| Statistica Sinica Preprint No: SS-2024-0305 | |
|---|---|
| **Title** | Inference of Community Numbers in Partial Networks |
| **Manuscript ID** | SS-2024-0305 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202024.0305 |
| **Complete List of Authors** | Xiyue Zhu,<br>Xiao Han and<br>Qing Yang |
| **Corresponding Authors** | Qing Yang |
| **E-mails** | yangq@ustc.edu.cn |
| Notice: Accepted author version. | |

# Inference of community numbers in partial networks

Xiyue Zhu*, Xiao Han* and Qing Yang†

*University of Science and Technology of China*

*Abstract:* Despite considerable efforts towards community detection from a comprehensive network perspective, the investigation of this problem becomes less explored when individuals only have access to their local view. In this paper, we propose an approach for testing and estimating the unknown number of communities $K$ in the global network model, but only with limited partial information available. Our procedure constructs a test statistic based on singular values and eigenvalues of partitioned matrices derived from a centered and rescaled partial adjacency matrix. We establish the asymptotic null distribution for testing and demonstrate consistency in estimating $K$ using results from random matrix theory. The effectiveness and usefulness of our proposed method are demonstrated through extensive simulations, including both directed and undirected graphs, as well as several real data examples.

*Key words and phrases:* Partial network, community structure, Tracy-Widom distribution.

*Co-first authors.

†Corresponding author. Email: yangq@ustc.edu.cn.

## 1. Introduction

Detecting the community memberships plays a critical role in the field of social network analysis. Communities allow the network to be naturally divided into subgroups with specific characteristics, which have been utilized in various applications, including aiding in product recommendations from retailers, assisting parties in conjecturing voter preferences and exploring genes with similar functional groups. A significant amount of research has been devoted to the community membership detection. Girvan and Newman (2002) put forward the Girvan-Newman hierarchical clustering method, where the cluster network nodes are constructed by repeatedly deleting the edge with the highest betweenness. Von Luxburg (2007) utilized the first few leading eigenvectors of the adjacency matrix or the graph Laplacian to classify by k-means clustering method. Rohe et al. (2011) presented the first high-dimensional clustering model that allows the number of clusters to grow with the number of nodes. More spectral based approaches can be found in Jin (2015), Sarkar and Bickel (2015), Chaudhuri et al. (2012), and Krzakala et al. (2013), among others.

The stochastic block model (SBM, Holland et al., 1983) is widely recognized as the most popular model for community structures, requiring only community membership vectors and a community probability matrix

to generate network models. Its variants, including the degree-corrected stochastic block model (DCSBM, Karrer and Newman, 2011) and the degree corrected mixed membership stochastic block model (DCMM, Airoldi et al., 2008), are also powerful tools with practical applications in various fields such as social science, biology, and information science. However, one common assumption in these models is that the number of communities $K$ is known a priori; yet in practice this value is often unknown and must be estimated through substantial efforts documented in literature. Examples include cross validation method (Chen and Lei, 2018; Li et al., 2020) and Bayesian inference techniques (Hofman and Wiggins, 2008; McDaid et al., 2013; Riolo et al., 2017). Likelihood-based methods have also gained popularity in estimating $K$, with notable contributions from Latouche et al. (2012), Saldana et al. (2017) and Wang and Bickel (2017), among others. Several attempts on using hypothesis testing have been made to automatically determine $K$ in Bickel and Sarkar (2016) and Banerjee and Ma (2017). Sequential testing frameworks for estimating of $K$ have been provided in Lei (2016) and Jin et al. (2023). Additionally, methods based on the spectrum of the Bethe Hessian matrices (Le and Levina, 2015; Hwang et al., 2024) have attracted attention for their computational efficiency. More recently, Ren et al. (2023) proposed a novel regularization term based on the popular

network embedding model, while Han et al. (2023) introduced a universal rank inference method via subsampling the residual matrix.

Despite extensive exploration, these works are based on a complete network. However, in reality, collecting a complete network can be challenging or costly and may raise privacy concerns. For example, the Aggregated Relational Data (ARD, Alidaee et al., 2020; Breza et al., 2023) captures information pertaining to a social network through the utilization of respondent queries in the format of "How many people with trait X do you know?" It is worth noting that non-responses are prevalent, thereby limiting our ability to obtain a comprehensive overview. In previous literature, partial networks are often associated with egocentric sampling (Freeman, 1982; Banerjee et al., 2013; Li et al., 2023) and respondent-driven sampling (Salganik and Heckathorn, 2004; Rohe, 2019; Yan et al., 2020). This paper, however, employs a different partial information framework that centers on individuals, as proposed by Han et al. (2024). This framework considers a particular individual with a path length of $L = 2$, meaning the network only contains information on nodes that are up to two steps away from the individual in question. In analogy to the global adjacency matrix, it features a partial adjacency matrix $\mathbf{B}$. Our research aims to estimate and draw inferences on the community numbers $K$ in a global SBM setup, solely utilizing

the perceived partial network $\mathbf{B}$. To achieve this, we first test whether an ensemble of stochastic block models with $K_0$ communities can suitably fit the partial adjacency matrix $\mathbf{B}$ by conducting the following hypothesis test:

$$H_0 : K = K_0 \quad \text{against} \quad H_1 : K > K_0. \tag{1.1}$$

Our proposed test statistic is made up of the sum of singular values and eigenvalues of partitioned matrices obtained from a centered and rescaled partial adjacency matrix. Making use of techniques in random matrix theory, we derive the asymptotic null distributions (Theorem 1) and provide asymptotic power guarantee under the alternative model (Theorem 3). Then by sequentially testing the hypothesis (1.1) for $K_0 = 1, \cdots, K_{\max}$ with some large enough positive integer $K_{\max}$, we can estimate $K$ as the first integer that makes our test fail to reject. The estimation algorithm is elaborated in Algorithm 1. The aforementioned theoretical developments lay the foundation for establishing the algorithm's consistency, as summarized in Theorem 4.

Previous studies, such as Bickel and Sarkar (2016) and Lei (2016), have suggested similar test statistics by leveraging the extreme eigenvalues of full adjacency matrices. However, our approach differs from these works as it relies on two submatrices of $\mathbf{B}$ along with their corresponding singular value and eigenvalue. Consequently, apart from analyzing the extreme

eigenvalues, we also need to establish the asymptotic distribution of the largest singular value and investigate joint distributions between the singular value and the extreme eigenvalues. Additionally, according to our procedure, we need to address scenarios where the dimensions of the submatrices are random, which depend on entries of the partial adjacency matrix $\mathcal{B}$.

The remaining sections of the paper are organized as follows: Section 2 presents the setup and our approach, while Section 3 establishes its asymptotic properties. In Sections 4 and 5, we demonstrate the effectiveness of our method through simulations and real data analysis. All proofs are relegated to the Supplementary Material.

**Notations.** We introduce some notations to be used throughout the paper. Let $K$ denote the true number of communities in the network, and $K_0$ its target counterpart. For a matrix $\mathbf{B} = (B_{ij})$, we write its largest and smallest eigenvalue values as $\lambda_1(\mathbf{B})$ and $\lambda_n(\mathbf{B})$ respectively and the largest singular value as $\sigma_1(\mathbf{B})$. The notation $\xrightarrow{d}$ means convergence in distribution. We use $|\cdot|$ to indicate the cardinality of a set. If two positive sequences $x_n$ and $y_n$ satisfy $\limsup_{n\to\infty}(x_n/y_n) < \infty$, we write $x_n = O(y_n)$. If $x_n = O(y_n)$ and $y_n = O(x_n)$, we write $x_n \asymp y_n$. While the notation $x_n = o(y_n)$ means $\limsup_{n\to\infty}(x_n/y_n) = 0$. Similarly, for a sequence of random variables $X_n$ and constants $a_n$, we write $X_n = O_p(a_n)$ if $X_n/a_n$ is stochastically bounded,

and $X_n = o_p(a_n)$ if $\lim_{n\to\infty} P\left(|X_n/a_n| \leq \varepsilon\right) = 1$ for any positive $\varepsilon$.

