "Large-Scale Multiple Testing", and "Johns Hopkins" is (almost) a subset of "Biostatistics (Citation)". The "Experimental Design" group has no strong preferences over all these three areas, so the nodes spread almost evenly to these three communities.

5.3.2. *Comparison with Coauthorship network (B).* We compare the community detection results by D-SCORE for the Citation network with those by SCORE for Coauthorship network (B) in Section 4.4. Note that for the

former, we have been focused on the weakly connected giant component of the Citation network (2654 nodes), and for the latter, we have been focused on the giant component of the Coauthorship network (B) (2263 nodes). The comparison of two sets of results is tabulated in Table 9.

Viewing the table vertically, we observe that Citation network provides additional insight into the Coauthorship network (B), and reveals structures we have not found previously. Below are the details.

First, the "Objective Bayes" community in Coauthorship network (B) contains two main parts. The first part consists of 55% of the nodes, and most of them are seen to be the researchers who have close ties to James Berger, including (sorted descendingly by the number of citers; same below) Alan Gelfand, Fernando Quintana, Steven MacEachern, Gary Rosner, Rui Paulo, Herbert Lee, Robert Gramacy, Athanasios Kottas, Pilar Iglesias, Daniel Walsh, Dongchu Sun. The second part consists of 25% of the nodes, and are assigned to the "Variable Selection" community in the Citation network by D-SCORE, including Carlos Carvalho, Feng Liang, Maria De Iorio, German Molina, Merlise Clyde, Luis Pericchi, Maria Barbieri, Nicholas Polson, Bala Rajaratnam, Edward George. For the second part, the result seems reasonable, as many nodes in the second part (e.g., Carlos Carvalho, Edward George, Feng Liang, Merlise Clyde) have an interest in model selection.

Second, the "Biostatistics (Coauthorship (B))" community in Coauthorship network (B) also has two main parts. The first part has 156 nodes (40% of the total, including high-degree nodes such as Joseph Ibrahim, Sudipto Banerjee, Hongtu Zhu, Ju-Hyun Park, Ming-Hui Chen, Yi Li, Montserrat Fuentes, Natesh Pillai, Andrew Finley, Amy Herring, Martin Schlather, Stuart Lipsitz, Jonathan Tawn, Siddhartha Chib, Alexander Tsodikov. The second part consists of 153 nodes (40% of the total). The high-degree nodes include Yi Lin, Dan Yu Lin, Ji Zhu, Helen Zhang, L J Wei, Wei Biao Wu, Donglin Zeng, Zhiliang Ying, David Dunson, Steve Marron, Anastasios Tsiatis, Wenbin Lu, Zhezhen Jin, Xiaotong Shen, Heping Zhang, Lu Tian, Jianwen Cai, Wing Hung Wong. The results are quite reasonable: many nodes in the second part (e.g., Dan Yu Lin, David Dunson, Helen Zhang, Steve Marron, Ji Zhu, Xiaotong Shen, Yi Lin) either have works in or have strong ties to the area of variable selection.

Last, the "High Dimensional Data Analysis" community in Coauthorship network (B) has three parts. The first part has 459 nodes (25%), including high-degree nodes such as Raymond Carroll, Gareth Roberts, Naisyin Wang, Adrian Raftery, Omiros Papaspiliopoulos, David Ruppert, Tilmann Gneiting, Jeffrey Morris, Michael Stein, Ciprian Crainiceanu, Marc Genton, Nicolas Chopin, Alan Welsh, Anthony OHagan, Fadoua Balabdaoui, N Reid.

The second part has 840 nodes (46%), including high-degree nodes such as Jianqing Fan, Hui Zou, Peter Hall, Nicolai Meinshausen, Peter Buhlmann, Ming Yuan, Runze Li, Peter Bickel, Trevor Hastie, Hans-Georg Muller, Emmanuel Candes, Cun-Hui Zhang, Heng Peng, Jian Huang, Tony Cai, Terence Tao, Jianhua Huang, Alexandre Tsybakov, Jonathan Taylor, Xihong Lin. The third part has 221 nodes (26%), including high-degree nodes such as Iain Johnstone, Larry Wasserman, Bradley Efron, John Storey, Christopher Genovese, David Donoho, Yoav Benjamini, David Siegmund, Peter Muller, Jiashun Jin, Felix Abramovich, David Cox, Daniel Yekutieli.

