High-Dimensional Network Inference (part I)

Jinchi Lv
Data Sciences and Operations Department
Marshall School of Business
University of Southern California

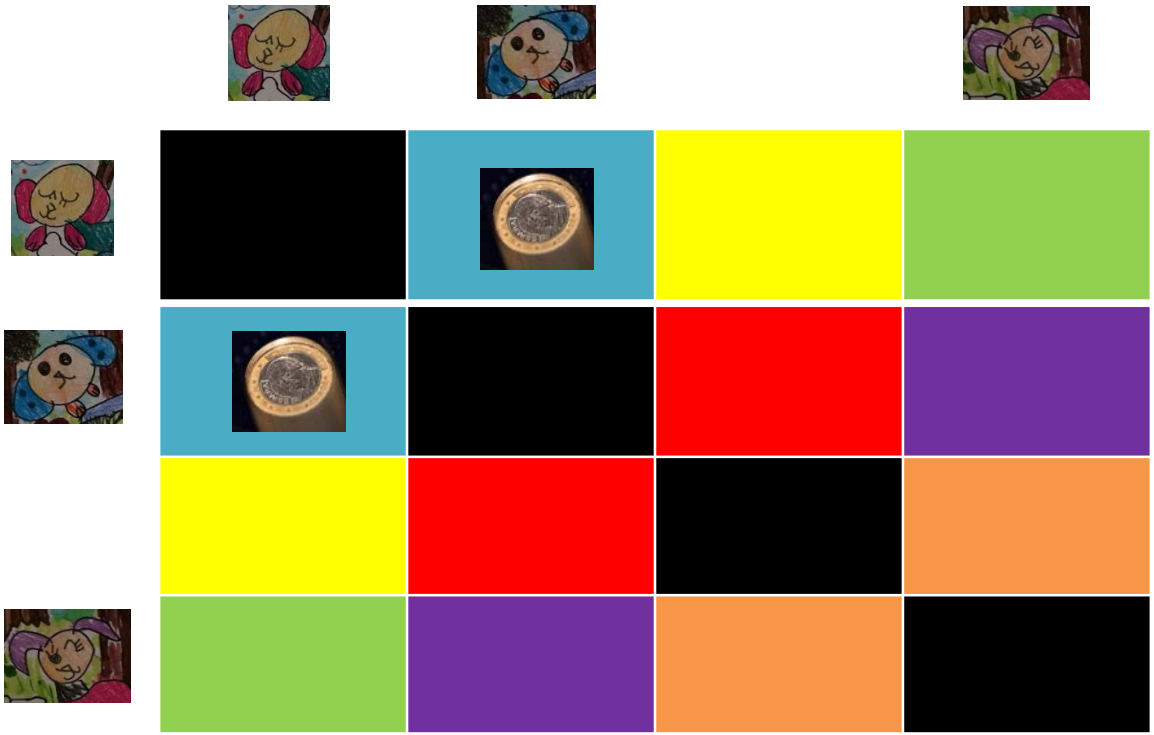
http://faculty.marshall.usc.edu/jinchi-lv

USC Summer School on Uncertainty Quantification (08/09/2024)

Outline of Fan, Fan, Han and L. (2022b)

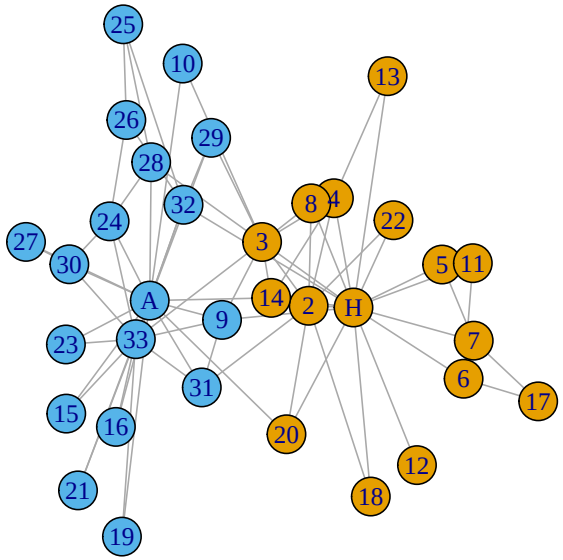
- A motivating example
- Flexible network inference
- SIMPLE for mixed membership models
- SIMPLE for degree-corrected mixed membership models
- Numerical examples