## 2. Methodology

### 2.1 Setup: adjacency matrix and partial adjacency matrix

A fully observed network with $(n+1)$ nodes is typically represented in the form of a symmetric adjacency matrix $\mathcal{A} = (a_{ij})_{(n+1)\times(n+1)}$, where 1 and 0 represent the presence and absence of pairwise interactions among $(n+1)$ individuals. Its expectation $\mathbb{E}\mathcal{A}$, in a SBM, can be expressed as

$$\mathbb{E}\mathcal{A} = \mathbf{\Pi}\mathbf{P}\mathbf{\Pi}^T,$$

where $\mathbf{P} = (P_{kl})$ is a $K \times K$ nonsingular symmetric matrix and each entry $P_{kl}$ represents the connection probability between communities $k$ and $l$, and $\mathbf{\Pi} = (\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_{n+1})^T \in \mathbb{R}^{(n+1)\times K}$ denotes the matrix of community membership vectors. Each $\boldsymbol{\pi}_i \in \{\mathbf{e}_1, ..., \mathbf{e}_K\}$ with $\mathbf{e}_k$ being a unit vector whose $k$th component is one and all others are zero. In this paper, our focus lies on networks without selfloops. Analysis of networks with selfloops can be conducted similarly by expressing the noise matrix as $(\mathcal{A} - \mathbb{E}\mathcal{A} + \text{diag}(\mathcal{A}))$.

We are interested in the partial network centers on an individual O with a path length of $L = 2$, as initially established in Han et al. (2024). This implies that we solely observe the connectivity patterns within the

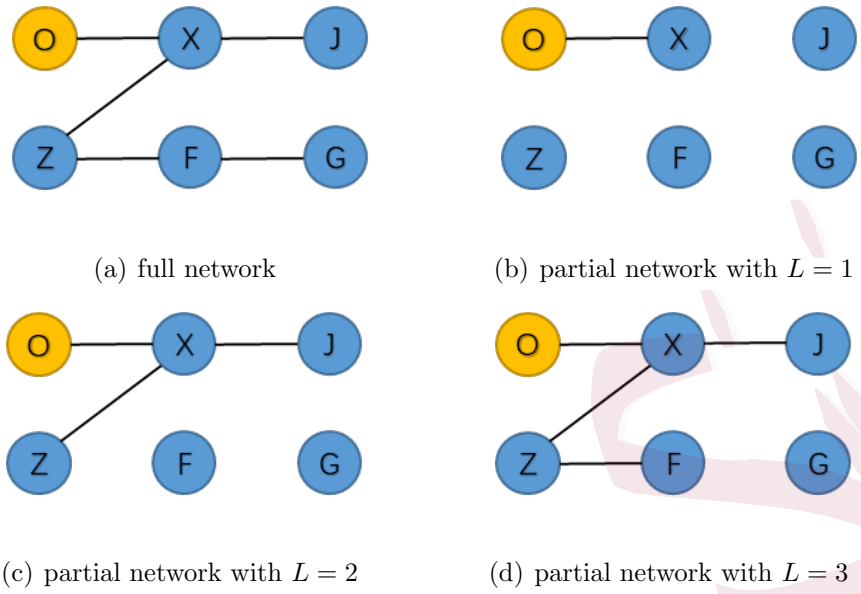
nodes that are up to two steps away from O. Figure 1 presents a toy example centered on the individual of interest, O, illustrating the network at various knowledge depths: (a) the full network, (b) partial network with $L = 1$, (c) $L = 2$, and (d) $L = 3$. Notably, Figure 1(c) demonstrates that connections beyond $L = 2$ are unobservable, while O's direct link to X ($L = 1$) and X's connections to J and Z ($L = 2$) remain visible. All connectivity among Z, F and G lies outside the specific knowledge depth. Without loss of generally, let O be the first node. The network is then recorded in a partial adjacency matrix $\mathcal{B} = (b_{ij})_{(n+1)\times(n+1)}$, which can be calculated by $b_{ij} = a_{ij}(1 - \mathbb{I}(a_{1i} = 0)\mathbb{I}(a_{1j} = 0))$. It follows that

$$\mathcal{B} = -\mathcal{S}\mathcal{A}\mathcal{S} + \mathcal{A}\mathcal{S} + \mathcal{S}\mathcal{A}, \tag{2.2}$$

where $\mathcal{S} = \text{diag}(a_{11}, ..., a_{1(n+1)})$ is a diagonal matrix formed from the first row of $\mathcal{A}$, representing the connectivity between the specific individual O and all other nodes.

## 2.2 Motivation and transformation of the adjacency matrices

We rewrite the first row of $\mathcal{A}$ (or $\mathcal{B}$) to $(a_{11}, \underbrace{1, \ldots, 1}_{n_1}, 0, \ldots, 0)$, where $n_1 = \sum_{i=2}^{n+1} a_{1i} = \sum_{i=2}^{n+1} b_{1i}$, and rearrange matrices $\mathcal{A}$ and $\mathcal{B}$ accordingly. By removing the first rows and columns of matrices $\mathcal{A}$ and $\mathcal{B}$, we obtain $n \times n$ matrices $\mathbf{A} = (A_{ij})_{n\times n}$ and $\mathbf{B} = (B_{ij})_{n\times n}$, thereby eliminating the inter-




(a) full network (b) partial network with $L = 1$

(c) partial network with $L = 2$ (d) partial network with $L = 3$

Figure 1: *Black lines represent connections and O denotes the individual of interest. (a) shows the full network, (b) shows the partial network with* $L = 1$*, (c)* $L = 2$*, (d)* $L = 3$*.*

connections between the specific individual and others. The motivation for such transformation stems from two aspects. Firstly, we observe that a section of the partial adjacency matrix does not contribute to our problem. To gain insights, let $\mathbf{S} = \begin{pmatrix} \mathbf{I}_{n_1} & \mathbf{0}_{n_1 \times m} \\ \mathbf{0}_{m \times n_1} & \mathbf{0}_{m \times m} \end{pmatrix}$ with $m = n - n_1$, a block matrix with the first diagonal block being a $n_1 \times n_1$ square matrix. Subsequently, throughout this paper, we consistently adhere to the same block structure as $\mathbf{S}$, unless otherwise specified. Denote the block matrices $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$

and $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$. By utilizing the relationship between full and partial adjacency matrix $B_{ij} = A_{ij}(1 - \mathbb{I}(A_{1i} = 0)\mathbb{I}(A_{1j} = 0))$, it follows that

$$\begin{aligned} \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} &= -\mathbf{SAS} + \mathbf{AS} + \mathbf{SA} \\ &= -\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &\quad + \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{0} \end{pmatrix}. \end{aligned} \tag{2.3}$$

In other words, $\mathbf{B}_{22} = \mathbf{0}$ contains no relevant information. Furthermore, the removal of interconnections involving the first node results in conditional independence between the remaining two blocks $\mathbf{B}_{11}$ and $\mathbf{B}_{12}$. This independence enables us to construct asymptotically independent test statistics using them, as demonstrated later in Theorem 2.

