Applied Network Science

RESEARCH

Open Access

Selecting a significance level in sequential testing procedures for community detection

Riddhi Pratim Ghosh[1*] and Ian Barnett[2*]

*Correspondence:
rpghosh@bgsu.edu;
ibarnett@pennmedicine.upenn.edu

[1] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA
[2] Department of Biostatistics, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

While there have been numerous sequential algorithms developed to estimate community structure in networks, there is little available guidance and study of what significance level or stopping parameter to use in these sequential testing procedures. Most algorithms rely on prespecifiying the number of communities or use an arbitrary stopping rule. We provide a principled approach to selecting a nominal significance level for sequential community detection procedures by controlling the tolerance ratio, defined as the ratio of underfitting and overfitting probability of estimating the number of clusters in fitting a network. We introduce an algorithm for specifying this significance level from a user-specified tolerance ratio, and demonstrate its utility with a sequential modularity maximization approach in a stochastic block model framework. We evaluate the performance of the proposed algorithm through extensive simulations and demonstrate its utility in controlling the tolerance ratio in single-cell RNA sequencing clustering by cell type and by clustering a congressional voting network.

Keywords: Community detection, Multiple testing, Sequential testing, Stochastic block model, Single cell RNA sequencing

Introduction

In the last few decades, there has been an increasing interest among physicists, computer and social scientists to study network data. Identifying community structure in a networks has gained particular attention: the vertices in networks are often found to cluster into related groups where vertices within a community are more likely to be connected [see, e.g., Newman and Girvan (2004), Newman (2006)]. The ability to detect such communities is crucial to understand the relationship between the structure and function of networks, such as the modeling of networks (Cheng et al. 2009), the evolution of networks (Zhang et al. 2008; Shen and Cheng 2010), the resilience of networks (Albert et al. 1999; Cheng et al. 2010), and the capacity of networks (Zhang et al. 2007a). The stochastic block model (Holland et al. 1983) is a popular model for community structures in network data where edge probabilities between and within communities are constant conditional on community membership.

Many community detection methods begin with a null model of no community structure. Historically, the most common approach involving a null model is the use of a node

partition score that is large when nodes within a partition are highly interconnected, relative to what is expected under the null of no structure (Newman 2006; Fortunato 2010). Many sequential community detection algorithms perform this task by first dividing the network into two communities, and subsequently subdividing each community hierarchically, ideally terminating when the true number of communities, K, has been reached. One such algorithm that is widely used in literature is based on modularity maximization proposed by Newman (2006) and its different variants including fast greedy modularity optimization (Clauset et al. 2004), exhaustive modularity optimization via simulated annealing (Guimera et al. 2004; Massen and Doye 2005; Medus et al. 2005; Guimera and Amaral 2005), fast modularity optimization (Blondel et al. 2008). Parallel community detection algorithms have garnered some attention over the last decade that modify existing algorithms to make them more suitable for the analysis of large networks. Riedy et al. (2011) modified the agglomerative community detection algorithm by choosing multiple contraction edges simultaneously as opposed to sequential contraction that is commonly done. Yang et al. (2016) compare several state-of-the-art algorithms on artificial networks in terms of accuracy and computing time. Que et al. (2015) proposed a parallel community detection algorithm derived from Louvain modularity maximization method using a novel graph mapping and data representation. A hypothesis testing framework based on modularity-based community detection has been studied by Zhang and Chen (2017) where they introduced a hypothesis testing procedure to determine the significance of the partitions obtained from maximizing the modularity function starting from a null model with no graph structure. However, this neglects the sequential nature of the test, and ignores correlations among test statistics which we incorporate in our approach. Bickel and Sarkar (2016) provides an algorithm for finding the number of clusters in a stochastic block framework using the Tracy-Widom distribution as the limiting distribution of the highest eigenvalue of the adjacency matrix, and therefore is not suitable for the small or moderate sized networks. While the Bayesian paradigm offers some remedy using the maximum aposteriori estimate of the number of clusters adjusting for underfitting in increased Bayesian hierarchies, however, this requires the suitable choice of the prior distribution on the number of communities (Peixoto 2019) which itself is a daunting task. To make a sequential community detection algorithm effective, the significance level for rejecting the null hypothesis needs to be specified for each test given by $H_0 : K = j$ community against $H_a : K > j$ starting with $j = 1$ and incrementing j over the integers until the test fails to reject H_0. The standard practice of setting the significance level arbitrarily to to 0.05 or 0.01 has drawbacks because it is susceptible to multiple testing leading to increased Type I error due to the repeated sequential tests.

