Journal of Advanced Statistics, Vol. 3, No. 2, June 2018
https://dx.doi.org/10.22606/jas.2018.32001 

# A Modified Method Using the Bethe Hessian Matrix to Estimate the Number of Communities

Laala Zeyneb

Department of Statistics and Mathematics, Central China Normal University
Email: zinebla91@hotmail.com

**Abstract** The community detection is one of the main problems in social network analysis. Many methods are proposed to recover the community labels of nodes by assuming the number of communities is known. There has been an increasing interest to explore the communities number. We propose a fast method based on spectral proprieties of the graph to estimate the number K of groups. The method performs well, especially when the number of groups in network is large, and we focus on when the network is unbalanced.

**Keywords:** Stochastic block model; community detection, Bethe Hessian matrix.

## 1 Introduction

An article was written by the mathematician Leonhard Euler, presented at the academy of Petersburg in 1735 and published in 1741, treated the problem of Seven Bridges of Königsberg [14], the problem was to find a walk from a given point that would return to this point by passing once and only one by each of the seven bridges of the city Königsberg. After that, several works are developed in the topic of random graphs in many fields. In mathematics, a random graph is a graph that is generated by a random process. The first model of random graphs was popularized by Paul Erdös and Alfréd Rényi in a series of articles published between 1959 and 1968 [4].

One of the important point in the study of the random graphs is the community detection. The community detection aims to divide the network to many subgroups that have nodes strongly connected. Until the existence of community detection, it was created several number of methods and algorithms to find the number of communities K in the network. It helps to find more useful information, that can't find in studying the network as a whole. For example if we have a number K of communities, from some parameters such as the average vertex degree, we can extract some information and conclusion about the members of these communities, which could not be done by looking at the statistics for the whole graph. In addition how the nodes in the same group are highly connected but with few links with other node so brings more information about the network. It used for different domain like biology, computer science, .... For example, in the metabolic network, the communities present the biological functions of the cell [15], in the web graph the topics of interest are communities [12,5]. In the paper of Girvan and Newman in 2002 [8], that the problem was proposed for the first time. The authors began from the principe that in the real small-world networks, it exists a community structure, so they ask how to do an automatic detection for this structure. Since 2002, and the introduction of the algorithm [8], several algorithms were proposed, the case of the paper Fortunato [6] that was published in 2009, with more than 50 methods. The paper of Xie et al [18] in 2011 did the comparison of 10 methods published after 2009.

There are many ways to model a community structure, the stochastic block model is one of them, is a generative model, generally a choice for a network position. This model is used in the study of random graphs, and have a long tradition of social sciences and computer science [11,7,1,17,9]. Stochastic block model tends to produce graphs containing subgraphs, that are connected with particular edge densities. Here, we use spectral methods to estimate the number of communities under the stochastic block model. Especially we can show how to use the bethe hessian matrix to find the number of communities

 

## 2 The Bethe Hessian Matrix and the Relation with the Non-Backtracking Matrix

First, let us introduce some notations that will be used below. For a connected network graph $G = (V, E)$, with number of nodes equal to $|V| = n$, the adjacency matrix $A$ is defined as $A_{ij} = 1$ if $(i, j)$ is an element in the set of edges $E$, and $A_{ij} = 0$ otherwise. In our work, we focused on that we call "undirected networks", on which edges don't have orientations (i.e., the edge (i,j) is identical to the edge (j,i)) so the adjacency matrix is symmetric, on the other side the "directed networks" have orientations. We denote also the degree matrix by $D = diag(d_1, \cdots, d_n)$, with $d_i$ is the vertex degree (i.e., the number of links connecting this vertex), and the Laplacian matrix $L = D - A$.

The Bethe Hessian, sometimes called deformed Laplacian is defined as

$$H(r) = (r^2 - 1)I - rA + D$$

where $r \in \mathbb{R}$ is a parameter, and $I$ is the $n \times n$ identity matrix.

