

SUPPLEMENT TO ‘ESTIMATION OF NUMBER OF COMMUNITIES IN ASSORTATIVE SPARSE NETWORKS’

Anonymous authors

Paper under double-blind review

1 APPENDIX

1.1 PRIOR RESULTS

Le et al. (2017) proved concentration results for regularized adjacency matrices generated by SBM and their symmetric normalized Laplacians in sparse regimes with $\tilde{d} = o(\log(N))$. Our main result relies on several of the results therein, so we recite them below. The first result relates the L_2 norm of a matrix to the L_1 norms of its rows and columns.

Lemma 1.1. *Consider a matrix \mathbf{B} in which each row has L_1 norm at most a , and each column has L_1 norm at most b . Then $\|\mathbf{B}\| \leq \sqrt{ab}$.*

Proof. See Lemma 2.7 in Le et al. (2017). ■

The next result shows that the number of high-degree nodes in \mathbf{A} is fixed.

Lemma 1.2. *Let $1 \leq m \leq N$ and $\alpha \geq \sqrt{m/N}$. Then for $r \geq 1$ the following holds with probability at least $1 - N^{-r}$. Consider a block $I \times J$ of size $m \times m$. Then all but $m/\alpha d$ rows of $\mathbf{A}_{I \times J}$ have at most $8r\alpha d$ ones.*

Proof. See Lemma 3.5 in Le et al. (2017). ■

The following is a main result from Le et al. (2017) and establishes that regularized adjacency matrices concentrate around their mean.

Theorem 1.3. *Consider a random graph from the inhomogeneous Erdős-Rényi model $G(N, (p_{ij}))$, and let $d = \max_{ij} Np_{ij}$. For any $r \geq 1$, the following holds with probability at least $1 - N^{-r}$. Consider any subset consisting of at most $10N/d$ vertices, and reduce the weights of the edges incident to those vertices in an arbitrary way. Let d' be the maximal degree of the resulting graph. Then the adjacency matrix \mathbf{A}' of the new (weighted) graph satisfies*

$$\|\mathbf{A}' - \mathbb{E}(\mathbf{A})\| \leq Cr^{3/2}(\sqrt{d} + \sqrt{d'}).$$

Moreover, the same bound holds for d' being the maximal L_2 norm of the rows of \mathbf{A}' .

Proof. See Theorem 2.1 in Le et al. (2017). ■

Next, we recite below a well-known result from linear algebra.

Theorem 1.4 (Sylvester’s law of inertia). *Let $\mathbf{A}, \mathbf{B} \in M_n$ (the set of all $n \times n$ complex matrices) be Hermitian. There is a nonsingular matrix $\mathbf{S} \in M_n$ such that $\mathbf{A} = \mathbf{SBS}^*$ if and only if \mathbf{A} and \mathbf{B} have the same inertia, that is, they have the same number of positive, negative, and zero eigenvalues.*

Proof. See Theorem 4.5.8 in Horn & Johnson (2012). ■

1.2 STATEMENTS AND PROOFS OF THE RESULTS IN SECTION 3

Lemma 1.5. Let $\mathbf{L}_\zeta := \frac{1}{\zeta}\mathbf{H}_\zeta = \tilde{\mathbf{D}}_\zeta - \mathbf{A}$, where $\tilde{\mathbf{D}}_\zeta = (\zeta - \frac{1}{\zeta})\mathbf{I}_N + \frac{1}{\zeta}\mathbf{D}$ and $\zeta > 1$ be the Laplacian form corresponding to the Bethe-Hessian matrix \mathbf{H}_ζ . Define the symmetric normalized Laplacian $\mathcal{L}(\mathbf{L}_\zeta) := \tilde{\mathbf{D}}_\zeta^{-1/2}\mathbf{L}_\zeta\tilde{\mathbf{D}}_\zeta^{-1/2}$. Then, \mathbf{H}_ζ and $\mathcal{L}(\mathbf{L}_\zeta)$ have the same number of negative eigenvalues.

Proof of Lemma 1.5. Writing $\mathcal{L}(\mathbf{L}_\zeta) := \tilde{\mathbf{D}}_\zeta^{-1/2}[\frac{1}{\zeta}\mathbf{H}_\zeta]\tilde{\mathbf{D}}_\zeta^{-1/2}$, note that $\frac{1}{\zeta}\mathbf{H}_\zeta$ is symmetric and $\tilde{\mathbf{D}}_\zeta^{-1/2}$ is non-singular. Then, by Theorem 1.4, the desired result follows. ■

Lemma 1.6. Let $\mathcal{L}(\mathbf{L}_\zeta)$ be defined as in Lemma 1.5. Analogously, define $\mathcal{L}(\bar{\mathbf{L}}_\zeta) := \tilde{\mathbf{D}}_\zeta^{-1/2}\bar{\mathbf{L}}_\zeta\tilde{\mathbf{D}}_\zeta^{-1/2}$, where $\bar{\mathbf{L}}_\zeta = \tilde{\mathbf{D}}_\zeta - \bar{\mathbf{A}}$ and $\tilde{\mathbf{D}}_\zeta = (\zeta - \frac{1}{\zeta})\mathbf{I}_N + \frac{1}{\zeta}\bar{\mathbf{D}}$. Then, there is a constant C such that for any $r \geq 1$ and $\zeta \in \omega(\sqrt{d})$,

$$\|\mathcal{L}(\mathbf{L}_\zeta) - \mathcal{L}(\bar{\mathbf{L}}_\zeta)\| \leq \frac{Cr^2\zeta d^{3/2}}{(\zeta^2 - 1)^2} \left(1 + \frac{d}{\zeta^2 - 1}\right)$$

with probability at least $1 - e^{-r}$.

