Mathematical Introduction For Data Analysis: Mini Project

Wanshan Li[*], Yuanzhi Li[*] ,Yuze Zhou[*]

[*]1300010695
[*]1300010682
[*]1300010669
[**]School of Mathematical Sciences, Peking University

2016.5

Contents

1 Introduction

We choose the data set about statisticians collected by Prof. Jiashun Jin to analyse in this mini-project. The data set contains 3607 authors and their 3248 papers published from 2003 to 2012 on the Annals of Statistics, the Journal of American Statistical Association, the Journal of Royal Statistical Society-series B and the Biometrika.

Our goal is to analyze the inner structure of the statisticians' community and the papers' set. More generally, we want to explore how to analyze the inner structure of a network with interactions between nodes.

From this data set we can obtain three types of data matrix(of interest), the coauthorship matrix, the paper-paper citation matrix, and the author-author citation matrix. These other three matrix give the adjacency matrix for a network or graph. The coauthorship matrix represents an undirected graph and the other two represent directed graph.

We intend analyzing the structure of the point set of statisticians and the network of papers and statisticians. Basically we use PCA and MDS to achieve our goal. Furthermore, considering the particular features of our data set, such as sparsity and high dimension, we try to apply some advanced methods such as sparse PCA, MDS with uncertainty and dimension reduction via random projection. However, due to the high dimension of the matrix, the first few principal components are a little small. Hence we use parallel analysis [2] to confirm the signals and noises in our principal components. We also have tried to enhance the signals.

To shed light on some deeper structures, we apply page-rank algorithm [6] to the graphs and do some clustering and hierarchical analysis on the data set. The page-rank algorithm can show the importance or influence these statisticians in the field in a way. The clustering and hierarchical analysis give a brief description of the papers and statisticians' community via clustering and finding exemplars. The results are reasonable, but still need to be improved.

Then to give detailed description of the statisticians' community, we apply some method of community detection on the coauthorship and author-author citation matrix. Several methods are briefly compared and we use the commonly preferred method Walktrap [8] and Louvain [1] to analyse. This topic has been discussed in Prof. Jin's work [5], and our result is a little different from his.

Furthermore, we extract the Chinese authors in the data set and analyze the features of this subset. We also anlayze the yearly change of the structure of the Chinese statisticians' community. The MDS and pagerank on this subgroup show us some new interesting things. By analyzing the yearly change of the this subgraph, we find some factors that influence the whole community's structure.

We try many method and analyze the data set in several different prospectives in this article, and the final goal is to describe the inner structure of these papers and statisticians while learn how to analyze similar network. During the whole process we find some points behind which there are some deeper theoretical results, such as Perron's of non-negative matrix and Fiedler's theory of adjacency matrix, and we will point out them when necessary.

2 Normal PCA and MDS

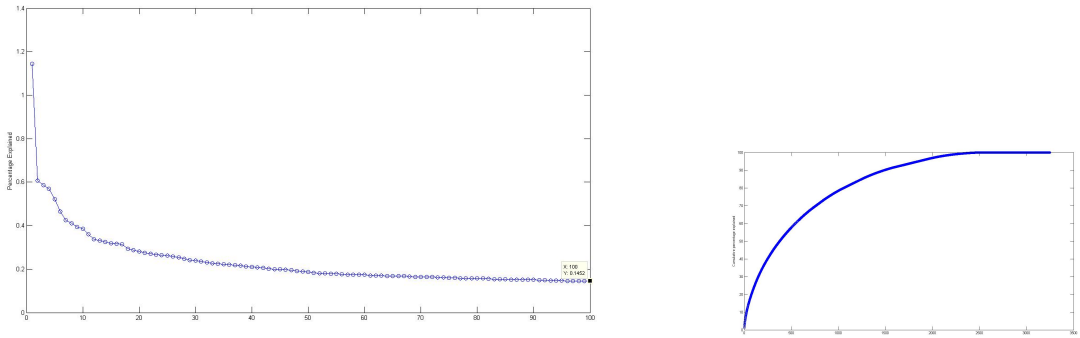
2.1 PCA for the Original Author-Paper Matrix

In this section, we apply normal PCA to the author-paper network, expecting to see how the authors are related with each other in writing academic papers. Furthermore, we found that if we multiply this matrix with its transpose, namely

$$authorpaperbiadj \times authorpaperdiadj^{T}$$

, we will get the co-authorship matrix, whose (i, j)th element represents the number of papers the ith author and the j author in the author list has cooperated on. It is obvious that PCA of the author-paper matrix is identical to that of the co-authorship matrix.

The scree plots of the top 100 principal components (12(a)) shows that the explanatory ability of the top principal components are somehow limited, the top 10 components only make for about 5% of the total variation, even the top 100 components explained only 0.23 of the total variation. And from the other plot, we can see a smooth curve showing the cumulative percentage of variation explained, indicating there are considerable variation occurs in nearly all the directions, thus normal PCA may not perform very well. The reason for such phenomenon may be that many statisticians have published only a few papers in the top statistical journals and co-operated with only a small group of other authors, the data of such authors may have worked as unnecessary noises in the data and made the major components not that distinguishable. Although each one of these authors accounts little for the variability explained, the total effects of them may be great.



(a) Scree plot(only the top 100 components) (b) Cumulative percentage explained

Figure 1: Scree plots of PCA

2.1.1 Authors of the first principal component

By sorting the authors according to the first eigenvector by a ascending order, we can find out the authors who get the lowest score in the first principal component, namely, the smallest values in the first eigenvector.

author	score
Peter Hall	-0.9553
Aurore Delaigle	-0.1711
Raymond J Carroll	-0.1311
Hans-Georg Muller	-0.1116
Qiwei Yao	-0.0530
Tapabrata Maiti	-0.0501
Yanyuan Ma	-0.0419
Jiashun Jin	-0.0403
Fang Yao	-0.0381
Peihua Qiu	-0.0376

There is a significant rise in value at the fifth author,**Qiwei Yao**, indicating that only the top fourth authors matters, which are **Peter Hall**, **Aurore Delaigle**, **Raymond J Carroll** and **Hans-Georg Muller**. The biggest ten values of the first eigenvector is comparably very small in absolute value, even the biggest among them is at around 10^{-16}, so the authors get the highest score according to the first eigenvector are not taken into consideration.

2.1.2 Authors of the second principal component

The authors of the second principal component is explored in the same way as the first principal component. However, different from the first one, effects of both the top part and the bottom part of the second eigenvector are very significant.The results are shown below:

author	score
Peter Hall	-0.1403
Hans-Georg Muller	-0.0682
Fang Yao	-0.0253
T Tony Cai	-0.0210
Jane-Ling Wang	-0.0146
Tapabrata Maiti	-0.0133
Jiashun Jin	-0.0118
Jing-Hao Xue	-0.0100
Hugh Muller	-0.0096
Yao-Ban Chan	-0.0096

author	score
Raymond J Carroll	0.7749
Jianqing Fan	0.4536
Yanyuan Ma	0.2002
Naisyin Wang	0.1167
Hua Liang	0.0916
Enno Mammen	0.0902
Arnab Maity	0.0869
Aurore Delaigle	0.0806
Jiancheng Jiang	0.0759
Xihong Lin	0.0769

According to the results in the tables, scores of authors whose scores are the smallest come to a rise at the fifth author, **Jane-Ling Wang**. **Peter Hall**, **Hans-Georg Muller**, **Fang Yao** and **T Tony Cai** are among the most noticeable authors in the top part of the second eigenvector. On the other hand, the scores of authors whose scores are the biggest come to a significant drop at the fourth author **Naisyin Wang** and **Raymond J Carroll**, **Jianqing Fan** and **Yanyuan Ma** are the most distinguishable authors.

2.1.3 Authors of the other principal components

The same practice is also applied to the third and the fourth principal component.**Jianqing Fan** and **Raymond J Carroll**, **Yanyuan Ma** are distinguished from the third principal component. **Joseph G Ibrahim**, **Hongtu Zhu**, **Donglin Zeng**, **Heping Zhang** and **Jianqing Fan** are distinguished from the fourth principal component.

2.1.4 Embedding with Top 3 PCA eigenvectors

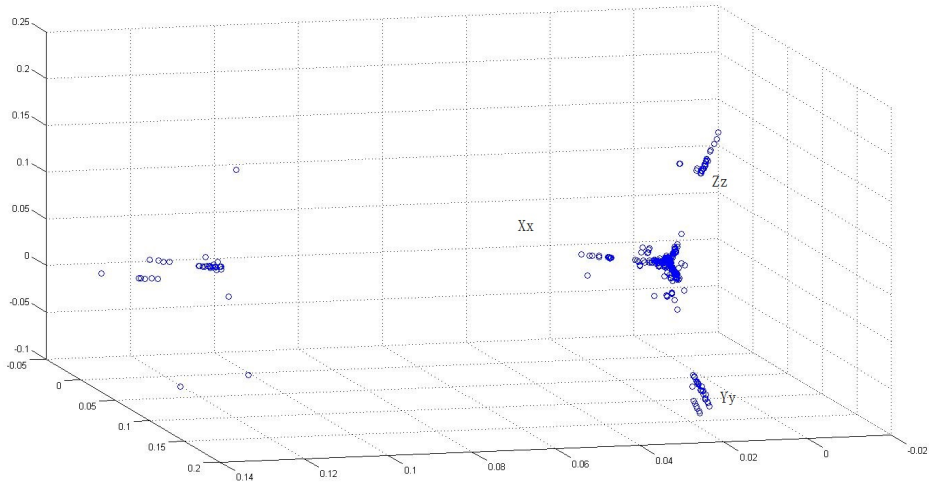


Figure 2: Principal component vector

After all the analysis with single principal components, we make a 3D plot of all the top three eigenvectors in (2), in the 3D plot, there seems to be three individual components (but not quite perpendicular), we name them as Xx, Yy and Zz respectively, in accordance with their direction in the 3-D coordinate system.

There are several distinguished authors highlighted in the plot, which are those stretches to the further end of the branches. We list several of them in table (1).

- The first group, Xx, leading by Peter Hall, is a 'Traditional statistics group'. In this group, authors are mainly interested in generalized traditional models in statistics, like the extension of linear model or multiple hypothesis testing.
- The second group, Yy, leading by Yanyuan Ma, is kind of a 'High dimensional statistics group', in which people are focusing on a branch of modern statistics, to deal with ultra-high dimensional data. In addition, their interest may also lie in biostatistics, which involves high-dimensional data generated from genomic data.
- The third group, Zz, leading by Jianqing Fan, is a 'non-parametric statistic group', which is also a staring branch of modern statistics. Their researches focus on non-parametric methods in statistics, and also its application in finance, economics and data mining.

Table 1: Distinguish authors in the three groups

Xx	Yy	Zz
Peter Hall	Yanyuan Ma	Jianqing Fan
Aurore Delaigle	Raymond J Carroll	Jiancheng Jiang
Hans-Georg Muller	Enno Mammen	Jianwen Cai
Tapabrata Maiti	Arnab Maity	Haibo Zhou
Jiashun Jin	Xihong Lin	Runze Li
Peihua Qiu	Nilanjan Chatterjee	Yang Feng
Fang Yao	Suojin Wang	Yacine Ait-Sahalia
Jing-Hao Xue	Jeffrey D Hart	Yingying Fan
Hugh Miller	Jianhua Z Huang	Heng Peng
Yao-Ban Chan	Bani Mallick	Yong Zhou
Alexander Meister	Yehua Li	Yichao Wu
Ingrid Van Keilegom	Yuedong Wang	Rui Song
Richard Samworth	Ying Wei	Clifford Lam
Joel L Horowitz	Lan Zhou	Wenyang Zhang
D M Titterington	Jeffrey S Morris	Tao Huang
Ryan T Elmore		
Amnon Neeman		

2.2 MDS and for co-authorship

2.2.1 MDS for the co-authorship matrix

According to the author-paper co-author matrix, each author takes co-ordinates in a 3248 dimensional- Euclidean space, the co-ordinate takes value 1 if the author has worked or cooperated on the i th published paper. The distance between two authors can therefore be defined as the Euclidean distance in that 3248-dimensional space. If the distance between two authors is smaller, then these two authors is more imitate in personal relationship and are more likely to co-operate.

We write

$$\mathrm{K} = \textit{authorpaperbiadj} \times \textit{authorpaperdiadj'}$$

$$\mathrm{D} = \textit{diag}(\mathbf{K}) \times \mathbf{1}^{\tau} + \mathbf{1} \times \textit{diag}(\mathbf{K}^{\tau}) \text{-} 2 \times \mathbf{K}$$

Then D is the defined distance matrix

Now that the distance matrix is already defined, we can operate MDS for co-authorship. Here the number of principal components used for MDS is three and the result is also visualized.

