Statistical Inference on Membership Profiles in Large Networks

Jianqing Fan

Princeton University

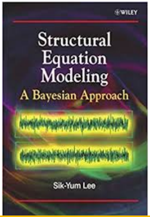
with **Yingying Fan, Xiao Han, Jinchi Lv**

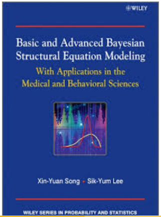
In Memory of Sik-Yum Lee

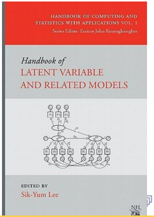
- ★ since 1995; colleagues for 5 years; bridge games;
- ★ Kind, generous, quiet, sporty
- ★ Great scholars: ASA fellow, ICSA award

Name	School	Year	Descendants
Shi, Jian Qing	Chinese University of Hong Kong	1996	
Song, Xin-Yuan	Chinese University of Hong Kong	2001	1
Zhang, Wenyang	Chinese University of Hong Kong	1999	6
Zhu, Hongtu	Chinese University of Hong Kong	2000	22

According to our current on-line database, Sik-Yum Lee has 4 students and 33 descendants.



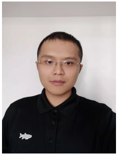




Outline

1. Introduction
2. Mixed Membership Models
3. Network Inference under degree **homogeneity**
4. Network Inference under degree **heterogeneity**
5. Numerical Studies

Yingying Fan

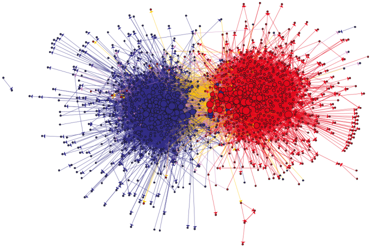


Xiao Han

Jinchi Lv

Introduction

A Networked World



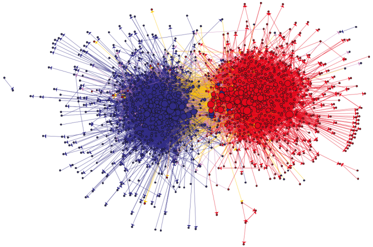
★citation ★social, ★trade ★economic ★gene regulatory, $\cdots$

Data: adjacency matrix $\mathbf{X} \in \{0, 1\}^{n \times n}$

How to quantify uncertainty

that a given pair of nodes are in the same community?

A Networked World



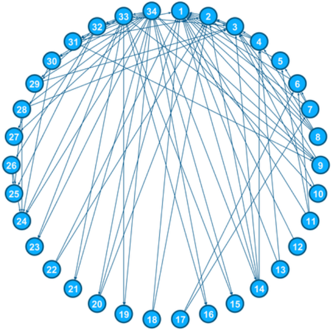
★citation ★social, ★trade ★economic ★gene regulatory, $\cdots$

Data: adjacency matrix $\mathbf{X} \in \{0,1\}^{n \times n}$

How to quantify uncertainty

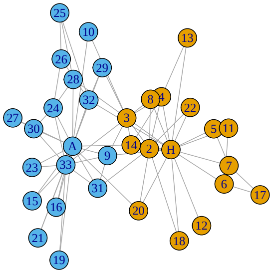
that a given pair of nodes are in the same community?

A Motivating Example



- A university *karate club network* data (Zachary, 1977) for 34 members (Girvan and Newman, 2002)
- Edge links two members spent much time together outside club meetings

A Network with Non-Overlapping Communities

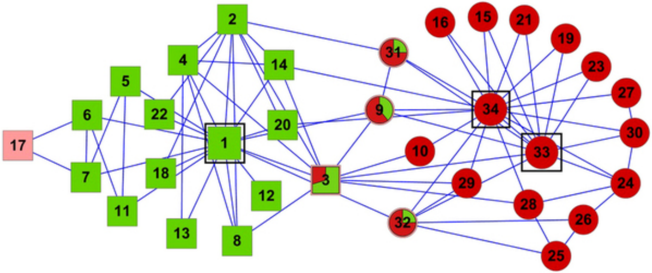


★Network structure obtained based on stochastic block model via spectral clustering

What if a different model is used?

