Community Detection In Sparse Networks

Laala Zeyneb[1] and Belguerna Abderrahmane[2]

[1,2]*Department of Mathematics and Computer Science, University of Salhi Ahmed, Naama,45000, Algeria*

Abstract

Spectral methods in which they are based on matrix eigenvectors are widely used in network data analysis, especially for community detection. These classical approaches are based on graph associated matrix (adjacency matrix) and related matrices, nevertheless go wrong with sparse networks, which they have a lot of interest in practice. The spectrum of the non-backtracking matrix, an alternate matrix representation of a network that shows a behavior in the sparse limit, has recently been presented as a solution to this problem. However, the use of this matrix was limited for a specified number of communities. We are presenting a matrix for the graph and showing that it can be using for different number of communities.

Key Words: Community Detection, Stochastic Block Model, Sparse Networks

1 Introduction

Statistical network data analysis is now a well-studied area of statistics and machine learning. Network datasets can be found in a variety of disciplines. Many examples of networks are in real-life data such as bioscience, protein-protein interaction (3), epidemiological networks (10); citation and collaboration networks (15); social media networks such as LinkedIn, Twitter, or Facebook (8); in which the network is a set of social actors,such as individuals or organizations, linked together by connections representing social interactions. The analysis of networks, based mainly on graph theory and aims to study various aspects of these networks. Developing statistical procedures for network data analysis as well as establishing the theoretical features of statistical methods are two ac-tive areas of research. In this research, we focus on identifying the number of communities in networks with arbitrary sparsity levels and networks with community structure.

The topic of ”community detection,” as it is commonly known, has seen a renaissance of interest over the last two decades. A frequent issue specification is to divide *N* nodes of a network into K communities with different edge densities inside and across communities, with K predetermined. The literature on estimating the number of communities (K) has recently become active. While the initial focus in the literature for estimating K was on developing algorithms and drawing support from domain-specific intuition and empirical studies using the Stochastic Block Model (SBM), which was first proposed in 1977 by Holland (5), the current focus is on using this Model (SBM) to estimate *K*. Many application later appear such as bayesian perspective (11) discussed priors for number of communities under the SBM and designed an Markov Chain Monte Carlo algorithm, Kemp et al. (12) presented a nonparametric Bayesian approach for detecting concept systems.

The rapid development of ”community discovery” approaches for solving the core problem of discovering modules inside an arbitrary network has been fueled by this diverse set of applications.

In recent years, methods based on the spectrum of a specific class of matrices have gained popularity as non-parametric alternatives that are more computationally efficient and applicable to a larger range of situations.

Spectral approaches based on the eigenvalues and eigenvectors of some matrix representation of the network are particularly common. These combine speed of execution with a wealth of information about the network that goes beyond the modular structure, such as the relative roles of each node and the characterization of the network’s dynamical properties. Spectral methods for the analysis of large graphs and networks have become a cornerstone in the study of empirical network data since their inception in the 1970s (16).

The structure of the network of interest is represented using one of various matrix forms, such as the adjacency matrix or the graph Laplacian, and the eigenvalues and eigenvectors are examined for information on the network topology. Experiments indicate (and some model networks can verify) that the eigenvalue spectrum consists of a dense ”spectral band” of tightly spaced eigenvalues, similar to an allowable energy band in condensed matter, plus a number of outlying eigenvalues separated from the band by a considerable band gap.

For a some real networks, spectral methods can fail. The eigenvalues are separated into two classes, with the great majority ”the bulk” following a well-defined distribution and the outliers of this distribution providing information about the community’s structure. Unfortunately, there are many real-world networks that are sparse.

Krzakala et al. (6) who propose focusing on the eigenvalues and eigenvectors of a different matrix representation of network structure, which they call the non-backtracking

matrix, also known as the Hashimoto edge matrix by previous authors (4; 1), have proposed an interesting solution to these problems.

