Spectral Clustering of Complex Networks: An Example of Chinese Herbal Medicine Network

Tony Guangyu Tong, Liqi Feng
Department of Sociology, Department of Biostatistics

Abstract

The cluster analysis on complex networks is of great interest for researchers in different fields. Graph Laplacian methods are of great importance in exploring this type of networks with weighed edges. This work adopts a localized graph Laplacian method and applied it on a self-entry dataset of co-prescription network of traditional Chinese medicine. The result shows that the symmetric Laplacian method achieves acceptable Type-I misclassification rate and outperforms other methods in terms of the Type-II misclassification rate. Future applications can be applied to other dense communication networks and have other improvement.

1 Introduction

1.1 Graph Laplacian Method

Graph Laplacian method is one of the most important tool for the spectrum clustering for complex networks [16]. For a weighted graph with the matrix W, the Laplacian matrix is always defined as $L = D - W$, where D is the diagonal matrix of a graph. The Laplacian matrix carries some of the most important properties of a graph. The cluster information is contained in the distribution of eigenvalues after an eigenvalue decomposition. Three types of graph Laplacians are currently popular in previous literature: unnormalized graph Laplacian, normalized graph Laplacian with symmetric specification, and with random walk specification.

The choice of these three methods are generally ad-hoc. Scholars briefly summarize that degree distributions usually imply which method of them should be used. For instance, if a graph is very regular with a very narrow degree distribution, the above three Laplacian methods are very similar to each other. Laplacians differ significantly, when the degree distribution is dispersed. The normalized Laplacian is generally more applicable than unnormalized Laplacian to this situation. Usually, eigenvectors of random-walk Laplacian outperforms those of symmetric Laplacian. However, for the degree distribution that does not conform to a typically power law distribution, symmetric Laplacian may outperform the random-walk Laplacian. Our example is also a case that support symmetric Laplacian.

1.2 ϵ-neighbour Network

ϵ-neighbour graph is a useful data structure that can connect points on different scales. Specifically, for a k nearest neighbor graph in a high dimensional Euclidean space, the distance of nodes can be measured through a Gaussian similarity function

$$s(x_i, x_j) = \exp\Big\{-\frac{||x_i - x_j||^2}{2\sigma^2}\Big\},$$

which is a common approach during the neighbor choice. It is easy to see that a general property of k nearest neighbor graph can be that k nearest neighbor graph can break a graph into several disconnected components based on the local density. It tends to connect vertices with similar local density, while disconnect regions of different densities from each other [16].

By incorporating the local network property into the the symmetric Graph Laplacian method, the clustering result of a network tend to keep a balance between community detection and structural equivalence. It tends to yield a better performance for the community detection especially for weighed networks, in that it is insensitive to the weight difference between low-weight edges.

2 An example of medicine networks

The Nobel Prize in Physiology or Medicine of 2015 is awarded to a Chinese pharmacist, Yoyo Tu, who was inspired by a herb prescription 1600 years ago and extracted an effective ingredient from the prescription treating Malaria. In addition to treatment of Malaria, the efficacy of Traditional Chinese herbal medicine (TCHM) has been proved in a lot of studies [4]. TCHM has a different perspective in the understanding of diseases compared to the modern western medicine; it believes that different synergies of the herbal ingredients can effectively target different parts of the body and arouse an instinct resistance to diseases. Often times, a TCHM doctor prescribes a compound mixture of 3-10 herbal ingredients according to their experience. Although the TCHM practice is highly popular in modern China and there are some theories on TCHM, TCHM prescriptions are rather subjective. Therefore, it is of particular interest to study the interrelationship between different herbal ingredients and learn rules for the prescription.

The interrelationship between herbal ingredients can be represented by a communication network [3, 6] where ingredients shown on the same prescription are connected. As most of the TCHM prescriptions rely heavily on the synergy effect of more than 3 herbal ingredients, the resulting interrelationship structure is complicated. Therefore, it is challenging to understand the synergy effect of herbal ingredients and to identify working combination of herbal ingredients.

Identification of working combinations of herbal ingredients can be viewed as a clustering problem based on communication network where herbal ingredients that have synergy effect are assigned to the same cluster. Spectral Clustering is used to approach the above goal as it is widely used in problem where the data is arranged in a complex and unknown shape [17]. Spectral clustering is one of the modern clustering method that outperforms the traditional clustering methods like k-means clustering [16]. It clusters data based on the similarity between points and the similarity between points is represented by the adjacency matrix.