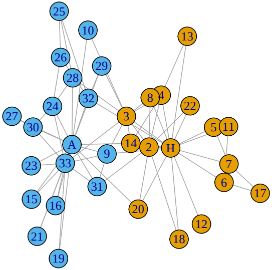
A Network with Overlapping Communities

Mixed membership model: Each node now equipped with a vector of **membership probabilities**

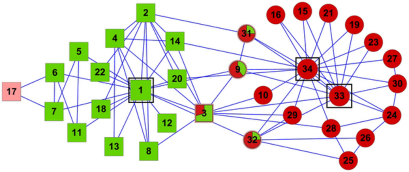


★Communities using mixed membership model

Uncertainty quantification



(a) Non-overlapping



(b) Overlapping

Can we quantify the uncertainties of links?

A Sneak Peek of Our Results

P-values for pairwise comparison

	7	8	9	10	27
7	1.0000	0.1278	0.0012	0.0685	0.0145
8	0.1278	1.0000	0.0026	0.0052	0.0000
9	0.0012	0.0026	1.0000	0.3308	0.0540
10	0.0685	0.0052	0.3308	1.0000	0.4155
27	0.0145	0.0000	0.0540	0.4155	1.0000

How to get these P-values?

Applications: ★Dim-reduction ★network centrality

Connections with Factor-adjusted sparsity

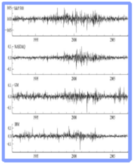
Data: $\{\mathbf{X}_t\}_{t=1}^n$

Factor model: $\mathbf{X}_t = \mu + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$

Assumption: $\boldsymbol{\Sigma}_u$ or $\boldsymbol{\Sigma}_u^{-1}$ sparse

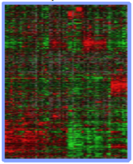
Time Points

Stocks



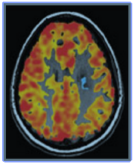
Experiments

Genes



Brain Scans

Voxels



Modeling sparsity: $\boldsymbol{\Sigma}_u^{-1} \equiv \Omega = \alpha \mathbf{I}_p + \beta \mathbf{L}_p$

Graph Laplancian: $\mathbf{L}_p = \mathbf{I}_p - D^{-1/2}\mathbf{X}D^{-1/2}$, $\quad \mathbf{D} = \mathrm{diag}(d_1, \cdots, d_p)$

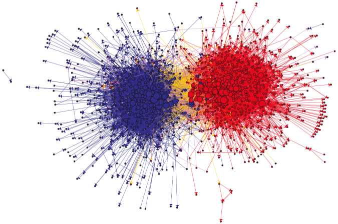
- $\omega_{ij} = 0 \quad \Longleftrightarrow \quad$ an edge
- Communities of nodes can be learned and inferenced.

Related Literature

- **Community detection**: ★**Algorithms**: Newman (2013a,b), Zhang and Moore (2014), ... ★**SMB**: Holland et al. (1983), Wang and Wong (1987), Bickel and Chen (09, 12), Abbe (2017), Li, Levina, Zhu (2019); ★**Degree-Corrected SBM** Karrer and Newman (2011); Zhao, Levina, and Zhu (2012), ★**Mixed Member**: Airoldi et al. (2008); ...
- **Spectral methods**: Rohe et al. (2011), Lei and Rinaldo (2015), Jin (2015), Abbe et al. (2017), ...
- **Hypothesis testing**: Bickel and Sarkar (2016), Lei (2016), Wang and Bickel (2017), ...
- **Link prediction** Liben-Nowell and Kleinberg (2007), Wu et al. (2018),...

Mixed Membership Models

Stochastic Block Model



K disjoint communities $\mathcal{C}_1, \cdots, \mathcal{C}_K$, with
$P(X_{ij} = 1) = p_{kl}$, for $i \in \mathcal{C}_k, j \in \mathcal{C}_l$, indep.
Edge probability: $\mathbf{P} = (p_{i,j})_{K \times K}$.

Degree-corrected: $P(X_{ij} = 1) = \theta_i \theta_j p_{kl}, \qquad i \in \mathcal{C}_k, j \in \mathcal{C}_l$.

Erdös-Rényi graph: $p_{ij} = p$, **degenerate**

Mixed Membership Profile

■ Each node *i* has

$$\mathbb{P}(\text{node } i \text{ belongs to community } \mathbf{C_k}) = \pi_\mathbf{i}(\mathbf{k})$$

★ Probability vector $\pi_\mathbf{i} = (\pi_i(1), \cdots, \pi_i(K))^T \in \mathbb{R}^K$ is the **membership profile**

★ $\pi_i = e_\ell$ reduces to communication detection.

