SIMPLE-RC: Group Network Inference with Non-Sharp Nulls and Weak Signals

Fan Yang

YMSC, Tsinghua University

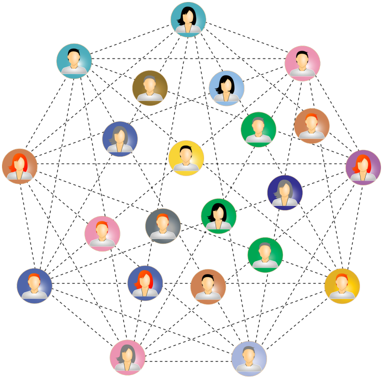
Joint work with Jianqing Fan, Yingying Fan, and Jinchi Lv

BU-Keio-Tsinghua conference

July 1, 2023

Overview

A World of Networks



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

P-Values for Networks

- It is also desirable to provide a *p-value table* for *network applications*.
- A *simple, natural* question is how to *test* whether a *pair* of social media users belong to *the same community*.
- The recent work of SIMPLE (*statistical inference on membership profiles in large networks*; Fan, Fan, Han and Lv, 2022b) provided *a first attempt* toward such a practical need.
- The approach can accommodate both *overlapping communities* and *degree heterogeneity*.

Beyond SIMPLE

- 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 requires *relatively strong assumptions* on both *network sparsity* and *signal strength*.
- Thus, it is important to enable *network inference with flexibility and theoretical guarantees* beyond SIMPLE.

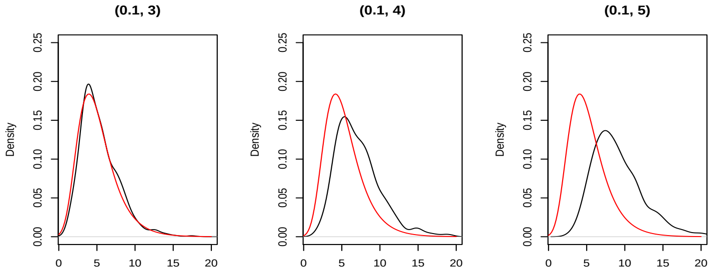
A Motivating Example

- Construct an adjacency matrix for the stocks in S&P 500 using the time series of the daily log returns. Performing network inference gives the following p-value table.

	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 can have *non-identical community membership profiles* even within the same sector of the stock market.
- 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 (# of signals) is crucial (the true # of communities = 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* (*instead of a pair*) might share *similar* (*not necessarily identical*) community membership profiles?

- How to deal with the challenging case of *sparse networks and weak signals*?

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

Overview

Background and motivation

Group network inference with SIMPLE-RC

Theoretical Justifications

Model Setting

A general network model

Consider a network with n nodes $\{1, \cdots, n\}$ and its *adjacency matrix* $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times n}$. $\mathbf{X}$ can be written as a signal-plus-noise matrix:

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

- Links x_{ij}'s independent Bernoulli random variables with means h_{ij}.
- $\mathbf{H} = \mathbb{E}\mathbf{X} = (h_{ij}) \in \mathbb{R}^{n\times n}$ is deterministic mean matrix (signal).
- $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n\times n}$ is symmetric random noise matrix with independent (up to symmetry) entries satisfying $\mathbb{E}w_{ij} = 0$. Known as a *Wigner-type matrix*.

Model Setting

A general network model

Consider a network with n nodes $\{1, \cdots, n\}$ and its *adjacency matrix* $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times n}$. $\mathbf{X}$ can be written as a signal-plus-noise matrix:

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

- Links x_{ij}'s independent Bernoulli random variables with means h_{ij}.
- $\mathbf{H} = \mathbb{E}\mathbf{X} = (h_{ij}) \in \mathbb{R}^{n \times n}$ is deterministic mean matrix (signal).
- $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{n \times n}$ is symmetric random noise matrix with independent (up to symmetry) entries satisfying $\mathbb{E}w_{ij} = 0$. Known as a *Wigner-type matrix*.

