In Memory of Sik-Yum Lee

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

<table>
<thead>
<tr>
<th>Name</th>
<th>School</th>
<th>Year</th>
<th>Descendants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shi, Jian Qing</td>
<td>Chinese University of Hong Kong</td>
<td>1996</td>
<td></td>
</tr>
<tr>
<td>Song, Xin-Yuan</td>
<td>Chinese University of Hong Kong</td>
<td>2001</td>
<td>1</td>
</tr>
<tr>
<td>Zhang, Wenyang</td>
<td>Chinese University of Hong Kong</td>
<td>1999</td>
<td>6</td>
</tr>
<tr>
<td>Zhu, Hongtu</td>
<td>Chinese University of Hong Kong</td>
<td>2000</td>
<td>22</td>
</tr>
</tbody>
</table>

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

Jianqing Fan (Princeton University)
Large Scale Network Inference
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

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

How to quantify uncertainty that a given pair of nodes are in the same community?

★citation ★social, ★trade ★economic ★gene regulatory, · · ·

Jianqing Fan (Princeton University) Large Scale Network Inference
A Networked World

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

How to quantify uncertainty that a given pair of nodes are in the same community?

- citation
- social
- trade
- economic
- gene regulatory
- ...
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?
**Mixed membership model**: Each node now equipped with a vector of *membership probabilities*

Starred Communities using mixed membership model
Can we quantify the uncertainties of links?
A Sneak Peek of Our Results

P-values for pairwise comparison

<table>
<thead>
<tr>
<th></th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1.0000</td>
<td>0.1278</td>
<td>0.0012</td>
<td>0.0685</td>
<td>0.0145</td>
</tr>
<tr>
<td>8</td>
<td>0.1278</td>
<td>1.0000</td>
<td>0.0026</td>
<td>0.0052</td>
<td>0.0000</td>
</tr>
<tr>
<td>9</td>
<td>0.0012</td>
<td>0.0026</td>
<td>1.0000</td>
<td>0.3308</td>
<td>0.0540</td>
</tr>
<tr>
<td>10</td>
<td>0.0685</td>
<td>0.0052</td>
<td>0.3308</td>
<td>1.0000</td>
<td>0.4155</td>
</tr>
<tr>
<td>27</td>
<td>0.0145</td>
<td>0.0000</td>
<td>0.0540</td>
<td>0.4155</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

How to get these P-values?

Applications: ★Dim-reduction ★network centrality
Connections with Factor-adjusted sparsity

**Data:** \( \{X_t\}_{t=1}^n \)

**Factor model:** \( X_t = \mu + Bf_t + u_t \)

**Assumption:** \( \Sigma_u \) or \( \Sigma_u^{-1} \) sparse

**Modeling sparsity:** \( \Sigma_u^{-1} \equiv \Omega = \alpha I_p + \beta L_p \)

**Graph Laplacian:** \( L_p = I_p - D^{-1/2} XD^{-1/2} \), \( D = \text{diag}(d_1, \ldots, d_p) \)

\( \omega_{ij} = 0 \iff \text{an edge} \)

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
$K$ disjoint communities $C_1, \cdots, C_K$, with $P(X_{ij} = 1) = p_{kl}$, for $i \in C_k, j \in C_l$, indep.

**Edge probability**: $P = (p_{i,j})_{K \times K}$.

**Degree-corrected**: $P(X_{ij} = 1) = \theta_i \theta_j p_{kl}$, $i \in C_k, j \in C_l$.

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

Each node $i$ has

$$P(\text{node } i \text{ belongs to community } \mathbf{C}_k) = \pi_i(k)$$

- Probability vector $\pi_i = (\pi_i(1), \cdots, \pi_i(K))^T \in \mathbb{R}^K$ is the membership profile
- $\pi_i = e_\ell$ reduces to communication detection.

