ESTIMATION OF NUMBER OF COMMUNITIES IN ASSORTATIVE SPARSE NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most community detection algorithms assume the number of communities, K, to be known *a priori*. Of the various approaches that have been proposed to estimate K, the non-parametric method based on the spectral properties of the Bethe Hessian matrix has garnered much popularity for its simplicity, computational efficiency, and robustness to sparsity of data. Recently, its consistency for networks in semi-dense regimes with the expected degree $\tilde{d} \gg \log(N)$ has been shown (N being the number of nodes in the network). In this paper, we show that the spectral method based on the Bethe Hessian matrix is in fact also consistent in sparse regimes with $\tilde{d} = o(\log(N))$, at any rate, thus establishing it as a method which is robust to a wide range of problem settings, regardless of the sparsity of networks.

1 INTRODUCTION

Statistical analysis of network data has now become an extensively studied field within statistics and machine learning (see (Goldenberg et al., 2010; Kolaczyk & Csárdi, 2014; Newman, 2018) for reviews). Network datasets show up in several disciplines. Examples include networks originating from biosciences such as gene regulation networks (Emmert-Streib et al. (2014)), protein-protein interaction networks (De Las Rivas & Fontanillo (2010)), structural (Rubinov & Sporns (2010)) and functional networks (Friston (2011)) of brain and epidemiological networks Reis et al. (2007); networks originating from social media such as Facebook, Twitter and LinkedIn (Faloutsos et al. (2010)); citation and collaboration networks (Lehmann et al. (2003)); information and technological networks such as internet-based networks (Adamic & Glance (2005)), power networks (Pagani & Aiello (2013)) and cell-tower networks (Isaacman et al. (2011)). There are several active areas of research in developing statistical methodologies for network data analysis and also deriving the theoretical properties of the statistical methods. *In this paper, we focus on networks with community structure and finding the number of communities in networks with arbitrary sparsity level.*

The last two decades saw a resurgence of interest in a problem popularly known as "community detection". A common problem definition is to partition N nodes in a graph into K communities such that there are differences in edge densities between within and between communities, where K is assumed to be known *a priori*. Estimating number of communities (K) has recently become active in the literature. While the initial focus in the literature for estimating K has been developing algorithms and drawing support from domain-specific intuition and empirical studies using the Stochastic Block Model (SBM), first proposed in Holland et al. (1983), (such as, Saade et al. (2014a), Yan et al. (2018)), there has been recent progress on attaining theoretical understanding of community numbers. Bickel & Sarkar (2015) and Lei et al. (2016) proposed hypothesis testing approaches based on principle eigenvalues or singular values. Some likelihood-based methods using the BIC criterion were proposed by Wang et al. (2017) and Hu et al. (2019). From a Bayesian perspective, Riolo et al. (2017) discussed priors for number of communities under the SBM and designed an Markov chain Monte Carlo (MCMC) algorithm, Cerqueira & Leonardi (2020) proposed an estimator based on integrated likelihood for the SBM. Chen & Lei (2018) and Li et al. (2020) proposed cross-validation based approaches with theoretical guarantees to estimate K. Yan et al. (2018) proposed a semi-definite programming approach, and Ma et al. (2018) proposed an estimator based on the loss of binary segmentation using pseudo-likelihood ratio. Both of these approaches had theoretical guarantees. *However, all the theoretical results were obtained under the assumption that mean density of the networks is greater than* $\log(N)$.

Methods based on the spectrum of a certain class of matrices have become increasingly popular in recent years as non-parametric alternatives that are more computationally efficient and applicable to a wider range of settings. Most notably the non-backtracking matrices (e.g., Krzakala et al. (2013), Saade et al. (2014b), Coste & Zhu (2019), Bordenave et al. (2015), Saade et al. (2016)) and the Bethe Hessian matrices (e.g., Saade et al. (2015b), Lelarge (2018), Dall'Amico et al. (2019), Saade et al. (2015a), Dall'Amico et al. (2020), Saade et al. (2014a), Le & Levina (2015)) have received much attention due to their non-parametric form and competitive performance in the presence of degree heterogeneity and sparsity. In particular, unlike the non-backtracking operator, the Bethe Hessian is a real symmetric operator and hence offers additional computational advantages. Le & Levina (2015) proved the consistency of the method based on the spectrum of the Bethe Hessian operator in semi-dense regimes, i.e., with the expected degree $\tilde{d} \gg \log(N)$, under certain choices for the scalar parameter for assortative and disassortative networks. *However, real-world networks are generally much more sparse and there is no theoretical result in the literature that guarantees the effectiveness of the Bethe Hessian operator in more sparse regimes*.

