The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009
Copyright © 2009 ORSC & APORC, pp. 337–344

A Linear Projection Approach for Resolving Community Structure *

Xiaoping Liao[1,†] Wei Ren[1] Guiying Yan[1]

[1] Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100190, China

Abstract Networks are widely used in the social, physical, and biological sciences as a concise representation of the topology of systems. In order to understand the structure of these networks, it can be helpful to decompose the network into communities. In this paper, we propose a linear projection approach for detecting community structure by transforming network community detection problem into a low-dimension vector clustering problem. Furthermore, the optimal number of communities can be inferred by using the gap statistic idea, if no prior information is provided.

Keywords Community structure, Linear projection, PCA, K-means clustering, Gap statistic

1 Introduction

In the last few years, complex networks have been studied extensively due to their relevance to many real systems such as the world-wide web, the Internet, the traffic network, social and biological networks [1]. So far, A few interesting properties have been identified in these networks, such as small world phenomenon[2] and power law distribution[3]. Recently, many considerable interests focus on another property called community structure property[4], which refers to the occurrence of groups of nodes in a network that are more densely connected internally than with the rest of the network.

A variety of methods[5, 6]have been developed to detect communities. And for evaluating the goodness of community structure, the modularity measure is proposed[5], which works very well in networks with balanced structure. Afterwards, many researchers propose their algorithms by using modularity as the quality function [7, 8, 9]. But unfortunately, it fails to work for unbalanced networks [10]. While, the information based method [11], the mixture model [12], the SPAEM [13], can partly solve this scale problem.

Intuitively, the community is similar to the cluster. In the data mining field, a cluster usually refers to a set of closely located vectors. Now, what is the relationship between network communities and vector clusters? A spectral clustering method [7] is proposed to find this kind of relationship, which first computes the embedding of the graph into a Euclidean space and then clusters these vectors by applying K-means clustering algorithm. However, this algorithm also uses modularity as the quality function to choose the optimal number of clusters.

*This paper is supported by the National Natural Science of Foundation of China under Grant Nos. 10531070, 10721101, and KJCX-YW-S7.

†Email: xpliao@amss.ac.cn

In this paper, we introduce a new way to transform network community detection problem into a common clustering problem, which turns out to work very well on both well known networks and simulating networks even if networks are embedded into a relative low dimension Euclidean space. And by using the idea of gap statistic, which is quite different from modularity function, the optimal number of communities can be inferred, if there is no prior knowledge about it.

2 Method

Our goal is to detect community structure of networks of which adjacent matrix A are given, where $A_{ij} = 1$ means an edge from node i to node j. The general framework of our model is following: first, calculate a similarity measure X between every node pair; second, do vectorization of the network by using linear projection method so that each node is represented by a vector; third, find clusters in these vectors by applying clustering algorithm, if the number of communities is predetermined. And then transform each cluster back to a node set, which is the final community detected; fourth, infer the optimal number of communities by using the gap statistic idea, if there is no information about it. The detail of the framework will be explained in the next few subsections.

2.1 Similarity matrix calculating

This step serves as a preliminary preparation. All algorithms which define similarity can be incorporated here, such as the shortest path similarity matrix or diffusion kernel. Different algorithms may impact the final results. Surprisingly, an interesting phenomenon happens in our linear projection method which dose not seem to be very sensitive to the choice of similarity matrix. This will be illustrated in the result section.

2.2 Vectorization by using Linear Projection Method

Given the similarity matrix X, the goal of vectorization is to represent each node by a p dimensional vector which maintains the similarity measure as much as possible. Here, treat every column of X as a variable, and every row as an observation. Then there are n observations $\{x_i, i = 1,2,...,n\}$.

Consider the following optimization problem:

$$\text{(MP)} \quad \begin{array}{ll} \min_{\mu,\{\lambda_i\},V_p} & f(x) = \sum_{i=1}^{n} \|x_i - \mu - \lambda_i V_p^T\|^2 \\ \text{s.t.} & V_p^T V_p = I, \end{array} \tag{1}$$

Here, both μ and λ_i are $1 \times p$ vectors, V_p is a $n \times p$ matrix. By setting the derivatives of $f(x)$ in Eq(1) to be zero, then

$$\hat{\mu} = \bar{x} - \frac{\sum_{i=1}^{n} \hat{\lambda}_i V_p^T}{n} \tag{2}$$

$$\hat{\lambda}_i = (x_i - \hat{\mu})V_p \tag{3}$$

and by substituting $\hat{\mu}$ in Eq(2) into Eq(3),

$$\hat{\lambda}_i = (x_i - \bar{x})V_p + \frac{\sum_{i=1}^{n} \hat{\lambda}_i}{n} \tag{4}$$