MDS for all authors It is shown clearly from the graph above that all authors are mainly scattered in three separate directions by MDS and most authors are scattered rather near to the original point because most of them have published only a few papers

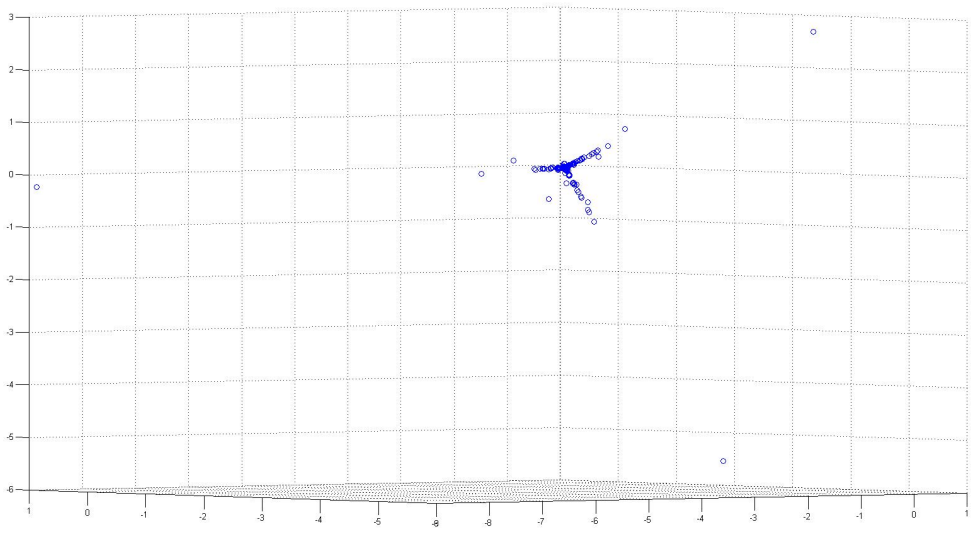


Figure 3: **MDS for all authors**

author	number of papers	numbers of authors cooperated
Jiancheng Jiang	8	6
Hans-Georg Muller	30	18
Yanyuan Ma	18	16
Aurore Delaigle	15	8
Raymond J Carroll	40	55
Jianqing Fan	40	38
Peter Hall	82	65

and thus their communal distances are short but their distances from those who have published plenty of papers are much more longer, so distinguishing the most outstanding ones among all the 3607 authors is quite reasonable.

Detecting the most outstanding authors from MDS As authors with a few published papers are likely to be scatter around the original point, in other words, the outstanding authors are likely to be scattered far away from the original point, the threshold of **outstanding authors** according to MDS is given as

$$d>1$$

where d is the distance between the co-ordinate of the author in MDS and the original point. 7 authors are selected as the most outstanding authors

All of the most outstanding seven authors all have published much more papers and co-operated many authors despite Jiancheng Jiang, who has published only 8 papers and has cooperated with only 6 other authors, it seems MDS still has some defects in detecting the authors.

2.2.2 MDS for only the valid authors

The definition for valid authors is given as before, and now MDS is only done on the valid authors, which is a subset of all authors. The graph of MDS for all authors

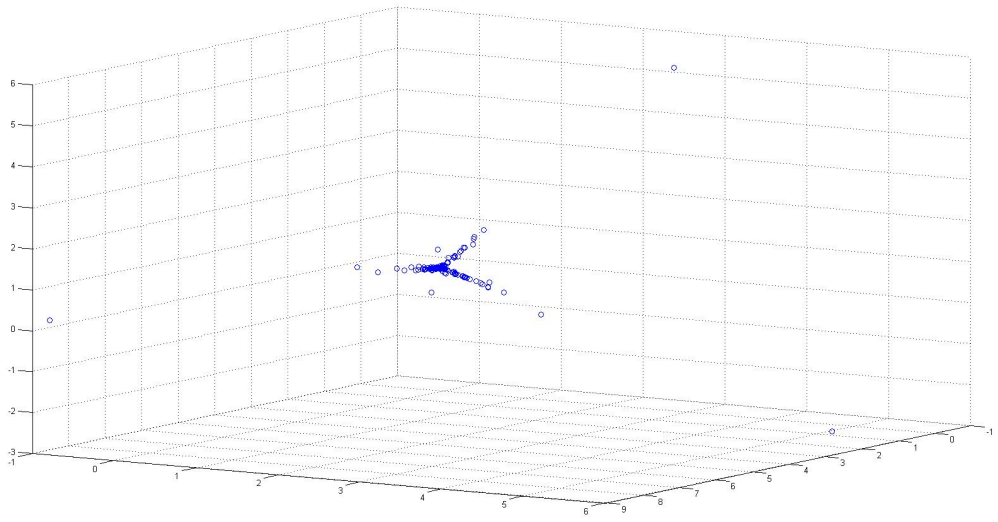


Figure 4: **MDS for valid authors**

much resemble the graph of MDS for all authors. It seems that the valid authors alone are adequate enough to represent all the information about the statistician community. To get more information about the role of the valid authors in the whole statistician community, the valid authors are plotted with a color different from the invalid ones in the figure of MDS for all authors.

In the figure above, the valid authors are marked with red dots but the invalid ones are marked with blue dots. Valid authors are primarily scattered in the outer part of the figure and authors whose co-ordinate is around zero are mostly invalid ones. This figure has well demonstrated out supposition that valid authors alone are enough to hold all the information about the whole statistician community, though the total number of them is indeed small, about $\frac{1}{8}$ of all authors.

2.3 Analysis of Paper-Paper Citation network

Firstly we analyse the distribution of the numbers of citations of these papers. There are 1693 papers who are not cited by others,1450 papers who don't cite others, and 778 papers who neither cite nor be cited by others. The main statistics are

Table 2: Statistical Description

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	1.762	2.000	75.000

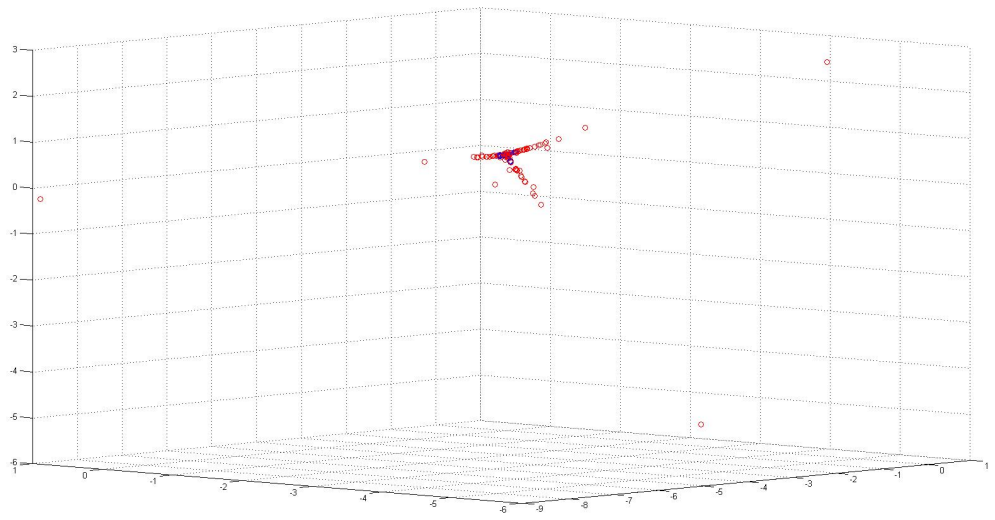
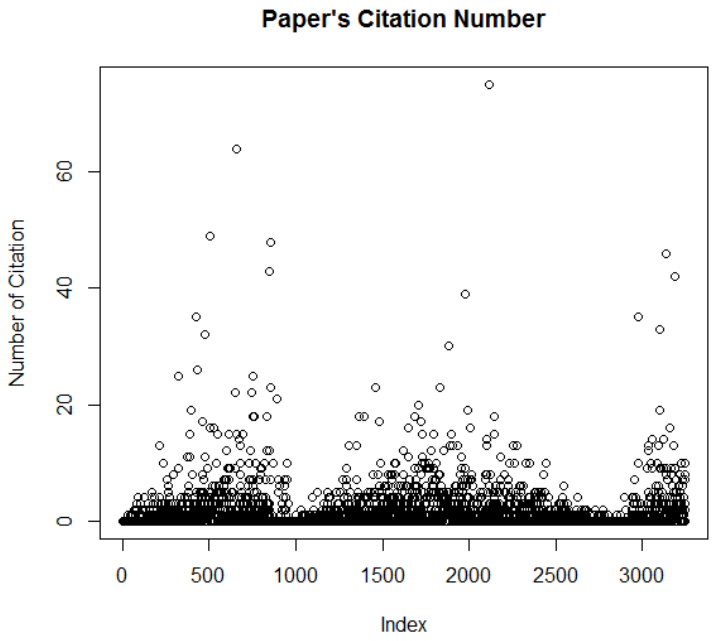


Figure 5: **valid authors in MDS for all authors**

from where we can conclude that the distribution is severely skewed. Then we can visualize this distribution by



But the above plot is too dense to see closely, so we only select papers whose number of citation is at least 10 and visualize their numbers of citations.

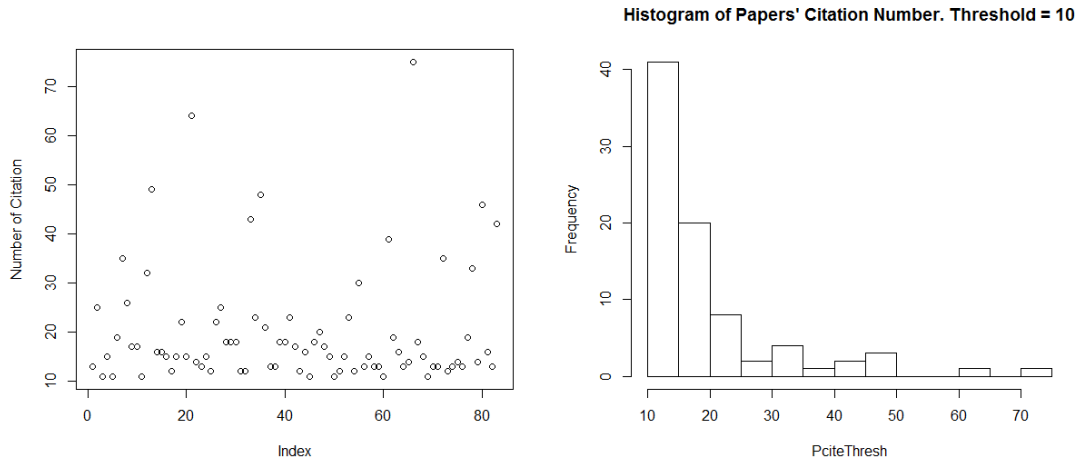


Figure 6: Scatter plot. The left figure shows the distribution of citations setting threshold $T = 10$. The right figure shows the corresponding histogram

Multi-Dimensional Scaling

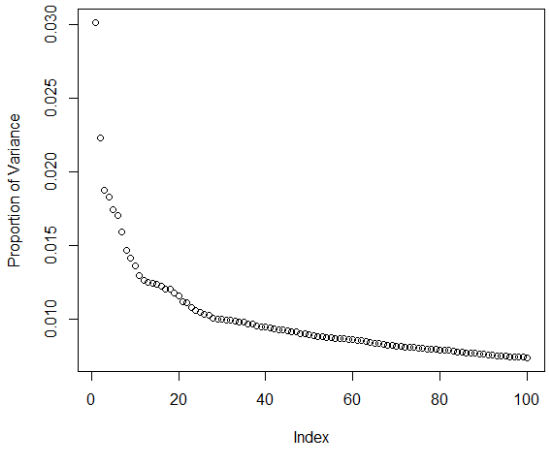
We can regard a paper as a vector whose entries show whether it is cited by other papers or not. Thus the unsymmetrical citation matrix of papers can be considered as a sample matrix $P = [p_1, p_2, \ldots, p_n]^T \in \mathbb{R}^{n \times p}$, where $n = 3248$ and $p = 3248$ is the number of papers in our dataset. For this matrix, we can apply MDS method on it and get a k-dimensional($k < p$) embedding for the data. We select $k = 100$ here, and the result of parallel analysis indicates that the first k principal analysis contain all the signals of interest in the matrix.

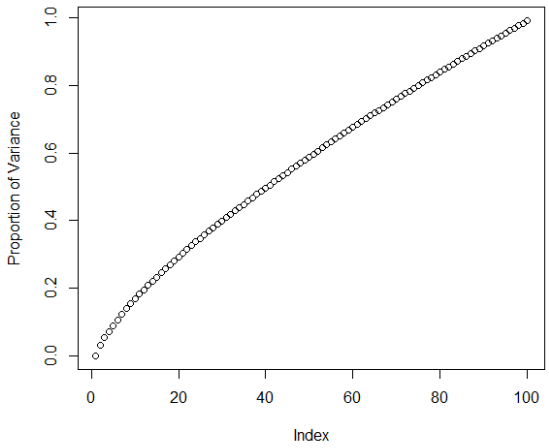
Precisely, we apply SVD on the centered matrix P and get

$$\tilde{P} = USV^T = [u_1, \ldots, u_k, u_{k+1}, \ldots, u_p] diag(\lambda_1, \ldots, \lambda_k, \lambda_{k+1}, \ldots, \lambda_p)[v_1, \ldots, v_k, v_{k+1}, \ldots, v_p]^T. \tag{1}$$

Then we can embed the data by taking the top k left singular vectors $P_k^{MDS} = U_k S_k^{1/2} \in \mathbb{R}^{n \times k}$.

To see the significance of MDS, we visualize the principal components by the proportion of variance each component explains, i.e., in the scatter plot the horizontal axis represents the index i of λ_i, and the vertical axis is the value $\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$.





We can see that there are no dominant principal components, say, even the first principal component can only explain 3% of the variance. When we visualize the accumulated proportion of variance the principal components explain, this phenomenon will become more evident, as is shown in the right scatter plot.