Proof of Lemma 1.6. We proceed in a manner similar to that outlined in the proof of Theorem 4.1 in Le et al. (2017). First, we decompose the deviation into two parts.

$$\begin{aligned} \mathcal{L}(\mathbf{L}_\zeta) - \mathcal{L}(\bar{\mathbf{L}}_\zeta) &= \tilde{\mathbf{D}}_\zeta^{-1/2}\bar{\mathbf{A}}\tilde{\mathbf{D}}_\zeta^{-1/2} - \tilde{\mathbf{D}}_\zeta^{-1/2}\mathbf{A}\tilde{\mathbf{D}}_\zeta^{-1/2} \\ &= \underbrace{\tilde{\mathbf{D}}_\zeta^{-1/2}(\bar{\mathbf{A}} - \mathbf{A})\tilde{\mathbf{D}}_\zeta^{-1/2}}_{\Phi} + \underbrace{\tilde{\mathbf{D}}_\zeta^{-1/2}\bar{\mathbf{A}}\tilde{\mathbf{D}}_\zeta^{-1/2} - \tilde{\mathbf{D}}_\zeta^{-1/2}\mathbf{A}\tilde{\mathbf{D}}_\zeta^{-1/2}}_{\Psi} \end{aligned}$$

Next, we compute the upper-bounds for Φ and Ψ .

Upper-bound for Φ

We use $\tilde{\mathbf{D}}_\zeta^{-1/2}$ to regularize \mathbf{A} and use the concentration results in Le et al. (2017) to compute the upper-bound. Define the diagonal matrix Δ as follows:

$$\Delta_{ii} := \begin{cases} 1 & \text{if } d_i \leq 8rd \\ \frac{d_i}{\zeta(\zeta - \frac{1}{\zeta})} + 1 & \text{otherwise} \end{cases}$$

With this notation, since each entry in $(\zeta - \frac{1}{\zeta})\Delta$ is upper-bounded by the corresponding entry in $\tilde{\mathbf{D}}_\zeta$, the bound for Φ can be decomposed into two parts as follows.

$$\left(\zeta - \frac{1}{\zeta}\right) \|\Phi\| \leq \underbrace{\left\|\bar{\mathbf{A}} - \Delta^{-1/2}\mathbf{A}\Delta^{-1/2}\right\|}_{R_1} + \underbrace{\left\|\Delta^{-1/2}\bar{\mathbf{A}}\Delta^{-1/2} - \bar{\mathbf{A}}\right\|}_{R_2}$$

For R_1 , note that $\Delta^{-1/2}$ reduces the weights of degrees of \mathbf{A} greater than $8rd$. Denoting $\mathbf{A}' := \Delta^{-1/2}\mathbf{A}\Delta^{-1/2}$, the maximal squared L_2 norm of its i -th row is given by

$$\|\mathbf{A}'_{i \cdot}\|_2^2 \leq \sum_{j=1}^N \frac{\mathbf{A}_{ij}^2}{\Delta_{ii}\Delta_{jj}} \leq \frac{d_i}{\Delta_{ii}} \leq \max\{8rd, \zeta^2 - 1\}$$

Hence, we can invoke Theorem 2.1 in Le et al. (2017) and obtain with probability $1 - N^{-r}$ the following upper-bound for R_1 :

$$R_1 = \left\|\bar{\mathbf{A}} - \Delta^{-1/2}\mathbf{A}\Delta^{-1/2}\right\| \leq C_1 r^2 \left(\sqrt{d} + (\zeta^2 - 1)^{1/4}\right)$$

For R_2 , denoting $I := \{i | d_i \leq 8rd\}$ to be the set of entries in $\Delta^{-1/2}\bar{\mathbf{A}}\Delta^{-1/2}$ that coincide with the corresponding entries in $\bar{\mathbf{A}}$, Lemma 3.5 in Le et al. (2017) guarantees that $|I^c| \leq N/d$ with

probability $1 - N^{-r}$. The entry-wise deviation where they do not coincide is bounded by the entries of $\bar{\mathbf{A}}$, i.e., d/N , and thus the L_1 norm of the row of $\bar{\mathbf{A}}_{I^c \times [N]}$ is $d/N \cdot N = d$ and that for the column is $d/N \cdot N/d = 1$. Similarly, the L_1 norm of the row and column of $\bar{\mathbf{A}}_{[N] \times I^c}$ is 1 and d , respectively. By Lemma 2.7 from Le et al. (2017), we have

$$R_2 \leq 2\sqrt{d}$$

Now, combining the two bounds together allows us to bound Φ as follows

$$\left\| \tilde{\mathbf{D}}_\zeta^{-1/2} (\bar{\mathbf{A}} - \mathbf{A}) \tilde{\mathbf{D}}_\zeta^{-1/2} \right\| \leq \frac{C_2 r^2}{\zeta - \frac{1}{\zeta}} (\sqrt{d} + (\zeta^2 - 1)^{1/4})$$

with probability at least $1 - 2N^{-r}$.

Upper-bound for Ψ

Using the Frobenius norm to bound the spectral norm, we have

$$\frac{1}{\zeta^2} \|\Psi\|^2 \leq \frac{1}{\zeta^2} \|\Psi\|_F^2 = \frac{1}{\zeta^2} \sum_{i,j=1}^N \Psi_{ij}^2$$

where

$$\begin{aligned} \frac{1}{\zeta} \Psi_{ij} &:= \bar{\mathbf{A}}_{ij} [1/\sqrt{\bar{\delta}_{ij}} - 1/\sqrt{\delta_{ij}}] \\ \bar{\delta}_{ij} &= (\bar{d}_i + \zeta^2 - 1)(\bar{d}_j + \zeta^2 - 1) \\ \delta_{ij} &= (d_i + \zeta^2 - 1)(d_j + \zeta^2 - 1) \end{aligned}$$

Note that $\bar{\mathbf{A}}_{ij} \leq \frac{d}{N}$ and

$$|1/\sqrt{\bar{\delta}_{ij}} - 1/\sqrt{\delta_{ij}}| \leq \frac{|\bar{\delta}_{ij} - \delta_{ij}|}{2(\zeta^2 - 1)^3}$$

where

$$\begin{aligned} \bar{\delta}_{ij} - \delta_{ij} &= (d_i + \zeta^2 - 1)(d_j + \zeta^2 - 1) - (\bar{d}_i + \zeta^2 - 1)(\bar{d}_j + \zeta^2 - 1) \\ &\quad + (d_i + \zeta^2 - 1)(\bar{d}_j + \zeta^2 - 1) - (d_i + \zeta^2 - 1)(\bar{d}_j + \zeta^2 - 1) \\ &= (d_i + \zeta^2 - 1)(d_j - \bar{d}_j) + (\bar{d}_j + \zeta^2 - 1)(d_i - \bar{d}_i) \end{aligned}$$