This leaves us to find the orthogonal matrix V_p to the following optimization problem:

$$\min_{\mu,\{\lambda_i\},V_p} f(x) = \sum_{i=1}^{n} \|x_i - \bar{x} - (x_i - \bar{x})V_pV_p^T\|^2 \tag{5}$$

It is clearly that the optimal $\hat{V}_p$ takes first p principle component directions as columns. Therefore the solution to Eq(5) can be expressed as follows. First of all, stack the (centered) observations($\tilde{x}_i = x_i - \bar{x}$) into the rows of an $n \times p$ matrix $\tilde{X}$, and then construct the singular value decomposition of $\tilde{X}$:

$$\tilde{X} = UDV^T \tag{6}$$

Here, D is a $n \times n$ diagonal matrix, with diagonal elements $d_{11} \geq d_{22} \geq ... \geq d_{nn} \geq 0$. Without much effort, it can be proven that the first p columns of V is exactly $\hat{V}_p$, the solution to Eq(5).

The vectors $\{(x_i - \bar{x})\hat{V}_p, i = 1, ..., n\}$ are the optimal projection of centered observations ($\tilde{x}_i = x_i - \bar{x}$) to the vector space spanned by the base $\{v_1, ..., v_p\}$, where v_i is the ith column of $\hat{V}_p$.

Our projection idea is based on Principle Component Analysis(PCA). As we know, PCA can be used for dimensionality reduction in a data set by keeping lower-order principal components and ignoring higher-order ones, which retain those characteristics of the data set that contribute most to its variance. Such low-order components often contain the "most important" aspects of the data. However, depending on the application this may not always be true. Fortunately, PCA turns out to be suitable for resolving community structure.

2.3 Clustering

After vectorization of the network, each node i is represented by a p-dimension vector

$$\hat{\eta}_i = (x_i - \bar{x})\hat{V}_p = (u_{i1}d_{11}, u_{i2}d_{22}, ..., u_{ip}d_{pp}) \tag{7}$$

and proportion of variance accounted for is

$$r_p = \frac{\sum_{i=1}^{p} d_{ii}^2}{\sum_{i=1}^{n} d_{ii}^2} \tag{8}$$

To detect community structure, we only need to cluster these n vectors $\{\hat{\eta}_1, \hat{\eta}_2, ..., \hat{\eta}_n\}$ by applying clustering algorithm, such as K-means clustering algorithm or hierarchical clustering algorithm, if a predetermined number of clusters is given. In this paper, K-means clustering algorithm is employed.

2.4 Model selection

The K-means clustering algorithm needs a predetermined number of clusters. In fact, there are a lot of methods for estimating the number of clusters[15, 16], besides the modularity method[5]. However, whether these methods are suitable for our projection idea is still unknown. Fortunately, the gap statistic idea turns out to be appropriate to infer the optimal number of clusters[16], which works very well when combining with the shortest path similarity matrix.

Gap statistic

Step 1: Cluster the n vectors, varying the total number k of clusters from 2 to K, giving within-cluster measure W_k

Step 2: Generate B reference data sets, and cluster each one giving within-cluster measures $W_{kb}, b = 1, 2, ..., B, k = 2, .., K$. Compute the gap statistic

$$Gap(k) = (1/B) \sum_b \log(W_{kb}) - \log(W_k)$$

Step 3: Let $ew = (1/B) \sum_b \log(W_{kb})$, compute the standard deviation

$$sd_k = [(1/B) \sum_b \{\log(W_{kb}) - ew\}^2]^{1/2}$$

and define $s_k = sd_k \sqrt{1 + 1/B}$

Step 4: Choose the number of clusters via

$\hat{k}$= smallest k such that $Gap(k) \geq Gap(k+1) - s_{k+1}$

2.5 Linear Projection Algorithm(LPA)

Now we can describe our algorithm as follows:

Input: Adjacent matrix A

1. Calculate similarity matrix X
2. Choose a suitable dimension p, and calculate projection vectors $\{\hat{\eta}_1, \hat{\eta}_2, ..., \hat{\eta}_n\}$
3. If the number of communities is predetermined, clustering $\{\hat{\eta}_1, \hat{\eta}_2, ..., \hat{\eta}_n\}$ by applying K-means clustering algorithm, and then transform each cluster back to a node set, which is the final community detected
4. Infer the optimal number of communities k by using the gap statistic idea, if there is no information about it

Output: Communities $A_1, A_2, ..., A_k$

3 Results

3.1 Experiment

First, test LPA on some classical networks, using the shortest path similarity matrix.

3.1.1 Zachary Club

Zachary club network is based on acquaintance relationship between 34 members of a Karate club. The club splits into two parts due to an internal disputation, so it has natural community structure.