A Networked World and A Simple Question



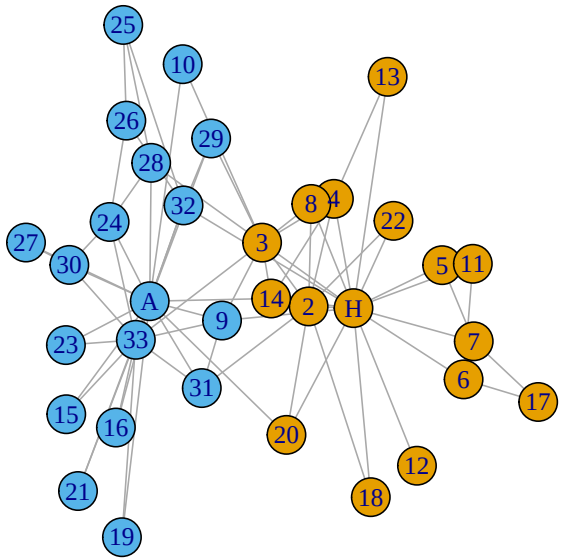
- How to *test* whether a *pair* of social media users or text documents belong to *the same community*?

A Motivating Example



- A university *karate club network* data (Zachary, 1977) for 34 members (Girvan and Newman, 2002)
- Edge meaning two members spent much time together outside club meetings
- At some point members split into *two communities* (one led by *H* and the other by *A*)

A Network with Non-Overlapping Communities



- Well understood network structure based on stochastic block model with *non-overlapping communities*

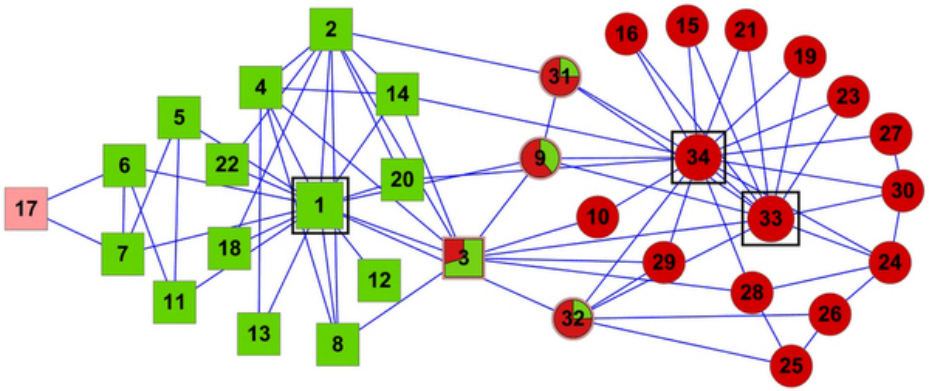
- Nodes 7, 8 belonging to one community (*H*) and nodes 9, 10, 27 belonging to the other (*A*)

P-Values?

	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

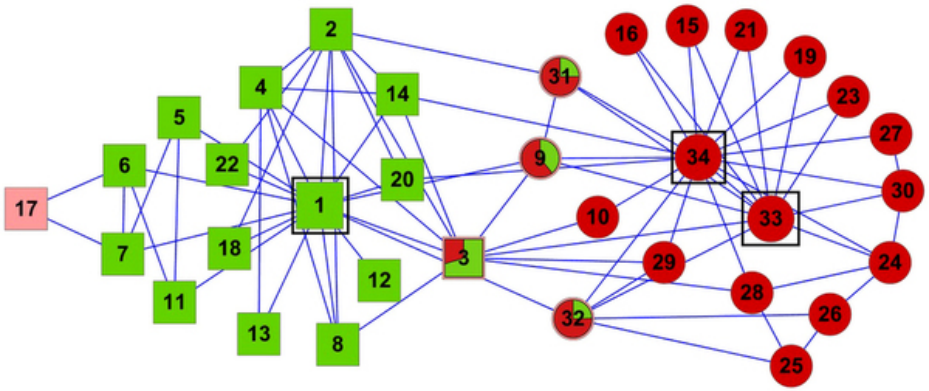
- Desired to *test* whether each *pair* of nodes belong to *the same community*
- Not only a *Y/N* but also precise *p-value for network inference* justifying significance of community labeling
- *How?*

A Network with Overlapping Communities



- *All models are wrong, but some are more useful than others* (George Box, 1979)
- *Model misspecification* has important implications (White, 1982; Cule, Samworth and Stewart, 2010; L. and Liu, 2014; Bühlmann and van de Geer, 2015; ...)
- What if switching to stochastic block model with *overlapping communities* (*various network structures*)

P-Values?



- Each node now equipped with a *membership probability vector* (*mixed membership*)
- *A different test needed?*
- *Different network models resulting in different testing results?*

A Surprise

	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

- The world of *network inference* can be much simpler than imagined (*same test*)
- *How?*

Questions of Interest

- How to design a tool for *flexible network inference* with precise p-values on testing whether two nodes share same membership profiles (*general network models with overlapping communities*)?
- How to allow for *degree heterogeneity* for more flexible network inference?
- How to develop a general framework of *asymptotic theory* on the *size* and *power* of the new tests?