Our contribution. In this paper, we contribute to the theoretical understanding of the Bethe Hessian operator in estimating K for networks generated from the SBM in any regime regardless of the sparsity. We have three main contributions.

- We show that the method of estimating K based on the spectral properties of the Bethe Hessian matrix is statistically consistent, even in sparse regimes with the expected degree $\tilde{d} = o(\log(N))$.
- We provide an interval from which the scalar parameter of the Bethe Hessian operator can be chosen such that it serves as a convenient sufficient condition for correctly estimating K asymptotically in network data.
- We propose an algorithm for estimating K based on our findings and demonstrate its performance in relation to existing methods with simulations.

The paper is arranged as follows. We present the definitions in Section 2. We present main theoretical results and empirical methods in Sections 3 and 4, respectively. The simulation results and concluding remarks are given in Sections 5 and 6, respectively.

2 PRELIMINARIES

2.1 NOTATION

Adjacency matrix, denoted by $\mathbf{A}$, is a random matrix whose rows and columns are labeled by nodes $i \in [N]$, where $\mathbf{A}_{ij} = 1$ if there is an edge between nodes i and j and 0 otherwise and $[N]$ denotes the set $\{1, \ldots, N\}$. The mean observed degree is denoted by $\bar{d} := \frac{1}{N}\mathbf{1}_N^T\mathbf{A}\mathbf{1}_N$ and the expected degree by $\tilde{d} := \frac{1}{N}\mathbf{1}_N^T\mathbb{E}\mathbf{A}\mathbf{1}_N$. $\lambda_\ell^{\downarrow}(\mathbf{A})$ denotes the ℓth largest eigenvalue of $\mathbf{A}$ and $\lambda_\ell^{\uparrow}(\mathbf{A})$ denotes the ℓth smallest eigenvalue of $\mathbf{A}$.

2.2 THE STOCHASTIC BLOCK MODEL

The stochastic block model (SBM) is a simple generative model for network data that embeds a community structure in adjacency matrix, $\mathbf{A}_{N\times N}$ of the randomly generated network. SBM has three parameters: (1) the number of communities K; (2) the membership vector $\vec{z} = (z_1, ..., z_N)$ that assigns a community to each node, i.e., $z_i \in [K]$; and (3) the connectivity probability matrix $\mathbf{B}_{K\times K}$ where the element B_{ab} represents the probability of an edge between nodes belonging to community a and b ($a, b \in [K]$). *Lastly, we consider that* $\mathbf{B}$ *is of full rank, that is, rank of* $\mathbf{B}$ *is* K. *Also we consider assortative networks, which can be defined as networks generated from SBM with* $\mathbf{B}$ *having all positive eigenvalues*.

$\mathbf{Z} \in \mathbb{Z}_{>0}^{N\times K}$ is defined as the community membership matrix such that $\mathbf{Z}_{ij} = 1$ if node i belongs to community j and 0 otherwise. We denote the maximum expected degree by $d_{max} := N \max_i[(\mathbf{ZBZ}^T)_{i\cdot} - Diag(\mathbf{ZBZ}^T)_{i\cdot}]$ and the maximum entry in matrix $\mathbf{B}$ by d/N, where $d := N \max_{i,j} \mathbf{B}_{ij}$. λ denotes the smallest eigenvalue of the normalized $\mathbf{B}$ matrix, $\lambda := \lambda_K^{\downarrow}\big(\frac{N}{d}\mathbf{B}\big)$. $\bar{\mathbf{A}}$

is the expectation of $\mathbf{A}$ and is computed as $\bar{\mathbf{A}} = \mathbf{ZBZ}^T - Diag(\mathbf{ZBZ}^T)$. $\bar{\mathbf{D}}$ is a diagonal matrix whose i-th diagonal entry is the sum of the i-th row of $\bar{\mathbf{A}}$. N_{min} denotes the number of nodes in the smallest community.

2.3 THE BETHE HESSIAN MATRIX

The Bethe Hessian matrix associated with an adjacency matrix $\mathbf{A}$ is defined as

$$\mathbf{H}_\zeta := (\zeta^2 - 1)\mathbf{I}_N + \mathbf{D} - \zeta\mathbf{A} \tag{2.1}$$

where $\zeta > 1$ is a real scalar parameter, $\mathbf{D} := Diag(\mathbf{A}\mathbf{1}_N)$ is a diagonal matrix whose i-th diagonal entry corresponds to the degree of i-th node, and $\mathbf{I}_N$ is an identity matrix of dimension $N \times N$.

