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.

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 findings are consistent with the belief that the statistics community has become increasingly more collaborative, competitive, and globalized.

For (b), we have identified the most prolific/collaborative/highly cited authors. We have also identified a handful of "hot" papers, suggesting "Variable Selection" as one of the "hot" areas.

For (c), we have identified about 15 meaningful communities or research groups, including large-size ones such as "Spatial Statistics", "Large-Scale Multiple Testing", "Variable Selection" as well as small-size ones such as "Dimensional Reduction", "Bayes", "Quantile Regression", and "Theoretical Machine Learning".

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

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: it may help administrators or funding agencies to prioritize research areas, and researchers to start a new topic or a new collaboration, and so on and so forth.

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.,

[‡] JJ was partially supported by NSF grant DMS-1208315 and DMS-1513414.

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.

Google Scholar), it is relatively convenient 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, and community structures) that can be extracted to understand many aspects of the scientific community.

Recent studies on such networks include but are not limited to the following: Grossman (2002) studied the Coauthorship network of mathematicians; Newman (2001a, 2004) and Martin et al. (2013) studied the Coauthorship networks of biologists, physicists and computer scientists; Ioannidis (2008) used the Coauthorship network to help assess the scientific impacts.

Unfortunately, as far as we know, Coauthorship and Citation networks for *statisticians* have not yet been studied. We recognize that people who are most interested in networks for statisticians are statisticians ourselves, and it is the statisticians' task to study our own networks. We also recognize that, as statisticians, we have the advantage of knowing something about many aspects of our own community; such "partial ground truth" can be very helpful in analyzing the networks and in interpreting the results.

With substantial time and efforts, we have collected two 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), *Biometrika*, *Journal of American Statistical Association* (JASA) and *Journal of Royal Statistical Society* (Series B) (JRSS-B).

The data sets provide a fertile ground for research on social networks. 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 community structures of statisticians. The data sets also serve as a starting point for a more ambitious project [Ji, Jin and Ke (2015)], where we collect a network data set that is similar in nature but is much larger: it covers about 30 journals and spans a time period of 40 years.

1.1. *Our findings.*

We have the following findings.

- *(a). Productivity, patterns and trends.* We identify noticeable productivity characteristics and publication patterns/trends for statisticians.
- *(b). Centrality.* We identify "hot" areas, authors who are most collaborative, and authors who are most highly cited.
- *(c). Community detection.* With possibly more sophisticated methods and analysis, we identify meaningful communities for statisticians.

We now discuss the three items separately.

(a). Productivity, patterns and trends. We have found the following.

- Between 2003 and 2012, the number of papers per author has been decreasing (Figure 1). The proportion of self-citations has been decreasing while the proportion of distant citations has been increasing (Figure 4). Possible explanations are: the statistics community has become increasingly more collaborative, competitive, and globalized.
- 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 [Barabasi and Albert (1999); Newman (2001b)].

(b). Centrality. We have 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 cited authors. See Table 2.

We have also identified 14 "hot" papers. See Table 3. 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".

(c). Community detection. Intuitively, communities in a network are groups of nodes that have more edges within than across (note that "community" and "component" are very different concepts); see Jin (2015) for example. The goal of community detection is to identify such groups (i.e., clustering).

We consider the Citation network and two versions of Coauthorship networks. In each of these networks, a node is an author.

- (c1). Coauthorship network (A). In this network, there is an (undirected) edge between two authors if and only if they have coauthored 2 or more papers in the range of our data sets.
- (c2). Coauthorship network (B). This is similar to Coauthorship network (A), but "2 or more papers" is replaced by "1 or more papers".
- (c3). Citation network. There is a (directed) edge from author i to j if author i has cited 1 or more papers by author j.

The first version of Coauthorship network is easier to analyze than the second version, and presents many meaningful research groups that are hard to find. We now discuss the three networks separately.

(c1). Coauthorship network (A). The network is rather fragmented. The giant component can be interpreted as the "High Dimensional Data Analysis (Coauthorship (A))" (HDDA-Coau-A) community, which has 236 nodes and may contain sub-structures; see Section 4.2. The next two largest components (Figure 8) can be interpreted as communities of "Theoretical Machine

Learning" (18 nodes) and "Dimension Reduction" (14 nodes), respectively. The next 5 components (Table 6) can be interpreted as communities of "Johns Hopkins", "Duke", "Stanford", "Quantile Regression", and "Experimental Design", respectively.

(c2). Coauthorship network (B). We have identified three meaningful communities as follows: "Bayes", "Biostatistics (Coauthorship (B))" (Biostat-Coau-B), "High Dimensional Data Analysis (Coauthorship (B))" (HDDA-Coau-B), presented in Figures 9, 10, and 11, respectively.

TABLE 1

The 14 communities introduced in Section 1.1. In Coauthorship Network (A), each community is a component of the network. In Coauthorship Network (B) and Citation Network, the communities are identified by SCORE and D-SCORE, respectively.

Network	Communities	#nodes	Visualization
Coauthor(A)	High-Dimensional Data Analysis (HDDA-Coau-A)	236	Figures 6, 7
	Theoretical Machine Learning	18	Figure 8
	Dimension Reduction	14	Figure 8
	Johns Hopkins	13	Table 6
	Duke	10	
	Stanford	9	
	Quantile Regression	9	
	Experimental Design	8	
Coauthor(B)	Bayes	64	Figure 9
	Biostatistics	388	Figure 10
	High-Dimensional Data Analysis (HDDA-Coau-B)	1181	Figure 11
Citation	Large-Scale Multiple Testing	359	Figure 13
	Variable Selection	1280	Figure 14
	Spatial & Semi-parametric/Non-parametric Statistics	1015	Figure 15

(c3). Citation network. We have identified three communities: "Large-Scale Multiple Testing", "Variable Selection" and "Spatial and semi-parametric/nonparametric Statistics", presented in Figures 13-15 respectively.

We present in Table 1 a road map for the 14 communities we just mentioned (some of these communities have sub-communities; see Sections 4-5). The communities or groups identified in each of the three networks are connected, intertwined, but are also very different. See Sections 5.2.1-5.2.2.

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.

It may be hard to understand why collecting such data is challenging: 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 data, the resources become surprisingly limited. For example, Google Scholar aggressively blocks any one 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 similar problems.

Other challenges we have faced are missing paper identifiers, ambiguous author names, etc.; we explain how we have overcome these in the Appendix.

1.3. *Experimental design and scientific relevance.* We have limited our attention to four journals (AoS, Biometrika, JASA, JRSS-B), which are regarded by many statisticians as the leading methodological journals (with a caveat for JASA applications). We recognize that we may have different results if we include in our data set 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).

Also, in our study, we are 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 here serve well for our targeted scientific problems: they provide many meaningful results in many aspects of our targeted community within the targeted time period. They also prepare us well for a more ambitious project [Ji, Jin and Ke (2015)] where we collect new data sets by downloading papers from about 30 journals in the last 40 years.

1.4. *Disclaimers.* Our primary goal in the paper is to present the data sets we collect, and to report our findings in such data sets. It is not our intention to rank one author/paper over the others. We wish to clarify that "highly cited" is not exactly the same as "important" or "influential". It is not our intention to rank one area over the other either. 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 not exactly an area that is exhausted either.

Also, 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.

Finally, social networks are about "real people" (and this time, "us"). To obtain interpretable results, we have to use real names, but we have not used any data that is not publicly available. The interest of the paper is on the statistics community *as a whole*, not on any individual statistician.

1.5. *Contents.* Section 2 studies the productivity, patterns and trends for statisticians. Section 3 discusses the network centrality. Sections 4-5 discuss community detection for the Coauthorship network and Citation network, respectively. Section 6 contains some discussion, and Section 7 is the Appendix, where we address the challenges in data collection and cleaning.

2. Productivity, patterns and trends.

2.1. *Productivity.* There are 3248 papers and 3607 authors in the data set (an average of .90 paper per author). To investigate how the productivity evolves over the years, we present in Figure 1 the total number of papers published in each year (left panel) and the yearly average productivity (per author)[1]. Over the 10-year period, the number of papers published in each year has been increasing, but the yearly average producibility 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.

It could also be the case that the productivity does not change much, but statisticians are publishing in a wider range of journals, and more younger ones have started making substantial contributions to the field.