Flexible Network Inference

Model Setting

- Consider a network with n nodes $\{1, \cdots, n\}$ and *adjacency matrix* $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times n}$ representing *connectivity structure* of network with $x_{ij} = 1$ for link and 0 for no link (Bhattacharyya and Bickel, 2016; Abbe, 2017; Le, Levina and Vershynin, 2018; Fan, Fan, Han and L., 2022a; ...)

- Assume adjacency matrix can be written generally as

$$\mathbf{X} = \mathbf{H} + \mathbf{W}$$

- $\mathbf{H} = (h_{ij}) \in \mathbb{R}^{n \times n}$ is deterministic mean matrix of *low rank* K

- Latent *network structure* encoded in the *eigen-structure* of mean matrix $\mathbf{H}$

Continued

- $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n \times n}$ is symmetric random noise matrix with independent diagonal and upper diagonal entries satisfying $\mathbb{E}w_{ij} = 0$ and $\max_{1 \le i,j \le n} |w_{ij}| \le 1$

- Noise matrix $\mathbf{W}$ known as *generalized Wigner matrix*

- Links x_{ij}'s independent Bernoulli random variables with means h_{ij}

- For the case of *without self loops*, we observe

$$\mathbf{X} - \mathrm{diag}(\mathbf{X}) = \mathbf{H} + (\mathbf{W} - \mathrm{diag}(\mathbf{X}))$$

- Our framework applicable to *both cases*

Community Structure

- Assume network can be decomposed into K communities $C_1, \cdots, C_K$

- Each node i has probability vector $\pi_i = (\pi_i(1), \cdots, \pi_i(K))^T \in \mathbb{R}^K$ with $\pi_i(k) \in [0,1]$, $\sum_{k=1}^K \pi_i(k) = 1$, and

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

- Assume number of communities K is finite but *unknown*

Hypothesis Testing

- For any given *pair* of nodes (i, j), our goal is to infer whether they share *same community identity* (*membership probability vector*) with quantified uncertainty level from observed adjacency matrix $\mathbf{X}$

- Interested in testing hypothesis

$$H_0 : \boldsymbol{\pi}_i = \boldsymbol{\pi}_j \quad \text{vs.} \quad H_1 : \boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j$$

- To make problem more explicit we first exploit mixed membership model

SIMPLE for Mixed Membership Models

Mixed Membership Model

- For now consider *mixed membership model* without degree heterogeneity (Airoldi, Blei, Fienberg and Xing, 2008)

$$\mathbf{H} = \theta \mathbf{\Pi P \Pi}^T$$

- Scalar $\theta > 0$ allowed to converge to 0 as $n \to \infty$
- $\mathbf{\Pi} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_n)^T \in \mathbb{R}^{n \times K}$ is matrix of membership probability vectors
- $\mathbf{P} = (p_{kl}) \in \mathbb{R}^{K \times K}$ is nonsingular irreducible symmetric matrix with $p_{kl} \in [0, 1]$
- Assume number of communities K is finite but *unknown*
- Including SBM (with $\boldsymbol{\pi}_i \in \{\mathbf{e}_1, \cdots, \mathbf{e}_K\}$) as a special case

Population and Empirical Eigenstructures

- Denote by $\mathbf{H} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ eigendecomposition of mean matrix
  - $\mathbf{D} = \mathrm{diag}(d_1, \cdots, d_K)$ with $|d_1| \geq \cdots \geq |d_K| > 0$ is matrix of nonzero eigenvalues of descending order in *magnitude*
  - $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_K) \in \mathbb{R}^{n \times K}$ is orthonormal matrix of corresponding eigenvectors
- Denote by $\widehat{d}_1, \cdots, \widehat{d}_n$ eigenvalues of $\mathbf{X}$ and $\widehat{\mathbf{v}}_1, \cdots, \widehat{\mathbf{v}}_n$ corresponding eigenvectors
  - Without loss of generality assume $|\widehat{d}_1| \geq \cdots \geq |\widehat{d}_n|$ and denote by $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \cdots, \widehat{\mathbf{v}}_K) \in \mathbb{R}^{n \times K}$
- Asymptotic distributions of *spiked eigenvectors and eigenvalues* for random matrices with Wigner-type noises (Fan, Fan, Han and L., 2022a)

An Ideal Test Statistic

- First consider the case of known K
- By a simple permutation argument,

$$\text{under } H_0 : \pi_i = \pi_j, \text{ we have } \mathbf{V}(i) = \mathbf{V}(j)$$

  with ith and jth rows viewed as column vectors

- Motivated by this we consider following *ideal test statistic*

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

- $\boldsymbol{\Sigma}_1 = \text{cov}((\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{W}\mathbf{V}\mathbf{D}^{-1})$ with $\mathbf{e}_i \in \mathbb{R}^n$ unit vector
- Need some basic conditions to derive asymptotic distributions of T_{ij} under H_0 and H_1