Assume the network can be decomposed into K communities $C_1, \cdots, C_K$ ($\operatorname{rank} \mathbf{H} = K$). Each node i has community membership probability vector $\boldsymbol{\pi}_i = (\boldsymbol{\pi}_i(1), \cdots, \boldsymbol{\pi}_i(K))^T$ with $\boldsymbol{\pi}_i(k) \in [0, 1]$, $\sum_{k=1}^K \boldsymbol{\pi}_i(k) = 1$, and

$$\mathbb{P}\{\text{node } i \text{ belongs to community } C_k\} = \boldsymbol{\pi}_i(k).$$

Let $K = O(1)$ be an *unknown* parameter. We can allow K to be *slowly diverging* ($\sim (\log n)^c$).

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 $\mathbf{X}$.
- We are 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

$$H_a : \max_{i,j \in \mathscr{M}} \left\| \boldsymbol{\pi}_i - \boldsymbol{\pi}_j \right\| > c_{2n}$$

with $c_{2n} > c_{1n}$ two positive sequences slowly converging to zero.

Mixed Membership Model

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

$$h_{ij} = \theta \sum_{k,l=1}^{K} \pi_i(k)\pi_j(l)p_{kl}.$$

$\mathbf{\Pi} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_n)^T \in \mathbb{R}^{n\times K}$ is *matrix of membership probability vectors*, $\mathbf{P} = (p_{kl})$ is a nonsingular matrix with $p_{kl} \in [0, 1]$, $n^{-1} \ll \theta \le 1$ is the *network sparsity parameter*.

(SBM is a special case with non-overlapping communities when each $\boldsymbol{\pi}_i$ has only one nonzero component.)

Mixed Membership Model

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

$$h_{ij} = \theta \sum_{k,l=1}^{K} \pi_i(k) \pi_j(l) p_{kl}.$$

$\mathbf{\Pi} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_n)^T \in \mathbb{R}^{n \times K}$ is *matrix of membership probability vectors*, $\mathbf{P} = (p_{kl})$ is a nonsingular matrix with $p_{kl} \in [0, 1]$, $n^{-1} \ll \theta \leq 1$ is the *network sparsity parameter*.

(SBM is a special case with non-overlapping communities when each $\boldsymbol{\pi}_i$ has only one nonzero component.)

- $\mathbf{H} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ is the eigendecomposition. $\mathbf{D} = \mathrm{diag}\{d_1, \cdots, d_K\}$ with $|d_1| \geq \cdots \geq |d_K| > 0$ is matrix of nonzero *eigenvalues* in descending order and $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_K)$ 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*.
- Let $|\widehat{d}_1| \geq \cdots \geq |\widehat{d}_K|$ 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 SIMPLE-RC, begin with the simple case $m = |\mathcal{M}| = 2$ (*testing a pair of given network nodes* $\{i, j\}$). Let K_0 be an integer with $1 \leq K_0 \leq K$, $\mathbf{V}_{K_0}$ 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.

- First observation: under mixed membership model, H_0 entails

$$\left\|\mathbf{D}_{K_0}\left[\mathbf{V}_{K_0}(i) - \mathbf{V}_{K_0}(j)\right]\right\| \leq c_{1n}\sqrt{d_1\,\theta_{\max}}$$

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

- Second observation: under mixed membership model, H_a entails

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

with $\theta_{\min} = \lambda_K(\mathbf{P})\theta$.

SIMPLE-RC for a Pair of Nodes

These observations suggest the following *ideal* SIMPLE-RC test statistic to assess membership profile information for the node pair $\{i, j\}$:

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

where $1 \leq K_0 \leq K$ is a pre-determined number, $\widehat{\mathbf{V}}_{K_0}$ is the $n \times K_0$ matrix formed by first K_0 columns of $\widehat{\mathbf{V}}$, and $\boldsymbol{\Sigma}_{i,j}(K_0) = \mathrm{cov}[(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{W}\mathbf{V}_{K_0}\mathbf{D}_{K_0}^{-1}]$ is the asymptotic covariance matrix ($\mathbf{e}_i$ standard unit vector in the ith direction).

- It reduces to original SIMPLE test statistic (Fan, Fan, Han and Lv, 2022b) for the case of sharp null $(c_{1n} = 0)$ and with choice of $K_0 = K$ (*for strong signals*).
- Choice of K_0 for SIMPLE-RC is crucial in network inference under weak signals (*one main difference from SIMPLE*).
- We can provide an estimate of $\boldsymbol{\Sigma}_{i,j}$ and specify the choice of K_0 with theoretical justifications (*more details later*).