Hence, we have

$$\frac{1}{\zeta^2} \|\Psi\|^2 \leq \frac{d^2}{4N^2(\zeta^2 - 1)^6} \tag{1.1}$$

$$\left[\sum_{i=1}^N (d_i + \zeta^2 - 1)^2 \sum_{j=1}^N (d_j - \bar{d}_j)^2 + N(d + \zeta^2 - 1)^2 \sum_{i=1}^N (d_i - \bar{d}_i)^2 \right] \tag{1.2}$$

Note that since $\text{Var}(d_i) \leq d$ for all $i \in [N]$, $\mathbb{E} \sum_{i=1}^N (d_i - \bar{d}_i)^2 \leq Nd$. Furthermore, denoting $X_i := (d_i - \bar{d}_i)$, by Bernstein's inequality, $\mathbb{P}\{X_i > Ct\sqrt{d}\} \leq e^{-t}$ for all $t \geq 1$. Note that the function $\psi_{1/2} : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$ defined $\psi_{1/2}(x) := \exp(x)^{1/2} - 1$ is convex, increasing, and satisfies $\psi_{1/2}(0) = 0$ and $\psi_{1/2}(x) \rightarrow \infty, x \rightarrow \infty$. Hence, $\psi_{1/2}$ is an Orlicz function and using Bernstein's inequality, we can define the Orlicz norm of the random variable X_i^2 as

$$\|X_i^2\|_{\psi_{1/2}} := \inf\{s > 0 : \mathbb{E} \exp(|X_i|/\sqrt{s}) \leq 1\} \leq C_2 d$$

Therefore, by triangle inequality, $\left\| \sum_{i=1}^N X_i^2 \right\|_{\psi_{1/2}} \leq C_2 Nd$ and this with Markov inequality implies

$$\sum_{i=1}^N (d_i - \bar{d}_i)^2 \leq C_2 r^2 Nd \tag{1.3}$$

with probability $1 - e^{-2r}$.

Next, (1.3) implies

$$\begin{aligned} \sum_{i=1}^N (d_i + \zeta^2 - 1)^2 &\leq 2 \sum_{i=1}^N (d_i - \bar{d}_i)^2 + 2 \sum_{i=1}^N (\bar{d}_i + \zeta^2 - 1)^2 \\ &\leq 2C_2 r^2 N d + 2N(d + \zeta^2 - 1)^2 \\ &\leq C_3 r^2 N(d + \zeta^2 - 1)^2 \end{aligned}$$

Plugging this into (1.1) yields

$$\begin{aligned} \|\Psi\|^2 &\leq \frac{\zeta^2 d^2}{4N^2(\zeta^2 - 1)^6} \left[\sum_{i=1}^N (d_i + \zeta^2 - 1)^2 \sum_{j=1}^N (d_j - \bar{d}_j)^2 + N(d + \zeta^2 - 1)^2 \sum_{i=1}^N (d_i - \bar{d}_i)^2 \right] \\ &\leq \frac{\zeta^2 d^2}{4N^2(\zeta^2 - 1)^6} (C_3 r^2 N d) (C_3 r^2 N(d + \zeta^2 - 1)^2 + N(d + \zeta^2 - 1)^2) \\ &\leq \frac{C_4 r^4 d^3 \zeta^2 (d + \zeta^2 - 1)^2}{(\zeta^2 - 1)^6} \leq \frac{C_4 r^4 \zeta^2}{\zeta^2 - 1} \left(\frac{d}{\zeta^2 - 1} \right)^3 \left(1 + \frac{d}{\zeta^2 - 1} \right)^2 \end{aligned}$$

which approaches 0 for $\zeta \in \omega(\sqrt{d})$.

Now, combining this with the bound for $\|\Phi\|$ above and its probability gives the desired result. ■

Proof of Theorem 3.1. Recall that $-\bar{\mathbf{L}}_\zeta := \bar{\mathbf{A}} - \tilde{\tilde{\mathbf{D}}}_\zeta$, where $\tilde{\tilde{\mathbf{D}}}_\zeta = (\zeta - \frac{1}{\zeta})\mathbf{I}_N + \frac{1}{\zeta}\bar{\mathbf{D}}$ and $\zeta > 1$, and $\mathcal{L}(\bar{\mathbf{L}}_\zeta) := \tilde{\tilde{\mathbf{D}}}_\zeta^{-1/2} \bar{\mathbf{L}}_\zeta \tilde{\tilde{\mathbf{D}}}_\zeta^{-1/2}$. Next, note that since $\tilde{\tilde{\mathbf{D}}}_\zeta^{-1/2}$ is non-singular and $\bar{\mathbf{L}}_\zeta$ is symmetric, it follows that $\mathcal{L}(\bar{\mathbf{L}}_\zeta)$ and $\bar{\mathbf{L}}_\zeta$ have the same inertia, as does $\zeta \bar{\mathbf{L}}_\zeta$ for $\zeta > 0$. Recall $\bar{\mathbf{A}} = \mathbf{Z}\mathbf{B}\mathbf{Z}^T - \text{Diag}(\mathbf{Z}\mathbf{B}\mathbf{Z}^T)$. So $-\bar{\mathbf{L}}_\zeta = \mathbf{Z}\mathbf{B}^{(t)}\mathbf{Z}^T - \mathbf{D}_1 - \mathbf{D}_2$ for diagonal matrices \mathbf{D}_1 and \mathbf{D}_2 .

Using Weyl's inequality,

$$\begin{aligned} \lambda_K(-\bar{\mathbf{L}}_\zeta) &\geq \lambda_K(\mathbf{Z}\mathbf{B}\mathbf{Z}^T) + \lambda_N(-\mathbf{D}_1 - \mathbf{D}_2) \\ &\geq d\lambda \frac{n_{\min}}{N} - \lambda_1(\mathbf{D}_1) - \lambda_1(\mathbf{D}_2) \\ &\geq d\lambda \frac{n_{\min}}{N} - \frac{d}{N} - \lambda_1(\mathbf{D}_2) > 0, \end{aligned}$$

when $\zeta + (d_{\max} - 1)/\zeta < d(\lambda N_{\min} - 1)/N$. On the other hand,

$$\begin{aligned} \lambda_{K+1}(-\bar{\mathbf{L}}_\zeta) &\leq \lambda_{K+1}(\mathbf{Z}\mathbf{B}\mathbf{Z}^T) + \lambda_1(-\mathbf{D}_1 - \mathbf{D}_2) \\ &\leq 0 - \lambda_N(\mathbf{D}_1) - \lambda_N(\mathbf{D}_2) < 0. \end{aligned}$$

So, $\bar{\mathbf{L}}_\zeta$ has exactly K negative eigenvalues.