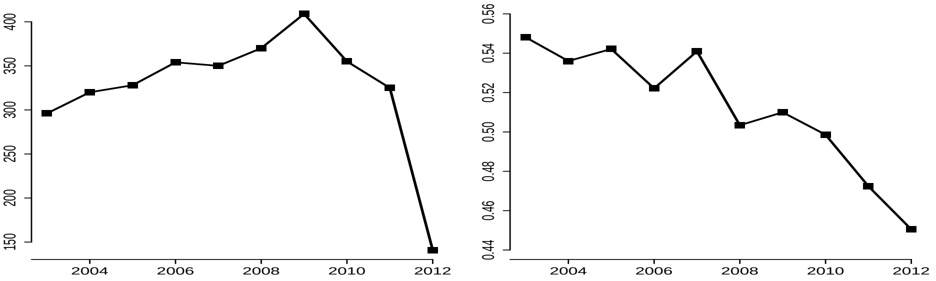

Fig 1. *Left: total number of papers published each year from* 2002 *to* 2012 *(for the year 2012, we have only data for the first half). Right: yearly average productivity per author.*

For any K-author paper, we may count each coauthor's contribution to this particular paper either as "divided" or as "non-divided", where we count every coauthor as has published 1 paper and $1/K$ paper, respectively.

For "non-divided" contribution, we have Figure 2 (left), 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. Figure 2 suggests that the distribution of the number of papers has a power law tail. For "divided"

[1] For each year, this is the ratio of the total number of papers in that year over the total number of authors who published at least once in that year.

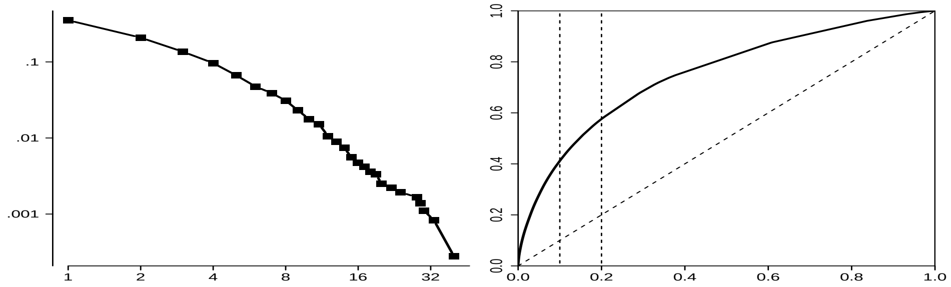

FIG 2. *Left: The proportion of authors who have written more than a certain number of papers (for a better view, both axes are evenly spaced on the logarithmic scale). Right: The Lorenz curve for the number of papers each author with divided contributions.*

contribution, we have the Lorenz curve for the number of papers by each author in Figure 2 (right), which suggests the distribution does not have a power law tail but is still very skewed. For example, the figure shows that the top 10% most prolific authors contribute 41% of the papers. Our findings are similar to that in Martin et al. (2013) for the physics community.

2.2. *Coauthor patterns and trends.* In the coauthorship network, the degrees (i.e., number of coauthors) range from 0 to 65, where Peter Hall (65), Raymond Carroll (55), Joseph Ibrahim (41), Jianqing Fan (38) have the highest degrees. Also, 154 authors have degree 0, and 913 authors have degree 1. The degree distribution (Figure 3, left) suggests a power law tail.

To investigate how the number of coauthors changes over time, we present in Figure 3 (right) the average number of coauthors in each year, where the average number of coauthors is steadily increasing. Again, this suggests that the statistics community has become increasingly more collaborative.

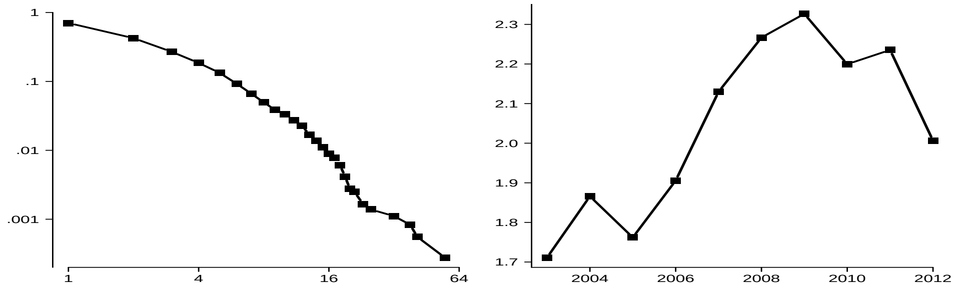

FIG 3. *Left: The proportion of authors with more than a given number of coauthors (for a better view, both axes are evenly spaced on the logarithmic scale). Right: The average number of coauthors for all authors who has published in these journals that year.*

2.3. *Citation patterns and trends.* For the 3248 papers (3607 authors) in our data sets, the average citation per paper is 1.76.[2] 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 set.

The distribution of the in-degree (the number of citations received by each paper) is highly skewed. For example, the top 10% highly cited papers receive about 60% of all citation counts. The Gini coefficient is .77 [Gini (1936)] suggesting that the in-degree is highly dispersed. The Lorenz curve (Figure 4, left) confirms that the distribution of the in-degrees is highly skewed.

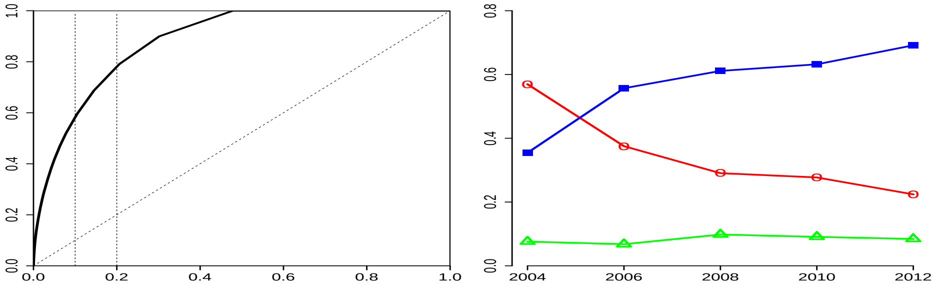

FIG 4. *Left: The Lorenz curve for the number of citation received by each paper. Right: The proportions of self-citations (red circles), coauthor citations (green triangles) and distant citations (blue rectangles) for each two-year block.*

It seems that authors tend to return a favor, 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%.

The overall proportions for self-citations, coauthor citations, and distant citations[3] are 27%, 9%, and 64%, respectively. Moreover, Figure 4 (right panel) suggests that over the 10-year period, the proportions of self-citations, coauthor citations, and distant citations have been slowly decreasing, roughly the same, and slowly increasing, respectively. The last item is a bit unexpected, but may due to that over the years, the publications have become increasingly more accessible. That the blue and red curves cross with each other on the left is probably due to the "boundary effect".[4]

[2] This 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 due to that we count only citations between papers in our data set.

[3] Citations from some one who is not oneself or a coauthor.

[4] Here is an example for boundary effect. For papers published in 2003, most papers

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 closeness centrality is defined as the reciprocal of the total distance to all others [Sabidussi (1966)]. The betweenness centrality measures the extent to which a node is located "between" other pairs of nodes [Freeman, Borgatti and White (1991)].

The degree centrality is conceptually simple, but the definitions vary from case to case. 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 *papers*, the centrality is the in-degree (i.e., the number of papers which cite this paper). For Citation network of *authors*, the centrality of an author is the number of citers (i.e., authors who cite his or her papers).

Table 2 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 2
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 3 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 (2006) (75 citations; adaptive lasso), Meinshausen and Bühlmann (2006) (64 citations; graphical lasso), and Candès and Tao (2007) (49 citations; Dantzig Selector). The three papers are all in a specific sub-area of high dimensional variable selection, where the theme is to extend the penalization methods (e.g., the lasso by Chen, Donoho and Saunders (1998) and Tibshirani (1996)) in various directions[5].

These results suggest "Variable Selection" as one of the "hot" areas. Other

they cite are probably published earlier than 2002 (so beyond the range of our data set).

[5] These fit well with the impression of many statisticians: in the past 10-20 years, there is a noticeable wave of research interest on the penalization approach to variable selection.

TABLE 3

Fourteen "hot" papers (alphabetically) identified by degree centrality (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 and Levina (2008a) (Covariance Estimation)			4
Candès and Tao (2007) (Variable Selection)	3		
Fan and Li (2004) (Variable Selection)		2	
Fan and Lv (2008) (Variable Selection)			1
Fan and Peng (2004) (Variable Selection)	4	1	
Huang et al. (2006) (Covariance Estimation)			3
Huang, Horowitz and Ma (2008) (Variable Selection)			5
Hunter and Li (2005) (Variable Selection)		4	
Johnstone and Silverman (2005) (Empirical Bayes)		5	
Meinshausen and Bühlmann (2006) (Variable Selection)	2		
Storey (2003) (Multiple Testing)		3	
Zou (2006) (Variable Selection)	1		
Zou and Hastie (2005) (Variable Selection)	5		
Zou and Li (2008) (Variable Selection)			2

"hot" areas may include "Covariance Estimation", "Empirical Bayes", and "Large-Scale Multiple Testing"; see Table 3 for details.

