Robust Spectral Clustering by Maximizing Similarity Modularity

Xiaohui Tian

School of Network Security and Informatization, Weinan Normal University, Weinan, China

Keywords: Graph; Spectral clustering; Similarity modularity

Abstract: Modularity is a metric of how well a given network is divided into communities. It has proven surprisingly effective in handling the ready-made real-world networks, including social networks, information networks, biological networks, and so forth. This paper, based on a notional instantiation of modularity to similarity modularity (s-mod), answers the question of how well the *s-mod* can measure the data-points clustering. This paper maximizes the *s-mod* by a new spectral algorithm. Compared to the existing typical spectral clustering techniques, the presented algorithm has two advantages: the higher robustness to the variation of input parameter and the capability of estimating the intrinsic clusters' number. This paper also analyzes theoretically why it performs robust to parameter variation and how to accurately estimate the intrinsic clusters' number. Experiments on typical UCI datasets show the good performance of the presented algorithm.

1. Introduction

The terminology “graph-cut clustering" that originated from the study community of machine learning and data mining, refers to achieving goal of data-points clustering by means of constructing a similarity graph over the points and then partitioning the constructed graph into sub-graphs. Spectral clustering algorithms [1, 2, 3, 4, 5, 6, 7, 24, 25] are a class of techniques that can solve the graph-cut clustering in a polynomial time via eigen-decomposition. They have attracted considerable research attention over the past decade due to their practical successes in handling both convex data and non-convex data. They have also been applied to a wide variety of areas, including computer vision [4, 9, 10], circuit placement [3, 11, 12], load balancing [13, 14], biological information [15, 16], and many others. Different from directly running the most popular *k*-means algorithm [8] that probably converges to local solutions [22], spectral clustering techniques can often guarantee the global relaxation solutions to the graph-cut optimization objectives. One can see [17, 18, 19, 22] for the surveys.

The typical spectral clustering techniques ([1] [4] [5] [6] and [7] etc.) have two *long-standing* drawbacks. Firstly, the optimization criterions can not directly aid the estimation of the clusters’ number. Most often, the clusters’ number must be prescribed before the clustering algorithms run. For many raw datasets, it is impractical to know in advance the intrinsic clusters’ number. Secondly, spectral clustering techniques are always sensitive to the selection of Gaussian kernel parameter. More concretely, the only input parameter of spectral clustering algorithms is the so-called similarity matrix in which the relations of pairwise data points are contained. However, the metric of similarity between pairwise data points typically requires a Gaussian kernel parameter whose variation might give rise to considerably different clustering outcomes. In unsupervised settings, currently, there is no good approach available to suggesting deterministically a right value of that parameter.

This situation motives us to consider new better spectral algorithms. We argue that a good spectral algorithm can perform significantly robust to the variation of parameter σ, and that an optimization objective can aid directly the estimation of intrinsic clusters’ number. In recent years, another type of modularity-based spectral method [20, 21] has been developed for detecting community structures of the real-world un-weighted networks. It is tempting that this new spectral method chooses a null model against which to compare the actual networks to random-generated networks, thereby skillfully lay a foundation that the optimization of modularity can be achieved by spectral decomposition of modularity matrices. The chief algorithm proposed by Newman [20, 21]

  DOI: 10.25236/systca.18.049

goes by repeatedly bi-partitioning a network into communities by making use of the leading eigenvectors of network and sub-networks. It performs surprisingly effective in finding the community structures underlying a wide variety of real networks, including social networks, information networks, biological networks, and so forth. This algorithm, however, is early proposed to deal with the ready-made 0-1 un-weighted networks instead of the weighted similarity graphs (the constructed similarity graphs with the purpose of graph-cut clustering). In this paper, the basic purpose is to extend the modularity metric to the data-points clustering, and to show that the derived algorithm can well overcome the two drawbacks in traditional spectral clustering algorithms. More concretely, we shall show that similarity modularity (*s-mod*) based spectral algorithm is significantly robust to the variation of values of Gaussian kernel parameter, and that the *s-mod* optimization objective can aid directly the estimation of intrinsic clusters' number. Our work first builds on a slight notion instantiation from modularity to *s-mod*, and then we interpret why this extension can give more robustness and show how it directly aids the estimation of intrinsic clusters' number. Experimentally, the algorithm outperforms the well-known NJW algorithm [7] on several most popular benchmark datasets.