Before speaking about the relation between the Bethe hessian matrix and the non-backtracking matrix, let us first define the non backtracking matrix :

The non-backtracking matrix was defined first by Kiichiro Hashimoto [10], and is a matrix that can represents the structure of links of a network. It can be used to identify non-backtracking walks on a network. The non-backtracking matrix is defined for directed links, but often used to undirected, in our case (undirected network), we replace each of undirected link between $i$ and $j$, with a pair of directed $i \to j$ and $j \to i$. The matrix encodes information about sequence of links that can follow in a walk through the network, specifically if you just traversed the link $i \to j$, the matrix helps to tells what links $k \to l$ are permissible as the next step in your walk with no possibility of immediately backtracking from $i \to j$.
Mathematically, it's given by :

$$B_{i\to j,k\to l} = \begin{cases} 1 \text{ if } \ j = k \ \text{ and } \ i \neq l \\ 0 \text{ otherwise} \end{cases}$$

The $2n \times 2n$ matrix $\tilde{B}$ is the spectrum of B, known in [2] and [13] is defined by

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

where, $0_n$ is the $n \times n$ matrix of all zeros, $I_n$ is the $n \times n$ identity matrix, and $D = diag(d_i)$ is $n \times n$ matrix with degrees $d_i$ on the diagonal.

In [5], it observed that in a network of $K$ groups (or clusters), the first $K$ largest eigenvalues in magnitude of $\tilde{B}$ are real-valued and well separated from the bulk, which is contained in a circle of radius $||\tilde{B}||^{1/2}$. It was also observed that the spectral norm of the non-backtracking matrix is approximated by :

$$\tilde{d} = \left( \sum_{i=1}^{n} d_i \right)^{-1} \left( \sum_{i=1}^{n} d_i^2 \right) - 1. \tag{1}$$

## 3 A Modified Bethe Hessian Matrix to Estimate the Number of Communities K

It was observed that in [3], the number of communities is presented by the number of negative eigenvalues of $H(r)$. The choice of parameter $r$ is different from one to another. It takes $r_\lambda = \sqrt{\lambda}$, where $\lambda$ is the average degree of the graph, that is defined mathematically as the double number of edges by the number of vertex. This choice was the best choice in the paper of [16], and for general works, they argued that the best choice is $|r| = ||\tilde{B}||^{1/2}$, and the informative eigenvalues of $H(r)$, $||\tilde{B}||$ can be approximated by 1. In [3], different choices are given for the parameter $r$, in our work we give another estimate value to this parameter.

 

It was argued in [16] that the informative eigenvalues of $H(r)$ are negative when $r = ||\tilde{B}||^{1/2}$. We see that a choice was given to $r$ with values of $r_a = \sqrt{(d_1 + \cdots + d_n)/n}$ which proposed in [3] and this method was noted by BHa.

When the network is unbalanced, using the method of BHa tends to underestimate the number of communities especially when the number of clusters becomes large, because when the network is balanced that means every community has the same number of community which is not always true in reality. The challenge is trying to find a solution that can solve this problem when the network is unbalanced, not necessary that every community had the same size as the other . In our method of estimation, we used a parameter $\alpha$, with $\alpha \in [0, 1]$, and the modified Bethe Hessian matrix becomes :

$$\hat{H}(r) = \alpha(r^2 - 1)I - rA + D.$$

Since the parameter $r$ is very important in this estimation, we give to $r$ a new value, that related with the parameter $r_a$ used in [3]. So the new value of the parameter is $r_{new}$, with $r_{new} = ((d_1 + \cdots + d_n)/n)^{1/4}$, and we denote this method by $Bha_{New}$.

Now, the Bethe Hessian matrix is not ready yet to use it to calculate the number of communities. We know that both of adjacency matrix and laplacian matrix give information about the graph, so we use the laplacian in the place of the adjacency matrix, and we correspond the number of communities by the number of positive eigenvalues of $\hat{H}(r)$.

$$\hat{H}(r) = \alpha(r_{new}^2 - 1)I - r_{new}L + D$$

where $\hat{H}$ is the new modified bethe hessian matrix, $L$ is the laplacian matrix that we defined before, and the parameter $\alpha$ depend of the number of communities, in the case of large number ($K > 5$), we take the values of $[0.5, 1]$, else we take from the interval $[0, 0.4]$.

## 4 Synthetic Networks

We generate a network under the stochastic block model with number of communities equal to K, we note a label vector $c \in \{1, \cdots, K\}^n$, so that $c_i = K$ if $n\pi_{K-1} + 1 \leq i \leq n\pi_K$, where $\pi_0 = 0$. We are interested in the networks that have communities of different sizes, so we take the proportions of nodes falling into each community $\pi$ by setting $\pi_1 = r/K$, $\pi_K = (2 - r)/K$, and $\pi_i = 1/K$ for $2 \leq i \leq K - 1$, where $r$ is the community-size ratio that varies in the range $[0.2, 1]$. As $r$ increases, the community sizes become more similar, and are all equal when $r = 1$. The $n \times K$ label matrix $Z$ is to encode $c$ by representing each node with a row of K elements, exactly, one of which is equal to 1 and the rest are equal to 0, the matrix $Z$ is presented as $Z_{iK} = 1_{c_i=K}$. We take also a $K \times K$ matrix $P$ with diagonal $w = (w_1, \cdots, w_k)$ that tends to control the relative edge densities within communities and off-diagonal entries $\beta$, that controls out-in probability and the matrix $M = ZPZ^T$.