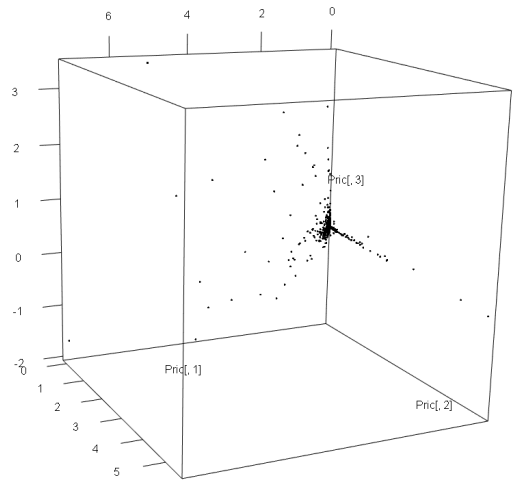
The reason of this, I think, is that papers can diverge in many aspects, such as topic, methodology, prospective, et al, thus we there should not be some dominant principal components to describe them.

From other prospective, we can say that it's hard to analysis the structure of the set of papers simply by analysing their citations.

So the first three principal components just explain 7.12% of the variance in total. Thus, they are not reasonably principal . The reason of this, I think, is that papers can diverge in many aspects, such as topic, methodology, prospective, et al, thus we there should not be some dominant principal components to describe them.

From other prospective, we can say that it's hard to analysis the structure of the set of papers simply by analysing their citations.

We draw the embedded data using the first three principal components and get the following result:



Thus using the first three left eigenvectors we can select some special papers from the dataset. For the first two left eigenvectors, papers' entries are whether 0 or positive, for the third left eigenvector both positive and negative entries exist. We check the papers whose corresponding entries(or after taking absolute value for the 3rd eigenvector) of the left eigenvectors are the largest.

Table 3: 1st Principal Component

Title	Year	Journal	Citation (Dataset)	Citation (Total)
The Adaptive Lasso and Its Oracle Properties	2006	JASA	75	2679
High-dimensional graphs and variable selection with the Lasso	2006	AoS	64	1669
The Dantzig selector: Statistical estimation when p is much larger than n	2007	AoS	49	2142
Nonconcave penalized likelihood with a diverging number of parameters	2004	AoS	48	525
Sure independence screening for ultrahigh dimensional feature space	2008	JRSSB	35	830
The sparsity and bias of the Lasso selection in high-dimensional linear regression	2008	AoS	35	443
Regularization and variable selection via the elastic net	2005	AoS	46	4629

Exploring more papers, we can confirm that the first principal component does describe the topic "High-dimensional Statistics" or "Variable Selection" .

Table 4: 2nd Principal Component

Title	Year	Journal	Citation (Dataset)	Citation (Total)
A stochastic process approach to false discovery control	2004	AoS	43	309
Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach	2003	JRSSB	42	953
Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis	2004	JASA	30	698
Higher criticism for detecting sparse heterogeneous mixtures	2004	AoS	23	365
False discovery and false nondiscovery rates in single-step multiple testing procedures	2006	AoS	14	73
The Positive False Discovery Rate: A Bayesian Interpretation and the Q-value	2003	AoS	21	1319
Adaptive linear step-up procedures that control the false discovery rate	2006	Bka	17	559

Similarly, by checking more papers we found that the second principal component describe the topic Hypothesis Testing , maybe Large-Scale Multiple Testing , as said in Jin's paper.

The third principal component has both positive and negative values.

Table 5: 3rd Principal Component

Title	Year	Journal	Citation (Dataset)	Citation (Total)
positive entries				
High-dimensional graphs and variable selection with the Lasso	2006	AoS	64	1669
Regularized estimation of large covariance matrices	2008	AoS	32	625
Functional Data Analysis for Sparse Longitudinal Data	2012	JASA	29	480
Covariance matrix selection and estimation via penalised normal likelihood	2006	Bka	23	80
Model selection and estimation in the Gaussian graphical model	2007	Bka	18	699
negative entries				
Nonconcave penalized likelihood with a diverging number of parameters	2004	AoS	48	525
The Adaptive Lasso and Its Oracle Properties	2006	JASA	75	2679
Tuning parameter selectors for the smoothly clipped absolute deviation method	2012	Bka	18	379
Regularization and variable selection via the elastic net	2005	AoS	46	4629
One-step sparse estimates in nonconcave penalized likelihood models	2008	Bka	26	628

However in the third principal components we cannot see any new common feature. They seem to have a common topic variable selection, as same as the first principal component. Thus we know that high-dimensional statistics or variable/model selection plays a fundamental role in today's statistical research.

Perron Theorem for Nonnegative Matrix Assume that $A \geq 0$. Then $\exists \lambda^* > 0, \|\nu^*\|_2 = 1$, s.t., $(\nu^*)^T A = \lambda^* (\nu^*)^T$. λ^* suffices that for any other eigenvalue λ of A, $|\lambda| \leq \lambda^*$. ν^* cannot be unique.

Perron theorem for nonnegative matrix explains why in our results of MDS, the coordinates of the embedded data with respect to the first principal component are all nonnegative.

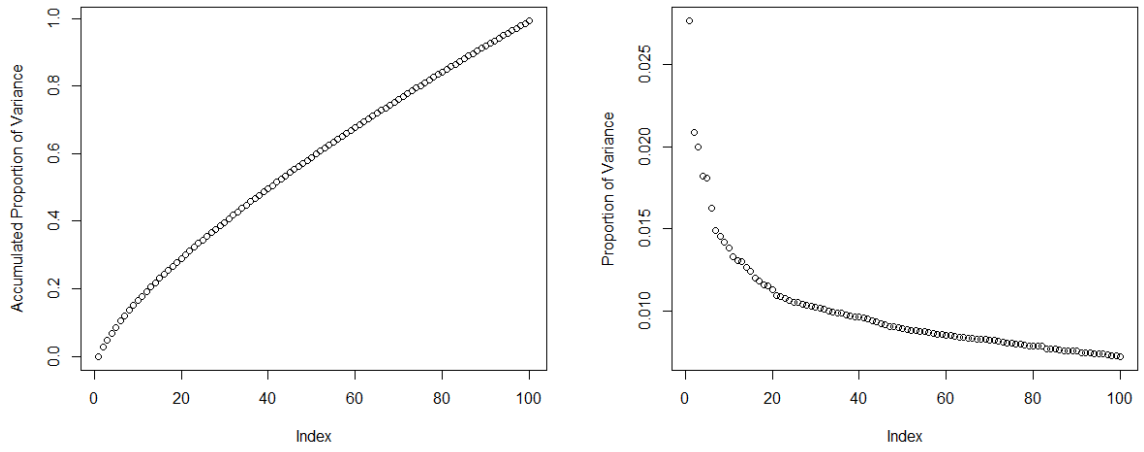
Guess

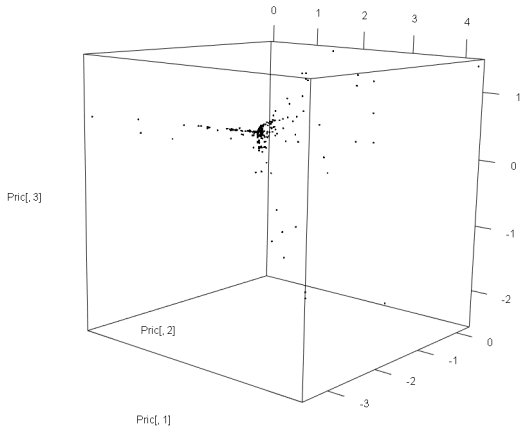
We think that the reason of the weak principal components is that in data, all the coauthorship relation and citation relationship are treated equally. But in fact, there should be some weights on the relation between data pairs. For instance, If we substitute

edges by weighted edges, maybe we can have better results. Futher exploration is still waiting.

Random Projection

As the matrices we study are all quite sparse, a natural question to ask is: whether we can use random projection on this data to reduce the dimension? The answer is yes, based on our numerical experiment. For the paper-paper citation matrix, if we firstly extract the largest maximum connected sub-graph, we shall get a 3248-by-2248 matrix. Then, if we apply a Bernoulli random projection to the matrix, we shall get a 3248-by-1000 matrix. For this matrix, the result of MDS are almost preserved. So random projection is valid in this problem.





Remark There are something interesting to remark:

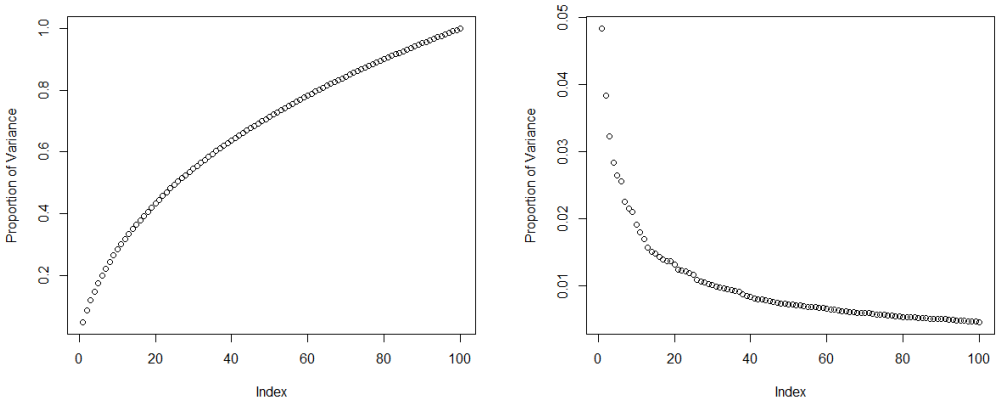
- A Bernoulli type random matrix does deduce a satisfying Random projection. Furthermore to get a comparatively well-interpret MDS result the minimum reduced dimension is 1000, i.e., 1/2 of the initial dimension.

- The phenomenon of phase transition can be obviously seen in the MDS result.

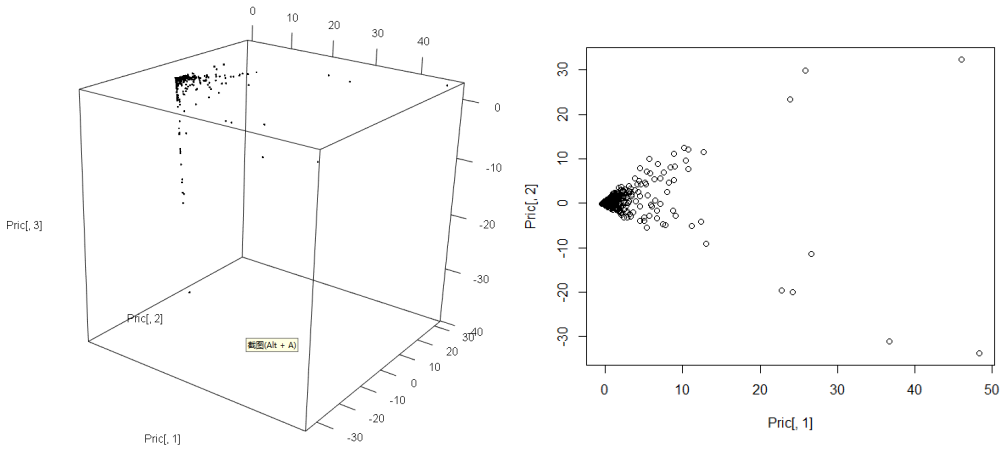
2.4 Analysis of Author-Author Citation Matrix

As we have seen, MDS of the coauthorship network and paper citation network are all not that satisfying, since there are some noises in the principal components. So we try to analyze the author citation matrix, and finally we get a slightly better result using this matrix.

The author citation matrix can be gotten easily by $A_C = A_P P_C A_P^T$. Then we perform MDS on A_C and get the plot for its principal components:



To visualize the MDS result, we have the following figures:



The first three principal components explains 11.91% of the variance, obviously stronger than they are in paper citation matrix, coauthorship matrix and author-paper adjacency matrix.

We sort the authors according to the first left eigenvectors in MDS and list the first 20 authors

Table 6: 1st Principal Component

1	Peter Hall	Jianqing Fan	Hans-Georg Muller
4	Raymond J Carroll	Hui Zou	Jane-Ling Wang
7	Runze Li	Fang Yao	Aurore Delaigle
10	Heng Peng	Naisyin Wang	T Tony Cai
13	Cun-Hui Zhang	Trevor J Hastie	Jian Huang
16	Yi Lin	Jiashun Jin	Nicolai Meinshausen
19	Ming Yuan	Peter J Bickel	

We can see that this principal component contains some evidences for finding the high-cited authors.

For the second principal component, or more precisely, in the second left eigenvector, there are both positive and negative entries. We list the first 30 authors according to the absolute value of the entries of the second left eigenvector for positive entries and negative entries separately and get the following results:

Surprisingly we find that the second left eigenvector of A_c approximately divide statisticians into two parts, for one part mainly study non-parametric, semi-parametric and other more traditional statistical topics, and the other part mainly study high dimensional statistics such as variable selection and lasso. This result is different from the result of the MDS for author-paper adjacency matrix and coauthorship matrix. We think the result here is more convincing because citation is more important an index to describe a statistician rather than coauthership.