3 THEORETICAL RESULTS

Now, we prove our main result in several steps. The proofs for the following results and relevant results from prior works in the literature are given in the Supplementary Materials. We start with the inertia of Bethe Hessian matrices. Note that the *inertia* of a matrix is a vector consisting of the number of positive, negative, and zero eigenvalues of the matrix. The first step in our analysis is showing that the Bethe Hessian matrix and its corresponding Laplacian form have the same inertia.

Lemma 3.1. *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 $\mathbf{H}_\zeta$. 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, $\mathbf{H}_\zeta$ and $\mathcal{L}(\mathbf{L}_\zeta)$ have the same number of negative eigenvalues.*

The next step involves showing that the symmetric normalized Laplacian form, $\mathcal{L}(\mathbf{L}_\zeta)$, of the Bethe Hessian matrix concentrates around its expectation.

Lemma 3.2. *Let $\mathcal{L}(\mathbf{L}_\zeta)$ be defined as in Lemma 3.1. Analogously, define $\mathcal{L}(\bar{\mathbf{L}}_\zeta) := \tilde{\bar{\mathbf{D}}}_\zeta^{-1/2}\bar{\mathbf{L}}_\zeta\tilde{\bar{\mathbf{D}}}_\zeta^{-1/2}$, where $\bar{\mathbf{L}}_\zeta = \tilde{\bar{\mathbf{D}}}_\zeta - \bar{\mathbf{A}}$ and $\tilde{\bar{\mathbf{D}}}_\zeta = (\zeta - \frac{1}{\zeta})\mathbf{I}_N + \frac{1}{\zeta}\bar{\mathbf{D}}$. Then, there is a constant C such that for any $r \geqslant 1$ and $\zeta \in \omega(\sqrt{d})$,*

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

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

Remark (Lemma 3.2). Note that the statement of Lemma 3.2 is a finite sample result. However the finite sample result can be converted into an asymptotic result by considering appropriate relationships between d, ζ, and r. The sufficient condition for the difference between sample and expected Laplacians, $\left\|\mathcal{L}(\mathbf{L}_\zeta) - \mathcal{L}(\bar{\mathbf{L}}_\zeta)\right\|$, to tend to zero with high probability is that $\sqrt{d} \ll \zeta$ or $1 \ll r^{2/3} \ll \frac{\zeta}{\sqrt{d}}$.

The final step of our main results is establishing a sufficient condition in terms of the scalar parameter ζ in order for the expectation of the Laplacian to have exactly K negative eigenvalues. This, combined with Lemmas 3.1 and 3.2, completes the argument.

Theorem 3.3. *Let $\beta := -d(\lambda N_{min} - 1)/N$. For any $\delta \in (0, 3/2)$, $\mathcal{L}(\bar{\mathbf{L}}_\zeta)$ (and hence $\mathbf{H}_\zeta$) has exactly K negative eigenvalues for all $\zeta \in \frac{1}{2}\big(-\beta \pm \sqrt{\beta^2 + 4 - 4d_{max}}\big)$ with probability at least $1 - \exp[-(\zeta/\sqrt{d})^{3/2-\delta}]$.*

Remark (Theorem 3.3). Note that again the statement of Theorem 3.3 is a finite sample result. However the finite sample result can be converted into an asymptotic result by considering appropriate relationships between d and ζ. Under the sufficient condition of $\sqrt{d} \ll \zeta$, we get a high probability result showing consistent estimation of the number of communities by using the number of negative eigenvalues of the Bethe Hessian matrix, $\mathbf{H}_\zeta$, with ζ chosen from the interval given in Theorem 3.3.

A sufficient threshold for detecting K is presented in Corollary 3.4 below.

Corollary 3.4. *K can be detected if the following threshold is satisfied:*

$$\lambda_K^{\downarrow}(\mathbf{B})N_{min} > 2\sqrt{d_{max} - 1} \tag{3.1}$$

There has been significant progress on the related question of identifying the necessary and sufficient condition for availability of an algorithm that can estimate the community memberships reliably. When $K = 2$, community sizes are equal, and $ij = a/N$ (resp. b/N) if $i = j$ (resp. $i \neq j$), it was conjectured in Decelle et al. (2011) that the necessary and sufficient condition is $(a-b)^2 > 2(a+b)$. The proof for the necessity of the condition was given in Mossel et al. (2015) and the sufficiency of the condition was proved independently in Massoulié (2014) and Mossel et al. (2018). For estimating K of the network generated by the simple SBM considered in Decelle et al. (2011), Lemma 3.3 implies the following interval for ζ:

$$\zeta \in \frac{1}{2}\big(-\beta \pm \sqrt{\beta^2 - 2(a+b-2)}\big) \quad \text{where } \beta = \frac{a}{N} - \frac{a-b}{2} \tag{3.2}$$