## 2.3 Test statistic and estimation algorithm

According to Section 2.2, the block matrices $\mathbf{B}_{11}$ and $\mathbf{B}_{12}$ obtained from the observed partial adjacency matrix provide valuable information for the

hypothesis testing (1.1). A natural approach to utilize this information is by subtracting the signal from the partial adjacency matrix, thereby yielding a residual matrix. To this end, we consider

$$\widetilde{B}^*_{ij} = \begin{cases} \dfrac{B_{ij} - M_{ij}}{\sqrt{(n_1 - 1)M_{ij}(1 - M_{ij})}} & i \neq j \text{ and } \min\{i, j\} \leq n_1, \\ 0 & i = j \text{ or } \min\{i, j\} > n_1, \end{cases} \tag{2.4}$$

where $M_{ij} = \mathbb{E}B_{ij} = P_{g_i g_j}$ for $i \neq j$ and $\min\{i, j\} \leq n_1$ with $g_i, g_j \in \{1, ..., K\}$. The community memberships can be expressed as a vector $\mathbf{g} = (g_1, \cdots, g_n)^T$. Write the block representation $\widetilde{\mathbf{B}}^* = (\widetilde{B}^*_{ij})_{n\times n} = \begin{pmatrix} \widetilde{\mathbf{B}}^*_{11} & \widetilde{\mathbf{B}}^*_{12} \\ \widetilde{\mathbf{B}}^*_{21} & \mathbf{0} \end{pmatrix}$. By leveraging results from random matrix theory, we have a comprehensive understanding of the matrix $\widetilde{\mathbf{B}}^*$, as stated in the following two lemmas.

**Lemma 1.** *Suppose Assumptions 1 and 2 stated in Section 3 hold. The matrix $\widetilde{\mathbf{B}}^*_{11}$ is a $n_1 \times n_1$ generalized Wigner matrix, satisfying $\mathbb{E}(\widetilde{B}^*_{ij}) = 0$ for all $(i, j)$ and $\sum_j Var(\widetilde{B}^*_{ij}) = 1$ for all $i$. When $n \to \infty$,*

$$T^*_{11} = n_1^{2/3}(\lambda_1(\widetilde{\mathbf{B}}^*_{11}) - 2) \xrightarrow{d} TW_1,$$

$$T^*_{12} = n_1^{2/3}(-\lambda_n(\widetilde{\mathbf{B}}^*_{11}) - 2) \xrightarrow{d} TW_1,$$

*where $TW_1$ denotes the standard Tracy-Widom distribution.*

**Lemma 2.** *Suppose Assumptions 1 and 2 stated in Section 3 hold. The matrix $(\widetilde{\mathbf{B}}^*_{12})^T\widetilde{\mathbf{B}}^*_{12}$ is a $m \times m$ sample covariance matrix, satisfying $\mathbb{E}(\widetilde{B}^*_{ij}) =$*

0 *for all* $(i, j)$ *and* $\sum_j Var(\widetilde{B}^*_{ij}) = 1$ *for all* $i$. *When* $n \to \infty$,

$$T_2^* = \frac{\lambda_1\left[(\widetilde{\mathbf{B}}^*_{12})^T\widetilde{\mathbf{B}}^*_{12}\right] - \mu_{nm}}{\sigma_{nm}} \xrightarrow{d} TW_1,$$

*where the parameters* $\mu_{nm}$ *and* $\sigma_{nm}$ *are defined by*

$$\mu_{nm} = \left(1 + \frac{\sqrt{m}}{\sqrt{n_1 - 1}}\right)^2, \tag{2.5}$$

$$\sigma_{nm} = \frac{1}{n_1 - 1}\left(\sqrt{n_1 - 1} + \sqrt{m}\right)\left(\frac{1}{\sqrt{n_1 - 1}} + \frac{1}{\sqrt{m}}\right)^{1/3}. \tag{2.6}$$

However, the matrix $\widetilde{\mathbf{B}}^*$ involves unknown connection probabilities $M_{ij} = P_{g_i g_j}$, the estimation of which necessitates community detection in the partial network. When a full adjacency matrix $\mathbf{A}$ obtained from $K$ communities is available, its signal term $\mathbb{E}\mathbf{A}$ satisfies $\text{rank}(\mathbb{E}\mathbf{A}) = K$. Then it is well-established that community membership can be estimated by applying k-means clustering to obtain $K$ clusters. Nevertheless, Han et al. (2024) demonstrated that when using a partial adjacency matrix $\mathbf{B}$, the rank of its signal term equals $2K$. Let $\mathbf{W}_{n\times 2K}$ be the matrix collecting the eigenvectors corresponding to the first $2K$ eigenvalues of $\mathbf{B}$. By applying the k-means algorithm to the non-zero rows of two groups $\mathbf{SW}$ and $(\mathbf{I} - \mathbf{S})\mathbf{W}$ ($\mathbf{S}$ is stated in (2.3)), they obtained $2K$ clusters denoted as $(\varepsilon_1, ..., \varepsilon_K)$ and $(\eta_1, ..., \eta_K)$, respectively. The two sets of clusters are equivalent up to a permutation of labels, where the first set corresponds to the nodes with $b_{1i} = 1$

and the second set corresponds to the nodes with $b_{1i} = 0$. Therefore, after merging the $2K$ clusters into $K$ clusters by the Algorithm 2 in Han et al. (2024), we propose to estimate unknown parameters $M_{ij}$ by $\widehat{M}_{ij} = \widehat{P}_{\widehat{g}_i\widehat{g}_j}$, where $\widehat{\mathbf{g}} = (\widehat{g}_1, ..., \widehat{g}_n)^T$ is an estimated community membership vector with the target number of communities being $K_0$. Then

$$\widehat{P}_{kl} = \frac{1}{\left|\widehat{\chi}_{k,l}^{(1)}\right| + \left|\widehat{\chi}_{k,l}^{(2)}\right|} \sum_{(i,j)\in\widehat{\chi}_{k,l}^{(1)}\cup\widehat{\chi}_{k,l}^{(2)}} B_{ij}, \tag{2.7}$$

where the sets $\widehat{\chi}_{k,l}^{(1)} = \{(i,j) : 1 \le i \le n_1, 1 \le j \le n_1, i \in \varepsilon_k, j \in \varepsilon_l\}$ and $\widehat{\chi}_{k,l}^{(2)} = \{(i,j) : 1 \le i \le n_1, (n_1+1) \le j \le n, i \in \eta_k, j \in \eta_l\}$. Plugging in the estimated parameters, we have

$$\widetilde{B}_{ij} = \begin{cases} \dfrac{B_{ij} - \widehat{M}_{ij}}{\sqrt{(n_1-1)\widehat{M}_{ij}(1-\widehat{M}_{ij})}}, & i \neq j \ and \ \min\{i,j\} \le n_1, \\ 0, & i = j \ or \ \min\{i,j\} > n_1, \end{cases} \tag{2.8}$$

and write the block representation $\widetilde{\mathbf{B}} = (\widetilde{B}_{ij})_{n\times n} = \begin{pmatrix} \widetilde{\mathbf{B}}_{11} & \widetilde{\mathbf{B}}_{12} \\ \widetilde{\mathbf{B}}_{21} & \mathbf{0} \end{pmatrix}$. Let

$$T_{11} = n_1^{2/3}\left(\lambda_1(\widetilde{\mathbf{B}}_{11}) - 2\right), \quad T_{12} = n_1^{2/3}\left(-\lambda_n(\widetilde{\mathbf{B}}_{11}) - 2\right),$$

and

$$T_2 = \frac{\lambda_1(\widetilde{\mathbf{B}}_{12}^T\widetilde{\mathbf{B}}_{12}) - \mu_{nm}}{\sigma_{nm}}.$$