For the third left eigenvectors, we similarly find that there are both positive and negative entries. We list the first authors just like we do for the 2nd left eigenvector as follows:

Fiedler Theorem Let L has n eigenvectors,

$$Lv_i = \lambda_i v_i,\ v_i \neq 0,\ i = 0, 1, \ldots, n-1,$$

where $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$. For the second smallest eigenvectors v_i, define

$$N_- = \{i : v_1(i) < 0\},$$

$$N_+ = \{i : v_1(i) > 0\},$$

$$N_0 = V - N_- - N_+.$$

We have the following results:

(1) $\#\{i, \lambda_i = 0\} = \#\{\text{connected components of G}\}$;

(2) If G is connected, then both N_- and N_+ are connected. $N_- \cup N_0$ and $N_+ \cup N_0$ might be disconnected if $N_0 \neq \emptyset$

Either by applying Fiedler theorem or using code "clusters" we can compute the con-

Table 7: 2nd Principal Component

positive entries			
1	Jianqing Fan	Hui Zou	Runze Li
4	Yi Lin	Trevor J Hastie	Heng Peng
7	Ming Yuan	R Dennis Cook	Jian Huang
10	Hansheng Wang	Nicolai Meinshausen	Peter Buhlmann
13	Bing Li	Cun-Hui Zhang	Hao Helen Zhang
16	Jinchi Lv	Lixing Zhu	Emmanuel J Candes
19	Chih-Ling Tsai	Terence Tao	Peter J Bickel
22	David R Hunter	Yingcun Xia	Robert J Tibshirani
25	Elizaveta Levina	Shuangge Ma	Ji Zhu
28	Mohsen Pourahmadi	Xiaotong Shen	L J Wei
negative entries			
1	Peter Hall	Hans-Georg Muller	Jane-Ling Wang
4	Fang Yao	Raymond J Carroll	Aurore Delaigle
7	Mohammad Hosseini-Nasab	T Tony Cai	Alexander Meister
10	David Ruppert	Naisyin Wang	Jeng-Min Chiou
13	Ulrich Stadtmuller	Gareth M James	John Staudenmayer
16	Daniel Gervini	Thomas C M Lee	Jeffrey S Morris
19	Yanyuan Ma	Jiashun Jin	Theo Gasser
22	Catherine A Sugar	Ciprian M Crainiceanu	Anastasios A Tsiatis
25	Bernard W Silverman	Tapabrata Maiti	Marina Vannucci
28	Damla Senturk	Philip J Brown	David L Donoho

Table 8: 3rd Principal Component

positive entries			
1	R Dennis Cook	Jianqing Fan	Bing Li
4	Lixing Zhu	Heng Peng	Runze Li
7	Jian Huang	Cun-Hui Zhang	Jinchi Lv
10	Larry Wasserman	Yi Lin	Jiashun Jin
13	Emmanuel J Candes	Francesca Chiaromonte	T Tony Cai
16	Nicolai Meinshausen	Christopher Genovese	Liqiang Ni
19	Yingcun Xia	Terence Tao	
negative entries			
1	David Dunson	Alan E Gelfand	Peter Muller
4	Gary L Rosner	Steven N MacEachern	Ju-Hyun Park
7	Fernando A Quintana	Mark F J Steel	Gareth Roberts
10	J E Griffin	Omiros Papaspiliopoulos	Athanasios Kottas
13	Michael I Jordan	Yee Whye Teh	Matthew J Beal
16	David M Blei	Natesh Pillai	Maria De Iorio
19	Chris C Holmes	Abel Rodriguez	

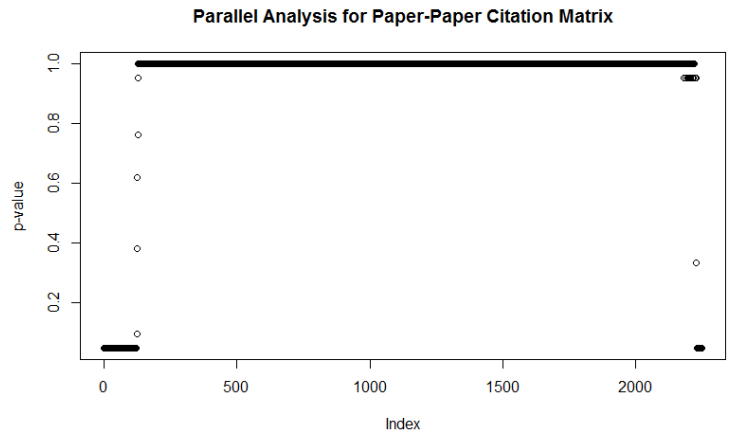
nected components in the coauthership graph. Without threshold, there are 523 connected components. If we set threshold T = 2 and let all the entries smaller than 2 in the coauthorship matrix to be 0, we can get a new coauthorship graph. In this graph there are 2985 connected components.

Furthermore, recall the results of parallel analysis. In MDS we actually conduct spectral decomposition on $M = \tilde{X}\tilde{X}^T$, where $\tilde{X}$ is the centered matrix for $X \in \mathbb{R}^{n\times p}$. We can check that M is a non-negative symmetric matrix, and is further an adjacency matrix for a connected graph G. Thus the second smallest eigenvalue of M contains information of the structure of G and is not noises, which can be seen in the result of parallel analysis.

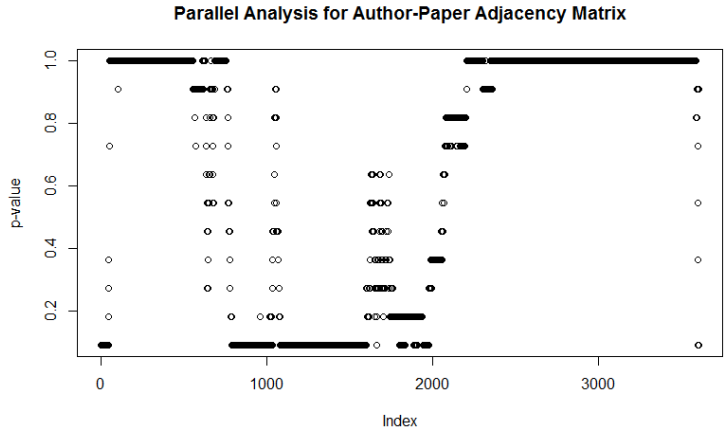
2.5 Parallel Analysis

In order to analyse whether the principal components we figure out are signals or noises, we conduct parallel analysis on the data matrix.

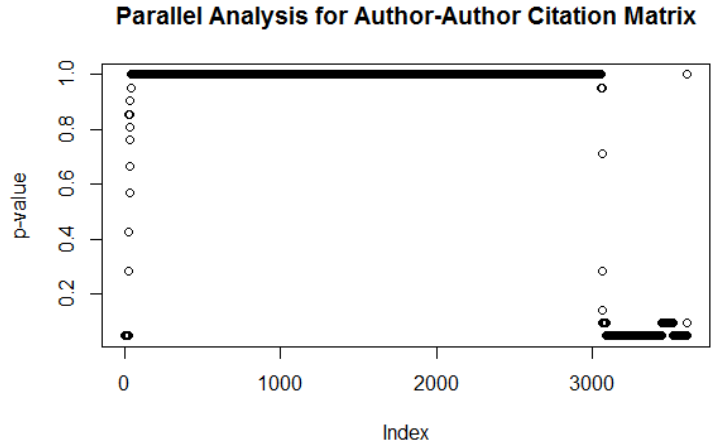
For the paper-paper citation matrix, after reducing the dimension by eliminating papers not in the maximal component in the citation graph, I set $R = 20$ and get the p-values of each principal components. The first 142 components are never detected to be noises, with p-values $p = 0.04761$. And the rest principal components are figured out to be noises. Further more, the sum of the first 142 principal components is 53.79% the sum of all principal components. The p-values can be seen in the following figure:



For the author-paper adjacency matrix, after reducing the dimension by eliminating papers not in the maximal component in the citation graph, I set $R = 10$ and get the p-values of each principal components. The first 142 components are never detected to be noises, with p-values $p = 0.0909$. And the rest principal components are figured out to be noises. Further more, the sum of the first 42 principal components is 13.82% the sum of all principal components. The p-values can be seen in the following figure:



For the author citation matrix, I set $R = 20$ and get the p-values of each principal components. The first 25 principal components are considered to be signals with p-value $p = 0.0476$, which means that they have never be considered to be noises in the parallel analysis. Furthermore, the first 25 principal components explain 66.13% of the variance, which is significantly larger than it was in former matrix. The p-values are



Remark The results of the parallel analysis make it sure that our MDS does make sense, since the first several principal components are actually judged to be signals. The weakness should be interpreted by the extraordinary large size of the data matrix. Remember that the first one hundred or 5% of the total dimension principal components almost explain 100% of the variance, so these principal components are strong and can be used to explain the structure of the dataset.

3 Sparse PCA

3.1 Introduction of sparse PCA

Now we consider a generalized PCA, sparse PCA, which focus to recover the sparse components of the data. And What we have now is exactly a sparse matrix! So it comes to us that a sparse recovery of the original data would be somehow attracting.

The algorithm of sparse PCA is motivated by SDP, Semi Definite Programming. Given Σ is the covariance matrix, recall that normal PCA is aimed to solve the following quadratic programming problem:

$$\begin{aligned} &\max\ x^T \Sigma x \\ &s.t.\ \|x\|_2 = 1 \end{aligned} \tag{2}$$

. which gives the maximal variation direction of covariance matrix Σ. Note that $x^T\Sigma = trace(\Sigma\, xx^T)$, (2) could also be written as

$$\begin{aligned} &\max\ trace(\Sigma X) \\ &s.t.\ trace(X) = 1 \\ &X \succeq 0 \end{aligned} \tag{3}$$

where $X \succeq 0$ means X is semi positive definite.

Now we are looing for sparse principal components ,namely $\#X_{ij} \neq 0$ are small while the explained variation remains as large as possible. Using 1-norm convexification [3], we have the following SDP formulation for sparse PCA:

$$\begin{aligned} &\max\ trace(\Sigma X) - \lambda \|X\|_2 \\ &s.t.\ trace(X) = 1 \\ &X \succeq 0 \end{aligned} \tag{4}$$

3.2 Large scale algorithm

This problem (4) could be solved by natural convex optimization algorithm in theory. However, with such a large matrix, it could be extremely time-consuming to calculate the exact solution. In consideration of this, we adopted an advanced algorithm called 'Augmented Lagrange Multiplier Method'. [7]

We choose $\lambda = \frac{max(diag(C))+min(diag(C))}{2}$, the results are quite surprising, all the top five sparse components are all concentrating on only one author in the Co-Authorship network, who are, probably coincidentally, just the most productive authors. To show this, we list them along with the number of paper published in (9). This phenomenon

might be caused by the choice of λ.

Table 9: Top sparse components in Co-authorship network

Paper published	Author
82	Peter Hall
40	Raymond J Carroll
40	Jianqing Fan
32	T Tony Cai
30	Hans-Georg Muller

4 Page-Rank Method

4.1 Rank the paper/author

First, consider the paper citation network (V, E), where V is the list of papers, and E is an adjacent matrix whose elements are:

$$E_{ij} = \begin{cases} 1, & \text{if the } j_{th} \text{ paper is cited by the } i_{th} \text{ paper} \\ 0, & \text{otherwise} \end{cases}$$

This graph measures the citations among papers. What we are seeking out it is the identify some influential papers, namely a single paper that is widely cited by other papers.

consider the Author citation network (G, F), where G is the list of all the authors, and E' is an adjacent matrix whose elements are:

$$F_{ij} = \begin{cases} k, & \text{if the } i_{th} \text{ author is cited by the } j_{th} \text{ author in k different papers} \\ 0, & \text{otherwise} \end{cases}$$

This graph measures the citations among authors. What we are seeking out of it is the find out some influential authors whose works are frequently cited by others.

From the first glance, these questions could be accomplished by calculating the 'in-degree' of a node, namely the total times cited for a specific paper/author. However, this may lead to a false detection if a small group of statisticians who frequently cite other people's work in that group even if they are not making big progress in the world of statistics.

4.2 Adopt Google's pagerank

Recall Google page-rank [6], we rank the web-pages in accordance to the links connecting each page, which is not only based on the 'quantity' of links, but also takes 'quality' into account. When a page is linked to some "influential" websites, even only a few, this page itself should be considered "influential" as well.

Similarly, this page-rank method could be applied to the directed graph of author-citation graph as well as paper-citation graph.

The algorithm is quite simple, just take the normalized adjacency matrix (as in) as a one step transition matrix P on the graph. And then consider a random walk on graph whose transition probability is given by P. This is exactly a Markov Chain according to the link structure of the papers. One would expect that stationary distributions of such a Markov Chain to reveal the relatively importance of the paper: the more cites, the more influential. So the stationary distribution will give the score for ranking to which we rely.

However, to assure that the stationary distribution is unique, which is guaranteed by the Perron-Frobenius theory, we have to make a small amendment on the matrix P, which is shown below in the detailed process:

Let E denote the original citation matrix as above, and

$$d = \mathbf{1}'E$$

$$D = diag(d)$$

so d_i denote the out-degree of the i_th node, which is the number of out-links.

Then, let

$$P = D^{-1}E' \tag{5}$$

, and we have P is a row-transition matrix, but not necessarily primitive. So again let

$$\tilde{P} = \alpha P + (1-\alpha)I_n$$

where α is a real number between $[0, 1]$, then $\tilde{P}$ is pair-wise connected, thus primitive. By Perron-Frobenius theory, the stationary distribution exists and is unique.

4.3 Results of paper-citation network

Taking sparsity into consideration, that is the citation relationship among the papers are quite sparse, thus the connectivity of the network should appear poor. Furthermore, about a half of the papers are never cited by others in this network, and a third who don't cite others. So the probability of a "random surf" should be augmented a little bit if we want a more exact recover of the citation relationship, for now we set $\alpha = 0.7$.