Here, test p from 1 to 34. When $p = 2$, the Scatter is shown in FIG. 1. Clearly, the 2 dimension vectors have obvious clusters. When p is bigger than 2, the original community structure can always be identified, see FIG. 2. Zachary club is divided into two communities(nodes with different shapes). Actually, the proportion r_p increases as p increases. Therefore, more information have been maintained after projection, which benefits the detecting community process. This result shows that LPA is suitable for detecting community, even if p is very small.

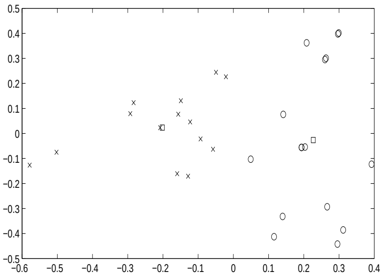

Figure 1: scatter: p=2. The projection vectors have obvious clusters(nodes with different shapes), and the square nodes denote two cluster centers

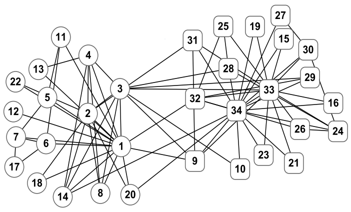

Figure 2: Zachary club: p=2, r=0.41.

3.2 Comparison

3.2.1 Dolphin Social Network

Dolphin social network reported by Lusseau [17] provides a natural example where communities vary in size. The original two subdivisions have different sizes, with one community 22 dolphins and the other 40. When $p \geq 1$, the result of LPA does not change, as shown by the left line in the FIG. 3. LPA outperforms the modularity method in this example, with only one node "SN89" misclassified.

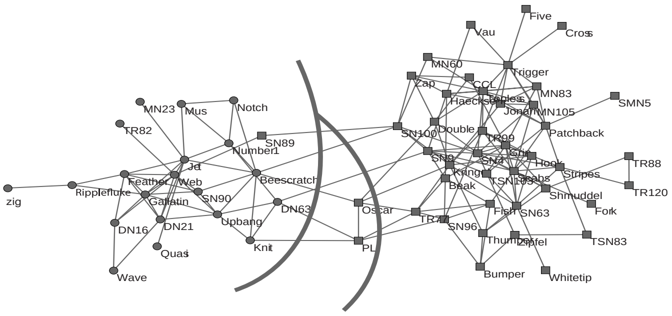

Figure 3: p=1, r=0.34. Nodes with different shapes denote the original partition. The left line is the split line by LPA, the right line is the split line by the modularity method[9]. LPA misclassifies only one node "SN89".

3.2.2 Simulating Network

Consider symmetric, node asymmetric, link asymmetric three cases as mentioned in [11]. In the symmetric test, each network is composed of 4 communities with 32 nodes, and each node have an average degree of 16. k_{out} is the average number of edges linking to nodes in different communities, and set $k_{out} = 6, 7, 8$. In the node asymmetric test, each network is composed of 2 communities with 96 and 32 nodes respectively, and k_{out} is set

Test	k_{out}	SP	DK	Info	Modu	Mixture	SPAEM	Spec
Symmetric	6	0.99	0.99	0.99	0.99	0.92	0.99	0.99
	7	0.95	0.96	0.97	0.97	0.81	0.95	0.95
	8	0.85	0.84	0.87	0.89	0.64	0.84	0.81
Link asymmetric	2	1.00	1.00	1.00	1.00	0.99	0.98	0.97
	3	1.00	0.99	1.00	0.96	0.94	0.94	0.88
	4	0.99	0.96	1.00	0.74	0.70	0.84	0.84
Node asymmetric	6	0.98	0.99	0.99	0.85	0.97	0.97	0.99
	7	0.94	0.95	0.96	0.80	0.92	0.92	0.95
	8	0.81	0.81	0.82	0.74	0.74	0.79	0.82

Table 1: Results for three simulation tests: Symmetric, Link Asymmetric and Node Asymmetric. SP and DK denote our projection method by using different similarity matrix when p is set to be 10. Other methods are as follows: the information based method [11], the modularity method [8], the mixture model [12], the SPAEM [13], the Spectral Clustering method [7]

same as the above case. In the link asymmetric test, 2 communities with 64 nodes each differ in their average degree sequence. Nodes in one community have average 24 edges while nodes in the other community have only 8 edges, $k_{out} = 2, 3, 4$.

The result is shown in the TABLE. 1. LPA outperforms almost all methods except the information based method[11]. Although the accuracy of our algorithm is 1 or 2 percent lower on average, LPA is still comparable to the information based method, as we choose p to be 10 which is much smaller than the number of nodes. And more interesting thing is that both SP and Dk work very well, so it is reasonable to conclude that LPA is not very sensitive to the choice of similarity matrix.

3.3 Model Selection: how to choose the optimal number c

A major challenge in detecting community is the estimation of the optimal number of communities. In this section, we test whether the gap statistic idea is suitable for it.