Hypothesis testing: For any two members,

$$H_0 : \pi_i = \pi_j \quad \text{vs.} \quad H_1 : \pi_i \neq \pi_j$$

Mixed Membership Model

Adjacency matrix $\mathbf{X} = (X_{ij}) \in \mathbb{R}^{n \times n}$,

(*Bhattacharyya and Bickel, 2016; Abbe, 2017; Le, Levina and Vershynin, 2018*)

$$X_{ij} \sim_{indep} \text{Bernoulli}(h_{ij}), \qquad \text{for } i > j$$

Connection Probability: (*Airoldi, Blei, Fienberg and Xing, 2008*)

$$P(X_{ij} = 1 | i \in \mathcal{C}_k, j \in \mathcal{C}_l) = \theta_i \theta_j p_{kl},$$

★$\mathbf{P} = (p_{kl}) \in \mathbb{R}^{K \times K}$ is nonsingular irreducible symmetric, $p_{kl} \in [0, 1]$.

Link with data

Edge probability

$$P(X_{ij} = 1) = \theta_i \theta_j \sum_{k=1}^{K} \sum_{l=1}^{K} \pi_i(k) \pi_j(l) p_{kl} = h_{ij}.$$

Mixed Membership Model: With $\mathbf{\Pi} = (\pi_1, \cdots, \pi_n)^T \in \mathbb{R}^{n \times K}$

$$\mathbf{X} = \mathbf{H} + \mathbf{W}, \qquad \mathbf{H} = \mathbf{\Theta \Pi P \Pi}^T \mathbf{\Theta},$$

★$\mathbf{\Theta} = \mathrm{diag}(\theta_1, \cdots, \theta_n)$, ★$\mathbf{W} = \mathbf{X} - E\mathbf{X}$ is generalized Wigner matrix

- Assume number of communities K is finite but **unknown**
- Including SBM as a special case

Flexible Network Inference

under degree homogeneity

Connections with spectral method

Assumption: $\boldsymbol{\Theta} = \sqrt{\theta}\mathbf{I}_n, \qquad \theta \to 0.$

$$\mathbb{E}\mathbf{X} = \mathbf{H} = \theta \underbrace{\overset{n\times K}{\boldsymbol{\Pi}} \ \mathbf{P} \ \boldsymbol{\Pi}^T}_{\text{rank } K} = \theta \begin{pmatrix} \boldsymbol{\pi}_1{}^T\mathbf{P}\pi_1 & \boldsymbol{\pi}_1{}^T\mathbf{P}\pi_2 & \cdots & \boldsymbol{\pi}_1{}^T\mathbf{P}\pi_n \\ \boldsymbol{\pi}_2{}^T\mathbf{P}\pi_1 & \boldsymbol{\pi}_2{}^T\mathbf{P}\pi_2 & \cdots & \boldsymbol{\pi}_2{}^T\mathbf{P}\pi_n \\ & & \cdots & \end{pmatrix}.$$

★ Eigenspace of $\mathbf{H}$ = column space spanned by $\boldsymbol{\Pi}$

Eigen-structures

★ **Population** Eigen-decomposition: $\mathbf{H} = \mathbf{V}\mathbf{D}\mathbf{V}^T$

- $\mathbf{D} = \text{diag}(d_1, \ldots, d_K)$ with $|d_1| \geq \cdots \geq |d_K| > 0$.
- $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_K) \in \mathbb{R}^{n \times K}$ is orthonormal matrix of eigenvectors

★ Rows of $\mathbf{V}$ are the same if $\pi_i = \pi_j$ by permutation

★ If $\{\pi_i\}_{i=1}^n$ has m clusters, rows of $\mathbf{V}$ have also m clusters.

↳ k-mean

★ **Sample** Eigen-decomposition: $\mathbf{X} = \widehat{\mathbf{V}}_n \widehat{\mathbf{D}}_n \widehat{\mathbf{V}}_n^T$

- WOLG, assume $|\widehat{d}_1| \geq \cdots \geq |\widehat{d}_n|$ and let $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_K) \in \mathbb{R}^{n \times K}$
- can have n nonzero eigenvalues

An Ideal Test Statistic

- By permutation argument, $\pi_\mathbf{i} = \pi_\mathbf{j} \Longleftrightarrow \mathbf{V(i)} = \mathbf{V(j)}$

- **Ideal test statistic**:

$$T_{ij} = (\widehat{\mathbf{V}}(i) - \widehat{\mathbf{V}}(j))^T \boldsymbol{\Sigma}_1^{-1} (\widehat{\mathbf{V}}(i) - \widehat{\mathbf{V}}(j))$$

- $\boldsymbol{\Sigma}_1$ is asymptotic variance — **challenge to derive**

$$\boldsymbol{\Sigma}_1 = \text{cov}((\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{WVD}^{-1})$$

Technical Conditions

A1) $\min_{1\le i\le K-1} \frac{|d_i|}{|d_{i+1}|} \ge 1 + c_0$, $\alpha_n^2 = \max_j \mathrm{var}(\sum_{i=1}^n X_{ij}) \to \infty$.

