JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–16: Theory and Methods
https://doi.org/10.1080/01621459.2023.2223793

On the Estimation of the Number of Communities for Sparse Networks

Neil Hwang[a], Jiarui Xu[b], Shirshendu Chatterjee[c], and Sharmodeep Bhattacharyya[d]

[a]Department of Business and Information Systems, City University of New York - Bronx Community College, New York, NY; [b]Meta Platforms Inc., Bellevue, WA; [c]Department of Mathematics, City University of New York, New York, NY; [d]Department of Statistics, Oregon State University, Corvallis, OR

ABSTRACT

Among the nonparametric methods of estimating the number of communities (K) in a community detection problem, methods based on the spectrum of the Bethe Hessian matrices ($\mathbf{H}_\zeta$ with the scalar parameter ζ) have garnered much popularity for their simplicity, computational efficiency, and robustness to the sparsity of data. For certain heuristic choices of ζ, such methods have been shown to be consistent for networks with N nodes with a common expected degree of $\omega(\log N)$. In this article, we obtain several finite sample results to show that if the input network is generated from either stochastic block models or degree-corrected block models, and if ζ is chosen from a certain interval, then the associated spectral methods based on $\mathbf{H}_\zeta$ is consistent for estimating K for the sub-logarithmic sparse regime, when the expected maximum degree is both $o(\log N)$ and $\omega(1)$, under some mild conditions even in the situation when K increases with N. We also propose a method to estimate the aforementioned interval empirically, which enables us to develop a consistent K estimation procedure in the sparse regime. We evaluate the performance of the resulting estimation procedure theoretically, also empirically through extensive simulation studies and application to a comprehensive collection of real-world network data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2022
Accepted May 2023

KEYWORDS
Community number estimation; Degree-corrected block model; Sparse networks; Spectral clustering; Stochastic block model; The Bethe Hessian operator

1. Introduction

Statistical analysis of networks has now become a well-studied field within statistics (see Goldenberg et al. 2010; Kolaczyk and Csárdi 2014 for reviews). Methods for network data analysis are being developed not only in the discipline of statistics but also in computer science, physics, and mathematics. Network datasets show up in several disciplines. Examples include networks originating from biosciences such as gene regulation networks (Emmert-Streib, Dehmer, and Haibe-Kains 2014), protein–protein interaction networks (De Las Rivas and Fontanillo 2010), structural (Rubinov and Sporns 2010) and functional networks (Friston 2011) of brain and epidemiological networks (Reis, Kohane, and Mandl 2007); networks originating from social media such as Facebook, Twitter and LinkedIn (Faloutsos, Karagiannis, and Moon 2010); citation and collaboration networks (Lehmann, Lautrup, and Jackson 2003); information and technological networks such as internet-based networks (Adamic and Glance 2005), power networks (Pagani and Aiello 2013) and cell-tower networks (Isaacman et al. 2011). There are several active research areas in developing statistical inference methods for network data analysis and for deriving the theoretical properties of the statistical methods. Examples of inferential questions that have received a lot of attention in current active research include fitting random graph models to the network datasets (Goldenberg et al. 2010), finding stochastic properties of summary statistics of networks like subgraph counts (Bickel, Chen, and Levina 2011), community detection (Fortunato 2010; Abbe 2017) and link prediction (Liben-Nowell and Kleinberg 2007).

The last two decades saw a resurgence of interest in a problem popularly known as "community detection." For this problem, the main task is to partition the nodes of a given graph into K communities so that the number of edges within communities is more than that between communities, where K is assumed to be known a priori. Such community structures are also known as *assortative* community structures. Estimating the number of clusters in a clustering problem has quite a bit of history (e.g., Rousseeuw 1987; Tibshirani, Walther, and Hastie 2001; Von Luxburg 2010). However, estimating the number of clusters in a clustering problem does not directly carry over to the problem of estimating the number of communities in a community detection problem.

Recently, a new line of research to estimate K for network datasets with community structure has become active in the literature. Typical approaches in this regard include designing suitable algorithms, testing the performance of the algorithms using simulated data, and proving desirable properties of the algorithms (e.g., Le and Levina 2015; Bordenave, Lelarge, and Massoulié 2015; Gulikers, Lelarge, and Massoulié 2016; Riolo et al. 2017; Wang and Bickel 2017; Yan, Sarkar, and Cheng 2018; Hu et al. 2019; Ma, Su, and Zhang 2021). Degree-Corrected Block Model (DCBM), Stochastic Block Model (SBM), and several of their variants have been widely used to generate random networks with community structure.

CONTACT Sharmodeep Bhattacharyya *bhattash@science.oregonstate.edu* Department of Statistics, Oregon State University, Corvallis, OR.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

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 SBM (e.g., Saade, Krzakala, and Zdeborová 2014; Riolo et al. 2017; Yan, Sarkar, and Cheng 2018), there has been recent progress in attaining a theoretical understanding of community number estimation using model selection techniques. Bhattacharyya and Bickel (2015) proposed a hypothesis testing approach with bootstrapping, Chen and Lei (2018) and Li, Levina, and Zhu (2020) used cross-validation, Yan, Sarkar, and Cheng (2018) proposed a semidefinite programming approach, Hu et al. (2019) used the BIC criterion, Chen and Hero (2018) considered a phase transition criterion, and Ma, Su, and Zhang (2021) presented an approach based on binary segmentation and profile likelihood ratio. Cerqueira and Leonardi (2018) introduced a novel method, albeit computationally inefficient, based on a penalized version of the Krichevsky-Trofimov (KT) mixture distribution to detect K for SBM networks, and the authors showed consistency of the method in sparse regimes provided the expected mean degree grows to infinity. A stepwise goodness of fit approach for the estimation of K was also provided in another recent work by Jin et al. (2022). But the method in Jin et al. (2022) was based on denser networks.

As nonparametric alternatives that are computationally more efficient and applicable to a broader range of settings, methods based on the spectrum of the non-backtracking operator[1] (Bordenave, Lelarge, and Massoulié 2015; Gulikers, Lelarge, and Massoulié 2016) and Bethe Hessian operator[2] have been considered (e.g., Saade, Krzakala, and Zdeborová 2014; Le and Levina 2015; Dall'Amico, Couillet, and Tremblay 2019, 2020). Bordenave, Lelarge, and Massoulié (2015) and Gulikers, Lelarge, and Massoulié (2016) analyzed the spectrum of the non-backtracking matrix theoretically in the sparse case for two-community networks. A review of the community number estimation methods related to the Bethe Hessian and the non-backtracking matrix is given in Section 2.4. *However, none of the proposed methods in the literature produces a computationally efficient (polynomial-time) estimator of the number of communities with mathematically rigorous proofs for correctness and consistency* in the sparse network regime when the expected degrees of the nodes are sub-logarithmic in the number of nodes for networks with community structures. We focus on this specific problem in this article.

The main contributions of this article are the following.

- We show that if ζ is chosen from a specific interval, then the number of communities K of networks generated from either the SBM or the DCBM can be estimated by counting the number of negative eigenvalues of the associated Bethe Hessian matrix $\mathbf{H}_\zeta$ having parameter ζ even when the network is heterogeneous in terms of the expected degrees of the nodes. We derive theoretical results in order to show that the estimated number of communities is consistent in the sparse regime, where the maximum among the expected degrees of all N nodes is both $\omega(1)$ and $o(\log N)$, under some mild conditions on the signal-to-noise parameters of the SBM and DCBM, even when K increases with N. *Results in this article are one of the very few rigorous results to show that it is possible to estimate the number of communities consistently and (computationally) efficiently for networks in the sparse regime.*
- We also provide a method to find an estimator within the aforementioned interval from which one needs to choose the parameter ζ in order to have a consistent estimator of K using $\mathbf{H}_\zeta$. We derive theoretical results in order to show that the estimator of K based on the empirically estimated ζ as mentioned above is consistent in the sparse regime under some mild conditions on the parameters of the SBM and DCBM. We note that there are two tuning parameters in the estimation procedure. We provide theoretical guidance in Lemma 4.3 and empirical guidance in Section 5.2.1 for tuning these hyperparameters. *To the best of our knowledge, the K-estimation procedure we have developed here is the first of its kind in that it is simultaneously (a) computationally efficient (polynomial-time), (b) provably consistent (a rigorous analytic proof is given in Section 4), and (c) works consistently even when the input network is in the sparse regime.*
- We demonstrate the efficacy of our methods via extensive simulation studies and the application of our approach to a comprehensive collection of real-world network data arising in diverse areas of interest.

The rest of the article is organized into five sections. In Section 2, we provide an introduction to the network data generating models, the Bethe Hessian matrices, and the relevant literature. In Section 3, we provide theoretical results on the Bethe Hessian matrices for networks in the sparse regime. In Section 4, we provide an algorithm for estimating K and provide theoretical results on the consistency of the estimator. In Section 5, we provide an extensive simulation study on various aspects of the proposed method to estimate K and compare our proposed method with existing ones. In Section 6, we apply our proposed method to a large collection of real-world network datasets.

2. Preliminaries

2.1. Notation

For some constant N, if for all $n \geq N$, we use the well-known rate notations,

$$f(n) = \begin{cases} O(g(n)) & \text{if } \limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty \\ o(g(n)) & \text{if } \limsup_{n\to\infty} \frac{f(n)}{g(n)} = 0 \\ \omega(g(n)) & \text{if } \liminf_{n\to\infty} \frac{f(n)}{g(n)} = \infty \\ \Theta(g(n)) & \text{if } \liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0 \text{ and} \\ & \qquad \limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty. \end{cases}$$

$\mathbb{R}^+$ (resp. $\mathbb{Z}^+$) denotes the set of positive real numbers (resp. integers). For $N \in \mathbb{Z}^+$, let $[N]$ denote the set $\{1, 2, \ldots, N\}$. The adjacency matrix, denoted by $\mathbf{A}$, is a random symmetric

[1] Given a set of edges E in a graph, the non-backtracking operator is a $2|E| \times 2|E|$ matrix indexed by directed edges such that the (ij, kl)th element takes on a value of 1 if $j = k, i \neq l$ and 0 otherwise.

[2] Given an $N \times N$ adjacency matrix $\mathbf{A}$, the Bethe Hessian operator is an $N \times N$ matrix defined as $(\zeta^2 - 1)\mathbf{I} - \zeta\mathbf{A} + \mathbf{D}$ with parameter $\zeta > 1$, a diagonal matrix $\mathbf{D}$ with the degree of node i in $\mathbf{D}_{ii}$ and identity matrix $\mathbf{I}$. See Section 2.3 for a detailed discussion.

matrix whose rows and columns are labeled by nodes in $[N]$, where A_{ij} equals 1 (resp. 0) if there is an (resp. no) edge between nodes i and j. $\mathbf{1}_N$ denotes a vector of N ones and $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. The average observed degree (the average row sum of $\mathbf{A}$) is denoted by $d^{\mathbf{A}} := \frac{1}{N}\mathbf{1}_N^T\mathbf{A}\mathbf{1}_N$. $\lambda_\ell^{\downarrow}(\mathbf{A})$ (resp. $\lambda_\ell^{\uparrow}(\mathbf{A})$) denotes the ℓth largest (resp. smallest) eigenvalue of $\mathbf{A}$. $\lambda_{\max}(\mathbf{A})$ (resp. $\lambda_{\min}(\mathbf{A})$) denotes the largest (resp. smallest) eigenvalue of $\mathbf{A}$. $\|\cdot\|$ (resp. $\|\cdot\|_F$, $\|\cdot\|_{\psi_{1/2}}$, and $\rho(\cdot)$) denotes the spectral norm (resp. Frobenius norm, Orlicz norm, and spectral radius) for matrices. Note that the spectral radius for a square matrix is its maximum absolute eigenvalue. Given a vector $\mathbf{x}$ having length n, let $\mathrm{diag}(\mathbf{x})$ be an $n \times n$ diagonal matrix whose diagonal entries are those of $\mathbf{x}$. For an $n \times n$ matrix $\mathbf{M}$, let $\mathrm{diag}(\mathbf{M})$ denote an $n \times n$ diagonal matrix which is obtained from $\mathbf{M}$ by zeroing out all off-diagonal entries. For a matrix $\mathbf{M}$, M_{ij} denotes its (i,j)th element. Similarly, x_i denotes the ith element of a vector $\mathbf{x}$. For a given parameter p, we use $\hat{p}$ to denote an estimator that will be specified. We denote the cardinality of a finite set S by $|S|$. For notational convenience, we use $\mathbf{P}$ to denote $\mathbb{E}(\mathbf{A})$. A summary of some more notations involving $\mathbf{A}$ and $\mathbf{P}$ is given in Table 1.