Respectively, the three parts are labeled as subsets of the "Biostatistics (Citation)", "Variable Selection", and "Large-Scale Multiple Testing" communities in the Citation network. This seems convincing: (a) most of the nodes in the first part are Biostatisticians or have a strong interest in Biostatistics (e.g., Ciprian Crainiceanu, Naisyin Wang, Raymond Carroll), (b) most of the nodes in the second part are leaders in variable selection, and (c) most nodes in the third part are leaders in Large-Scale Multiple Testing and in the topic of control of FDR.

Viewing the table horizontally gives similar claims but also reveals some additional insight. For example, "Large-Scale Multiple Testing" contains three main parts. One part consists of 221 nodes and is a subset of the "High Dimensional Data Analysis" community in Coauthorship network (B). The second consists of 115 nodes and falls outside the giant component of Coauthorship network (B). A significant fraction of nodes in this part are from Germany and have close ties to Helmut Finner, a leading researcher in Multiple Testing. Another significant part (17 nodes) are researchers in Bioinformatics (e.g., Terry Speed) who do not publish many papers in these four journals for the time period.

TABLE 9

Sizes of the intersections of the communities identified by D-SCORE ($K = 3$) in the Citation network (rows; "other" stands for nodes outside the weakly connected giant component) and the communities identified by SCORE in Coauthorship network (B) (columns; "other" stands for nodes outside the giant component).

	Obj. Bayes	Biostat-Coau-B	HDDA-Coau-B	other	
Biostat-Cita	35	156	459	360	1010
Var. Selection	16	153	840	276	1285
Multiple Tests	6	17	221	115	359
other	7	62	291		360
	64	388	1811	751	3014

5.3.3. *Comparison of D-SCORE and LNSC.* We have also applied LNSC to the Citation network, with $K = 3$. The communities are very different

from those identified by D-SCORE, and maybe interpreted as follows.

- "Semi-parametric and non-parametric" (434 nodes). We find this community hard to interpret, but it could be the community of researchers on semi-parametric and non-parametric models, functional estimation, etc.. The hub nodes include (sorted descendingly by the number of citers; same below) Peter Hall, Raymond Carroll, Hans-Georg Muller, Xihong Lin, Fang Yao, Naisyin Wang, Marina Vannucci, David Ruppert, Gerda Claeskens, Wolfgang Hardle, Jeffrey Morris, Enno Mammen, Ciprian Crainiceanu, James Robins, Anastasios Tsiatis, Catherine Sugar, Zhezhen Jin, Alan Welsh, Sunil Rao, Philip Brown.
- "High Dimensional Data Analysis" (HDDA-Cita-LNSC) (614 nodes). The second one can be interpreted as the "High Dimensional Data Analysis" community, where the high-degree nodes include (sorted descendingly by the number of citers) Jianqing Fan, Hui Zou, Nicolai Meinshausen, Peter Buhlmann, Ming Yuan, Yi Lin, Iain Johnstone, Runze Li, Peter Bickel, Trevor Hastie, Larry Wasserman, Emmanuel Candes, Cun-Hui Zhang, Heng Peng, Bradley Efron, John Storey, Jian Huang, Tony Cai, Christopher Genovese, Terence Tao.
- "Biostatistics" (Biostat-Cita-LNSC) (1605 nodes). The community is hard to interpret, but could be the Biostatistics community. The high-degree nodes include Xuming He, Gareth Roberts, Joseph Ibrahim, Adrian Raftery, Peter Muller, Omiros Papaspiliopoulos, Alan Gelfand, L J Wei, Tilmann Gneiting, James Berger, Michael Stein, Zhiliang Ying, David Dunson, Nils Lid Hjort, Marc Genton, David Cox, Hao Zhang, Fernando Quintana, Nicolas Chopin, Zhiyi Chi.