Under the stochastic block model, the adjacency matrix $A$ is generated according to an edge probability matrix $A = \mathbb{E}A$ proportional to $M$[3]. In our estimation, the number of nodes is given by $n = 2000$, the out-in probability ration $\beta = 0.2$, the average degree is take the value of $\lambda = 18$ for all figures, and the parameter $w$ is fixed by 1, $w_i = 1$, for all $1 \leq i \leq K$. We consider different values for the number of communities $K = 4, 5, 8, 10$, and 12. For each setting, we generate 200 replications of the network and record the accuracy, the accuracy is defined as the fraction of times that the method is correctly estimate the number of communities K by the length of total replication. First, we varied the community-size ratio to see the result for different clustering, so see the performance of our method when the network is unbalanced. We can see that the method performs well, with different number of groups. The figures 1 and 2 present the results of the method, especially when the number of groups K becomes large, for example when K=10, the method BHa cannot calculate the number, which can be done with the new method. In figure 3, we remark that the method gives some information when the others method did not.

## 5 Real World Network

In this section, we talk about the result of the performance of the new method for real network. We applied our algorithm to the college football network [8], where nodes are the 115 teams of US college

 

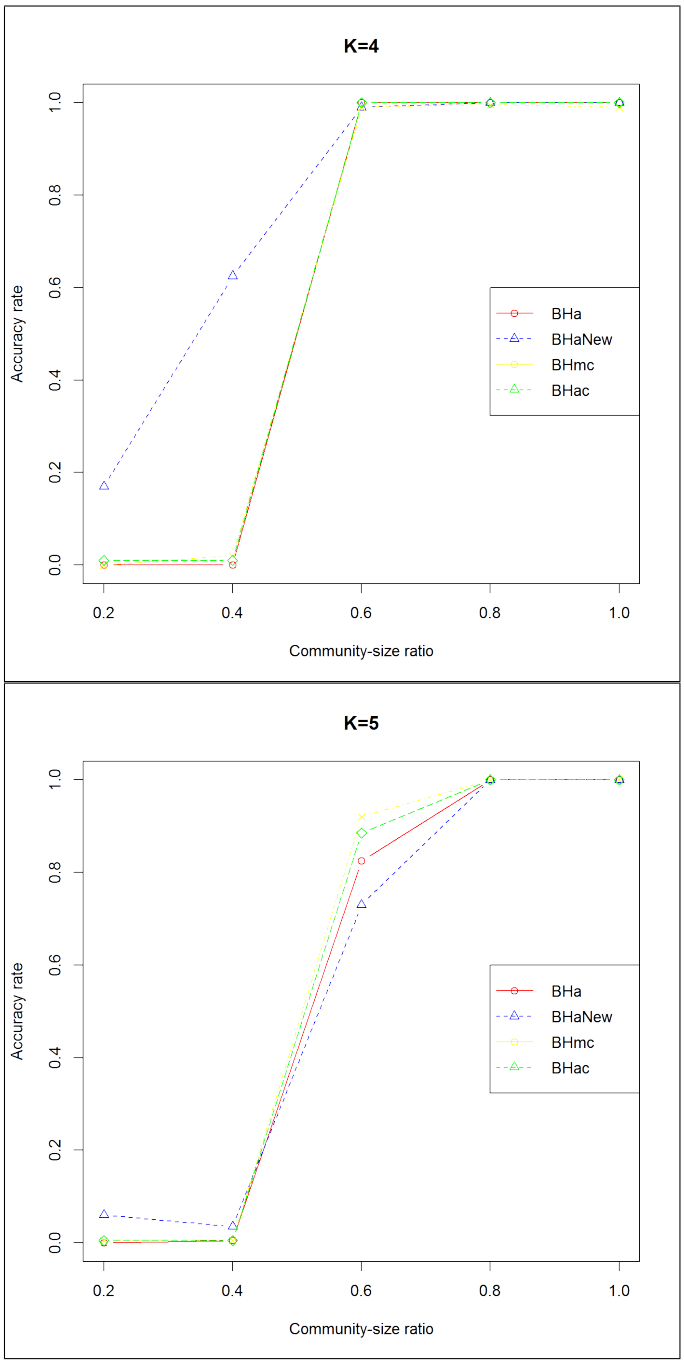


**Figure 1.** The accuracy of estimating K as a function of the community-size ratio r for K=4 and K=5, with $\alpha = 0.25$ and $\alpha = 0.29$ respectively.