SIMPLE-RC for a Group of Nodes

Consider group testing for the case of diverging $m = |\mathcal{M}| \to \infty$ and assume $m \in 2\mathbb{N}$ (*for simplicity*). A natural idea would be to investigate test statistic $\max_{\{i,j\}\subset\mathcal{M}} T_{ij}$, but doing so is rather challenging because of potentially high correlations among all individual T_{ij}'s.

SIMPLE-RC for a Group of Nodes

Consider group testing for the case of diverging $m = |\mathscr{M}| \to \infty$ and assume $m \in 2\mathbb{N}$ (*for simplicity*). A natural idea would be to investigate test statistic $\max_{\{i,j\} \subset \mathscr{M}} T_{ij}$, but 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, i.e., the *SIMPLE-RC* method:

- Randomly pick pairs of nodes in group $\mathscr{M}$ without replacement until all nodes are coupled. Denote by $\mathscr{P}$ the set of 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, T converges to a Gumbel distribution under H_0 (*more details later including power analysis*).

The finite m case is simpler due to the fact that individual test statistics T_{ij} based on random coupling are asymptotically independent.

SIMPLE-RC with Degree Heterogeneity

We also consider the more general case with *degree heterogeneity*.

- *Degree-corrected mixed membership model* for degree heterogeneity assuming $\mathbb{E}\mathbf{X} = \mathbf{H} = \boldsymbol{\Theta\Pi P\Pi}^T\boldsymbol{\Theta}$ (Zhang, Levina and Zhu, 2014; Jin, Ke and Luo, 2017)
- $\boldsymbol{\Theta} = \mathrm{diag}\{\vartheta_1, \cdots, \vartheta_n\}$ with $\vartheta_i > 0$ being degree heterogeneity.
- Suggest 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* by exploiting *eigenvector ratio statistics*:

$$\widehat{v}_k(i) \to \frac{\widehat{v}_k(i)}{\widehat{v}_1(i)}.$$

- More details and comprehensive theory can be found in Fan, Fan, Lv, and Y., 2022.

Overview

Background and motivation

Group network inference with SIMPLE-RC

Theoretical Justifications

Technical Conditions

Suppose for some $1 \le K_0 \le K$ (K_0 *can be random*), the following conditions hold.

(i) (*Network sparsity*) $\theta \gg (\log n)^8/n$

(ii) (*Spiked eigenvalues*) $|d_k| \gg q \log n$ for $1 \le k \le K_0$, where $q = \sqrt{n\theta}$.

(iii) (*Eigengap*) The spiked eigenvalues are non-degenerate:

$$\min_{1 \le k \le K_0} |d_k|/|d_{k+1}| > 1 + \varepsilon_0.$$

No eigengaps required for smaller eigenvalues $|d_k|$ with $k > K_0$.

(iv) (*Mean matrix*) $0 < \lambda_K(\mathbf{P}) \le \cdots \le \lambda_1(\mathbf{P}) \le C$ for some large constant $C > 0$.

(v) (*Covariance matrix*) $\mathbf{D}_{K_0} \boldsymbol{\Sigma}_{i,j}(K_0) \mathbf{D}_{K_0} \sim \theta$ in the sense of eigenvalues.

- q is a key parameter: *CLT fluctuation of the node degrees* and *typical size of eigenvalues of noise matrix* $\mathbf{W}$.
- Relax $\theta \ge n^{-1+\varepsilon}$ to $\theta \gg (\log n)^8/n$ (i.e., *much sparser networks*).
- Relax $|d_k|/q \ge n^{\varepsilon}$ to $\gg \log n$ (i.e., *much weaker signals*).

SIMPLE-RC for A Pair of Nodes

Theorem (SIMPLE-RC for a pair)

Under the above technical conditions, the test statistic $T_{ij}(K_0)$ satisfies;

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

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left\{ T_{ij}(K_0) \leq x \right\} - F_{K_0}(x) \right| \to 0,$$

where conditional on K_0, F_{K_0} is $\chi^2_{K_0}$ distribution.

(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 large constant $C > 0$.

SIMPLE-RC for A Pair of Nodes

Theorem (SIMPLE-RC for a pair)

Under the above technical conditions, the test statistic $T_{ij}(K_0)$ satisfies;

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

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

where conditional on K_0, F_{K_0} is $\chi^2_{K_0}$ distribution.

(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 large constant $C > 0$.

