Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 10, October -2017

COMMUNITY DETECTION BASED ON LAPLACIAN SPECTRAL PARTITIONIG METHOD

Harekrishna Mili

[1]*Department of Mathematics, Dibrugarh University*

Abstract — *Many methods have been proposed for community detection in networks. Community detection is the division of a network into dense sub networks with only sparse connections between them, has been a topic of vigorous study in recent year. In this paper, show that inference methods can be reduced to minimum cut partitioning problem, which helps to solve solution of the community detection problem. Here community inference, testing the resulting algorithm on computer generated and real world networks are perform by adapting the Laplacian spectral partitioning method both running time and quality of the results gives rival the best previous methods.*

Keywords*-community detection, network, Laplacian spectral partitioning method, community*

I. INTRODUCTION

In recent years, the problem of community detection in a network is one of the most famous problems in research area. Many methods are used to determine community structure, but in recent years statistical inference methods used widely, because they give excellent results. In this paper, two fundamental methods is used, based on the stochastic block model or its degree corrected variant. It is possible to map both methods the minimum cut graph partitioning problem, for which we can take any available methods for graph partitioning to the community detection problem. Here Laplace spectral partitioning method is apply to derive a community detection method to get better results than the best currently available algorithms.

II. LIKLIHOOD MAXIMIZATION FOR THE STOCHASTIC BLOCK MODEL

Our first method is based on the stochastic block model, also known as the Planted partition model, is well known model of community structure in networks. Here we consider a network of n vertices and make some small groups or communities and there are different probabilities for different connections within and between groups. For our simplicity, we consider only two groups of any size. In the model, edges are placed randomly between vertex pairs with probability pin for the pairs of same group and the probability pout for the pair of different groups. In this paper, a Poisson distributed number of edges are placed for the pairs of in the same groups with mean ωin and for the pairs of different groups with ωout. The fraction of possible edges that are actually present in the real world network is very small, for which the model is describe in [7] and the Poisson version of the model are indistinguishable, but the analysis of Poisson version is preferred, because its analysis is more straightforward. The statistical inference of community structure is matter of answering the problem that what were the values if win and ωout used to generate the network and more importantly, which vertices fell in which group? We use maximum likelihood method to the question. Labeling the two communities or groups in our model by group1 and group2 and gi denote the group in which vertex i belongs.
In the network, edges are represented by an adjacency matrix having elements

$$A=\begin{cases} 1 \text{ if there is an edge between vertices } i, j, \\ \qquad 0 \ otherwise \end{cases}$$

Then in a network or graph of the likelihood G. Given a complete set of group membership, denoted by g and the poisson parameter denoted by ω and is given by

$$P(G|g,\omega) = \prod_{i<j} \frac{\omega_{ij}^{A^{ij}}}{A_{ij}!} e^{-\omega_{ij}} \qquad \text{…. (2)}$$

Where, ωij is expected number of edges between I and j vertices is equal to ωin or ωout, depending on whether the vertices are belongs to the same group or different groups. Here we consider that in the network, there is no self loop. So Aii = 0 for all i.
By differentiating the likelihood was maximized with respect to the parameter ωin and ωout [7],

In equation (2), we can apply this method which gives most likely values of ωin and ωout such that

$$\mathrm{l}\omega_{in} = \frac{2m_{in}}{n_1{}^2+n_2{}^2}, \omega_{out} = \frac{m_{out}}{n_1 n_2} \qquad \text{……………………………….(3)}$$

Where, min= observed number of edges with the groups. mout= observed number of edges between groups.

And n1 and n2 are the number of vertices in each group, group1 and group2. Putting these values in equation (2), we get the proper likelihood and it depends on the group labels only.

Neglecting some unimportant constants, the logarithm of the profile likelihood is Q= m_{in} $\ln\frac{2m_{in}}{n_1{}^2+n_2{}^2} + m_{out}\ln\frac{m_{out}}{n_1{}^2+n_2{}^2}$ (4)

By maximizing this quantity, the communities are identified over all possible assignments of the vertices to the groups.

In equation (4), at first the likelihood is maximized over ω, for fixed group assignment, and then group assignments. Considering reverse approach, for given ω , maximizing first over the group assignments and then ω at the end. This approach helps us in such a way that as we will show the standard of minimum cut graph partitioning is equivalent to the problem of maximizing with respect to the group assignments when ω is given for which many methods are available.

And the remaining problem of maximizing with respect to ω after maximizing with respect to the group assignments is a one parameter optimization which can be trivially solved.

The resulting algorithm is that the problem of maximum likelihood community detection is reduced to a well known method, graph partitioning with extra one step.