2.2. Network Generative Model

The degree-corrected block model (DCBM), which includes the standard stochastic block model (SBM) as a special case, is a generative model for network data that embeds a community structure in $\mathbf{A}$. DCBM has four sets of parameters: (a) the number of communities K; (b) the community allocation probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \in (0,1)^K$ satisfying $\sum_{i\in[K]} \pi_i = 1$; (c) the degree parameter vector $\Psi = (\psi_1, \ldots, \psi_N)$ with $\psi_i \in (0,1], \forall i \in [N]$; and (d) the $K \times K$ connectivity probability matrix $\mathbf{B}$ of rank K with all positive eigenvalues[3]. Given the parameters, the community labels $z_1, \ldots, z_N$ for nodes $1, \ldots, N$, respectively are generated independently from Multinomial$(1; \boldsymbol{\pi})$ distribution. Having obtained the community labels, nodes i and j are connected with probability $\psi_i\psi_j\mathbf{B}_{z_i,z_j}$. The latent membership vector $\mathbf{z} = (z_1, \ldots, z_N) \in [K]^N$ consisting of community labels for all nodes is a set of unknown latent variables for the model. For each $k \in [K]$, $C_k \subseteq [N]$ denotes the set of all nodes having community label k.

In our theoretical results in Section 3, we take Ψ to be a set of fixed (nonrandom) numbers in $(0,1)$. In the supplement (see Section B.2), we extend our theoretical results to the case where Ψ is allowed to be a random vector. In order to maintain the identifiability of the DCBM parameters, we assume that Ψ are normalized so that the maximum in each community is 1 (as done in Lei and Rinaldo 2015). That is,

$$\max_{i\in C_k} \psi_i = 1 \text{ for all } k \in [K] \tag{1}$$

Suppose $(\psi_i')_i$ are unnormalized degree parameters, then we take for all i, $\psi_i = \psi_i'/\max_{j\in C_k}\psi_j', i \in C_k$, so that the identifiability assumption is satisfied.

We let $N_k := \sum_{i\in C_k}[\psi_i^2/d_i^{\mathbf{P}}]$, where $d_i^{\mathbf{P}}$ is the expected degree of node i, for each $k \in [K]$. For $k \in [K]$, N_k is a measure of the contribution of degree parameters to the expected degree in the kth community. Without loss of generality, we consider that community sizes are in descending order, that is, $|C_1| \geq |C_2| \geq \cdots \geq |C_K|$. Community sizes are assumed to be balanced, that is, $|C_1|/|C_K| = \Theta(1)$. Let $\mathbf{Z}$ be the $N \times K$ community membership matrix, such that $Z_{ik} = 1$ if node $i \in C_k$. Then, the above network generating mechanism tells us that

$$\begin{aligned}\mathbf{Z}_{1,*}, \ldots, \mathbf{Z}_{N,*} &\overset{\text{iid}}{\sim} \text{Multinomial}(1; \boldsymbol{\pi}),\\ \mathbf{P} &:= \mathrm{diag}(\Psi)\big(\mathbf{Z}\mathbf{B}\mathbf{Z}^T - \mathrm{diag}(\mathbf{Z}\mathbf{B}\mathbf{Z}^T)\big)\mathrm{diag}(\Psi),\\ A_{ij} &\overset{\text{indep}}{\sim} \text{Bernoulli}(P_{ij}) \text{ for } 1 \leq i < j \leq N.\end{aligned}$$

We denote the maximum expected degree by $d_{\max}^{\mathbf{P}} := \max_{i\in[N]}\sum_{j=1}^N P_{ij}$. $d_{\min}^{\mathbf{P}}$ is defined in a similar way. $b_{\max}$ and λ denote the largest element of $\mathbf{B}$ and the minimum eigenvalue of the normalized matrix $\frac{\mathbf{B}}{b_{\max}}$, respectively.

The SBM is obtained from the DCBM as a special case by setting $\psi_i = 1$ for all $i \in [N]$.

We summarize our notations in Table 1.

2.3. The Bethe Hessian Matrix

The Bethe Hessian matrix for a network is defined as

$$\mathbf{H}_\zeta := (\zeta^2 - 1)\mathbf{I}_N + \mathbf{D}_1^{\mathbf{A}} - \zeta\mathbf{A} \tag{2}$$

where $\zeta > 1$ is a scalar parameter, $\mathbf{D}_1^{\mathbf{A}} := \mathrm{diag}(\mathbf{A}\mathbf{1}_N)$ and $\zeta > 1$. ζ is often referred to as the "radius" of $\mathbf{H}_\zeta$. Once an appropriate value for ζ is chosen, the discrete set of negative eigenvalues of the resulting matrix $\mathbf{H}_\zeta$ are isolated from the contiguous bulk. Then, estimating K entails counting the number of these isolated negative eigenvalues. For this reason, the negative eigenvalues are referred to as the "informative eigenvalues" of the matrix, while the contiguous bulk is referred to as uninformative. Next, we discuss several heuristics that have been proposed in the literature for choosing ζ so that the negative eigenvalues of the associated Bethe Hessian matrix are informative. We also discuss some other approaches for estimating K.

2.4. Review of Other K-Estimation Methods

What follows is an overview of various approaches that have been proposed in the literature for estimating K. While some are based on the Bethe Hessian matrix, many others are not. Some of the methods have been used for comparative empirical studies in Sections 5 and 6.

Several methods have been proposed based on the Bethe Hessian matrices. First, on the question of what values to use for the parameter ζ, it was first empirically shown in Saade, Krzakala, and Zdeborová (2014) that the number of negative eigenvalues of the Bethe Hessian matrix directly estimates K when ζ is assigned either $\sqrt{d^{\mathbf{P}}}$ for assortative networks or $\sqrt{\rho(\mathbf{NB})}$ for sparse networks with the bounded expected degree, where $\mathbf{NB}$ is the non-backtracking matrix.[4] The consistency of these estimators was proved in Le and Levina (2015) for regimes $d^{\mathbf{P}} = \omega(\log N)$, in which estimator $\hat{\zeta}_a := \left(\frac{1}{N}\sum_{i\in[N]} d_i^{\mathbf{A}}\right)^{1/2}$ was

[3] This is a standard condition to ensure assortative community structures (Lei and Rinaldo 2015; Bhattacharyya and Chatterjee 2020).

[4] See footnote 1 for definition.

Table 1. Summary of notations.

Sample		Population	
Notation	Definition	Notation	Definition
$\mathbf{A}$	$N \times N$ adjacency matrix	$\mathbf{P}$	$\mathbb{E}(\mathbf{A})$
$d_i^{\mathbf{A}}$	Observed degree of node i (ith row sum of $\mathbf{A}$)	$d_i^{\mathbf{P}}$	Expected degree of node i (ith row sum of $\mathbf{P}$)
$d^{\mathbf{A}}$	Average observed degree (average of row sums of $\mathbf{A}$)	$d^{\mathbf{P}}$	Average *expected* degree (average of row sums of $\mathbf{P}$)
$d_{\max}^{\mathbf{A}}$	Maximum observed degree (maximum row sum of $\mathbf{A}$)	$d_{\max}^{\mathbf{P}}$	Maximum expected degree (maximum row sum of $\mathbf{P}$)
$d_{\min}^{\mathbf{A}}$	Minimum observed degree (minimum row sum of $\mathbf{A}$)	$d_{\min}^{\mathbf{P}}$	Minimum expected degree (minimum row sum of $\mathbf{P}$)
$\hat{d}_{\min}^{\mathbf{P}}$	Estimator of $d_{\min}^{\mathbf{P}}$	$b_{\max}$	Largest element of matrix $\mathbf{B}$
$\lambda_{\min}(\mathbf{A})$	Smallest eigenvalue of $\mathbf{A}$	λ	$\lambda_{\min}\left(\mathbf{B}/b_{\max}\right)$
$\mathbf{D}_1^{\mathbf{A}}$	$\operatorname{diag}(\mathbf{A}\mathbf{1}_N)$ (diagonal matrix where the observed degrees are on the diagonal)	$\mathbf{D}_1^{\mathbf{P}}$	$\operatorname{diag}(\mathbf{P}\mathbf{1}_N)$ (diagonal matrix where the expected degrees are on the diagonal)
$\mathbf{D}_\zeta^{\mathbf{A}}$	$(\zeta - \frac{1}{\zeta})\mathbf{I}_N + \frac{1}{\zeta}\mathbf{D}_1^{\mathbf{A}}$ for $\zeta \geq 1$	$\mathbf{D}_\zeta^{\mathbf{P}}$	$(\zeta - \frac{1}{\zeta})\mathbf{I}_N + \frac{1}{\zeta}\mathbf{D}_1^{\mathbf{P}}$ for $\zeta \geq 1$
		N_k	$\sum_{i \in C_k}[\psi_i^2/d_i^{\mathbf{P}}]$ where C_k is kth largest community

Table 2. Definitions of ζ estimates and interval bounds.

Method	Notation	Definition
1	$\hat{\zeta}$	An estimate chosen pursuant to Step (3) in Algorithm 4.1
2	$\hat{\zeta}_a$	Defined as $\left(\frac{1}{N}\sum_{i\in[N]} d_i^{\mathbf{A}}\right)^{1/2}$ as an estimator for $\sqrt{Nb_{\max}}$
3	$\hat{\zeta}_m$	Defined as $\left(\left(\sum_{i\in[N]}(d_i^{\mathbf{A}})^2\right)/\left(\sum_{i\in[N]} d_i^{\mathbf{A}}\right) - 1\right)^{1/2}$ as an approximation of $\sqrt{\rho(\mathbf{NB})}$
4	$\hat{\zeta}_{\mathbf{NB}}$	$\sqrt{\rho(\mathbf{NB})}$ calculated using the approximation of $\mathbf{NB}$ in eq. (27) in Dall'Amico, Couillet, and Tremblay (2021)
5	ζ_L	c/ν_K in Claim 1 in Dall'Amico, Couillet, and Tremblay (2021) as an approximation of the lower bound for the interval for ζ
6	ζ_U	$\sqrt{c\Phi}$ in Claim 1 in Dall'Amico, Couillet, and Tremblay (2021) as an approximation of the upper bound for the interval for ζ[1]

NOTE: Methods 2 and 3 are from (Le and Levina 2015). Method 4 is from (Dall'Amico, Couillet, and Tremblay 2021).
[1] $c = \mathbf{1}^T \frac{1}{\mathbb{E}[\psi_i]^2}\mathbf{B}\operatorname{diag}(|C_1|,\ldots,|C_K|)\mathbf{1}/K$, $\nu_p = \lambda_p^{\downarrow}(\frac{1}{\mathbb{E}[\psi_i]^2}\mathbf{B}\operatorname{diag}(|C_1|,\ldots,|C_K|))$, and $\Phi = \frac{1}{\mathbb{E}[\psi_i]^2}\mathbb{E}[\psi_i^2]$.

Table 3. K-estimation methods.

	Algorithm	Description and source
1	$\mathrm{BH}_{\hat{\zeta}}$	An estimate for K based on Algorithm 4.1
2	BHa	Counting the negative eigenvalues of $\mathbf{H}_{\hat{\zeta}_a}$
3	BHm	Counting the negative eigenvalues of $\mathbf{H}_{\hat{\zeta}_m}$
4	BHac	BHa with a correction for small positive eigenvalues
5	BHam	BHm with a correction for small positive eigenvalues
6	$\mathrm{BH}_{\mathbf{NB}}$	Counting the negative eigenvalues of $\mathbf{H}_{\hat{\zeta}_{\mathbf{NB}}}$
7	LRBIC	Penalized likelihood from Wang and Bickel (2017)
8	NCV	Cross-validation based on node-pair splitting procedure from Chen and Lei (2018)
9	ECV	Cross-validation for approximately low-rank networks from Li, Levina, and Zhu (2020).

NOTE: Methods 2 through 5 are from Le and Levina (2015). Method 6 is from (Dall'Amico, Couillet, and Tremblay 2021).

Table 4. Simulation settings.

Setting	N	η	γ	ϕ	Ψ
1	15,000	25	1	1	Unif(0.5, 1)
2	{1000, 5000, 15,000, 25,000, 35,000}	25	1	1	Unif(0.5, 1)
3	15,000	{15, 17.5, 20, 22.5, 25}	1	1	Unif(0.5, 1)
4	15,000	25	{0.8, 0.85, 0.9, 0.95, 1}	1	Unif(0.5, 1)
5	15,000	25	1	{0.8, 0.85, 0.9, 0.95, 1}	Unif(0.5, 1)
6	15,000	25	1	1	$\{I, U_A, U_B, P_A, P_B\}^*$
7	1000	{15, 17.5, 20, 22.5, 25}	1	1	Unif(0.5, 1)

NOTE: $K \in \{3, 7, 10, 25\}$ for all settings. *For notational convenience, we denote $I = \mathrm{Unif}(1,1)$, $U_A = \mathrm{Unif}(0.75, 1)$, $U_B = \mathrm{Unif}(0.5, 1)$, $P_A = \mathrm{Pareto}(0.58, 10)$, and $P_B = \mathrm{Pareto}(1.16, 10)$.

proposed for $\sqrt{d^{\mathbf{P}}}$ and computationally efficient approximation $\hat{\zeta}_m := \left(\left(\sum_{i\in[N]}(d_i^{\mathbf{A}})^2\right)/\left(\sum_{i\in[N]} d_i^{\mathbf{A}}\right) - 1\right)^{1/2}$ was proposed for $\sqrt{\rho(\mathbf{NB})}$. Le and Levina (2015) then proposed estimating K by counting the number of negative eigenvalues of $\mathbf{H}_{\hat{\zeta}_a}$, that is, the Bethe Hessian matrix based on $\hat{\zeta}_a$ as the parameter value, and similarly $\mathbf{H}_{\hat{\zeta}_m}$. To overcome the observed tendency of these methods to underestimate K in unbalanced networks, the authors proposed a correction whereby some of the small positive eigenvalues sufficiently isolated within the bulk are counted as informative.