football, and edges represents the game played in 2000. And with $\alpha = 0.4$, we found the number of communities that is equal to 12.

## 6 Discussion

Finally, we can remark that in this paper we have talked about how to find the method, or to find the best way to estimate the number of communities in a network. Also the spectral method as is known

 

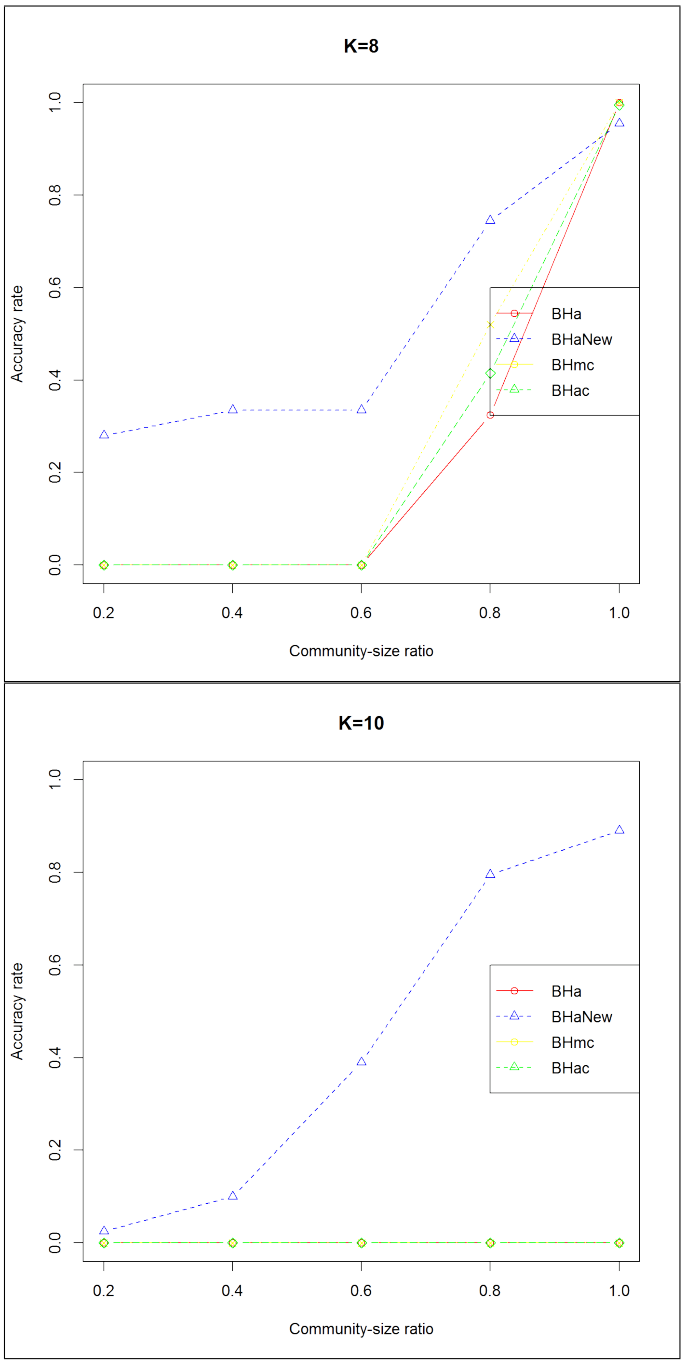


**Figure 2.** The accuracy of estimating K as a function of the community-size ratio $r$ for K=8 and K=10, with $\alpha = 0.62$ and $\alpha = 0.65$ respectively.

performs well and fast. Our method is especially used for the unbalanced networks that is the case of many network in real. Our suggestion can help to find the clusters. The method of approximation of the value $\alpha$ will be treated on future work.