For detectability of K for the SBM considered in Decelle et al. (2011), Corollary 3.4 implies the following threshold, which interestingly turns out to be quite similar to the one conjectured in Decelle et al. (2011) for a seemingly very different task of identifying a non-trivial community assignment for each node.

$$(a-b)^2 > 32(a+b) - 64 \tag{3.3}$$

For SBMs with $a + b > \frac{31}{15}$, Corollary 3.4 imposes a higher threshold for the task of estimating the number of communities compared to the community-detection threshold. Although one would expect K-estimation to be an easier task at first glance, the comparison of the thresholds of the two tasks suggests they may be of a similar degree of difficulty.

4 Empirical Methods

4.1 Estimation of number of communities with the Bethe Hessian matrix

Theorem 3.3 indicates that we can estimate the number of communities of a network by simply counting negative spectrum of the Bethe Hessian matrix with properly choosen scalar parameter ζ. Similar to the BHm and BHa method proposed by Le & Levina (2015), we propose the algorithm 4.1.

Algorithm 4.1 $\mathbf{BH_{sparse}}$ Estimation of number of communities for sparse network with the Bethe Hessian matrix

Input: Adjacency matrix $\mathbf{A}$; Scalar parameter ζ;

Output: Community size estimate $\hat{K}$;

1: $\mathbf{D} \leftarrow Diag(\mathbf{A}\mathbf{1}_N)$.

2: $\mathbf{H}_\zeta \leftarrow (\zeta^2 - 1)\mathbf{I}_N + \mathbf{D} - \zeta\mathbf{A}$.

3: Obtain $\lambda_1^{\uparrow}, ..., \lambda_N^{\uparrow}$ as sorted eigenvalues of $\mathbf{H}_\zeta$ from the smallest to the largest.

4: $\hat{K} \leftarrow max\{k : \lambda_k^{\uparrow} < 0\}$.

Compared to the choices of scalar parameter ζ in Le & Levina (2015), we choose ζ based on Theorem 3.3. By corollary 3.4, any values of scalar parameter $\zeta \in \frac{1}{2}\big(-\beta \pm \sqrt{\beta^2 + 4 - 4d_{max}}\big)$ under the assumption $\lambda_K^{\downarrow}(\mathbf{B})N_{min} > 2\sqrt{d_{max} - 1}$. In practice, oracle values of d, d_{max}, λ, and N_{min} need to be estimated in order to check the assumption as well as to calculate the bounds for ζ. Here we propose estimators for d, d_{max}, λ, and N_{min}.

4.2 Estimating the interval for Bethe Hessian scalar parameter

We observe that calculations of d, d_{max}, λ, and N_{min} all depends on knowledge of the probability connectivity matrix $\mathbf{B}$ and the community membership matrix $\mathbf{Z}$. One intuition is to estimate community memberships by regularized spectral clustering (Amini et al. (2013); Le et al. (2017)) and then calculate the maximum likelihood estimate of $\mathbf{B}$ using the estimated $\mathbf{Z}$. Consistency of these estimates have been established in prior literature (Le et al. (2017)).

Procedure 4.2 **MLESPC** Maximum likelihood estimates of SBM with spectral clustering

Input: Adjacency matrix $\mathbf{A}$; Number of communities K_0;

Output: Community size vector estimate $\hat{\mathbf{N}}_{K_0}$; Probability connectivity matrix estimate $\hat{\mathbf{B}}_{K_0}$;

1: Obtain $\hat{\mathbf{U}} \in \mathbb{R}^{N \times K_0}$ consisting of K_0 eigenvectors of $\mathbf{A}$ corresponding to the largest eigenvalues.

2: Obtain $\hat{\mathbf{Z}}$ by applying K-means clustering algorithm on column vectors of $\hat{\mathbf{U}}$.

3: $\hat{\mathbf{N}}_{K_0} \leftarrow \hat{\mathbf{Z}}\mathbf{1}_N$.

4: $\hat{\mathbf{B}}_{K_0} \leftarrow \hat{(}1/\mathbf{N}_{K_0})\mathbf{I}_{K_0}\hat{\mathbf{Z}}^T\mathbf{A}\hat{\mathbf{Z}}\mathbf{I}_{K_0}(1/\mathbf{N}_{K_0})$.

