Community Detection by SCORE

with applications to Statisticians' Networks

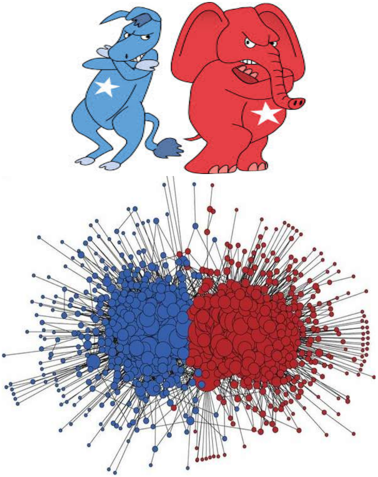
Jiashun Jin

Statistics Department
Carnegie Mellon University

Collaborators: Pengsheng Ji (Univ. of Georgia)
Zheng Tracy Ke (Univ. of Chicago)

April 6, 2015

Network community detection



Political web blogs (Adamic and Glance; 2005)

- $n = 1222$ web blogs (nodes)
- 16714 hyperlinks (edges)
- $\#\{edges\} \ll n^2$: adjacency matrix X is very sparse
- Two perceivable communities
- **Goal**. Find the (unknown) community labels

Abstraction (undirected)

Data: adjacency matrix A of a network $\mathcal{N} = (V, E)$

- $V = \{1, 2, \ldots, n\}$: nodes

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

- K perceivable "communities"

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

Goal. For each node, predict the community label.

Diagonals of A are 0 for convenience

Signal and noise decomposition

Adjacency matrix : $A = E[A] + W, \quad W \equiv (A - E[A]), \quad$ "*signal*" + "*noise*"

- $W = A - E[A]$: generalized Wigner matrix
  - upper triangles: independent centered-Bernoulli
- **Question:** How to model Ω if we write

$$E[X] = \Omega - \text{diag}(\Omega)$$

Box's wisdom

George E.P. Box (1919–2013)

"All models are wrong, but some are useful"

Degree Corrected Block Model (DCBM)

$$\frac{\Omega(i,j)}{\theta(i)\cdot\theta(j)} = P(k,\ell) \qquad \Longleftrightarrow \qquad \Omega = \Theta L \Theta$$

$$P = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \qquad \Theta = \begin{bmatrix} \theta(1) & & & \\ & \theta(2) & & \\ & & \ddots & \\ & & & \theta(7) \end{bmatrix}$$

$$L = \begin{bmatrix} a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \end{bmatrix} \xrightarrow{permute} \begin{bmatrix} a & a & a & a & b & b & b \\ a & a & a & a & b & b & b \\ a & a & a & a & b & b & b \\ a & a & a & a & b & b & b \\ b & b & b & b & c & c & c \\ b & b & b & b & c & c & c \\ b & b & b & b & c & c & c \end{bmatrix}$$

Karrer and Newman (2010)

Tukey's suggestion

John W. Tukey (1915–2000)

"Which part of the sample contains the information"

Tukey (1965), PNAS

Where is the information?

$$A = \Omega - \text{diag}(\Omega) + W \ \approx \ \Omega$$

$$SVD: \qquad \Omega = \Theta L \Theta = U_{n,K} D_{K,K} (U_{n,K})'$$

$$U_{n,K} = \Theta T_{n,K} = \begin{bmatrix} \theta(1) & & & \\ & \theta(2) & & \\ & & \ddots & \\ & & & \theta(n) \end{bmatrix} \begin{bmatrix} s_1 & t_1 \\ s_2 & t_2 \\ s_1 & t_1 \\ \vdots & \vdots \\ s_1 & t_1 \\ s_2 & t_2 \end{bmatrix}$$