So maximizing the logarithm L of the likelihood,

L= ln P(G|g,ω) = $\sum_{i<j}[A_{ij}\ln\omega_{ij} - \omega_{ij}$ - lnA_{ij}!] …………………(5)

Which gives the same result of the problem of maximizing the likelihood, equation (2) with respect to the group labels gi for given values of parameter ωin and ωout.

For further steps, we write ωin and lnωij as ωij = δgigjωin + (1 – δgigj)ωout ……(6)

lnωij= δgigjlnωin + (1 – δgigj) lnωout ……. (7)

Here δij is Kronecker delta.

Substituting these results in equation (5) and neglecting some terms, which have no effect in position of maximum, the likelihood can be written as

L=$\sum_{i<j}(1 - \delta_i\delta_j)$(γ-$A_{ij}$) ……………(8)

Where γ=$\frac{\omega_{in}-\omega_{out}}{ln\omega_{in}+ln\omega_{out}}$. ……………….(9)

Which gives the position value whenever ωin - ωout > ϒ, which means in our network, we have traditional community structure.

And, $\sum$i <j (1 - δgigj) is called cut size, which is the number of edges connecting vertices in different communities

$\sum$i <j (1 - δgigj) = n1n2 …. (10)

We previously denoted the cut size by mout, so the log likelihood is in the form

L = - mout + ϒn1n2 …. (11)

For maximizing equation (11), its dependent on the value of ϒ, which depend on ωin and ωout via (9). Then for simplicity, we consider maximization of (11), where the sizes n1 and n2 are considered fixed and so the term ϒn1n2 is a constant and we can neglect this term. Then in equation (11), -mout is the only term to be maximized. Now this problem is equivalent to the standard minimum cut problem of graph partitioning.

In the two groups, there are (n+1) possible choices of the sizes. For each of these (n+1) possible choices, if we solve the minimum cut problem, then we will get (n+1) solutions and it is obvious that out of these (n+1) solutions, one of solution must be the solution to our overall maximum likelihood problem. To find that one solution, we simply calculate the profile likelihood equation (4), for each solution and in turn and find the one that gives the largest result.

A. Spectral Algorithm:-

For this approach, we consider spectral algorithm which is based on Laplace spectral bisection method of graph partitioning introduced by Fiedler [8, 9].

In this method we describe how to search an edge separator of a graph G. We can get two parts of the network G with the network G with the number of vertices n1 and n2 and also require the size of cutting edges (cut-size) to be small.

This method shown that two parts of a network of specified sizes can be found with minimum cut-size by calculating the Fiedler vector which is a Eigen vector of Laplacian matrix of the network corresponding to the second smallest Eigen value. To get the Fiedler vector, the network is divided into group of required size n1 and n2 and assigning the n1 vertices to group1 and the remaining n2 to group2.

A good feature is that in a single calculation in this approach, we can be known the entire one parameter family of minimum cut division of the network. For calculate the profile likelihood of the resulting divisions of the network, we calculate the Fiedler vector only once, and then sort the elements of Fiedler vector in decreasing order, then cut them into two groups in each of the n+1 possible ways. The maximum likelihood community division of the network is the one with the highest score.

B. Degree Corrected Block model:-

In case of most of real world networks, the standard block models gives the poor results because the degree distributions that real world network posses, it fails to explain.
By replacing the expected number ωij of edges between i and j by the term kikjωij where ki is the degrees of vector I, the problem of the standard models can be solved. Then log likelihood and by profile likelihood is given as

$$L= -m_{out} + \gamma\kappa_1\kappa_2 \text{ , } Q = m_{in} \ln\frac{2m_{ij}}{\kappa_1^2+\kappa_2^2} + m_{out} \ln\frac{m_{out}}{\kappa_1\kappa_2} \quad \text{……………………..(12)}$$

Where k1 and k2 are the sum of the degrees of the vertices in the two groups.

The maximized of L is reduced to a generalizes minimum cut partitioning problem, which again favors g balanced groups with a term proportional to k1k2. By the equivalent of our previous approach, holding k1 and k2 constant, without knowing the value of ϒ, we reduced the problem to a variant of the minimum cut problem. Again based on the graph Laplacian, a spectral algorithm for this problem can be derived. We can show for the standard spectral method that a good approximation of the minimum cut problem with fixed k1, k2 is given by the second Eigen vector of the generalized Eigen system Lv = λDv, not by the second Eigen vector of Lv. Again we find out the vector and make two groups of the vertices according to their vector elements and sizes and once again from this we have n+1, one parameter family of solutions which can be choose an overall winner by finding the one with the highest profile likelihood, equation (2).

iii. Results:-

This method is tested on a variety of networks and a good result is obtained. In first figure, figure1 it shows the result from tests on a large group of computer-generated networks- which were using the standard stochastic block model to generate themselves. In figure1, there are two parts. In our part, the result of profile likelihood for the families of n+1 solutions generated by the spectral calculation for networks with equally sized groups (top) and in second part, with unequal groups (bottom).