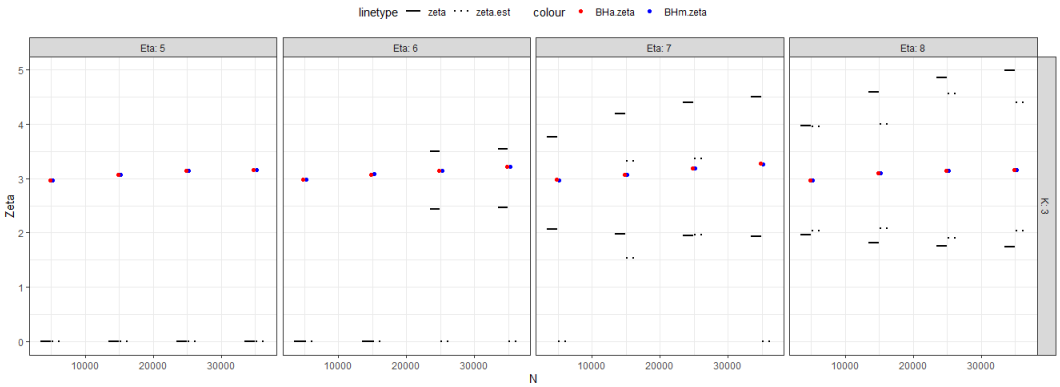

Figure 4.1: Numerical comparison of oracle and estimated intervals for ζ proposed in lemma 3.3 with choices of ζ proposed in Le & Levina (2015). Network data is simulated from SBM with $K = 3$, $N = (5000, 15000, 25000, 35000)$, $\eta = (5, 6, 7, 8)$, and $\tilde{d} = 3\sqrt{log(N)}$. For each setting there are 20 repetitions. Oracle intervals are plotted as solid lines, mediums of estimated intervals are shown as dotted lines. Intervals are set to be 0 when the assumption in corollary 3.4 is not met. Mediums of estimated values of ζ from BHm and BHa methods are shown as red and blue points.

We use procedure 4.2 to get $\hat{\mathbf{N}}_{K_0}$ and $\hat{\mathbf{B}}_{K_0}$ with candidate number of communities $K_0 \in [1, ..., K_{max}]$, where K_{max} serves as a tuning parameter in the estimation. Then we estimate the minimal community size as $\hat{N}_{min} = \min\{\hat{\mathbf{N}}_2\}$, since it acts as an upper bound of N_{min} with high probability. By simulation we found that this turns out to be a good estimate of N_{min}. Estimating d, d_{max}, and λ is slightly more complex, as it involves a prior estimation of K_0. Trying to restore the community membership with the maximum estimated community signal, we propose $\hat{K_0} = \arg\max_{K_0}(\lambda_{K_0}(\hat{\mathbf{B}}_{K_0}))$ as a prior estimate of K_0. Then the estimates of d, d_{max}, and λ are $\hat{d} = Nmax_{a,b}\{(\hat{\mathbf{B}}_{\hat{K_0}})_{a,b}\}$, $\hat{d}_{max} = max\{\hat{\mathbf{B}}_{\hat{K_0}}\hat{\mathbf{N}}_{\hat{K_0}}\}$, and $\hat{\lambda} = \lambda_{\hat{K_0}}(\hat{\mathbf{B}}_{\hat{K_0}})$ respectively. Numerically, we found $\hat{d}$ and $\hat{\lambda}$ good estimates while $\hat{d}_{max}$ tends to over estimate. Under the assumption that $\mathbf{B} = \rho(\eta - 1)b[\mathbf{I}_K + \frac{1}{\eta-1}\mathbf{1}_K\mathbf{1}_K^T]$, another estimate can be $\hat{d}'_{max} = \bar{d}$. Simulation results based on the performance of these estimators are included in the Supplementary Materials.

We evaluate performances of estimated intervals and oracle intervals of Bethe-Hessian scalar parameter numerically, along with the choices of ζ proposed in Le & Levina (2015). Under the settings of large network size (N) and assortativity condition, Figure 4.1 shows that estimated intervals turn out to capture the oracle intervals well. If the assumption in corollary 3.4 is met, choices of ζ proposed in Le & Levina (2015) (BHm and BHa) turn out to be good picks. In the simulation section (Section 5), we will show that compared to BHm and BHa, other choices of ζ within the given intervals can improve performances of the community size estimation, especially when N is large in the sparse regime.