To circumvent the multiple testing problem in sequential community detection procedures, analogous to controlling family-wise error rate, specifying a nominal significance level accounting for multiple tests is necessary. We aim to instead control for the underfitting (overfitting) probability, defined as the probability that the estimated number of communities obtained by a sequential testing procedure is less than (greater than) the true number of communities K present in the network. Any given contexts specific tolerance for overfitting and underfitting probabilities ultimately dictates the nominal significance level that should be used. We address the problem of finding the nominal significance level and aim to provide an algorithm to determine it aligns with a context-specific user-specified

tolerance ratio, defined as the ratio of underfitting probability to overfitting probability in a generic sequential testing framework. Our algorithm hinges on finding a suitable estimate of the number of communities at a significance level that preserves the prespecified tolerance ratio.

The rest of this article is organized as follows. In "Sequential community detection" section, we first describe sequential community detection procedures and subsequently introduce our algorithm to choose a significance level guided by a pre-specified tolerance ratio. In "Example sequential community detection algorithm" section, we provide an example of our approach applied to Newman's modularity maximization for sequential community detection to select an appropriate significance level. "Simulations" section describes the performance of our algorithm through extensive simulation studies in stochastic block model frameworks. We derive appropriate significance levels in two real applications in "Real data analysis" section. Finally "Discussion" section concludes with a discussion of limitations and next directions for our approach.

Sequential community detection

In this section, we first describe a generalizable sequential testing procedure to detect the number of communities in a network. Secondly, we describe the estimation of the tolerance ratio by deriving the expressions of underfitting and overfitting probabilities using an estimate of the number of communities. This tolerance ratio estimate is a function of the nominal significance level, which we can then solve for to arrive at a desired prespecified level.

Sequential testing procedure

Assuming a network of size n, the sequential testing procedure can be described by the following hypotheses:

$$H_0 : K = j, \qquad \text{against} \qquad H_A : K > j, \tag{1}$$

for each integer $j \geq 1$ until a test fails to reject.

Significance level from tolerance ratio

A common problem faced in community detection is the choice of an appropriate significance level α. Analogous to multiple testing problem (Benjamini and Hochberg 1995) where the goal is to control the family-wise error rate (FWER) through some procedures such as Bonferroni correction, Tukey's range test etc., we focus on sequential community detection algorithms, where tests of the null hypothesis $H_0 : K = j$ against the alternative $H_A : K > j$ are performed sequentially for $j = 1, 2, \ldots$ until a test fails to reject H_0. We let $p(j)$ be the p value of the jth such test $T(j; \alpha)$ defined as:

$$T(j; \alpha) = \begin{cases} 1 & \text{for } p(j) \leq \alpha \\ 0 & \text{for } p(j) > \alpha \end{cases}$$

Using this sequential procedure, the estimated number of communities is:

$$\hat{K}(\alpha) = \inf\{k \in \mathbb{N} : T(k+1; \alpha) = 0\}, \tag{2}$$

where $\hat{K}(\alpha)$ is a non-decreasing (step) function of α. Note that this details a generic sequential testing procedure, and applies generally to all sequential community detection methods that may have different forms of the test statistic that lead to the generation of pvalues $p(j)$. For example, the test statistic could range from modularity (Zhang and Chen 2017) to eigenvalues from a spectral decomposition (Bickel and Sarkar 2015).

We define the underfitting probability to be pr $(\hat{K}(\alpha) < K) = \eta_u$ and the overfitting probability to be pr $(\hat{K}(\alpha) > K) = \eta_o$. The *tolerance ratio* is defined as $\gamma = \eta_u/\eta_o$, where K is the true number of communities. One can note that $\gamma \in [0, \infty)$. In particular, $\gamma = 1$ implies underfitting and overfitting probabilities are equally likely. For unknown K, this also suggests one approach to estimate K that is independent of α: select $\hat{K}$ to be the value of K that results from the widest subinterval of α in [0, 1]. We call this α*-free* estimator K^*.

We assume a stochastic block model (SBM) which is a network of n nodes divided into K communities where B_{uv} is the probability of an edge between a node from community u and a node from community v, where $u, v \in \{1, 2, \ldots, K\}$. If g_i is the community label of node i ($g_i \in \{1, 2, \ldots, K\}$), then the probability of an edge between nodes i and j is the (g_i, g_j) entry of the matrix $\boldsymbol{B}$. The adjacency matrix $\boldsymbol{A} = (A_{ij})$ is given by $A_{ij} = B_{g_i g_j}, i \in \{1, 2, \ldots, n\}$.

We propose the following iterative procedure to identify the correct marginal significance level α to use from the user-specified tolerance ratio γ.