The method perhaps closest to our proposed approaches is discussed in Dall'Amico, Couillet, and Tremblay (2021). In that paper, the authors claim, with empirical support and non-rigorous but intuitive arguments from statistical physics, that K can be inferred by (a) finding a value for ζ such that the largest eigenvalue of $\mathbf{H}_\zeta$ that is isolated from the bulk is zero; then (b) concluding that K is the number of isolated eigenvalues. The authors in Dall'Amico, Couillet, and Tremblay (2021) proposed $\sqrt{\rho(\mathbf{NB})}$ as an estimator of ζ. The value of $\sqrt{\rho(\mathbf{NB})}$ was calculated using the largest eigenvalue in modulus of a variation of the non-backtracking matrix (eq. (27) in Dall'Amico, Couillet, and Tremblay 2021). In Section 5, we provide an empirical comparison of the approach proposed in Dall'Amico, Couillet, and Tremblay (2021) and our methods with extensive simulations.

Other than the Bethe Hessian-based methods, there have been several other approaches proposed in the literature. The more commonly-studied approaches have centered around the model selection. Wang and Bickel (2017) proposed a model selection criterion for choosing K in the form of a penalized likelihood. The authors proved its asymptotic consistency under

both SBM and DCBM in the $O(\log N)$ regime. Ma, Su, and Zhang (2021) used pseudo-likelihood ratio and binary segmentation to estimate the number of communities in DCBM. But these likelihood-based methods are not computationally efficient. Another class of methods for model selection uses network cross-validation, which is based on the idea that the community structure can still be recovered even when a small subset (validation set) of the edges is removed for use in selecting the optimal K based on the performance on this validation set. In Chen and Lei (2018), a cross-validation strategy was proposed adapted to SBM and DCBM data based on a node-pair splitting procedure. Li, Levina, and Zhu (2020) proposed a more general cross-validation strategy applicable to a broader class of networks that are approximately low-rank. However, none of these proposed estimators for K has asymptotic consistency in the $o(\log N)$ regime. Also, compared to the methods based on the Bethe Hessian matrices, the model selection methods are computationally more expensive due to multiple iterations of calculations of likelihood and loss functions, which renders them impractical when network data are large.

In Sections 5 and 6, we evaluate the performance of our methods against two groups of benchmarks using simulated and real-world data. First, we test our method for choosing the parameter ζ consisting of our oracle interval derived in Section 3 and their empirical estimators. Second, we compare the accuracy of our algorithm for estimating K with other Bethe Hessian-based methods and non-Bethe Hessian approaches in Section 5.

3. Theoretical Results on the Bethe Hessian Matrices

Our main theoretical contribution is twofold. First, we show that even in a sparse regime when $Nb_{\max}$ is $o(\log N)$ and $\omega(1)$, the number of informative (negative) eigenvalues of $\mathbf{H}_\zeta$ directly estimates K consistently. Second, we propose a novel interval of appropriate values such that for any ζ chosen from this interval, the number of informative eigenvalues of the associated matrix $\mathbf{H}_\zeta$ directly estimates K. We show that this interval also conveniently serves as a sufficient condition for the correct estimation of K. Below, we walk through our theoretical results in several steps and build intuition with intermediate results.

In this section, we make the assumption that the degree parameters Ψ are fixed. In supplement section B.2, we relax this assumption and extend our theoretical results in this section by allowing Ψ to be random. Relevant prior results from the literature are given in supplement section A and full proofs for all of our results are presented in supplement section C.

The first step in our analysis involves exploiting the connection between the Bethe Hessian matrices and normalized Laplacian matrices. The connection between these two matrices was pointed out in Dall'Amico, Couillet, and Tremblay (2019). In order to prove the properties of the eigenvalues of the Bethe Hessian matrices, we first move from the Bethe Hessian matrices to their Laplacian counterparts. We show that the Bethe Hessian matrix and its Laplacian have the same inertia for the same adjacency matrices.

Lemma 3.1. Let $\mathbf{L}_\zeta^{\mathbf{A}} := \frac{1}{\zeta}\mathbf{H}_\zeta = \mathbf{D}_\zeta^{\mathbf{A}} - \mathbf{A}$ and define the symmetric normalized Laplacian $\mathcal{L}(\mathbf{L}_\zeta^{\mathbf{A}}) := (\mathbf{D}_\zeta^{\mathbf{A}})^{-1/2}\mathbf{L}_\zeta^{\mathbf{A}}(\mathbf{D}_\zeta^{\mathbf{A}})^{-1/2}$. Then, $\mathbf{H}_\zeta$ and $\mathcal{L}(\mathbf{L}_\zeta^{\mathbf{A}})$ have the same number of negative eigenvalues.

The proof of the result uses Sylvester's Law of Inertia (Horn and Johnson 2012) and is given as Theorem A.4 in supplement section A.

Now, we focus on the symmetric normalized Laplacian matrix, $\mathcal{L}(\mathbf{L}_\zeta^{\mathbf{A}})$. The next result shows that the symmetric normalized Laplacian of the Bethe Hessian matrix concentrates around its expectation.

Lemma 3.2. Let $\mathcal{L}(\mathbf{L}_\zeta^{\mathbf{A}})$ be as defined in Lemma 3.1, and we similarly define its population counterpart as $\mathcal{L}(\mathbf{L}_\zeta^{\mathbf{P}}) := (\mathbf{D}_\zeta^{\mathbf{P}})^{-1/2}\mathbf{L}_\zeta^{\mathbf{P}}(\mathbf{D}_\zeta^{\mathbf{P}})^{-1/2}$, where $\mathbf{L}_\zeta^{\mathbf{P}} := \mathbf{D}_\zeta^{\mathbf{P}} - \mathbf{P}$. Suppose $Nb_{\max}$ is $o(\log N)$ and $\omega(1)$. Then, for any $r \geq 1$, with probability at least $1 - e^{-r}$,

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

for some constant C and $\zeta = \omega(\sqrt{Nb_{\max}})$.[5]

Lemma 3.2 gives a concentration bound for $\mathcal{L}(\mathbf{L}_\zeta^{\mathbf{A}})$ around its population counterpart, and one can see that with an increase in r, the bound becomes a high probability event. Note that this result holds for sparse regimes of expected degrees that are $o(\log N)$. The proof uses concentration results from random matrix theory stated as Theorem A.3 in Supplement section A.

Since for certain choices for the scalar parameter ζ, the population counterpart of the Laplacian, $\mathcal{L}(\mathbf{L}_\zeta^{\mathbf{P}})$, has exactly K negative eigenvalues, combined with Lemmas 3.1 and 3.2, the final result is obtained which says that $\mathbf{H}_\zeta$ has exactly K negative eigenvalues with high probability for certain values of the scalar parameter ζ.

Theorem 3.3. For a network generated from the DCBM with parameters $(K, \boldsymbol{\pi}_{K\times 1}, \Psi_{N\times 1}, \mathbf{B}_{K\times K})$, suppose that community sizes are balanced, that is, $\frac{|C_1|}{|C_K|} = \Theta(1)$. Let $\beta := b_{\max}\lambda d_{\min}^{\mathbf{P}} N_K$, where $b_{\max}, \lambda, d_{\min}^{\mathbf{P}}$, and N_K are as defined in Table 1. Then, under the condition that $Nb_{\max}$ is both $o(\log N)$ and $\omega(1)$, $\mathbf{H}_\zeta$ has exactly K negative eigenvalues for all

$$\zeta \in \frac{1}{2}\left(\beta \pm \sqrt{\beta^2 + 4 - 4d_{\min}^{\mathbf{P}}}\right) \tag{3}$$

with probability at least $1 - \exp\left[-(\zeta/\sqrt{Nb_{\max}})^{3/2-\delta}\right]$ for any $\delta \in (0, 3/2)$. The last probability bound is $1 - o(1)$ if either

$$\lambda = \omega\left(\frac{K}{\sqrt{Nb_{\max}}} \cdot \frac{d_{\max}^{\mathbf{P}}}{d_{\min}^{\mathbf{P}}}\right) \tag{4}$$

or (assuming $d_{\max}^{\mathbf{P}}/d_{\min}^{\mathbf{P}} = \Theta(1)$)

$$\lambda = \omega\left(\frac{K}{\sqrt{Nb_{\max}}}\right) \tag{5}$$

[5] As $\zeta \to 1$, the left-hand side of the inequality in Lemma 3.2 approaches the value $\left\|(\mathbf{D}_1^{\mathbf{P}})^{-1/2}\mathbf{P}(\mathbf{D}_1^{\mathbf{P}})^{-1/2} - (\mathbf{D}_1^{\mathbf{A}})^{-1/2}\mathbf{A}(\mathbf{D}_1^{\mathbf{A}})^{-1/2}\right\|$. Here, note that the term $(\mathbf{D}_1^{\mathbf{A}})^{-1/2}$ explodes in sparse settings. For this reason, the hypothesis of the Lemma includes the condition that $\zeta = \omega(\sqrt{Nb_{\max}})$.

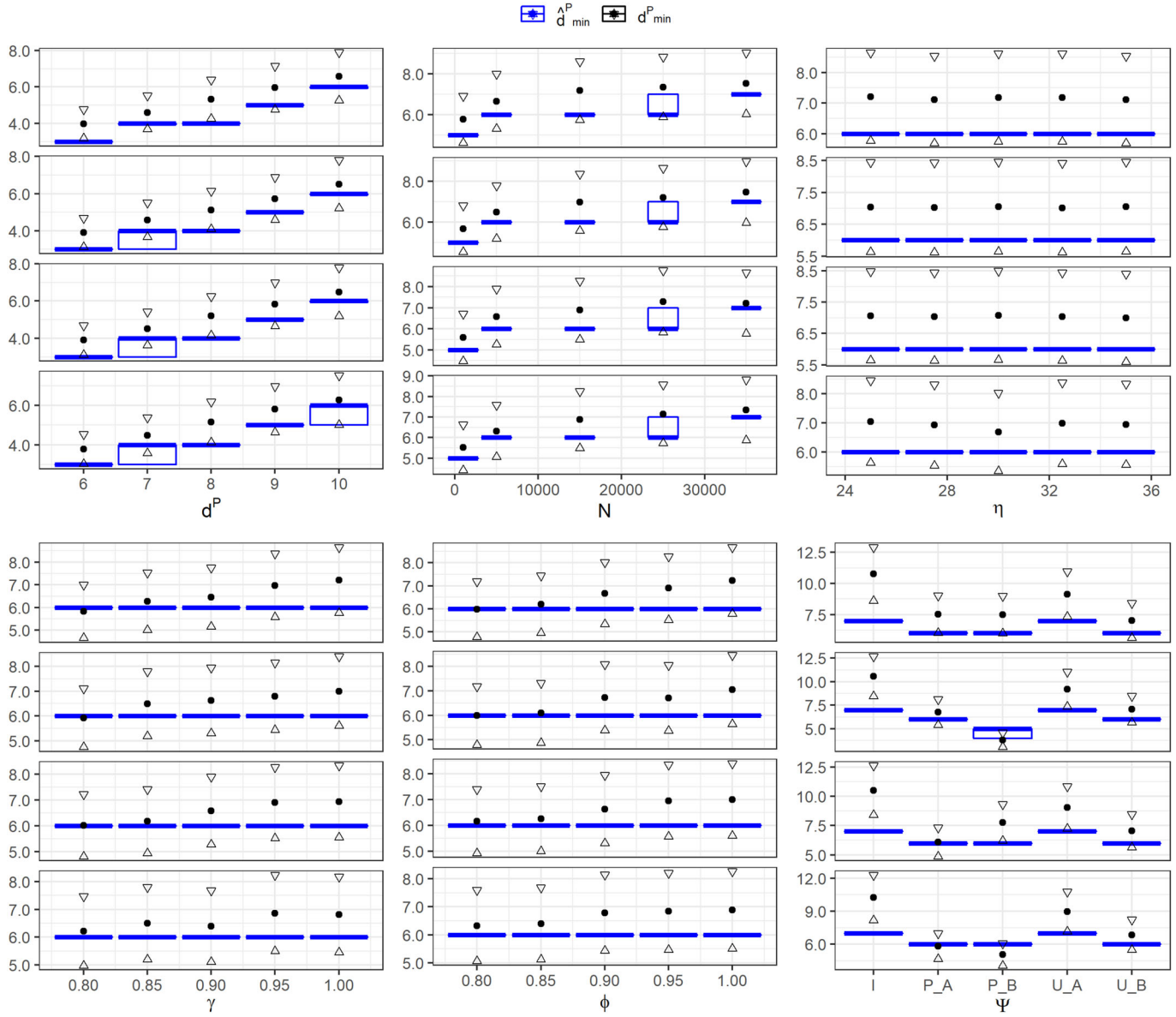

Figure 1. (Sparse networks) $\hat{d}^{\mathbf{A}}_{\min}$ versus $d^{\mathbf{P}}_{\min}$ for networks with $d^{\mathbf{P}} = 3.5\sqrt{\log N}$ under the settings listed in Table 4. Each panel shows four subplots corresponding to $K \in \{3(\text{top}), 7, 10, 25\}$. Blue boxes denote {5%, 50%, 95%}-percentiles of $\hat{d}_{\min}(\mathbf{P})$ based on 20 replications and $\{0.8d^{\mathbf{P}}_{\min}, d^{\mathbf{P}}_{\min}, 1.2d^{\mathbf{P}}_{\min}\}$ are shown in black.

