Mini-project

ZHU, Weizhi HKUST
FAN, Min HKUST

December 28, 2017

1 Coauthorship detection

In this part, we focus on the data of Coauthership matrix as follow:

CA coauthorAdj.txt - 3607x3607 adjacency matrix of the coauthorship network. More preciously, the (i, j)-th entry of CA is 1 if authors i and j had coauthored at least one paper, 0 otherwise.

CAG coauthorAdjGiant.txt - 2263x2263 adjacency matrix of the giant component of the coauthorship network where each edge denotes at least t=1 paper coauthored. That is to say, the graph CA has many disjiont subgraphs. And CAG is the giant component subgraph of CA.

We let CAT denote the Threshold adjacency matrix whose (i, j)-th entry is 1 if authors i and j had coauthored at least two papers, 0 otherwise. CATG coauthorAdjThreshGiant.txt - 236x236 adjacency matrix of the giant component of the graph CAT.

ACA authorCiteAdj.txt - 3607x3607 adjacency matrix whose (i, j)-th entry is 1 if author i cites author j at least once, 0 otherwise. And ACAG authorCitAdjGiant.txt - 2654x2654 adjacency matrix is a giant component of graph of ACA.

We will firstly detect the connect component of original coauthorship adjancency matrix CA by **Fiedler Theroy**, the disconnection distinguishes in most apparent way: people in each connect component do not cooperate.

Then we further dectect community in the largest connect component, with adjacency matrix CATG(or CAG), lots of clustering method could be applied here, such as **Minimizing Cut**,**Minimizing distance/probability of transition in different clusters**,**SCORE method** and so on. We try to give some explantion in different community dection and compare each result.

1.1 Connect component dectection

1.2 Community dectection

As we have introduced at the begin of this section, we will do community dectection of the largest connenct component CATG. The first method we apply is to **minimize** cut, for a given k,

$$\begin{aligned} RatioCut(A_1, ..., k) &= \sum_{i=1}^{k} \frac{cut(A_i, A_i^c)}{|A_i|} \\ Ncut(A_1, ..., A_k) &= \sum_{i=1}^{k} \frac{cut(A_i, A_i^c)}{Vol(A_i)} \end{aligned} \tag{1.1}$$

However, the combination optimization is NP-hard, we using **Spectral Method** as a relaxation to dectect the community, such as **Normalized Laplacian**, **Unnormalized Laplacian** and **Cheeger bipartition**.

Before we introduce several algorithms, we firstly give a breif introduction on graph Laplacian. Given an adjacent matrix $A^{n\times n}$, firstly we consider the case of undirected graph, i.e.$a_{ij} = a_{ji}$:

Let $D = diag(d_i), d_i = \sum_{j=1}^{n} A_{ij}$, then unnormalized Laplacian is defined as $L = D - A$, normalized Laplacian is defined as $\mathcal{L} = D^{-1/2}LD^{-1/2}$, and the probability transition matrix is defined as $P = D^{-1}L$.

Algorithm of Unnormalized Laplacian Spectral Clustering is given below,

Algorithm 1: Unnormalized spectral clustering

```
Input Similarity matrix $S \in \mathrm{R}^{n\times n}$, number $k$ of clusters to construct initial $r_0 = b$ , $x_0 = 0$ , $S_0 = \emptyset$
Construct a similarity graph by one of the ways described in Section 2. Let $W$ be its weighted
  adjacency matrix.
Compute the unnormalized Laplacian $L$.
Compute the first $k$ eigenvectors $v_1, \cdots, v_k$ of $L$.
Let $V \in \mathrm{R}^{n\times k}$ be the matrix containing the vectors $v_1, \cdots, v_k$ as columns.
For $i = 1, \cdots, n$, let $y_i \in \mathrm{R}^k$ be the vector corresponding to the $i$-th row of $V$.
Cluster the points $(y_i)_{i=1,\cdots,n}$ in $\mathrm{R}^k$ with the $k$-means algorithm into clusters $C_1, \cdots, C_k$.
Output Clusters $A_1, \cdots, A_k$ with $A_i = \{j|y_j \in C_i\}$.
```

The unnormalized one is basically similar which we omit. Cheeger's bipartition will be done in the final project.