A2) $\lambda_K(\mathbf{\Pi}^T\mathbf{\Pi}) \ge c_1 n$, $\lambda_K(\mathbf{P}) \ge c_1$, and $\theta \ge n^{-c_2}$, $0 < c_1, c_2 < 1$.

A3) All eigenvalues of $n^2\theta\mathbf{\Sigma}_1$ are bounded away from 0 and ∞.

★ α_n measures sparsity of network

★ Node degree is of order $n\theta \ge n^{1-c_2}$ and A2) ensures

$$d_k \sim n\theta, \quad k = 1, \cdots, K$$

Asymptotic Distributions

Theorem 1: Assume A1)–A3).

a) Under **Null hypothesis** H_0,

$$T_{ij} \xrightarrow{d} \chi^2_K, \qquad \text{as } n \to \infty$$

b) Under **contiguous alternative** $\sqrt{n\theta}\|\pi_i - \pi_j\| \to \infty$, then

$$T_{ij} \xrightarrow{p} \infty.$$

c) If $\|\pi_i - \pi_j\| \sim \frac{1}{\sqrt{n\theta}}$, and $(\mathbf{V}(i) - \mathbf{V}(j))^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{V}(i) - \mathbf{V}(j)) \to \mu$, then

$$T_{ij} \xrightarrow{d} \chi^2_K(\mu)$$

Practical Test Statistic

■Replace K and $\mathbf{\Sigma}_1$ in $\mathbf{T_{ij}}$ by $\widehat{K}$ and $\widehat{\mathbf{S}}_1$ $\Longrightarrow$ $\widehat{\mathbf{T}}_{\mathbf{ij}}$.

Theorem 2: Assume that the following accuracy:

$$P(\widehat{K} = K) = 1 - o(1) \quad \text{and} \quad n^2\theta \|\widehat{\mathbf{S}}_1 - \mathbf{\Sigma}_1\|_2 = o_p(1).$$

Then, the same results as in Theorem 1 continue to hold for $\widehat{\mathbf{T}}_{\mathbf{ij}}$.

How to estimate K and Σ_1?

Practical Test Statistic

■Replace K and $\mathbf{\Sigma}_1$ in $\mathbf{T_{ij}}$ by $\widehat{K}$ and $\widehat{\mathbf{S}}_1$ $\Longrightarrow$ $\widehat{\mathbf{T}}_{\mathbf{ij}}$.

Theorem 2: Assume that the following accuracy:

$$P(\widehat{K} = K) = 1 - o(1) \quad \text{and} \quad n^2\theta\|\widehat{\mathbf{S}}_1 - \mathbf{\Sigma}_1\|_2 = o_p(1).$$

Then, the same results as in Theorem 1 continue to hold for $\widehat{\mathbf{T}}_{\mathbf{ij}}$.

How to estimate K and $\mathbf{\Sigma}_1$?

Estimation of Unknown Parameters

$$\widehat{K} = \#\left\{\widehat{d}_i : \quad \widehat{d}_i^2 > 2.01(\log n)\max_i \sum_{j=1}^n X_{ij}, \right\}$$

Proposition: The (a,b) entry of matrix $\mathbf{\Sigma}_1$ is

$$\frac{1}{d_a d_b}\left\{\sum_{t\in\{i,j\}}\sum_{l\notin\{i,j\}}\sigma_{tl}^2\mathbf{v}_a(l)\mathbf{v}_b(l)+\sigma_{ij}^2[\mathbf{v}_a(j)-\mathbf{v}_a(i)][\mathbf{v}_b(j)-\mathbf{v}_b(i)]\right\}$$

■Plug in: estimating $\sigma_{ab}^2 = \mathrm{var}(X_{ab})$ is somewhat complicated.