For the 30 most cited papers, see `http://faculty.franklin.uga.edu/psji/sites/faculty.franklin.uga.edu.psji/files/top-cited-30.xlsx`. These papers account for 16% of the total number of citation counts. The list further shows that the most highly cited papers are on the penalization approach to variable selection (e.g., adaptive lasso, group lasso).

On the other hand, note that some important and innovative works in the area of variable selection have significantly fewer citations. These include but are not limited to the phenomenal paper by Efron et al. (2004) on least angle regression, which has received a lot of attention from a broader scientific community[6]. A similar claim can be drawn on other areas or topics.

The fact 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.

4. Community detection for Coauthorship networks.

In this section, we study community detection for Coauthorship networks (A) and (B).

4.1. *Community detection methods (undirected networks).*

Community detection is a problem of major interest in network analysis [Goldenberg et al. (2009)]. Consider an *undirected* and *connected* network $\mathcal{N} = (V, E)$

[6] The paper has 4900 citations on Google Scholar, but is only cited 11 times by papers in our data set (in comparison, the adaptive lasso paper Zou (2006) has received 75 citations).

with n nodes. We think V as the union of a few (disjoint) subsets which we call the "communities":

$$V = V^{(1)} \cup V^{(2)} \ldots \cup V^{(K)},^{7}$$

where "$\cup$" stands for the conventional union in set theory (same below). Intuitively, we think communities as subsets of nodes where there are more edges "within" than "across" [e.g., Bickel and Levina (2008b)]. The goal of community detection is clustering: for each $i \in V$, decide to which of the K communities it belongs.

There are many community detection methods for undirected networks. In this paper, we consider the Spectral Clustering approach (NSC) by Newman (2006), the Profile Likelihood approach (BCPL) by Bickel and Chen (2009) and Zhao, Levina and Zhu (2012), the Pseudo Likelihood approach (APL) by Amini et al. (2013), and the SCORE by Jin (2015).

NSC is a spectral method, based on the key observation is that Newman and Girvan's modularity matrix can be approximated by the leading eigenvectors of the matrix. Following Newman (2006), we cluster by using the signs of the first leading eigenvectors when $K = 2$, and use the recursive bisection approach when $K \geq 3$.

BCPL is a penalization method proposed by Bickel and Chen (2009) which uses greedy search to maximize the profile likelihood. For large networks, BCPL may be computationally inefficient. In light of this, Amini et al. (2013) modified BCPL and proposed APL as a new Profile Likelihood approach. APL ignores some dependence structures in the modeling so the resultant profile likelihood has a simpler form and is easier to compute.

SCORE, or **S**pectral **C**lustering **O**n **R**atios of **E**igenvectors, is a spectral method motivated by the recent Degree Corrected Block Model [DCBM, Karrer and Newman (2011)]. SCORE recognizes that, the degree heterogeneity parameters in DCBM are nearly ancillary, and can be conveniently removed by taking entry-wise ratios between the eigenvectors of the adjacency matrix; see Jin (2015). SCORE is a flexible idea and is highly adaptable. 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.

Remark. For different methods, the vectors of predicted labels can be very different. For a pair of the predicted label vectors, we measure the similarity by the Adjusted Rand Index (ARI) [Hubert and Arabie (1985)] and the Variation of Information (VI) [Meila (2003)]; a large ARI or a small VI suggests that two predicted label vectors are similar to each other.

[7] For simplicity, we assume the communities are non-overlapping in this paper.

4.2. *Coauthorship network (A).* In this network, by definition, there is an edge between two nodes (i.e., authors) if and only if they have coauthored 2 or more papers (in the range of our data sets). The network is very much fragmented: the total of 3607 nodes split into 2985 different components, where 2805 (94%) of them are singletons, 105 (3.5%) of them are pairs, and the average component size is 1.2.

The giant component (236 nodes) is seen to be the "High Dimensional Data Analysis (Coauthorship (A))" community (HDDA-Coau-A), including (sorted descendingly by the degree) Peter Hall, Raymond Carroll, Jianqing Fan, Joseph Ibrahim, Tony Cai, David Dunson, Hua Liang, Jing Qin, Donglin Zeng, Hans-Georg Müller, Hongtu Zhu, Enno Mammen, Jian Huang, Runze Li, etc. It seems that the giant component has sub-structures. In Figure 5 (left), we plot the scree-plot of the adjacency matrix associated with 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 for $K \geq 3$, the results by different methods are largely inconsistent with each other: the maximum of ARI and the minimum VI (see the remark in Section 4.1) across different pairs of methods are 0.15 and 1.19, respectively.

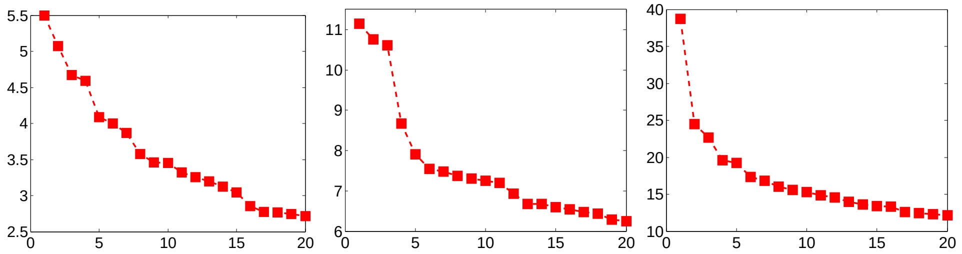

FIG 5. *Scree plots. From left to right: the giant component of Coauthorship network(A), Coauthorship network(B), Citation network (asymmetric; plotted are singular values).*

We now focus on the case of $K = 2$. In Table 4, we present the ARI and VI for each pair of the methods. The table suggests that: the 4 methods split into two groups where SCORE and APL are in one of the groups with an ARI of 0.72 between them, and NSC and BCPL are in the other group with an ARI of 0.21. The results for methods in each group are moderately consistent to each other, but those for methods in different groups are rather inconsistent. That BCPL and APL have rather different results is unexpected, as APL is a variant of BCPL. Possible explanation is that both methods use random starting points; they do not necessary converge even for a long time, and so may produce different results from run to run. See Table 5, which compares the sizes of the communities identified by the 4 methods.

In Figures 6-7, we further compare the community detection results by each of the 4 methods ($K = 2$). In each panel, nodes are marked with either blue circles or red squares, 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.

Comparing the results by different methods, one of the major discrepancies lies in the "Fan" group: SCORE and APL cluster the "Fan" group (with Jianqing Fan being the hub) into the "Carroll-Hall" community, and N-SC and BCPL cluster it into the"North Carolina" community. A possible explanation is that, the "Fan" group has strong ties to both communities. Another explanation is that there are ≥ 3 communities. However, the results by all 4 methods are rather inconsistent if we assume $K \geq 3$; see discussions before. How to obtain a more convincing explanation is an interesting but challenging problem. We omit further discussions for reasons of space.

Table 4

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

Other noteworthy discrepancies are as follows:

- SCORE includes the "Dunson" branch in the "North Carolina" group, but APL 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 "Dunson" branch and two small branches connecting to Jianqing Fan. In comparison, the results by NSC seem more meaningful.

Table 5
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

Moving away from the giant component, the next two largest components are the "Theoretical Machine Learning" group (18 nodes) and the "Dimension Reduction" group (14 nodes); see Figure 8. The first one is a research group who work on Machine Learning topics using sophisticated statistical theory, including Peter Bühlmann, Alexandre Tsybakov, Jon Wellner, and Bin Yu. The second one is a research group on Dimension Reduction, including Francesca Chiaromonet, Dennis Cook, Bing Li and their collaborators.

A conversation with Professor Qunhua Li (Statistics Department at Penn State) helps to illuminate why these groups 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 Wellner 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 (Peter Bühlmann, Markus Kalische, Sara van de Geer). Nocolai Meinshausen could be one of the "bridging nodes" between ETH and Berkeley: he was a Ph.D student of Peter Bühlmann and then a postdoc at Berkeley. In the second group, Ms. Chiaromonet obtained her Ph.D from University of Minnesota, where Dennis Cook served as the supervisor. She then went on to work in the Statistics Department at Pennsylvania State University, and started to collaborate with Bing Li on Dimension Reduction.