Now, we turn to SCORE, that is, Spectral Clustering On Ratios of Eigenvectors.SCORE is a recent spectral method proposed by Jin [24]. This method tends to be reasonable because the ratios of eigenvectors are proportion to the heterogeneity parameters. Assume K (number of communities) as known and let A be the adjacency matrix associated with $\mathcal{N}$:

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

note that A is symmetric. SCORE consists of the following simple steps.

- Let $\hat{\xi}_1, \hat{\xi}_2, \cdots, \hat{\xi}_K$ be the first K (unit-norm) eigenvectors of A. Obtain the $n \times (K-1)$ matrix $\hat{R}$ by $\hat{R}(i,k) = \hat{\xi}_{k+1}(i)/\hat{\xi}_1(i), 1 \leq i \leq n, 1 \leq k \leq K-1$.

- Clustering by applying the classical k-means to $\hat{R}$, assuming there are $\leq K$ communities.

Besides NSC,BCPL,APL method will be tried in final project.

Now we give the experiment result:

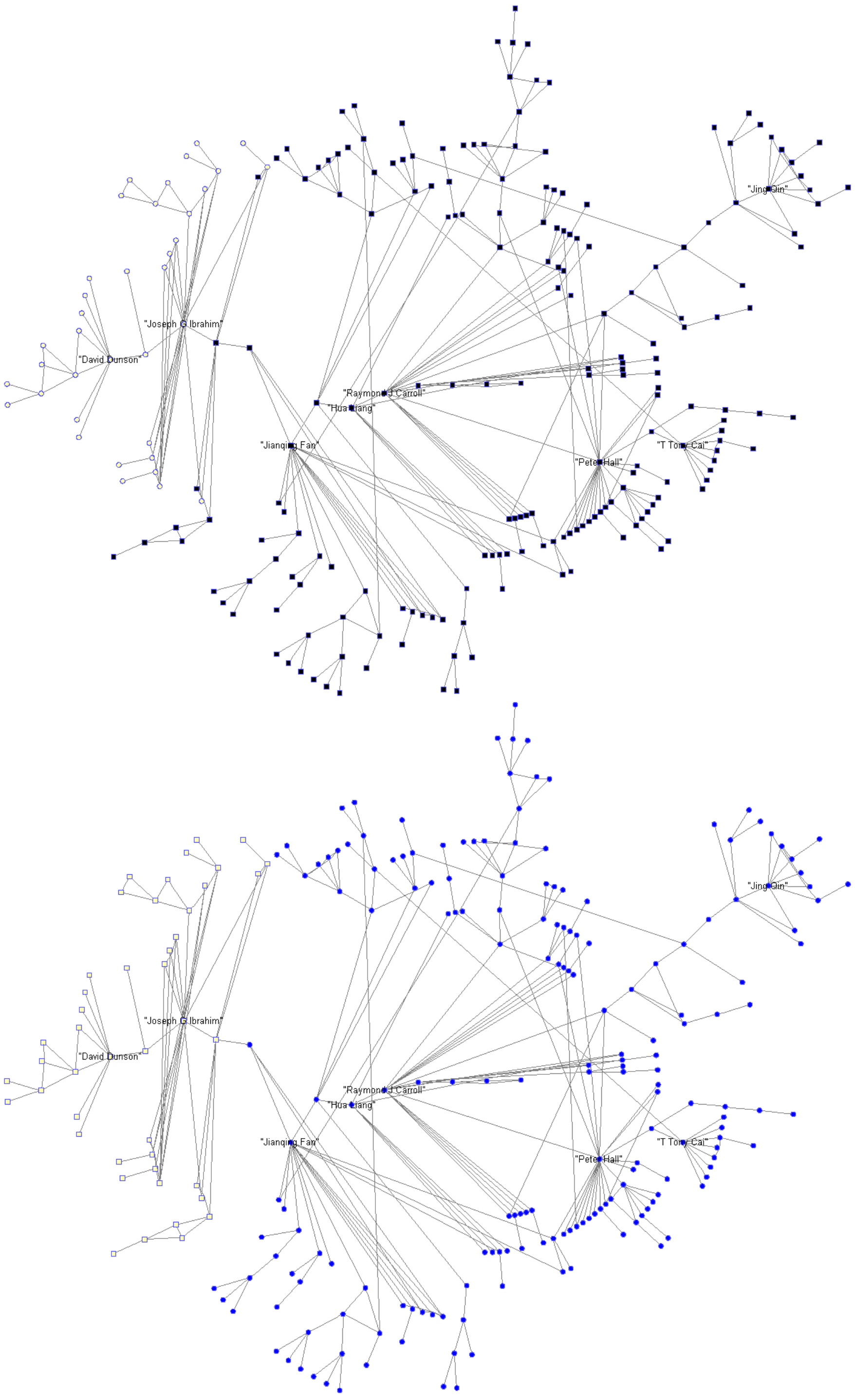

Figure 1: Setting the cluster numbers $K = 2$. Comparing Spectral method (top) and SCORE method (bottom), different color implies different clusters. Note that different Spectral method coincide when $K = 2$.

Firstly, we set the cluster number to be $K = 2$ (Figure 1), in this case, three different spectral methods give exactly same clusterings. But SCORE method performs a little different at the edge of two clusterings. More explanation may be further explored in final project.