These results are rather inconsistent to those obtained by D-SCORE: the ARI and VI between two the vectors of predicted community labels by LNSC and SCORE are 0.07 and 1.68, respectively. Moreover, it seems that

- LNSC merges part of the nodes in the "Variable Selection" (1285 nodes) and "Large-Scale Multiple Testing" (359 nodes) communities identified by D-SCORE into a new HDDA-Cita-LNSC community, but with a much smaller size (614 nodes).
- The Biostat-Cita-LNSC community (1605 nodes) is much larger than the Biostat-Cita community identified by D-SCORE (1010 nodes).

Our observations here somehow agree with [28] that LNSC can not properly distinguish the directions of the edges and can not detect communities representing directionality patterns among the nodes.

6. Discussions. We have collected, cleaned, and analyzed two network data sets: the Coauthorship network and Citation network for statisticians. We investigate several different aspects of these networks: productivity, patterns and trends, centrality, community structures, with an array of different tools, ranging from Exploratory Data Analysis (EDA) [44] tools such as Lorenz curve to rather sophisticated methods for community detection. Some of these tools are relatively recent (e.g., SCORE, NSC, BCPL, APL, LNSC), and some are even new (e.g., we propose D-SCORE as a new method for community detection). We have also presented an array of interesting results. For example, we find the statistics community has become increasingly more collaborative, competitive, and globalized, and identify about 15 meaningful communities such as "Biostatistics", "Dimension Reduction", "Large-Scale Multiple Testing", "Objective Bayes", "Quantile Regression", "Theoretical Machine Learning", and "Variable Selection".

The paper also has several limitations that need further explorations. First of all, constrained by time and resources, the two data sets we collect are limited to the papers published in four "core" statistical journals: AoS, JASA, JRSS-B, and Biometrika in the 10 year period from 2003 to 2012. We recognize that many statisticians not only publish in so-called "core" statistical journals but also publish in a wide variety of journals of other scientific disciplines, including but not limited to Nature, Science, PNAS, IEEE journals, journals in computer science, cosmology and astronomy, economics and finance, probability, and social sciences. We also recognize that many statisticians (even very good ones, such as David Donoho, Steven Fienberg) do not publish often in these journals in this specific time period. For these reasons, some of the results presented in this paper may be biased and they need to be interpreted with caution.

Still, the two data sets and the results we presented here serve well for our purpose of understanding many aspects of the networks of statisticians who have USA as their home base; see Section 1.3. They also serve as a good starting point for a much more ambitious project on social networks for statisticians with a more "complete" data set for statistical publications.

Second, for reasons of space, we have primarily focused on data analysis in this paper, and the discussions on models, theory, and methods have been kept as brief as we can. On the other hand, the data sets provide a fertile ground for modeling and development of methods and theory, and there are an array of interesting problems worthy of exploration in the near future. For example, what could be a better model for either of the two data sets, what could be a better measure for centrality, and what could be a better method for community detection. In particular, we propose D-SCORE as a

new community detection method for directed network, but we only present the idea underlying the methods, without careful analysis. We address the latter in a forthcoming paper [24]. Also, sometimes, the community detection results by different methods (e.g., SCORE, D-SCORE, NSC, BCPL, APL, LNSC) are inconsistent with each other. When this happens, it is hard to have a conclusive comparison or interpretation. In light of this, it is of great interest to set up a theoretical framework and use it to investigate the weaknesses and strengths of these methods.

Last but not the least, there are many other problems we have not addressed here: link prediction, relationship between citations and recognitions (e.g., receiving an important award, elected to National Academy of Science), relationship and differences between "important work", "influential work", and "popular work". It is of interest to explore these in the future.

7. Appendix. In this section, we describe how the data were collected and preprocessed, and how we have overcome the challenges we have faced.

We focus on all papers published in AoS, JASA, JRSS-B, and Biometrika from 2003 to the first half of 2012. For each paper in this range, we have extracted the Digital Object Identifier (DOI), title, information for the authors, abstract, keywords, journal name, volume, issue, and page numbers, and the DOIs of the papers in the same range that have cited this paper. The raw data set consists all such entries of (about) 3500 papers and 4000 authors.

Among these papers, we are only interested in those for original research, so we have removed items such as the book reviews, erratum, comments or rejoinders, etc. Usually, these items contain signal words such as "Book Review", "Corrections" etc. in the title. Removing such items leaves us with a total of 3248 papers (about 3950 authors) in the range of interest.