Establishes important extensions of SIMPLE (Fan, Fan, Han and Lv, 2022a and 2022b):

- non-sharp nulls;
- allow for slowly diverging number K.

SIMPLE-RC for a Group of Nodes: Null

Theorem (SIMPLE-RC for a group: Null)

Suppose $m \to \infty$. *If* $c_{1n} \ll [d_1 \lambda_1(\mathbf{P})]^{-\frac{1}{2}} (\log n)^{-\frac{1}{2}}$, *then the SIMPLE-RC test statistic* $T = \max_{\{i,j\} \in \mathscr{P}} T_{ij}$ *satisfies that under null hypothesis* H_0,

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

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.*

SIMPLE-RC for a Group of Nodes: Null

Theorem (SIMPLE-RC for a group: Null)

Suppose $m \to \infty$. If $c_{1n} \ll [d_1 \lambda_1(\mathbf{P})]^{-\frac{1}{2}} (\log n)^{-\frac{1}{2}}$, then the SIMPLE-RC test statistic $T = \max_{\{i,j\} \in \mathscr{P}} T_{ij}$ satisfies that under null hypothesis H_0,

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

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.

- Individual test statistics T_{ij} based on random coupling are asymptotically independent. (So when m is bounded, asymptotic distribution of T becomes maximum of $m/2$ independent $\chi^2_{K_0}$ under H_0.)
- When $m \to \infty$, the maximum of $m/2$ "almost independent" random variables with exponential tail leads to the Gumbel distribution.

SIMPLE-RC for a Group of Nodes: Power

Theorem (SIMPLE-RC for a group: power)

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

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

The key observation is that with high probability (as $m \to \infty$),

$$\max_{\{i,j\}\in\mathscr{P}} \left\|\mathbf{D}_{K_0}\left[\mathbf{V}_{K_0}(i) - \mathbf{V}_{K_0}(j)\right]\right\| \geq \frac{1}{3}\max_{\{i,j\}\subset\mathscr{M}} \left\|\mathbf{D}_{K_0}\left[\mathbf{V}_{K_0}(i) - \mathbf{V}_{K_0}(j)\right]\right\|.$$

SIMPLE-RC for a Group of Nodes: Power

Theorem (SIMPLE-RC for a group: power)

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

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

The key observation is that with high probability (as $m \to \infty$),

$$\max_{\{i,j\}\in\mathscr{P}} \left\|\mathbf{D}_{K_0}\left[\mathbf{V}_{K_0}(i) - \mathbf{V}_{K_0}(j)\right]\right\| \ge \frac{1}{3}\max_{\{i,j\}\subset\mathscr{M}} \left\|\mathbf{D}_{K_0}\left[\mathbf{V}_{K_0}(i) - \mathbf{V}_{K_0}(j)\right]\right\|.$$

Given a set of points $\{x_i : 1 \le i \le m\}$ with metric d. Let $\ell = d(x_{i_0}, x_{j_0})$ be the maximum distance between pairs of points.

$$A = \{x_i : d(x_i, x_{i_0}) \le \ell/3\}, \quad B = \{x_i : d(x_j, x_{j_0}) \le \ell/3\}.$$

Consider the two cases: (1) $A = o(m)$ or $B = o(m)$; (2) $A \ge cm$, $B \ge cm$.

Empirical Versions of SIMPLE-RC

Need to provide an estimate of covariance matrix $\boldsymbol{\Sigma}_{i,j}$ and specify the choice of K_0.

- Suggest consistent estimator $\widehat{\boldsymbol{\Sigma}}_{i,j}(K_0)$ of covariance matrix $\boldsymbol{\Sigma}_{i,j}(K_0)$ 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 $K_0 < k \le K$.)
- For K_0, suggest a simple thresholding estimator

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

  with $\check{q}^2 := \max_{j\in[n]} \sum_{l=1}^{n} X_{lj}$ maximum node degree ($\check{q} \sim q$), $C_n \to \infty$ a deterministic parameter (e.g., $C_n = \log\log n$).
- *Consistency of covariance matrix* and corresponding *asymptotic null distributions* and *power analysis* for *SIMPLE-RC test with estimates $\widehat{K}_0$ and $\widehat{\boldsymbol{\Sigma}}_{i,j}(\widehat{K}_0)$* are rigorously established (Fan, Fan, Lv, and Y., 2022).