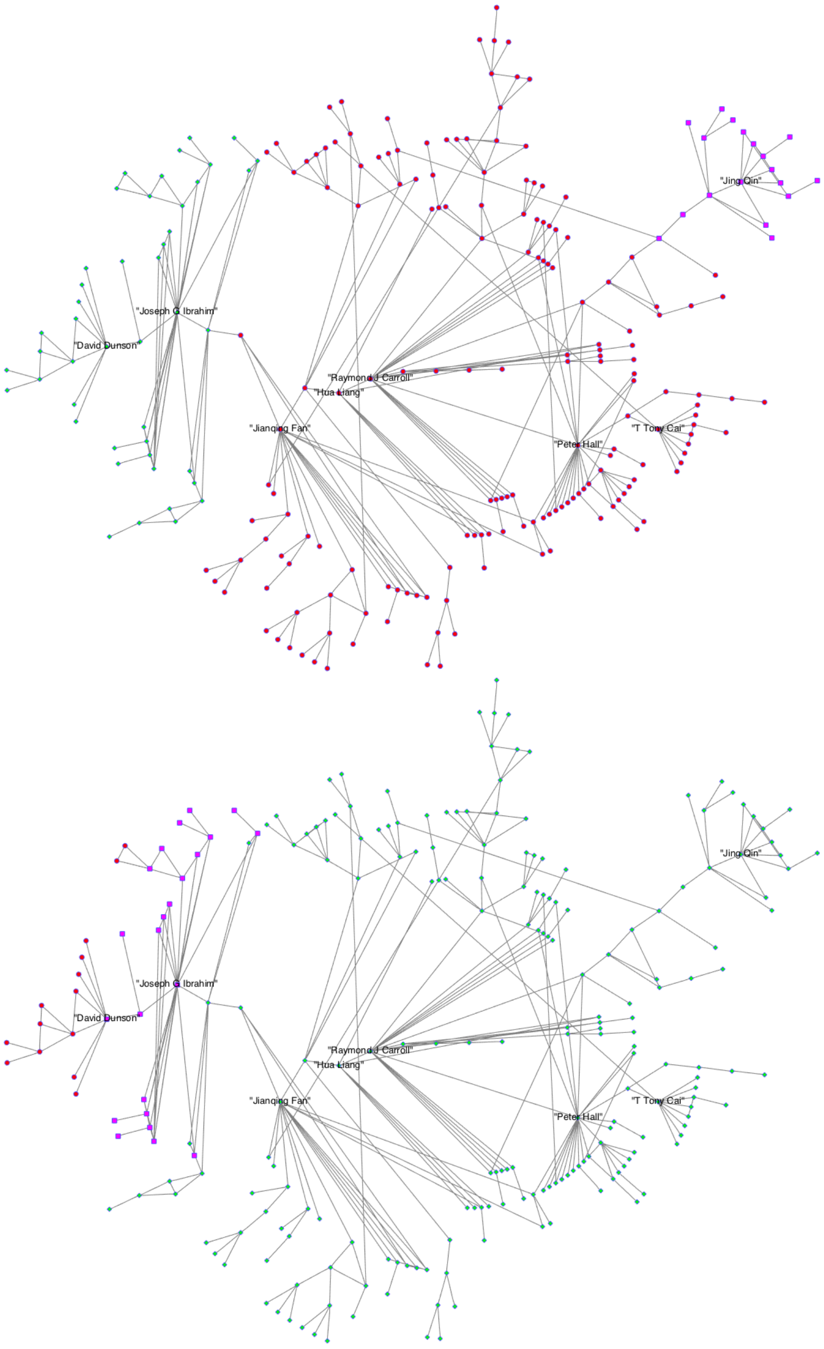

Figure 2: Setting the cluster numbers $K = 3$. Comparing Spectral method with Normalized Laplacian (top) and SCORE method (bottom), different color implies different clusters.Authors with more than 8 cooperators has been named in the figure

Then, the case of $K = 3$. Different methods present different incline for clustering, not only different in the edge. For example (Figure 2), spectral method(Normalized Laplacian) detect ”Divid Dunson”, ”Jianqing Fan” and ”Jing Qin” communities, as we explain for intuitive meaning of cutting, we may regard this three communities do not intend to cooperate with others. However, SCORE method further bipartition ”Divid Dunson” community instead of dectect ”Jing Qin” community, which means SCORE method are inclined to regard ”Jing Qin” and ”Jianqing Fan” as one.

We can introduce Adjuested Rand Index (ARI) and Variation Information (VI) to measure the similarity

Table 1: K = 2

	SCORE	SpCUL	SpCNL-sym	SpCNL-rw
SCORE	1.00/0.00	0.8399/27.3041	0.8399/27.3041	0.8399/27.3041
SpCUL		1.00/0.00	1.00/0.00	1.00/0.00
SpCNL-sym			1.00/0.00	1.00/0.00
SpCNL-rw				1.00/0.00

Table 2: K =3

	SCORE	SpCUL	SpCNL-sym	SpCNL-rw
SCORE	1.00/0.00	0.6023/45.4977	0.5924/46.2624	-0.0901/64.5352
SpCUL		1.00/0.00	0.9859/3.3394	0.3517/40.9162
SpCNL-sym			1.00/0.00	0.3422/44.2556
SpCNL-rw				1.00/0.00

of different clusterings. ARI and VI are given as follow:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}$$

$$VI(X;Y) = -\sum_{i,j} r_{ij}[\log(r_{ij}/p_i) + \log(r_{ij}/q_j)]$$

Lager ARI and less VI will ensure the consistency of the two partitions.