**Hypothesis testing**: For any two members,

$$H_0 : \pi_i = \pi_j \text{ vs. } H_1 : \pi_i \neq \pi_j$$
Adjacency matrix $X = (X_{ij}) \in \mathbb{R}^{n \times n}$,

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

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

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

$P(X_{ij} = 1 | i \in C_k, j \in C_l) = \theta_i \theta_j p_{kl}$,

$\star P = (p_{kl}) \in \mathbb{R}^{K \times K}$ 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 \( \Pi = (\pi_1, \cdots, \pi_n)^T \in \mathbb{R}^{n \times K} \)

\[ \mathbf{X} = \mathbf{H} + \mathbf{W}, \quad \mathbf{H} = \Theta \Pi \mathbf{P} \mathbf{P}^T \Theta, \]

\[ \star \Theta = \text{diag}(\theta_1, \cdots, \theta_n), \quad \star \mathbf{W} = \mathbf{X} - E\mathbf{X} \text{ is generalized Wigner matrix} \]

- 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:** $\Theta = \sqrt{\theta} I_n$, \( \theta \to 0 \).

$$\mathbf{E} \mathbf{X} = \mathbf{H} = \theta \left( \prod \mathbf{P} \prod^T \right) = \theta \left( \begin{array}{ccc} \pi_1^T \mathbf{P} \pi_1 & \pi_1^T \mathbf{P} \pi_2 & \cdots & \pi_1^T \mathbf{P} \pi_n \\ \pi_2^T \mathbf{P} \pi_1 & \pi_2^T \mathbf{P} \pi_2 & \cdots & \pi_2^T \mathbf{P} \pi_n \\ \cdots & \cdots & \cdots & \cdots \end{array} \right) .$$

★ Eigenspace of $\mathbf{H} =$ column space spanned by $\prod$
Eigen-structures

**Population** Eigen-decomposition: $\mathbf{H} = \mathbf{VDV}^T$

- $\mathbf{D} = \text{diag}(d_1, \ldots, d_K)$ 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.

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

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

$k$-mean
By permutation argument, \( \pi_i = \pi_j \iff V(i) = V(j) \)

**Ideal test statistic:**

\[
T_{ij} = (\hat{V}(i) - \hat{V}(j))^T \Sigma_1^{-1} (\hat{V}(i) - \hat{V}(j))
\]

\( \Sigma_1 \) is asymptotic variance — challenge to derive

\[
\Sigma_1 = \text{cov}((e_i - e_j)^T WVD^{-1})
\]
A1) \[ \min_{1 \leq i \leq K-1} \left| \frac{d_i}{d_{i+1}} \right| \geq 1 + c_0, \quad \alpha_n^2 = \max_j \text{var} \left( \sum_{i=1}^n X_{ij} \right) \to \infty. \]

A2) \[ \lambda_K(\Pi^T \Pi) \geq c_1 n, \quad \lambda_K(P) \geq c_1, \quad \text{and} \quad \theta \geq n^{-c_2}, \quad 0 < c_1, c_2 < 1. \]

A3) All eigenvalues of \( n^2 \theta \Sigma_1 \) are bounded away from 0 and \( \infty \).

★ \( \alpha_n \) measures sparsity of network

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

\[ d_k \sim n\theta, \quad k = 1, \ldots, K \]
Theorem 1: Assume A1)–A3).

a) Under **Null hypothesis** $H_0$,
\[ T_{ij} \xrightarrow{d} \chi^2_K, \quad \text{as } n \to \infty \]

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 $(V(i) - V(j))^T \Sigma_1^{-1}(V(i) - V(j)) \to \mu$, then
\[ T_{ij} \xrightarrow{d} \chi^2_K(\mu) \]
Replace $K$ and $\Sigma_1$ in $T_{ij}$ by $\hat{K}$ and $\hat{S}_1$ \implies \hat{T}_{ij}.