Remark 1. The high probability statement in Theorem 3.3 comes from the hypothesis of Lemma 3.2 that $\zeta = \omega(\sqrt{Nb_{\max}})$, and the conditions (3) and $\delta \in (0, 3/2)$. In order to have $\zeta = \omega(\sqrt{Nb_{\max}})$ and (3), one must also have $\zeta = \Theta(\beta) = \omega(\sqrt{Nb_{\max}})$. This implies the second assertion of Theorem 3.3 involving (4) and (5). In the rest of this article, we denote the above interval in (3) by "the oracle ζ interval" or simply "the oracle interval" when the context is clear. We call (3) the "oracle interval" since the interval is based on population parameters and thus is not observable. In Section 4, we propose an empirical ζ estimator that lies within this oracle interval with high probability[6].

Note that since the concentration result given in Lemma 3.2 holds for the sparse regime $Nb_{\max} = o(\log N)$, the oracle interval appearing in (3) can be nonempty for the same regime $Nb_{\max} = o(\log N)$ as well.

We also provide a sufficient condition on λ for the existence of the oracle interval (3). The existence of this interval depends on whether the radicand in (3) is positive and in turn, guarantees detectability of K with a high probability for any ζ chosen from (3). Thus, we obtain a sufficient condition involving the

[6]It is worth noting that a sufficient condition for the detection of K is that the interval appearing in (3) is nonnegative, that is, $\beta^2 + 4 \geq 4d^{\mathbf{P}}_{\min}$. Solving this inequality by substituting in the expression for β and rearranging, one can obtain the expression $(b_{\max} - b_{\min})^2 + \Theta(1) > 8(b_{\max} + b_{\min})$. Although this expression appears on the surface similar to the threshold for weak recovery in the related problem domain of community detection, one should use caution in making such comparisons. While our objective in this article is to consistently estimate K as the degree tends to infinity at a rate $\omega(1)$ and $o(\log(N))$, the goal in weak recovery in community detection is to detect community labels in the constant degree setting better than a certain accuracy level.

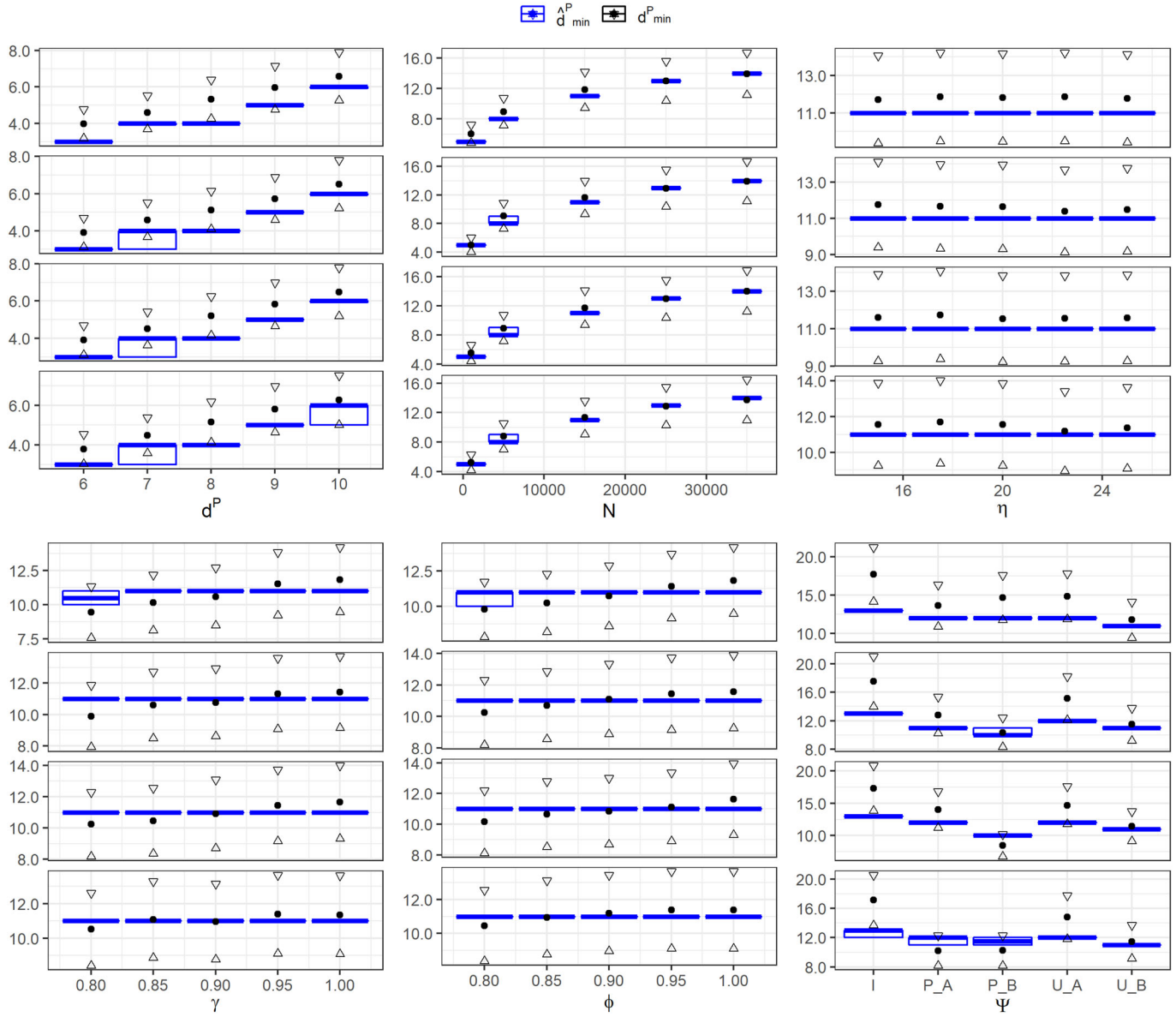

Figure 2. (Dense networks) $\hat{d}^{\mathbf{P}}_{\min}$ versus $d^{\mathbf{P}}_{\min}$ for networks with $d^{\mathbf{P}} = 0.193(\log N)^2$ under the settings listed in Table 4. Each panel shows four subplots corresponding to $K \in \{3(\text{top}), 7, 10, 25\}$. Blue boxes denote {5%, 50%, 95%}-percentiles of $\hat{d}^{\mathbf{P}}_{\min}$ based on 20 replications and $\{0.8d^{\mathbf{P}}_{\min}, d^{\mathbf{P}}_{\min}, 1.2d^{\mathbf{P}}_{\min}\}$ are shown in black.

parameters of the DCBM that generates the input network for detecting K correctly.

Corollary 3.4. In the setup of Theorem 3.3, with high probability, K can be detected if the following holds:

$$\lambda = \omega\left(\frac{K}{\sqrt{d^{\mathbf{P}}_{\min}}}\right). \tag{6}$$

Remark. An equivalent statement to (6) is $K = o(\lambda\sqrt{d^{\mathbf{P}}_{\min}})$.

Note that (6) is a *sufficient*, not *necessary*, condition. We discuss this point in light of our empirical studies in Section 7. Nonetheless, these conditions provide intuition about the situations when K can be estimated in the sparse regime $Nb_{\max} = o(\log N)$.

4. Empirical Method for Estimation of K

In this section, we propose an approach for empirically estimating a ζ value that lies within the oracle interval (3). In order to get the empirical ζ estimate, we proceed in two steps.

First, we propose an estimator $\hat{d}^{\mathbf{P}}_{\min}$ for the minimum expected degree $d^{\mathbf{P}}_{\min}$ in the following lemma using the quantile function based on the empirical distribution of the degrees defined as follows.

Definition 4.1. The empirical cumulative distribution function F of degrees $\{d^{\mathbf{A}}_i\}_i$ is defined as

$$F(y) := \frac{1}{N}\sum_{i=1}^{N} d^{\mathbf{A}}_i \tag{7}$$

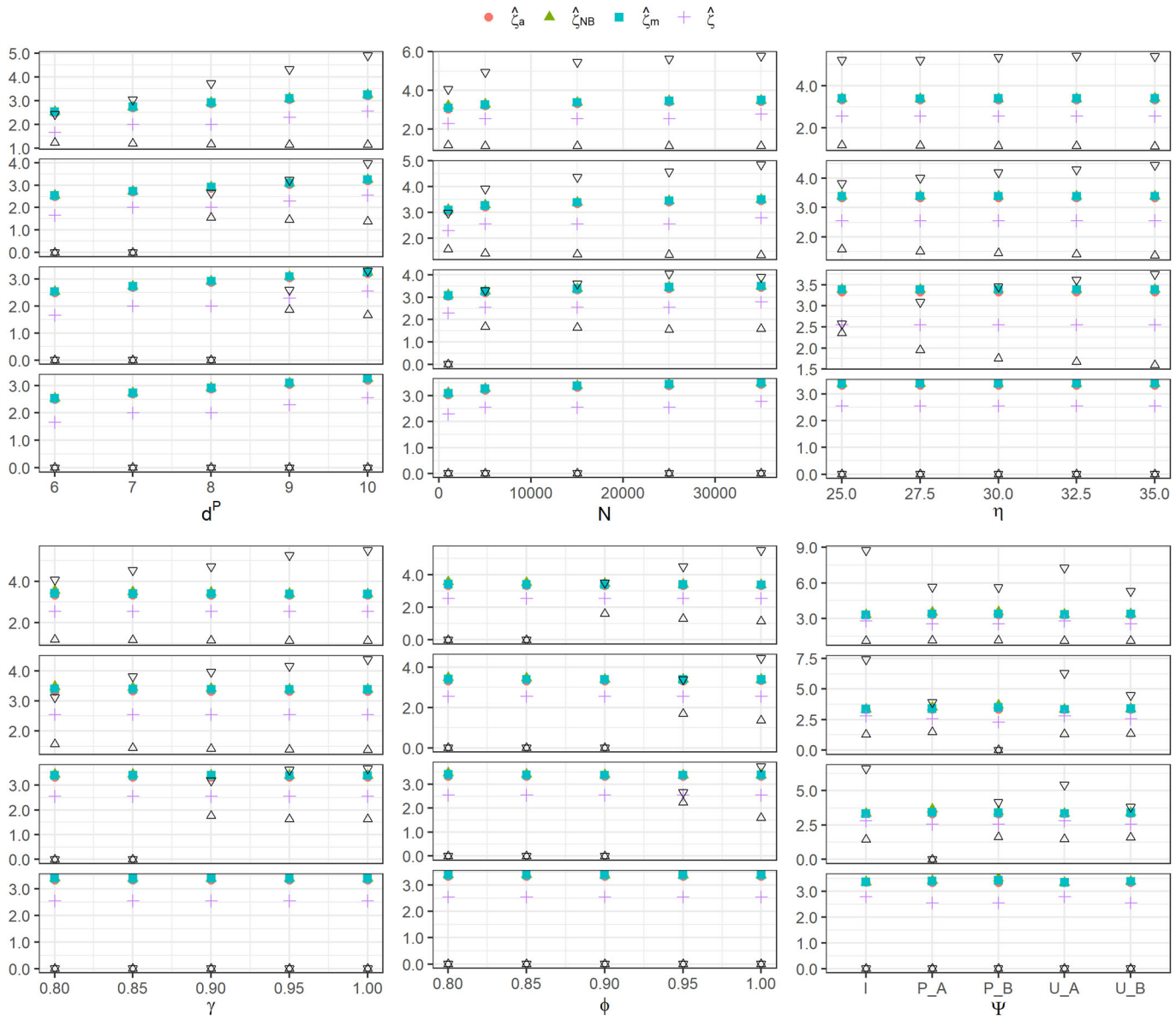

Figure 3. (Sparse networks) Oracle ζ intervals (black triangles) versus ζ estimated pursuant to alternative Bethe Hessian-based methods under Simulation Settings in Table 4. Endpoints of the oracle interval are denoted by triangles. Each estimated ζ is denoted by a colored dot representing the median of 20 replications. $\hat{\zeta}$ was computed with $c = 0.3$ and $\hat{\epsilon}_\delta = 0.2$ (see Lemma 4.3). Each panel shows four subplots corresponding to $K \in \{3(\text{top}), 7, 10, 25\}$.