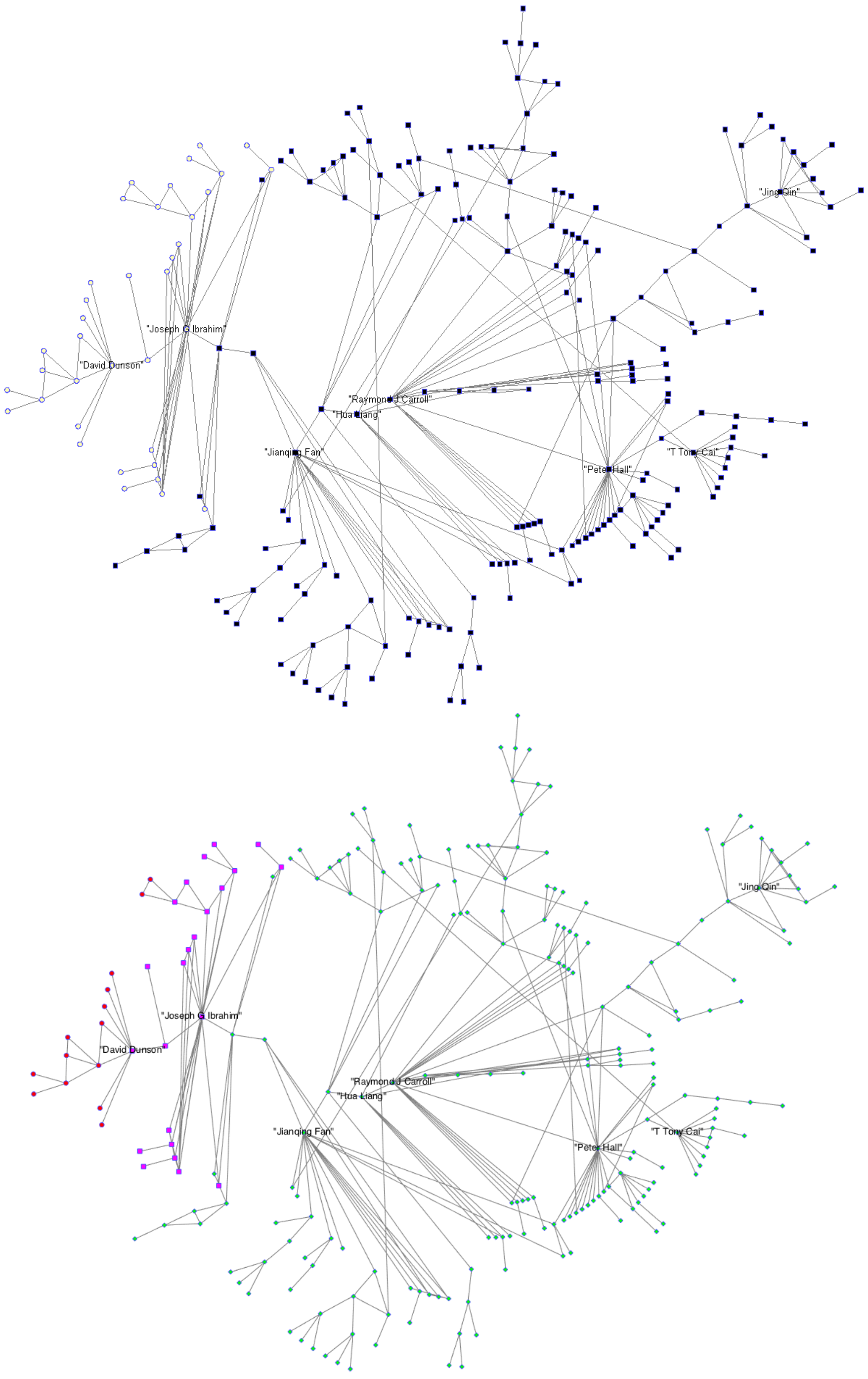

Figure 3: Focusing on SCORE method, $k = 2$ (top) and $k = 3$ (bottom) are shown.