Estimating σ^2_{ab}

- $\widehat{w}^2_{0,ab}$ with $\widehat{\mathbf{W}}_0 = (\widehat{w}_{0,ab}) = \mathbf{X} - \underbrace{\sum_{k=1}^{\widehat{K}} \widehat{d}_k \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^T}_{\widehat{H}}$ is **not good** enough.

Refined estimator: Inspired by the expansion of $\widehat{d}_k$.

1. Calculate the initial estimator $\widehat{\mathbf{W}}_0$
2. Update the estimator of d_k by

$$\widetilde{d}_k = \left(\frac{1}{\widehat{d}_k} + \frac{\widehat{\mathbf{v}}_k^T \operatorname{diag}(\widehat{\mathbf{W}}_0^2) \widehat{\mathbf{v}}_k}{\widehat{d}_k^3} \right)^{-1}$$

shrinkage

3. Update the estimator of $\mathbf{W}$ as $\widehat{\mathbf{W}} = \mathbf{X} - \sum_{k=1}^{\widehat{K}} \widetilde{d}_k \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^T$.
   Estimate σ^2_{ab} as $\widehat{\sigma}^2_{ab} = \widehat{w}^2_{ab}$

Consistency of estimated parameters

Proposition: Under Conditions A1)–A3), we have

$$P(\widehat{K} = K) \to 1, \quad \text{and} \quad n^2\theta\|\widehat{\mathbf{S}}_1 - \boldsymbol{\Sigma}_1\|_2 = o_p(1).$$

Corollary: The critical region

$$\{\widehat{T}_{ij} \geq \chi^2_{\widehat{K},1-\alpha}\}$$

is asymptotic **size** α and asymptotic **power one** when $\sqrt{n\theta}\|\pi_i - \pi_j\| \to \infty$

Flexible Network Inference

under degree hoterogeneity

Degree Corrected Mixed Membership

Model: *(Zhang, Levina and Zhu, 2014; Jin, Ke and Luo, 2017, ...)*

$$\mathbf{H} = \boldsymbol{\Theta}\boldsymbol{\Pi}\mathbf{P}\boldsymbol{\Pi}^T\boldsymbol{\Theta}, \qquad \boldsymbol{\Theta} = \mathrm{diag}(\theta_1, ..., \theta_n)$$

Eigen-ratio: $\mathbf{V}/\mathbf{v}_1$ gets rid of heterogeneity. *(Jin, 2015)*

$$\bigstar \pi_i = \pi_j \quad \text{iff} \quad \frac{\mathbf{v}_k(i)}{\mathbf{v}_1(i)} = \frac{\mathbf{v}_k(j)}{\mathbf{v}_1(j)}, \quad 2 \leq k \leq K$$

Ratio Statistics: $Y(i,k) = \frac{\widehat{\mathbf{v}}_k(i)}{\widehat{\mathbf{v}}_1(i)}$ with $0/0$ defined as 1

★Build test by **comparing** $\mathbf{Y}_i = (Y(i,2), \cdots, Y(i,K))^T$ with $\mathbf{Y}_j$

An Ideal Test for $H_0 : \pi_i = \pi_j$

$$\mathbf{G_{ij}} = (\mathbf{Y}_i - \mathbf{Y}_j)^T \mathbf{\Sigma}_2^{-1} (\mathbf{Y}_i - \mathbf{Y}_j)$$

- $\mathbf{\Sigma}_2 =$ asymp. var. matrix of $\mathbf{Y}_i - \mathbf{Y}_j$
- $\mathbf{\Sigma}_2 = \mathrm{cov}(\mathbf{f})$ with $\mathbf{f} = (f_2, \cdots, f_K)^T$ with

$$f_k = \frac{\mathbf{e}_i^T \mathbf{W} \mathbf{v}_k}{t_k \mathbf{v}_1(i)} - \frac{\mathbf{e}_j^T \mathbf{W} \mathbf{v}_k}{t_k \mathbf{v}_1(j)} - \frac{\mathbf{v}_k(i) \mathbf{e}_i^T \mathbf{W} \mathbf{v}_1}{t_1 \mathbf{v}_1^2(i)} + \frac{\mathbf{v}_k(j) \mathbf{e}_j^T \mathbf{W} \mathbf{v}_1}{t_1 \mathbf{v}_1^2(j)}.$$

Technical Conditions

A4) $\min_{1\le k\le K}|\mathcal{N}_k| \ge c_2 n$, $\theta_{\min}^2 \ge n^{-c_3}$ for $c_2, c_3 \in (0,1)$, and $\theta_{\max} \le c_4\theta_{\min}$.