SCORE: algorithm

SCORE: **S**pectral **C**lustering **O**n **R**atios-of-**E**igenvectors

Input: A and K

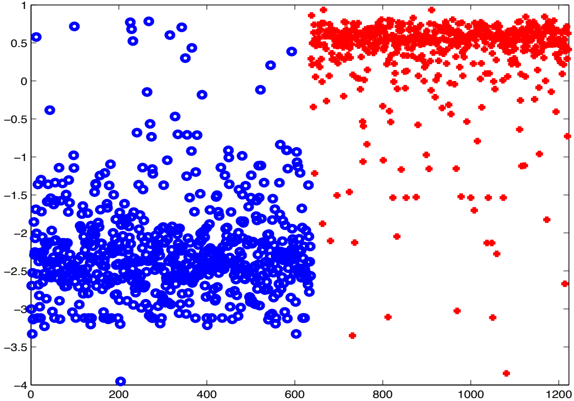
- Obtain leading eigenvectors $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_K$
- Obtain $n \times (K-1)$ matrix of *entry-wise ratios*

$$\hat{R}(i,k) = \frac{\hat{\eta}_{k+1}(i)}{\hat{\eta}_1(i)}, \qquad 1 \le i \le n, \;\; 1 \le k \le K-1$$

- Apply k-means to $\hat{R}$ (assume $\le K$ clusters)

Political weblog network ($K = 2$)

x-axis: $i = 1, 2, \ldots, n$; y-axis: $\hat{R}(i)$; 58 errors (lowest in literature)



Methods	SCORE	PCA	normalized PCA	NSC	BCPL
Errors	58	437	600	69	104.5 (SD: 145.4)

Newman (2016), Bickel and Chen (2009), Zhao et al (2012)

Regularity conditions

$$A = \Omega - \text{diag}(\Omega) + W, \qquad \Omega = \Theta L \Theta, \qquad L = \sum_{k,\ell=1}^{K} P(k,\ell) \mathbf{1}_k \mathbf{1}_\ell'$$

- *(a)*. Eigen-spacing of DPD is $\geq$ a constant C

$$D(k,k)^2 = \big[\sum_{i \in V^{(k)}} \theta(i)^2 \big] / \|\theta\|^2$$

- *(b)*. $\log(n)\theta_{max}\|\theta\|_1/\|\theta\|^4 \to 0$, so that

$$\|W\| \ll \|\Omega\|, \qquad \text{with prob. } 1 - o(n^{-3})$$

- *(c)*. $\log(n)\theta_{max}^2/\theta_{min} \leq \|\theta\|_3^3$, so matrix-form Bernstein inequality holds (for the sum of random matrices)

Consistency of SCORE

$$\mathrm{Hamm}_p(\hat{\ell}, \ell) = n^{-1} \min_{\pi} \sum_{i=1}^{n} P\big(\hat{\ell}_i \neq \pi(\ell_i)\big), \quad err_n = \frac{\|\theta\|_3^3}{\|\theta\|^4} \max\{\sum_{i=1}^{n} \frac{1}{\theta(i)}, \frac{1}{\theta_{min}} (\frac{\|\theta\|_1}{\|\theta\|^2})^2\}$$

Theorem. Consider DCBM where (a)-(c) hold. As $n \to \infty$, if $n_*^{-1} \log(n) err_n \to 0$, where n_* is the minimum community size, then $\mathrm{Hamm}_p(\hat{\ell}^{score}, \ell) \leq C n^{-1} \log^3(n) err_n$.

Proof. Full analysis of $\Theta^{-1}(\hat{\eta}_k - \eta_k)$

- Spectral perturbation theory
- Classical large deviations inequalities
- Matrix-form Bernstein inequality (Tropp, 2012)

Remark. *If we assume* $\theta(i) \stackrel{iid}{\sim} F$ *as in Zhao et al (2012), then* $\mathrm{Hamm}_p(\hat{\ell}^{score}, \ell) \leq C n^{-1} \log^3(n)$

Coauthor/Citation Networks (statisticians)

- People most interested: statisticians/friends
- We know "inside information" N/A to outsiders