Empirical Versions of SIMPLE-RC

Need to provide an estimate of covariance matrix $\boldsymbol{\Sigma}_{i,j}$ and specify the choice of K_0.

- Suggest consistent estimator $\widehat{\boldsymbol{\Sigma}}_{i,j}(K_0)$ of covariance matrix $\boldsymbol{\Sigma}_{i,j}(K_0)$ 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 $K_0 < k \leq K$.)
- For K_0, suggest a simple thresholding estimator

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

  with $\check{q}^2 := \max_{j\in[n]} \sum_{l=1}^{n} X_{lj}$ maximum node degree ($\check{q} \sim q$), $C_n \to \infty$ a deterministic parameter (e.g., $C_n = \log\log n$).
- *Consistency of covariance matrix* and corresponding *asymptotic null distributions* and *power analysis* for *SIMPLE-RC test with estimates $\widehat{K}_0$ and $\widehat{\boldsymbol{\Sigma}}_{i,j}(\widehat{K}_0)$* are rigorously established (Fan, Fan, Lv, and Y., 2022).

Asymptotic null distributions and *power analysis* for *SIMPLE-RC test statistics $\mathscr{T}_{ij}$ and $\mathscr{T}$ with degree heterogeneity* are formally justified (Fan, Fan, Lv and Yang, 2022).

We *lose one degree of freedom* in asymptotic null distributions due to the use of eigenvector ratio statistics

A RMT Framework

- Our technical analyses empowered by *novel asymptotic expansions of spiked eigenvector entries* for large random matrices with *weak spikes*:

$$\widehat{v}_k(i) = v_k(i) + \frac{1}{d_k}(\mathbf{W}\mathbf{v}_k)_i + \text{error}.$$

A RMT Framework

- Our technical analyses empowered by *novel asymptotic expansions of spiked eigenvector entries* for large random matrices with *weak spikes*:

$$\widehat{v}_k(i) = v_k(i) + \frac{1}{d_k}(\mathbf{W}\mathbf{v}_k)_i + \text{error}.$$

- Exploit the Cauchy integral formula to extract the information of eigenvectors:

$$\mathbf{x}^T \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^* \mathbf{y} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathscr{C}_k} \mathbf{x}^T (\mathbf{X} - z)^{-1} \mathbf{y} dz, \quad \mathscr{C}_k \text{ encloses } \widehat{d}_k \text{ only}.$$

A RMT Framework

- Our technical analyses empowered by *novel asymptotic expansions of spiked eigenvector entries* for large random matrices with *weak spikes*:

$$\widehat{v}_k(i) = v_k(i) + \frac{1}{d_k}(\mathbf{W}\mathbf{v}_k)_i + \text{error}.$$

- Exploit the Cauchy integral formula to extract the information of eigenvectors:

$$\mathbf{x}^T \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^* \mathbf{y} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathscr{C}_k} \mathbf{x}^T (\mathbf{X} - z)^{-1} \mathbf{y} dz, \quad \mathscr{C}_k \text{ encloses } \widehat{d}_k \text{ only}.$$

- Reduce to the study of the *Green's function* $\mathbf{G}(z) = (\mathbf{W} - z)^{-1}$ of the *noise matrix* $\mathbf{W}$. Need to characterize *asymptotic behavior of* $\mathbf{x}^T \mathbf{G}(z) \mathbf{y}$ for any deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (*convergence to a deterministic limit named anisotropic local law*).

Key challenge is to derive a *sharper anisotropic local law* for $\mathbf{G}(z)$ under *weaker* conditions on sparsity level and signal strength.

A RMT Framework

• Our technical analyses empowered by *novel asymptotic expansions of spiked eigenvector entries* for large random matrices with *weak spikes*:

$$\widehat{v}_k(i) = v_k(i) + \frac{1}{d_k}(\mathbf{W}\mathbf{v}_k)_i + \text{error}.$$

• Exploit the Cauchy integral formula to extract the information of eigenvectors:

$$\mathbf{x}^T \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^* \mathbf{y} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathscr{C}_k} \mathbf{x}^T (\mathbf{X} - z)^{-1} \mathbf{y} dz, \quad \mathscr{C}_k \text{ encloses } \widehat{d}_k \text{ only}.$$

• Reduce to the study of the *Green's function* $\mathbf{G}(z) = (\mathbf{W} - z)^{-1}$ of the *noise matrix* $\mathbf{W}$. Need to characterize *asymptotic behavior of* $\mathbf{x}^T \mathbf{G}(z) \mathbf{y}$ for any deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (*convergence to a deterministic limit named anisotropic local law*).