Technical Conditions

- **Condition 1.** Assume that $\min_{1\le i\le K-1} \frac{|d_i|}{|d_{i+1}|} \ge 1 + c_0$ for some constant $c_0 > 0$ and $\alpha_n \to \infty$ as $n \to \infty$ with $\alpha_n^2 = \max_j \sum_{i=1}^n \mathbb{E} w_{ij}^2$

- **Condition 2.** There exist some small constants $0 < c_1, c_2 < 1$ such that $\lambda_K(\mathbf{\Pi}^T \mathbf{\Pi}) \ge c_1 n$, $\lambda_K(\mathbf{P}) \ge c_1$, and $\theta \ge n^{-c_2}$ with $\lambda_j(\cdot)$ standing for jth largest eigenvalue or singular value

- **Condition 3.** Assume that all eigenvalues of $n^2\theta\mathbf{\Sigma}_1$ are bounded away from 0 and ∞

Interpretations

- α_n measures the noise level
- Condition 1 allows for sparse network model
- Condition 2 ensures that

$$\alpha_n^2 \leq n\theta, \quad d_k \sim n\theta, \quad k = 1, \cdots, K$$

- Condition 2 also entails that average node degree is roughly of order $n\theta \geq n^{1-c_2}$

Asymptotic Distributions

Theorem 1

a). *Under Conditions 1–3 and null hypothesis H_0, it holds that*

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

b). *Under Conditions 1–2 and alternative hypothesis H_1, if $\sqrt{n\theta}\|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\| \to \infty$ then with asymptotic probability one,*

$$T_{ij} \to \infty.$$

If in addition Condition 3 holds, $\|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\| \sim \frac{1}{\sqrt{n\theta}}$, and $(\mathbf{V}(i) - \mathbf{V}(j))^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{V}(i) - \mathbf{V}(j)) \to \mu$ with μ some constant, then

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

Practical Test Statistic

- Ideal test statistic T_{ij} not directly applicable due to *unknown* population quantities K and covariance matrix $\mathbf{\Sigma}_1$
- Consider practical test statistic $\widehat{T}_{ij}$ by replacing K and $\mathbf{\Sigma}_1$ in T_{ij} with $\widehat{K}$ and $\widehat{\mathbf{S}}_1$, respectively

Theorem 2
Assume that

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

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

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

Estimation of Unknown Parameters

- A *simple thresholding* estimator for estimating K:

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

Proposition 1

The (a, b)th entry of matrix $\mathbf{\Sigma}_1$ takes the form

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

- A plug-in estimator of $\mathbf{\Sigma}_1$ can be constructed
  - $\mathbf{v}_a$ and d_a can be estimated by $\widehat{\mathbf{v}}_a$ and $\widehat{d}_a$, respectively
  - Estimation of σ_{ab}^2 is a bit more complicated...

An Iterative Algorithm for Estimating σ_{ab}^2

- Recall $\sigma_{ab}^2 = E[w_{ab}^2]$
- With $\widehat{K}$, the naive estimator $\widehat{w}_{0,ab}^2$ with $\widehat{\mathbf{W}}_0 = (\widehat{w}_{0,ab}) = \mathbf{X} - \sum_{k=1}^{\widehat{K}} \widehat{d}_k \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^T$ is not accurate enough
- We propose an iterative estimation procedure
  - Calculate the initial estimator $\widehat{\mathbf{W}}_0$
  - With $\widehat{\mathbf{W}}_0$, update the estimator of d_k as

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

  - Update the estimator of $\mathbf{W}$ as $\widehat{\mathbf{W}} \equiv (\widehat{w}_{ij}) = \mathbf{X} - \sum_{k=1}^{\widehat{K}} \widetilde{d}_k \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^T$. Estimate σ_{ab}^2 as $\widehat{\sigma}_{ab}^2 = \widehat{w}_{ab}^2$
- The above iterative procedure is motivated from high-order asymptotic expansion of sample eigenvalue $\widehat{d}_k$

Consistency of estimated parameters

Proposition 2

Under Conditions 1–3, we have $P(\widehat{K} = K) = 1 - o(1)$ *and* $n^2\theta\|\widehat{\mathbf{S}}_1 - \mathbf{\Sigma}_1\|_2 = o_p(1)$

Corollary 1

The asymptotic size of the rejection region

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

is α *and the asymptotic power is one as* $n \to \infty$

- Note that the rejection region is pivotal to unknown parameters

SIMPLE for Degree-Corrected Mixed Membership Models

Degree Heterogeneity

- Now consider the more general scenario of *degree heterogeneity*

$$\mathbf{H} = \mathbf{\Theta \Pi P \Pi}^T \mathbf{\Theta}$$

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

- $\mathbf{\Theta} = \text{diag}(\theta_1, \cdots, \theta_n)$ with $\theta_i > 0$ is degree heterogeneity matrix