The organization of the rest paper is as follows. In the next Section we first review the spectral method of modularity. In Section 3 we show the generalization of modularity-based spectral algorithm to data clustering is feasible, thereby lay a foundation for building the new spectral clustering algorithm. In Section 4 we describe and analyze the new algorithm. In Section 5 we test the algorithm using a variety of most popular benchmark datasets and discuss the results. In the last Section we draw the conclusions.

2. Modularity Method

An algorithm for detecting the communities of an un-weighted network generally dedicates to divide the network into *k* sub-networks such that the edges within sub-networks have a high density and between sub-networks have a low density. The benefit function of modularity [20, 21] holds another viewpoint that a natural community should contain an *exceeded-expected* fraction of edges, and the partitioning task then amounts to maximizing the modularity. The general expression of the modularity function can be written as

$$Q = \sum_{i=1}^{k} (f_i - \overline{f}_i), \tag{1}$$

where Q denotes the value of modularity and f_i denotes the fraction of edges within the kth community and $\overline{f}_i$ denotes the *expected* fraction within them. Let k_i and k_j denote respectively the degrees of vertex V_i and V_j, and m denotes the total number of the edges contained in a network. The extended form of Eq. (1) is

$$Q = \frac{1}{2m} \sum_{i=1}^{n} \sum_{i=1}^{n} (A_{ij} - \frac{k_i k_j}{2m}) \delta(g_i, g_j), \tag{2}$$

where $\delta(g_i, g_j)=1$ if $g_i=g_j$ and 0 otherwise. Let B denotes an n-by-n matrix (called modularity matrix) and its entries are given by $B_{ij}= A_{ij}-k_ik_j/2m$. The aforementioned *null model* is exactly the following equation:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} B_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} (A_{ij} - \frac{k_i k_j}{2m}) = 0, \tag{3}$$

which is the basis of the spectral algorithms in [20, 21]. Let s denote an indicator vector of size n in which each entry just can take -1 or +1, meaning that the nodes associated with the same sign belong to a same community. For the bi-sections problem, Eq. (2) can then be formulated as

$$Q = \frac{1}{4m} s^T B s \,. \tag{4}$$

In the following we mention the basic properties of the modularity matrix B by summarizing the past works in [20, 21]:

1) The sum of the elements in each row or column is zero.
2) $\mathbf{1}$ = [1, 1, ..., 1] is an eigenvector of B associated with the eigenvalue 0.
3) All the eigenvectors of the modularity matrix B are linear independence and orthogonal each other, meaning that the eigen-matrix $U = [u_1|u_2|...|u_n]$ of B is an orthogonal matrix satisfying $U^{-1} = U^T$ where u_i denotes one of the column eigenvectors.

Hence B can be further written as $U^T \Sigma U$ where Σ is a diagonal matrix with the diagonal elements: $\lambda_1, \lambda_2, ..., \lambda_n$, the eigen-values of B. Q can then be extended to the form as follows in terms of *spectral decomposition* as follows:

$$Q = \frac{1}{4m} \sum_{i=1}^{n} \lambda_i \left| s^T u_i \right|^2 \,. \tag{5}$$

So far it implies bi-partitioning a network is equivalently to maximizing Q subject to $tr(ss^T)=n$. Then, a straightforward way for bi-partitioning a network goes by finding the eigenvector u_1 associated with the largest eigenvalue of B and then making s as close as possible parallel to u_1 [20,21]. However, for the discrete optimization objective, the entries of s are just limited to ± 1. Hence, to achieve the goal of maximizing the general modularity, the entries of s should have the same signs as the associated entries in u_1. For the network contained in multiple communities, it requires a repeatedly bi-partitioning procedure over the sub-networks until the general modularity can not be improved.