Definition 4.2. Given the empirical cumulative distribution function F defined in (7), the quantile function $Q : [0, 1] \to \mathbb{Z}^+$ is defined as

$$Q(x) := \inf\{y \in \mathbb{R}^+ : x \leq F(y)\} \tag{8}$$

Lemma 4.3. Under the framework of Theorem 3.3 and given $f(N) = o(\log N)$, let $\hat{d}^{\mathbf{P}}_{\min} = \max\left\{d^{\mathbf{A}}_i | d^{\mathbf{A}}_i < Q(\frac{c}{f(N)}), i \in [N]\right\}$, for some $c \in (0, f(N))$ and quantile function Q as defined in (8). Then, with probability at least $1 - 2\delta$,

$$\hat{d}^{\mathbf{P}}_{\min} \in \left((1 - \varepsilon_\delta)d^{\mathbf{P}}_{\min}, (1 + \varepsilon_\delta)d^{\mathbf{P}}_{\min}\right) \tag{9}$$

for any

$$\varepsilon_\delta \geq \sqrt{\frac{-2\log(c\delta/f(N))}{d^{\mathbf{P}}_{\min}}}. \tag{10}$$

Remark 1. In Lemma 4.3, the quantity δ can be arbitrarily small depending on the concentration bound of the estimate $\hat{d}^{\mathbf{P}}_{\min}$ to its population counterpart $d^{\mathbf{P}}_{\min}$. So, the estimate $\hat{d}^{\mathbf{P}}_{\min}$ can be close to $d^{\mathbf{P}}_{\min}$ with high probability depending on the extent of their proximity.

Remark 2. The inequality (10) says that the lower bound for ϵ_δ increases as $c\delta/f(N)$ gets smaller. That is to say, in addition to the constant c being sufficiently large, $f(N)$ also can be taken to be arbitrarily small to obtain an estimate $\hat{d}^{\mathbf{P}}_{\min}$ that is sufficiently close to $d^{\mathbf{P}}_{\min}$ with high probability. Essentially, we need to take $f(N) = \omega(1)$ and $d^{\mathbf{P}}_{\min} = \omega(\log(f(N))$ to have $\epsilon_\delta = o(1)$.

Next, we use $\hat{d}^{\mathbf{P}}_{\min}$ to estimate the oracle interval (3) in Theorem 3.3 which, if it exists, provides a sufficient condition for the detection of K with high probability. To do this, we choose the

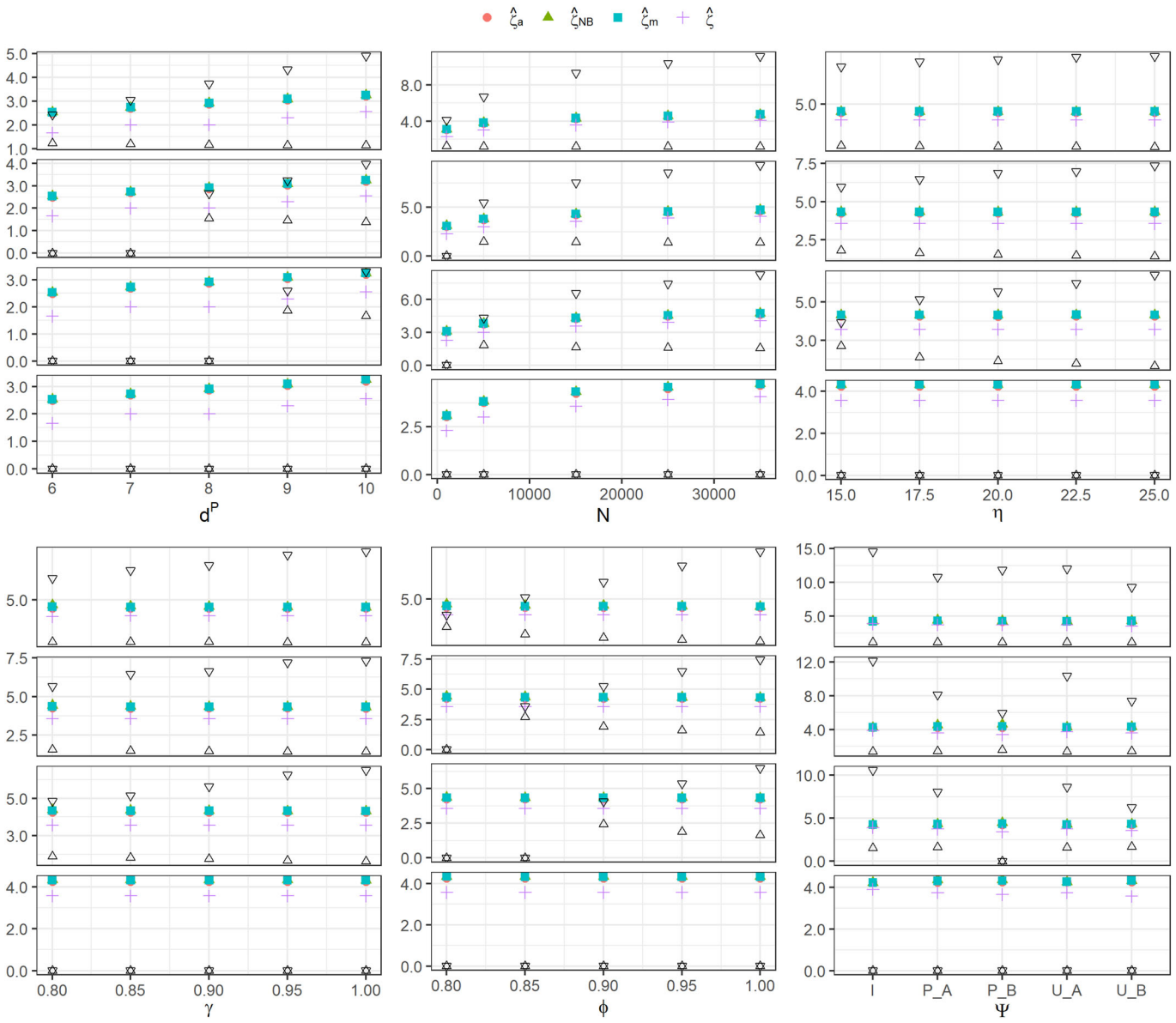

Figure 4. (Dense networks) Oracle ζ intervals (black triangles) versus ζ estimated pursuant to alternative Bethe Hessian-based methods under Simulation Settings in Table 4. Endpoints of the oracle interval are denoted by triangles. Each estimated ζ is denoted by a colored dot representing the median of 20 replications. $\hat{\zeta}$ was computed with $c = 0.3$ and $\hat{\epsilon}_\delta = 0.2$ (see Lemma 4.3). Each panel shows four subplots corresponding to $K \in \{3(\text{top}), 7, 10, 25\}$.

estimated mid-point of the oracle interval in (3) as the empirical estimate for ζ, denoted $\hat{\zeta}$.

The mid-point of the oracle interval in (3) is given by $\frac{\beta}{2}$. We provide $\frac{\hat{\beta}}{2} = \sqrt{\hat{d}^{\mathbf{P}}_{\min} - 1 + \hat{\epsilon}_\delta \hat{d}^{\mathbf{P}}_{\min}/(1-\hat{\epsilon}_\delta)}$ as an estimate of $\frac{\beta}{2}$ in Step 3 of Algorithm 4.1. The proof of Theorem 4.4 provides a rigorous justification that $\frac{\hat{\beta}}{2}$ lies in the oracle interval around $\frac{\beta}{2}$ with high probability. The Bethe Hessian $\mathbf{H}_\zeta$ is constructed with $\hat{\zeta}$ and $\hat{\mathbf{K}}$ is simply the number of the negative eigenvalues of $\mathbf{H}_{\hat{\zeta}}$. The steps delineated above are summarized in Algorithm 4.1.

Remark. There are two tuning parameters in Algorithm 4.1: c and $\hat{\epsilon}_\delta$, both of which emanate from Lemma 4.3. A larger c in Step 2 leads to a larger $\hat{d}_{\min}(\mathbf{A}_0)$ and $\hat{\epsilon}_\delta$ in Step 3 remedies the bias in its estimation of $d^{\mathbf{P}}_{\min}$ by increasing ζ. See Section 5.2.1 for empirical findings on tuning these parameters. Hereinafter, we denote the ζ parameter obtained in Step 3 in Algorithm 4.1 by $\hat{\zeta}$ and Algorithm 4.1 by $\mathbf{BH}_\zeta$.

Finally, we end this section with the main summarizing result on the efficacy of estimator $\hat{K}$ from Algorithm $\mathbf{BH}_\zeta$.

Theorem 4.4. Under the framework of Theorem 3.3, with probability at least $1 - \delta$ as defined in Lemma 4.3, $\hat{\mathbf{K}}$ obtained from Algorithm 4.1 is the true $\mathbf{K}$.

The proof of Theorem 4.4 is given in supplement C.

5. Simulations

We first illustrate the existence of the oracle ζ intervals and compare it to the ζ interval $[\zeta_L, \zeta_U]$ proposed in (Dall'Amico,

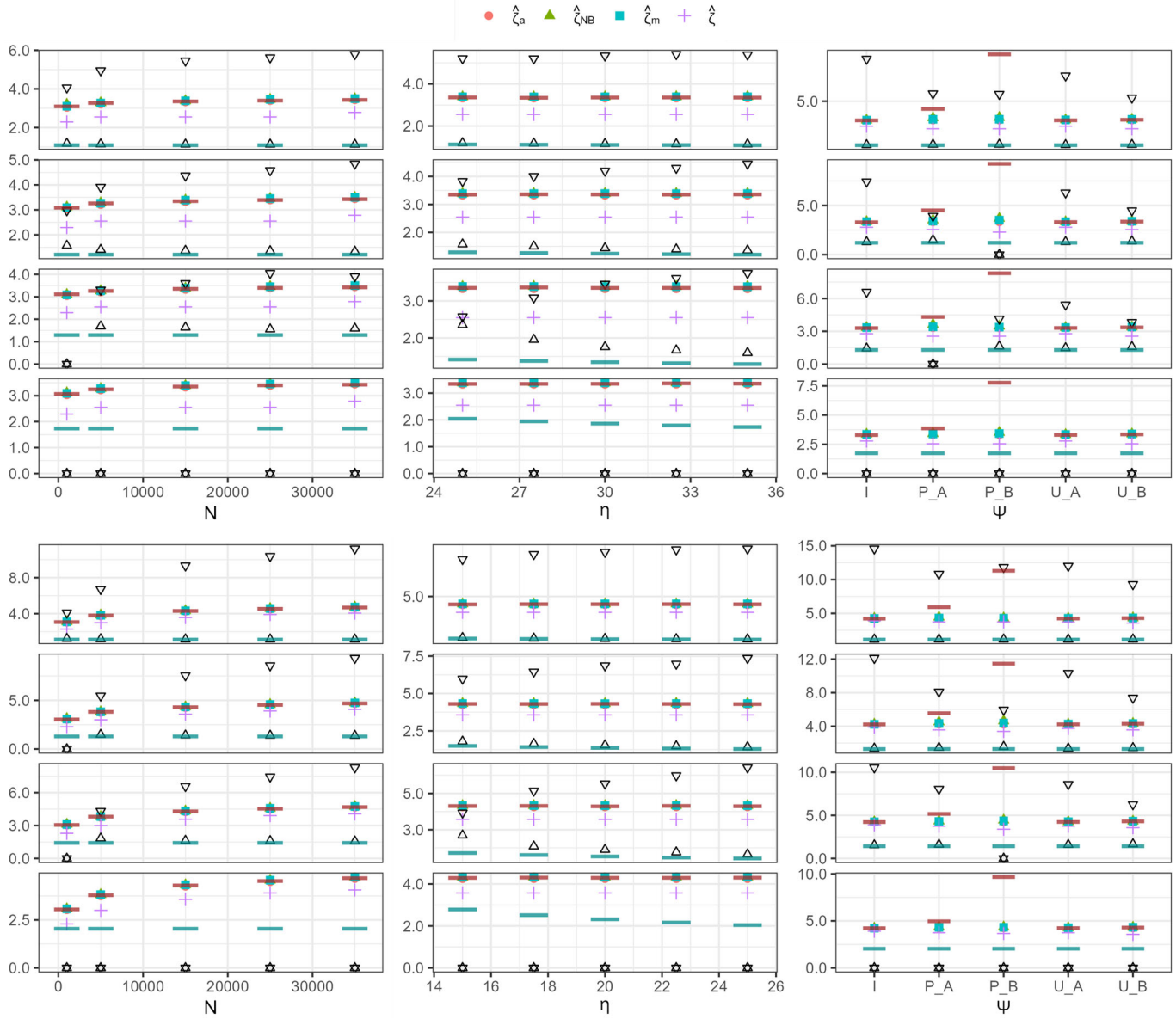

Figure 5. (Up: Sparse networks; Down: Dense networks) Oracle ζ intervals (black triangles), ζ_U (brown solid lines), and ζ_L (cyan solid lines) versus ζ estimated pursuant to alternative Bethe Hessian-based methods under Simulation Settings 2, 3, and 6 in Table 4. Endpoints of the oracle interval are denoted by triangles. Each estimated ζ is denoted by a colored dot representing the median of 20 replications. $\hat{\zeta}$ was computed with $c = 0.3$ and $\hat{\epsilon}_\delta = 0.2$ (see Lemma 4.3). Each panel shows four subplots corresponding to $K \in \{3(\text{top}), 7, 10, 25\}$.