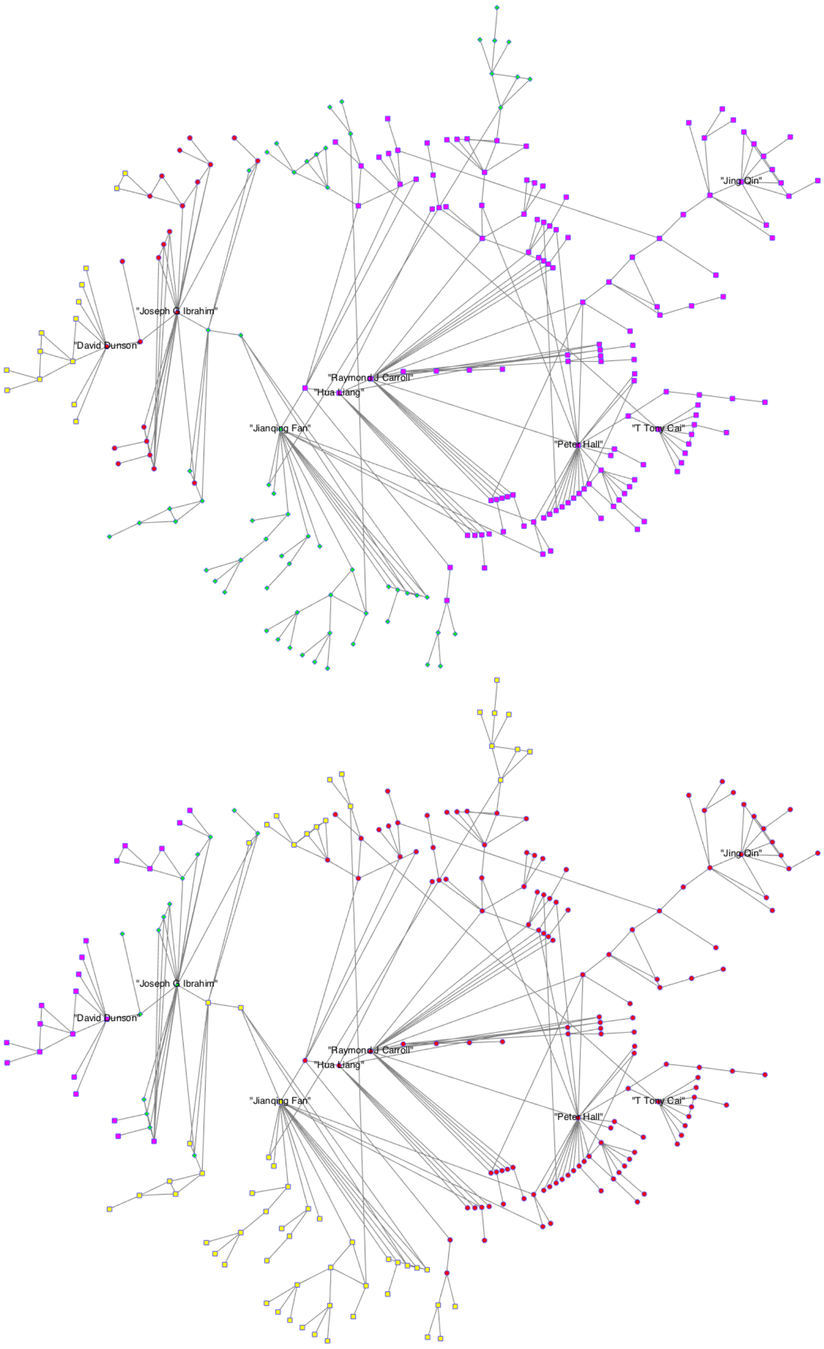

Figure 4: Focusing on SCORE method, $k = 4$ (top) and $k = 5$ (bottom) are shown.

Finally, we focus the SCORE Method to see the trand as k increase(Figure 3 and Figure 4). In fact, more clusterings are dectected, continue to $k = 3$, "Jianqing Fan" community is been furthered partitioned.

2 Citation community detection

Now we focus on the citation network CAG which is a directed graph. A efficient method for the communities detection will be introduced briefly below. Direct SCORE (D-SCORE) is commuity detection method which constructs two associated undirected graphs and then uses SCORE method to do the futher community clustering. For detailed explanations, see the forthcoming manuscript [23]. Given a directed network $\mathcal{N} = (V, E)$, assume $\mathcal{N}$ has K communities. Let A be the adjacency matrix, and let $\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_K$ and $\hat{v}_1, \hat{v}_2, \cdots, \hat{v} - K$ be the first K left singular vectors and the first K right singular vectors of A, respectively. Also, define two associated (undirected) networks with the same set of nodes as follows

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

Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be the giant components of the Citer network and Citee network, respectively. Define two $n \times (K-1)$ matrices $\hat{R}^{(l)} \hat{R}^{(r)}$ by

