High-Dimensional Network Inference (part II)

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

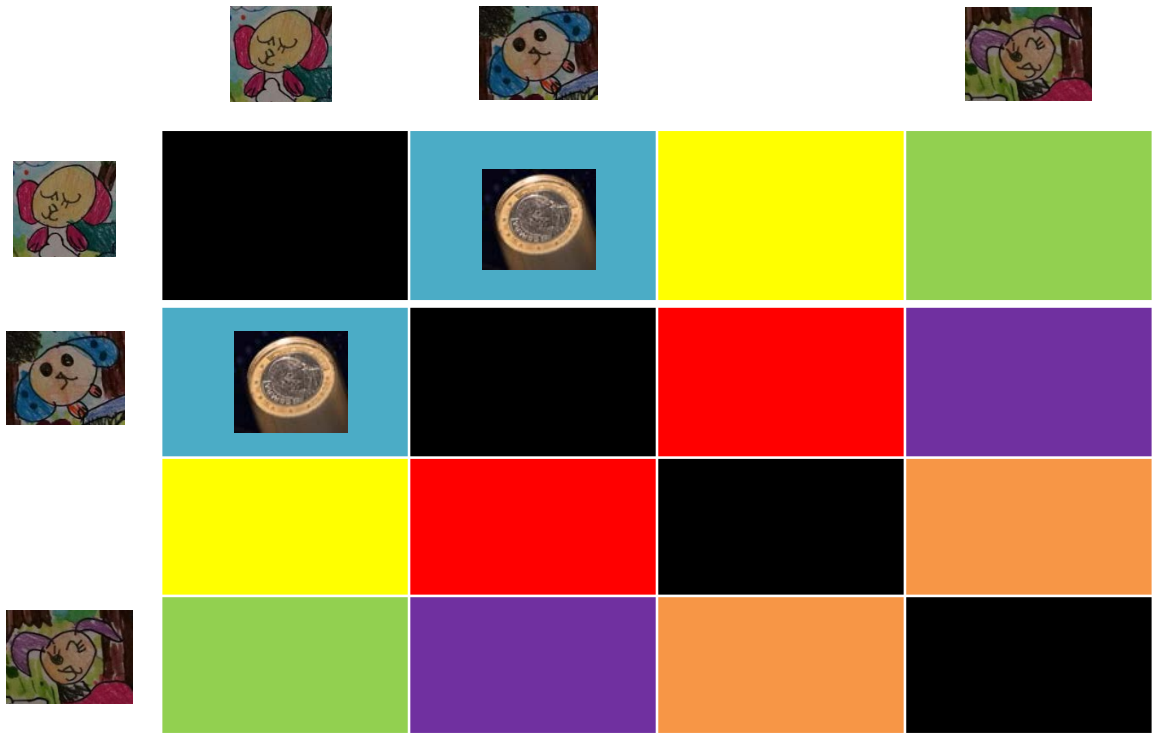
http://faculty.marshall.usc.edu/jinchi-lv

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

Outline of Fan, Fan, L. and Yang (2024)

- A motivating example
- Group network inference with SIMPLE-RC
- Theoretical justifications
- Numerical studies

A Networked World



- Individual nodes of a network (e.g. social media users or text documents) may share *similarities in the latent space*
- Common to provide *binary answers* (*i.e. Y/N)* based on community labeling given by *clustering*

Network P-Values

- *P-value tables* routinely produced and utilized for *linear and logistic regression* applications
- Important to provide a *p-value table* for *network applications*
- A *simple, natural* question is how to *test* whether a *pair* of social media users or text documents belong to *the same community*
- The recent work of SIMPLE (*statistical inference on membership profiles in large networks*; Fan, Fan, Han and L., 2022b) provided *a first attempt* toward such a practical need
- Accommodates *overlapping communities* and *degree heterogeneity*