Our test statistic takes the following form

$$T = \max\left\{|T_2| + |T_{11}|, |T_2| + |T_{12}|\right\}. \tag{2.9}$$

In Section 3, we prove that the asymptotic null distributions of $T_{11}$ $T_{12}$ and $T_2$ are $TW_1$ distributions, and that $T_2$ is asymptotically independent of $T_{11}$ and $T_{12}$. Then the rejection rule for testing problem (1.1) is:

$$\text{Reject} \quad H_0 : K = K_0 \quad \text{when} \quad T \geq t_\alpha, \tag{2.10}$$

where $t_\alpha$ is the critical value at a given significance level $\alpha \in (0, 1)$. By selecting $t_\alpha = t(\alpha/2)$, the upper $\alpha/2$ quantile of the absolute sum of two independent $TW_1$ distributions, we can ensure that the asymptotic type I error is bounded by

$$\begin{aligned}\mathbb{P}\left[T \geq t(\alpha/2)\right] &\leq \mathbb{P}\left[|T_2| + |T_{11}| \geq t(\alpha/2)\right] + \mathbb{P}\left[|T_2| + |T_{12}| \geq t(\alpha/2)\right] \\ &= \alpha/2 + o(1) + \alpha/2 + o(1) = \alpha + o(1).\end{aligned}$$

By setting $t_\alpha = t_n$, which will be specified in Section 3, we sequentially test the hypotheses (1.1) for $K_0 = 1, 2, ..., K_{\max}$ using rule (2.10). Once the test statistic fails to reject for a value of $K_0$, we stop and use it as the estimate of the rank. Denote by $\widehat{K}$ our resulting estimate. The estimation procedure is summarized in Algorithm 1. In practice, when prior knowledge is available—for example, biological constraints in genetic networks or physical limits in social networks—these factors often provide a natural upper bound for $K$. In the absence of such information, we adopt $K_{\max} = \lceil Cn^{1/6} \log n \rceil$, where $C$ is a generic positive constant and $\lceil \cdot \rceil$ de-

notes the ceiling function. This choice, motivated by the growth rate bound for $K$ in Theorem 1, ensures that $K_{\max} > K$ and retains model flexibility.

**Algorithm 1** Estimation of $K$.

**Require:** Partial adjacency matrix $\mathcal{B}$

Transfer $\mathcal{B} \to \mathbf{B}$ as in Section 2.2 and take matrices $\mathbf{S}, \mathbf{W}$ as in Section 2.3.

**for** $K_0 = 1, 2, ..., K_{\max}$ **do**

  Apply k-means($\mathbf{SW}$)$\to$ $(\varepsilon_1, ..., \varepsilon_K)$, k-means($(\mathbf{I}-\mathbf{S})\mathbf{W}$)$\to$ $(\eta_1, ..., \eta_K)$.

  Estimate $M_{ij}$ by (2.7) and construct test statistic $T$ via (2.9).

  **if** $T < t_\alpha$ **then**

    stop

  **else**

    $K_0 \leftarrow K_0 + 1$

  **end if**

**end for**

**Ensure:** $\widehat{K} \leftarrow K_0$.

## 3. Asymptotic properties

Theoretical properties of our approach are discussed in this section, which requires the incorporation of several assumptions.

**Assumption 1.** *The dimension of* $\mathbf{B}_{11}$ *is represented by* $n_1 = \sum_{i=2}^{n+1} b_{1i}$. *The ratio* $\gamma_n = \sum_{i=2}^{n+1} \mathbb{E}b_{1i}/n$ *satisfies* $\lim_{n\to\infty} \gamma_n = \gamma \in (0, 1)$.

**Assumption 2.** *The entries of the community connection probability matrix* $\mathbf{P}$ *are uniformly bounded away from 0 and 1. There exists some positive constant* $c_0$ *such that* $\min_{1\le k\le K} \sum_{i=1}^{n} \mathbb{I}(\boldsymbol{\pi}_i = \mathbf{e}_k) \ge c_0 n/K$.

**Assumption 3.** *The community membership estimator* $\widehat{\mathbf{g}}$ *is consistent in the sense that* $\mathbb{P}(\widehat{\mathbf{g}} = \mathbf{g}) \to 1$.

Note that in the partial framework, the dimension of $\mathbf{B}_{11}$ is random. Assumption 1 places a constraint on its expectation. Assumptions 2- 3 are also imposed in Lei (2016) to establish their main theorems. The first part of Assumption 2 ensures that the entries of $\mathbf{P}$ are of constant order. The second part guarantees that each community contains a sufficiently large number of nodes, which is essential for the consistent estimation of the connection probability matrix. We present Assumption 3 as a condition here; however, its validity has been demonstrated in Han et al. (2024) (refer to the exact recovery result in Section C of their Appendix). The asymptotic null distributions of the statistics employed in our approach are established in Theorem 1 and Theorem 2.

**Theorem 1.** *(Asymptotic null distributions) Suppose that Assumptions 1-*

*3 hold and* $K = O(n^{\frac{1}{6}-\tau})$ *for some* $\tau \in (0, 1/6)$. *Consider the matrix* $\widetilde{\mathbf{B}}$ *generated by* (2.8). *Under the null hypothesis* $K = K_0$, *we have*

$$T_{11} = n_1^{2/3}(\lambda_1(\widetilde{\mathbf{B}}_{11}) - 2) \xrightarrow{d} TW_1, \quad T_{12} = n_1^{2/3}(-\lambda_n(\widetilde{\mathbf{B}}_{11}) - 2) \xrightarrow{d} TW_1,$$

$$T_2 = \frac{\lambda_1(\widetilde{\mathbf{B}}_{12}^T\widetilde{\mathbf{B}}_{12}) - \mu_{nm}}{\sigma_{nm}} \xrightarrow{d} TW_1,$$

*where*

$$\mu_{nm} = \left(1 + \frac{\sqrt{m}}{\sqrt{n_1 - 1}}\right)^2,$$

*and*

$$\sigma_{nm} = \frac{1}{n_1 - 1}\left(\sqrt{n_1 - 1} + \sqrt{m}\right)\left(\frac{1}{\sqrt{n_1 - 1}} + \frac{1}{\sqrt{m}}\right)^{1/3}.$$

**Theorem 2.** *Suppose that Assumptions 1-3 hold. Under the null hypothesis* $K = K_0$, *when* $n \to \infty$, *the statistic* $T_2$ *is asymptotically independent of* $T_{11}$ *and* $T_{12}$.

Considering our test statistic $T$ defined in (2.9), Theorems 1 and 2, along with the subsequent discussion following rule (2.10), ensure that its type I error is asymptotically bounded by the given significance level $\alpha$. Theorem 3 below further provides a lower bound on the growth rate of $T$ under the alternative hypothesis $K_0 < K$.

**Theorem 3.** *(Asymptotic power guarantee) Suppose that Assumptions 1-3 hold. The matrix* $\widetilde{\mathbf{B}}_{12}$ *is generated by (2.8). Let* $\delta_n$ *be the smallest* $l_\infty$

*distance among all pairs of distinct rows of* $\mathbf{P}$. *For any* $K_0 < K$ *and any estimated community membership vector* $\widehat{\mathbf{g}}$, *we have*

$$\sigma_1\left(\widetilde{\mathbf{B}}_{12}\right) \geq \frac{\delta_n c_0 \sqrt{n_1}}{2K^2} + O_p(1).$$

According to Theorem 3, it can be inferred that the growth rate of $T$ is at least $\delta_n^2 n K^{-4} \cdot n^{2/3}$. A similar lower bound has been established in Lei (2016) for $\lambda_1(\widetilde{\mathbf{B}}_{11})$ and $\lambda_n(\widetilde{\mathbf{B}}_{11})$. By combining these results with Theorem 1, we observe a clear distinction between the asymptotic behavior of $T$ under null and alternative hypotheses. Consequently, through an appropriate selection of $t_n$, we can obtain a consistent estimator of $K$.