## References

1. C. J. Anderson, S. Wasserman and K. Foust , *Social networks.* 14,135 (1992).

 

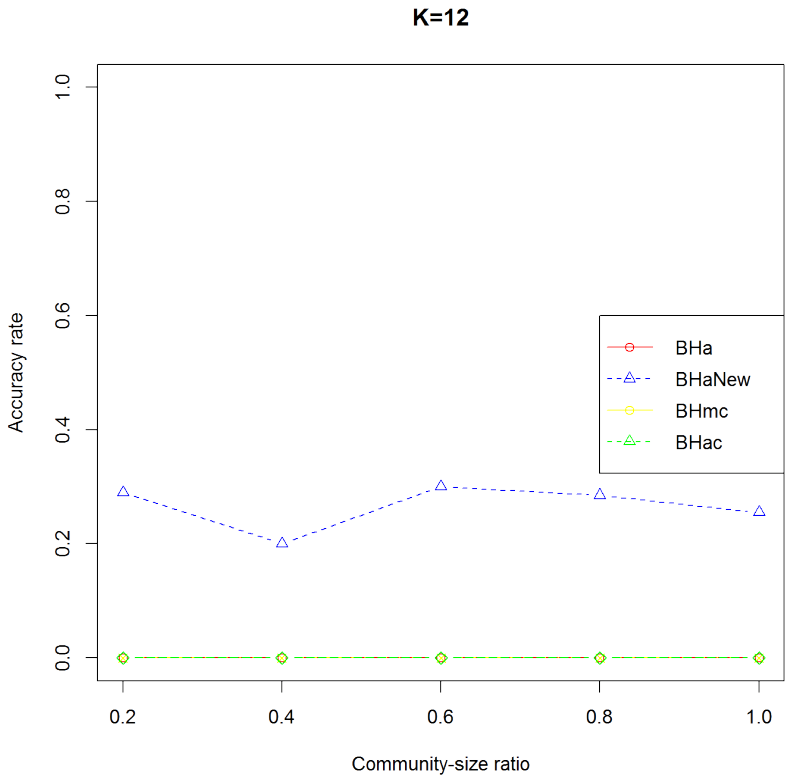


**Figure 3.** The accuracy of estimating K as a function of the community-size ratio $r$ for K=12, with $\alpha = 0.64$ .


2. O. Angel, J. Friedman, and S. Hoory, *The non-backtracking spectrum of the universal cover of a graph.* arxiv:0712.0192, 2007.
3. Can M. Le, Elizaveta Levina, *Estimating the number of communities in networks by spectral methods.* arxiv:1507.00827v1, 3 Jul 2015.
4. Erdös, P. Rényi, *On Random Graph. I.* Publicationes mathematicae. 6, 1959, 290-297.
5. 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.
6. S. Fortunato. *Community detection in graphs. Physics Reports.* 486(3): 75-174, 2010.
7. K. Foust and S. Wasserman , *Social networks.* 14,5(1992).
8. M. Girvan and M.J. Newman. *Community structure in social and biological networks.* Proceedings of the national Academy of sciences, 99(12) : 7821-7826, 2002.
9. A. Goldenberg, A. X. Zheng, S.E. Feinberg, and E. M. Airoldi , *Foundations and Trends in machine learning.*2, 1(2009).
10. K. Hashimoto, *Zeta functions of finite graphs and representations of p-adic groups.* Adv.Stud. Pure Math. 15. 211-280 (1989).
11. P. W. Holland, K. B. Laskey, and S. Leinhardth , *Social networks.* 5,109(1983).
12. Jon Kleinberg and Steve Lawrence, *The structure of the web.* Science, 294(5548):1849Ű1850, 2001.
13. F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborov, and P. Zhang, *Spectral redemption in clustering sparse networks.* Proceedings of the National Academy of sciences, 110(52):20935-20940, 2013.
14. Leonhard Euler. *Commentarii academiae scientirum.* Petropolitanae 8, 1741, page 128-140)
15. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabàsi, *Hierarchical Organization of Modularity in Metabolic Networks.* Science, 297(5586):1551Ű 1555, 2002.
16. A. Saade, F. Krzakala, and L. Zdeborová, *Sectral clustering of graphs with the Bethe Hessian,* September 9, 2014.
17. T. A. Snijders and K. Nowicki , *Journal of classification.* 14, 75(1997).


 

18. J. Xie, S. Kelley, and B.K Szymanski. *Overlapping community detection in networks: the state of the art and comparative study.* arxiv: 1110. 5813, 2011.

 