5 SIMULATION RESULTS

In this section we evaluate the performance of algorithm 4.1 with empirical accuracy rate (ACR) by simulation. Recent literature (Le & Levina (2015), Yan et al. (2018), Cerqueira & Leonardi (2020)) show that Bethe Hessian methods (BHm, BHa, BHmc, BHac) proposed in Le & Levina (2015) are competitive in performance and it is computationally efficient in semi-dense regimes. However, sizes of synthetic networks used from those simulations are relatively small compared to real life networks. Also, the empirical comparisons had been carried in the semi-dense regime of expected degree $d \gg O(log(N))$. In our simulation, we compare performances of our method (BHsparse) with BHm, BHa, BHmc and BHac on large networks with node size up to 35000. Also, our simulation setting is based on sparse networks and we find that in the sparse regime where $d = o(log(N))$, BHsparse outperforms BHm and BHa methods, especially when node size is large and assortativity is strong. For comparison, we only use the BHm and BHa methods, as these methods have been favorably compared empirically with most of the existing methods of estimation of number of communities.

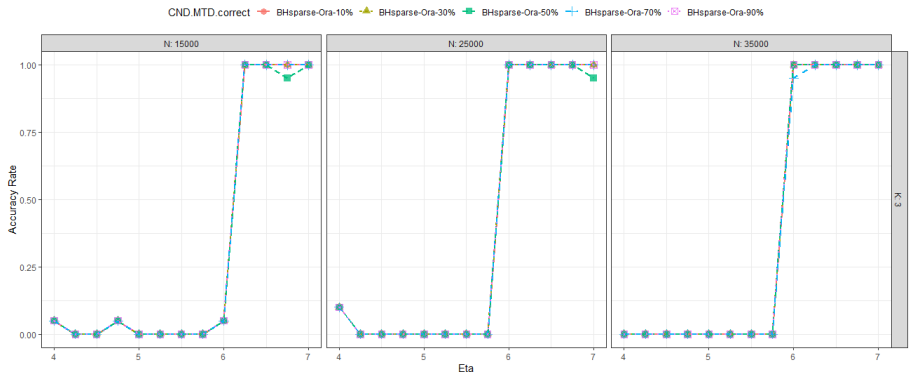

Figure 5.1: Numerical performance of BHsparse method with choices ζ at five quantiles (10%, 30%, 50%, 70%, 90%) of oracle intervals proposed in Theorem 3.3. Each subplot shows accuracy rates of the community size detection with increasing η for three network sizes ($N = 15000, 25000, 35000$).

5.1 DATA GENERATION AND SIMULATION SETTINGS

We simulate the network data under the framework of SBM, where we set up the probability connectivity matrix as $\mathbf{B} := \rho(\eta - 1)b[\mathbf{I}_K + \frac{1}{\eta-1}\mathbf{1}_K\mathbf{1}_K^T] := \rho\mathbf{B}_0$. ρ controls the expected degree of the network with the equation $\rho = N\tilde{d}/(\mathbf{1}_N^T(\mathbf{Z}\mathbf{B}_0\mathbf{Z}^T - tr(\mathbf{Z}\mathbf{B}_0\mathbf{Z}^T))\mathbf{1}_N)$, where $\tilde{d}$ denotes the expected expected degree. η controls the in-out ratio of the probability connectivity matrix, which also determines the assortativity strength. b is the baseline value in the probability connectivity matrix which we set as 0.1. To generate the data, we first simulate the membership vector $z \sim Multinomial(1; (\frac{1}{K}, ..., \frac{1}{K}))$. With a fixed membership vector and probability connectivity matrix under each setting, we then generate 20 repetitions of adjacency matrices.

We set the expected expected degree as $\tilde{d} = 3\sqrt{\log(N)}$ for all simulations. For the simulation shown in Figure 4.1, we consider three settings of node size $N \in \{15000, 25000, 35000\}$, 13 settings of in-out ratio η varying from 4 to 7 by 0.25, and one setting of community size $K = 3$. For simulations shown in Figures 5.2 and 5.3, we consider $N \in \{5000, 15000, 25000, 35000\}$, $\eta \in \{5, 6, 7, 8\}$ and $K \in \{3, 4\}$.

5.2 NUMERICAL RESULTS

We use empirical accuracy rate (ACR) to evaluate the performances of different community size estimation methods. Empirical accuracy rate (ACR) is the ratio of number of accurate estimation of community size out of 20 repetitions in each of the simulation setting.