Couillet, and Tremblay 2021) (see definitions in Table 2). We also compare the effectiveness of $\hat{\zeta}$ computed in Algorithm 4.1 with $\hat{\zeta}_a$, $\hat{\zeta}_m$, and $\hat{\zeta}_{\mathbf{NB}}$ in aligning the number of negative eigenvalues with K. Then, we evaluate the performance of $\mathbf{BH}_{\hat{\zeta}}$ in estimating K versus other Bethe Hessian-based methods and other approaches discussed in Section 2.4 and summarized in Table 3. Lastly, in Section 6, we apply our approach to 15 real-world datasets. Notations and definitions are summarized in Table 2.

5.1. Network Data Generation

Network data were generated under the DCBM in (2.2) and operationalized as follows:

$$\mathbf{Z} \sim \text{Mult}(1; (\frac{\gamma}{K}, \ldots, \frac{\gamma}{K} + (i-1)\frac{2(1-\gamma)}{K(K-1)}, \ldots, \frac{2-\gamma}{K}))$$

$$\mathbf{B} := \rho\mathbf{B}_0 := \rho[b(\mathbf{1}_K\mathbf{1}_K^T - \mathbf{I}_K) + \text{diag}(\phi\eta b, \ldots, \phi\eta b + (i-1)\frac{2(1-\phi)\eta b}{K-1}, \ldots, (2-\phi)\eta b)_K]$$

where γ determines the extent of the imbalance in relative community sizes, ρ controls the overall degree density of the network, ϕ determines the variation in within-community degree densities, and η sets the in-out ratio of degrees. η is monotonically related to λ and thus controls the community structure signal strength. Like, under the case $\phi = 1$, that is we have no variations among the within community probabilities, $\lambda = 1 - \eta^{-1}$. Each model parameter was varied incrementally in simulating the adjacency matrices as shown in Table 4.

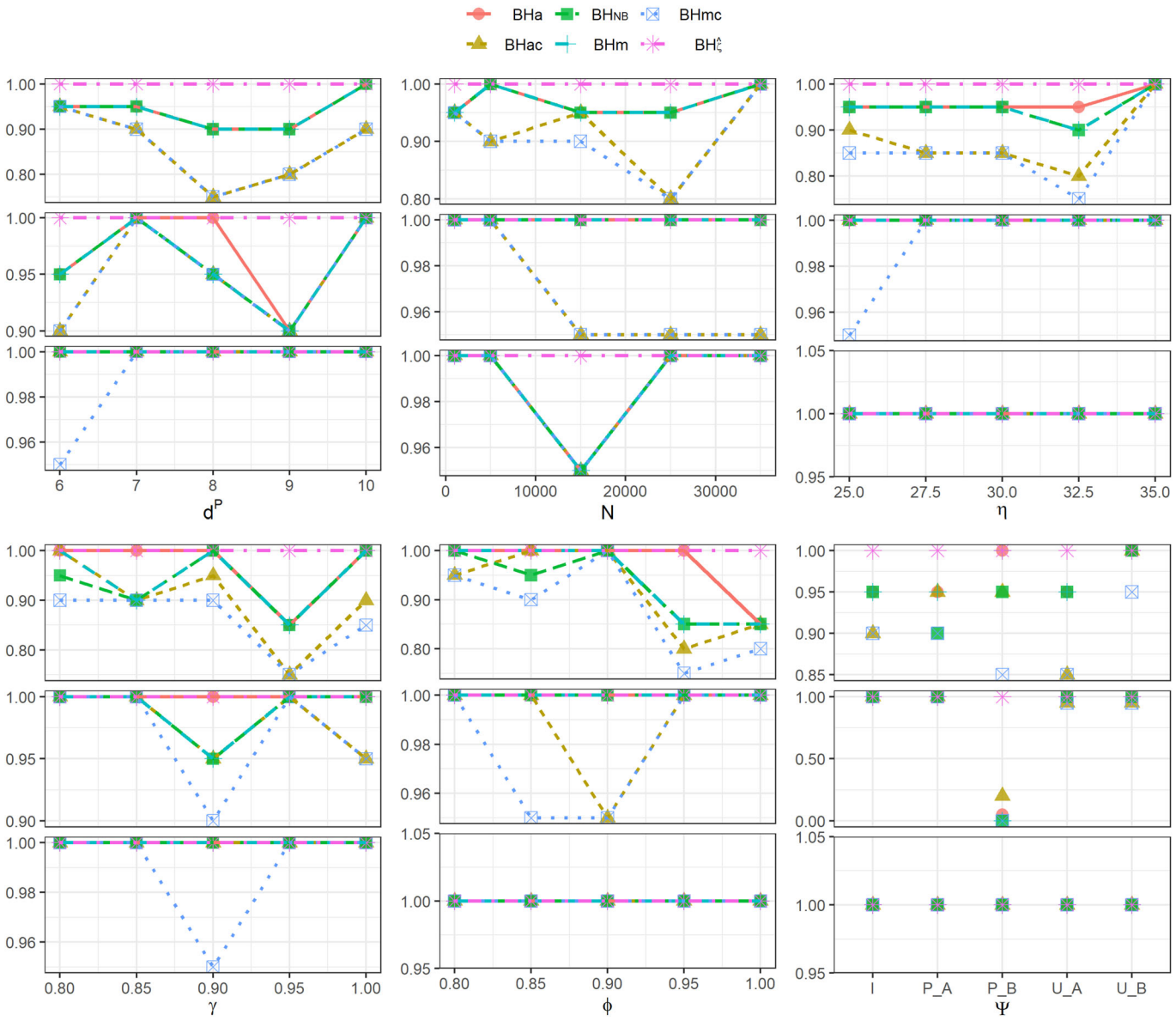

Figure 6. (Sparse networks) Accuracy of $\hat{K}$ computed pursuant to alternative Bethe Hessian-based methods under Simulation Settings in Table 4. $\mathbf{BH}_{\hat{\zeta}}$ was computed with $c = 0.3$ and $\hat{\epsilon}_\delta = 0.2$ (see Lemma 4.3). Each panel shows three subplots corresponding to $K \in \{3(\text{top}), 7, 10\}$.

Let $d^{\mathbf{P}} \in \{6, 7, 8, 9, 10\}$ for Setting 1 and $d^{\mathbf{P}} \in \{3.5\sqrt{\log N}, 0.193(\log N)^2\}$ for Settings 2 through 7, where the specific constants were chosen such that $d^{\mathbf{P}} = 9.2$ for $N = 1000$. For Settings 1 through 6, η is increased by 10 when $d^{\mathbf{P}} = 3.5\sqrt{\log N}$ since a stronger signal is needed to recover K in more sparse networks. The parameter values in Pareto(0.58, 10) and Pareto(1.16, 10) were chosen to effect the same variance as Unif(0.75, 1) and Unif(0.5, 1), respectively. Each setting was simulated with 20 replications.

5.2. Results

5.2.1. Hyperparameter Tuning

Figures 1 and 2 show the performance of $\hat{d}^{\mathbf{P}}_{\min}$ in estimating $d^{\mathbf{P}}_{\min}$ in networks of varying levels of sparsity. Networks have $d^{\mathbf{P}} = 3.5\sqrt{\log N}$ in Figure 1 and $d^{\mathbf{P}} = 0.193(\log N)^2$ in Figure 2. $\hat{d}_{\min}(\mathbf{P})$ was computed using Lemma 4.3 with $c = 0.3$ and $\hat{\epsilon}_\delta = 0.2$ for the reasons we discuss next. Observe that while $\hat{d}^{\mathbf{A}}_{\min}$ generally lies in the interval $[0.8d^{\mathbf{P}}_{\min}, 1.2d^{\mathbf{P}}_{\min}]$, it tends to underestimate when: (a) the network is sparse; (b) the degree distribution is uniform; and (c) community sizes and densities are more balanced, that is, γ and ϕ are closer to 1. In our empirical analyses, $c = 0.3$ worked well in the estimation of K regardless of the density regime. For better results, one can consider adjusting c higher (lower resp.) for networks in more sparse (dense resp.) regimes and have homogeneous (heterogeneous resp.) degree distributions. Letting $\hat{\epsilon}_\delta = 0.2$ seemed to allow $\hat{d}_{\min}(\mathbf{P})$ enough flexibility to closely approximate $d^{\mathbf{P}}_{\min}$. Results based on other values of the hyperparameters are presented in supplement Section D.

5.2.2. Existence of Oracle ζ intervals

The sufficient conditions stated in (4) and (5) for detecting K in Corollary 3.4 depends on the following:

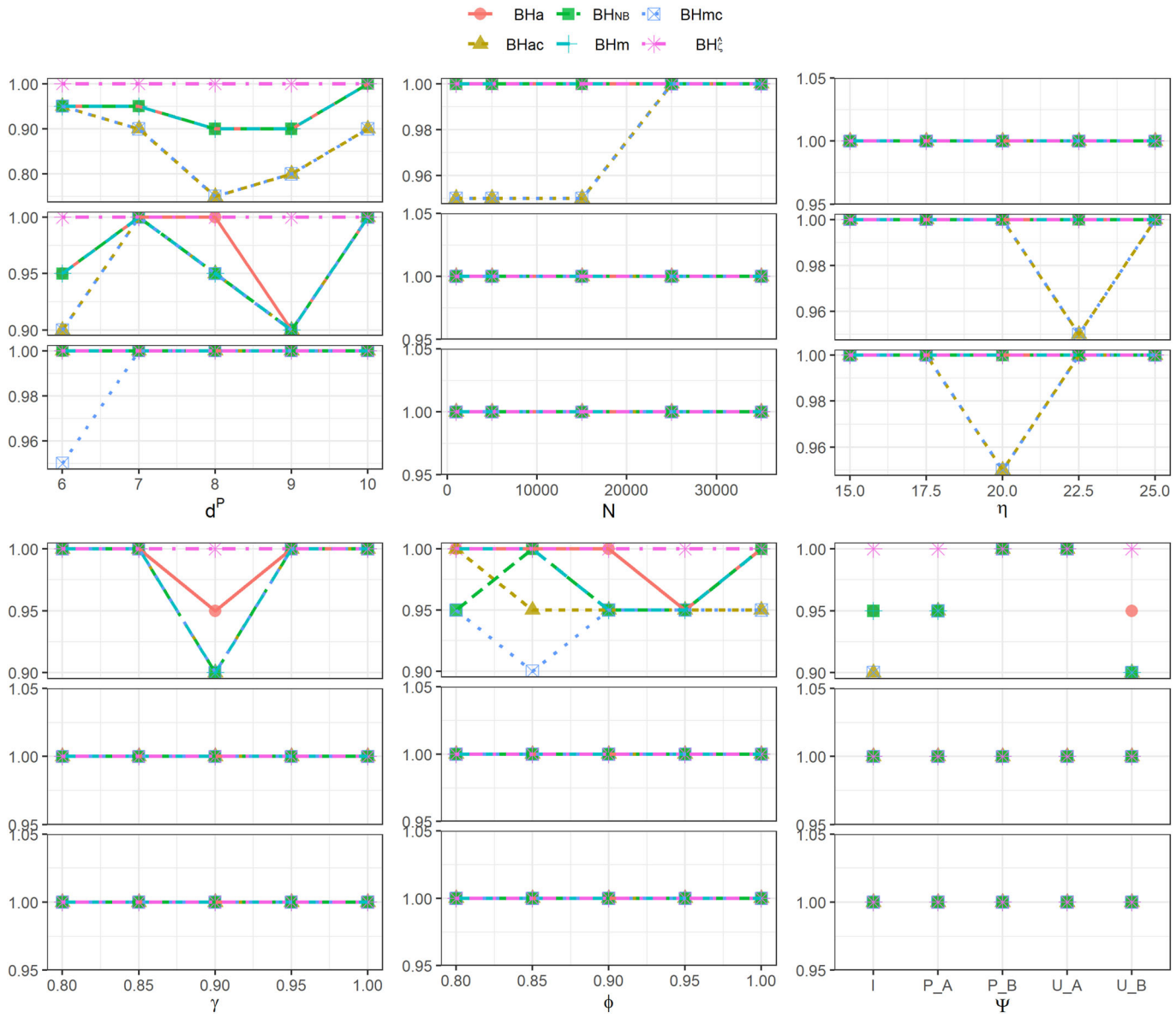

Figure 7. (Dense networks) Accuracy of $\hat{K}$ computed pursuant to alternative Bethe Hessian-based methods under Simulation Settings in Table 4. $\mathbf{BH}_{\hat{\zeta}}$ was computed with $c = 0.3$ and $\hat{\epsilon}_\delta = 0.2$ (see Lemma 4.3). Each panel shows three subplots corresponding to $K \in \{3(\text{top}), 7, 10\}$.