For the probability statement, let E_r be the event of Theorem 3.3, where $r = (\zeta/\sqrt{d})^{3/2-\delta}$. It is not difficult to see that the event $\mathcal{L}(\bar{\mathbf{L}}_\zeta)$ has K negative eigenvalues holds on E_r if d is large enough. So, the bound for $\mathbb{P}(E_r)$ from Theorem 3.3 gives the probability estimate of the theorem. ■

Proof of Corollary 3.4. The desired result is immediately apparent by setting the radicand in the interval for ζ in Theorem 4.3 to positive, and rearranging the terms. ■

1.3 ADDITIONAL SIMULATION RESULTS

We use the same simulation settings introduced in the main paper. We re-state the setting here for the purpose of continuity. We simulate network data from the SBM. In the Simulation Setting (1), the probability connectivity matrix is defined as $\mathbf{B} := \rho \mathbf{B}_0 := \rho(\eta - 1)b[\mathbf{I}_K + \frac{1}{\eta-1}\mathbf{1}_K \mathbf{1}_K^T]$. ρ controls the expected degree of the network by $\tilde{d} = \rho(\mathbf{1}_N^T(\mathbf{Z}\mathbf{B}_0\mathbf{Z}^T - \text{Diag}(\mathbf{Z}\mathbf{B}_0\mathbf{Z}^T))\mathbf{1}_N)/N$. η is the in/out

ratio based on \mathbf{B} and determines the degree of assortativity. b is the baseline value in \mathbf{B} , which is set to 0.1. To generate data, we first simulate the membership vector $\mathbf{Z} \sim \text{Mult}(1; (\frac{1}{K}, \dots, \frac{1}{K}))$. We set $\tilde{d} \in \{3\sqrt{\log(N)}, 0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$ by varying ρ , to assess the performance of the algorithms under different sparsity regimes. The constants in the rates of \tilde{d} are chosen in way that \tilde{d} is same at $N = 1000$ for all the rates. With a fixed \mathbf{Z} and \mathbf{B} , and given model parameters K , N , \tilde{d} , and η , we then generate \mathbf{A} with 20 repetitions. Table 1.1 summarises the combinations of model parameter settings used in the simulations.

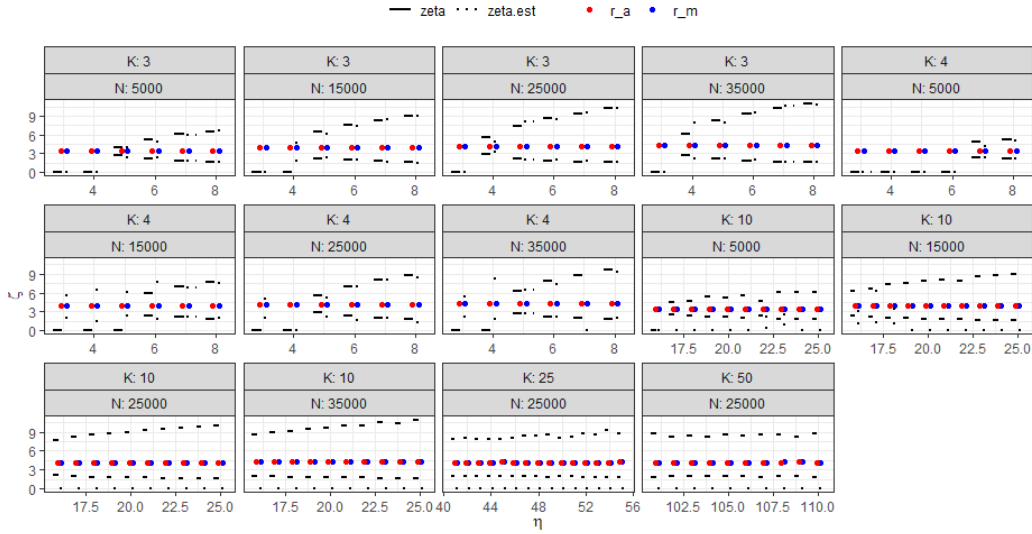
Table 1.1: Model Parameters for Simulation Setting (1)

| K | N | \tilde{d} | η |
|-----|-----------------------------|---|----------------------|
| 3 | {5000, 15000, 25000, 35000} | $\{3\sqrt{\log(N)}, 0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$ | {3, 4, ..., 8} |
| 4 | {5000, 15000, 25000, 35000} | $\{3\sqrt{\log(N)}, 0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$ | {3, 4, ..., 8} |
| 10 | {5000, 15000, 25000, 35000} | $\{3\sqrt{\log(N)}, 0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$ | {16, 17, ..., 25} |
| 25 | {25000} | $\{3\sqrt{\log(N)}, 0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$ | {41, 42, ..., 55} |
| 50 | {25000} | $\{3\sqrt{\log(N)}, 0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$ | {101, 102, ..., 110} |

In Simulation Setting (2), we use a more general probability connectivity matrix as defined in equation 1.4, where $\eta \in \{2.5 + (m - 1)0.25 : m = 1, \dots, 9\}$, and set other parameters as follows: $\tilde{d} = 3\sqrt{\log(N)}$; $K = 3$; and $N \in \{5000, 15000, 25000, 35000\}$.

$$\mathbf{B} := \rho \begin{pmatrix} 1 + \eta & 0.5 & 0.3 \\ 0.5 & 2 + \eta & 0.1 \\ 0.3 & 0.1 & 0.5 + \eta \end{pmatrix} \quad (1.4)$$

Figures 1.1, and 1.2 show oracle intervals for ζ and its estimates are shown with two popular heuristic choices for ζ (r_m and r_a) commonly used in literature and discussed in Le & Levina (2015). Network data was simulated from Simulation Setting (1) with $\tilde{d} \in \{0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$, each simulated with 20 repetitions. Intervals are shown as zeros when the threshold of detection is not met.