- The previous test T_{ij} is no longer applicable because of degree heterogeneity

- We build a new test using the ratio statistic (Jin, 2015)

A Ratio Statistic

- Consider the following componentwise ratio to correct the degree heterogeneity

$$Y(i,k) = \frac{\widehat{\mathbf{v}}_k(i)}{\widehat{\mathbf{v}}_1(i)}, \quad 1 \le i \le n,\, 2 \le k \le K \tag{1}$$

with $0/0$ defined as 1

- Under the null hypothesis, due to the exchangeability of nodes i and j

$$\frac{\mathbf{v}_k(i)}{\mathbf{v}_1(i)} = \frac{\mathbf{v}_k(j)}{\mathbf{v}_1(j)}, \qquad 2 \le k \le K \tag{2}$$

- We build our test by comparing $\mathbf{Y}_i = (Y(i,2), \cdots, Y(i,K))^T$ with $\mathbf{Y}_j = (Y(j,2), \cdots, Y(j,K))^T$

An Ideal Test under Degree Heterogeneity

- For now assume that K is known
- We propose the test statistic to test $H_0 : \pi_i = \pi_j$

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

- $\boldsymbol{\Sigma}_2 = \mathrm{cov}(\mathbf{f})$ with $\mathbf{f} = (f_2, \cdots, f_K)^T$ and

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

- $\boldsymbol{\Sigma}_2$ is asymptotic covariance matrix of $\mathbf{Y}_i - \mathbf{Y}_j$

Conditions

- **Condition 4.** There exist some constants $c_2, c_3 \in (0,1)$ and constant $c_4 > 0$ such that $\min_{1\le k\le K} |\mathcal{N}_k| \ge c_2 n$, $\theta_{\max} \le c_4 \theta_{\min}$, and $\theta^2_{\min} \ge n^{-c_3}$

- **Condition 5.** Matrix $\mathbf{P} = (p_{kl})$ is positive definite, irreducible, and has unit diagonal entries. Moreover $n \min_{1\le k\le K,\, t=i,j} \mathrm{var}(\mathbf{e}_t^T \mathbf{W} \mathbf{v}_k) \to \infty$

- **Condition 6.** It holds that all the eigenvalues of $n\theta^2_{\min}\mathrm{cov}(\mathbf{f})$ are bounded away from 0 and ∞

- Condition 4 requires that there are enough pure nodes from each community and degree heterogeneity cannot be too extreme

- Degree density measured by $\theta^2_{\min}$ (converging to zero)

Asymptotic Distributions

Theorem 3

a). *Under Conditions 1 and 4–6 and null hypothesis H_0, it holds that*

$$G_{ij} \xrightarrow{d} \chi^2_{K-1} \text{ as } n \to \infty$$

b). *Under Conditions 1 and 4–6 and alternative hypothesis H_1, if $\lambda_2(\boldsymbol{\pi}_i \boldsymbol{\pi}_i^T + \boldsymbol{\pi}_j \boldsymbol{\pi}_j^T) \gg \frac{1}{n\theta^2_{\min}}$, then with asymptotic probability one,*

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

Practical Test Statistic

- Ideal test statistic G_{ij} is not directly applicable due to *unknown* population quantities K and covariance matrix $\mathbf{\Sigma}_2$

- Similarly, consider practical test statistic $\widehat{G}_{ij}$ by replacing K and $\mathbf{\Sigma}_2$ in G_{ij} with $\widehat{K}$ and $\widehat{\mathbf{S}}_2$, respectively

Theorem 4
Assume that

$$P(\widehat{K} = K) = 1 - o(1) \ \ \textit{and} \ \ n\theta_{\min}^2 \|\widehat{\mathbf{S}}_2 - \mathbf{\Sigma}_2\|_2 = o_p(1).$$

Then the same results as in Theorem 3 continue to hold for $\widehat{G}_{ij}$ under the same conditions

Estimation of Unknown Parameters

- We use the same thresholding estimator to estimate K

Proposition 3

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

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

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

$$+ \sigma_{ij}^2 \left[\frac{t_1 \mathbf{v}_{a+1}(j)}{t_{a+1}\mathbf{v}_1(i)} - \frac{\mathbf{v}_{a+1}(i)\mathbf{v}_1(j)}{\mathbf{v}_1(i)^2} - \frac{t_1 \mathbf{v}_{a+1}(i)}{t_{a+1}\mathbf{v}_1(j)} + \frac{\mathbf{v}_{a+1}(j)\mathbf{v}_1(i)}{\mathbf{v}_1(j)^2}\right]$$

$$\times \left[\frac{t_1 \mathbf{v}_{b+1}(j)}{t_{b+1}\mathbf{v}_1(i)} - \frac{\mathbf{v}_{b+1}(i)\mathbf{v}_1(j)}{\mathbf{v}_1(i)^2} - \frac{t_1 \mathbf{v}_{b+1}(i)}{t_{b+1}\widehat{\mathbf{v}}_1(j)} + \frac{\mathbf{v}_{b+1}(j)\mathbf{v}_1(i)}{\mathbf{v}_1(j)^2}\right]\Bigg\}$$