Continued

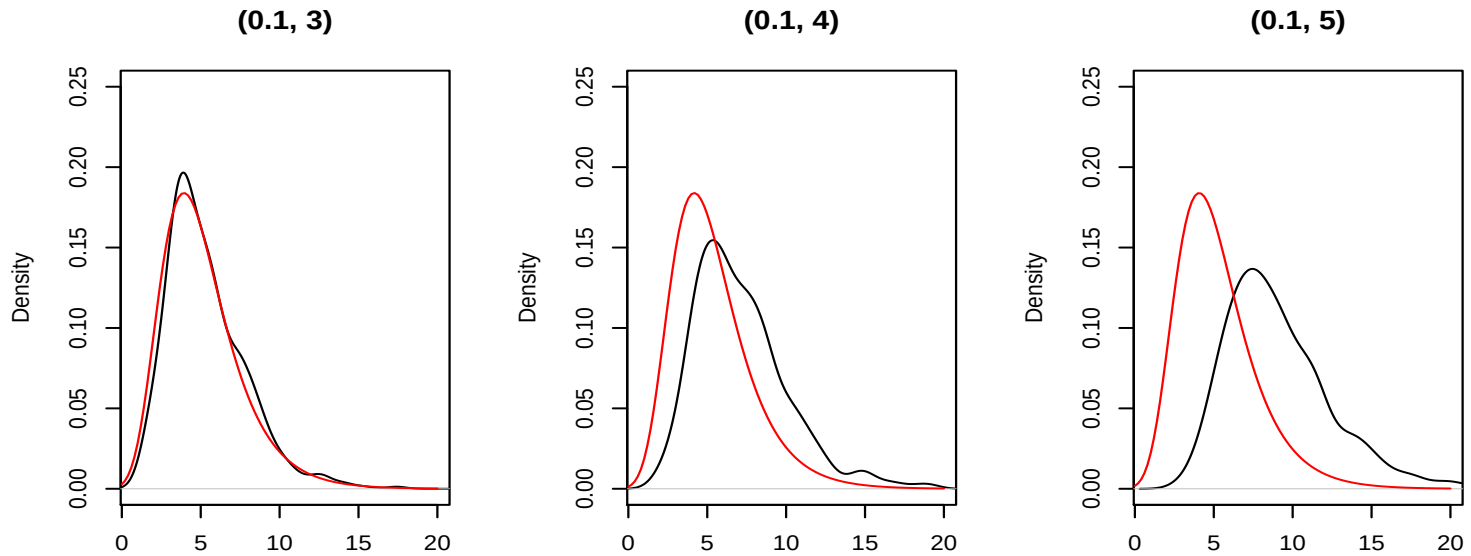
- In practice, we are often interested in investigating *a group of individuals* as opposed to *a pair of nodes*
- The group of individuals might share *similar* (*but not necessarily identical*) *community membership profiles*
- Real network applications may exhibit *much more network sparsity* and *much lower signal strength*, while SIMPLE required *relatively strong assumptions* on both *network sparsity* and *signal strength*
- Important to enable *network inference with flexibility and theoretical guarantees*

A Motivating Example

	Technology	Healthcare	Financial	Energy	Communication
Technology	0.1246	0.0247	0.0000	0.0001	0.0000
Healthcare	0.0247	0.0658	0.0279	0.0337	0.0000
Financial	0.0000	0.0279	0.7726	0.0004	0.0000
Energy	0.0001	0.0337	0.0004	0.8033	0.0000
Communication	0.0000	0.0000	0.0000	0.0000	0.7220

- Stocks in S&P 500 list can have *non-identical community membership profiles* even within the same sector of stock market (*due to complicated structures*)
- Desired to *test* whether a *group of individuals* (*network nodes*) might share *similar* (*not necessarily identical*) community membership profiles

An Interesting Phenomenon



- *Empirical null distributions* of *SIMPLE-RC test* (*to be introduced*) may deviate from *limiting distributions* under *weak signals*
- Choice of *parameter* K_0 is crucial (though true $K = 5$)
- *Any theoretical justifications under the lens of random matrix theory?*

Questions of Interest

- How to design a tool for *flexible group network inference* with precise p-values on testing whether a *group of nodes* might share *similar* (*not necessarily identical*) community membership profiles

- How to deal with the challenging case of *sparse networks and weak signals* and accommodate *popularly used* network models?

- How to develop a more general framework of asymptotic theory on *spiked eigenvectors and eigenvalues* for large *structured* random matrices powering *group network inference with non-sharp nulls and weak signals*?

Group Network Inference with SIMPLE-RC

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., 2022b; ...)