The non-backtracking was used to estimate the number of communities in a networks. Newman (9) has proposed a solution method based on this matrix to detect communities for sparse networks, this kind of graphs have an interesting application. However, his proposed methods can works only for networks with two communities using modularity. After that, a paper proposed by Singh and Humphries (14) in which they generalize for more than two communities. The authors proposed Reluctant backtracking and it gives better estimation than the flow operator proposed earlier.

In this paper, we propose a matrix that can work better than the non-backtracking matrix. The proposed matrix can estimate the number of communities in a sparse networks.

We start our paper by giving background and notation about graphs, used model and matrices. In second section, we present our main result. As last point, we give numerical results for our findings.

2 Preliminaries

Notation

With a long history, graphs are the tools that can represent a set of data visually, solve scientific problems and explain more about the real world. Any network graph is presented by points that called "nodes", and relations between these nodes, which are "edges".

A graph $G = (V, E)$, is composed by a set $V = \{1, ..., n\}$ of nodes and a set E of edges with $\{i, j\} \in E$ if it has an edge between i and j. Mathematically, the graph is represented by an adjacency matrix, with elements a_{ij} equal to 1 when there is an edge from vertex i to j, and 0 when there is no edge.

$$a_{ij} = \begin{cases} 1 & \text{if}(v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}$$

Stochastic Block Model

With a long history, the stochastic block model was created for detecting communities. For the SBM model, we denote a set of integers $\{1, \cdots, n\}$ with a binary symmetric matrix A (the adjacency matrix), and we set a number of blocks K. We denote a label vector $Z = (Z_i)_{i\in[n]}$ such that $Z_i = k$ if the nodes are k-labeled with a vector of block proportion $\alpha = (\alpha_1, \cdots, \alpha_K)$. The variables A_{ij} are independent Bernoulli variables. We note the

probability between two communities k and q is denoted by p_{kq}. Then we obtain the matrix of probabilities $P = (p_{kq})_{k,q} \in [K]$. Finally the model will be denoted $G(n, p, \alpha)$.

Non-backtracking matrix

Let m be the number of edges in an undirected network, and $2m = \sum_{i,j=1}^{n} A_{ij}$ be the number of edges in a directed network. We express the edge between nodes i and j by two directed edges, one from i to j and the other from j to i to generate the non-backtracking matrix.

The $2m \times 2m$ matrix, indexed by these directed edges, is defined by

$$\tilde{B} = \begin{cases} 1 & \text{if } j = k \text{ and } i \neq l \\ 0 & \text{otherwise} \end{cases}$$

The spectrum of the non-backtracking matrix was given in (6), (1) by

$$B = \begin{pmatrix} 0_n & D - I_n \\ -I_n & A \end{pmatrix}$$

Where 0_n is the $n \times n$ matrix of all zeros, I_n is the $n \times n$ identity matrix, and D is $n \times n$ diagonal matrix with degrees d_i on the diagonal. The non-backtracking matrix's informative eigenvalues are real-valued and isolated from the bulk of the radius under the SBM. The first K greatest eigenvalues in magnitude of $\tilde{B}$ in a network of K groups (or clusters) are real-valued and well separated from the bulk, which is contained in a circleof radius $\|\tilde{B}\|^{1/2}$.

3 Main results

In our proposed method, we take a matrix similar to the spectrum of the non-backtraking matrix, called the informative matrix given by

$$F = \begin{pmatrix} 0_n & (D - I_n)^2 \\ -I_n & H \end{pmatrix}$$

where H is the modified bethe-hessian matrix proposed by laala (7) when the author replace the adjacency matrix by using the laplacian. For the purpose to estimate the number of cluster in a network, especially sparse networks we need to create our infor- mative matrix. The informative matrix with the elements can estimate the number of cluster with a small average degree. The principal component analysis tend alwas to ex- tract principal elements. We note that we used the principal component matrix for our

matrix to take first principals from the adjacency. It is shown that, for some reason, the non-backtracking matrix and even the bethe-hessian matrix do not perform well in some condition. In the paper of Laala, when the author add parameters to the bethe hessian and taking the laplacian instead of the adjacency because as we know that it give information about network. We used that equation in the informative matrix with specified number of α, the parameter in which control the results.