The next 5 largest components in Coauthorship network (A) are the "Johns Hopkins" group (13 nodes; including faculty at Johns Hopkins University and their collaborators; similar below), "Duke" group (10 nodes; including Mike West, Jonathan Stroud, Carlos Caravlaho, etc.), "Stanford" 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 groups are presented in Table 6.

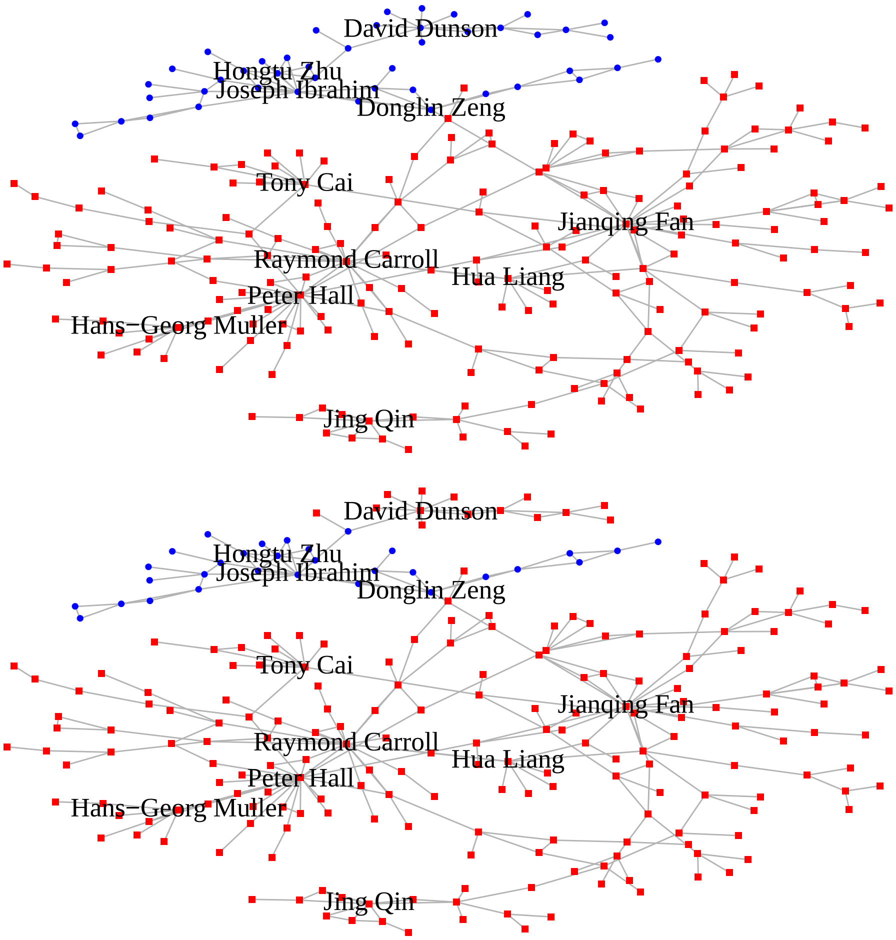

FIG 6. *Community detection results by SCORE (top) and APL (bottom) for the giant component of Coauthorship network (A), assuming $K = 2$. Nodes in blue circles and red squares represent two different communities.*

4.3. *Coauthorship network (B).* In this network, 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 has 2263 nodes (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.

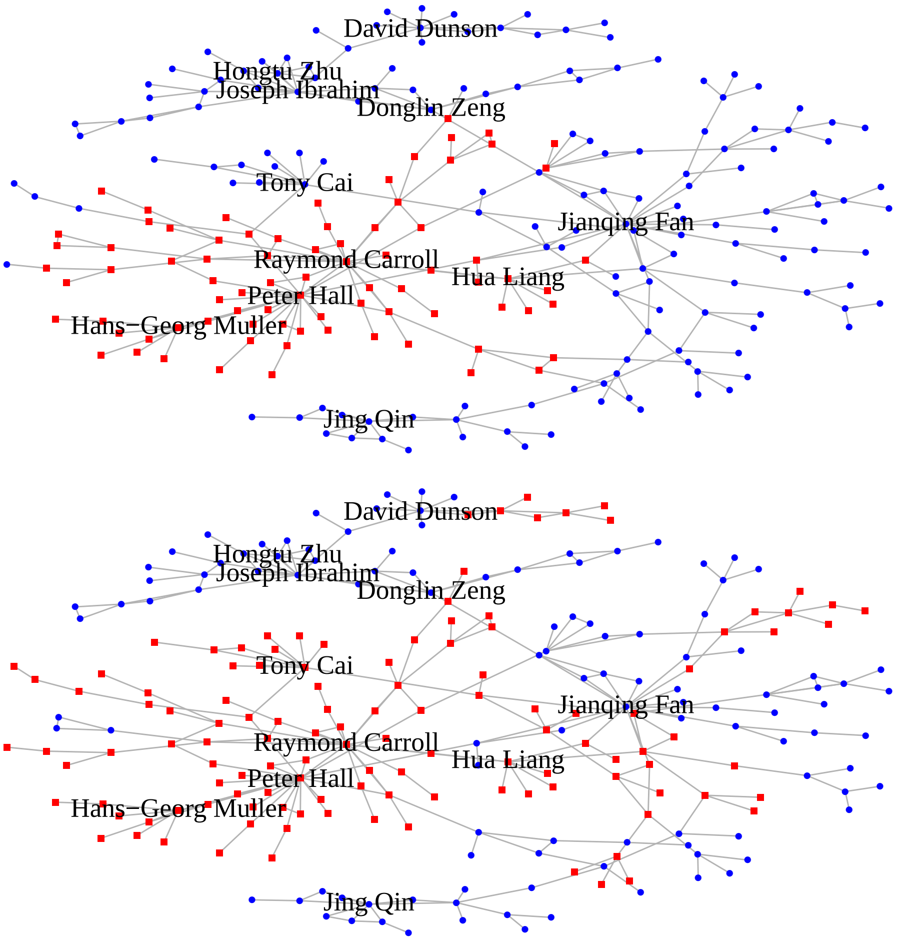

FIG 7. *Community detection results by NSC (top) and BCPL (bottom) for the giant component of Coauthorship network (A), assuming $K = 2$. Nodes in blue circles and red squares represent two different communities.*

Figure 5 (middle) presents the scree plot for the adjacency matrix of Coauthorship network (B), suggesting 3 or more communities. Assuming $K = 3$, we apply SCORE, NSC, BCPL, and APL, and below are the findings.

First, somewhat surprisingly, the results 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. See Table 7, where we compare all 4 methods pair-wise and

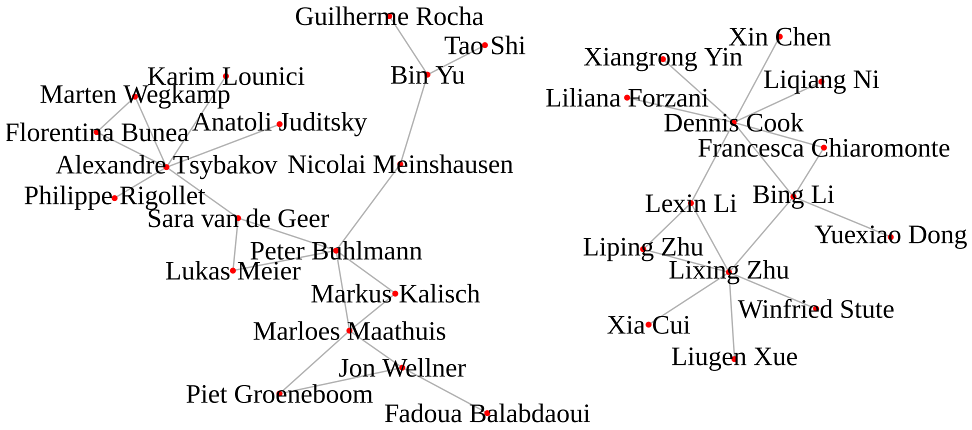

Fig 8. *The second largest (left) and third largest (right) components of Coauthorship network (A). They can be possibly interpreted as the "Theoretical Machine Learning" and "Dimension Reduction" communities, respectively.*

Table 6

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 Müller	
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	

tabulate the corresponding ARI and VI (see Remark 2).

Second, the results by SCORE, NSC, and APL are reasonably consistent with each other: the ARI between the vector of predicted labels by SCORE

TABLE 7

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

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 7 for details. In particular, the three methods seem to agree on that there are three communities which can be interpreted as follows (arranged ascendingly in size).