**Theorem 2**: Assume that the following accuracy:

\[ P(\hat{K} = K) = 1 - o(1) \quad \text{and} \quad n^2 \theta \| \hat{S}_1 - \Sigma_1 \|_2 = o_p(1). \]

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

How to estimate $K$ and $\Sigma_1$?
Replace $K$ and $\Sigma_1$ in $T_{ij}$ by $\hat{K}$ and $\hat{S}_1$ \[ \Rightarrow \hat{T}_{ij}. \]

**Theorem 2:** Assume that the following accuracy:

\[
P(\hat{K} = K) = 1 - o(1) \quad \text{and} \quad n^2 \theta \| \hat{S}_1 - \Sigma_1 \|_2 = o_p(1).
\]

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

How to estimate $K$ and $\Sigma_1$?
Estimation of Unknown Parameters

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

**Proposition**: The \((a, b)\) entry of matrix \(\Sigma_1\) is

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

Plug in: estimating \(\sigma_{ab}^2 = \text{var}(X_{ab})\) is somewhat complicated.
Estimating $\sigma^2_{ab}$

\[ \hat{w}_{0,ab}^2 \text{ with } \hat{W}_0 = (\hat{w}_{0,ab}) = x - \sum_{k=1}^{\hat{K}} \hat{d}_k \hat{v}_k \hat{v}_k^T \text{ is not good enough.} \]

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

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

\[
\tilde{d}_k = \left( \frac{1}{\hat{d}_k} + \frac{\hat{v}_k^T \text{diag}(\hat{W}_0^2) \hat{v}_k}{\hat{d}_k^3} \right)^{-1}
\]

3. Update the estimator of $W$ as $\hat{W} = x - \sum_{k=1}^{\hat{K}} \tilde{d}_k \hat{v}_k \hat{v}_k^T$.

Estimate $\sigma^2_{ab}$ as $\hat{\sigma}^2_{ab} = \hat{w}_{ab}^2$
**Proposition**: Under Conditions A1)–A3), we have

\[ P(\hat{K} = K) \to 1, \quad \text{and} \quad n^2 \theta \| \hat{S}_1 - \Sigma_1 \|_2 = o_p(1). \]

**Corollary**: The critical region

\[ \{ \hat{T}_{ij} \geq \chi^2_{\hat{K}, 1-\alpha} \} \]

is asymptotic **size** \( \alpha \) and asymptotic **power one** when

\[ \sqrt{n \theta} \| \pi_i - \pi_j \| \to \infty \]
Flexible Network Inference under degree heterogeneity
Degree Corrected Mixed Membership

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

\[ H = \Theta \Pi \Pi^T \Theta, \quad \Theta = \text{diag}(\theta_1, \ldots, \theta_n) \]

**Eigen-ratio:** \( V / v_1 \) gets rid of heterogeneity. (Jin, 2015)

\[ \star \pi_i = \pi_j \quad \text{iff} \quad \frac{v_k(i)}{v_1(i)} = \frac{v_k(j)}{v_1(j)}, \quad 2 \leq k \leq K \]

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

\[ \star \text{Build test by comparing } Y_i = (Y(i, 2), \ldots, Y(i, K))^T \text{ with } Y_j \]
An Ideal Test for $H_0: \pi_i = \pi_j$

$$G_{ij} = (Y_i - Y_j)^T \Sigma_2^{-1}(Y_i - Y_j)$$

- $\Sigma_2 = \text{asympt. var. matrix of } Y_i - Y_j$

- $\Sigma_2 = \text{cov}(f)$ with $f = (f_2, \cdots, f_K)^T$ with

$$f_k = \frac{e_i^T W v_k}{t_k v_1(i)} - \frac{e_j^T W v_k}{t_k v_1(j)} - \frac{v_k(i) e_i^T W v_1}{t_1 v_1^2(i)} + \frac{v_k(j) e_j^T W v_1}{t_1 v_1^2(j)}.$$
A4) $\min_{1 \leq k \leq K} |\mathcal{N}_k| \geq c_2 n$, $\theta_{\min}^2 \geq n^{-c_3}$ for $c_2, c_3 \in (0, 1)$, and $\theta_{\max} \leq c_4 \theta_{\min}$.