In this work, a coordinate system of herbal ingredients is proposed by considering the interrelationship between herbal ingredients. The adjacency matrix of the communication network is created by the approach of *k-nearest* neighbor. Spectral clustering with normalized symmetric graph Laplacian is then employed to perform the clustering.

3 Method

3.1 Data preparation

In this example, I use a self-entry dataset of TCHM extracted from the most authoritative TCHM text book for college students in China. The dataset includes 603 ancient prescriptions from around 1000 B.C. to 1800 A. D. that are documented to cure various diseases from infectious diseases to chronic diseases. Each prescription comprises of 2-8 different types of herbal medicines. Each pair of medicines in each prescription carries a weight of 1, and the total 603 prescriptions yield a co-prescription network of 215 medicines and more than 2900 connections.

It is an relatively dense network with a density score of 0.07. The degrees of the medicines range from 1 to 142, but do not follow the power law. Rather, is appears to conform to a Poisson distribution with a mean around 20. I also arbitrarily decide to drop the top two medicines that have more than 100 degrees, since they are well connected and compatible with most types of medicines and may threaten the performance of spectrum clustering method. The network ends up with 213 vertices and 2633 edges. The medicine network with with vertices of degrees above 25 is presented in Figure 1. In order to test the performance of our method and make effective comparison to other methods, I created another test dataset with 58 pairs of medicines that traditional medicine books would usually recommend or forbid to co-prescribe (each category about half). Both the performance of Type-I and Type-II classification errors are documented and used to evaluate the performance of different clustering method. The data processing and analytic procedures are completed in R with the help of the packages `igraph` and `kknn`.

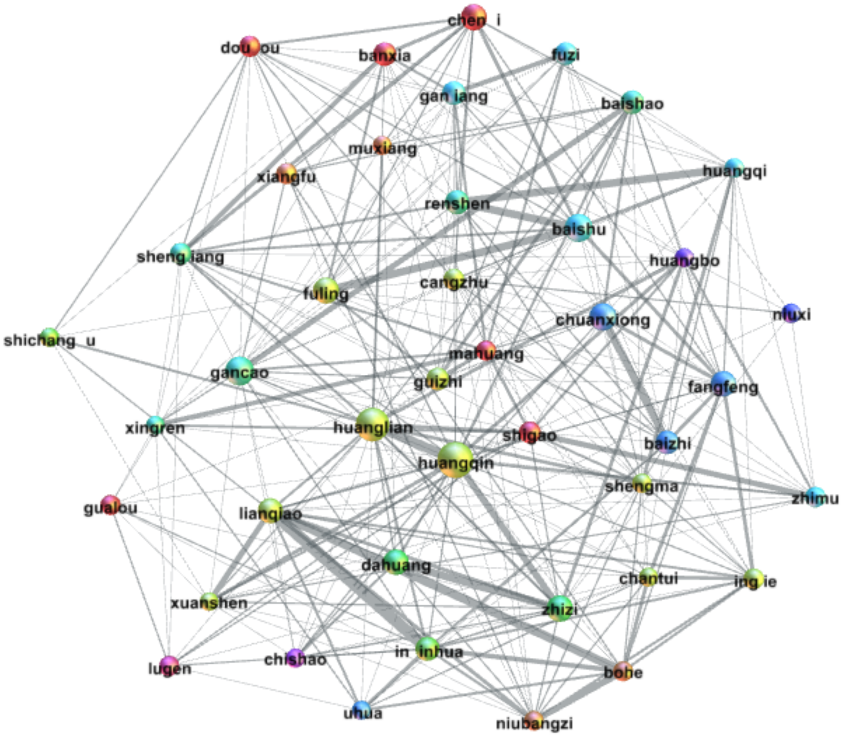

Figure 1: Medicine network of 603 prescriptions from a THCM text book ($V = 213$, $E = 2633$). Only degree > 25 presented.