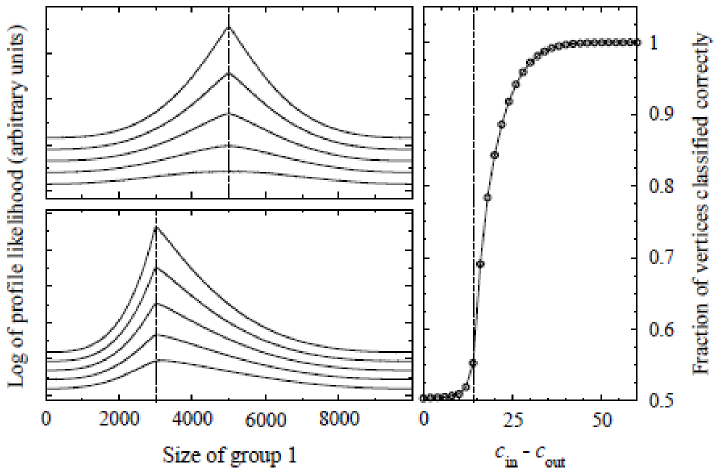

FIG. 1: (a) Profile likelihood as a function of group size for candidate solutions generated from the spectral method for single network of n = 10000 vertices, generated using the standard (uncorrected) stochastic block model with equal group sizes of 5000 vertices each and a range of strengths of the community structure. Defining cin = nωin, cout = nωout, the curves are (top to bottom) cin = 80, 75, 70, 65, and 60, and cout = 100 – cin. The dashed vertical line indicates the true size of the planted communities. The curves have been displaced from one another vertically for clarity. The vertical axis units are arbitrary because additive and multiplicative constants have been neglected in the definition of the log likelihood. (b) Profile likelihoods for the same parameter values but unequal groups of size 3000 and 7000. (c) The average fraction of vertices classified correctly for networks of 10000 vertices each and two equally sized groups. Each point is an average over 100 networks. Statistical errors are smaller than the points in all cases. The vertical dashed line indicates the position of the “detectability threshold” at which community structure becomes formally undetectable [6, 14–16].

In the profile likelihood in each case, there is a clear peak at the correct group sizes, from which can say that the group membership of most vertices is correctly identified by the algorithm. The third part of figure1 gives the conclusion of the community structure by calculating the fraction of correctly identified vertices as a function of the strength of equally sized groups. The “detectability threshold” is represented by the dashed vertical line, discussed in [6, 14- 17], every method of community detection must fail, below it. The algorithm is also fail below this point, but essentially works well all the way down to the transition, and the result is exact for the dense network [16].

Figure2 shows the result of the Zachary's "karate club" network [6] and Adamic and Glance's network of political blogs [9], that algorithm is working fast. Both have pronounced community structure and the division is found by spectral algorithm, the algorithm is working fast. The Lanczos method, an iteration methods are used to find Eigen vector which take time O(m) per iteration in which maximum likelihood profile can be achieved, where m is the number of edges in the network. The exact number of iteration is although not known; the number of iteration should be small.
For the (n+1) different group, each group differs from other just because of single vertex movement between the groups.

By changing the quantities appearing in equation (2) according to $K1 \Rightarrow k1- ki,\ \ k2\Rightarrow k2+ ki$ .... (13)

$$min \Rightarrow min-\Delta m\ ,\qquad mout \Rightarrow mout + \Delta m \qquad .... (14)$$

the vertex i moves between groups. Where Δm is the number of edges between the group1 and i minus the edges between vertices of group2 and i. The changes of profile likelihood and these quantities can be found in time proportional to the degree of the vertex. So all the n vertices can be moved in time proportional to 2m, which is the sum of the degrees in the network. So the algorithm takes time m time of Lanczos iteration, which is treated as small, and so the method is fast as the best competing algorithm.