Scientific Problem: Dynamics of US-based statisticians in theory & methods of the HDDA era

HDDA: High-Dimensional Data Analysis

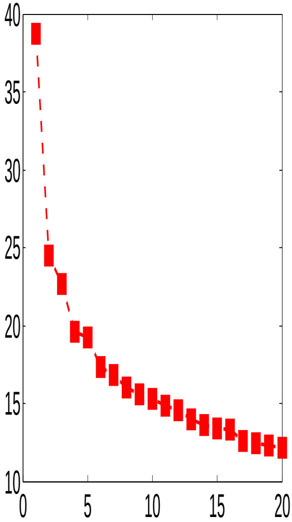
Data: All published research papers in AoS, Biometrika, JASA, and JRSS-B, 2003–2012

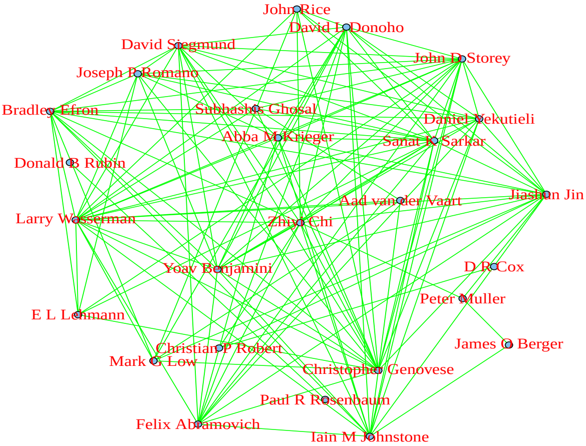
Disclaimer

- Data and scope of scientific interests: limited
- It is not our intention to
  - rank one author/paper/area over the others
  - label an author/paper to a certain area
- We have to use real names because the networks are for real people ("us")

Citation Network, I

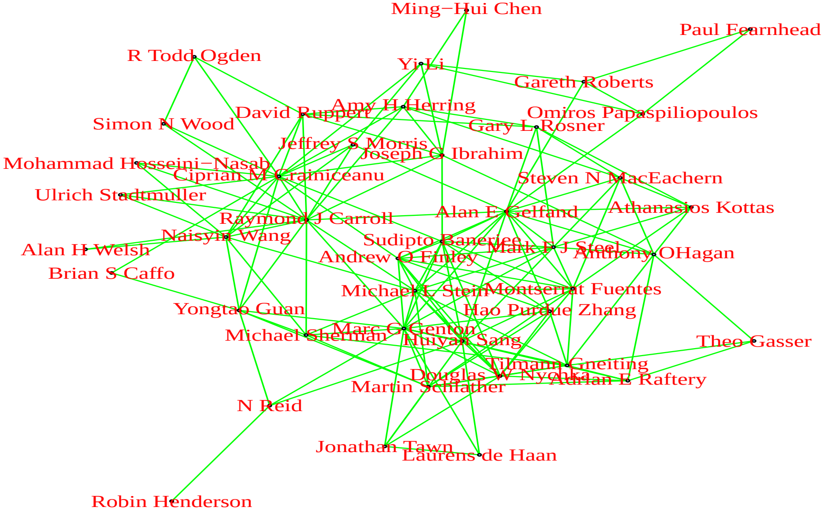
Large-Scale Multiple Testing by SCORE (359 nodes; 26 shown)