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

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

Note that all nodes split into four disjoint subsets:

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

D-SCORE clusters nodes in each subset separately.

1. $(\mathcal{N}_2 \cap \mathcal{N}_1)$. Restricting the rows of $\hat{R}^{(l)}$ and $\hat{R}^{(r)}$ to the set $\mathcal{N}_2 \cap \mathcal{N}_1$ and obtaining two matrices $\tilde{R}^{(l)}$ and $\tilde{R}^{(r)}$, we cluster all nodes in $\mathcal{N}_2 \cap \mathcal{N}_1$ by applying the k-means to the matrix $[\tilde{R}^{(l)}, \tilde{R}^{(r)}]$ assuming there are $\leq K$ communities.
2. $(\mathcal{N}_2 \backslash \mathcal{N}_1)$. 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 $\mathcal{N}_2 \backslash \mathcal{N}_1$, 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. $(\mathcal{N}_2 \backslash \mathcal{N}_1)$. 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. $(\mathcal{N} \backslash (\mathcal{N}_1 \cup \mathcal{N}_2))$. We say there is a weak-edge between i and j if there is an edge between i and j in the weakly connected citation network. By 1-2, all nodes in $\mathcal{N}_1 \cup \mathcal{N}_2$ partition into $\leq K$ communities. For each node in $\mathcal{N} \backslash (\mathcal{N}_1 \cup \mathcal{N}_2)$, we assign it to the community to which it has the largest number of weak-edges.