- 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

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

Continued

- Noise random matrix **W** known as *generalized Wigner matrix*
- Links x_{ij}'s independent Bernoulli random variables with means h_{ij}
- Assume network can be decomposed into K communities $C_1, \cdots, C_K$
- Each node i has community membership 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

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

- Allow number of communities K to be *slowly diverging* (of order $(\log n)^c$) and *unknown*

Group Network Inference with Non-Sharp Nulls

- For any given *group* of nodes $\mathscr{M} \subset \{1, \cdots, n\}$, our goal is to infer whether they share *similar* (*but not necessarily identical*) membership profiles (i.e. *probability vectors*) with quantified uncertainty level from observed adjacency matrix $\mathbf{X}$

- Interested in testing *non-sharp* null hypothesis

$$H_0 : \max_{i,j \in \mathscr{M}} \left\| \boldsymbol{\pi}_i - \boldsymbol{\pi}_j \right\| \leq c_{1n}$$

versus alternative hypothesis

$$H_a : \max_{i,j \in \mathscr{M}} \lambda_{\min}^{1/2} \left\{ \left(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j \right)^{\top} \left(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j \right) \right\} > c_{2n}$$

with $\lambda_{\min}\{\cdot\}$ denoting smallest eigenvalue and $c_{2n} > c_{1n}$ two positive sequences slowly converging to zero

Continued

- Formulation of alternative hypothesis motivated by our theoretical analyses

- H_0 gives an upper bound on $\max_{i,j\in\mathscr{M}} \lambda_{\min}^{1/2}\{\left(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j\right)^{\top}\left(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j\right)\}$, while H_a implies a lower bound on $\max_{i,j\in\mathscr{M}} \|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\|$

- To make problem more explicit, we first exploit *mixed membership model* without degree heterogeneity by assuming $\mathbb{E}\mathbf{X} = \mathbf{H} = \theta\boldsymbol{\Pi}\mathbf{P}\boldsymbol{\Pi}^T$ (Airoldi, Blei, Fienberg and Xing, 2008)

- We further exploit *degree-corrected mixed membership model* for degree heterogeneity assuming $\mathbb{E}\mathbf{X} = \mathbf{H} = \boldsymbol{\Theta}\boldsymbol{\Pi}\mathbf{P}\boldsymbol{\Pi}^T\boldsymbol{\Theta}$ (Zhang, Levina and Zhu, 2014; Jin, Ke and Luo, 2017)

Mixed Membership Model

- For now focus on *mixed membership model* without degree heterogeneity by assuming $\mathbb{E}\mathbf{X} = \mathbf{H} = \theta\mathbf{\Pi P \Pi}^T$ (Airoldi, Blei, Fienberg and Xing, 2008)

- Scalar $\theta > 0$ (*network sparsity parameter*) 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 and $\mathbf{P} = (p_{kl}) \in \mathbb{R}^{K \times K}$ is nonsingular matrix with $p_{kl} \in [0, 1]$

- Including stochastic block model with non-overlapping communities (when each $\boldsymbol{\pi}_i$ has one nonzero component)

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* in descending order in *magnitude* and $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_K) \in \mathbb{R}^{n \times K}$ 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 $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \cdots, \widehat{\mathbf{v}}_K) \in \mathbb{R}^{n \times K}$ (*consisting of top K empirical spiked eigenvectors*)

SIMPLE-RC for A Pair of Nodes

- To motivate our method SIMPLE-RC, begin with case of $m = |\mathscr{M}| = 2$ (*testing a pair of given network nodes* $\{i, j\}$)

- Let K_0 be an integer with $1 \le K_0 \le K$, $\mathbf{V}_{K_0}$ an $n \times K_0$ matrix formed by first K_0 columns of $\mathbf{V}$, and $\mathbf{D}_{K_0}$ a $K_0 \times K_0$ principal minor of $\mathbf{D}$ containing its first K_0 diagonal entries

- A simple observation is that under mixed membership model, H_0 entails

$$\|\mathbf{D}_{K_0}[\mathbf{V}_{K_0}(i) - \mathbf{V}_{K_0}(j)]\| \le c_1 n \sqrt{d_1 \theta_{\max}}$$

with $\theta_{\max} = \lambda_1(\mathbf{P})\theta$ (*ith and jth rows viewed as column vectors*)