A5) $P = (p_{kl}) > 0$ irreducible, $n \min_{1 \leq k \leq K, t=i,j} \text{var}(e_t^T Wv_k) \to \infty$

A6) All eigenvalues of $n \theta_{\min}^2 \text{cov}(f)$ are bounded away from 0 and $\infty$

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

\[ G_{ij} \to \infty \]

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

\[
P(\hat{K} = K) = 1 - o(1) \text{ and } n\theta^2_{\text{min}}\|\hat{S}_2 - \Sigma_2\|_2 = o_p(1),
\]

the same results as in Theorem 3 hold.
Proposition: The \((a, b)\) entry of matrix \(\Sigma_2\) takes the form

\[
\frac{1}{t_1^2} \left\{ \sum_{l=1, l\neq j}^n \sigma_{jl}^2 \left[ \frac{t_1 v_{a+1}(l)}{t_{a+1} v_1(i)} - \frac{v_{a+1}(i) v_1(l)}{v_1(i)^2} \right] \left[ \frac{t_1 v_{b+1}(l)}{t_{b+1} v_1(i)} - \frac{v_{b+1}(i) v_1(l)}{v_1(i)^2} \right] \right. \\
+ \sum_{l=1, l\neq i}^n \sigma_{il}^2 \left[ \frac{t_1 v_{a+1}(l)}{t_{a+1} v_1(j)} - \frac{v_{a+1}(j) v_1(l)}{v_1(j)^2} \right] \left[ \frac{t_1 v_{b+1}(l)}{t_{b+1} v_1(j)} - \frac{v_{b+1}(j) v_1(l)}{v_1(j)^2} \right] \\
\left. + \sigma_{ij}^2 \left[ \frac{t_1 v_{a+1}(j)}{t_{a+1} v_1(i)} - \frac{v_{a+1}(i) v_1(j)}{v_1(i)^2} - \frac{t_1 v_{a+1}(i)}{t_{a+1} v_1(j)} + \frac{v_{a+1}(j) v_1(i)}{v_1(j)^2} \right] \times \left[ \frac{t_1 v_{b+1}(j)}{t_{b+1} v_1(i)} - \frac{v_{b+1}(i) v_1(j)}{v_1(i)^2} - \frac{t_1 v_{b+1}(i)}{t_{b+1} \hat{v}_1(j)} + \frac{v_{b+1}(j) v_1(i)}{v_1(j)^2} \right] \right\}.
\]

\(t_k\) very complicated, estimated by \(\hat{d}_k\)
Asymptotic size and test

**Proposition**: The rejection region

\[ \{ \hat{G}_{ij} \geq \chi^2_{K - 1, 1 - \alpha} \} \]

has asymptotic size \( \alpha \) and the asymptotic power one when

\[ \lambda_2(\pi_i \pi_i^T + \pi_j \pi_j^T) \gg \frac{1}{n \theta_{\text{min}}^2} \]

\( \hat{G}_{ij} \) can be used under degree **homogeneity**, but \( \hat{T}_{ij} \) has **better** practical performance in this case.
Numerical Studies
Simulations: $K$ Known

- **Model**: $K = 3$, $\star$ 3 pure nodes, $\star$ 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 $r^2 \in \{0.2, 0.3, \cdots, 0.9\}$
- $\Sigma_1$ and $\Sigma_2$ are estimated from data
### Size and Power

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

<table>
<thead>
<tr>
<th>Model 1</th>
<th>$\theta$</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>0.058</td>
<td>0.046</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.058</td>
<td>0.036</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>0.734</td>
<td>0.936</td>
<td>0.986</td>
<td>0.998</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### Model 2

<table>
<thead>
<tr>
<th>$r^2$</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>0.076</td>
<td>0.062</td>
<td>0.072</td>
<td>0.062</td>
<td>0.074</td>
<td>0.046</td>
<td>0.044</td>
<td>0.056</td>
</tr>
<tr>
<td>Power</td>
<td>0.426</td>
<td>0.562</td>
<td>0.696</td>
<td>0.77</td>
<td>0.89</td>
<td>0.93</td>
<td>0.952</td>
<td>0.976</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model 1</th>
<th>$\theta$</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>0.082</td>
<td>0.066</td>
<td>0.052</td>
<td>0.052</td>
<td>0.044</td>
<td>0.042</td>
<td>0.038</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>0.936</td>
<td>0.994</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### Model 2