As mentionned earlier, spectral methods can perform well. Our proposed matrix can work same role as the spectrum of non-backtracking matrix, but for sparse network when the last can't estimate. Our main result is based on the eigenvalues of the informative matrix, when we find that can gives better results by applying PCA for the matrix. The number of communities is calculated by counting the number of eigenvalues separated from the bulk.

4 Numerical results

We compare the empirical accuracy of estimating the number of communities by using the proposed matrix in this section. Using stochastic block model, we generate a network with a number of communities equal to K by noting a label vector $c \in \{1, \cdots, K\}^n$, so that $c_i = K$ if $n\pi_{K-1} + 1 \leq i \leq n\pi_K$, with $\pi_0 = 0$. We present communities with different sizes (different number of nodes in each cluster that we call unbalanced network). The proportion of nodes in each community π is given by $\pi_1 r/K$, $\pi_K = (2 - r)/K$, and $\pi_i = 1/K$ for $2 \leq i \leq K - 1$, where r is the ration of community size in which vary in the range [0.2, 1]. As the ration r increases, the community size become more similar, and become equal whe it is equal to 1. We note a matrix Z with dimension $n \times K$ that encode the label c. This matrix is given by $Z_{iK} = 1_{ci=k}$, it represent each node with a row of K element. Note by P , the matrix with diagonal element that control the edge densities within communities and and other that control out-in probabilities. We compare this method by different approaches to estimate the number of communities in network. For different values of number of communities, we generate with a network of 500 nodes under stochastic block model with a small average degree, a sparse network. We repeat the operation more than 100 times and we record the results.

Figures , show the result of estimating the number of communities by the informative matrix with red line, however the accurancy for the non-backtraking is zero among the plot, which means for different community size.

In figure , we generate the network under degree stochastic block model (DCSBM), with a smaller average degree, and we ramark that can also perform.

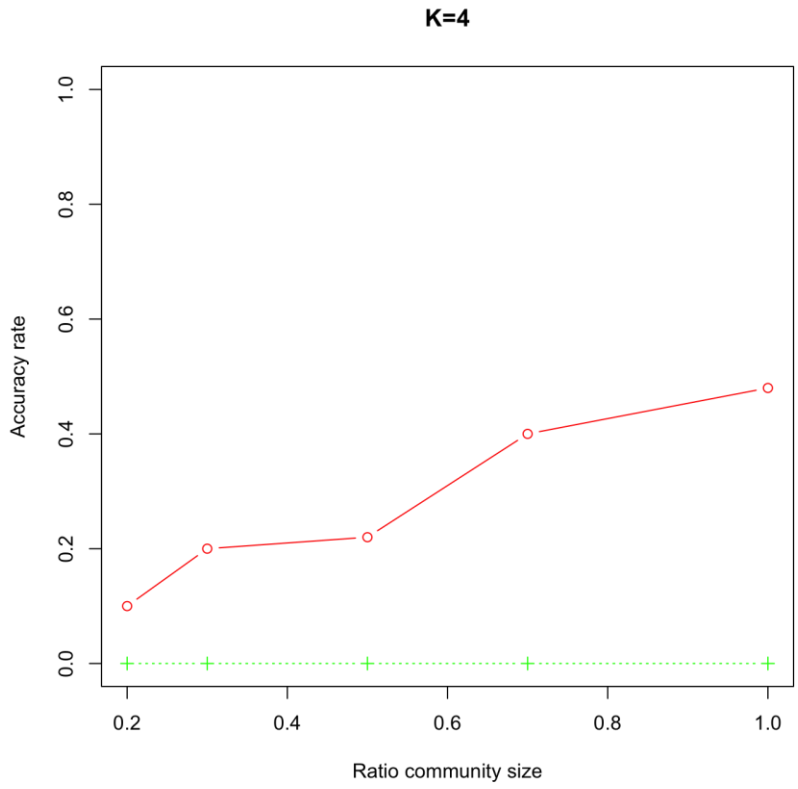

Figure 1: Estimate number of communities (for 4 communities) under SBM

5 Real-world network application

PolBooks network data is a network with 105 nodes. Nodes represent books about US politics published around the 2004 presidential election and sold by Amazon.com, an online bookstore. Edges represent frequent book co-purchases by the same buyers, as indicated by Amazon's customers who bought this book also bought these other books feature.