3.2 Spectral clustering notations and procedures

The medicine network is a $n \times n$ off-diagonal matrix $\boldsymbol{C}$ where n denotes the number of herbal ingredients. C_{ij} is the number of prescriptions that ingredients i and ingredients j both show up. The coordinate of ingredient i is the inverse of the corresponding row $\boldsymbol{C}_{i\cdot}$ with the i^{th} element be 0.

The *adjacency matrix* of herbal ingredients is constructed by taking the nearest k neighbors of every herbal ingredient: $\boldsymbol{A} = (a_{ij})_{i,j=1,\ldots,n}$ where ingredients ν_i and ν_j disconnected $\rightarrow w_{ij} = 0$. Otherwise, a_{ij} is 1.

Graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ where $\boldsymbol{D}$ is a diagonal degree matrix, $D_{ii} = \sum_{j=1}^{n} a_{ij}$. Because herbal ingredients have similar degrees of connections as a prescription often contains a lot of ingredients, symmetric graph Laplacian is used and has the following form:

$$\boldsymbol{L}_n = \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{L} \boldsymbol{D}^{-\frac{1}{2}}.$$

The number of small eigenvalues (≈ 0) of $\boldsymbol{L}_{sym}$ indicates the number of clusters in the original communication network. Thus, Spectral clustering determines the appropriate number of clusters by finding the eigenvalue λ_l of of $\boldsymbol{L}_{sym}$ and λ_{l+1} is relatively large. l is the number of clusters used in the analysis.

3.3 Algorithm

The following algorithm is adapted and modified based on the spectrum clustering method by Ng et al. [7].

1) Construct a similarity graph by connecting two herbal ingredients based on the distance in the communication network and the k nearest neighbor specified

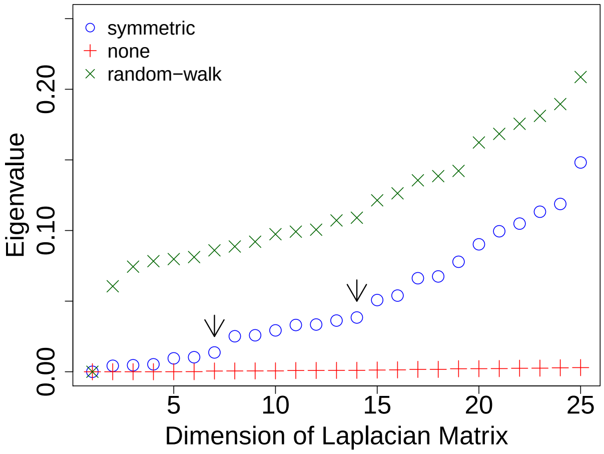

Figure 2: Eigenvalue plot of the Laplacian matrices. Two arrows indicate the optimal number of cluster for the algorithm (eigengap heuristic)

2) Compute the normalized Laplacian $\boldsymbol{L}_{sym}$.

3) Compute the first l eigenvectors (ascending order) $u_1, ..., u_l$ of L_{sym}.

4) Let $\boldsymbol{U} \in \mathbb{R}^{n \times l}$ contains $u_1, ..., u_l$ as columns.

5) Form the matrix $\boldsymbol{T} \in \mathbb{R}^{n \times l}$ from U by normalizing the rows to norm 1 ($t_{ij} = u_{ij}/(\sum_l u_{il}^2)^{\frac{1}{2}}$).

6) For $i = 1, ..., n$, let $y_i \in \mathbb{R}^l$ be the vector corresponding to the i^{th} row of $\boldsymbol{T}$.

7) Clusters the points $(y_i)_{i=1,...,n}$ with the k-means algorithm into clusters $\mathcal{C}_1, ..., \mathcal{C}_l$.

4 Result

Choosing the number of k clusters is a general and opened-up problem for all clustering methods. For a model with two tuning parameters the nearest neighbour k and the number of clusters k in the k-mean step, the model selection for our method seems to be very complicated. Based on diferent assumptions, several methods have been adopted by previous scholars, such as the log-likelihood of the data, information criteria of within-cluster and between-cluster similarity [13], the gap statistic a [14] and stability approach [1]. However, another method is suggested extremely efficient for spectrum clustering–eigengap heuristic map. This method is also compatible for the comparison over the three different types of graph Laplacian methods. Figure 2 shows the parameter slection for k in the k mean. By sorting the values of eigenvalues, if kth eigenvalue is significantly smaller than that of the $k+1$th, k is likely to be the optimal number of clusters. Multiple gaps map be identified in the procedure. It is worth noting that for networks with very regular structure, this method usually cannot work well. However, the eigen heuristic shows perfect performance in our example.