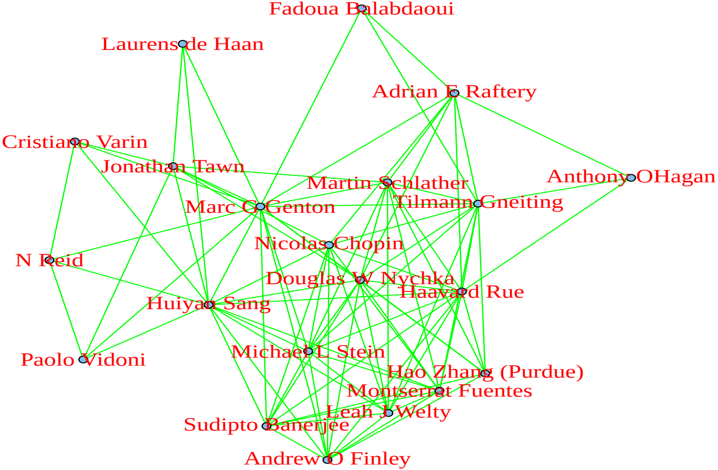
Citation Network, II

Spatial stat./nonparametric stat. by SCORE (1010 nodes; 42 shown)



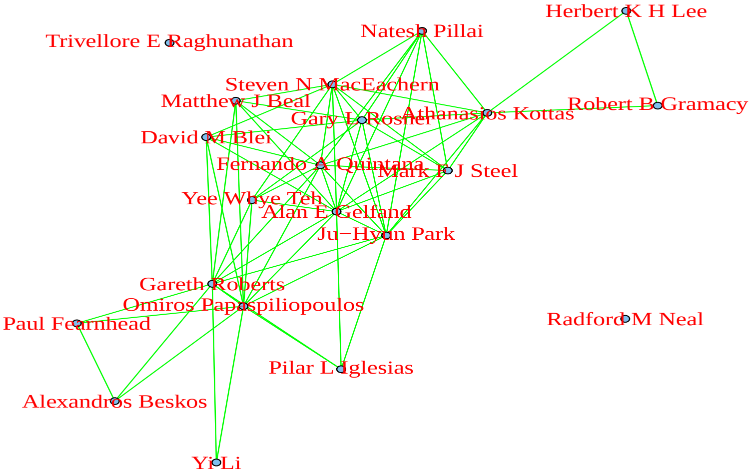
Citation Network, II (further split, I)

Parametric Spatial Statistics by SCORE (304 nodes; 21 shown)



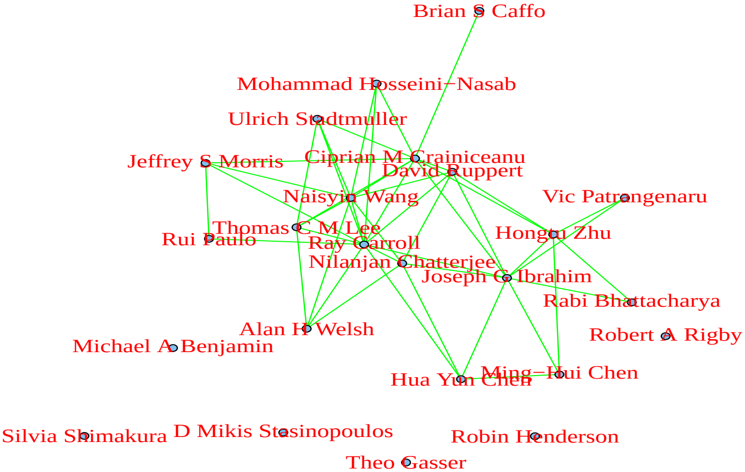
Citation Network, II (further split, II)

Nonparametric Spatial Statistics by SCORE (212 nodes; 21 shown)