Input: The original or estimated adjacency matrix *A* of a graph and user-specified tolerance γ

1. For a given α, perform sequential community detection to obtain $\hat{K}(\alpha)$: For each $k \in \{1, 2, \ldots, n\}$, perform community detection given k communities to estimate community membership labels, and empirically estimate the k by k matrix of edge probabilities ($\tilde{\boldsymbol{P}}$) between and within these communities. Next, we simulate networks of size n from $\tilde{\boldsymbol{P}}$ as under an SBM while preserving the number of nodes in each of the k communities, and repeat this generation $\mathcal{B}$ times as in a parametric bootstrap to calculate p values from the empirical null distribution of the test statistic of choice in (2). For the bth bootstrapped adjacency matrix, the resulting estimator of K that results from the sequential testing procedure at the α level is defined to be $\hat{K}^{(b)}(\alpha)$ for $b = 1, 2, \ldots, \mathcal{B}$.
2. Determining K^*: $\hat{K}(\alpha)$ is a non-decreasing step function of α which can take integer values between 1 and n. Let $\alpha_1 < \alpha_2 < \ldots < \alpha_m$ denote the values of α in [0, 1] which yield distinct values $\hat{K}(\alpha_1) < \hat{K}(\alpha_2) < \ldots < \hat{K}(\alpha_m)$, where $1 \leq m \leq n$. Let $l_j = \{\alpha \in [0, 1] : \hat{K}(\alpha) = \hat{K}(\alpha_j)\}$ for $j = 1, 2, \ldots, m$, and $M = \underset{1 \leq j \leq m}{\text{argmax}}\ \text{length}(l_j)$.
3. For a given α, compute the tolerance ratio $\gamma(\alpha)$: $\gamma(\alpha) = \frac{\frac{1}{\mathcal{B}} \sum_{b=1}^{\mathcal{B}} I_{\{\hat{K}^{(b)}(\alpha) < K^*\}}}{\frac{1}{\mathcal{B}} \sum_{b=1}^{\mathcal{B}} I_{\{\hat{K}^{(b)}(\alpha) > K^*\}}}$
4. Return an $\alpha \in [0, 1]$ such that $|\gamma(\alpha) - \gamma|$ is minimized. When the minimizer is not unique, return the range of α that minimize $|\gamma(\alpha) - \gamma|$.

remark In Step 1 of the algorithm, here we use the spectral clustering algorithm (Rohe et al. 2011) using the *reg.SP* function in R that performs a community detection with a given adjacency matrix and a given number of clusters. However, the above algorithm is general and applies equally to any sequential community detection method. It is possible that the denominator of $\gamma(\alpha)$ is 0, in which case $\gamma(\alpha)$ is undefined. In practice by adding a small $\epsilon > 0$ to both the numerator and denominator of $\gamma(\alpha)$, the procedure will remain stable and well defined by adding a small bias towards $\gamma = 1$.

Below we present a brief proof of the convergence of the algorithm which assumes the Lipschitz condition on γ, and exploits some key characteristics about the change of underfitting probability with respect to α.

Theorem 1 *Suppose the target value of the significance is α_0 that corresponds to the tolerance ratio γ_0. Also assume that the function $\gamma(\alpha)$ satisfies the Lipschitz condition*

$$|\gamma(\alpha) - \gamma(\alpha^*)| \leq c|\alpha - \alpha^*|,$$

where $c > 0$ is the Lipschitz constant, $\alpha, \alpha^* \in (0, 1)$, and $\mathcal{B}$ tends to ∞. Then if the algorithm results in precision ϵ' for the significance α, then the precision of the tolerance ratio is $c\epsilon'$.

Proof First note that $\gamma(\alpha) = P(\hat{K}(\alpha) < K^*)/P(\hat{K}(\alpha) > K^*)$ is an increasing function of the underfitting probability $P(\hat{K}(\alpha) < K^*)$, which in turn is a decreasing function of α. This implies $\gamma(\alpha)$ is a decreasing function of α. So, there exists a c (could be very large) so that the Lipschitz condition holds. Therefore, if we use the precision ϵ' for α, the precision for γ is $c\epsilon'$. □

Two remarks are in order.

Remark 1 In Theorem 1, we assume that $\gamma(\alpha)$ is a continuous decreasing function of α. However, in Step 3 of our algorithm, $\gamma(\alpha)$ is guaranteed to be a non-increasing step function of α because we are estimating it empirically and $\hat{K}^{(b)}(\alpha)$ can take finitely many values in $\{1, 2, \ldots, n\}$. Therefore the difference $|\gamma(\alpha) - \gamma(\alpha^*)|$ can range over the entire real line, taking only finite values. The difference $|\alpha - \alpha^*|$ can range in the interval (0, 1). It is instructive to note that when the difference in α is zero or small, the corresponding difference in γ is also zero. Therefore, one can always pick $c > 0$ so that the Lipschitz condition is satisfied.

Remark 2 Our algorithm takes the user-specified tolerance as input and is expected to return a significance level is close to the tolerance ratio as possible. By fixing ϵ' beforehand the algorithm returns a significance level that lies within the ϵ'- neighborhood of the optimizing significance level.