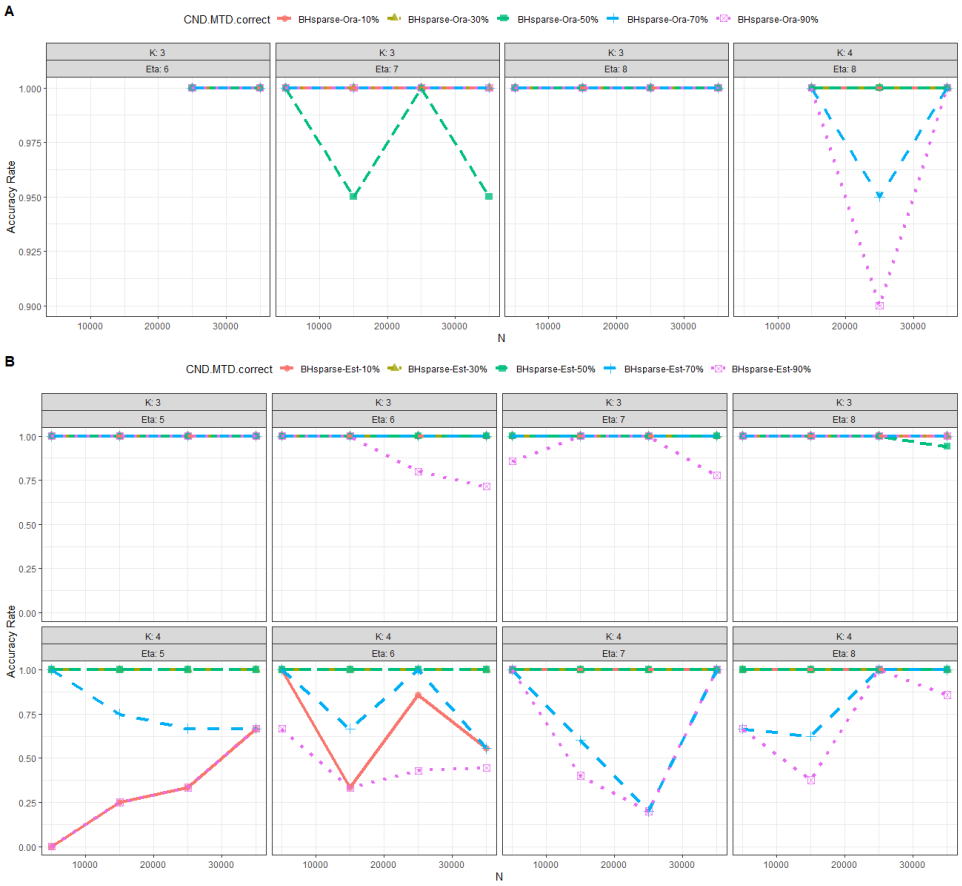

Figure 5.2: Numerical comparisons of BHsparse method with choices ζ as different quantiles of oracle and estimated intervals proposed in Lemma 3.3. Each subplot shows accuracy rates of the community size detection with increasing node sizes (N). Five quantiles (10%, 30%, 50%, 70%, 90%) of both (A) oracle intervals and (B) estimated intervals are compared. (A) only includes cases where oracle thresholds 3.4 are satisfied and (B) only includes cases where estimated thresholds 3.4 are satisfied.

Figure 5.1 shows the ACR performances of BHsparse method with different quantiles of the oracle interval with varying community assotativity strength (η). It is clear from Figure 5.1 that there is a sharp change in performance of the method of estimation of number of communities based on η. There is a threshold based on η such that below the threshold, detection of number of communities fails and above which it succeeds. In Figure 5.1, it is seen that the threshold decreases with minimum community size (N_{min}). The same result is also shown theoretically in the threshold equation 3.1.

Figure 5.2 shows the ACR performances of BHsparse method with different quantiles of the intervals (oracle or estimated). Only those cases where the oracle intervals or estimated intervals exist are shown in the plot. For the oracle and 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.

We then compare performances of BHsparse of ζ equals 30% and 50% quantiles of the intervals (both oracle and estimated) with BHmc and BHac. Figure 5.3 shows the ACR performances of Bethe Hessian matrix-based community size estimation method with these choices of ζ. When bounds exist (i.e., the assumption in Lemma 3.4 is satisfied), choices of ζ as 30% to 50% quantiles of the intervals we proposed perform better compared to the methods proposed in Le & Levina (2015). In particular, the relative accuracy of our method is higher when the number of nodes, N, is large in the sparse regime.

Figure 5.3: Numerical comparisons of BHsparse, BHmc and BHac. Each subplot shows accuracy rates (ACR) of the community size detection with the increasing node sizes (N). As choices of ζ for BHsparse, 30% quantile and 50% quantile of (A) oracle intervals and (B) estimated intervals are considered. (A) only includes cases where oracle thresholds 3.4 are satisfied and (B) only includes cases where estimated thresholds 3.4 are satisfied.