- "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.
- "Biostatistics (Coauthorship (B))" (Biostat-Coau-B) community. The sizes of three versions of this community (corresponding to three methods) are quite different and range from 50 to 388. While it is probably not exactly accurate to call this community "Biostatistics", the community consists of a number of statisticians and biostatisticians in the Research Triangle Park of North Carolina. It also includes many statisticians and biostatisticians from Harvard University, University of Michigan at Ann Arbor, University of Wisconsin at Madison.
- "High Dimensional Data Analysis (Coauthorship (B))" (HDDA-Coau-B) community. The sizes of this community identified by three different methods range from 1811 to 2193. The community includes researchers from a wide variety of research areas in or related to high dimensional data analysis (e.g., Bioinformatics, Machine Learning).

Figures 9-11 present these 3 communities (by SCORE) respectively.

In Table 8, 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

Table 8

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.

	Bayes	Biostat-Coau-B	HDDA-Coau-B
SCORE	64	388	1811
NSC	68	163	2032
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

estimate of the "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 those by APL, SCORE, and NSC, respectively). A similar claim applies to the Biostat-Coau-B community identified by each of the methods (sizes are 50, 388, and 169 for those by APL, SCORE, and NSC, respectively). This suggests that APL may have underestimated these two communities but overestimated the HDDA-Coau-B community.[8]

It is also interesting to compare these results with those we obtain in Section 4.2 for Coauthorship network (A). Below are three noteworthy points.

First, recall that in Figure 8 and Table 6, 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 6) 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

[8] In Column 1 of Table 8, the authors in "SCORE ∩ NSC \ APL" are mostly Bayesian statisticians, including Steven MacEachern, Alan Gelfand, Bruno Sanso, Gary Rosner, Nicholas Polson, Herbert Lee, Edward George, etc. In Column 2 of Table 8, the authors in the subset of "SCORE ∩ NSC \ APL" are mostly biostatisticians including Trivellore Raghunathan, Jun Liu, L J Wei, Louise Ryan, Ram Tiwari, Joseph Lucas, Nathaniel Schenker, etc.

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 "Bayes" community where James Berger is a hub node.

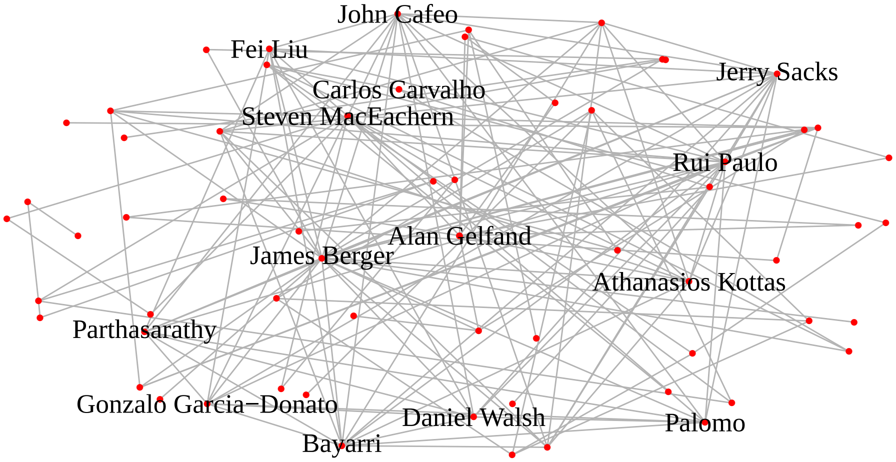

FIG 9. *The "Bayes" community in Coauthorship network (B) identified by SCORE (*64 *nodes). Only names for* 14 *nodes with a degree of* 9 *or larger are shown.*

5. Community detection for Citation network. The Citation network is a directed network. As a result, the study in this section is very different from that in Section 4, and provides additional insight.

5.1. *Community detection methods (directed networks).* In the Citation network, each node is an author and there is a directed edge from node i to node j if and only if node i has cited node j at least once. To analyze the Citation network, one usually focuses on the *weakly connected giant component.*[9] From now on, when we say the Citation network, we mean the weakly connected giant component of the original Citation network.

For community detection of directed networks, we consider two methods: LNSC and Directed-SCORE (D-SCORE). See the remark in Section 5.2.3 for discussions on other methods.

[9] I.e., the giant component of the *weakly connected citation network*, where there is an (undirected) edge between nodes i and j if one has cited the other at least once [Bang-Jensen and Gutin (2009)].

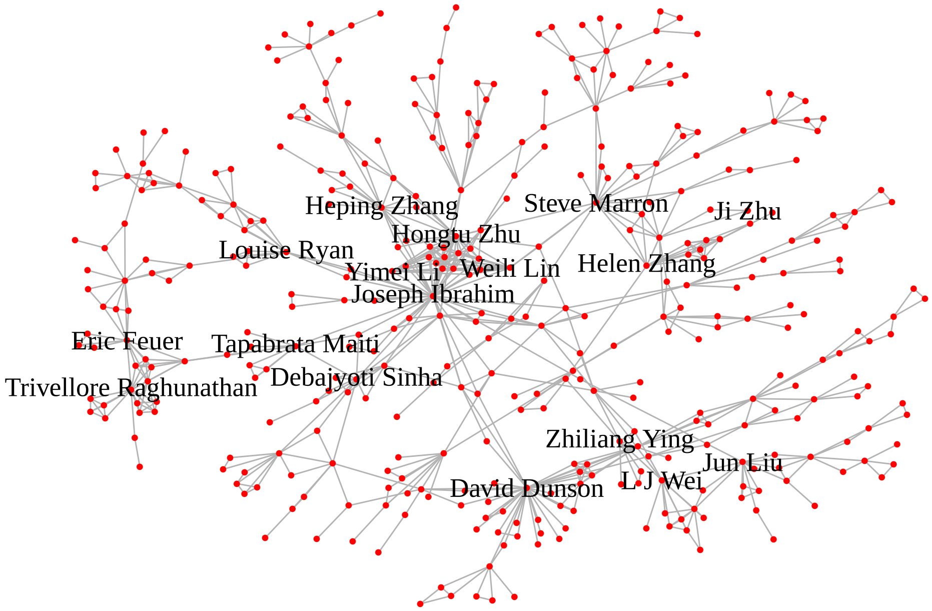

FIG 10. *The "Biostatistics" community (Biostat-Coau-B) in Coauthorship network (B) identified by SCORE (388 nodes). Only names for 17 nodes with a degree of 13 or larger are shown. A "branch" in the figure is usually a research group in an institution or a state.*

LNSC stands for the Spectral Clustering approach proposed in Leicht and Newman (2008): the authors extended the spectral modularity methods by Newman (2006) for undirected networks to directed networks, using the so-called generalized modularity [Arenas et al. (2007)]. However, it is pointed out in Kim, Son and Jeong (2010) 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 adaption of SCORE [Jin (2015)] (see Section 4.1) to directed networks. Let A be the adjacency matrix, and let $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_K$ and $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$ be the first K left singular vectors and the first K right singular vectors of A, respectively. Also, let $\mathcal{N}_1$ be the support of $\hat{u}_1$ and $\mathcal{N}_2$ be the support of $\hat{v}_1$. Define two $n \times (K-1)$ matrices $\hat{R}^{(l)}$ and $\hat{R}^{(r)}$ by

$$(5.1) \quad \hat{R}^{(l)}(i,k) = \begin{cases} \mathrm{sgn}(\hat{u}_{k+1}(i)/\hat{u}_1(i)) \cdot \min\{|\frac{\hat{u}_{k+1}(i)}{\hat{u}_1(i)}|, \log(n)\}, & i \in \mathcal{N}_1, \\ 0, & i \notin \mathcal{N}_1, \end{cases}$$