Figure 1.1: Oracle intervals for ζ and its estimation are shown with two heuristic choices for ζ , r_m and r_a .

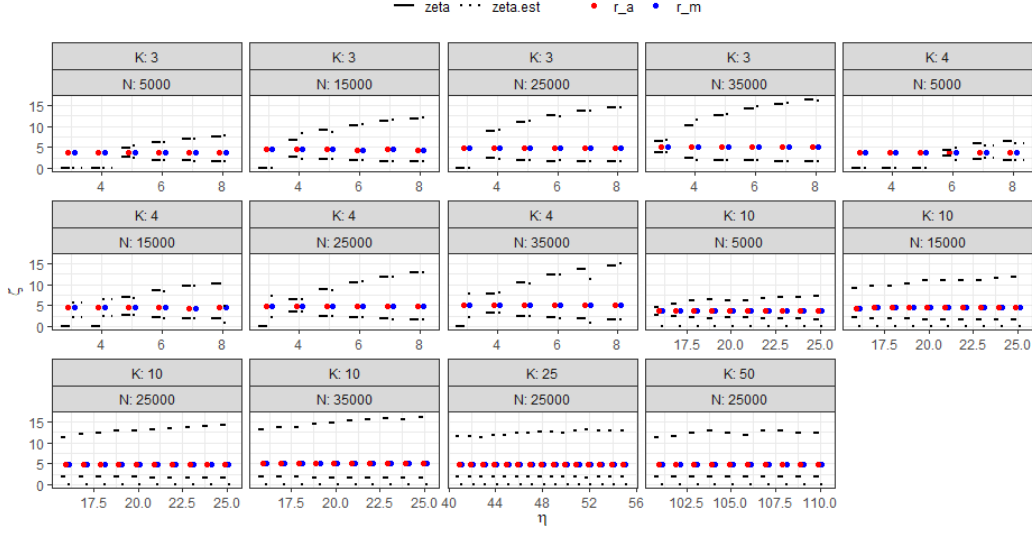


Figure 1.2: Oracle intervals for ζ and their estimates are shown with $\zeta \in \{r_m, r_a\}$.

Procedure 4.2 in the main paper outputs $\hat{\mathbf{N}}_{K_0}$ and $\hat{\mathbf{B}}_{K_0}$ with candidate number of communities $K_0 \in [1, \dots, K_{\max}]$ as an input, where K_{\max} is a tuning parameter. Then, the minimal community size is estimated with $\hat{N}_{\min} = \min\{\hat{\mathbf{N}}_2\}$. \hat{N}_{\min} is an upper bound of N_{\min} with high probability and has shown in simulations to be a good estimate of N_{\min} . Next, estimating d , d_{\max} , and λ requires K_0 as an input. We propose $\hat{K}_0 = \arg \max_{K_0} (\lambda_{K_0}(\hat{\mathbf{B}}_{K_0}))$ to recover community membership with maximum signal. Then the estimators of d , d_{\max} , and λ are $\hat{d} = N \max_{a,b} \{(\hat{\mathbf{B}}_{\hat{K}_0})_{a,b}\}$, $\hat{d}_{\max} = \max\{\hat{\mathbf{N}}_{\hat{K}_0} \hat{\mathbf{B}}_{\hat{K}_0}\}$, and $\hat{\lambda} = \lambda_{\hat{K}_0}(\hat{\mathbf{B}}_{\hat{K}_0})$ respectively. In simulations, we found \hat{d} and $\hat{\lambda}$ to be good estimates while \hat{d}_{\max} tended to overestimate. Hence, we derived another estimator $\hat{d}'_{\max} = \bar{d}$ based on our modeling assumption for \mathbf{B} (see §5 in main paper).

Figure 1.3 below shows performances of estimators for n_{\min} , λ , d , and d_{\max} discussed in §5.2 in the main manuscript. Network data is simulated from the SBM under Simulation Setting (1) with $\tilde{d} = 3\sqrt{\log(N)}$.