**Theorem 4.** *(Consistent estimator) Under the conditions in Theorem 1 and Theorem 3, if* $\delta_n$ *is bounded away from* 0 *and* $\widehat{K}$ *is estimated by Algorithm 1 with threshold* $t_\alpha = t_n \asymp n^\epsilon$ *for some* $\epsilon \in (0, 5/6)$, *we have*

$$\mathbb{P}\left(\widehat{K} = K\right) \to 1.$$

Theorem 4 guarantees the consistency of the estimator for $K$ provided that $\delta_n$ is bounded away from zero. Moreover, the practical selection of $t_\alpha$ will be discussed in detail in Section 4.2.

## 4. Numerical experiments

In this section, we employ simulations to demonstrate the efficacy of our proposed method in testing and estimating $K$. Section 4.1 verifies the proposed null distributions and Section 4.2 examines the performance in individual-centered partial networks. An extension to directed graphs is investigated in Section 4.3.

### 4.1 Asymptotic null distribution

Prior studies (Bickel and Sarkar, 2016; Lei, 2016) have noted the efficiency of the bootstrap correction method. We extend this idea to the partial adjacency matrix. The bootstrap test statistic is then defined by

$$\begin{aligned} BT = \max\Bigg\{ & \left| u_{tw} + s_{tw} \frac{\lambda_1(\widetilde{\mathbf{B}}_{12}^T \widetilde{\mathbf{B}}_{12}) - \widehat{\mu}_{nm}}{\widehat{\sigma}_{nm}} \right| + \left| u_{tw} + s_{tw} \frac{\lambda_1(\widetilde{\mathbf{B}}_{11}) - \widehat{\mu}_1}{\widehat{s}_1} \right|, \\ & \left| u_{tw} + s_{tw} \frac{\lambda_1(\widetilde{\mathbf{B}}_{12}^T \widetilde{\mathbf{B}}_{12}) - \widehat{\mu}_{nm}}{\widehat{\sigma}_{nm}} \right| + \left| u_{tw} + s_{tw} \frac{-\lambda_n(\widetilde{\mathbf{B}}_{11}) + \widehat{\mu}_n}{\widehat{s}_n} \right| \Bigg\} \\ := \max & \left\{ |BT_2| + |BT_{11}|, |BT_2| + |BT_{12}| \right\}, \end{aligned}$$

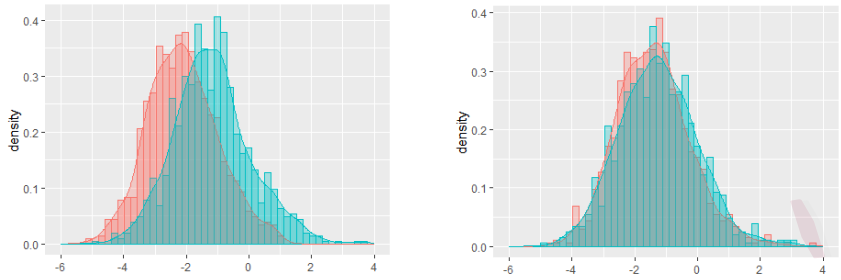
where $u_{tw}$ and $s_{tw}^2$ are the mean and variance of $TW_1$. The values $(\widehat{\mu}_1, \widehat{s}_1^2)$, $(\widehat{\mu}_n, \widehat{s}_n^2)$ and $(\widehat{\mu}_{nm}, \widehat{\sigma}_{nm}^2)$ are the sample means and variances of the largest eigenvalues of $\widetilde{\mathbf{B}}_{11}^{(h)}$, the smallest eigenvalues of $\widetilde{\mathbf{B}}_{11}^{(h)}$ and the largest eigenvalues of $(\widetilde{\mathbf{B}}_{12}^T \widetilde{\mathbf{B}}_{12})^{(h)}$, respectively, $h = 1, ..., H$. Throughout the simulations,

we select a bootstrap sample size of $H = 50$, which has been proved reasonable in Efron and Tibshirani (1994). The fifty partial adjacency matrices are generated from SBM with $\widehat{\mathbf{P}}$ calculated using (2.7).

We generate 1000 independent partial adjacency matrices, each consisting of $n = 200$ nodes and two communities of equal size, following the stochastic block model with $P_{11} = P_{22} = 0.65$ and $P_{12} = P_{21} = 0.2$. The estimated densities of $T_{11}$, $T_{12}$ and $T_2$ from 1000 repetitions are displayed in Figures 2, 3 and 4, respectively, both with and without bootstrap corrections. We can observe that the disparity between the finite empirical distribution and the theoretical limiting distribution ($TW_1$ as established in Theorem 1) can be significantly reduced through the bootstrap correction, leading to a convergence of two density curves. Similar phenomenon is observed in Figure 1 of Lei (2016). With an increased node size of $n = 1600$, as shown in Figure 5, this difference may become negligible even without employing bootstrap.

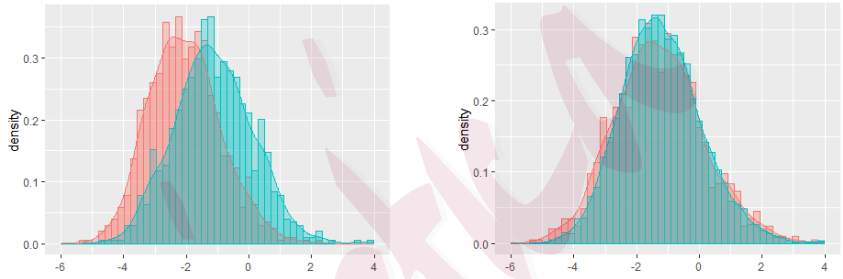
## 4.2 Performance in partial networks

This section is to evaluate the performance of our approach as outlined in Algorithm 1. Set the community connection probability matrix $\mathbf{P} = (P_{kl})_{K\times K}$ with $P_{kl} = r(1 + 2 \times \mathbb{I}(k = l))$, which means that the intra-




(a) $T_{11}$ (b) $T_{11}$ *with bootstrap correction*

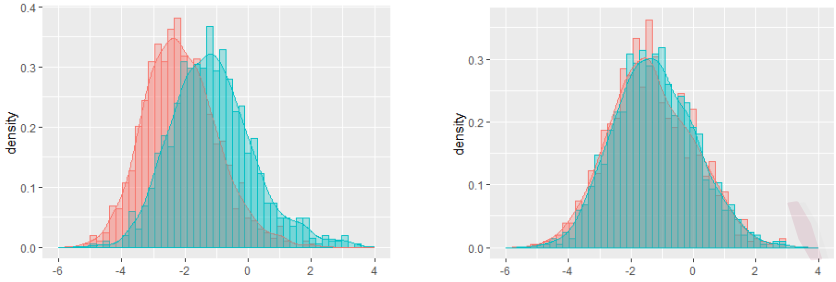
Figure 2: *The histograms of* $T_{11}$ *without and with bootstrap* ($n = 200$)*, where the blue parts represent the standard Tracy-Widom distribution* ($\beta = 1$) *and the red parts represent the empirical null distributions.*




(a) $T_{12}$ (b) $T_{12}$ *with bootstrap correction*

Figure 3: *The histograms of* $T_{12}$ *without and with bootstrap* ($n = 200$)*, where the blue parts represent the standard Tracy-Widom distribution* ($\beta = 1$) *and the red parts represent the empirical null distributions.*

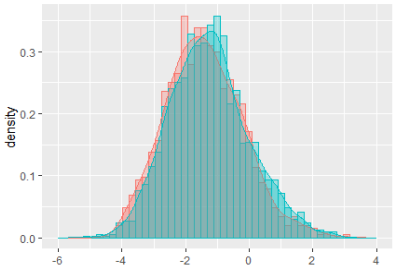
community connection probability is $3r$, while the inter-community connection probability is $r$. We generate 200 independent partial adjacency