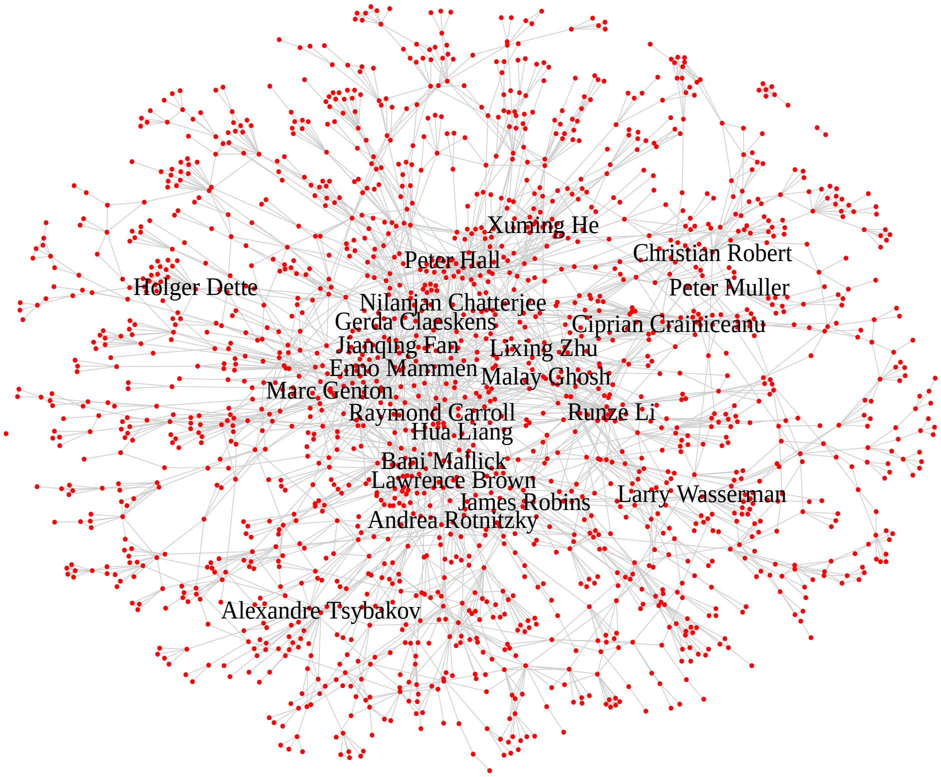

FIG 11. *The "High Dimensional Data Analysis" community (HDDA-Coau-B) in Coauthorship network (B) identified by SCORE (*1181 *nodes). Only names for* 22 *nodes with degree of* 18 *or larger are shown.*

$$(5.2)\quad \hat{R}^{(r)}(i,k) = \begin{cases} \mathrm{sgn}(\hat{v}_{k+1}(i)/\hat{v}_1(i)) \cdot \min\{|\frac{\hat{v}_{k+1}(i)}{\hat{v}_1(i)}|, \log(n)\}, & i \in \mathcal{N}_2, \\ 0, & i \notin \mathcal{N}_2. \end{cases}$$

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. $(\mathcal{N}_1 \cap \mathcal{N}_2)$. Restricting the rows of $\hat{R}^{(l)}$ and $\hat{R}^{(r)}$ to the set $\mathcal{N}_1 \cap \mathcal{N}_2$ and obtaining 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. $(\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2))$. In Step 1-3, 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.

We don't need a sophisticated clustering method for nodes in $\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2)$ as we assume this set is small. See Section 5.2 for an example.

Figure 12 illustrates how D-SCORE works using the statistical citation network data set with $K = 3$. Two panels show similar clustering patterns, suggesting that there are three communities; see Section 5.2 for details.

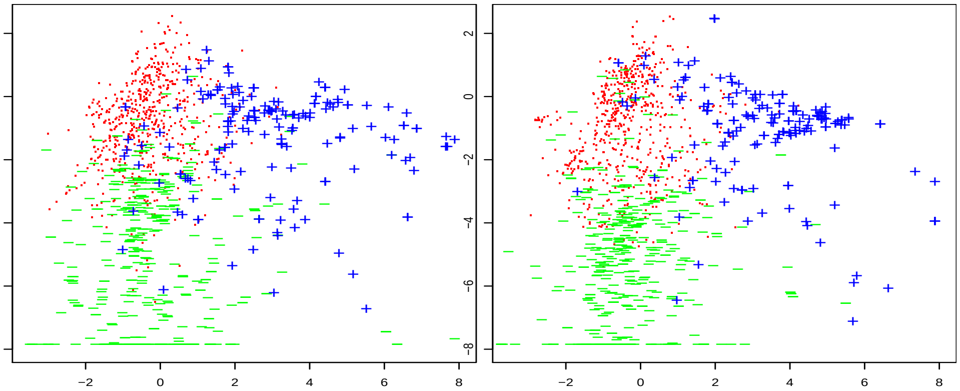

FIG 12. *Left: each point represents a row of $\hat{R}^{(l)}$ (the matrix has only two columns since $K = 3$) associated with the statistical Citation network (x-axis: first column, y-axis: second column). Only rows with indices in $\mathcal{N}_1$ are shown. Blue pluses, green bars, and red dots represent 3 different communities identified by SCORE, which can be interpreted as "Large-Scale Multiple testing", "Spatial and Semi-parametric/Nonparametric Statistics" and "Variable Selection", Right: similar but with $(\hat{R}^{(l)}, \mathcal{N}_1)$ replaced by $(\hat{R}^{(r)}, \mathcal{N}_2)$.*

5.2. *Community detection of the Citation network by D-SCORE.*

The original citation network data set has 3607 nodes (i.e., authors). The associated weakly connected network has 927 components. The giant component has 2654 nodes, and all other components have no more than 5 nodes.

We restrict our attention to the weakly connected giant component $\mathcal{N} = (V, E)$. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be as in defined in Section 5.1. We have $|\mathcal{N}_1| = 2126$,

$|\mathcal{N}_2| = 1790$, $|\mathcal{N}_1 \cap \mathcal{N}_2| = 1276$, and $|\mathcal{N} \setminus (\mathcal{N}_1 \cup \mathcal{N}_2))| = 14$. Let A be the adjacency matrix of $\mathcal{N}$. Figure 5 (right) presents the scree plot of A. The plot suggests that there are $K = 3$ communities in $\mathcal{N}$.

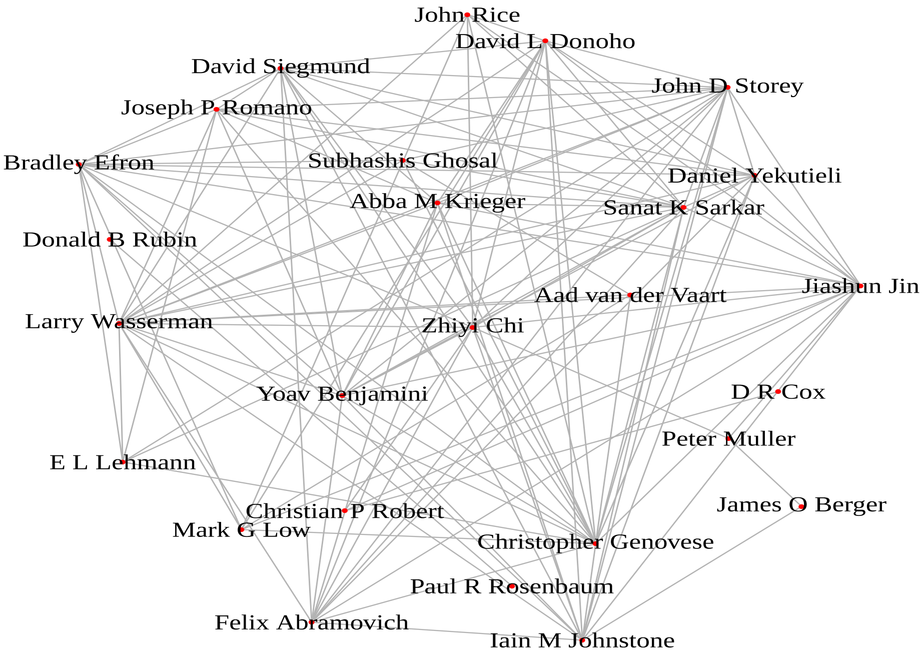

FIG 13. *The "Large-Scale Multiple Testing" community identified by D-SCORE ($K = 3$) in the Citation network (359 nodes). Only 26 nodes with 24 or more citers are shown.*

We now present the results by D-SCORE. The results of LNSC are very different and are only discussed briefly in Section 5.2.3). Assuming $K = 3$, D-SCORE identifies three communities as follows.

- "Large-Scale Multiple Testing" community (359 nodes). This consists of researchers in multiple testing and control of False Discovery Rate. It includes a Bayes group (James Berger, Peter Müller), three Berkeley-Stanford groups (Bradley Efron, David Siegmund, John Storey; David Donoho, Iain Johnstone, Mark Low,[10] John Rice; Erich Lehmann, Joseph Romano), a Carnegie Mellon group (e.g., Christopher Genovese, Jiashun Jin, Isabella Verdinelli, Larry Wasserman), a

[10] He and Abba Krieger below are at University of Pennsylvania

Causal Inference group (Donald Rubin, Paul Rosenbaum), and a Tel Aviv group (Felix Abramovich, Yoav Benjamini, Abba Krieger, Daniel Yekutieli).