Example sequential community detection algorithm

While our approach for identifying an α that corresponds with a prespecfied tolerance ratio is agnostic to which sequential community detection algorithm is used, we detail one example use case here. Aside various community detection algorithms such as spectral clustering (White and Smyth 2005; Zhang et al. 2007b), random walks (Pons and Latapy 2005), a popular approach to community detection is based on the idea of optimizing *modularity*. Modularity metrics were introduced by Newman and Girvan (2004), and the idea of detecting communities by optimizing a modularity function was proposed by Newman (2004) Nowadays, there are many variants of the modularity-based community detection approach to deal with directed or weighted networks (Leicht and Newman 2008). Also, some variants of the modularity-based community detection approach use modularity functions with a somewhat modified mathematical structure (Reichardt and Bornholdt 2006; Waltman et al. 2010; Traag et al. 2011).

Here we revisit Newman's sequential algorithm (Newman 2006) of community detection which begins by first dividing the network into two communities and then subdividing into further communities by maximizing additional modularity; and we implement our approach to selecting an appropriate significance level in this context.

For a network with n vertices, let $\boldsymbol{A}$ denote the $n \times n$ adjacency matrix and $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)^\top \in \{-1, 1\}^n$ where $s_i = 1$ if the i-th vertex belongs to group 1 and -1 otherwise. Let k_i denote the degree of vertex i and $m = \sum_{i=1}^{n} k_i/2$ be the total number of edges in the network. Then the modularity of the network is defined as

$$Q = \frac{1}{4m}\boldsymbol{s}^\top \boldsymbol{B}\boldsymbol{s}, \tag{3}$$

where the matrix $\boldsymbol{B} = (B_{uv})$ is defined as $B_{uv} = A_{uv} - \frac{k_u k_v}{2m}$, a symmetric matrix of order n.

Let $u_1, u_2, \ldots, u_n$ be the eigenvectors of $\boldsymbol{B}$ corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Then Q in (3) is maximized if $s_i = 1$ if the corresponding element in u_1 is positive and $s_i = -1$ otherwise rendering a network divided into two communities.

For further dividing a group j of size n_j, the additional contribution to the modularity is

$$\delta Q_j = \frac{1}{4m}\boldsymbol{s}^\top \boldsymbol{B}^{(j)}\boldsymbol{s} \tag{4}$$

is maximized in the similar way for Q in (3), where $B^{(j)}_{uv} = B_{uv} - \delta_{uv}\sum_{l\in j} B_{ul}$, and δ_{uv} is the Kronecker δ-symbol.

If the total modularity of the network after splitting the network into j communities is $Q^{(j)}$, then the gain in the modularity is defined by $\Delta Q^{(j)} = Q^{(j+1)} - Q^{(j)}$. Again, while we use this quantity $\Delta Q^{(j)}$ as our test statistic for the jth step ($H_0 : K = j$ vs $H_A : k > j$), we stress that any sequential community detection algorithm can be adopted to this framework.

Simulations

Homogenous case: stochastic block model

We perform extensive simulation study in various directions to assess the performance of the proposed algorithm. In each set-up, networks of size n and $2n$ are simulated through SBM with K_0 number of balanced communities of size n/K_0. We vary $n \in \{100, 200\}$ corresponding to $K_0 = 5, 10$ respectively for symmetric edge probability matrix $\boldsymbol{P}$ of dimension K_0 of the form

$$\boldsymbol{P} = 2\epsilon \boldsymbol{I}_{K_0} + (0.5 - \epsilon)\mathbf{1}_{K_0}\mathbf{1}_{K_0}^\top,$$

where $\boldsymbol{I}_{K_0}$ is the identity matrix of order K_0, $\mathbf{1}_{K_0}$ is the vector of 1 s of dimension K_0, so that the diagonal and off-diagonal entries of $\boldsymbol{P}$ are $0.5 + \epsilon$ and $0.5 - \epsilon$ respectively implying that the difference between edge probability within and between community is 2ϵ. We vary $\epsilon = 0.195, 0.010$ to represent two cases of (S) strong and (W) weak community structure, respectively (Table 1).

Table 1 Mode of 100 independent replications ($\hat{K}_\alpha$), and proportion of times true number of communities correctly estimated ($\text{pr}(\hat{K} = K_0)$) shown in parenthesis for different choice of α, $\boldsymbol{P}$, and $n = 100$

(n, K_0)	Signal	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
(100, 5)	S	5 (0.82)	5 (0.85)	5 (0.85)	6 (0.45)
	W	3 (0.10)	3 (0.15)	4 (0.35)	4 (0.20)
(100, 10)	S	8 (0.25)	8 (0.35)	10 (0.40)	10 (0.55)
	W	2 (0.00)	2 (0.00)	3 (0.00)	3 (0.00)