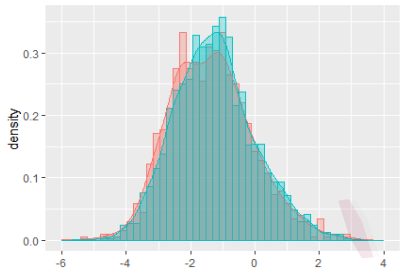

(a) $T_2$ (b) $T_2$ *with bootstrap correction*

Figure 4: *The histograms of* $T_2$ *without and with bootstrap* ($n = 200$)*, where the blue parts represent the standard Tracy-Widom distribution* ($\beta = 1$) *and the red parts represent the empirical null distributions.*

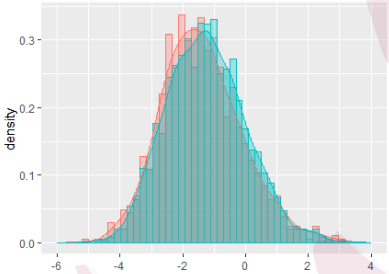
matrices, each consisting of $n + 1$ nodes with $n = 1000$ and $K$ equal-sized communities. Algorithm 1 is executed for different sparsity levels $r \in \{0.1, 0.2\}$ and various numbers of communities $K \in \{2, 3, 4, 5, 6\}$ with $K_{\max} = \lceil Cn^{1/6} \log n \rceil = 10$. For each of $10^5$ simulations, we generate three independent random variables from the $TW_1$ distribution and compute the test statistic $T$ according to Equation (2.9). The empirical distribution of $T$ is then constructed from these $10^5$ replicates, and its upper $\alpha/2$ quantile is used as the critical threshold. For the choice of $\alpha$, a conventional significance level such as 0.05 is sufficient for single hypothesis testing with $K = K_0$. However, for sequential testing, a more conservative value of $\alpha = 2 \times 10^{-4}$ is recommended to effectively control the overall Type I error




(a) $T_{11}$




(b) $T_{12}$




(c) $T_2$

Figure 5: *The histograms of* $T_{11}$, $T_{12}$ *and* $T_2$ *without bootstrap* ($n = 1600$), *where the blue parts represent the standard Tracy-Widom distribution* ($\beta = 1$) *and the red parts represent the empirical null distributions.*

rate. A similar choice is adopted in Lei (2016).

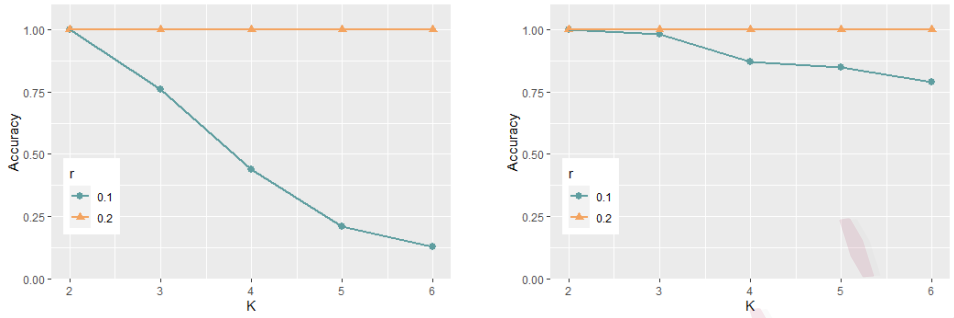
The proportions of correct estimations are recorded in Table 1 and visually presented in Figure 6. It is evident that when $r = 0.2$, the proportion approaches nearly 100%. As $r$ decreases to 0.1 and $K$ increases, there is a noticeable decline in the rate of correct estimations. This observation

Table 1: *Proportion of correct estimations of $K$ over 200 simulations under different sparsity levels controlled by $r$. The network size is $n = 1000$ with $K$ equal-sized communities.*

| | $T$ | | $BT$ | |
|---|---|---|---|---|
| $r$ | 0.1 | 0.2 | 0.1 | 0.2 |
| $K = 2$ | 1 | 1 | 1 | 1 |
| $K = 3$ | 0.76 | 1 | 0.98 | 1 |
| $K = 4$ | 0.44 | 1 | 0.87 | 1 |
| $K = 5$ | 0.21 | 1 | 0.85 | 1 |
| $K = 6$ | 0.13 | 1 | 0.79 | 1 |

aligns with expectations since the size $n$ remains constant at 1000; larger values of $K$ result in smaller communities. Additionally, the accuracy using bootstrap-corrected statistics significantly surpasses that without bootstrap correction. This discrepancy can be attributed to the faster convergence demonstrated in Section 4.1 for statistics incorporating bootstrap correction.

An interesting question pertains to how tailored methods for full networks perform in our partial network framework. In order to address this, we employ two well-established techniques: Universal Singular Value




(a) $T$ (b) $T$ *with bootstrap correction*

Figure 6: *Proportion of correct estimations of* $K$ *over 200 simulations under different sparsity levels controlled by* $r$*, where the cadetblue lines represent the results when* $r = 0.1$ *and the goldenrod lines represent the results when* $r = 0.2$*. The network size is* $n = 1000$ *with* $K$ *equal-sized communities.*

Thresholding (USVT) proposed by Chatterjee et al. (2015) and goodness of fit tests with ($T_{n,K}(boot)$) or without ($T_{n,K}$) bootstrap as suggested by Lei (2016). Considering the same model as described above, Table 2 records the proportions of correct estimations for those approaches when $r = 0.1$, where $T$ and $BT$ denote our method without and with bootstrap respectively. One may see that the well-known methods, which demonstrate excellent performance in a full network, exhibit unsatisfactory performance when applied to a partial network. This observation supports the necessity of our proposed approach. We argue that these methods are expected to fail since they are designed based on the entire adjacency matrix, implying that the thresholds

Table 2: *Proportion of correct estimations of $K$ over 200 simulations by different methods when $r = 0.1$. The network size is $n = 1000$ with $K$-equal sized communities.*

| | USVT | $T_{n,K}$ | $T_{n,K}(boot)$ | $T$ | $BT$ |
|---|---|---|---|---|---|
| $K = 2$ | 0.3 | 0 | 1 | 1 | 1 |
| $K = 3$ | 0.05 | 0 | 0 | 0.76 | 0.98 |

employed in Chatterjee et al. (2015) and estimators utilized in Lei (2016) may become unreliable when only partial information is available.

### 4.3 An extension to directed graphs

A directed graph is a network in which each edge has a specific direction, indicating the flow from one node to another. In general the adjacency matrix of a directed network $\mathbf{D} = (D_{ij})$ is asymmetric, since the presence of an edge from $i$ to $j$ does not necessarily imply the existence of an edge from $j$ to $i$. Wang and Wong (1987) proposed stochastic block models for directed graphs and used some asymmetrical connection probability matrices to construct adjacency matrices.

Our approach can be readily extended to handle directed graphs by utilizing the block matrix $\widetilde{\mathbf{B}}_{12}$. Assuming that both the receiver and sender are partitioned into $K$ communities, we employ steps 2-10 in Algorithm 1

with three adjustments. Firstly, we use the co-clustering method proposed by Rohe et al. (2016) to discover community memberships in step 3. Secondly, in step 4, we estimate the unknown connection probabilities using following $\widehat{P}_{kl}$:

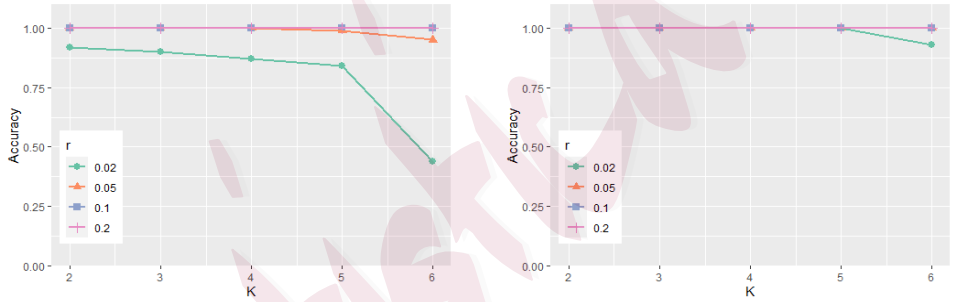
$$\widehat{P}_{kl} = \frac{\sum_{i\in\widehat{N}_k^S, j\in\widehat{N}_l^R} D_{ij}}{\widehat{n}_k^S \widehat{n}_l^R},$$

where $\widehat{N}_k^S = \{i : \widehat{g}_i^S = k\}$, $\widehat{n}_k^S = |\widehat{N}_k^S|$, $\widehat{N}_l^R = \{i : \widehat{g}_i^R = l\}$ and $\widehat{n}_l^R = |\widehat{N}_l^R|$. The superscripts $S$ and $R$ indicate the sender and the receiver. Lastly, we substitute the test statistic $T$ with $|T_2|$ and its bootstrap version $|BT_2|$.