1. Sparsity: $Nb_{\max}$ and $d^{\mathbf{P}}_{\min}$
2. Signal strength (in/out ratio): λ
3. Number of communities: K
4. Network and relative community sizes, and degree heterogeneity: N and N_K

We investigate the effect of each on the size of oracle intervals whose positive value indicates the existence of the interval. Figure 3 shows oracle intervals for sparse networks of expected average degree $d^{\mathbf{P}} = 3.5\sqrt{\log N}$ and Figure 4 for denser networks with $d^{\mathbf{P}} = 0.193(\log N)^2$.

The length of the oracle intervals tends to increase in networks that are (a) denser (higher average degree); (b) larger (greater N); (c) fewer communities (lower K); (d) more assortative (higher in/out ratio); (e) more balanced in either community sizes or community sparsity; and (f) less heterogeneity or less skewness in the degree distribution.

5.2.3. Performance Comparison

In this section, we use simulated networks to assess the performance of Algorithm 4.1 in relation to alternative approaches. First, we demonstrate that whenever the oracle interval exists, $\hat{\zeta}$ lies inside the interval while $\hat{\zeta}_a$, $\hat{\zeta}_m$, and $\hat{\zeta}_{\mathbf{NB}}$ fall outside the interval in networks that are sparse, imbalanced, or lacking signal strength. $\hat{\zeta}$ also lies within the boundaries $[\zeta_L, \zeta_U]$ proposed in Dall'Amico, Couillet, and Tremblay (2021). Second, we show empirically that with $\hat{\zeta}$, Algorithm 4.1 outputs $\hat{K}$ that is more accurate than those output by other methods.

Figures 3 and 4 depict $\hat{\zeta}$ versus $\hat{\zeta}_a$, $\hat{\zeta}_m$, and $\hat{\zeta}_{\mathbf{NB}}$. Networks are generated according to simulation settings in Table 4. Figure 3 is based on sparse networks with $d^{\mathbf{P}} = 3.5\sqrt{\log N}$ and Figure 4 is based on more dense networks with $d^{\mathbf{P}} = 0.193(\log N)^2$. Figure 5 compares ζ_L and ζ_U proposed in Dall'Amico, Couillet, and Tremblay (2021) to our oracle intervals and ζ estimates based on Settings 2, 3, and 6. Figures 6 and 7 show accuracy

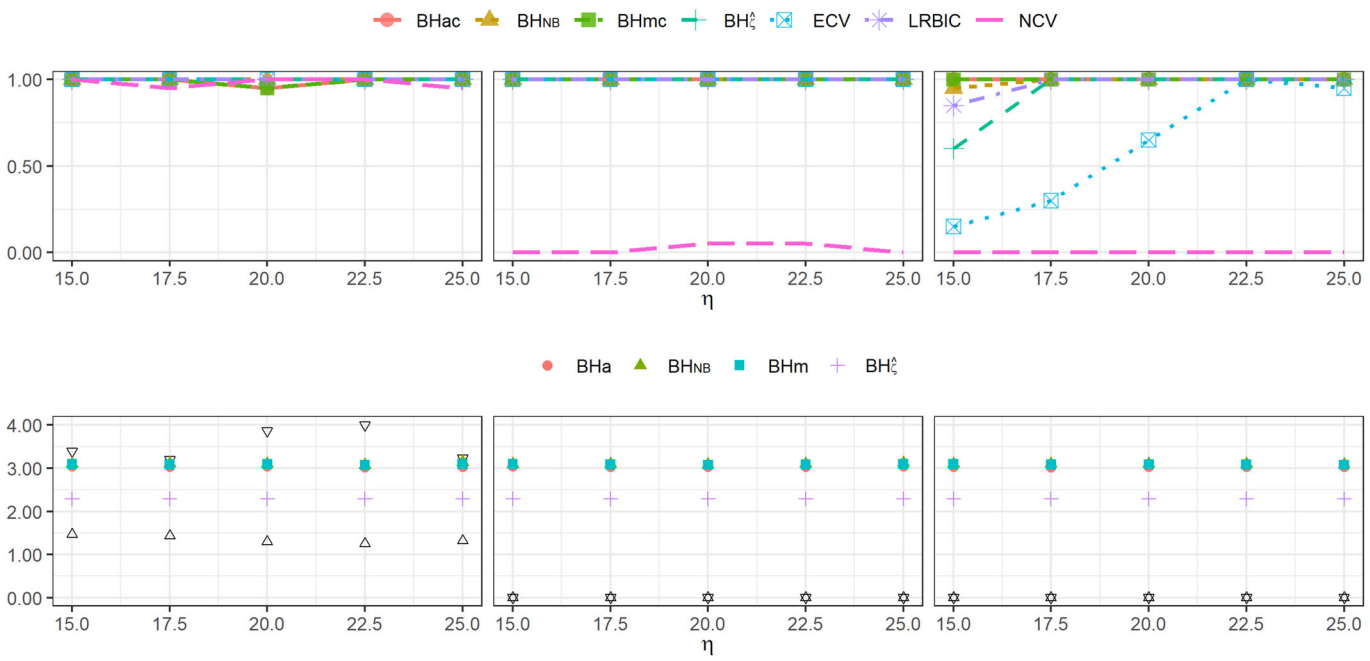

Figure 8. (Top Row) Accuracy of $\hat{K}$ computed pursuant to alternative methods under Setting 7 in Table 4. (Bottom Row) $\hat{K}$ using **BHa, BHm,** and $\mathbf{BH}_{\hat{\zeta}}$ with $K = 3$. In both rows, $\mathbf{BH}_{\hat{\zeta}}$ was computed with $c = 0.3$ and $\hat{\epsilon}_\delta = 0.2$ (see Lemma 4.3). (Bottom Row) Oracle ζ intervals (black triangles) versus ζ estimated pursuant to alternative Bethe Hessian-based methods with $K = 3$. In both rows, $\mathbf{BH}_{\hat{\zeta}}$ was computed with $c = 0.3$ and $\hat{\epsilon}_\delta = 0.2$ (see Lemma 4.3).

Algorithm 4.1 $\mathbf{BH}_\zeta$: Empirical K-estimation

Input: Adjacency matrix $\mathbf{A}$, threshold parameter c, upscaling parameter $\hat{\epsilon_\delta}$, function $f(N) = o(\log N)$

Output: Estimated number of communities $\hat{\mathbf{K}}$

1: Obtain $\mathbf{A}_0$ by removing zero-degree nodes from $\mathbf{A}$

2: $\hat{d}^{\mathbf{P}}_{\min}(\mathbf{A}_0) = \max\{d_i(\mathbf{A}_0) | d_i(\mathbf{A}_0) < Q(\frac{c}{f(N)}), i \in [N]\}$, where $c \in (0, f(N))$ and Q is the quantile function as defined in (8)

3: Choose $\hat{\zeta} = \sqrt{\hat{d}^{\mathbf{P}}_{\min}(\mathbf{A}_0) - 1 + \hat{\varepsilon}_\delta \hat{d}^{\mathbf{P}}_{\min}(\mathbf{A}_0)/(1 - \hat{\varepsilon}_\delta)}$ where $\hat{\varepsilon}_\delta$ can be computed with $\hat{d}^{\mathbf{P}}_{\min}(\mathbf{A}_0)$ in (10)

4: Compute $\mathbf{H}_{\hat{\zeta}} := (\hat{\zeta}^2 - 1)\mathbf{I}_N + \mathbf{D}_1^{\mathbf{A}_0} - \hat{\zeta}\mathbf{A}_0$ where $\mathbf{D}_1^{\mathbf{A}_0}$ is the diagonal degree matrix of $\mathbf{A}_0$

5: Perform eigen-decomposition of $\mathbf{H}_{\hat{\zeta}}$ and let $\hat{\mathbf{K}}$ be the number of negative eigenvalues of $\mathbf{H}_{\hat{\zeta}}$

6: $\hat{\mathbf{K}}$ is the estimate of $\mathbf{K}$.

rates of $\hat{K}$ defined as the proportion of the number of correct estimates from 20 network replicates. The networks considered here are the same as those in Figures 3 and 4 with the exception of $K = 25$, which was excluded since the oracle ζ intervals did not exist. Figure 8 compares the performance of $\mathbf{BH}_{\hat{\zeta}}$ with other methods based on 20 replications of smaller, sparse networks ($N = 1000, d^{\mathbf{P}} = 9.2$).

We make a few observations regarding the performance of the Algorithm 4.1 from Figures 3–8:

1. In Figures 3 and 4, we observe that when the oracle interval does exist, $\hat{\zeta}$ from Algorithm 4.1 lies inside the interval and closer to the center of the interval while $\hat{\zeta}_a$, $\hat{\zeta}_m$, and $\hat{\zeta}_{\mathbf{NB}}$ are larger and lie outside the interval in networks that are made more challenging by more imbalance in community sizes, lower signal strength, and higher sparsity. This is illustrated, for example, in the subplot in row 3 column 3 in Figure 3 corresponding to sparse networks of balanced community sizes and densities with $N = 15{,}000$, $K = 10$, $\eta = 25$, and uniformly distributed degrees.

   Lemma B.2 in the supplement provides an intuition for why $\hat{\zeta}$ is closer to the center of the oracle interval. According to the Lemma, the quantity $\sqrt{d^{\mathbf{P}}_{\min}\lambda_K((\mathbf{D}_1^{\mathbf{A}'})^{-1/2}\mathbf{A}'(\mathbf{D}_1^{\mathbf{A}'}))}$ (where $\mathbf{A}'$ is a regularized version of $\mathbf{A}$ and $\mathbf{D}_1^{\mathbf{A}'}$ its diagonal degree matrix) lies between the center and the upper bound of the oracle interval under more signal-rich conditions (i.e., fewer communities, more assortative structure, or more balanced in either community sizes or community sparsity) compared to the conditions for the existence of the interval. For instance, consider the network corresponding to the subplot in row 3 column 3 in Figure 3 where low assortativity due to the small η makes it a challenging network. Since $\hat{\zeta}_a^2 = d^{\mathbf{A}} \approx \hat{\zeta}_m^2 = \frac{\sum d_i^2}{\sum d_i} - 1 \approx Nb_{\max} \geq \lambda_1(\mathbf{A}') \geq d^{\mathbf{P}}_{\min}\lambda_K((\mathbf{D}_1^{\mathbf{A}'})^{-1/2}\mathbf{A}'(\mathbf{D}_1^{\mathbf{A}'}))$, it is more likely for $\hat{\zeta}_a$, $\hat{\zeta}_m$, and $\hat{\zeta}_{\mathbf{NB}}$ to be farther away from the center of the interval, or even outside of it, compared to $\hat{\zeta}$. In general, $\hat{\zeta}_a$, $\hat{\zeta}_m$, and $\hat{\zeta}_{\mathbf{NB}}$ are not ideal choices for ζ in challenging networks.
2. It is evident in Figure 5 that, for networks of diverse settings of size, signal strength, and degree distribution, $\hat{\zeta}$ also lies within the boundaries $[\zeta_L, \zeta_U]$ proposed in Dall'Amico, Couillet, and Tremblay (2021). It can also be observed that, for easy settings where the ζ oracle interval exists, ζ_L is close to the lower bound of the oracle interval, while ζ_U is smaller than the upper bound of the oracle interval. Dall'Amico, Couillet, and Tremblay (2021) also provided estimators for ζ_L and

Table 5. *K*-estimation in real-world network data.

Dataset	Nodes	Edges	c	$\bar{d}$	Ground truth* K	$\hat{K}$
polblogs	1500	19,000	0.8	12.8	2 (Chen et al. 2016)	2
Political books	105	441	0.3	8.4	3 (Chen and Lei 2018)	3
TerroristRel	881	8600	0.3	9.9	21 (Yang et al. 2022)	21
fb-CMU-Carnegie49	6637	249,967	0.2	75.3	3 (Rossi and Ahmed 2015)	4
Citation3Core	707	3285	0.5	9.3	15 (Li et al. 2020)	15
Email-Enron	36,692	183,831	0.5	10.0	267 (Chen et al. 2016)	68
karate	34	78	1.5	4.6	2 (Zachary 1977)	2
Red Hot Jazz Archive	198	2742	0.08	55.4	4 (Jazz-Archive 2002)	4
NCAA Football	115	613	0.9	10.7	12 (Newman and Girvan 2004)	7
Santa Fe Coauthorship	271	200	0.95	3.4	3 (Girvan and Newman 2002)	3
Dolphins	62	159	0.95	5.1	2 (Lusseau et al. 2003)	2
Les Miserable	77	254	0.95	6.6	6 (Newman and Reinert 2016)	4
Statistics Coauthorship	635	1204	0.95	3.8	7 (Ji and Jin 2016)	7
C. elegans	453	4585	0.35	20.2	10 (Duch and Arenas 2005)	10
Email-Tarragona	1133	5449	0.7	9.6	13 (Duch and Arenas 2005)	13

NOTE: *Ground truth" values are given if known; otherwise, estimated values are reported in the literature. $\hat{\epsilon_\delta}$ was set to 0.05 for all datasets. Adjusting $\hat{\epsilon_\delta}$ to 0.194 for TerroristRel and 0.3 for fb-CMU-Carnegie49 improves the number of accurate estimates to 12. See supplement section E for detailed discussions of each dataset and the reference sources and explanations for ground truth K.