Continued

- Another useful observation is that under mixed membership model, H_a entails

$$\|\mathbf{D}_{K_0}[\mathbf{V}_{K_0}(i) - \mathbf{V}_{K_0}(j)]\| \gtrsim c_{2n}\sqrt{d_K\theta_{\min}}$$

  with $\theta_{\min} = \lambda_K(\mathbf{P})\theta$, provided that

$$\|\mathbf{D}_{K_0}[\mathbf{V}_{K_0}(i) - \mathbf{V}_{K_0}(j)]\| \geq c\,\|\mathbf{D}[\mathbf{V}(i) - \mathbf{V}(j)]\|$$

  for some constant $c > 0$

- Under the above assumption, using only $\mathbf{V}_{K_0}$ (instead of $\mathbf{V}$) can still capture a significant fraction of difference between π_i and π_j

- Important for achieving high power using SIMPLE-RC

Continued

- Motivated by these observations, we suggest following *ideal* SIMPLE-RC test statistic to assess membership profile information for node pair $\{i, j\}$

$$T_{ij}(K_0) := \left[\widehat{\mathbf{V}}_{K_0}(i) - \widehat{\mathbf{V}}_{K_0}(j)\right]^T \left[\boldsymbol{\Sigma}_{i,j}(K_0)\right]^{-1} \left[\widehat{\mathbf{V}}_{K_0}(i) - \widehat{\mathbf{V}}_{K_0}(j)\right]$$

  with $1 \leq K_0 \leq K$ some pre-determined number and $\widehat{\mathbf{V}}_{K_0}$ the $n \times K_0$ matrix formed by first K_0 columns of $\widehat{\mathbf{V}}$

- $\boldsymbol{\Sigma}_{i,j}(K_0) = \mathrm{cov}[(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{W}\mathbf{V}_{K_0}\mathbf{D}_{K_0}^{-1}]$ with $\mathbf{e}_i \in \mathbb{R}^n$ unit vector with ith component 1

Continued

- Reduces to original SIMPLE test statistic (Fan, Fan, Han and L., 2022b) for the case of sharp null (*i.e.* $c_{1n} = 0$) and with choice of $K_0 = K$ (*for strong signals*)

- Choice of parameter $1 \leq K_0 \leq K$ for SIMPLE-RC plays a key role in network inference under weak signals (*one of major distinctions from the work of SIMPLE*)

- We can provide an estimate of covariance matrix $\mathbf{\Sigma}_{i,j}$ and specify choice of K_0 with theoretical justifications (*more details later*)

SIMPLE-RC for A Group of Nodes

- Now consider group testing for the case of diverging $m = |\mathscr{M}|$ and assume $m \in 2\mathbb{N}$ (*for simplicity*)
- A natural idea would be to investigate test statistic $\max_{\{i,j\} \subset \mathscr{M}} T_{ij}$
- Yet doing so is rather challenging because of potentially high correlations among all individual T_{ij}'s
- To deal with such a challenging issue, we suggest a *random coupling* strategy for group network inference
- This gives rise to the name of our *SIMPLE-RC* method

Continued

- Randomly pick pairs of nodes in group $\mathscr{M}$ without replacement until all nodes are coupled
- Denote by $\mathscr{P}$ set of resulting pairs of such random coupling
- Given random coupling set $\mathscr{P}$, formally define our SIMPLE-RC test statistic T as

$$T = \max_{\{i,j\} \in \mathscr{P}} T_{ij}$$

- We show formally that under suitable centering and rescaling, SIMPLE-RC test statistic T converges to a Gumbel distribution under H_0 (*more details later including power analysis*)

SIMPLE-RC with Degree Heterogeneity

- We further investigate more general case with *degree heterogeneity*
- Exploit *degree-corrected mixed membership model* for degree heterogeneity assuming $\mathbb{E}\mathbf{X} = \mathbf{H} = \mathbf{\Theta\Pi P\Pi}^T\mathbf{\Theta}$ (Zhang, Levina and Zhu, 2014; Jin, Ke and Luo, 2017)
- $\mathbf{\Theta} = \mathrm{diag}\{\theta_1, \cdots, \theta_n\}$ with $\theta_i > 0$ is degree heterogeneity matrix
- Also suggested another form of SIMPLE-RC test statistics $\mathscr{T}_{ij}$ and $\mathscr{T}$ (*similar flavor but different form*) and established parallel *asymptotic distributions* as well as *power analysis* (*exploiting eigenvector ratio statistics*)
- More details and comprehensive theory (Fan, Fan, L. and Yang, 2024)

What is the theory behind such a procedure?