A5) $\mathbf{P} = (p_{kl}) > 0$ irreducible, $n\min_{1\le k\le K,\, t=i,j}\mathrm{var}(\mathbf{e}_t^T\mathbf{W}\mathbf{v}_k) \to \infty$

A6) All eigenvalues of $n\theta_{\min}^2\,\mathrm{cov}(\mathbf{f})$ are bounded away from 0 and ∞

■A4)-A5) are similar to those in Jin et al. (2017)

Asymptotic Distributions

Theorem 3: Assume A1), A4)–A6)

a) Under H_0, $G_{ij} \xrightarrow{d} \chi^2_{K-1}$

b) If $\lambda_2(\pi_i\pi_i^T + \pi_j\pi_j^T) \gg \frac{1}{n\theta^2_{\min}}$, then

$$G_{ij} \to \infty$$

Theorem 4: For substitution test $\widehat{G}_{ij}$ with

$$P(\widehat{K} = K) = 1 - o(1) \text{ and } n\theta^2_{\min}\|\widehat{\mathbf{S}}_2 - \mathbf{\Sigma}_2\|_2 = o_p(1),$$

the same results as in Theorem 3 hold.

Estimation of Unknown Parameters

★Use the same thresholding estimator for K

Proposition: The (a,b) entry of matrix $\mathbf{\Sigma}_2$ takes the form

$$\frac{1}{t_1^2}\Bigg\{\sum_{l=1,l\neq j}^{n}\sigma_{il}^2\left[\frac{t_1\mathbf{v}_{a+1}(l)}{t_{a+1}\mathbf{v}_1(i)}-\frac{\mathbf{v}_{a+1}(i)\mathbf{v}_1(l)}{\mathbf{v}_1(i)^2}\right]\left[\frac{t_1\mathbf{v}_{b+1}(l)}{t_{b+1}\mathbf{v}_1(i)}-\frac{\mathbf{v}_{b+1}(i)\mathbf{v}_1(l)}{\mathbf{v}_1(i)^2}\right]$$
$$+\sum_{l=1,l\neq i}^{n}\sigma_{jl}^2\left[\frac{t_1\mathbf{v}_{a+1}(l)}{t_{a+1}\mathbf{v}_1(j)}-\frac{\mathbf{v}_{a+1}(j)\mathbf{v}_1(l)}{\mathbf{v}_1(j)^2}\right]\left[\frac{t_1\mathbf{v}_{b+1}(l)}{t_{b+1}\mathbf{v}_1(j)}-\frac{\mathbf{v}_{b+1}(j)\mathbf{v}_1(l)}{\mathbf{v}_1(j)^2}\right]$$
$$+\sigma_{ij}^2\left[\frac{t_1\mathbf{v}_{a+1}(j)}{t_{a+1}\mathbf{v}_1(i)}-\frac{\mathbf{v}_{a+1}(i)\mathbf{v}_1(j)}{\mathbf{v}_1(i)^2}-\frac{t_1\mathbf{v}_{a+1}(i)}{t_{a+1}\mathbf{v}_1(j)}+\frac{\mathbf{v}_{a+1}(j)\mathbf{v}_1(i)}{\mathbf{v}_1(j)^2}\right]$$
$$\times\left[\frac{t_1\mathbf{v}_{b+1}(j)}{t_{b+1}\mathbf{v}_1(i)}-\frac{\mathbf{v}_{b+1}(i)\mathbf{v}_1(j)}{\mathbf{v}_1(i)^2}-\frac{t_1\mathbf{v}_{b+1}(i)}{t_{b+1}\widehat{\mathbf{v}}_1(j)}+\frac{\mathbf{v}_{b+1}(j)\mathbf{v}_1(i)}{\mathbf{v}_1(j)^2}\right]\Bigg\}.$$

★t_k very **complicated**, estimated by $\widehat{d}_k$

Asymptotic size and test

Proposition: The rejection region

$$\{\widehat{G}_{ij} \geq \chi^2_{\widehat{K}-1,1-\alpha}\}$$

has asymptotic size α and the asymptotic power one when $\lambda_2(\pi_i\pi_i^T + \pi_j\pi_j^T) \gg \frac{1}{n\theta^2_{\min}}$

■ $\widehat{G}_{ij}$ can be used under degree **homogeneity**, but $\widehat{T}_{ij}$ has **better** practical performance in this case.