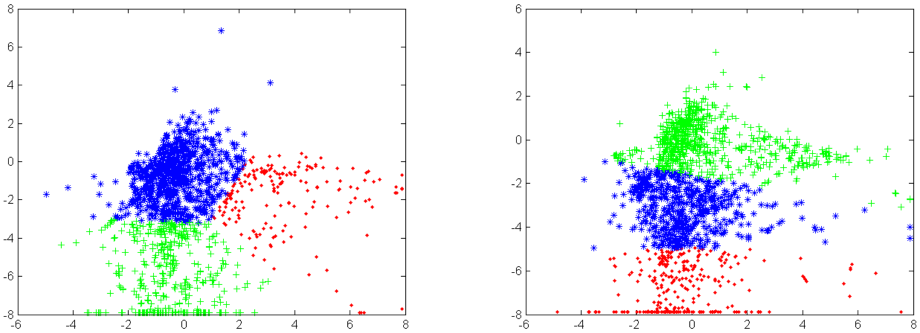

Figure 5: Left, do clustering by the left singular vector ratios; Right, do clustering by the left singular vector ratios.

From Figure 5, we see the communities detected by left and right singular vector ratios have no significant difference. This also confirms the D-SCORE method is reasonable.

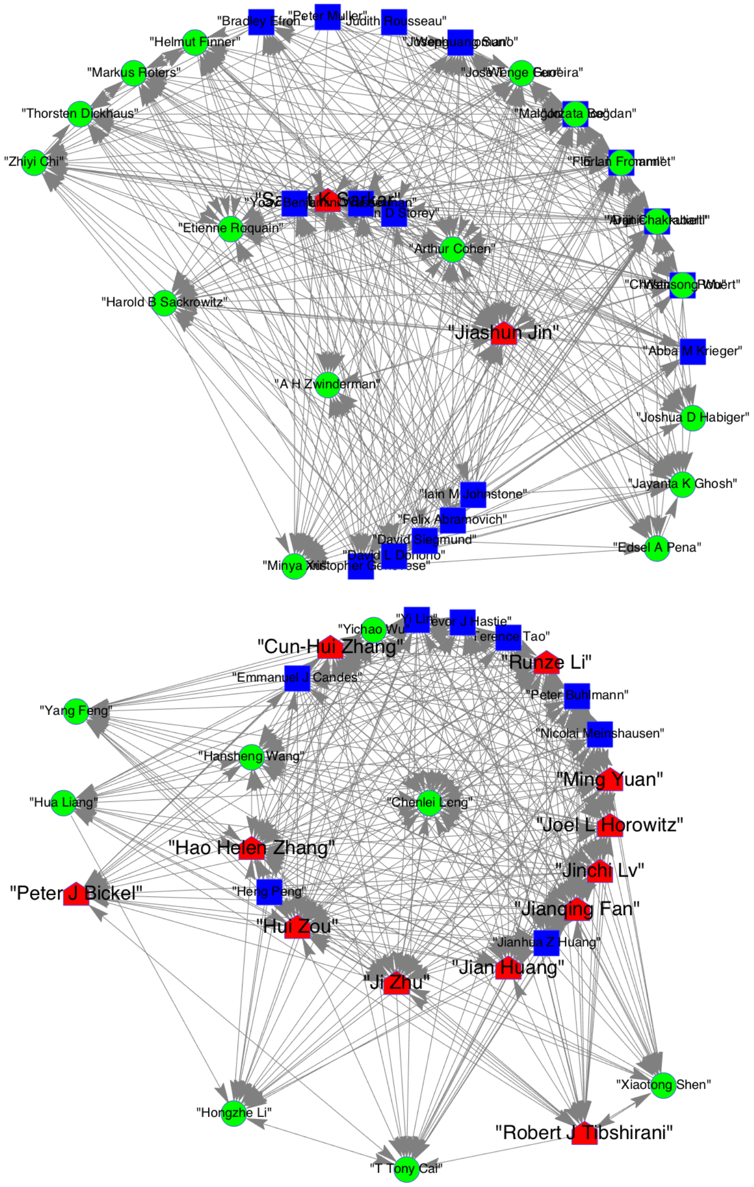

Figure 6: K = 3, Communities with 1285 nodes, 359 nodes and 1010 nodes respectively. Top, community with 1285 nodes, only nodes with most 10 authority and hub are shown; Bottom, community with 1285 nodes, only nodes with most 10 authority and hub are shown.

1-community top20aut: ”Jianqing Fan”, ”Chenlei Leng”, ”Ji Zhu”, ”Hui Zou”, ”Hua Liang”, ”Runze Li”, ”T Tony Cai”, ”Hansheng Wang”, ”Cun-Hui Zhang”, ”Hongzhe Li”, ”Ming Yuan”, ”Hao Helen Zhang”, ”Robert J Tibshirani”, ”Yang Feng”, ”Jian Huang”, ”Joel L Horowitz”, ”Jinchi Lv”, ”Xiaotong Shen”, ”Yichao Wu”, ”Peter J Bickel”.

top20hub:”Jianqing Fan”, ”Hui Zou”, ”Peter Buhlmann”, ”Nicolai Meinshausen”, ”Yi Lin”, ”Ming Yuan”, ”Trevor J Hastie”, ”Runze Li”, ”Emmanuel J Candes”, ”Cun-Hui Zhang”, ”Heng Peng”, ”Jian Huang”, ”Peter J Bickel”, ”Terence Tao”, ”Jinchi Lv”, ”Robert J Tibshirani”, ”Ji Zhu”, ”Hao Helen Zhang”, ”Joel L Horowitz”, ”Jianhua Z Huang”.

2-community

top20aut: "Jayanta K Ghosh", "Arijit Chakrabarti", "Florian Frommlet", "Malgorzata Bogdan", "Jiashun Jin", "Etienne Roquain", "Sanat K Sarkar", "Thorsten Dickhaus", "Helmut Finner", "Markus Roters", "Zhiyi Chi" "Minya Xu", "Harold B Sackrowitz", "Arthur Cohen", "Wensong Wu", "Joshua D Habiger", "Edsel A Pena", "Wenge Guo", "A H Zwinderman", "Jose T A S Ferreira".