Theoretical Justifications

Technical Conditions

- Condition 1:

  (i) (*Network sparsity*) It holds that $q \gg (\log n)^4$ with $q = \sqrt{n\theta}$

  (ii) (*Spiked eigenvalues*) It holds that $|d_k| \geq q \log\log n$ for all $1 \leq k \leq K_0$

  (iii) (*Eigengap*) There exists some constant $\varepsilon_0 > 0$ such that

  $$\min_{1 \leq k \leq K_0} \frac{|d_k|}{|d_{k+1}|} > 1 + \varepsilon_0,$$

  where we do not require eigengaps for smaller eigenvalues $|d_k|$ with $K_0 + 1 \leq k \leq K$

  (iv) (*Mean matrix*) There exists some constant $\varepsilon_1 > 0$ such that $\max_{i,j \in [n]} h_{ij} \leq 1 - \varepsilon_1$ and $\max_{i \in [n]} \sum_{j \in [n]} h_{ij} \geq \varepsilon_1 n\theta$, and the eigenvalues of $\mathbf{P}$ satisfy that $0 < \lambda_K(\mathbf{P}) \leq \cdots \leq \lambda_1(\mathbf{P}) \leq C$ for some large constant $C > 0$

Continued

- Condition 1:

  (v) (*Covariance matrix*) There exists some constant $0 < \varepsilon_2 < 1$ such that all the eigenvalues of $\theta^{-1}\mathbf{D}_{K_0'}\mathbf{\Sigma}_{i,j}(K_0')\mathbf{D}_{K_0'}$ are between ε_2 and ε_2^{-1} for all $\{i,j\} \subset \mathscr{M}$ and $1 \le K_0' \le K_0$

- For fixed ε_0 and ε_2, denote by $K_{\max} \equiv K_{\max}(n, \varepsilon_0, \varepsilon_2) \le K$ the largest K_0 such that parts (ii), (iii), and (v) above hold

- Parameter q is key to our technical study (*giving typical size of eigenvalues of noise random matrix* $\mathbf{W}$)

- *Much weaker* assumptions on *network sparsity* and *signal strength* (*i.e. signal-to-noise ratio*)

SIMPLE-RC for A Pair of Nodes

Theorem 1. Assume that some regularity conditions hold, K_0 is a random variable such that $1 \le K_0 \le K_{\max} \wedge C_0$ almost surely for some large constant $C_0 > 0$, and $1 \le K \ll \frac{q}{(\sqrt{n}\|\mathbf{V}\|_{\max})\log n} \wedge \frac{|d_{K_0}|^2}{(\sqrt{n}\|\mathbf{V}\|_{\max})^2 q^2}$. Then test statistic $T_{ij}(K_0)$ satisfies that

(i) If $c_{1n} \ll [d_1\lambda_1(\mathbf{P})]^{-\frac{1}{2}}$, it holds that under null hypothesis H_0,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\mathbb{P}\left\{T_{ij}(K_0) \le x\right\} - F_{K_0}(x)\right| = 0,$$

where conditional on K_0, F_{K_0} is chi-square distribution with K_0 degrees of freedom

(ii) If $c_{2n} \gg [d_K\lambda_K(\mathbf{P})]^{-\frac{1}{2}}$, it holds that under alternative hypothesis H_a,

$$\lim_{n\to\infty} \mathbb{P}\left\{T_{ij}(K_0) > C\right\} = 1$$

for each arbitrarily large constant $C > 0$

Continued

- Establishes important extensions of recent results (Fan, Fan, Han and L., 2022a and 2022b)
- Considers hypothesis testing with non-sharp nulls
- Allows for slowly diverging number of communities K
- Relaxes lower bound on parameter q from earlier $q \geq n^{\varepsilon}$ to $q \gg (\log n)^4$ (*much sparser networks accommodated in our setting*)
- Relaxes lower bound on signal-to-noise ratio $|d_{K_0}|/q$ from earlier n^{ε} to $\sqrt{\log n}$

Theorem 2. Assume that some regularity conditions hold. Then SIMPLE-RC test statistic T satisfies that under null hypothesis H_0,