Numerical Studies

Simulations: K Known

- **Model**: ★$K = 3$, ★3 pure nodes, ★4 mixed membership;
- $n \in \{1500, 3000\}$, $N_{sim} = 500$, sig. level 0.05
- For mixed membership model, $\theta \in \{0.2, 0.3, \cdots, 0.9\}$
- For degree corrected mixed membership model, $\theta_i^{-1} \sim U[r^{-1}, 2r^{-1}]$ with $r^2 \in \{0.2, 0.3, \cdots, 0.9\}$
- $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are estimated from data

Size and Power

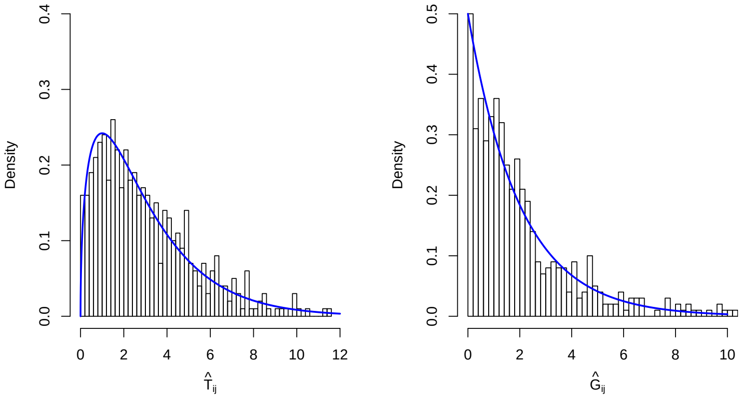
$n = 1500$, size at $\pi_0 = (0.2, 0.6, 0.2)$, power at $\pi_a = (0, 1, 0)$

Model 1	θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.058	0.046	0.06	0.05	0.05	0.058	0.036	0.05
	Power	0.734	0.936	0.986	0.998	1	1	1	1
Model 2	r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.076	0.062	0.072	0.062	0.074	0.046	0.044	0.056
	Power	0.426	0.562	0.696	0.77	0.89	0.93	0.952	0.976

$n = 3000$, size at $\pi_0 = (0.2, 0.6, 0.2)$, power at $\pi_a = (0, 1, 0)$

Model 1	θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.082	0.066	0.052	0.052	0.044	0.042	0.038	0.062
	Power	0.936	0.994	1	1	1	1	1	1
Model 2	r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.082	0.06	0.062	0.058	0.062	0.066	0.064	0.06
	Power	0.67	0.842	0.918	0.972	0.99	1	1	1

Asymptotic Null Distributions



★Left: Dist of $\widehat{T}_{ij}$ with $\theta = 0.9$ (Blue curve is χ^2_3). $n = 3000$.

★Right: Dist of $\widehat{G}_{ij}$ with $r^2 = 0.9$ (Blue curve is χ^2_2).

Simulations: K Unknown

Estimation accuracy of K, $n = 3000$

	θ (r^2)	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MM	$P(\widehat{K} = K)$	1	1	1	1	1	1	1	1
	$P(\widehat{K} \leq K)$	1	1	1	1	1	1	1	1
DCMM	$P(\widehat{K} = K)$	0	0	0	1	1	1	1	1
	$P(\widehat{K} \leq K)$	1	1	1	1	1	1	1	1

Size and power, size at $\pi_0 = (0.2, 0.6, 0.2)$, power at $\pi_a = (0, 1, 0)$

Model 1	θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.082	0.066	0.052	0.052	0.044	0.042	0.038	0.062
	Power	0.936	0.994	1	1	1	1	1	1
Model 2	r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.054	0.058	0.062	0.058	0.062	0.066	0.064	0.06
	Power	0.074	0.042	0.918	0.972	0.99	1	1	1

U.S. Political Data

- 105 political books sold online in 2004 (*V. Krebs, source: http://www.orgnet.com*)
- Links between two books represent frequency co-purchasing of books by the same buyers
- Books have been assigned manually three labels (conservative, liberal, and neutral) by M. E. J. Newman
- Such labels may not be accurate (e.g. mixed members)

Comparisons of selected books

- Consider mixed memberships with $K = 2$ communities
- Consider the same 9 books reported in Jin et al. (2017)

Title	Label (by Newman)	Node index
Empire	neutral	1
The Future of Freedom	neutral	2
Rise of the Vulcans	conservative	3
All the Shah's Men	neutral	4
Bush at War	conservative	5
Plan of Attack	neutral	6
Power Plays	neutral	7
Meant To Be	neutral	8
The Bushes	conservative	9