Continued

- The expansion involves population parameter t_k, whose definition is too complicated to include here
- We have results showing that $t_k/d_k \to 1$ and t_k is indeed the asymptotic mean of $\widehat{d}_k$
- A plug-in estimator $\widehat{\mathbf{S}}_2$ can be constructed for estimating $\mathbf{\Sigma}_2$
  - t_k can be estimated by $\widehat{d}_k$
  - $\mathbf{v}_a$ can be estimated by $\widehat{\mathbf{v}}_a$
  - σ_{ab}^2 can be estimated by one-step estimator $\widehat{\sigma}_{ab}^2$

Proposition 4

Under Conditions 1 and 4–6, estimator $\widehat{\mathbf{S}}_2$ *achieves the desired estimation accuracy, i.e.,*

$$n\theta_{\min}^2 \|\widehat{\mathbf{S}}_2 - \mathbf{\Sigma}_2\|_2 = o_p(1)$$

Rejection Region

Corollary 2

The asymptotic size of the rejection region

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

is α *and the asymptotic power is one as* $n \to \infty$

- The above rejection region is pivotal
- $\widehat{G}_{ij}$ can be used with or without degree heterogeneity
- Due to the ratio $\widehat{T}_{ij}$ has better practical performance without degree heterogeneity

Numerical Examples

Simulation Setting

- $n \in \{1500, 3000\}$ and $K = 3$ with significance 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

Simulation Results

Table 1: The size and power of test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ when the true value of K is used. The nominal level is 0.05 and sample size is $n = 1500$.

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

Continued

Table 2: The size and power of test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ when the true value of K is used. The nominal level is 0.05 and sample size is $n = 3000$.