ζ_U. Comparisons with their estimators for ζ_L and ζ_U are presented in supplement section D.

3. One can readily see in Figures 6 and 7 that $\mathbf{BH}_{\hat{\zeta}}$ outperforms alternative methods, especially when K is small and even when the sufficiency condition in Corollary 3.4 is not satisfied. An example is the case with $K = 10$ and $d^{\mathbf{P}} \in \{6, 7, 8\}$ demonstrated in the subplots in row 3 column 1 in both figures.
4. Figure 8 demonstrates that while $\mathbf{BH}_{\hat{\zeta}}$ generally outperforms the other methods, it can underperform when the oracle interval does not exist due to, for example, K being too large and η too small. Nevertheless, we note that only $\mathbf{BH}_{\hat{\zeta}}$ is shown to work in sparse networks while enjoying computational efficiency. In contrast, **LRBIC, NCV,** and **ECV** are computationally expensive especially for large networks.

6. Real-World Network Applications

We test the efficacy of our algorithm on 15 real-world networks discussed in the literature. Our results and summary numbers of the datasets are presented in Table 5. Ground truth values of K are shown if known; otherwise, estimated values cited in the literature are shown along with citations. The values used for the hyperparameter c in $\mathbf{BH}_{\hat{\zeta}}$ and $\hat{\epsilon_\delta}$ are also shown. In 10 of the 15 networks, our algorithm correctly detects K that is consistent with either the ground truth or what is reported in the literature. The number of accurate estimates increased to 12 after making adjustments to $\hat{\epsilon_\delta}$ for two of the datasets. We note that our algorithm correctly detects K even in cases where the ground truth K is higher than the sufficiency threshold in Corollary 3.4. Examining the remaining four networks for which our algorithm did not detect the ground truth K, we note that K is significantly larger than the sufficiency threshold in Corollary 3.4 compared to the other 11 networks. For instance, the "Email-Enron" has an average degree of only 10, implying a sufficient condition K value of less than $\sqrt{10}$. However, its "ground truth K" per Chen et al. (2016) is much larger at 267. We refer to supplement section E for a detailed discussion of each dataset.

7. Discussion

Despite the effectiveness and efficiency of the Bethe Hessian-based methods for estimating K, the potential for their widespread adoption has been hampered by the uncertainty around the parameter ζ. Although several heuristics for picking a value for ζ have been proposed with some empirical support and principles from statistical physics, the question of precisely what values of ζ allow for K detection with rigorous theoretical guarantees for correctness has remained open. In this regard, our main contribution in this article is a precise, theoretically rigorous characterization of the interval of values from which ζ can be chosen to ensure with a high probability that true K is detected by using the corresponding Bethe Hessian matrix. We provide an algorithm to empirically estimate this interval and the parameter $d^{\mathbf{P}}_{\min}$ that is needed in the interval calculation and prove their correctness and consistency. Both extensive simulation studies and test applications to real-world networks attest to the efficacy of our method.

We note that we have not addressed the *necessary* condition for detecting K in this article, rather we have focused on finding some *sufficient* conditions. In other words, the lack of existence of the oracle interval does not preclude the detectability of K. Based on the success of our algorithm in detecting K in networks not meeting the sufficiency condition in Corollary 3.4, we conjecture that the oracle interval we presented in this article can be widened. The extent to which it can be widened would depend on the necessary condition for K detection, and we believe it to be a promising future research topic.

Supplementary Materials

The supplementary information contains Appendices A–E, R code files related to the paper, and the real world network datasets used in the paper. Appendices A–C contain the supplementary theoretical results. Appendix D contains additional simulation results. Appendix E contains the details of the real-world network data analysis. In the code files, README.html and README.rmd files provide the workflow details. Separately, simulation.html and simulation.rmd provide the workflow of the simulation study, and realdata_analysis.html and realdata_analysis.rmd provide the workflow of the real-life network data analysis.

Acknowledgments

The authors would also like to thank Peter Bickel, Liza Levina, and the anonymous reviewers for helpful discussions and meaningful suggestions.

Funding

S. Chatterjee has been partially supported by NSF-DMS grant # 2154564 and the PSC-CUNY Enhanced Research Grant # 62781-00 50 for this work. N. Hwang was partially supported by the Rich Internship awarded by the Department of Mathematics, City College of New York, CUNY, in the Summer of 2020. S. Bhattacharyya has been partially supported by USDA-NIFA grant # 20226701538059.

References

Abbe, E. (2017), "Community Detection and Stochastic Block Models: Recent Developments," *The Journal of Machine Learning Research*, 18, 6446–6531. [1]

Adamic, L. A., and Glance, N. (2005), "The Political Blogosphere and the 2004 US Election: Divided They Blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 36–43. [1]

Bhattacharyya, S., and Bickel, P. J. (2015), "Subsampling Bootstrap of Count Features of Networks," *The Annals of Statistics*, 43, 2384–2411. [2]

Bhattacharyya, S., and Chatterjee, S. (2020), "Consistent Recovery of Communities from Sparse Multi-Relational Networks: A Scalable Algorithm with Optimal Recovery Conditions," in *Complex Networks XI*, pp. 92–103, Springer. [3]

Bickel, P. J., Chen, A., and Levina, E. (2011), "The Method of Moments and Degree Distributions for Network Models," *The Annals of Statistics*, 39, 2280–2301. [1]

Bordenave, C., Lelarge, M., and Massoulié, L. (2015), "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. [1,2]

Cerqueira, A., and Leonardi, F. (2018), "Strong Consistency of Krichevsky-Trofimov Estimator for the Number of Communities in the Stochastic Block Model," arXiv preprint arXiv:1804.03509. [2]

Chen, K., and Lei, J. (2018), "Network Cross-validation for Determining the Number of Communities in Network Data," *Journal of the American Statistical Association*, 113, 241–251. [2,4,5,14]

Chen, P.-Y., and Hero, A. O. (2018), "Phase Transitions and a Model Order Selection Criterion for Spectral Graph Clustering," *IEEE Transactions on Signal Processing*, 66, 3407–3420. [2]

Chen, Y., Zhao, P., Li, P., Zhang, K., and Zhang, J. (2016), "Finding Communities by Their Centers. *Scientific Reports*, 6, 1–8. [14]

Dall'Amico, L., Couillet, R., and Tremblay, N. (2019), "Revisiting the Bethe-hessian: Improved Community Detection in Sparse Heterogeneous Graphs," in *Advances in Neural Information Processing Systems*, pp. 4037–4047. [2,5]

Dall'Amico, L., Couillet, R., and Tremblay, N. (2020), "Community Detection in Sparse Time-Evolving Graphs with a Dynamical Bethe-hessian," in *Advances in Neural Information Processing Systems* (Vol. 33), pp. 7486–7497. [2]

Dall'Amico, L., Couillet, R., and Tremblay, N. (2021), "A Unified Framework for Spectral Clustering in Sparse Graphs," *Journal of Machine Learning Research*, 22, 9859–9914. [4,10,12,13]

De Las Rivas, J., and Fontanillo, C. (2010), "Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks," *PLoS Computational Biology*, 6, e1000807. [1]

Duch, J., and Arenas, A. (2005), "Community Detection in Complex Networks Using Extremal Optimization," *Physical Review E*, 72, 027104. [14]

Emmert-Streib, F., Dehmer, M., and Haibe-Kains, B. (2014), "Gene Regulatory Networks and their Applications: Understanding Biological and Medical Problems in Terms of Networks," *Frontiers in Cell and Developmental Biology*, 2, 38. [1]

Faloutsos, M., Karagiannis, T., and Moon, S. (2010), "Online Social Networks," *IEEE Network*, 24, 4–5. [1]

Fortunato, S. (2010), "Community Detection in Graphs," *Physics Reports*, 486, 75–174. [1]

Friston, K. J. (2011), "Functional and Effective Connectivity: A Review," *Brain Connectivity*, 1, 13–36. [1]

Girvan, M., and Newman, M. E. (2002), "Community Structure in Social and Biological Networks," *Proceedings of the National Academy of Sciences*, 99, 7821–7826. [14]

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010), "A Survey of Statistical Network Models," *Foundations and Trends in Machine Learning*, 2, 129–233. [1]

Gulikers, L., Lelarge, M., and Massoulié, L. (2016), "Non-Backtracking Spectrum of Degree-Corrected Stochastic Block Models," arXiv preprint arXiv:1609.02487. [1,2]

Horn, R. A., and Johnson, C. R. (2012), *Matrix Analysis*, Cambridge: Cambridge University Press. [5]

Hu, J., Qin, H., Yan, T., and Zhao, Y. (2019), "Corrected Bayesian Information Criterion for Stochastic Block Models," *Journal of the American Statistical Association*, 115, 1771–1783. [1,2]

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. (2011), "Identifying Important Places in People's Lives from Cellular Network Data," in *International conference on Pervasive Computing*, pp. 133–151. Springer. [1]

Jazz-Archive (2002), "The Red Hot Jazz Archive." Available at *https://snap.stanford.edu/data/com-Amazon.html*. [14]

Ji, P., and Jin, J. (2016), "Coauthorship and Citation Networks for Statisticians," *The Annals of Applied Statistics*, 10, 1779–1812. [14]

Jin, J., Ke, Z. T., Luo, S., and Wang, M. (2022), "Optimal Estimation of the Number of Network Communities," *Journal of the American Statistical Association*, 1–16. [2]

Kolaczyk, E. D., and Csárdi, G. (2014), *Statistical Analysis of Network Data with R* (Vol. 65), New York: Springer. [1]

Le, C. M., and Levina, E. (2015), "Estimating the Number of Communities in Networks by Spectral Methods," arXiv preprint arXiv:1507.00827. [1,2,3,4]

Lehmann, S., Lautrup, B., and Jackson, A. D. (2003), "Citation Networks in High Energy Physics," *Physical Review E*, 68, 026113. [1]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [3]

Li, T., Lei, L., Bhattacharyya, S., Van den Berge, K., Sarkar, P., Bickel, P. J., and Levina, E. (2020), "Hierarchical Community Detection by Recursive Partitioning," *Journal of the American Statistical Association*, 117, 951–968. [14]

Li, T., Levina, E., and Zhu, J. (2020), "Network Cross-Validation by Edge Sampling," *Biometrika*, 107, 257–276. [2,4,5]

Liben-Nowell, D., and Kleinberg, J. (2007), "The Link-Prediction Problem for Social Networks," *Journal of the American Society for Information Science and Technology*, 58, 1019–1031. [1]

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003), "The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations," *Behavioral Ecology and Sociobiology*, 54, 396–405. [14]

Ma, S., Su, L., and Zhang, Y. (2021), "Determining the Number of Communities in Degree-Corrected Stochastic Block Models," *Journal of Machine Learning Research*, 22, 1–63. [1,2,5]

Newman, M. E., and Girvan, M. (2004), "Finding and Evaluating Community Structure in Networks," *Physical Review E*, 69, 026113. [14]

Newman, M. E., and Reinert, G. (2016), "Estimating the Number of Communities in a Network," *Physical Review Letters*, 117, 078301. [14]

Pagani, G. A., and Aiello, M. (2013), "The Power Grid as a Complex Network: A Survey," *Physica A: Statistical Mechanics and its Applications*, 392, 2688–2700. [1]

Reis, B. Y., Kohane, I. S., and Mandl, K. D. (2007), "An Epidemiological Network Model for Disease Outbreak Detection," *PLoS Medicine*, 4, e210. [1]

Riolo, M. A., Cantwell, G. T., Reinert, G., and Newman, M. E. (2017), "Efficient Method for Estimating the Number of Communities in a Network," *Physical Review E*, 96, 032310. [1,2]

Rossi, R. A., and Ahmed, N. K. (2015), "The Network Data Repository with Interactive Graph Analytics and Visualization," in *AAAI*. Available at *https://networkrepository.com*. [14]

Rousseeuw, P. J. (1987), "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, 20, 53–65. [1]

Rubinov, M., and Sporns, O. (2010), "Complex Network Measures of Brain Connectivity: Uses and Interpretations," *Neuroimage*, 52, 1059–1069. [1]

Saade, A., Krzakala, F., and Zdeborová, L. (2014), "Spectral Clustering of Graphs with the Bethe Hessian," *Advances in Neural Information Processing Systems*, 27, 406–414. [2,3]

Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *Journal of the Royal Statistical Society*, Series B, 63, 411–423. [1]

Von Luxburg, U. (2010), "Clustering Stability: An Overview," *Foundations and Trends® in Machine Learning*, 2, 235–274. [1]

Wang, Y. R., and Bickel, P. J. (2017), "Likelihood-based Model Selection for Stochastic Block Models," *The Annals of Statistics*, 45, 500–528. [1,4]

Yan, B., Sarkar, P., and Cheng, X. (2018), "Provable Estimation of the Number of Blocks in Block Models," in *International Conference on Artificial Intelligence and Statistics*, pp. 1185–1194. [1,2]

Yang, Y., Shi, P., Wang, Y., and He, K. (2022), "Quadratic Optimization based Clique Expansion for Overlapping Community Detection," *Knowledge-Based Systems*, 247, 108760. [14]

Zachary, W. W. (1977), "An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33, 452–473. [14]