Our data collection process has three main steps. In the first step, we identify all papers in the range of interest. In the second step, we figure out all citations between the papers of interest (note that the information for *citation relationship between any two authors* is not directly available). In the third step, we identify all the authors for each paper.

In the first step, recall that the goal is to identify every paper in our range of interest, and for each of them, to collect information for the title, author, DOI, keywords, abstract, journal name, etc. In this step, we face two main challenges.

First, all popular online resources have strict limits for high-quality high-volume downloads; we have explained this in Section 1.2 with details. Eventually, we manage to overcome the challenge by downloading the desired

data and information from Web of Science and MathSciNet little by little, each time in the maximum amour that is allowed. Overall, it has taken us a few months to download and combine the data from two different sources.

Second, it is hard to find a good identifier for the papers. While the titles of the papers could serve as unique identifiers, they are difficult to format and compare. Also, while many online resources have their own paper identifiers, they are either unavailable or unusable for our purpose. Eventually, we decide to use the DOI as the identifier. The DOI has been used as a unique identifier for papers by most publishers for statistical papers since 2000.

Using DOI as the identifier, with substantial time and efforts, we have successfully identified all paper in the range of interest with Web of Science and MathSicNet. One more difficulty we face here is that Web of Science does not have the DOIs of (about) 200 papers and MathSciNet does not have the DOIs of (about) 100 papers, and we have to combine these two online sources to locate the DOI for each paper in our range of interest.

We now discuss the second step. The goal is to figure out the citation relationship between any two papers in the range of interest. MathSciNet does not allow automated downloads for such information, but, fortunately, such information is retrievable from Web of Science, if we parse the XML pages in R at a small amount each time. One issue we encounter in this step is that (as mentioned above) Web of Science misses the DOIs of about 200 papers, and we have to deal with these papers with extra efforts.

Consider the last step. The goal is to uniquely identify all authors for each paper in the range of interest. This is the most time consuming step, and we have faced many challenges. First, for many papers published in Biometrika, we do not have the first name and middle initial for each author, and this causes problems. For instance, "L. Wang" can be any one of "Lan Wang", "Li Wang", "Lianming Wang", etc. Second, the name of an author is not listed consistently in different occasions. For example, "Lixing Zhu" may be also listed as "Li Xing Zhu", "L. X. Zhu", and "Li-Xing Zhu". Last but not the least, different authors may have the same name: at least three authors (from Univ. of California at Riverside, Univ. of Michigan at Ann Arbor and Iowa State Univ., respectively) have the same name of "Jun Li".

Note that every service has its own internal identification system, but, unfortunately, none of them is willing to reveal the system to the end users. Also, people have been trying hard to create a universal author identification system, in a similar spirit to that of using DOI as an universal identifier for each paper. Among these are ResearcherID introduced by Thomson Reuters in 2008 and Open Researcher and Contributor ID (ORCID) introduced in 2012. However, the use of such systems is still very limited.

Eventually, we have to solve the problem on our own. First, roughly saying, we have written a program which mostly uses the author names (e.g., first, middle, and last names; abbreviations) to correctly identify all except 200 (approximately) authors, about whom we may have problems in identification. We then manually identify each of these 200 authors using additional information (e.g., affiliations, email addresses, information on their websites). After all such cleaning, the number of authors is reduced from about 3950 to 3607.

For reproducibility purpose, we have prepared the data files and a demo for readers who are interested in exploring the data sets. All these can be found at `www.stat.uga.edu/~psji/` once the paper is accepted for publication. In particular, the data files include the following.

- `4Journals.bib`: the raw bibtex data for about 3500 items including papers, book reviews, corrections, etc
- `4Journals_cleaned.bib`: the cleaned bibtex data for 3248 papers after removing the book reviews, paper corrections and clustering the author names
- `author-cluster.txt`: the final clustering rules for the author names
- `author-cluster-man.txt`: the manually defined clustering rules for the author names
- `author_list.txt`: the list of the 3607 valid authors after disambiguation
- `author-paper-adjacency.txt`: the 3607x3248 bipartite adjacency matrix
- `coauthor-adjacency.txt`: the 3607x3607 coauthor adjacency matrix
- `citation-adjacency.txt`: the 3607x3607 adjacency matrix for the Citation network of authors

Acknowledgements. JJ thanks David Donoho and Jianqing Fan; the paper was inspired by a lunch conversation with them in 2011 on H-index. PJ and JJ thank Yunpeng Zhao for helpful pointers.