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

Continued

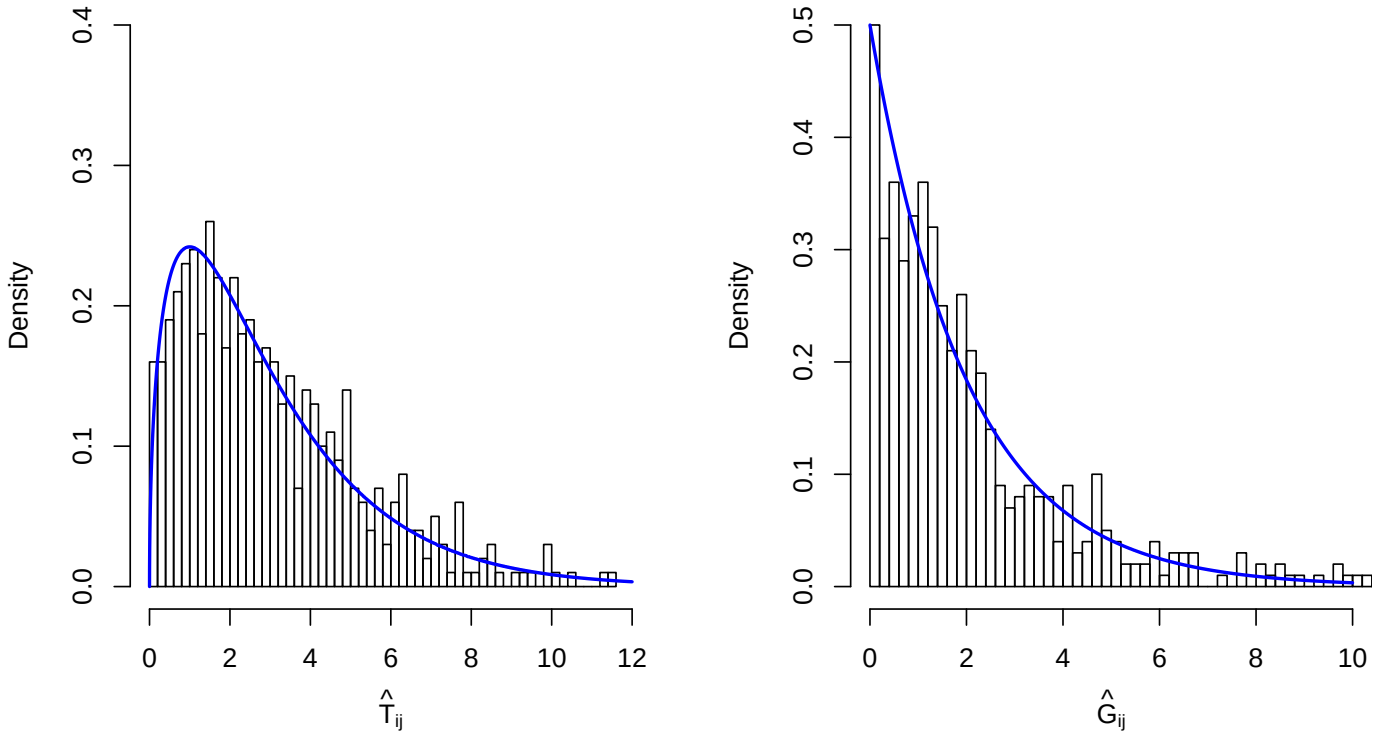


Figure 1: Left: the histogram of test statistic $\widehat{T}_{ij}$ under null hypothesis with known K when $\theta = 0.9$. Blue curve is the density function of χ^2_3. Right: the histogram of test statistic $\widehat{G}_{ij}$ under null hypothesis with known K when $r^2 = 0.9$. Blue curve is the density function of χ^2_2. Here sample size $n = 3000$.

Table 3: Estimation accuracy of K using the thresholding rule (27)

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

Continued

Table 4: The size and power of test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ when the estimated value of K is used. The nominal level is 0.05 and sample size is $n = 3000$.

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*)
- Each book is represented by a node and links between nodes represent frequent 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 extremely accurate
- In fact, as argued in multiple papers (e.g., Koutsourelakis and Eliassi-Rad (2008)), the mixed membership model may better suit this data set

Continued

- We will view the network as having $K = 2$ communities and treat neutral ones as having mixed memberships
- Consider the set of 9 books (Jin et al., 2017)

Table 7: Political books with labels

Title	Label (by Newman)	Node index
Empire	Neutral	105
The Future of Freedom	Neutral	104
Rise of the Vulcans	Conservative	59
All the Shah's Men	Neutral	29
Bush at War	Conservative	78
Plan of Attack	Neutral	77
Power Plays	Neutral	47
Meant To Be	Neutral	19
The Bushes	Conservative	50

Table 8: P-values based on test statistics $\widehat{T}_{ij}$. The labels provided by Newman are in the parentheses.

Node No.	105(N)	104(N)	59(C)	29(N)	78(C)	77(N)	47(N)	19(N)	50(C)
105(N)	1.0000	0.6766	0.0298	0.3112	0.0248	0.0000	0.0574	0.1013	0.0449
104(N)	0.6766	1.0000	0.0261	0.2487	0.0204	0.0000	0.0643	0.1184	0.0407
59(C)	0.0298	0.0261	1.0000	0.1546	0.2129	0.0013	0.0326	0.0513	0.9249
29(N)	0.3112	0.2487	0.1546	1.0000	0.3206	0.0034	0.0236	0.0497	0.2121
78(C)	0.0248	0.0204	0.2129	0.3206	1.0000	0.0991	0.0042	0.0084	0.2574
77(N)	0.0000	0.0000	0.0013	0.0034	0.0991	1.0000	0.0000	0.0000	0.0035
47(N)	0.0574	0.0643	0.0326	0.0236	0.0042	0.0000	1.0000	0.9004	0.0834
19(N)	0.1013	0.1184	0.0513	0.0497	0.0084	0.0000	0.9004	1.0000	0.1113
50(C)	0.0449	0.0407	0.9249	0.2121	0.2574	0.0035	0.0834	0.1113	1.0000

Table 9: P-values based on test statistics $\widehat{G}_{ij}$. The labels provided by Newman are in the parentheses.

Node No.	105(N)	104(N)	59(C)	29(N)	78(C)	77(N)	47(N)	19(N)	50(C)
105(N)	1.0000	0.4403	0.1730	0.4563	0.8307	0.5361	0.0000	0.0000	0.1920
104(N)	0.4403	1.0000	0.0773	0.9721	0.3665	0.6972	0.0000	0.0000	0.1144
59(C)	0.1730	0.0773	1.0000	0.0792	0.1337	0.0885	0.0000	0.0000	0.8141
29(N)	0.4563	0.9721	0.0792	1.0000	0.4256	0.7624	0.0000	0.0000	0.1153
78(C)	0.8307	0.3665	0.1337	0.4256	1.0000	0.5402	0.0000	0.0000	0.1591
77(N)	0.5361	0.6972	0.0885	0.7624	0.5402	1.0000	0.0000	0.0000	0.1294
47(N)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.9778	0.0000
19(N)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9778	1.0000	0.0000
50(C)	0.1920	0.1144	0.8141	0.1153	0.1591	0.1294	0.0000	0.0000	1.0000