The k in the k-nearest neighbour is then the only parameter to turn. with multiple trials on the prediction power of testing dataset, the results in the Table 1 shows the optimal selected k and comparison of performance with other methods, which includes the unnormalized Laplacian [16], the betweenness-based method [5], the information map method [12, 11], the label propogation method [9], the leading eigenvector method [6], the multilevel method [2], the statistic mechanics method [10, 6, 15], and the short random walk method [8].

The results shows symmetric Laplacian method achieves reliable performance in terms of the both ends of misclassification error rate. When $k = 4$ for the nearest neighbour, and 14 centers, the Type-I misclassification rate can be low; when $k = 10$ for the nearest neighbour with 4 centers, the Type-II misclassification rate is 10 percent. Although some other methods, such as information map and betweenness based method can achieves even lower Type-I misclassification rate, the symmetric Laplacian outperforms all other methods in terms of the Type-II misclassification rate. This performance is especially important for the clustering of co-prescription network in that lowering Type-II error is more important than lowering Type-I error.

Table 1: Comparison of performance between different clustering methods.

Method	# Clusters	Misclassification Rates (%)
Lap-Symm	14 ($K = 4$)	(14.3, 34.2)
Lap-Symm	4 ($K = 10$)	(38.8, 10.0)
Lap-None	12 ($K = 3$)	(40.0, 36.8)
Betweenness	84	(8.0, 23.5)
InfoMap	68	(0.0, 50.0)
Label Propogation	1	(47.5, 100.0)
Lead Eigenvalue	10	(13.0, 30.6)
Multilevel	6	(13.0, 30.6)
Stat Mechanics	7	(13.0, 30.6)
Short Ran-Walk	14	(17.9, 25.8)

⋆ misclassification rates is computed as the Type I error and Type II error

5 Discussion

In summary, the symmetric Laplacian spectral clustering achieves reliable performance in obtaining both small Type I and Type II mis-classification rates. More importantly, it outperforms other methods in terms of the performance with the lowest Type II misclassification rate, which helps probe wrong prescription combinations. Symmetric approach also outperforms the unnormalized approach and the random walk approach. With a larger dataset and other information, such as illness, age gender and other demographic background of patients, the performance of this method can be further improved. This application also has promising perspective in the scientific management on traditional herbic medicine in China.

This work explores an example of complex network using a localized graph Laplacian method. It incorporates the power of community detection and structural equivalence in the clustering procedure. Similar applications can be applied on email networks and other types of communication networks with dynamic vertices. Most importantly, the k-nearest neighbour localized graph Laplacian method potentially can be further incorporated with the boosting method or other reweighting methods on weak learners to achieve even better performance in different classification tasks.

References

[1] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 17(FEBRUARY 2002):6–17, 2002.

[2] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *J. Stat. Mech*, page P10008, 2008.

[3] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004.

[4] K.L.E. Hon, T.F. Leung, P.C. Ng, M.C.A. Lam, W.Y.C. Kam, K.Y. Wong, K.C.K. Lee, Y.T. Sung, K.F. Cheng, T.F. Fok, K.P. Fung, and P.C. Leung. Efficacy and tolerability of a chinese herbal medicine concoction for treatment of atopic dermatitis: a randomized, double-blind, placebo-controlled study. *British Journal of Dermatology*, 157(2):357–363, 2007.

[5] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006.

[6] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69(026113), 2004.

[7] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. pages 849–856, 2002.

[8] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218, 2005.

[9] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E*, 76:036106, Sep 2007.

[10] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, Jul 2006.

[11] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[12] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[13] Susanne Still and William Bialek. How many clusters? An information theoretic perspective. *Neural Computation*, 2506:26, 2003.

[14] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*, 63(2):411–423, 2001.

[15] V. A. Traag and Jeroen Bruggeman. Community detection in networks with positive and negative links. *Phys. Rev. E*, 80:036115, Sep 2009.

[16] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[17] Lihi Zelnik-manor and Pietro Perona. Self-tuning spectral clustering. pages 1601–1608, 2005.