<table>
<thead>
<tr>
<th>$r^2$</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>0.082</td>
<td>0.06</td>
<td>0.062</td>
<td>0.058</td>
<td>0.062</td>
<td>0.066</td>
<td>0.064</td>
<td>0.06</td>
</tr>
<tr>
<td>Power</td>
<td>0.67</td>
<td>0.842</td>
<td>0.918</td>
<td>0.972</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Asymptotic Null Distributions

★ Left: Dist of $\hat{T}_{ij}$ with $\theta = 0.9$ (Blue curve is $\chi^2_3$). $n = 3000$.  
★ Right: Dist of $\hat{G}_{ij}$ with $r^2 = 0.9$ (Blue curve is $\chi^2_2$).
### Estimation accuracy of $K$, $n = 3000$

<table>
<thead>
<tr>
<th></th>
<th>$\theta (r^2)$</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>$P(\hat{K} = K)$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$P(\hat{K} \leq K)$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DCMM</td>
<td>$P(\hat{K} = K)$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$P(\hat{K} \leq K)$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>$\theta$</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>Size</td>
<td>0.082</td>
<td>0.066</td>
<td>0.052</td>
<td>0.052</td>
<td>0.044</td>
<td>0.042</td>
<td>0.038</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.936</td>
<td>0.994</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$r^2$</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>Model 2</td>
<td>Size</td>
<td>0.054</td>
<td>0.058</td>
<td>0.062</td>
<td>0.058</td>
<td>0.062</td>
<td>0.066</td>
<td>0.064</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Power</td>
<td>0.074</td>
<td>0.042</td>
<td>0.918</td>
<td>0.972</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>Title</th>
<th>Label (by Newman)</th>
<th>Node index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empire</td>
<td>neutral</td>
<td>1</td>
</tr>
<tr>
<td>The Future of Freedom</td>
<td>neutral</td>
<td>2</td>
</tr>
<tr>
<td>Rise of the Vulcans</td>
<td>conservative</td>
<td>3</td>
</tr>
<tr>
<td>All the Shah’s Men</td>
<td>neutral</td>
<td>4</td>
</tr>
<tr>
<td>Bush at War</td>
<td>conservative</td>
<td>5</td>
</tr>
<tr>
<td>Plan of Attack</td>
<td>neutral</td>
<td>6</td>
</tr>
<tr>
<td>Power Plays</td>
<td>neutral</td>
<td>7</td>
</tr>
<tr>
<td>Meant To Be</td>
<td>neutral</td>
<td>8</td>
</tr>
<tr>
<td>The Bushes</td>
<td>conservative</td>
<td>9</td>
</tr>
</tbody>
</table>
P-values Based on $\hat{T}_{ij}$