Continued

- Our results based on $\widehat{G}_{ij}$ mostly consistent with labels provided by Newman and also consistent with those in Table 5 of Jin (2015)
- Books 3 and 9 are both labeled as “conservative” by Newman and our tests return large p-values between them
- These two books generally have much smaller p-values with books labeled as “neutral”
- Book 5, which was labeled as “conservative” by Newman, seems to be more similar to some neutral books

Continued

- This phenomenon also observed in Jin et al. (2017), who interpreted this as a result of having a liberal author
- Book 4 has relatively larger p-values with conservative books
- This book has even larger p-values with some other neutral books such as book 2
- Consistent with results in Jin et al. (2017) who reported that these two books have very close membership probability vectors

Visualization

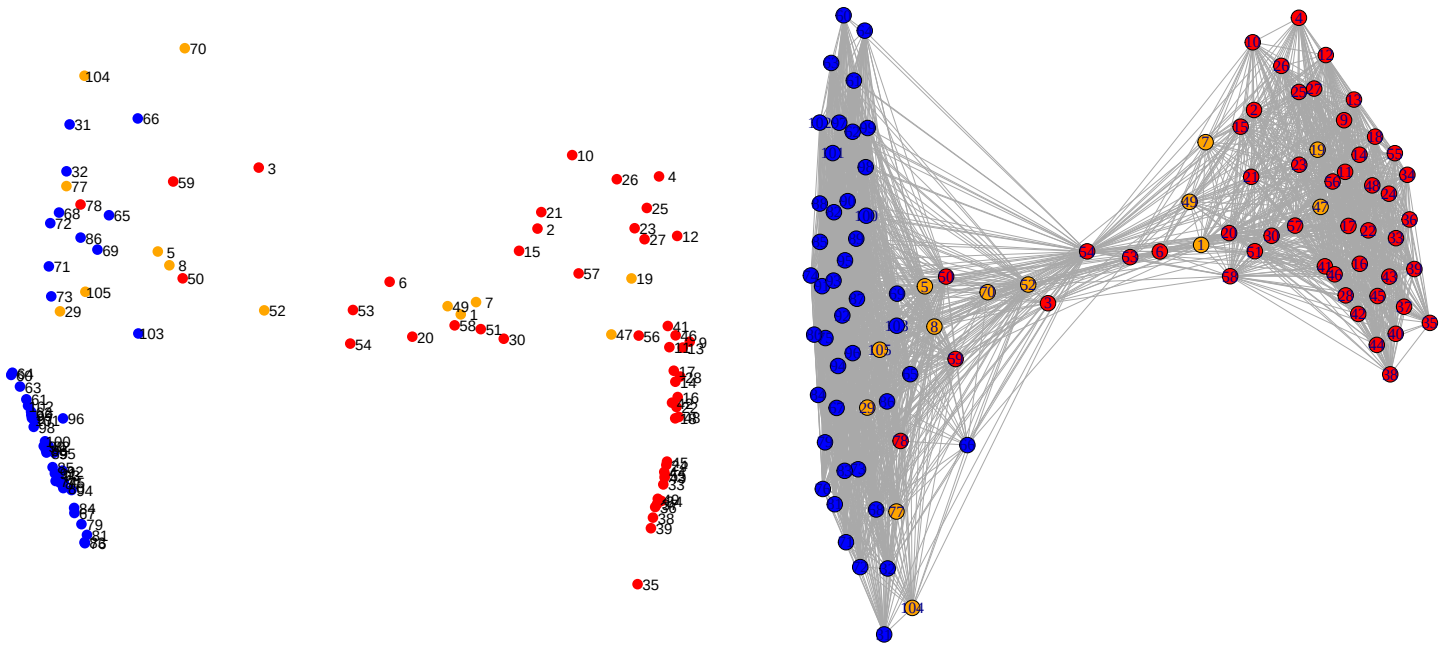


Figure 3: Left panel: the multidimensional scaling map of the nodes based on test statistics $\widehat{G}_{ij}$. Right panel: the connectivity graph generated from the thresholded p-valuate matrix based on $\widehat{G}_{ij}$. The nodes are color coded according to Newman's labels, with red representing "conservative," blue representing "liberal," and orange representing "neutral."

Conclusions

- Suggested a tool for *flexible network inference* with precise p-values on testing whether two nodes share same membership profiles
- Generally applicable to networks *with or without overlapping communities* allowing for *degree heterogeneity*
- Our SIMPLE framework *pivotal* to *unknown* parameters including K
- Provided *theoretical justifications* of our tests (both *size* and *power*)

References

- Fan, J., Fan, Y., Han, X. and Lv, J. (2022a). Asymptotic theory of eigenvectors for random matrices with diverging spikes. *Journal of the American Statistical Association* **117**, 996–1009.

- Fan, J., Fan, Y., Han, X. and Lv, J. (2022b). SIMPLE: statistical inference on membership profiles in large networks. *Journal of the Royal Statistical Society Series B* **84**, 630–653.