The second is Karate, the Zachary karate club network is well-known and widely used. Wayne Zachary collected data from members of a university karate club in 1977. Each node represents a club member, and each edge represents a tie between two club members. The network is undirected. Finding the two groups of persons into whom the karete club separated after an argument between two teachers is a frequently discussed problem utilizing this information.

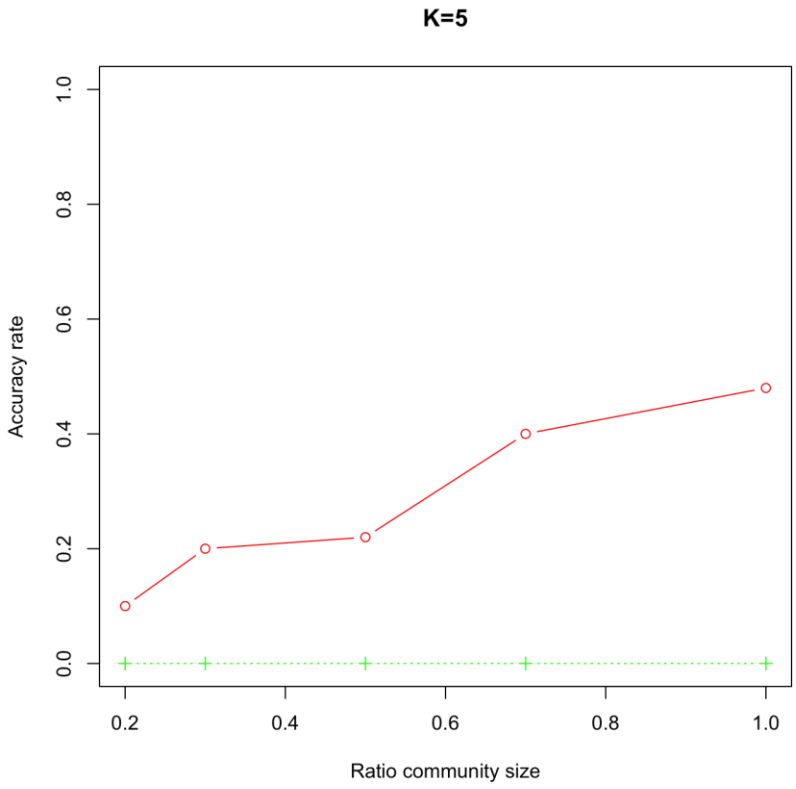

Figure 2: Estimate number of communities (for 5 communities) under SBM

6 Discussion

In this paper, we contribute results of estimating the number of community detection for sparse networks with low average degree under stochastic block model and for one case in degree corrected block model (DCSBM). To the best of our knowledge, this is the first study using the called informative matrix which is derivative from the non-backtracking matrix and which estimate better the number of communities. We note that the non-backtracking matrix used to estimate the number K in general but when the average degree is small it underestimate K. We support our results with numerical studies for synthetic networks, and for real network too.

In this paper, we only give for network under stochastic block model the estimation for the number of cluster. An important future work will be about a high accuracy rate, and studies under Degree corrected stochastic block model.