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{T(K_0) - b_m(K_0)}{2} \le x \right\} - \mathscr{G}(x) \right| = 0,$$

where $\mathscr{G}(x) = \exp(-e^{-x})$ denotes the Gumbel distribution and

$$b_m(K_0) = 2\log\frac{m}{2} + (K_0 - 2)\log\log\frac{m}{2} - 2\log\Gamma\left(\frac{K_0}{2}\right)$$

with $\Gamma(\cdot)$ representing the gamma function

Continued

- Individual test statistics T_{ij} based on random coupling shown to be asymptotically independent
- When m is bounded, asymptotic distribution of SIMPLE-RC test statistic T becomes maximum of $m/2$ independent $\chi^2_{K_0}$ random variables under H_0
- Focus on more interesting case of *diverging m* in this work
- Interesting to see that limiting null distribution is *free of random variable* K_0
- Helpful in deriving asymptotic null distribution when we replace K_0 with its sample counterpart later

Theorem 3. Assume that some regularity conditions hold and

$$\max_{\{i,j\}\subset\mathscr{M}} \|\mathbf{D}_{K_0}[\mathbf{V}_{K_0}(i)-\mathbf{V}_{K_0}(j)]\| \geq c \max_{\{i,j\}\subset\mathscr{M}} \|\mathbf{D}[\mathbf{V}(i)-\mathbf{V}(j)]\|$$

for a constant $c>0$ almost surely. If $c_{2n} \gg [d_K \lambda_K(\mathbf{P})]^{-1/2}\sqrt{\log n}$, then SIMPLE-RC test statistic T satisfies that under alternative hypothesis H_a, for each arbitrarily large constant $C>0$,

$$\lim_{n\to\infty} \mathbb{P}\left\{\frac{T(K_0)-b_m(K_0)}{2} > C\right\} = 1$$

Estimation of Covariance Matrices and Empirical Versions

- Need to provide an estimate of covariance matrix $\mathbf{\Sigma}_{i,j}$ and specify choice of K_0
- Suggest consistent estimator $\widehat{\mathbf{\Sigma}}_{i,j}(K_0)$ of covariance matrix $\mathbf{\Sigma}_{i,j}(K_0)$ based on residual matrix $\widehat{\mathbf{W}} := \mathbf{X} - \sum_{k=1}^{K_0} \widehat{d}_k \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^\top$
- Such estimator disregards completely weak signals $\widehat{d}_k$ with $K_0 + 1 \le k \le d$
- For testing a pair of nodes, suggest to use estimate

$$\widehat{K}_0 := \max\left\{k \in [n] : |\widehat{d}_k| \ge \check{q}(\log n)^{1/2} \cdot \log\log n\right\}$$

with $\check{q} > 0$ and $\check{q}^2 := \max_{j\in[n]} \sum_{l=1}^{n} X_{lj}$ maximum node degree

Continued

- For the group test, suggest to use estimate

$$\widehat{K}_0 := \max\left\{k \in [n] : |\widehat{d}_k| \geq \check{q}(\log n)^{3/2} \cdot \log\log n\right\}$$

- One can replace factor $\log\log n$ with another sequence $C_n \to \infty$ as $n \to \infty$

- *Consistency of covariance matrix estimator* and corresponding *asymptotic null distributions* and *power analysis* for *SIMPLE-RC test with estimates* $\widehat{\mathbf{\Sigma}}_{i,j}(K_0)$ *and* $\widehat{K}_0$ rigorously established (Fan, Fan, L. and Yang, 2024)

SIMPLE-RC with Degree Heterogeneity

- Further investigate more general case with degree heterogeneity
- Additional technical conditions required for dealing with more challenging case of degree heterogeneity
- *Asymptotic null distributions* and *power analysis* for *SIMPLE-RC test statistics $\mathscr{T}_{ij}$ and $\mathscr{T}$ with degree heterogeneity* formally justified (Fan, Fan, L. and Yang, 2024)
- In contrast, we *lose one degree of freedom* in asymptotic null distributions due to use of eigenvector ratio statistics

What is the key tool powering the theory of flexible group network inference?