References.

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

[2] Arenas, A., Duch, J., Fernandez, A. and Gomez, S. (2007). Size reduction of complex networks preserving modularity. *New J. Phys.* **9(6)** 176.

[3] Bang-Jensen, J. and Gutin, G. (2009). *Digraphs: Theory, Algorithms and Applications*. Springer.

[4] Barabasi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286** 509-512.

[5] Bickel, P. and Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Nat. Acad. Sci.* **106** 21068-21073.
[6] Bickel, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227.
[7] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604.
[8] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313–2351.
[9] Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61.
[10] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499.
[11] Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99** 710–723. MR2090905 (2005d:62053)
[12] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. B* **70** 849–911. MR2530322
[13] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194 (2005g:62047)
[14] Freeman, L., Borgatti, S. and White, D. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Soc. Networks* **13** 141–154.
[15] Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series* **208** 73-79.
[16] Goldenberg, A., Zheng, A., Fienberg, S. and Airoldi, E. (2009). A survey of statistical network models. *Foundations and Trends in machine learning* **2** 129-233.
[17] Grossman, J. (2002). The evolution of the mathematical research collaboration graph. *Congressus Numerantium* **158** 201-212 .
[18] Horn, R. and Johnson, C. (1990). *Matrix Analysis*. Cambridge University Press.
[19] Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613.
[20] Huang, J., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98.
[21] Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classif.* **2** 193-218.
[22] Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557
[23] Ioannidis, J. (2008). Measuring co-authorship and networking-adjusted scientific impact. *PLOS ONE* **3**.
[24] Ji, P., Jin, J. and Ke, Z. (2014). Joint community detection for Coauthorship and Citation networks of statisticians by D-SCORE. *Manuscript*.
[25] Jin, J. (2014). Fast community detection by SCORE. *Ann. Statist. To appear*.
[26] Johnstone, I. and Silverman, B. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752. MR2166560
[27] Karrer, B. and Newman, M. (2011). Stochastic blockmodels and community structures in network. *Phys. Rev.* **83** 1436–1462.
[28] Kim, Y., Son, S.-W. and Jeong, H. (2010). Finding communities in directed networks. *Phys. Rev. E* **81** 016103.
[29] Kolaczyk, E. (2009). *Statistical analysis of network data: methods and models* **200**. Springer, NY.
[30] Leicht, E. and Newman, M. (2008). Community structure in directed networks. *Phys. Rev. Lett.* **100** 118703.

[31] Li, Q. (2013). Personal communications.
[32] Martin, T., Ball, B., Karrer, B. and Newman, M. (2013). Coauthorship and citation patterns in the Physical Review. *Phys. Rev. E* **88**.
[33] Meila, M. (2003). Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines: 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop* (B. Scholkopf and M. K. Warmuth, eds.) Springer.
[34] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462.
[35] Newman, M. (2001). The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98** 404-409.
[36] Newman, M. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E* **64** 016131.
[37] Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci. USA* **101** 5200-5205.
[38] Newman, M. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103** 8577-8582.
[39] Newman, M. (2010). *Networks: an introduction*. Oxford University Press.
[40] Newman, M. and Girvan, M. (2004). Finding and evaluating community structures in networks. *Phys. Rev. E.* **69** 026113.
[41] Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika* **31** 581–683.
[42] Storey, J. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398 (2004k:62055)
[43] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.
[44] Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley.
[45] Wasserman, S. (1994). *Social network analysis: methods and applications* **8**. Cambridge University Press.
[46] Zhao, Y., Levina, E. and Zhu, J. (2011). Community extraction for social networks. *Proc. Nat. Acad. Sci.* **108** 7321-7326.
[47] Zhao, Y., Levina, E. and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266-2292.
[48] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.
[49] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327
[50] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 (2010a:62222)

Pengsheng Ji
Department of Statistics
University of Georgia
Athens, GA 30602
E-mail: psji@uga.edu

Jiashun Jin
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
E-mail: jiashun@stat.cmu.edu