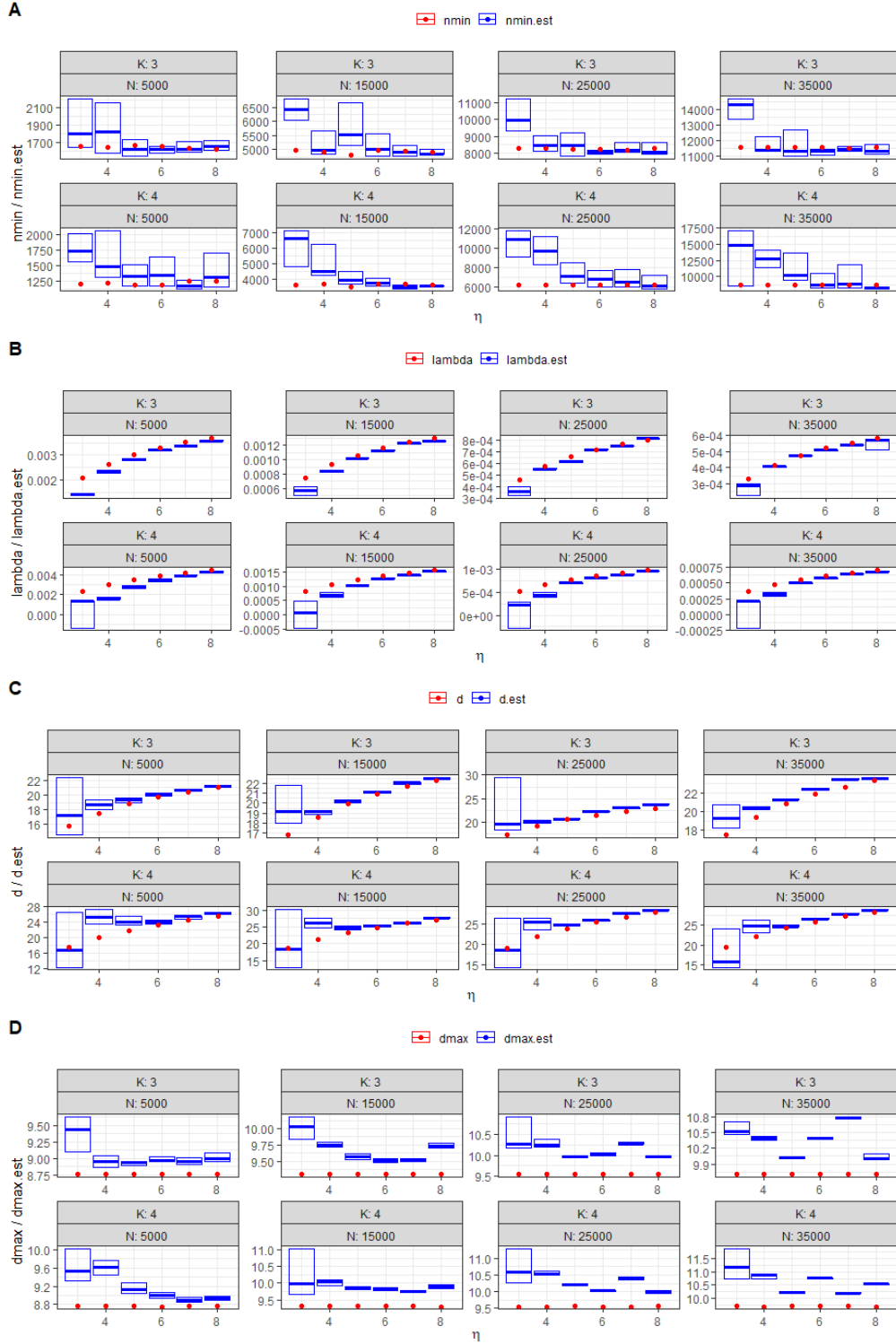


Figure 1.3: Oracle values of (A) N_{\min} , (B) λ , (C) d , and (D) d_{\max} with their estimates. Red points are the oracle values of parameters, blue boxes are estimates of the oracle parameters out of 20 repetitions, where the upper line represents the 75% quantile of those estimations, the lower line represents the 25% quantile, and the middle line represents the median.

Figure 1.4 below shows ACR of BHsparse with different quantiles of the oracle intervals with varying community assortativity levels, i.e., η , for more dense synthetic networks with $\tilde{d} \in \{0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$. Network data was generated from Simulation Setting (1). It is clear that there is a threshold-like point in η at which the algorithm’s performance changes sharply. It is shown to be an empirical property that the threshold in η decreases with increase in the expected degree (\tilde{d}).

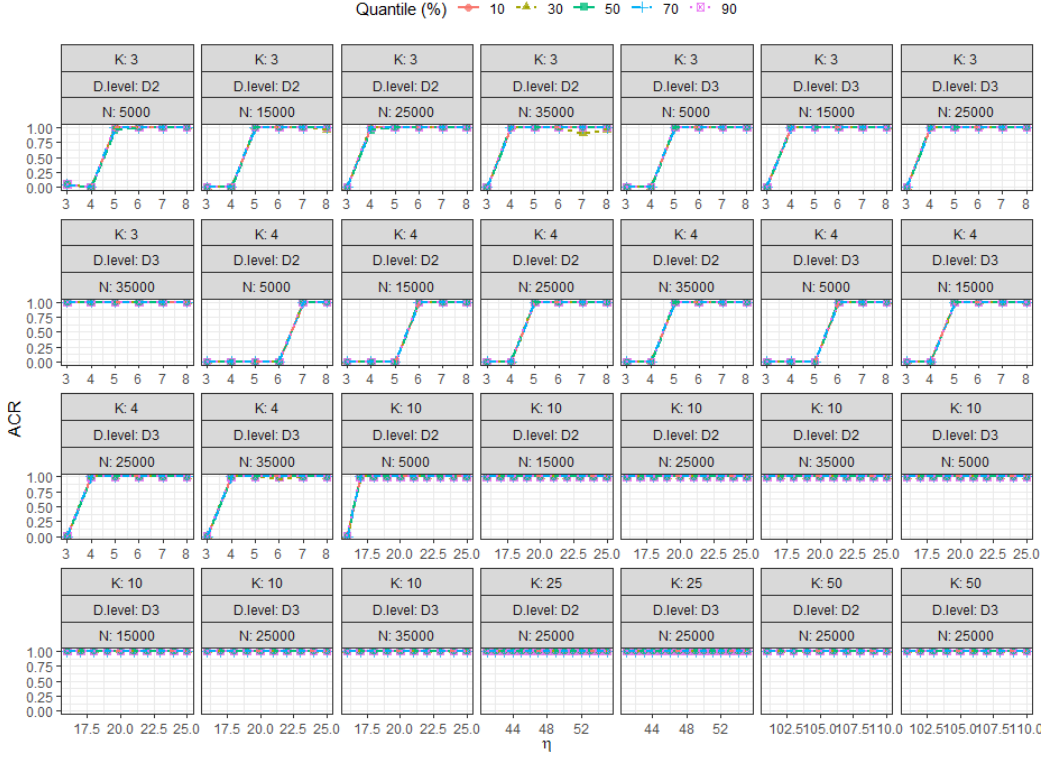


Figure 1.4: ACR of BHsparse versus η using oracle intervals with ζ set to five quantiles (10%, 30%, 50%, 70%, 90%) of the intervals. The following two settings for \tilde{d} were used: (D2) $\tilde{d} = 0.165(\log(N))^2$ and (D3) $\tilde{d} = 0.788(N)^{(1/3)}$.

Figure 1.5 below shows ACR of BHsparse with different quantiles of the intervals (oracle or estimated) for more dense synthetic networks with two choices of the expected degree $\tilde{d} \in \{0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$. Only those cases where the oracle intervals or estimated intervals exist in the plot. For the estimated choices of ζ , performances of the BHsparse algorithm are worse when ζ are close to the end-points of the intervals. Generally 30% to 50% quantiles within the intervals turn out to be the best picks for ζ .