Table 2 Mode of 100 independent replications ($\hat{K}_\alpha$), and proportion of times true number of communities correctly estimated ($\text{pr}(\hat{K} = K_0)$) shown in parenthesis for different choice of α, $\boldsymbol{P}$, and $n = 200$

(n, K_0)	Signal	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
(200, 5)	S	5 (0.85)	5 (0.88)	5 (0.89)	5 (0.55)
	W	4 (0.30)	4 (0.45)	5 (0.58)	5 (0.45)
(200, 10)	S	9 (0.40)	9 (0.45)	10 (0.55)	10 (0.60)
	W	3 (0.00)	6 (0.00)	6 (0.05)	7 (0.10)

Estimated number of communities ($\hat{K}_\alpha$)

For a fixed α, we simulate 1000 parametric bootstrap sample values of the null test statistic and calculate p values by comparing them with the observed test statistic value. We start from the number of communities $K = 1$, and proceed by incrementing Kuntil the p value is greater than α. We replicate the procedure 100 times and finally report the value of the estimated number of communities $\hat{K}_\alpha$ by taking the mode of the 100 replications.

Next, we vary $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$ and the corresponding $\hat{K}_\alpha$s are reported in Sect. 1 and Table 2. Further, the estimates of $\text{pr}(\hat{K}_\alpha = K_0)$ are reported in parenthesis by taking the proportion of times $\hat{K}_\alpha$ is equal to K_0 over 100 replications. Note that the probability of recovery of the true number of communities is bounded from above by $(1 - \alpha)$. This follows because $\text{pr}(\hat{K}_\alpha = K_0) \leq \text{pr}(\hat{K}_\alpha = K_0 | \hat{K}_\alpha \geq K_0) = 1 - \alpha$. In the presence of strong differences in communities, estimated communities are close to the true number for $\alpha = 0.01, \ldots, 0.2$. For weak signals, the number of communities is under estimated for the aforementioned α. However, the number of communities is over estimated for larger value of the significance level. This indicates that the choice of α can greatly influence $\hat{K}$, which provides further incentivize for developing a rigorous approach to selecting an appropriate α.

Choice of significance level (α)

In each simulation set-up, we use 1000 bootstrap samples for a wide range of α (typically in the range [0.001, 0.5]) and store the values of $\hat{\gamma}(\alpha)$ according to Step 3 of the algorithm.

We consider the value of tolerance ratio $\gamma = 0.5, 1, 2$ corresponding to the cases where underfitting probability is half, equal, and twice of overfitting probability. In

Table 3 Choice of α for different choices of tolerance ratio γ and network size $n(n' = 2n)$

(n, K_0)	Signal	$\gamma = 1/2$	$\gamma = 1$	$\gamma = 2$
(100, 5)	S	0.06 (0.06)	0.01 (0.02)	0.005 (0.006)
	W	0.10 (0.09)	0.05 (0.04)	0.02 (0.01)
(200, 10)	S	0.07 (0.06)	0.01 (0.01)	0.005 (0.006)
	W	0.10 (0.10)	0.05 (0.04)	0.01 (0.01)

each case, we find the $\hat{\gamma}(\alpha)$ such that $|\hat{\gamma}(\alpha) - \gamma|$ is the minimum among the stored values, and report the corresponding value of α in Table 3. One can note that as γ is increasing (i.e., the overfitting probability is increasing relative to the underfitting probability), the significance level decreases. This is consistent with the fact that for a smaller value of α, the test is getting accepted at an early step than a larger value of α.

Heterogenous case: degree-corrected block model

Here we perform another simulation study for the case of heterogenous networks following the framework of Gao et al. (2018). We generate a network of n nodes and 2 communities where the sizes are n_1 and n_2, and we vary $n \in \{300, 600\}$, $n_1 \in \{100, 200\}$, and $n_2 \in \{200, 400\}$ respectively. The off-diagonal entries of the adjacency matrix $\boldsymbol{A} = (A_{ij})$ are simulated as

$$\begin{cases} A_{ij} = A_{ji} \overset{indep}{\sim} \text{Ber}(\theta_i\theta_j p), \text{ if } i \text{ and } j \text{ belongs to the same community,} \\ A_{ij} = A_{ji} \overset{indep}{\sim} \text{Ber}(\theta_i\theta_j q), \text{ otherwise,} \end{cases}$$

where $p = 0.1, q = 3p/10$ for strong signal (S) and $7q/10$ for weak signal (W), $\theta_i = |Z_i| + 1 - (2\pi)^{1/2}, Z_i \overset{i.i.d}{\sim} N(0, 0.25)$ for $i = 1, 2, \ldots, n$ so that $E(\theta_i) = 1$.

Using 1000 parametric bootstraps as mentioned in the algorithm, we present the proportion of times the true number of communities correctly estimated and the values of α for different choices of tolerance ratio in the following two tables. Like the previous case, Tables 4 and 5 also demonstrate that while for strong signals the number of communities is almost correctly estimated, however, for weak signals, they are underestimated. This provides incentivize for developing a rigorous approach of selecting α.