- "Variable Selection" community (1280 nodes). This includes (sorted descendingly by the number of citers) Jianqing Fan, Hui Zou, Peter Hall, Nicolai Meinshausen, Peter Bühlmann, Ming Yuan, Yi Lin, Runze Li, Peter Bickel, Trevor Hastie, Hans-Georg Müller, Emmanuel Candès, 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.
- "Spatial and Semi-parametric/Nonparametric Statistics" (for short, "Spatial Statistics") community (1015 nodes). See discussions below.

The first two communities are presented in Figures 13 and 14, respectively. The last community is harder to interpret and seems to contain sub-structures. For further investigation, we first restrict the network to this community (i.e., ignoring all edges to/from outside) and obtain a sub-network. We then apply D-SCORE with $K = 3$ to the giant component (908 nodes) of this sub-network, and obtain three meaningful sub-communities as follows.

- Non-parametric spatial/Bayes statistics (212 nodes), including David Blei, Alan Gelfand, Yi Li, Steven MacEachern, Omiros Papaspiliopoulos, Trivellore Raghunathan, Gareth Roberts.
- Parametric spatial statistics (304 nodes), including Marc Genton, Tilmann Gneiting, Douglas Nychka, Anthony O'Hagan, Adrian Raftery, Nancy Reid, Michael Stein.
- Semi-parametric/Non-parametric statistics (392 nodes), including Raymond Carroll, Nilanjan Chatterjee, Ciprian Crainiceanu, Joseph Ibrahim, Jeffrey Morris, David Ruppert, Naisyin Wang, Hongtu Zhu.

These sub-communities are presented in Figure 15.

5.2.1. *Comparison with Coauthorship network (A).* In Section 4.2, we present 8 different components of Coauthorship network (A). In Table 9, 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 "Spatial Statistics and Semi-parametric/Non-parametric statistics", including (sorted descendingly by the number of citers; same below) Raymond Carroll, Joseph Ibrahim, Naisyin Wang, Alan Gelfand, Jeffrey Morris, Marc Genton, Sudipto

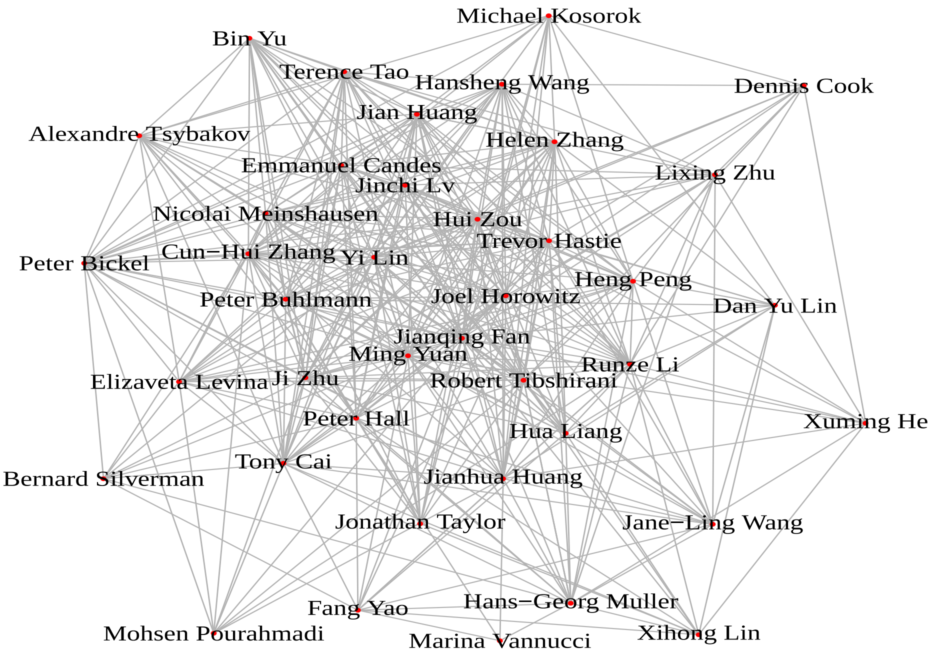

FIG 14. *The "Variable Selection" community identified by D-SCORE ($K = 3$) in the Citation network (1280 nodes). Only 40 nodes with 54 or more citers are shown here.*

TABLE 9

*Sizes of the intersections of the communities identified by D-SCORE ($K = 3$) in the Citation network (rows) and the 8 largest components of Coauthorship network (A) as presented in Section (columns). "Other": nodes outside the weakly connected giant component; *: 9 out of 12 are in the "Semi-parametric/Non-parametric" sub-community of the "Spatial Statistics" community.*

		Mach.	Dim.	Johns			Quant.	Exp.
	giant	Learn.	Reduc.	Hopkins	Duke	Stanford	Reg.	Design
Spatial	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

Banerjee, Hongtu Zhu, Jeng-Min Chiou, Ju-Hyun Park, Ulrich Stadtmuller, Ming-Hui Chen, Yi Li, 166 nodes in "Variable Selection" including Jianqing Fan, Hui Zou, Peter Hall, Ming Yuan, Yi Lin, Runze Li, Trevor Hastie, Hans-

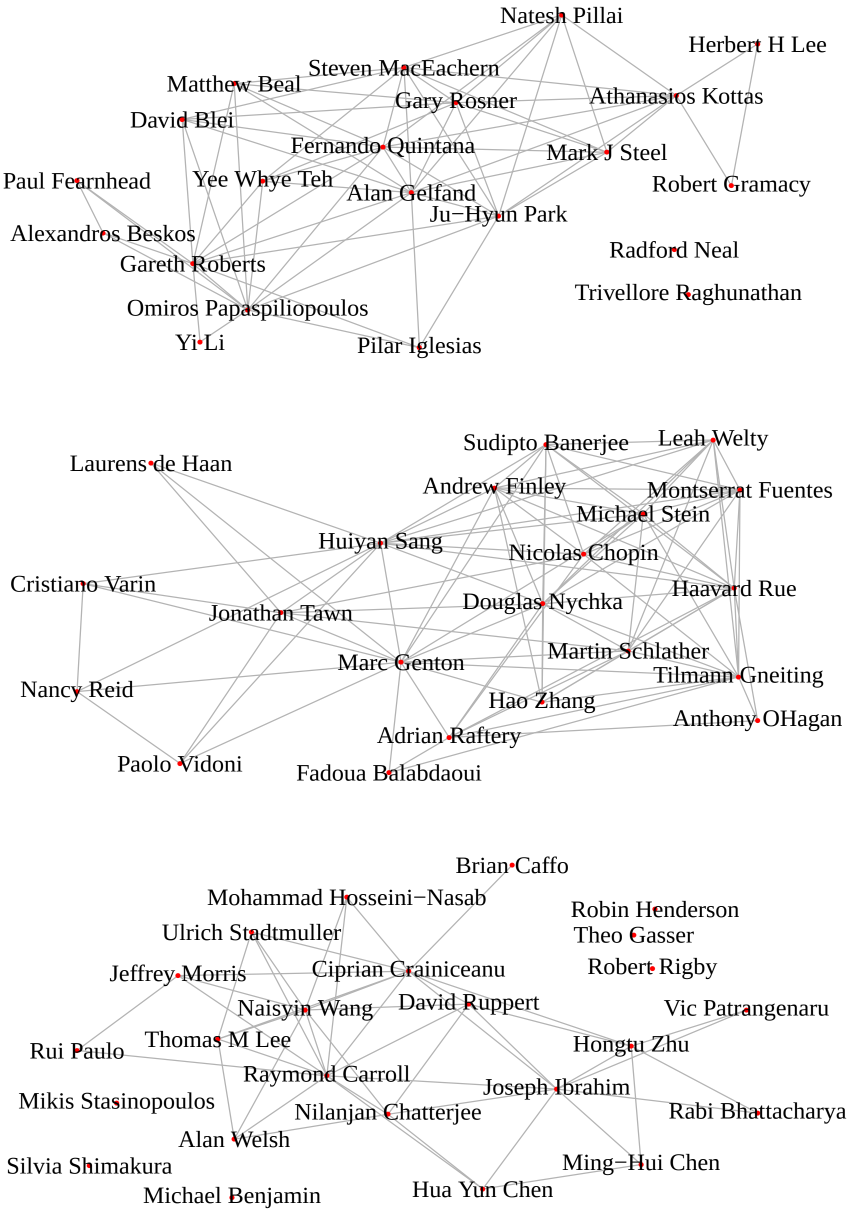

FIG 15. *The "Spatial and Semi-parametric/Non-parametric Statistics" community has sub-communities: Non-parametric Spatial/Bayes (upper), Parametric Spatial (middle), Semi-parametric/Non-parametric (lower). In each, only about 20 high-degree nodes are shown.*

Georg Müller, Emmanuel Candès, 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" group and a "North Carolina" community: The "Carroll-Hall" group has strong ties to the area of variable selection, and the "North Carolina" group has strong ties to Biostatistics. Raymond Carroll has close ties to both of the two groups: it is not surprising that S-CORE assigns him to the "Carroll-Hall" group (Coauthorship network (A)) but D-SCORE assigns him to the "Spatial" community (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 "Spatial Statistics". The "Experimental Design" group has no stronger relation to one area than to the others, so the nodes spread almost evenly to these three communities.