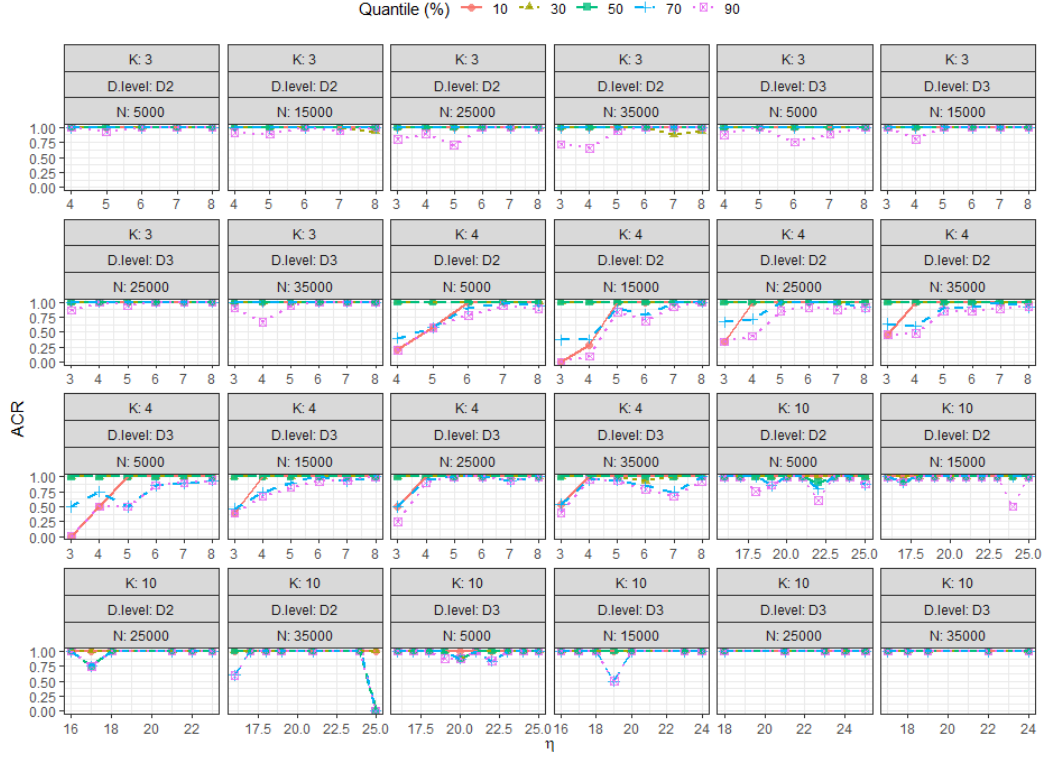


Figure 1.5: ACR of BHsparse as η varies, using the estimated interval and ζ set to quantiles (10%, 30%, 50%, 70%, 90%) of the estimated intervals using our proposed method and based on networks satisfying the threshold in Corollary 3.2. Network data was generated from Simulation Setting (1) with two levels of \tilde{d} : (D2) $0.165(\log(N))^2$ and (D3) $0.788(N)^{(1/3)}$.

Figure 1.6 (Figure 1.7 resp.) below show ACR of BHsparse with ζ set equal to 30% and 50% quantiles of the oracle intervals (estimated intervals resp.). For more dense synthetic networks, two values were used for the expected degree: $\tilde{d} \in \{0.165(\log(N))^2, 0.788(N)^{(1/3)}\}$. When the threshold in Corollary 3.2 is satisfied, these choices for ζ perform better compared to the heuristic choices.

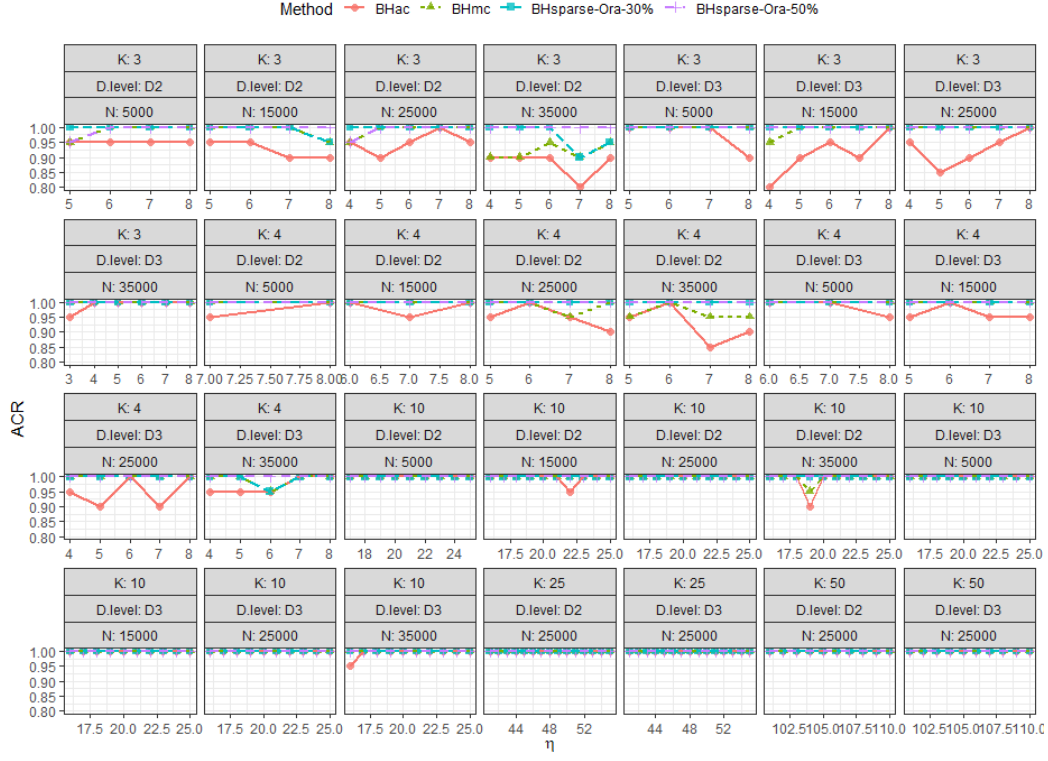


Figure 1.6: A comparison of ACRs of BHsparse, BHmc and BHac at varying levels of η . Network data was generated from Simulation Setting (1) with two levels of \tilde{d} : D2 = $0.165(\log(N))^2$, and D3 = $0.788(N)^{(1/3)}$. ζ was set to 30% and 50% quantiles of the oracle intervals. Only cases where the oracle threshold in Corollary 3.2 is satisfied are considered.