We plot the sorted scores on (7), which is divided by the standard deviation in advance to see the relative scores. From the scores, we can see that top 3 papers have much higher score that the other papers, and there is another "gap" between the top 7 papers and

those after them. We list the "top" 7 papers along with their authors and total times cited in table(10).

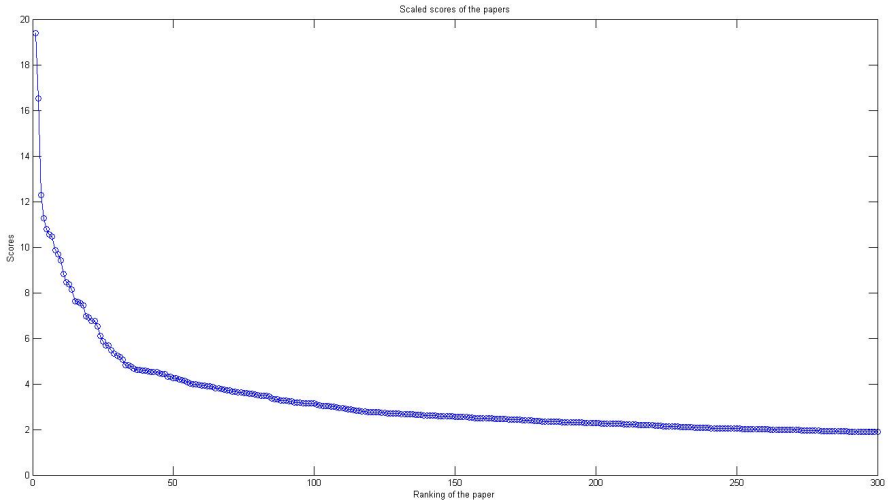


Figure 7: Sorted score for paper citation network (only the top 300)

Note that not all the "popular" papers are ranked high by Page-rank, there are some papers who have only been cited several times, like the 7th paper, but still on the top list. On the other hand, there exist such paper who have been cited all the times in the network but still get a rank off top 10. The most cited paper, *The Adaptive Lasso and Its Oracle Properties* by H.Zou et al., is only ranked 14 in spite of its 74 times of citation.

4.4 Results of author-citation network

Similarly, the author-citation network is also sparse, about a third of the authors are never cited in others' work, and a half never cite others work in their own paper, restricted in this small network though. Now, we set $\alpha = 0.7$ as well.

We plot the sorted scores on (8), which is divided by the standard deviation in advance to see the relative scores. From the scores, we can see that top 2 authors have much higher score that others, and there is a distinguish "gap" between the top 6 authors and those after them, which make these 6 authors a influential group. We list the "top" 6 authors along with the total times cited in table(11).

We can see that the top authors are not the most cited ones as well. Ryan Martin and Surya T. Tokdar have only published 2 or 3 papers and have been cited limited times, but if we focus on the ratio of citations per paper, their potential influence appears to be pretty considerable, or maybe they are cited by some influential authors, so they get a high ranking.

But a higher chance of getting cited do ensure a higher ranking in general. For those authors who have been cited 5 times standard deviation more than average, their scaled score are all more than 4, which guarantee a ranking within top 100.

Table 10: Add caption

Paper	Authors				Num
Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach	David Siegmund'	John D Storey'	Jonathan E Taylor'	""	42
Nonconcave penalized likelihood with a diverging number of parameters	Heng Peng'	Jianqing Fan'	""	""	48
Frequentist Model Average Estimators	Gerda Claeskens'	Nils Lid Hjort'	""	""	17
High-dimensional graphs and variable selection with the Lasso	Nicolai Meinshausen'	Peter Buhlmann'	""	""	64
Rankbased inference for the accelerated failure time model	Dan Yu Lin'	L J Wei'	Zhezhen Jin'	Zhiliang Ying'	18
Clustering for Sparsely Sampled Functional Data	Catherine A Sugar'	Gareth M James'	""	""	15
The Focused Information Criterion	Gerda Claeskens'	Nils Lid Hjort'	""	""	8

4.5 Further discuss of the leading authors

In addition to the scores, we would like to see how the distinguished authors are related to each other, say the top 11 authors (where there are also a gap at the 11th author in the ranking). The graph visualizing this relation is shown in (9).

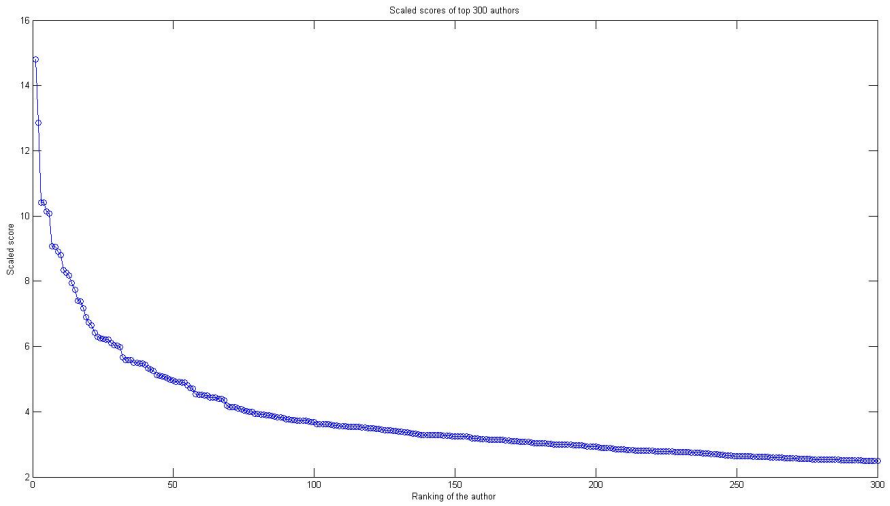


Figure 8: Sorted scores for author citation network (only the top 300)

Table 11: Distinguished authors in author-citation network

Author	Paper published	Times Cited	Score(scaled)
Peter Hall	82	134	14.28283
David Dunson	29	110	12.37133
Jianqing Fan	40	143	12.11169
Ryan Martin	3	12	11.80052
T Tony Cai	33	95	10.23443
Surya T Tokdar	2	8	9.95911

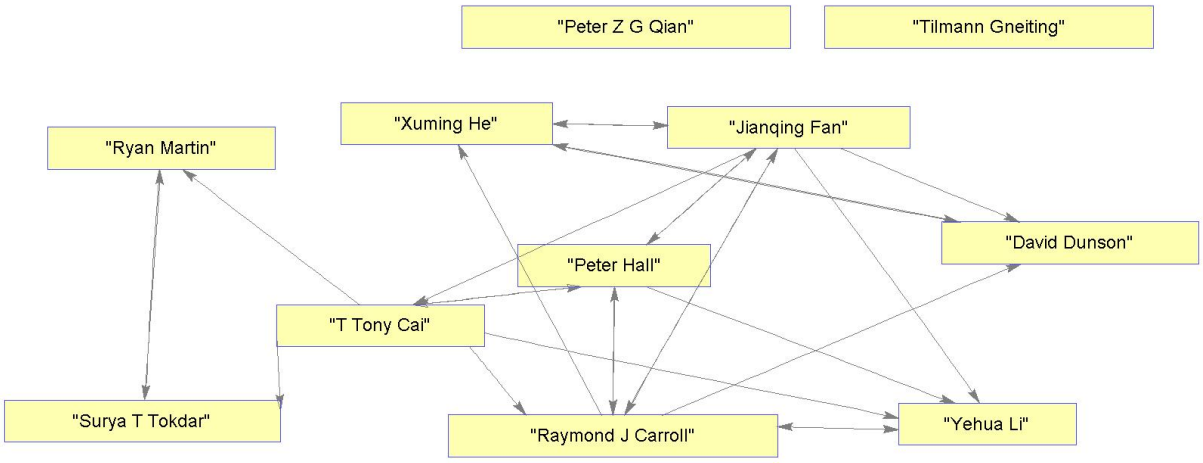


Figure 9: Citation network among 11 leading authors

Generally speaking, there are quite a complex citation relationship among these elite authors, where Peter Hall and Jianqing Fan serve as two centres connecting all the authors. We can see that there are two authors, Peter Qian and Tilmann Gneiting, not involved in any of the citation relationship of the other leading authors, which may implies that their research interest not overlap with other mainstream authors in the statistician's world.

4.6 Some flaws in Page-rank

However, note that this is only one rough way to measure the relatively importance of the paper and could be biased. There are ways to cheat Page-rank, that is if there are many cross links between a small set of nodes, those nodes would appear to be ranked much higher in Page-rank. Furthermore, the sparsity of the adjacency matrix may also lead to a biased ranking, for there are less connections among the nodes. Thus, this ranking is neither authorized nor formally authenticated, we adopt this just to show one possible ranking and to see the power or Paper-rank.

5 Community Detection

There are many community detection methods for undirected networks. In Jin's paper [5], they considered Newman's Spectral Clustering approach (NSC), Bickel and Chen's Prole Likelihood approach (BCPL), Armini et al.'s Pseudo Likelihood approach (APL), and Jin's SCORE.

Here we use the R package *igraph* and mainly try two methods, walktrap [8] and Louvain method [**?**]. Before applying the algorithm, we should extract a subset of the whole dataset so that the detected community is reasonable and the visualization of the result is handleable. We firstly let all the entries lower than 2 in the coauthorship matrix to be 0 and get a new adjacency matrix M'. Then we take the largest connected component G_1 from the graph G' which is represented by M'.

We apply these two methods on this connected component G_1. The following figure is the result gotten by Walktrap method, where the step is 10:

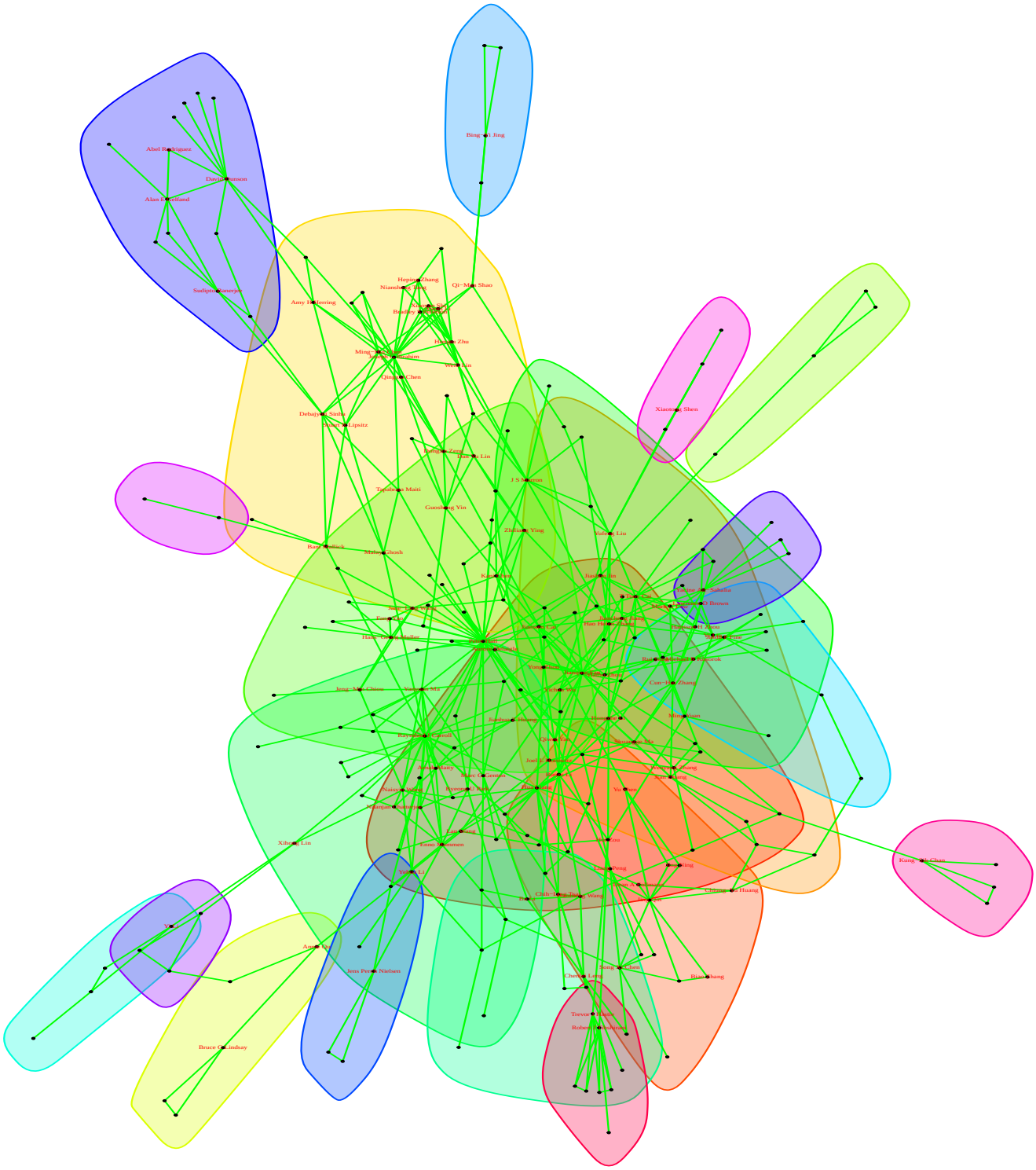


Figure 10: Community detection via walktrap algorithm. Only the nodes that have no less than 7 coauthors are labelled out with names.

From the above figure three obvious large communities can be seen, which central nodes are Raymond J Carroll, Petter Hall, and Jianqing Fan respectively.