3.3.1 Journal Citation Network

The Journal Citation Network has four different kinds of journals(Physics, Chemistry, Biology, Ecology), clearly, the optimal number of communities is four. Next, use gap statistic idea to choose the number of communities.

When $p \geq 3$, the optimal number four is always obtained. The results for $p = 3, 10$ are shown in FIG. 4 and FIG. 5. In turn, the results confirm that gap statistic idea is suitable for inferring the optimal number of communities.

3.3.2 American Football League

In American Football League, the nodes represent the 115 teams, while the links represent 613 games played. The teams are divided into 12 conferences.

The result is shown in the TABLE. 2. When $p \geq 35$, the community number 11 is always obtained. The result seems to be wrong since there should be 12 communities. However, there is one community which in fact is not a community because mostly it

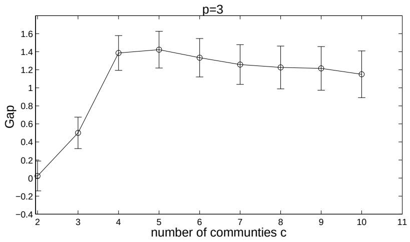

Figure 4: p=3. $r_p = 0.55$

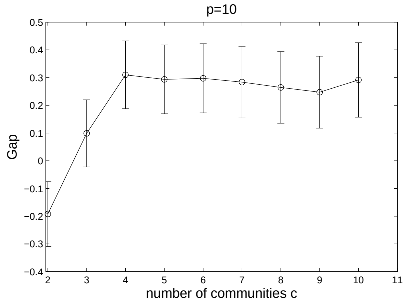

Figure 5: p=10

Test	$p = 15$	$p = 25$	$p = 30$	$p = 35$	$p = 40$	$p = 45$
r_p	0.78	0.87	0.89	0.92	0.94	0.95
Community number k	12	12	12	11	11	11

Table 2: Test on different p

plays games with adjacent communities. So according to the definition of community, it may be more sensible to divide the American Football League into 11 communities. Therefore our result about the optimal number does make sense.

3.4 How to choose p

Besides estimation of the optimal number of communities, there is another issue need to be concerned about LPA, how to choose a appropriate dimension p. There have been a few traditional methods about how to choose p, such as the eigenvalue-one criterion, the scree test, proportion of variance accounted for, the interpretability criterion. Here, the proportion method is used, as shown in Eq(8).

The goal of mapping vectors is to extract the main structure while excluding noisy information from the original similarity matrix X. To get good mapping, an equilibrium between the two aspects should be reached, in other words, r_p should be set to be in a reasonable value range to determine p. Empirically, setting $r_p \geq 0.5$ can capture the main structure of the network, if the number of community is known. On the other hand, if no information about the network is obtained, $r_p \geq 0.9$ is enough. All our test results support the above two criterions.

4 Conclusion

In this paper, we come up with a linear projection algorithm(LPA) for resolving community structure by transforming network community detection problem into a common clustering problem. The results show that it works very well on the test sets, even if p is very small. So it is reasonable to conclude that the main features of the community structure are actually captured by just the low-dimension vectors, which allows us to reduce the computational cost.

References

[1] R.Albert, A.-L.Barabási, Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47-97 (2002).

[2] R. Albert, H. Jeong, A.-L. Barabási, The diameter of the World Wide Web, Nature 401, 130 (1999).

[3] A.-L. Barabási, R. Albert, Emergence of scaling in random networks Science 286, 509-512 (1999).

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

[5] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113 (2004).

[6] M. E. J. Newman, Detecting community structure in networks, Eur. Phys. J. B 38, 321 (2004)

[7] S. White and P. Smyth, A Spectral Clustering Approach To Finding Communities in Graphs, in SIAM International Confer- ence on Data Mining (2005).

[8] M. E. J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. USA 103, 8577 (2006).

[9] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74, 036104 (2006).

[10] S. Fortunato, M. Barthlemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. USA 104, 36 (2007).

[11] M. Rosvall, C. T. Bergstrom, An information-theoretic framework for resolving community structure in complex networks, Proc. Natl. Acad. Sci. USA 104, 7327 (2007).

[12] M. E. J. Newman, E. A. Leicht, Mixture models and exploratory analysis in networks,Proc. Natl. Acad. Sci,104, 9564 (2007).

[13] W. Ren, G.Y. Yan, X.P. Liao, A Simple Probabilistic Algorithm for Detecting Community Structure in Social Networks, Phys. Rev. E 79, 036111 (2009)

[14] T. Hastie, R. Tibshirani, J. Friedman, The Element of Stastical Learning, Springer (2001)

[15] A.Gordan, Classification, 2nd edn. London: Chapman and Hall-CRC (1999)

[16] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology),63,411-423 (2001)

[17] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, Behavioral Ecology and Sociobiology 54, 396 (2003).