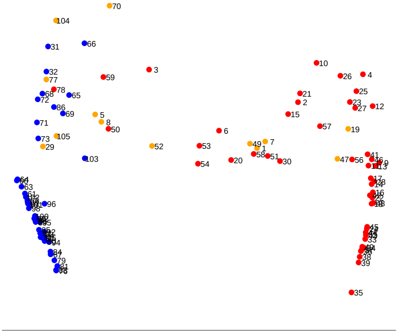
P-values Based on $\widehat{T}_{ij}$

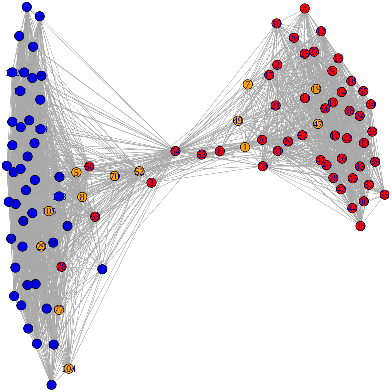
Node	1(N)	2(N)	3(C)	4(N)	5(C)	6(N)	7(N)	8(N)	9(C)
1(N)	**1.0000**	**0.6766**	**0.0298**	**0.3112**	**0.0248**	0.0000	**0.0574**	**0.1013**	**0.0449**
2(N)	0.6766	1.0000	0.0261	0.2487	0.0204	0.0000	0.0643	0.1184	0.0407
3(C)	**0.0298**	**0.0261**	**1.0000**	0.1546	**0.2129**	**0.0013**	**0.0326**	0.0513	**0.9249**
4(N)	0.3112	0.2487	0.1546	1.0000	0.3206	0.0034	0.0236	0.0497	0.2121
5(C)	0.0248	0.0204	0.2129	0.3206	1.0000	0.0991	0.0042	0.0084	0.2574
6(N)	0.0000	0.0000	0.0013	0.0034	0.0991	1.0000	0.0000	0.0000	0.0035
7(N)	0.0574	0.0643	0.0326	0.0236	0.0042	0.0000	1.0000	0.9004	0.0834
8(N)	0.1013	0.1184	0.0513	0.0497	0.0084	0.0000	0.9004	1.0000	0.1113
9(C)	0.0449	0.0407	0.9249	0.2121	0.2574	0.0035	0.0834	0.1113	1.0000

P-values Based on $\widehat{G}_{ij}$

Node	1(N)	2(N)	3(C)	4(N)	5(C)	6(N)	7(N)	8(N)	9(C)
1(N)	**1.0000**	**0.4403**	0.1730	**0.4563**	0.8307	**0.5361**	0.0000	0.0000	0.1920
2(N)	0.4403	1.0000	0.0773	0.9721	0.3665	0.6972	0.0000	0.0000	0.1144
3(C)	0.1730	0.0773	**1.0000**	0.0792	0.1337	0.0885	**0.0000**	**0.0000**	**0.8141**
4(N)	0.4563	0.9721	0.0792	1.0000	0.4256	0.7624	0.0000	0.0000	0.1153
5(C)	0.8307	0.3665	0.1337	0.4256	1.0000	0.5402	0.0000	0.0000	0.1591
6(N)	0.5361	0.6972	0.0885	0.7624	0.5402	1.0000	0.0000	0.0000	0.1294
7(N)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.9778	0.0000
8(N)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9778	1.0000	0.0000
9(C)	0.1920	0.1144	0.8141	0.1153	0.1591	0.1294	0.0000	0.0000	1.0000

Test-distance and P-values based clustering





★distances $\widehat{G}_{ij}$ ★used P-values of $\widehat{G}_{ij}$ as weights; ★no links when P-value $< 5\%$.

★red: C; Blue: Liberal; yellow: Neutral

Consistent w/ Newman's labels

Summary

- Our work represents a first attempt to address community detection with statistical significance.
- We proposed two tests for equality of membership profiles any given pair of nodes (MMM w/ and w/o degree corr.)
- Our method is pivotal to unknown parameters including K.
- We have provided theoretical justifications of our results and illustrated the method with estimated K.

The End

Thank  *You*

- Fan, J., Fan, Y., Han, X. and Lv, J. (2018). Asymptotic theory of eigenvectors for large random matrices. *Manuscript*.
- Fan, J., Fan, Y., Han, X. and Lv, J. (2019). SIMPLE: Statistical Inference on Membership Profiles in Large Networks. *Manuscript*.