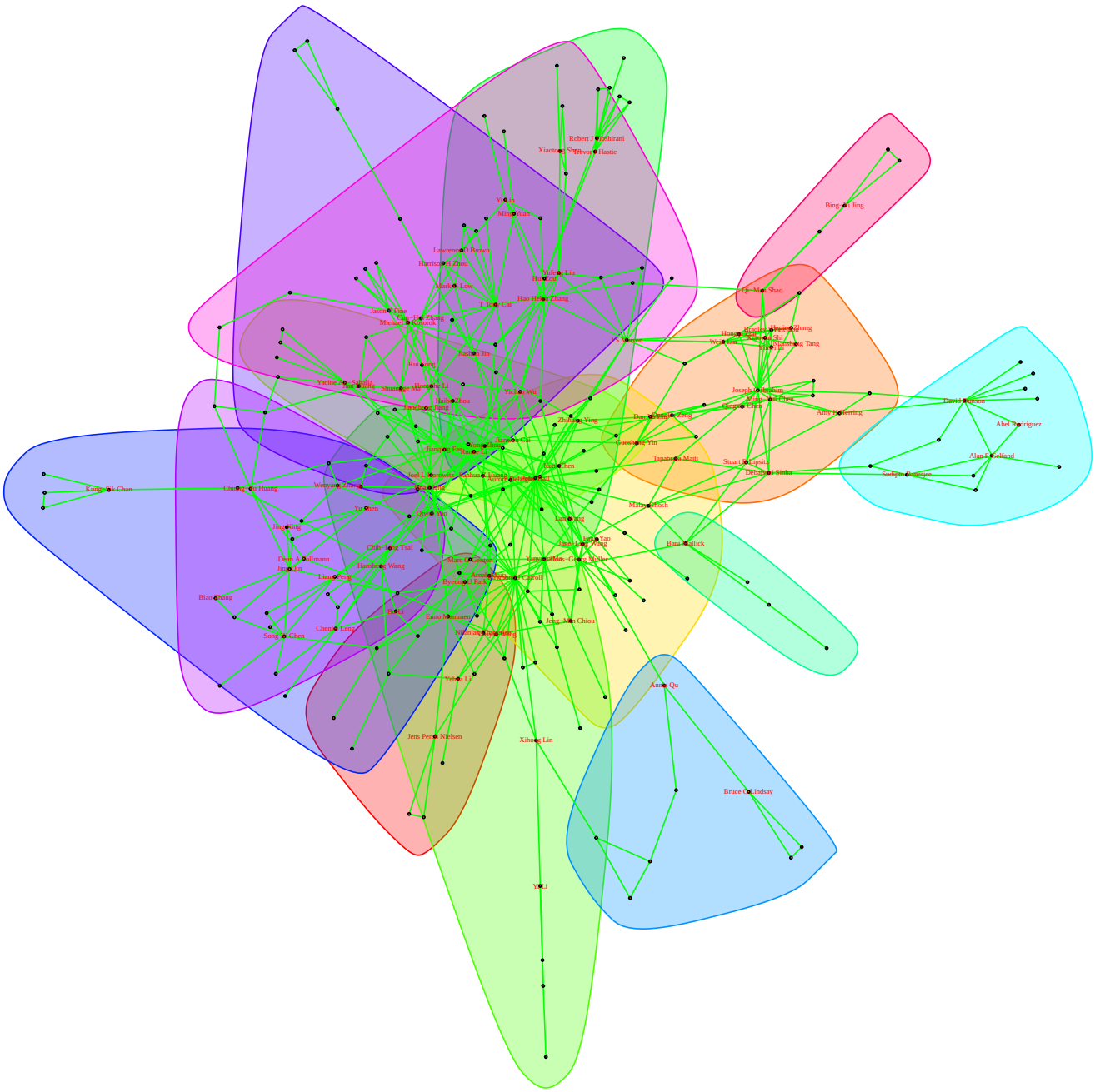


Figure 11: Community detection via Louvain algorithm. Only the nodes that have no less than 7 coauthors are labelled out with names.

The result here is similar as but slightly different from the result by Waltrap method. For instance, the Walktrap method divide Trevor J Hastie and Robert J Tibshirani into two communities, both different from the communities exempted by Raymond J Carroll, Petter Hall, and Jianqing Fan. But the Louvain method divides the two authors into the

same community as Peter Hall.

Another interesting thing to notice is the belongingness of Jianqing Fan. Jiashun Jin says in his article that maybe Fan's group is actually belong to both the "North-Carolina" and "Caroll-Hall" group. This phenomenon can also be seen in our detection results. Both the two methods detected three different communities for the three high-degree nodes Raymond J Carroll, Petter Hall, and Jianqing Fan. But both of them also give another community where there are at two of the three authors and Jianqing Fan is respectively be together with Raymond J Carroll and Petter Hall in the two methods' results. Thus our results coincide with Jin's and make its sense in a way.

Both the two results are quite different from the one gotten by Jiashun Jin. This inconsistency of the results by different methods is pointed out in Jin's paper [5]. They tried NSC, BCPL, APL and SCORE and found that when the input number of communities is $3 \leq K \leq 7$, the reults are quite different from each other. So they set $K = 2$ to get "more consistent" results. Here both the Walktrap method and the Louvain method does not need an input number of communities, thus the inconsistent result does make sense.

6 Hierarchical Analysis

Basically, there should be a hierarchical structure in the citation dataset. This hierarchical structure is actually an analogue structure as clustering structure in dataset where distance can be defined. For instance, the papers can be clustered in terms of how similar or closely connected they are, and the authors can be clustered according to how closely they cooperated or how similar their study fields are.

We use Affinity Propagation algorithm to figure out the clustering result in author-author citation dataset, then employ agglomerative clustering to join clusters.

6.1 Propagation

This algorithm is a clustering algorithm first published on Science in 2007 by Brendan J. Frey and Delbert Dueck [4]. The main difference between affinity propagation and other classical clustering algorithm is that affinity propagation does not need one to give the number of clusters before clustering.

A-P algorithm uses a heuristic frame to select exemplars for all the points, so cluster the points. Assume we have n data points to cluster. Firstly a similarity matrix $S \in \mathbb{R}^{n \times n}$ should be given, where $s(i, j)$ indicates how well the data point with index j is suited to be the exemplar for data point i. Then, a preference vector $p \in \mathbb{R}^{n \times 1}$ should be selected(which often takes value of diag(S)). A larger p_k indicates a larger possibility for data point k to be chosen as an exemplar.

Given S and p, the A-P algorithm uses an iterative scheme to get the final clusters. There are two key variables during the algorithm, responsibility $R \in \mathbb{R}^{n \times n}$ and availability $A \in \mathbb{R}^{n \times n}$. These two matrix describe the information past between points. One thing to

note is that A-P algorithm admits a insymmetric similarity matrix S, which makes the clustering to be more flexible.

The responsibility $R(i,j)$,sent from data point i to candidate exemplar point j, reflects the accumulated evidence for how well-suited point j is to serve as the exemplar for point i, taking into account other potential exemplars for point i. The availability $A(i,j)$, sent from candidate exemplar point j to point i, reflects the accumulated evidence for how appropriate it would be for point i to choose point j as its exemplar, taking into account the support from other points that point j should be an exemplar. $R(i,j)$ **and** $A(i,j)$ **can be viewed as log-probability ratios.**

To initialize, we can set $A = 0$. Then R and A are renewed by

$$R(i,j) \longleftarrow S(i,j) - \max_{j' \text{s.t.} j' \neq j} A(i,j') + S(i,j');$$

$$A(i,j) \longleftarrow R(i,j) - \min 0, R(j,j) + \sum_{i' \text{s.t.} i' \notin \{i,j\}} \max\{0, R(i',j)\}, \quad i \neq j;$$

$$A(j,j) \longleftarrow \sum_{i' \text{s.t.} i' \neq k} \max\{0, R(i',k)\}, \quad i = j.$$

This is the main recursion of A-P algorithm.

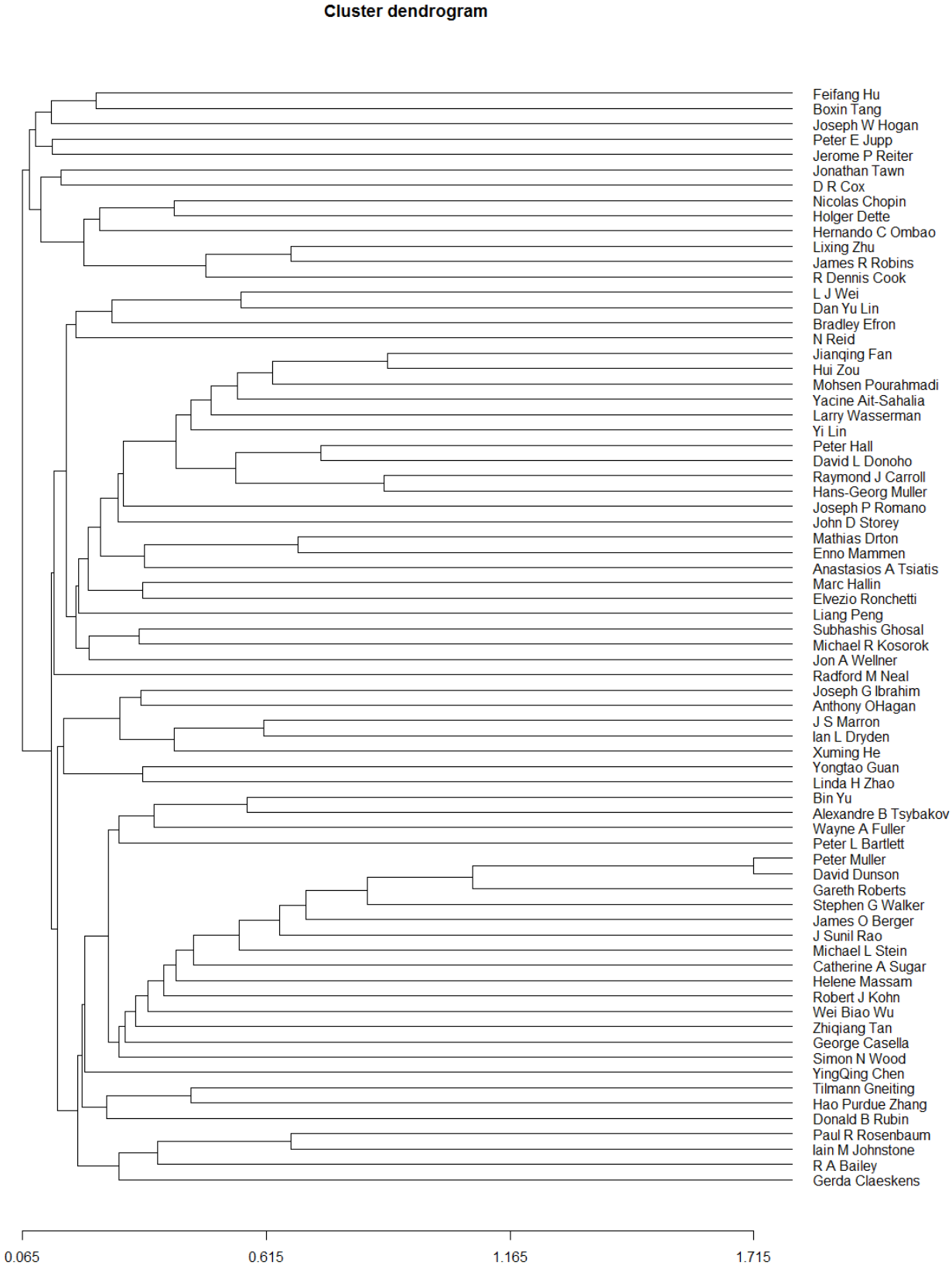
Thus the two citation matrix is quite appropriate for this algorithm. Since the citation number from i to j gives similarity $S(i,j)$ and the in-symmetric similarity matrix is permitted.

6.2 Agglomerative Clustering

The agglomerative clustering is used for hierarchical data analysis. It aggregate data points or clusters according to their similarities or distances to establish a hierarchical structure for the data set. In one turn of aggregation, it just cluster two data points or clusters. A commonly used rule to compute the distance between two clusters is the average distances between the data points in these two clusters.