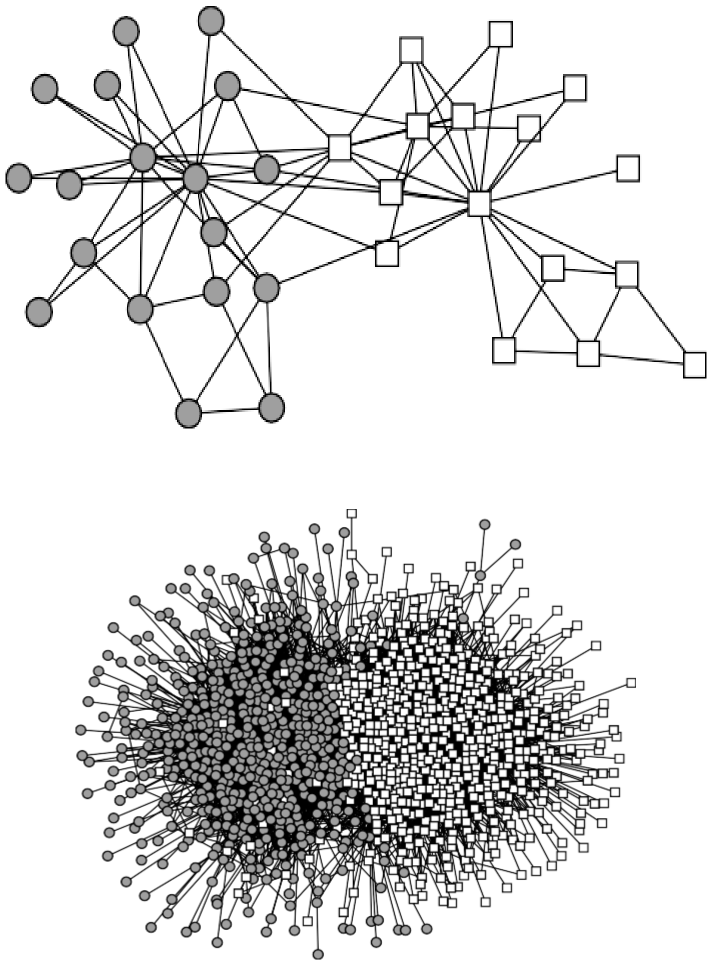

FIG. 2: The division into two groups of two well-known networks from the literature. Top: the karate club network of Zachary [18]. Bottom: the network of political blogs compiled by Adamic and Glance [19]. Vertices colors and shapes indicate the group membership and both divisions are qualitatively similar to the accepted ones.

CONCLUSION

On this paper shown that, community detection in a network, maximum likelihood method can be reduced to a small family candidate solutions, where each of which itself a solution of well studied maximum cut graph partitioning problem. As an example, to test its performance on both the real work and synthetic networks, the Laplacian spectral partitioning method is used.
In future, we can study more general forms of the parameter ω, for different values of ωin and ωout for more than two groups.

REFERENCES

[1] M. Girvan and M. E. J. Newman, Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002).

[2] S. Fortunato, Community detection in graphs. Phys. Rep. 486, 75–174 (2010).

[3] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, Mixed membership stochastic block models. Journal of Machine Learning Research 9, 1981–2014 (2008).

[4] A. Clauset, C. Moore, and M. E. J. Newman, Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101(2008).

[5] P. J. Bickel and A. Chen, A nonparametric view of network models and Newman–Girvan and other modularities. Proc. Natl. Acad. Sci. USA 106, 21068–21073 (2009).

[6] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborova´, Inference and phase transitions in the detection of modules in sparse networks. Phys. Rev. Lett. 107, 065701 (2011).

[7] B. Karrer and M. E. J. Newman, Stochastic block models and community structure in networks. Phys. Rev. E 83, 016107 (2011).

[8] M. Fiedler, Algebraic connectivity of graphs. Czech. Math. J. 23, 298–305 (1973).

[9] A. Pothen, H. Simon, and K.-P. Liou, Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. 11, 430–452 (1990).

[10] A. Condon and R. M. Karp, Algorithms for graph partitioning on the planted partition model. Random Structures and Algorithms 18, 116–140 (2001).

[11] U. Elsner, Graph partitioning—a survey. Technical Report 97-27, Technische Universit¨at Chemnitz (1997).

[12] P.-O. Fj¨allstr¨om, Algorithms for graph partitioning: A survey. Link¨oping Electronic Articles in Computer and Information Science 3(10) (1998).

[13] M. E. J. Newman, Networks: An Introduction. Oxford University Press, Oxford (2010).

[14] J. Reichardt and M. Leone, (Un) detectable cluster structure in sparse networks. Phys. Rev. Lett. 101, 078701 (2008).

[15] D. Hu, P. Ronhovde, and Z. Nussinov, Phase transitions in random Potts systems and the community detection problem: Spin-glass type and dynamic perspectives. Phil. Mag 92, 406–445 (2012).

[16] R. R. Nadakuditi and M. E. J. Newman, Graph spectra and the detectability of community structure in networks. Phys. Rev. Lett. 108, 188701 (2012).

[17] E. Mossel, J. Neeman, and A. Sly, Stochastic block models and reconstruction. Preprint arXiv:1202.1499 (2012).