6 Discussion

In this paper, we contribute theoretical results on the selection of Bethe Hessian scalar parameter, ζ, for a consistent estimation of number of communities (K) in networks that are generated from the SBM with arbitrary degree of sparsity. To the best of our knowledge, this is the first study to theoretically prove the consistency of the Bethe Hessian spectral method to estimate K in sparse regimes with $\tilde{d} = o(\log(N))$. We also rigorously derive the oracle interval and provide a convenient way to empirically estimate the intervals for selecting ζ to construct the Bethe Hessian operator to consistently estimate K. We demonstrated our theoretical results with simulation studies.

Lastly, we related our result on the threshold for the estimation of K to the threshold for community recovery proposed in Decelle et al. (2011) and proved in Mossel et al. (2015) and Massoulié (2014). It suggests possible avenues for future research on this topic of thresholds for these seemingly unrelated tasks. Also in this paper, we only prove an upper bound of the hypothesized threshold for estimation of number of communities. An important future work will be to prove the lower bound results such that the existence of the threshold can be properly established.

REFERENCES

Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43. ACM, 2005.

Arash A Amini, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

Peter J Bickel and Purnamrita Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1(78): 253–273, 2015.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 1347–1357. IEEE, 2015.

Andressa Cerqueira and Florencia Leonardi. Estimation of the number of communities in the stochastic block model. *IEEE Transactions on Information Theory*, 2020.

Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

Simon Coste and Yizhe Zhu. Eigenvalues of the non-backtracking operator detached from the bulk. *arXiv preprint arXiv:1907.05603*, 2019.

Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. Revisiting the bethe-hessian: improved community detection in sparse heterogeneous graphs. In *Advances in Neural Information Processing Systems*, pp. 4037–4047, 2019.

Lorenzo Dall'Amico, Romain Couillet, and Nicolas Tremblay. Community detection in sparse time-evolving graphs with a dynamical bethe-hessian. *arXiv preprint arXiv:2006.04510*, 2020.

Javier De Las Rivas and Celia Fontanillo. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807, 2010.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, 2:38, 2014.

Michalis Faloutsos, Thomas Karagiannis, and Sue Moon. Online social networks. *IEEE network*, 24(5):4–5, 2010.

Karl J Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. *A survey of statistical network models*. Now Publishers Inc, 2010.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Jianwei Hu, Hong Qin, Ting Yan, and Yunpeng Zhao. Corrected bayesian information criterion for stochastic block models. *Journal of the American Statistical Association*, pp. 1–13, 2019.

Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. In *International Conference on Pervasive Computing*, pp. 133–151. Springer, 2011.

Eric D Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*, volume 65. Springer, 2014.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

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.

Sune Lehmann, Benny Lautrup, and Andrew D Jackson. Citation networks in high energy physics. *Physical Review E*, 68(2):026113, 2003.

Jing Lei et al. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1): 401–424, 2016.

Marc Lelarge. Community detection with the triplet loss, 2018. URL `https://openreview.net/forum?id=HkTisP1DM`.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020.

Shujie Ma, Liangjun Su, and Yichong Zhang. Determining the number of communities in degree-corrected stochastic block models. *arXiv preprint arXiv:1809.01028*, 2018.

Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 694–703, 2014.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

Mark Newman. *Networks*. Oxford university press, 2018.

Giuliano Andrea Pagani and Marco Aiello. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688–2700, 2013.

Ben Y Reis, Isaac S Kohane, and Kenneth D Mandl. An epidemiological network model for disease outbreak detection. *PLoS medicine*, 4(6):e210, 2007.

Maria A Riolo, George T Cantwell, Gesine Reinert, and Mark EJ Newman. Efficient method for estimating the number of communities in a network. *Physical review e*, 96(3):032310, 2017.

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

Alaa Saade, Florent Krzakala, and Lenka Zdeborová. Spectral clustering of graphs with the bethe hessian. In *Advances in Neural Information Processing Systems*, pp. 406–414, 2014a.

Alaa Saade, Florent Krzakala, and Lenka Zdeborová. Spectral density of the non-backtracking operator on random graphs. *EPL (Europhysics Letters)*, 107(5):50005, 2014b.

Alaa Saade, Florent Krzakala, and Lenka Zdeborová. Matrix completion from fewer entries: Spectral detectability and rank estimation. In *Advances in Neural Information Processing Systems*, pp. 1261–1269, 2015a.

Alaa Saade, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Spectral detection in the censored block model. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1184–1188. IEEE, 2015b.

Alaa Saade, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Clustering from sparse pairwise measurements. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 780–784. IEEE, 2016.

YX Rachel Wang, Peter J Bickel, et al. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2):500–528, 2017.

Bowei Yan, Purnamrita Sarkar, and Xiuyuan Cheng. Provable estimation of the number of blocks in block models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1185–1194, 2018.