Consider $r \in \{0.02, 0.05, 0.1, 0.2\}$ for varying levels of network sparsity when $K \in \{2, 3, 4, 5, 6\}$ and $K_{\max} = 10$. The threshold $t_n = t_\alpha$ with the nominal type I error bound $\alpha = 2 \times 10^{-4}$. The dimension of $\widetilde{\mathbf{B}}_{12}$ is $n \times n$. We set the diagonal elements of the community connection probability matrix $\mathbf{P}$ to $3r$, and the off-diagonal elements to either $r$ or $1.5r$ with equal probability. The simulated proportions of correct estimations are shown in Table 3, and a visual representation is provided in Figure 7. It can be observed that the performance is generally satisfactory, except for very sparse models with a large $K$ value. One may see that the results in Table 3 are better than those in Table 1. It is reasonable because with probability one, the dimension of the adjacency matrix in the directed network is greater than the dimension of $\widetilde{\mathbf{B}}_{12}$ in the partial network.

Table 3: *Proportion of correct estimations of $K$ over 200 simulations under different sparsity levels controlled by $r$. The network size is $n = 1000$.*

| | $\lvert T_2\rvert$ | | | | $\lvert BT_2\rvert$ | | | |
|---|---|---|---|---|---|---|---|---|
| $r$ | 0.02 | 0.05 | 0.1 | 0.2 | 0.02 | 0.05 | 0.1 | 0.2 |
| $K = 2$ | 0.92 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $K = 3$ | 0.9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $K = 4$ | 0.87 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $K = 5$ | 0.84 | 0.99 | 1 | 1 | 1 | 1 | 1 | 1 |
| $K = 6$ | 0.44 | 0.95 | 1 | 1 | 0.93 | 1 | 1 | 1 |




(a) $\lvert T_2\rvert$ (b) $\lvert BT_2\rvert$

Figure 7: *Proportion of correct estimations of $K$ over 200 simulations under different sparsity levels controlled by $r$. The network size is $n = 1000$ with $K$ equal-sized communities.*

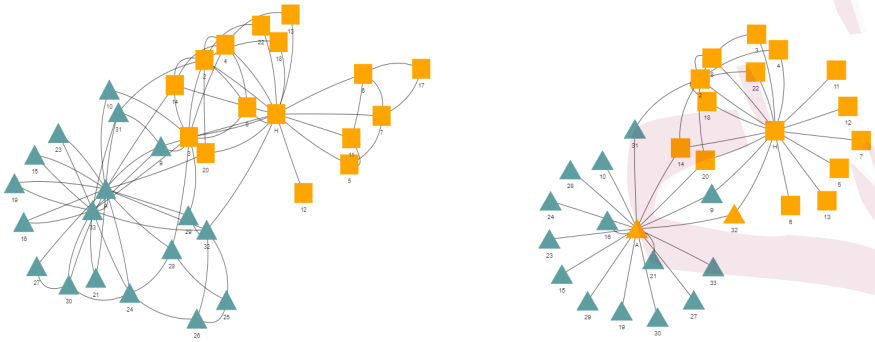
## 5. Real data analysis

Our partial network design is motivated by real-world investigative scenarios, such as criminal networks and epidemic contact tracing, where investigators typically begin with a targeted individual and gradually uncover their immediate and extended connections. Furthermore, compared to approaches that require complete network information, our method significantly reduces computational costs and storage requirements, particularly for large networks. In this section, we analyze two real datasets; full networks are used as the ground truth for performance comparison. The threshold value for the rejection area is 8.282 when setting $\alpha = 2 \times 10^{-4}$. Sequential hypothesis testing (1.1) is conducted using the bootstrap test statistic $BT$, and we will record this value for both datasets. The estimator is chosen as the first occurrence of $BT$ becoming less than 8.282.

The first one is Zachary's karate club data (Zachary, 1977), which records association within and between academic classes at the university. It comprises 34 members and 78 edges. This dataset has been extensively used in community detection research and is widely recognized to comprise two distinct communities, one led by the instructor and another led by the administrator. Following the discussion by Han et al. (2024), we also select individual 20 as the individual of interest, who possesses no specific

Table 4: *Estimation results for Zachary's karate club data.*

| $H_0$ | $K = 1$ | $K = 2$ | estimator |
|---|---|---|---|
| $BT$ | 15.10 | 2.44 | $K = 2$ |



(a) true communities with a full network (b) estimated communities with a partial network

Figure 8: *Orange nodes and cadetblue nodes represent two different communities respectively. (a): the full karate club network with nodes colored by ground truth community labels. (b): the partial karate club network centers on individual 20 with observed nodes colored by estimated community labels.*

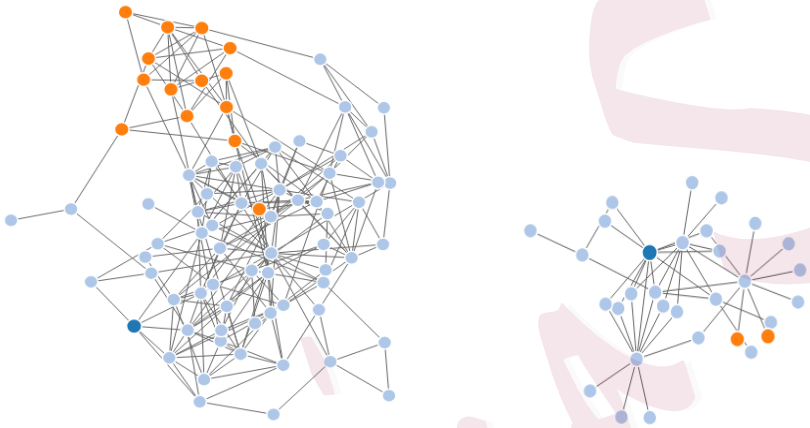
advantage in terms of observed edges but exhibits accurate membership estimation. The outcomes obtained through the implementation of our approach are documented in Table 4 and Figure 8, which estimate the number of communities to be 2, aligning with the prevailing consensus in the literature.

The second dataset pertains to a microfinance program implemented

across 43 Indian villages, as introduced by Banerjee et al. (2013). In each village, we take households as nodes and caste information as the truth community membership. Preprocessing of the data is conducted, which involves removing households with missing caste information and deleting unsurveyed data. Three villages with 2-4 balanced communities are chosen to examine our algorithm. Village 63 contains 77 nodes and 2 communities (OBC and SCHEDULE CASTE), Village 56 contains 68 households and 3 communities (GENERAL, OBC and MINORITY), and Village 50 has 114 nodes and 4 communities (OBC, GENERAL, SCHEDULE TRIBE and SCHEDULE CASTE). The households with ID 63001, ID 56001 and ID 50002 are selected as the individuals of interest for the three villages, respectively. The results are reported in Table 5. We can see that for all villages, our approach consistently estimated the true number of communities. The topology of Village 63 is depicted in Figure 9, where the left plot shows the full network while the right one displays the partial network consisting only of nodes visible to the individual with ID 63001. In comparison to Figure 8(b) from the first dataset, this particular individual has a more restricted perspective on the network. Nevertheless, we are still able to accurately estimate the number of communities. The additional topologies for Village 56 and Village 50 can be found in the Supplementary Material.