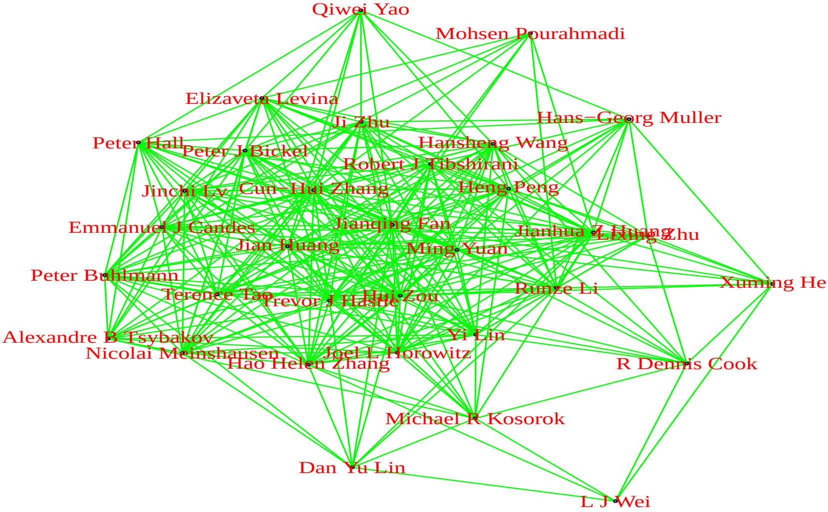
Citation Network, II (further split, III)

Non-parametrics/semi-parametrics by SCORE (392 nodes; 24 shown)



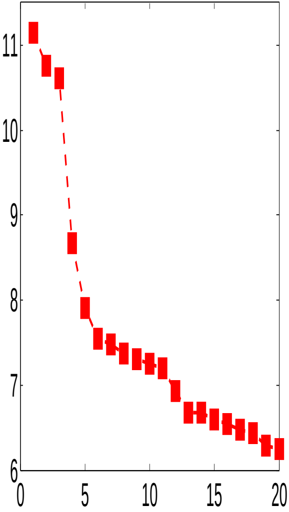
Citation Network, III

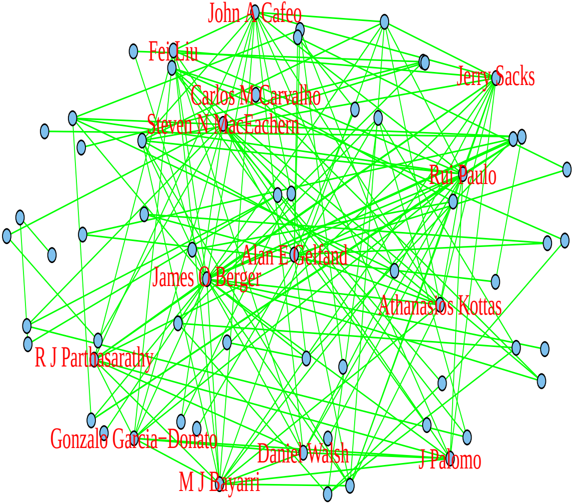
Variable Selection by SCORE (1285 nodes; 40 shown)



Coauthorship Network, I

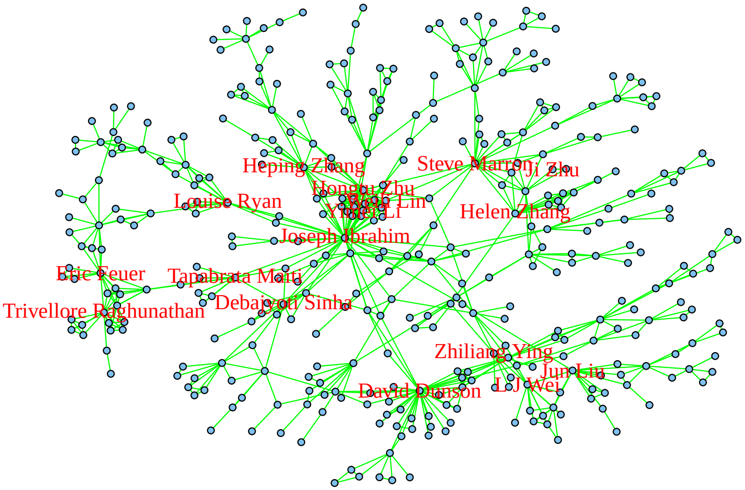
Objective Bayes by SCORE (64 *nodes;* 14 *shown*)