Table 4 Mode of 100 independent replications ($\hat{K}_\alpha$), and proportion of times true number of communities correctly estimated $(\text{pr}(\hat{K} = K_0))$ shown in parenthesis for different choice of α, $\boldsymbol{P}$, and (n, n_1, n_2) for DCBM

(n, n_1, n_2, K_0)	Signal	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
(300, 100, 200, 2)	S	2 (0.70)	2 (0.70)	2 (0.75)	2 (0.78)
	W	1 (0.20)	1 (0.25)	1 (0.45)	1 (0.45)
(600, 200, 400, 2)	S	2 (0.25)	2 (0.35)	2 (0.40)	2 (0.55)
	W	1 (0.03)	1 (0.05)	1 (0.15)	2 (0.50)

Table 5 Choice of α for different choices of tolerance ratio γ and network size (n, n_1, n_2)

(n, n_1, n_2, K_0)	Signal	$\gamma = 1/2$	$\gamma = 1$	$\gamma = 2$
(300, 100, 200, 2)	S	0.08	0.05	0.01
	W	0.15	0.10	0.08
(600, 200, 400, 2)	S	0.09	0.05	0.02
	W	0.15	0.11	0.09

Real data analysis

Single cell RNA (scRNA-seq) data

We apply our algorithm to the scRNA-seq data generated from the retina cells of two healthy adult donors using the 10X Genomics Chromium^{TM} system. We should expect some clustering by cell type in networks derived from this data. Detailed preprocessing and donor characteristics of the scRNA-seq data can be found in Lyu et al. (2019). The data consists of 33694 genes sequenced over 92385 cells. The sequencing data were initially analyzed with R package *Seurat* (Satija et al. 2015) and each of the cells was identified as a particular cell-type. The virtual representation of the data in the t-SNE plot is given in Fig. 1.

Among different clusters in *Seurat*, we consider the data pertaining to five hierarchical clusters: "Astrocytes", "Endothelium", "Ganglion", "Horizontal", "Pericytes". Before we perform the analysis, we process the data in three steps. First, genes whose variability was less than the 50th quantile are filtered out, and then cells whose total cell counts across all genes are less than 500 and greater than 2500 are also filtered out. Second, we compute the normalized score (row wise) and perform a log transformation ($log_2(1 + x/10000)$) as done in Booeshaghi and Pachter (2021) to convert the data

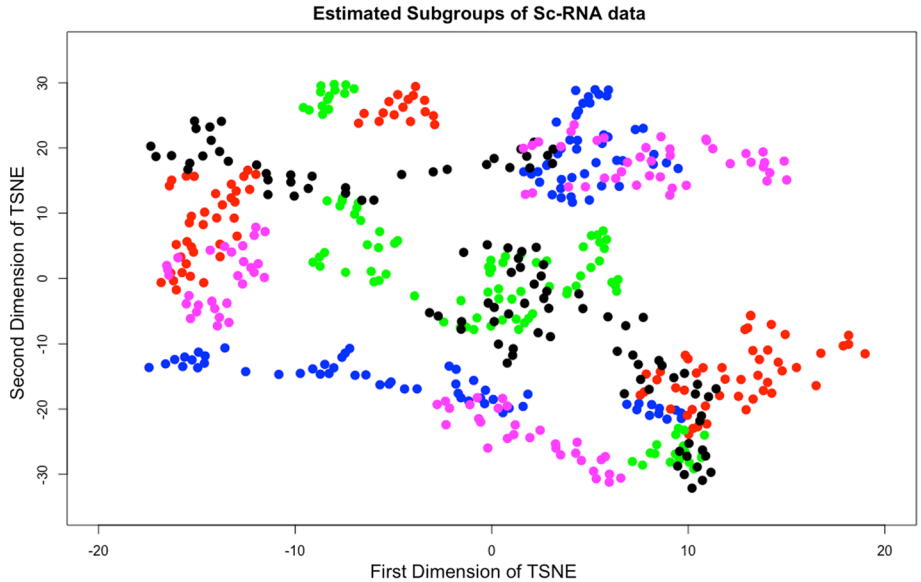

Fig. 1 Virtual representation of the estimated number of clusters of the analyzed scRNA-seq data of human retina cells in the t-SNE plot obtained by selecting Seurat classified cell types namely: Astrocytes, Endothelium, Ganglion, Horizaontal, Pericytes in an equal manner of roughly 100 cells per cell type

Table 6 Choice of α and corresponding estimated number of communities $\hat{K}_\alpha$ for different values of tolerance ratio η across various choices of correlation threshold γ for the scRNA-seq data