<table>
<thead>
<tr>
<th>Node</th>
<th>1(N)</th>
<th>2(N)</th>
<th>3(C)</th>
<th>4(N)</th>
<th>5(C)</th>
<th>6(N)</th>
<th>7(N)</th>
<th>8(N)</th>
<th>9(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(N)</td>
<td>1.0000</td>
<td>0.6766</td>
<td>0.0298</td>
<td>0.3112</td>
<td>0.0248</td>
<td>0.0000</td>
<td>0.0574</td>
<td>0.1013</td>
<td>0.0449</td>
</tr>
<tr>
<td>2(N)</td>
<td>0.6766</td>
<td>1.0000</td>
<td>0.0261</td>
<td>0.2487</td>
<td>0.0204</td>
<td>0.0000</td>
<td>0.0643</td>
<td>0.1184</td>
<td>0.0407</td>
</tr>
<tr>
<td>3(C)</td>
<td>0.0298</td>
<td>0.0261</td>
<td>1.0000</td>
<td>0.1546</td>
<td>0.2129</td>
<td>0.0013</td>
<td>0.0326</td>
<td>0.0513</td>
<td>0.9249</td>
</tr>
<tr>
<td>4(N)</td>
<td>0.3112</td>
<td>0.2487</td>
<td>0.1546</td>
<td>1.0000</td>
<td>0.3206</td>
<td>0.0034</td>
<td>0.0236</td>
<td>0.0497</td>
<td>0.2121</td>
</tr>
<tr>
<td>5(C)</td>
<td>0.0248</td>
<td>0.0204</td>
<td>0.2129</td>
<td>0.3206</td>
<td>1.0000</td>
<td>0.0991</td>
<td>0.0042</td>
<td>0.0084</td>
<td>0.2574</td>
</tr>
<tr>
<td>6(N)</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0013</td>
<td>0.0034</td>
<td>0.0991</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0035</td>
</tr>
<tr>
<td>7(N)</td>
<td>0.0574</td>
<td>0.0643</td>
<td>0.0326</td>
<td>0.0236</td>
<td>0.0042</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.9004</td>
<td>0.0834</td>
</tr>
<tr>
<td>8(N)</td>
<td>0.1013</td>
<td>0.1184</td>
<td>0.0513</td>
<td>0.0497</td>
<td>0.0084</td>
<td>0.0000</td>
<td>0.9004</td>
<td>1.0000</td>
<td>0.1113</td>
</tr>
<tr>
<td>9(C)</td>
<td>0.0449</td>
<td>0.0407</td>
<td>0.9249</td>
<td>0.2121</td>
<td>0.2574</td>
<td>0.0035</td>
<td>0.0834</td>
<td>0.1113</td>
<td>1.0000</td>
</tr>
<tr>
<td>Node</td>
<td>1(N)</td>
<td>2(N)</td>
<td>3(C)</td>
<td>4(N)</td>
<td>5(C)</td>
<td>6(N)</td>
<td>7(N)</td>
<td>8(N)</td>
<td>9(C)</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>1(N)</td>
<td>1.0000</td>
<td>0.4403</td>
<td>0.1730</td>
<td>0.4563</td>
<td>0.8307</td>
<td>0.5361</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1920</td>
</tr>
<tr>
<td>2(N)</td>
<td>0.4403</td>
<td>1.0000</td>
<td>0.0773</td>
<td>0.9721</td>
<td>0.3665</td>
<td>0.6972</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1144</td>
</tr>
<tr>
<td>3(C)</td>
<td>0.1730</td>
<td>0.0773</td>
<td>1.0000</td>
<td>0.0792</td>
<td>0.1337</td>
<td>0.0885</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.8141</td>
</tr>
<tr>
<td>4(N)</td>
<td>0.4563</td>
<td>0.9721</td>
<td>0.0792</td>
<td>1.0000</td>
<td>0.4256</td>
<td>0.7624</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1153</td>
</tr>
<tr>
<td>5(C)</td>
<td>0.8307</td>
<td>0.3665</td>
<td>0.1337</td>
<td>0.4256</td>
<td>1.0000</td>
<td>0.5402</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1591</td>
</tr>
<tr>
<td>6(N)</td>
<td>0.5361</td>
<td>0.6972</td>
<td>0.0885</td>
<td>0.7624</td>
<td>0.5402</td>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1294</td>
</tr>
<tr>
<td>7(N)</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.9778</td>
<td>0.0000</td>
</tr>
<tr>
<td>8(N)</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.9778</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>9(C)</td>
<td>0.1920</td>
<td>0.1144</td>
<td>0.8141</td>
<td>0.1153</td>
<td>0.1591</td>
<td>0.1294</td>
<td>0.0000</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Test-distance and P-values based clustering

★distances $\hat{G}_{ij}$ ★used P-values of $\hat{G}_{ij}$ as weights; ★no links when P-value < 5%.
★red: C; Blue: Liberal; yellow: Neutral
Consistent w/ Newman’s labels
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 $K$. 