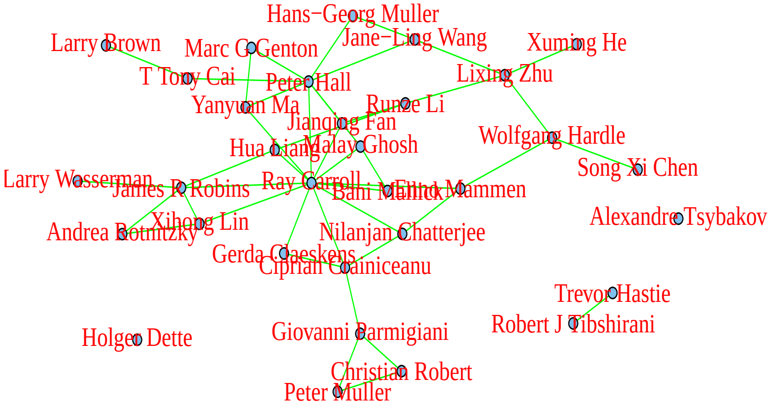
Coauthorship Network, II

*Biostatistics by SCORE (*388 *nodes;* 16 *shown)*



Coauthorship Network, III

*HDDA by SCORE (*1811 *nodes,* 32 *shown)*



Comparisons, I

Undirected networks:

- Newman's Spectral Clustering (NSC)
- Bickel and Chen's Profile Likelihood (BCPL)
- Amini et al's Pseudo Likelihood (APL)

Directed networks: Leicht & Newman's Spectral Clustering

Comparisons, II

Adjusted Rand Index (ARI); larger means more similar

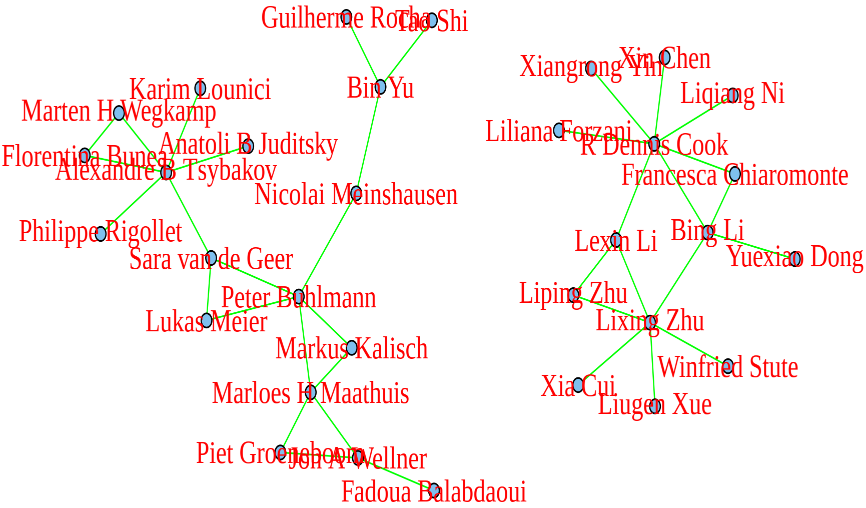
	SCORE	NSC	APL	BCPL
SCORE	1.00	.55	.19	.00
NSC		1.00	.41	.00
APL			1.00	0.00
BCPL				1.00

Sizes of the 3 communities identified by SCORE, NSC, and APL

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

More on Coauthorship Network, I

"Theo. Statist. Learning" (15 nodes) and "Dim. Reduction" (14 nodes)



More on Coauthorship Network, II

"Johns Hopkins", "Duke", "Stanford", "Quant. Reg.", "Exp. Design"

Barry Rowlingson
Brian S Caffo
Chong-Zhi Di
Ciprian M Crainiceanu
David Ruppert
Dobrin Marchev
Galin L Jones
James P Hobert
John P Buonaccorsi
John Staudenmayer
Naresh M Punjabi
Peter J Diggle
Sheng Luo