Key challenge is to derive a *sharper anisotropic local law* for $\mathbf{G}(z)$ under *weaker* conditions on sparsity level and signal strength.

• Anisotropic local laws enable us to derive *precise asymptotic expansions* of *spiked eigenvectors* that hold uniformly for all entries with high probability.

A RMT Framework

• Our technical analyses empowered by *novel asymptotic expansions of spiked eigenvector entries* for large random matrices with *weak spikes*:

$$\widehat{v}_k(i) = v_k(i) + \frac{1}{d_k}(\mathbf{W}\mathbf{v}_k)_i + \text{error}.$$

• Exploit the Cauchy integral formula to extract the information of eigenvectors:

$$\mathbf{x}^T \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^* \mathbf{y} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathscr{C}_k} \mathbf{x}^T (\mathbf{X} - z)^{-1} \mathbf{y} dz, \quad \mathscr{C}_k \text{ encloses } \widehat{d}_k \text{ only}.$$

• Reduce to the study of the *Green's function* $\mathbf{G}(z) = (\mathbf{W} - z)^{-1}$ of the *noise matrix* $\mathbf{W}$. Need to characterize *asymptotic behavior of* $\mathbf{x}^T \mathbf{G}(z) \mathbf{y}$ for any deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (*convergence to a deterministic limit named anisotropic local law*).

Key challenge is to derive a *sharper anisotropic local law* for $\mathbf{G}(z)$ under *weaker* conditions on sparsity level and signal strength.

• Anisotropic local laws enable us to derive *precise asymptotic expansions* of *spiked eigenvectors* that hold uniformly for all entries with high probability.

The *uniform results* on *asymptotic distributions* of empirical spiked eigenvectors are key to *random coupling* for *group network inference*.

A RMT Framework

- Our technical analyses empowered by *novel asymptotic expansions of spiked eigenvector entries* for large random matrices with *weak spikes*:

$$\widehat{v}_k(i) = v_k(i) + \frac{1}{d_k}(\mathbf{W}\mathbf{v}_k)_i + \text{error}.$$

- Exploit the Cauchy integral formula to extract the information of eigenvectors:

$$\mathbf{x}^T \widehat{\mathbf{v}}_k \widehat{\mathbf{v}}_k^* \mathbf{y} = \frac{1}{2\pi \mathrm{i}} \oint_{\mathscr{C}_k} \mathbf{x}^T (\mathbf{X} - z)^{-1} \mathbf{y} dz, \quad \mathscr{C}_k \text{ encloses } \widehat{d}_k \text{ only}.$$

- Reduce to the study of the *Green's function* $\mathbf{G}(z) = (\mathbf{W} - z)^{-1}$ of the *noise matrix* $\mathbf{W}$. Need to characterize *asymptotic behavior of* $\mathbf{x}^T \mathbf{G}(z) \mathbf{y}$ for any deterministic vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (*convergence to a deterministic limit named anisotropic local law*).

Key challenge is to derive a *sharper anisotropic local law* for $\mathbf{G}(z)$ under *weaker* conditions on sparsity level and signal strength.

- Anisotropic local laws enable us to derive *precise asymptotic expansions* of *spiked eigenvectors* that hold uniformly for all entries with high probability.

The *uniform results* on *asymptotic distributions* of empirical spiked eigenvectors are key to *random coupling* for *group network inference*.

- *More comprehensive theory* (Fan, Fan, Lv and Y., 2022).

Conclusions

Reference: Fan, J., Fan, Y., Lv, J. and Yang, F. (2022+). SIMPLE-RC: group network inference with non-sharp nulls and weak signals. *arXiv:2211.00128*.

- 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 weaker spikes*.
- Revealed an interesting phenomenon of *eigen-selection* for *valid network inference*.

Conclusions

Reference: Fan, J., Fan, Y., Lv, J. and Yang, F. (2022+). SIMPLE-RC: group network inference with non-sharp nulls and weak signals. *arXiv:2211.00128*.

- 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 weaker spikes*.
- Revealed an interesting phenomenon of *eigen-selection* for *valid network inference*.

Thank you!