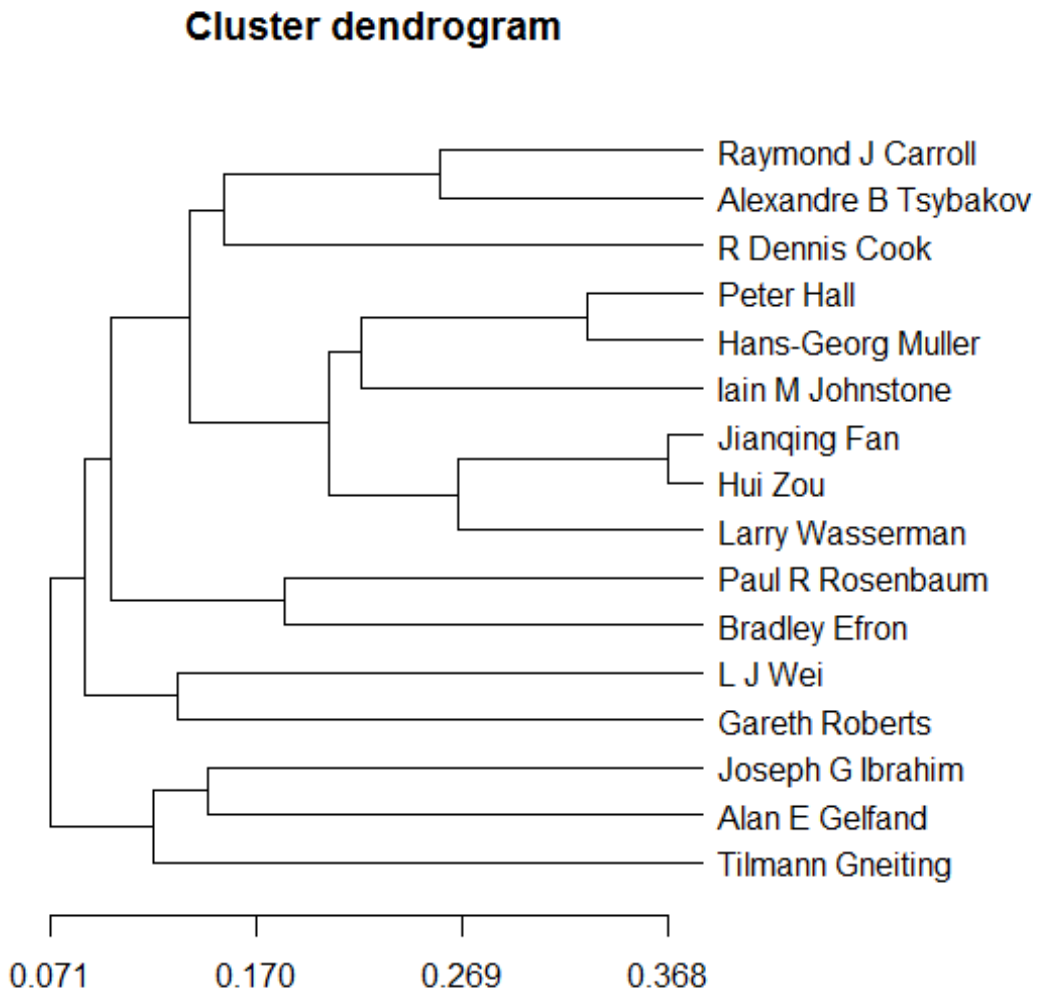
For the paper-paper citation matrix, the performance of the A-P algorithm is not very good. The selected exemplars for clusters are not papers that are really important in terms of the number of citation.

For the author-author citation matrix, we can get 72 clusters after using A-P algorithm when we choose a common preference $p = -10$. After applying the agglomerative clustering algorithm, the hierarchical structure we get is as follows:


Cluster dendrogram
Feifang Hu
Boxin Tang
Joseph W Hogan
Peter E Jupp
Jerome P Reiter
Jonathan Tawn
D R Cox
Nicolas Chopin
Holger Dette
Hernando C Ombao
Lixing Zhu
James R Robins
R Dennis Cook
L J Wei
Dan Yu Lin
Bradley Efron
N Reid
Jianqing Fan
Hui Zou
Mohsen Pourahmadi
Yacine Ait-Sahalia
Larry Wasserman
Yi Lin
Peter Hall
David L Donoho
Raymond J Carroll
Hans-Georg Muller
Joseph P Romano
John D Storey
Mathias Drton
Enno Mammen
Anastasios A Tsiatis
Marc Hallin
Elvezio Ronchetti
Liang Peng
Subhashis Ghosal
Michael R Kosorok
Jon A Wellner
Radford M Neal
Joseph G Ibrahim
Anthony OHagan
J S Marron
Ian L Dryden
Xuming He
Yongtao Guan
Linda H Zhao
Bin Yu
Alexandre B Tsybakov
Wayne A Fuller
Peter L Bartlett
Peter Muller
David Dunson
Gareth Roberts
Stephen G Walker
James O Berger
J Sunil Rao
Michael L Stein
Catherine A Sugar
Helene Massam
Robert J Kohn
Wei Biao Wu
Zhiqiang Tan
George Casella
Simon N Wood
YingQing Chen
Tilmann Gneiting
Hao Purdue Zhang
Donald B Rubin
Paul R Rosenbaum
Iain M Johnstone
R A Bailey
Gerda Claeskens
0.065
0.615
1.165
1.715

As we can see, the clustering result is reasonable since some highly cited authors are chosen to be exemplars, such as Peter Hall, Jianqing Fan, Bin Yu, Hui Zou, et al.

Then we set the common preference p to be -50 and get the following result, where there are 16 clusters:



We can see that with a lower p the number of clusters decrease from 72 to 16. But the highly performed authors are preserved to be exemplars. Though there are some unreasonable exemplars, e.g., with a comparatively low citation number or without papers of huge influence, we can say that the hierarchical analysis does make sense and can help us to find highly active statisticians, or more generally, highly active nodes, in a network. In details, here it find out Raymond J Carroll, Petter Hall, and Jianqing Fan and lay Raymond J Carroll and Petter Hall in the closest position. Also it place Jianqing Fan and Hui Zou in the closest position. All these facts show the validation of this method.

7 A Subgroup of Chinese Author

Note that there are four Chinese authors in the top 11 authors given by Page-rank, a natural question would be how the Chinese authors cooperate? After analysing the whole dataset, we focus our attention to a relatively small dataset, which contains information of all Chinese authors.

7.1 Identifying Chinese Authors

To recognize as more Chinese authors as possible in the whole network, we decide to proceed with the family names. We notice that some Chinese may change their first names in their lives, but the family names often remain unchanged. Luckily, common Chinese family names are given in a list includes only several hundreds of family names.

In this way, we can identify those authors whose last name spell as a typical Chinese character. There are about 780 ethnic-Chinese authors out of all the authors, which makes up about one-fourth of the whole group. Even for those who has published more than one paper, there are still 330 left.

7.2 PCA for the Chinese statisticians

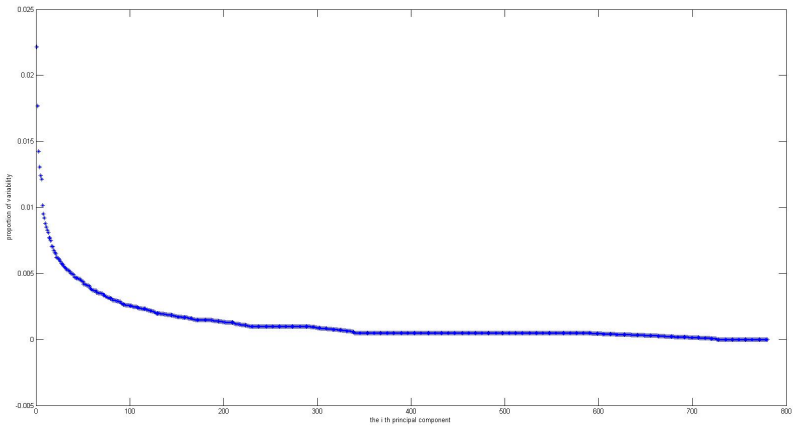
Since the Chinese is selected from all the authors, the Chinese co-authorship can therefore be generated, and normal PCA is done to reveal how the Chinese statisticians co-operated with each other. The result is plotted below.

However, from the figures above, the top principal components are not that effective, the top five principal components only accounts for 8 percent of the total variability. It can also be found in the first figure that the explanatory ability suddenly comes to a significant drop at the seventh principal component.

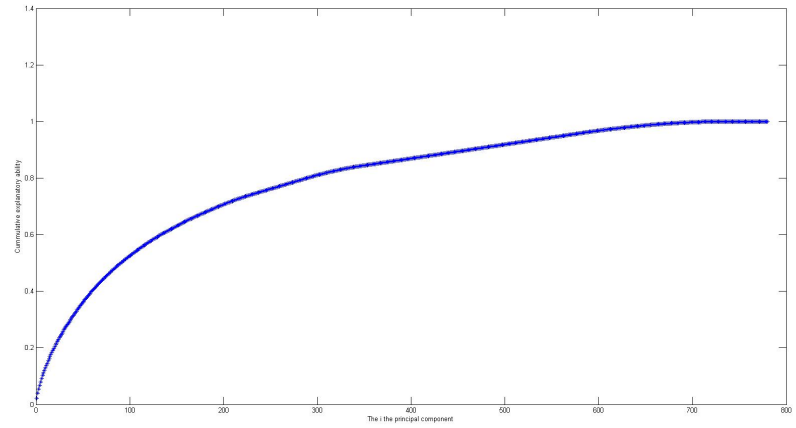
Though normal PCA itself is not very effective, a bunch of distinguished Chinese authors can still be found. Among which, we can distinguish **Jianqing Fan** from the first principal component; **T Tony Cai** from the second component; **Lu Tian**, **L J Wei** and **Tian Xi Cai** from the third component, as well as **Jing Qin** from the fourth component;**Lixing Zhu** from the fifth component and **Hongtu Zhu**, **Heping Zhang** from the sixth component.

7.3 Statisticians of Chinese origins in the results of MDS

In the figures below, authors with Chinese origins are marked with red dots and others are marked with blue dots.



(a) Scree plot(only the top 100 components)



(b) Explanatory ability of the principal components

Figure 12: Cumulative Explanatory ability of the principal components

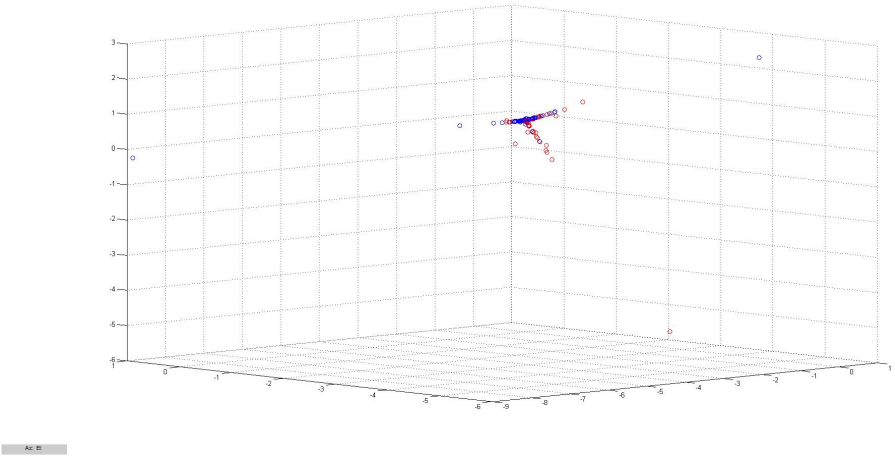


Figure 13: **Chinese in MDS for all authors**

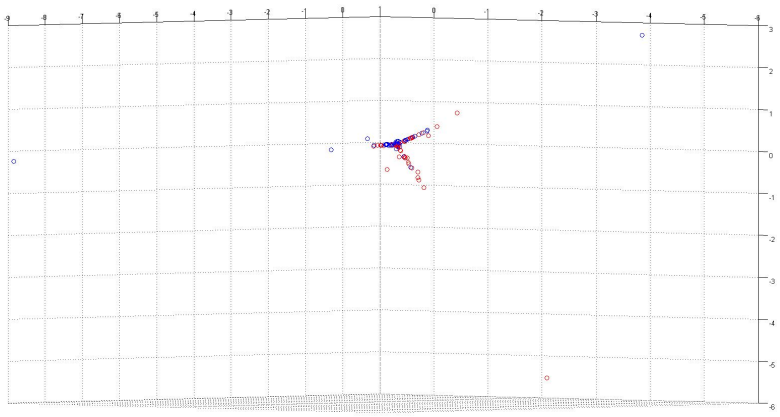


Figure 14: **Chinese in MDS for valid authors**

Nothing special is found in the first graph as the Chinese authors are scattered evenly among all the others, but the second graph concerning only about valid authors leads to some findings. The valid Chinese authors nearly make up the whole part in one of the three main directions, but in the other two directions, there are much fewer Chinese.

7.4 Page-rank

We are now aimed at rank the Chinese authors depending only on the citation within this small network. The procedure is the same as above, and the process is simplified due to the size of the group is restricted. The sorted scores are shown below in (15). We can see that there are about eight distinguished authors in the Chinese statistician network, whose score is clearly higher than the others given by the Page-rank (about 5 standard deviations higher than the average), their names are listed in (**??**) along with the total times of being cite.

We can see that they are indeed talented and outstanding statisticians in the world, many of them are ranked top 20 among all the authors by Page-rank. Chinese statistician is now a influential group to the world and have contributed a lot to the academic field of statistic.