Table 5: *Estimation results for Indian village data.*

| | $BT$ $(K=1)$ | $BT$ $(K=2)$ | $BT$ $(K=3)$ | $BT$ $(K=4)$ | estimator |
|---|---|---|---|---|---|
| Village 63 | 12.47 | 4.61 | | | $K=2$ |
| Village 56 | 16.58 | 10.58 | 4.39 | | $K=3$ |
| Village 50 | 18.17 | 14.19 | 15.41 | 7.8 | $K=4$ |



(a) true communities with a full network

(b) estimated communities with a partial network

Figure 9: *Lightblue nodes and orange nodes represent OBC and SCHEDULE CASTE in Village 63 respectively. The individual with ID 63001 is visually distinguished by the darkblue color. (a): the full network with nodes colored by ground truth community labels. (b): the partial network centers on individual ID 63001 with observed nodes colored by estimated community labels.*

## 6. Discussion

This paper introduces a sequential testing framework for estimating the global number of communities $K$ using only partially observed network data. The proposed algorithm not only achieves provable consistency under the SBM but is also validated through extensive numerical simulations and real-data analyses. There are several promising directions for future research. While our focus is on the SBM, the proposed approach may be adaptable to more complex network models, such as the DCSBM, the DCMM and weighted networks. However, challenges include estimating additional unknown parameters or characterizing the relationship between partial and full networks in these extended models. Additionally, our methodology could potentially be generalized to sparse network settings, which represents an important avenue for future investigation.

## Supplementary Materials

The Supplementary Material contains all technical proofs and some additional real data analysis results.

## Acknowledgements

Xiyue Zhu and Xiao Han are co-first authors. Qing Yang is the corresponding author. This work was supported by NSF of China (No.12371278), NSF of China (No.12571297), National Key R & D Program of China-2022YFA1008000, the CAS Talent Introduction Program (Category B), and the Young Elite Scientist Sponsorship Program by Cast (No. YESS20220125).

The authors would like to thank the editorial team and reviewers for their invaluable insights and constructive suggestions, which have significantly improved the quality of this manuscript.

## References

Airoldi, E. M., D. Blei, S. Fienberg, and E. Xing (2008). Mixed membership stochastic blockmodels. *Advances in neural information processing systems 21*.

Alidaee, H., E. Auerbach, and M. P. Leung (2020). Recovering network structure from aggregated relational data using penalized regression. *arXiv preprint arXiv:2001.06052*.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144), 1236498.

Banerjee, D. and Z. Ma (2017). Optimal hypothesis testing for stochastic block models with growing degrees. *arXiv preprint arXiv:1705.05305*.

Bickel, P. J. and P. Sarkar (2016). Hypothesis testing for automated community detection

in networks. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 253–273.

Breza, E., A. G. Chandrasekhar, S. Lubold, T. H. McCormick, and M. Pan (2023). Consistently estimating network statistics using aggregated relational data. *Proceedings of the National Academy of Sciences 120*(21), e2207185120.

Chatterjee, S. et al. (2015). Matrix estimation by universal singular value thresholding. *Annals of Statistics 43*(1), 177–214.

Chaudhuri, K., F. Chung, and A. Tsiatas (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pp. 35–1. JMLR Workshop and Conference Proceedings.

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

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Freeman, L. C. (1982). Centered graphs and the structure of ego networks. *Mathematical Social Sciences 3*(3), 291–304.

Girvan, M. and M. E. Newman (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences 99*(12), 7821–7826.

Han, X., Y. R. Wang, Q. Yang, and X. Tong (2024). Individual-centered partial information in social networks. *Journal of Machine Learning Research 25*(230), 1–60.

# REFERENCES

Han, X., Q. Yang, and Y. Fan (2023). Universal rank inference via residual subsampling with application to large networks. *The Annals of Statistics 51*(3), 1109–1133.

Hofman, J. M. and C. H. Wiggins (2008). Bayesian approach to network modularity. *Physical review letters 100*(25), 258701.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Hwang, N., J. Xu, S. Chatterjee, and S. Bhattacharyya (2024). On the estimation of the number of communities for sparse networks. *Journal of the American Statistical Association 119*(547), 1895–1910.

Jin, J. (2015). Fast community detection by score. *The Annals of Statistics 43*(1), 57–89.

Jin, J., Z. T. Ke, S. Luo, and M. Wang (2023). Optimal estimation of the number of network communities. *Journal of the American Statistical Association 118*(543), 2101–2116.

Karrer, B. and M. E. Newman (2011). Stochastic blockmodels and community structure in networks. *Physical review E 83*(1), 016107.

Krzakala, F., C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences 110*(52), 20935–20940.

Latouche, P., E. Birmele, and C. Ambroise (2012). Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling 12*(1), 93–115.

# REFERENCES

Le, C. M. and E. Levina (2015). Estimating the number of communities in networks by spectral methods. *arXiv preprint arXiv:1507.00827*.

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

Li, T., E. Levina, and J. Zhu (2020). Network cross-validation by edge sampling. *Biometrika 107*(2), 257–276.

Li, T., Y.-J. Wu, E. Levina, and J. Zhu (2023). Link prediction for egocentrically sampled networks. *Journal of Computational and Graphical Statistics*, 1–24.

McDaid, A. F., T. B. Murphy, N. Friel, and N. J. Hurley (2013). Improved bayesian inference for the stochastic block model with application to large networks. *Computational Statistics & Data Analysis 60*, 12–31.

Ren, M., S. Zhang, and J. Wang (2023). Consistent estimation of the number of communities via regularized network embedding. *Biometrics 79*(3), 2404–2416.

Riolo, M. A., G. T. Cantwell, G. Reinert, and M. E. Newman (2017). Efficient method for estimating the number of communities in a network. *Physical review E 96*(3), 032310.

Rohe, K. (2019). A critical threshold for design effects in network sampling. *The Annals of Statistics 47*(1), 556–582.

Rohe, K., S. Chatterjee, and B. Yu (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics 39*(4), 1878–1915.

# REFERENCES

Rohe, K., T. Qin, and B. Yu (2016). Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences 113*(45), 12679–12684.

Saldana, D. F., Y. Yu, and Y. Feng (2017). How many communities are there? *Journal of Computational and Graphical Statistics 26*(1), 171–181.

Salganik, M. J. and D. D. Heckathorn (2004). 5. sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology 34*(1), 193–240.

Sarkar, P. and P. J. Bickel (2015). Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics 43*(3), 962–990.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing 17*(4), 395–416.

Wang, Y. J. and G. Y. Wong (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association 82*(397), 8–19.

Wang, Y. R. and P. J. Bickel (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics 45*(2), 500–528.

Yan, Y., B. Hanlon, S. Roch, and K. Rohe (2020). Asymptotic seed bias in respondent-driven sampling. *Electronic Journal of Statistics 14*(1).

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research 33*(4), 452–473.

# REFERENCES

Xiyue Zhu

School of Management, University of Science and Technology of China.

E-mail: zhuxiyue@mail.ustc.edu.cn

Xiao Han

International Institute of Finance, School of Management, University of Science and Technology of China.

E-mail: xhan011@ustc.edu.cn

Qing Yang

International Institute of Finance, School of Management, University of Science and Technology of China.

E-mail: yangq@ustc.edu.cn