top20hub: "John D Storey", "Larry Wasserman", "Christopher Genovese", "Yoav Benjamini", "David Siegmund", "David L Donoho", "Bradley Efron", "Iain M Johnstone", "Jiashun Jin", "Sanat K Sarkar", "Daniel Yekutieli" "Felix Abramovich", "Abba M Krieger", "Joseph P Romano", "E L Lehmann", "John Rice", "Wenguang Sun", "Christian P Robert", "Judith Rousseau", "Peter Muller".

As weve shown in Figure 8, three communities are detected, we do some remarks as follow:

1. Almost all Top-20-aut and Top-20-hub are in the same community, community with 359 nodes. We conclude this community as a Self-Loop community. They merely cite researcher out of this community for their Authority in this area.

2. As [2] suggests, three communities detected here could regarded as Large-Scale Multiple Testing ,Variable Selection and Spatial and Semi-parametric/Nonparametric Statistics communities.

3. Citation community detection base on citee and citer, while coauthorship community detection base on operation, they differs a lot.

References

[1] Bickel, P. J., and A. Chen, *A nonparametric view of network models and Newman-Girvan and other modularities*, Proceedings of the National Academy of Sciences of the United States of America 106.50(2009):21068-73.

[2] Jin, Jiashun, *Fast community detection by SCORE*, Annals of Statistics 43.2(2012):672-674.

[3] Ji, Pengsheng, and J. Jin, *Coauthorship and Citation Networks for Statisticians*, Eprint Arxiv (2014).