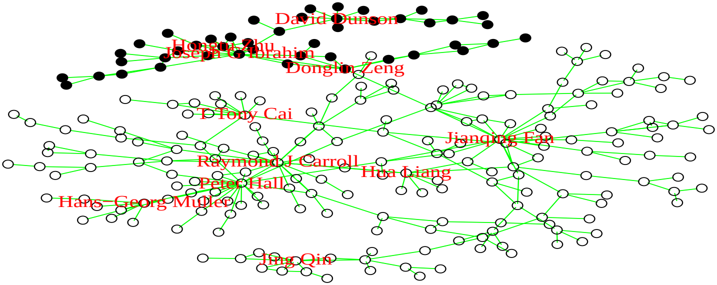
Carlos M Carvalho
Gary L Rosner
Gerard Letac
Helene Massam
James G Scott
Jonathan R Stroud
Maria De Iorio
Mike West
Nicholas G Polson
Peter Muller

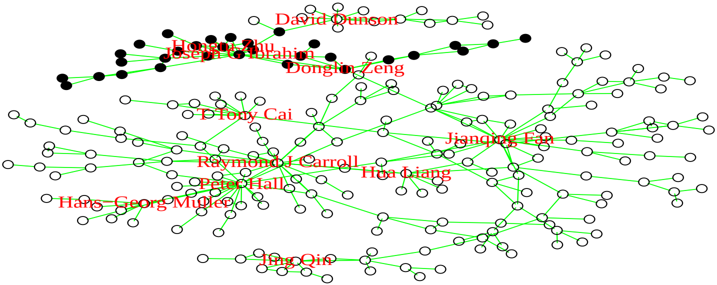
Armin Schwartzman
Benjamin Yakir
David Siegmund
F Gosselin
John D Storey
Jonathan E Taylor
Keith J Worsley
Nancy Ruonan Zhang
Ryan J Tibshirani

Hengjian Cui
Huixia Judy Wang
Jianhua Hu
Jianhui Zhou
Valen E Johnson
Wing K Fung
Xuming He
Yijun Zuo
Zhongyi Zhu

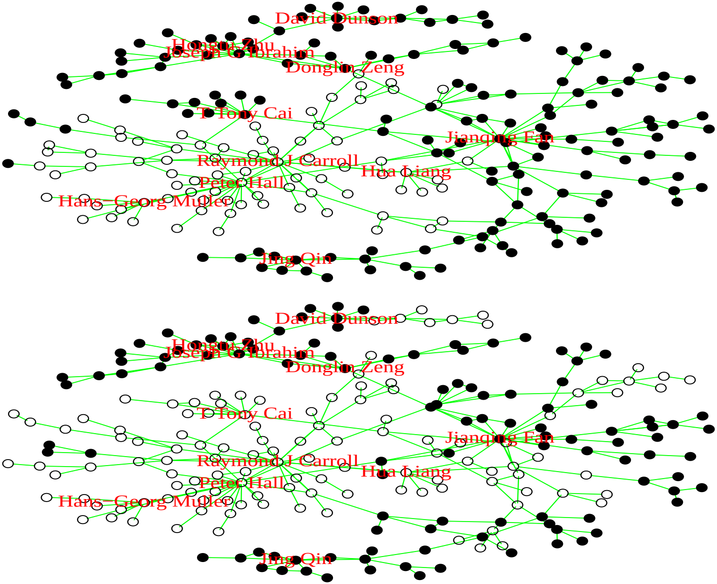
Andrey Pepelyshev
Frank Bretz
Holger Dette
Natalie Neumeyer
Stanislav Volgushev
Stefanie Biedermann
Tim Holland-Letz
Viatcheslav B Melas

More on Coauthorship Network, III





More on Coauthorship Network, IV



Take home messages

- Proposed a fast, flexible, easy-to-implement, yet effective, method: SCORE
- Successfully applied to Statisticians' networks and found many meaningful communities
- Data sets: a fertile ground for future research (many results are not reported here)

References:

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

Ji P, Jin J (2014) Coauthorship and Citation networks for statisticians. *arXiv.1410.2840*.