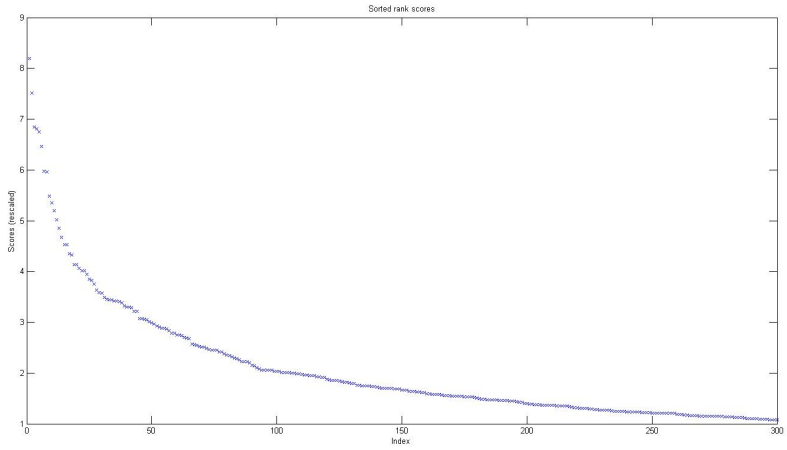


Figure 15: Sorted score within Chinese group (only top 300)

Table 12: 8 Distinguished authors in the Chinese network

Author	Paper Published	Times cited	Score
Jianqing Fan	40	59	0.0084
Peter Z G Qian	9	11	0.0077
Jianhua Z Huang	9	27	0.0070
Song Xi Chen	14	37	0.0069
T Tony Cai	33	30	0.0069
Xuming He	17	32	0.0066
Hansheng Wang	11	38	0.0061
Hua Liang	15	43	0.0061

Furthermore, we want to see how these distinguished authors cite each other within this small group of 'elite' authors. A network is drawn in (16) to show this. We can see that these authors are connected quite closely to each others, in which Jianqing Fan, who cites as well as being cited the most, works as a center, and Songxi Chen is also widely cited. Compared to the citation in the international group in(9), this Chinese group clearly has a more complex structure. This means that top Chinese statisticians are united and they value each other's work. Note that there is one author 'Peter Qian' missing, strangely, he never cites or being cited by other authors in this list.

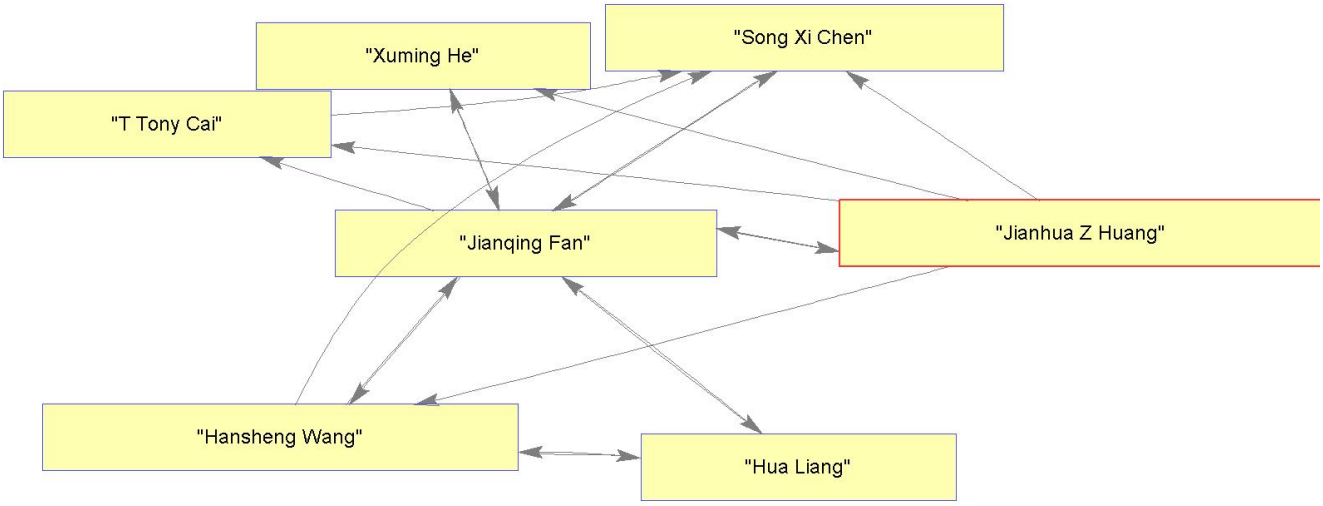


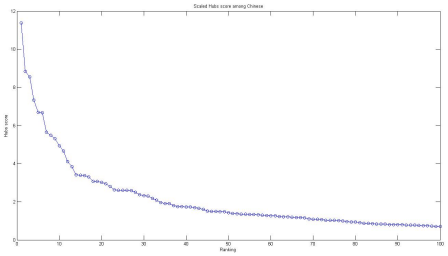
Figure 16: Network among the distinguished authors

Also, we noticed that nearly all the distinguished authors work in top universities in the U.S., only Hansheng Wang and Songxi Chen are working in Guanghua School of Management, Peking University.

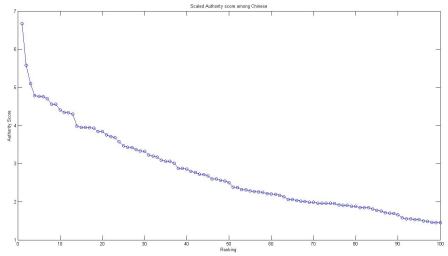
7.5 Authority and Hub

According to the concept by [**?**], we want to identify the authorities and hubs in this small linked network. The authorities represent those with large out-degree, and the hubs are those with large in-degree. The algorithms are simple, namely α is the primary eigenvector of $W'W$, gives the "authority score", and γ is the primary eigenvector of WW', gives the "hubs score".

For the Chinese group, the scores for the top 100 authorities and top 100 hubs are plotted in (7.5). Note that the top 6 hubs and top 13 authorities stand out from the others, so we list the distinguished hub and authority authors in table (13) along with their numbers of cites and cited.



(a) Hubs score (top 100)



(b) Authority score(top 100)

Table 13: Outstanding authority and hub authors

Authority	Cited	Hub	Cites
Jianqing Fan	59	Jianqing Fan	128
Runze Li	42	Hui Zou	79
Hansheng Wang	38	Runze Li	77
Hua Liang	43	Heng Peng	63
Hao Helen Zhang	35	Ming Yuan	63
Lixing Zhu	38	Yi Lin	57
Yang Feng	29		
Hongzhe Li	29		
Lan Wang	30		
Yichao Wu	26		
Hui Zou	28		
Song Xi Chen	37		
Cheng Yong Tang	27		

Notice that there is a overlapping in the two groups, that are Jianqing Fan, Hui Zou and Runze Li, indicating they are both citing the most and cited the most, make them a center group of the whole Chinese network.

Again, we would like to visualize the relationships between those authors, the network is shown in (17), we can see that there are such a frequent citation between those authors. In this smaller network, wee see that Jianqing Fan, Hui Zou and Runze Li do serve as the productive centres, connecting with many others in this network. Yan Feng, Hansheng Wang, and Lixing Zhu are located on the periphery, are cited frequently as authorities, while Heng Peng, Ming Yuan and Yi Lin cites others frequently, serve as information hubs who spread others' work.

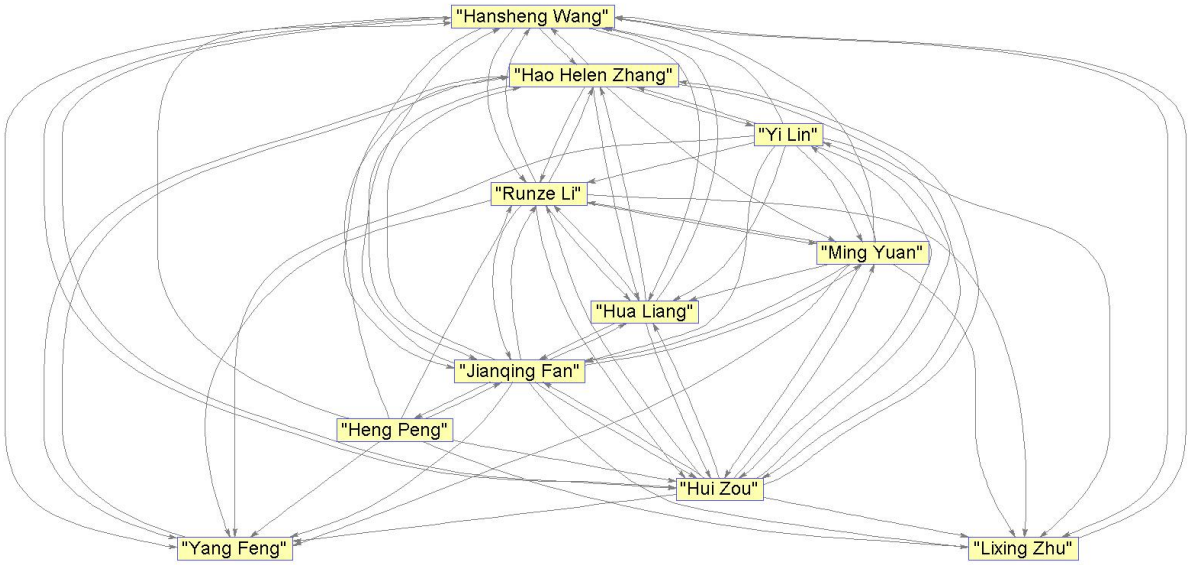


Figure 17: Network among hubs and authorities

8 The yearly change of the statistician community

8.1 The change of the number of statisticians with years

The total number of statisticians of each year is decided by the number of authors who has published papers before that year, and the number of new comers of each year is decided by the difference of the total number of statisticians of the next year and the total number of that year.

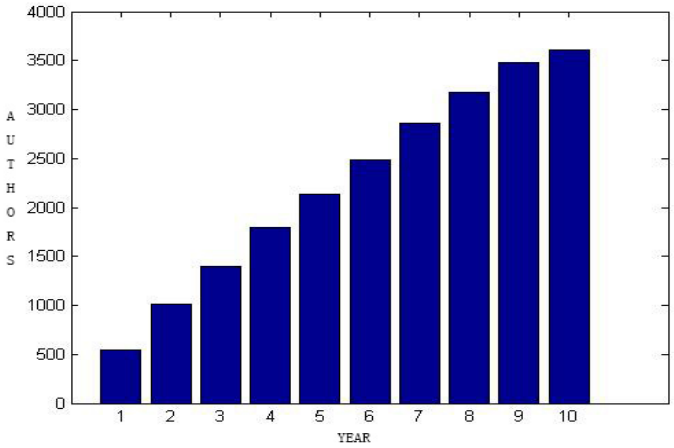


Figure 18: **The number of authors in each year**

From the figure above, the number of statisticians who have already published papers

has been on the rise continually at an almost constant pace, indicating a steady growth and development in the statistician community. The new comers to the statistician community is at around 400 each year, though the pace of rise has greatly dropped in 2012, but the paper published in 2012 is also significantly fewer. Only 164 were published in 2012, less than half the number of the other years. The reason for fewer papers published in 2012 may be that the data collector had ended data collection work in the middle of 2012, and only papers in the first half year in 2012 were collected.

8.2 The change of the modularity of the statistician community network with years

The definition for modularity of a community network is

$$Modularity = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(i,j)$$

Where i,j represents the i th and the j th author in the author list. A_{ij} takes value 1 if the i th and the j th author have co-operated, otherwise, $A_{ij} = 0$. k_i is the number of authors the i th author has co-operated with, namely, $k_i = sum_j A_{ij}$. $\delta(i,j)$ takes value 1 if the i th and the j th author are divided into the same community, otherwise, $\delta(i,j) = 0$. If Modularity is bigger, then the network shows a stronger characteristics of communitxxies. Generally, experiences show that if $Modularity > 0.3$, the network is regarded to have clear and significant structures of communities. The modularity for each year can therefore be calculated through walktrap algorithm.

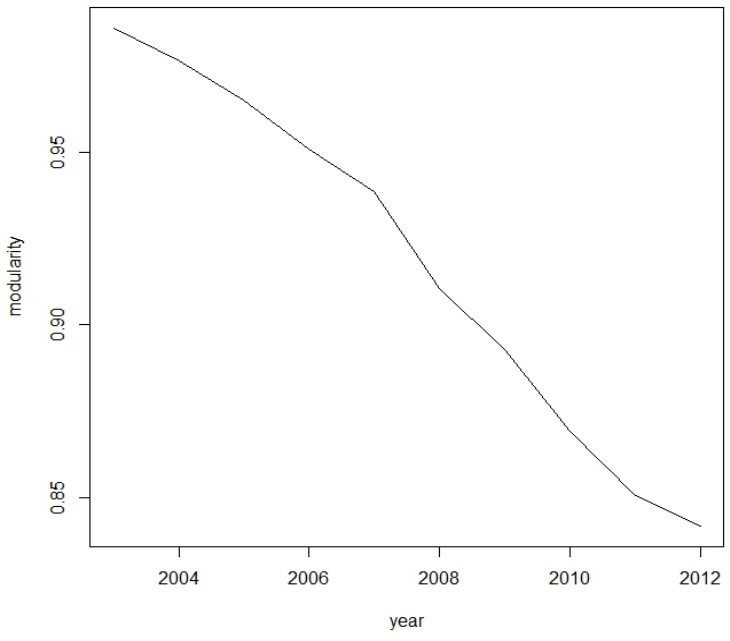


Figure 19: **Modularity in each year**

According to the graph above, the value for modularity has been always larger than 0.80, and the statistician's co-authorship network shows a strong sign of communities structure. The value of modularity has been noticeably continually on the drop at an almost constant pace, which behaves like the pattern of random data. It can be partially explained by the new comers as they can form new communities when new fields of statistics have merged as well as destroy the old pattern of communities when they frequently co-operated with only a part of the old guys.

References

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[2] Buja and Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 1992.

[3] Alexandre d'Aspremont, Laurent El Ghaoul, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation of sparse pca using semidefinite programming. *SIAM Review*.

[4] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 2007.

[5] Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *Annals of Applied Statistics*, 2015.

[6] Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[7] Nikhil Naikal, Allen Y. Yang, and S. Shankar Sastry. Informative feature selection for object recognition via sparse pca. *2011 International Conference on Computer Vision*, 2011.

[8] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, Journal of Graph Algorithms and Applications.