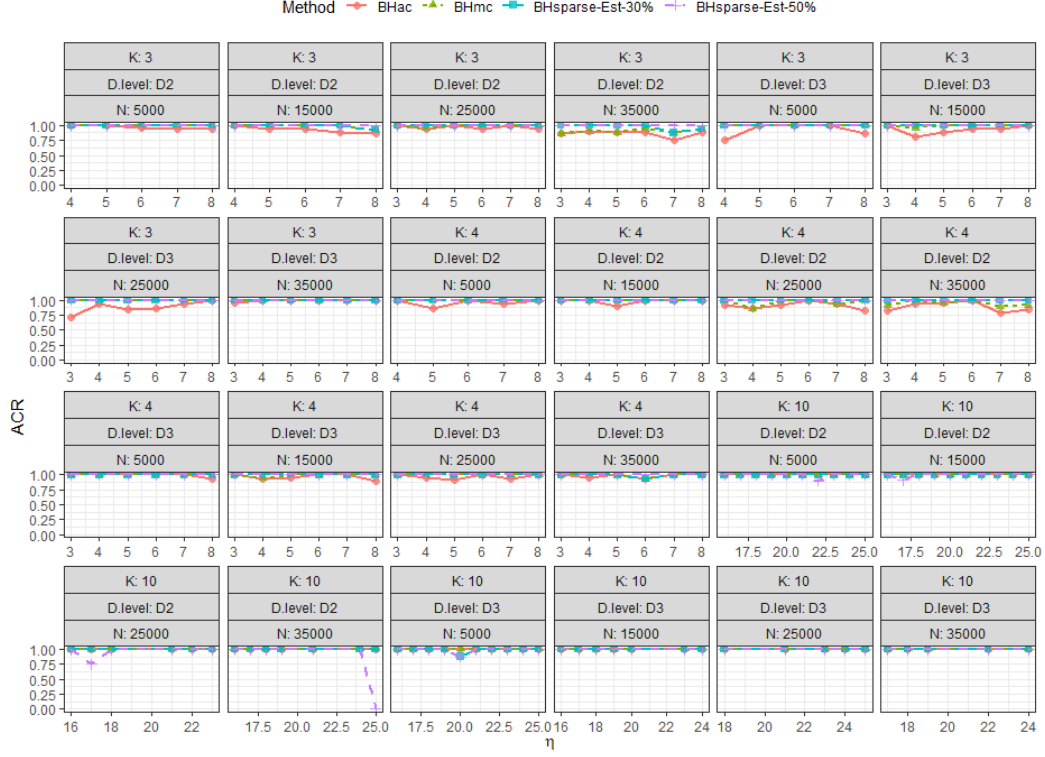


Figure 1.7: A comparison of ACRs of BHsparse, BHmc and BHac at varying levels of η . Network data was generated from Simulation Setting (1) with two levels of \tilde{d} : D2 = $0.165(\log(N))^2$, and D3 = $0.788(N)^{(1/3)}$. ζ was set to 30% and 50% quantiles of the estimated intervals. Only cases where the estimated threshold in Corollary 3.2 is satisfied are considered.

For Simulation Setting (2), Figure 1.8 below shows the performance of estimators for n_{\min} , λ , d , and d_{\max} discussed in §5.2. Figure 1.9 shows ACR of BHsparse with different quantiles of the oracle interval with varying η . It can be seen that in a more general setting of the probability connectivity structure, when K is small, the sharp change in performance of the method still exists, showing an empirical property of the detection threshold. The estimations of the interval of ζ proposed in Procedure 4.1 also perform well.

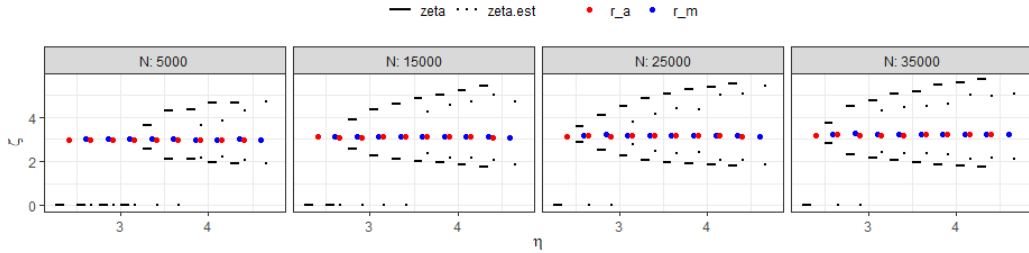


Figure 1.8: The oracle interval for ζ and its estimation are shown with the two popular heuristic choices for ζ (r_m and r_a). Network data was generated from Simulation Setting (2), where $\tilde{d} = 3\sqrt{\log(N)}$ and $K = 3$.

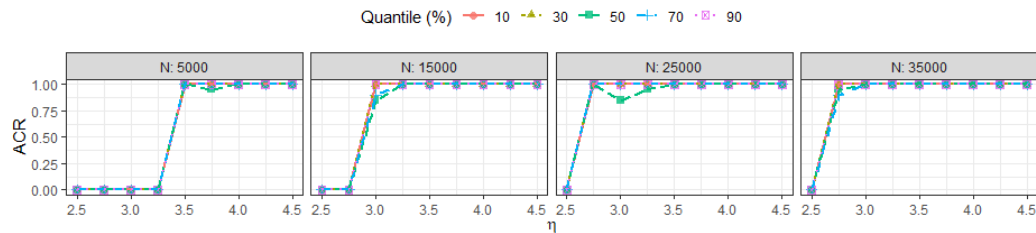


Figure 1.9: Performance of BHsparse with ζ set equal to quantiles (10%, 30%, 50%, 70%, 90%) of the oracle intervals. Network data was generated from Simulation Setting (2) with $\tilde{d} = 3\sqrt{\log(N)}$ and $K = 3$.

REFERENCES

- Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.
- Can M Le and Elizaveta Levina. Estimating the number of communities in networks by spectral methods. *arXiv preprint arXiv:1507.00827*, 2015.
- Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.