A General Theoretical Foundation

- Our technical analyses empowered by *novel asymptotic expansions of spiked eigenvectors* for large random matrices with *weak spikes*
- Exploits tools of the Cauchy integral formula related to random matrix $\mathbf{X}$ and the *Green function* $\mathbf{G}(z)$ (*i.e., resolvent*) for *noise random matrix* $\mathbf{W}$
- Need to characterize *asymptotic behavior of* $\mathbf{x}^\top \mathbf{G}(z)\mathbf{y}$ for any deterministic unit vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (*convergence to a deterministic limit named anisotropic local law*)
- Key innovation in deriving *sharper anisotropic local law* for resolvent $\mathbf{G}(z)$ under *weaker* conditions on sparsity level and signal strength

Continued

- *Much weaker signals* with signal-to-noise ratio as low as $\sqrt{\log n}$ and *much sparser networks* with sparsity as low as $(\log n)^8/n$
- *Much finer and more delicate combinatorial arguments* needed for evaluating huge products of random matrices (*exploiting quadratic vector equation (QVE) instead of series expansion*)
- Anisotropic local laws enable us to further derive *precise asymptotic expansions* of empirical *spiked eigenvectors*
- *Uniform results* on *asymptotic distributions* of empirical spiked eigenvectors key to *random coupling* for *group network inference*
- *More comprehensive theory* (Fan, Fan, L. and Yang, 2024)

Numerical Studies

Simulation Settings

- First example considers *mixed membership model*
- Set $n = 3000$ and $K = 5$ (*each having* $n_0 = 300$ *pure nodes*)
- Divide remaining $n - Kn_0$ nodes into *four groups* of equal size with community membership probability vectors $\mathbf{a}_l$'s
- Set $\mathbf{a}_1 = (0.1, 0.6, 0.1, 0.1, 0.1)^T$, $\mathbf{a}_2 = (0.6, 0.1, 0.1, 0.1, 0.1)^T$, $\mathbf{a}_3 = (0.1, 0.1, 0.6, 0.1, 0.1)^T$, and $\mathbf{a}_4 = (1/K, \cdots, 1/K)^T$
- $n \times K$ matrix of community membership probability vectors $\mathbf{\Pi}$
- Choose matrix $\mathbf{P}$ as a $K \times K$ nonsingular matrix with diagonal entries one and (i, j)th entry $\rho/|i - j|$ for $1 \le i \ne j \le n$ with $\rho = 0.2$

Continued

- Vary *sparsity parameter* θ in $\{0.1, 0.2, \cdots, 0.8\}$ (*smaller value for lower average node degree and weaker signal strength*)
- For *null hypothesis* H_0, choose a representative group $\mathscr{M}$ of $m = |\mathscr{M}| = 10$ or 20 nodes from *non-pure membership profile group* with community membership probability vector $\mathbf{a}_1$
- Apply SIMPLE-RC test with parameter K_0 chosen to be in $\{3, 4, 5\}$ and repeat 500 times
- Further consider *DCMM model* (*with network sparsity parameter* r^2) and *power analysis* (*with parameter* δ *measuring distance between two subgroups*)

Simulation Results

m	K_0	θ							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
10	3	0.052	0.028	0.034	0.04	0.032	0.02	0.022	0.028
	4	0.148	0.094	0.086	0.078	0.086	0.06	0.056	0.076
	5	0.328	0.188	0.204	0.182	0.194	0.142	0.132	0.138
20	3	0.038	0.026	0.028	0.018	0.024	0.032	0.018	0.024
	4	0.108	0.064	0.06	0.048	0.044	0.056	0.04	0.064
	5	0.246	0.15	0.13	0.116	0.116	0.104	0.09	0.104

TABLE 1

The empirical sizes of the SIMPLE-RC test with test statistic T under different values of (m, K_0, θ) and with nominal level $\alpha = 0.05$ for simulation example 1 in Section 5.1.

Continued

m	K_0	r^2							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
10	3	0.102	0.072	0.044	0.036	0.046	0.038	0.034	0.036
	4	0.198	0.128	0.098	0.09	0.1	0.086	0.09	0.072
	5	0.43	0.246	0.196	0.19	0.16	0.174	0.162	0.136
20	3	0.118	0.044	0.05	0.05	0.052	0.044	0.03	0.042
	4	0.212	0.088	0.094	0.084	0.092	0.058	0.058	0.078
	5	0.372	0.192	0.174	0.15	0.162	0.12	0.116	0.124

TABLE 2

The empirical sizes of the SIMPLE-RC test with test statistic $\mathcal{T}$ under different values of (m, K_0, r^2) and with nominal level $\alpha = 0.05$ for simulation example 2 in Section 5.1.