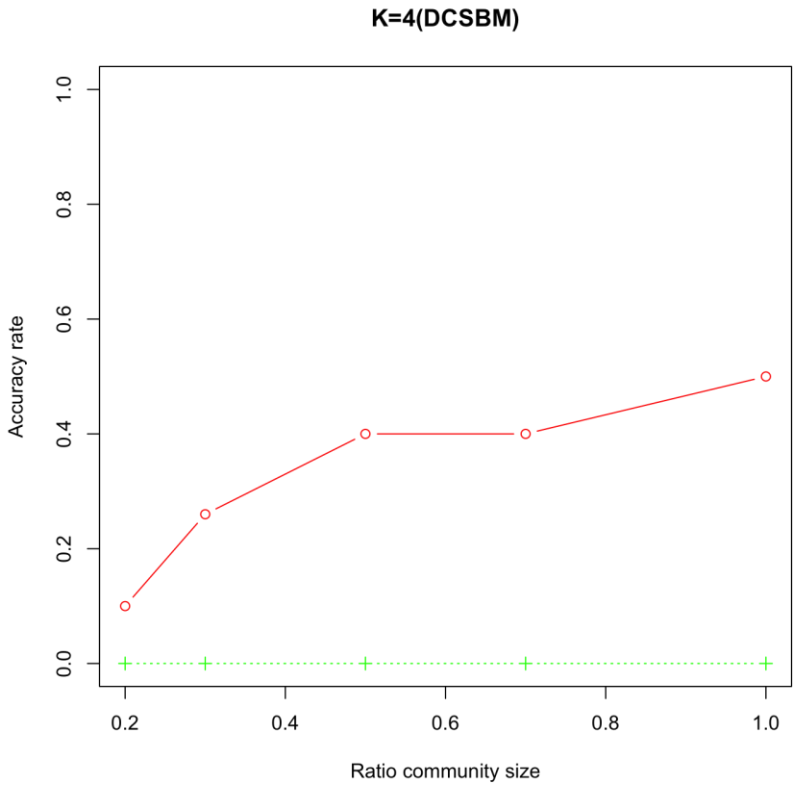

Figure 3: Estimate number of communities (for 4 communities) under DCSBM

References

[1] O. Angel, J. Friedman, and S. Hoory, The non- backtracking spectrum of the universal cover of a graph. *Preprint arXiv:0712.0192* (2007).

[2] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, Self-organization and identification of web communities. *Computer*, 35(3):66-71 (2002).

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

[4] K. Hashimoto, Zeta functions of finite graphs and representations of p-adic groups. *Adv. Stud. Pure Math.* 15, 211-280 (1989).

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

[6] C. Kemp, J. B. Tenenbaum, et al. Learning systems of concepts with an infinite relational model. *In AAAI*, 3, 5 (2006).

[7] F. Krzakala et al., *Spectral redemptionin clustering sparse networks*. *Proc. Natl. Acad.Sci*., 110(52):20935-20940 (2013).

[8] Laala Zeyneb, A Modified Method Using the Bethe Hessian Matrix to Estimate theNumber of Communities, *Journal of Advanced Statistics*. (2018)

[9] Michalis Faloutsos, Thomas Karagiannis, and Sue Moon. Online social networks. *IEEEnetwork*, 24(5):4-5, (2010).

[10] M. E. J. Newman, Spectral community detection in sparse networks. *arXiv:1308.6494* (2013).

[11] B. Y. Reis, S. K. Isaac, and D. M. Kenneth. An epidemiological network model fordisease outbreak detection. *PLoS medicine*, 4(6):e210, (2007).

[12] M. A. Riolo, G. T. Cantwell et al., Efficient method for estimating the number of communities in a network. *Physical review e*, 96(3):032310 (2017).

[13] A. R. Rossi and N. K, Ahmed, The network data repository with interactive graph analytics and visualization, AAAI, http://networkrepository.com (2015).

[14] A. Singh and M. Humphries, Finding communities in sparse networks. *Sci Rep 5*, 8828 (2015).

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

[16] U. von Luxburg, A tutorial on spectral clustering. *Statisics and Computing* 17, 395-416 (2007).

[17] W. W. Zachary, An information flow model for conflict and fission in small groups,
Journal of Anthropological Research 33 , 452-473 (1977).