5.2.2. *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 4.3. For the former, we have been focused on the weakly connected giant component of the Citation network (2654 nodes). 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 10 (for each of the 16 cells, the complete name list can be found at http://faculty.franklin.uga.edu/psji/sites/faculty.franklin.uga.edu.psji/files/Table10_Expanded.zip).

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 "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, etc. The second part consists of 25% of the nodes, and is 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, etc.

Table 10

*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). *:* 14 *and* 17 *are in the "Non-parametric Spatial/Bayes" and "Semi-parametric/Non-parametric" sub-communities of the "Spatial and Semi-parametric/Non-parametric Statistics" community, respectively.*

	Bayes	Biostat-Coau-B	HDDA-Coau-B	other	
Spatial	35*	156	462	362	1015
Var. Selection	16	153	837	274	1280
Multiple Tests	6	17	221	115	359
other	7	62	291	593	953
	64	388	1811	1344	3067

The results are reasonable for many nodes in the second part (e.g., Carlos Carvalho, 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, etc. 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, etc. The results are quite reasonable: many nodes in the second part (e.g., Dan Yu Lin, David Dunson, Helen Zhang, Steve Marron, Ji Zhu, 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, etc. The second part has 840 nodes (46%), including high-degree nodes such as Jianqing Fan, Hui Zou, Peter Hall, Nicolai Meinshausen, Peter Bühlmann, etc. 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, etc.

Respectively, the three parts are labeled as subsets of the "Spatial and Semi-parametric/Non-parametric Statistics", "Variable Selection", and "Large-Scale Multiple Testing" communities in the Citation network. This seems convincing: (a) most of the nodes in the first part have a strong interest in spatial statistics or 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.

5.2.3. *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 may be 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 Müller, Xihong Lin, Fang Yao, Naisyin Wang, Marina Vannucci, etc.
- “High Dimensional Data Analysis” (HDDA-Cita-LNSC) (615 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 Bühlmann, Ming Yuan, Yi Lin, Iain Johnstone, Runze Li, Peter Bickel, Trevor Hastie, Larry Wasserman, Emmanuel Candès, Cun-Hui Zhang, Heng Peng, Bradley Efron, etc.
- “Biostatistics” (Biostat-Cita-LNSC) (1605 nodes). The community is hard to interpret and includes researchers from several different areas. For example, it includes researchers in biostatistics (e.g., Joseph Ibrahim, L J Wei), in nonparametric (Bayes) methods (e.g., Peter Müller, David Dunson, and Nils Hjort, Fernando Quintana, Omiros Papaspiliopoulos), and in spatial statistics and uncertainty quantification (e.g., Mac Genton, Tilmann Gneiting, Michael Stein, Hao Zhang).

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" (1280 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 "Spatial" community identified by D-SCORE (1015 nodes), and hard to interpret.

Our observations here somehow agree with Kim, Son and Jeong (2010) on that LNSC can not properly distinguish the directions of the edges and can not detect communities representing directionality patterns among the nodes.

Remark. There are some other approaches to community detection for directed networks. One possibility is classical hierarchical approach, but the challenge there is how to cut the clustering tree and how to interpret the results [Newman (2004)]. The other possibility is the EM approach by Newman and Leicht (2007). However, as pointed out by Ramasco and Mungan (2008) that this approach fails to detect obvious community structures if there are some nodes with zero out-degree or zero in-degree (this is the case for our data set as many junior researchers have no citations within the range of our data set). For reasons of space, we omit further discussions.

6. Discussions.

We have collected, cleaned, and analyzed a data set for the network of statisticians. We have investigated the productivity, patterns and trends, centrality, and community structures for the statisticians with many different tools, ranging from Exploratory Data Analysis [EDA; Tukey (1977)] tools to rather sophisticated methods. Some of these tools are relatively recent (e.g., SCORE, NSC, BCPL, APL, LNSC), and some are even new (e.g., D-SCORE for directed networks). We have presented an array of interesting results. For example, we have identified the "hot" authors and papers, and about 15 meaningful communities such as "Spatial Statistics", "Dimension Reduction", "Large-Scale Multiple Testing", "Bayes", "Quantile Regression", "Theoretical Machine Learning", and "Variable Selection".

The paper 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, Biometrika, JASA, and JRSS-B 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 working on methodology and theory; see Section 1.3. They also serve as a good starting point for a more ambitious project [Ji, Jin and Ke (2015)] where we are collecting and cleaning 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 the 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 algorithm in this paper without careful analysis. 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 interpretation. It is therefore of interest to compare the weaknesses and strengths of these methods theoretically.

Third, there are many other interesting problems we have not addressed here: the issue of mixed membership, link prediction, relationship between citations and professional 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.

Last but not the least, coauthorship and citation networks only provide limited information for studying the research habits, trends, topological patterns, etc. of the statistical community. There are more informative approaches (say, using other information of the paper: abstract, author affiliations, key words, or even the whole paper) to studying such characteristics. Such study is beyond the scope of the paper, so we leave it to the future.

7. Appendix: Data collection and cleaning.

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 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 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; see Section 1.2. 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 amount 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, which 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”.

To solve this problem, 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 carefully documented the data files and R codes that produced the results in our paper. The data files include the raw and cleaned bibtex files for all papers in the range of our study, and also the author lists, paper lists and adjacency matrices, etc. These files (with detailed instructions) can be found at http://faculty.franklin.uga.edu/psji/sites/faculty.franklin.uga.edu.psji/files/SCC2015.zip.

Acknowledgements. JJ thanks David Donoho and Jianqing Fan; the paper was inspired by a lunch conversation with them in 2011 on H-index. The authors thank Arash Amini, Stephen Fienberg, Qunhua Li, Douglas Nychka, and Yunpeng Zhao for helpful pointers.

References.

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.

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

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

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

Bickel, P. and Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Nat. Acad. Sci. USA* **106** 21068-21073.

Bickel, P. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227.

Bickel, P. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313–2351.

Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499.

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)

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70** 849–911. MR2530322

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194 (2005g:62047)

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.

Gini, C. (1936). On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series* **208** 73-79.

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.

Grossman, J. (2002). The evolution of the mathematical research collaboration graph. *Congressus Numerantium* **158** 201-212 .

Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613.

Huang, J., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classif.* **2** 193-218.

Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

Ioannidis, J. (2008). Measuring co-authorship and networking-adjusted scientific impact. *PLOS ONE* **3**.

Ji, P., Jin, J. and Ke, Z. (2015). Social networks for statisticians, new data and new perspectives *Manuscript*.

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

Johnstone, I. and Silverman, B. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752. MR2166560

Karrer, B. and Newman, M. (2011). Stochastic blockmodels and community structures in network. *Phys. Rev.* **83** 1436–1462.

Kim, Y., Son, S.-W. and Jeong, H. (2010). Finding communities in directed networks. *Phys. Rev. E* **81** 016103.

Leicht, E. and Newman, M. (2008). Community structure in directed networks. *Phys. Rev. Lett.* **100** 118703.

Martin, T., Ball, B., Karrer, B. and Newman, M. (2013). Coauthorship and citation patterns in the Physical Review. *Phys. Rev. E* **88**.

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.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462.

Newman, M. (2001a). The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98** 404-409.

Newman, M. (2001b). Scientific collaboration networks. I. Network construction and fundamental results. *Phys. Rev. E* **64** 016131.

Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci. USA* **101** 5200-5205.

Newman, M. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577-8582.

Newman, M. E. J. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proc. Natl. Acad. Sci. USA* **104** 9564-9569.

Ramasco, J. J. and Mungan, M. (2008). Inversion method for content-based networks. *Phys. Rev. E* **77** 036122.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika* **31** 581–683.

Storey, J. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398 (2004k:62055)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley.

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.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320. MR2137327

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