Correlation threshold (τ)		**0.3**			**0.5**			**0.7**	
Tolerance ratio (γ)	0.5	1	2	0.5	1	2	0.5	1	2
Significance level (α)	0.05	0.03	0.01	0.04	0.04	0.01	0.05	0.03	0.02
Estimated # communities ($\hat{K}_\alpha$)	7	7	6	9	8	7	9	9	7

into a continuous scale. The rationale behind such a transformation is that that different genes have different variances implying that genes that are highly expressed will have high variance whereas the genes that are barely expressed at all, will have almost zero variance. The transformed data is now used to compute correlations between the cells. Finally, for each cluster, we randomly select 100 cells ensuring that the within and between cluster correlations do not differ by more than 0.1 from those of the composite data. We use the correlation threshold (τ) to construct an adjacency matrix $\boldsymbol{A}$, and vary $\tau \in \{0.3, 0.5, 0.7\}$, and report the significance level along with estimated number of communities in Table 6. We observe that estimated number of communities is larger as we increase the value of τ which gives rise to a denser network. The significance level (α) ranges over [0.01, 0.05] depending on the tolerance ratio. Also, for each choice of τ, the estimated number of communities is increasing with α.

It can also be noted from Table 6 that different values of α lead to different number of estimated communities. If one were to arbitrarily pick α as, say, 0.05, this choice can have a large impact on the analysis. For example, corresponding to $\tau = 0.3$, α changes from 0.05 to 0.01 leading to different value of $\hat{K}$. Thus the choice of α is an impactful decision, and the tolerance ratio presents an intuitive measure that allows the practitioner to place a value of overfitting relative to underfitting when performing community detection.

United States House Votes 1984 (USHV) data

In this example, we consider a data set of 267 democrats and 167 republican congressmen who has voted in 16 issues in 1984 in the United States of America. The data contains yes/no answer for each congressman on 16 different questions with some missing values. After removing the congressman who has not voted in more than three of the sixteen questions, the data is represented by a 417×17 matrix where the first column represents the political affiliation-republican and democrat. The adjacency matrix $\boldsymbol{A}$ is calculated by thresholding the correlations among congressmen by τ, i.e., if the correlation of voting between two congressmen is as high as τ, we assume they are connected by an edge and hence the corresponding entry of the adjacency matrix 1, and 0 otherwise. Finally we vary $\tau \in \{0.3, 0.5, 0.7\}$. It is instructive to note that smaller values of τ leads to a more dense network.

In this data, the number of distinct communities is not expected to go below 2 because of the two party affiliations. However, in our analysis, the estimated number of communities varies in $\{3, 4, 5\}$ (depending on the desired tolerance ratio) implying potential further subdivisions among political parties. Here too, the significance level has a large impact on the analysis. For example, when $\tau = 0.7$, changing α from 0.06 to 0.04 drops the number of communities from 5 to 4. Therefore, a judicious choice of the significance

Table 7 Choice of α and corresponding estimated number of communities $\hat{K}_\alpha$ for different values of tolerance ratio γ across various choices of correlation threshold τ for the USVA data

Correlation threshold (τ)		**0.3**			**0.5**			**0.7**	
Tolerance ratio (γ)	0.5	1	2	0.5	1	2	0.5	1	2
Significance level (α)	0.07	0.05	0.02	0.06	0.05	0.01	0.06	0.04	0.01
Estimated # communities ($\hat{K}_\alpha$)	3	3	3	4	3	3	5	4	4

level is necessary, and the tolerance ratio again provides a means of guiding this choice in an intuitive manner (Table 7).

Discussion

We have proposed an algorithm to provide guidance to the practitioner in order to obtain a nominal significance level that matches their desired balance between overfitting and underfitting probabilities. Traditional approaches to estimate the number of communities often arbitrarily set the significance level, and the tolerance ratio presents an intuitive alternative. To construct the test statistic in a sequential testing framework, we used Newman's modularity maximization approach. Even though our approach is demonstrated on modularity based test, it will also work on any general sequential testing approach, for example, spectral clustering method (Ng et al. 2001) and its variants, model-based approaches (Lee and Wilkinson 2019). Our proposed method can be adapted to such settings by replacing the modularity by between clusters sum of squares variability at each step of sequential testing.

Although here we have assumed a stochastic block model, a feasible extension of this approach would be to apply it to dynamic stochastic block models Matias and Miele (2015) in order to allow a time varying network structure. It is instructive to note that we proposed the solution using the sequential tests, and implemented the algorithm via bootstrap due to the lack of the analytic expression of the test statistic. A potential bottleneck that the proposed algorithm will face is when the network size is very large because bootstrapping will be computationally expensive. However, in case an analytic expression of the test statistic is available in closed form, the algorithm can be adapted trivially to use it in place of bootstrapping. This would further increase algorithmic stability by removing stochasticity introduced through the bootstrap.