m	δ	θ							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
10	0.5	0.922	0.998	1	1	1	1	1	1
	0.4	0.496	0.756	0.888	0.966	0.992	0.994	1	0.998
	0.3	0.132	0.18	0.218	0.288	0.42	0.56	0.598	0.722
	0.2	0.098	0.088	0.088	0.1	0.116	0.176	0.202	0.254
	0.1	0.088	0.05	0.058	0.05	0.066	0.082	0.08	0.082
	0	0.09	0.044	0.05	0.044	0.042	0.058	0.04	0.028
20	0.5	0.978	0.998	1	0.998	1	0.998	0.998	1
	0.4	0.596	0.884	0.976	0.996	1	0.998	0.998	1
	0.3	0.146	0.166	0.252	0.384	0.462	0.604	0.678	0.794
	0.2	0.106	0.09	0.076	0.116	0.116	0.172	0.216	0.248
	0.1	0.09	0.066	0.054	0.064	0.044	0.07	0.06	0.094
	0	0.07	0.06	0.038	0.036	0.028	0.032	0.022	0.042

TABLE 3

The empirical powers of the SIMPLE-RC test with test statistic T under different values of (m, δ, θ) and with nominal level $\alpha = 0.05$ for simulation example 3 in Section 5.2, where parameter K_0 is chosen as 3.

Continued

m	δ	r^2							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
10	0.5	0.804	0.96	0.986	1	1	1	1	1
	0.4	0.326	0.478	0.666	0.788	0.872	0.944	0.952	0.97
	0.3	0.138	0.122	0.148	0.178	0.186	0.216	0.296	0.282
	0.2	0.134	0.078	0.064	0.066	0.068	0.068	0.076	0.066
	0.1	0.132	0.068	0.056	0.05	0.058	0.06	0.072	0.05
	0	0.132	0.056	0.052	0.044	0.044	0.06	0.07	0.042
20	0.5	0.908	0.994	0.998	1	1	0.996	1	1
	0.4	0.48	0.596	0.766	0.888	0.966	0.984	0.992	1
	0.3	0.224	0.098	0.16	0.168	0.238	0.272	0.292	0.354
	0.2	0.19	0.062	0.092	0.066	0.062	0.064	0.058	0.092
	0.1	0.202	0.07	0.058	0.048	0.046	0.044	0.042	0.046
	0	0.186	0.054	0.07	0.056	0.044	0.042	0.042	0.052

TABLE 4

The empirical powers of the SIMPLE-RC test with test statistic $\mathcal{T}$ under different values of (m, δ, r^2) and with nominal level $\alpha = 0.05$ for simulation example 4 in Section 5.2, where parameter K_0 is chosen as 3.

Continued

- The empirical sizes of both forms of SIMPLE-RC test generally controlled at the nominal level of $\alpha = 0.05$ across different model settings (*in line with our theoretical results*)
- The power of SIMPLE-RC test generally enhances as the *distance between the two subgroups* increases and the *signal strength* becomes stronger
- A *larger value of m* can boost the power of group network inference with SIMPLE-RC under *weaker signals*

Continued

- An interesting phenomenon that a *lower value of parameter* K_0 may be needed to alleviate the practical issue of *rather weak signals*
- In contrast, the choice of K_0 as the true value of $K = 5$ can render the *sizes much inflated* at the presence of *weak signals*
- More comprehensive numerical studies on group network inference (as well as *a financial application*) investigated in the paper (*similar empirical findings*)

Conclusions

- Suggested a tool for *group network inference* with precise p-values on testing whether two groups of nodes share *similar* membership profiles
- Generally applicable to networks *with or without overlapping communities* and *degree heterogeneity*
- Established simple-to-use *asymptotic null distributions* and *power analysis* empowered by our new theory for *random matrices with weak spikes*
- Revealed an interesting phenomenon of *eigen-selection* for *valid network inference*

References

- Fan, J., Fan, Y., Lv, J. and Yang, F. (2024). SIMPLE-RC: group network inference with non-sharp nulls and weak signals. *Manuscript*.

- Han, X., Tong, X. and Fan, Y. (2023). Eigen selection in spectral clustering: a theory guided practice. *Journal of the American Statistical Association* **118**, 109–121.

- Han, X., Yang, Q. and Fan, Y. (2023). Universal rank inference via residual subsampling with application to large networks. *The Annals of Statistics* **51**, 1109–1133.