Acknowledgements
The authors (RG & IB) would like to thank Mingyao Li, Professor of Biostatistics at the University of Pennsylvania for the scRNA-seq data. IB is supported by R01MH116884.

Author contributions
RG performed the analysis, wrote, and edited the manuscript. IB edited the manuscript

Funding
IB is supported by R01MH116884.

Availability of data and materials
The data set is not available publicly.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
Not applicable.

Received: 4 February 2023 Accepted: 27 June 2023
Published online: 01 August 2023

References

Albert R, Jeong H, Barabási A-L (1999) Diameter of the world-wide web. Nature 401(6749):130–131
Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 57(1):289–300
Bickel PJ, Sarkar P (2015) Hypothesis testing for automated community detection in networks. J R Stat Soc Ser B (Stat Methodol) 1(78):253–273
Bickel PJ, Sarkar P (2016) Hypothesis testing for automated community detection in networks. J R Stat Soc Ser B (Stat Methodol) 78(1):253–273
Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
Booeshaghi AS, Pachter L (2021) Normalization of single-cell RNA-seq counts by log (x+ 1) or log (1+ x). Bioinformatics 37(15):2223–2224
Cheng X-Q, Ren F-X, Zhou S, Hu M-B (2009) Triangular clustering in document networks. New J Phys 11(3):033019
Cheng X-Q, Ren F-X, Shen H-W, Zhang Z-K, Zhou T (2010) Bridgeness: a local index on edge significance in maintaining global connectivity. J Stat Mech Theory Exp 2010(10):P10011
Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111
Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174
Gao C, Ma Z, Zhang AY, Zhou HH (2018) Community detection in degree corrected block models
Guimera R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433(7028):895–900
Guimera R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70(2):025101
Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: first steps. Soc Netw 5(2):109–137
Lee C, Wilkinson DJ (2019) A review of stochastic block models and extensions for graph clustering. Appl Netw Sci 4(1):1–50
Leicht EA, Newman ME (2008) Community structure in directed networks. Phys Rev Lett 100(11):118703
Lyu Y, Zauhar R, Dana N, Strang CE, Wang K, Liu S, Miao Z, Pan N, Gamlin P, Kimble JA, Messinger JD, Curcio CA, Stambolian D, Li M (2019) Integrative single-cell and bulk RNA-seq analysis in human retina identified cell type-specific composition and gene expression changes for age-related macular degeneration. bioRxiv, 768143
Massen CP, Doye JP (2005) Identifying communities within energy landscapes. Phys Rev E 71(4):046101
Matias C, Miele V (2015) Statistical clustering of temporal networks through a dynamic stochastic block model. arXiv preprint arXiv:1506.07464
Medus A, Acuña G, Dorso CO (2005) Detection of community structures in networks via global optimization. Physica A 358(2–4):593–604
Newman ME (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69(6):066133
Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582
Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. In: Advances in neural information processing systems, vol 14
Peixoto TP (2019) Bayesian stochastic block modeling. In: Advances in network clustering and block modeling, pp 289–332
Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: International symposium on computer and information sciences, pp 284–293
Que X, Checconi F, Petrini F, Gunnels JA (2015) Scalable community detection with the Louvain algorithm. In: 2015 IEEE international parallel and distributed processing symposium, pp 28–37
Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74(1):016110
Riedy EJ, Meyerhenke H, Ediger D, Bader DA (2011) Parallel community detection for massive graphs. In: International conference on parallel processing and applied mathematics, pp 286–296
Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic block model
Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33(5):495–502
Shen H-W, Cheng X-Q (2010) Spectral methods for the detection of network community structure: a comparative analysis. J Stat Mech Theory Exp 2010(10):P10020
Traag VA, Van Dooren P, Nesterov Y (2011) Narrow scope for resolution-limit-free community detection. Phys Rev E 84(1):016114
Waltman L, Van Eck NJ, Noyons EC (2010) A unified approach to mapping and clustering of bibliometric networks. J Informetr 4(4):629–635
White S, Smyth P (2005) A spectral clustering approach to finding communities in graphs. In: Proceedings of the 2005 SIAM international conference on data mining, pp 274–285
Yang Z, Algesheimer R, Tessone CJ (2016) A comparative analysis of community detection algorithms on artificial networks. Sci Rep 6(1):1–18
Zhang J, Chen Y (2017) A hypothesis testing framework for modularity based network community detection. Stat Sin 27:437–456

Zhang G-Q, Wang D, Li G-J (2007a) Enhancing the transmission efficiency by edge deletion in scale-free networks. Phys Rev E 76(1):017101
Zhang S, Wang R-S, Zhang X-S (2007b) Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A 374(1):483–490
Zhang G-Q, Zhang G-Q, Yang Q-F, Cheng S-Q, Zhou T (2008) Evolution of the internet and its cores. New J Phys 10(12):123027

Publisher's Note