3. Generalized Modularity to Data Clustering

In this section, we shall generalize the modularity notion from the 0-1 un-weighted networks to arbitrary non-negative weighted graphs such that the generalized modularity can serve as measuring the graph-cut clustering. We show mathematically that the so-called null model also holds in the non-negative weighted graphs such as the constructed similarity graphs, and also show that the data clustering problem can be solved as a spectral optimization problem of generalized modularity.

A formal description about the problem in question is as follows. Given a set of vector data $X_1, X_2, \ldots, X_n \in R^d$ ($d \in \mathbf{N}$), the graph-based clustering methods proceed by first constructing a similarity graph G, and then partitioning it into a set of components $G^{(1)}, G^{(2)}, \ldots, G^{(k)}$, in each of which all the data points are treated as a cluster. For all similarity graphs, we define them without loops and multiple edges. In G, the similarity between X_i and X_j , is measured conventionally by the expression

$$W_{ij} = \exp\left(-\frac{\left\| X_i - X_j \right\|_2^2}{\sigma^2} \right), \tag{6}$$

where W denotes the similarity matrix and σ denotes a tunable parameter. With the decrease of parameter σ,

the constructed similarity graphs tend to be sparse due to the limitation of computing accuracy. Let d_i be the

weighted degree of V_i, and let t be the total weights of G given by

$$t = \frac{1}{2} \sum_{V_i \subseteq G} d_i \,. \tag{7}$$

We first give the following two definitions.

Definition 1. (Similarity modularity matrix) Denote by H the n-by-n similarity modularity matrix

(SMM). Its entries are given by the following equation:

$$H_{ij} = W_{ij} - \frac{d_i d_j}{2t}. \tag{8}$$

Definition 2. (Similarity modularity) Denote by Q_s the similarity modularity, meaning that the actual similarity within clusters *minus* the *expected* similarity that can be given by the following equation:

$$Q_s = \frac{1}{2t}\sum_{ij} H_{ij}\frac{s_i s_j + 1}{2}. \tag{9}$$

For a dataset containing two clusters, the goal of data clustering is then transformed to the problem of maximizing Q_s for finding the indicator vector s. The data points associated with the same signs of eigenvector's entries can be acted as inside the same cluster.

In the following, we shall show that for all non-negative weighted acyclic graphs, there must be a similar null model like Eq. (3). First of all, we can easily find that the sum of the elements in each row or column is zero according to the following deductions. Note that

$$t = \frac{1}{2}\sum_i d_i = \frac{1}{2}\sum_j d_j. \tag{10}$$

We then have

$$H_i = \sum_j H_{ij} = \sum_j W_{ij} - \sum_j \frac{d_i d_j}{2t} = d_i - d_i \frac{2t}{2t} = 0. \tag{11}$$

Further we have

$$\sum_{ij} H_{ij} = \sum_i H_i = 0. \tag{12}$$

It means that the sum of all the elements of H is zero. This null model agrees well with the formulation of un-weighted networks. Hence for any such graphs, we have

$$Q_s = \frac{1}{4t} s^T H s. \tag{13}$$

This expression agrees well with Eq. (4). One then can check that bi-partitioning a similarity graph can be given by the spectral composition of similarity modularity matrix H. One can also check that k-partitioning a network can be given by the trace maximization of Eq. (14) for finding the indicator matrix S, in which the elements are binary values, with each column vector for one cluster. The data points assigned into the same cluster are associated with the entries ones.

$$Q_s = \frac{1}{4t} tr(S^T H S) = \frac{1}{4t}\sum_{i=1}^{n}\sum_{j=1}^{k} \lambda_i \left| u_i^T s_j \right|^2 \tag{14}$$

4. Algorithm

In this section, we shall present a simple k-way spectral algorithm of generalized modularity. The general framework of the derived algorithm originates from the work of Ng.et al. [7]. We just do two aspects of modifications. On one hand, we substitute the normalized similarity matrix for the generalized modularity matrix, both the properties of matrices and the optimization objective of them being discriminative each other. On the other hand, the algorithm proceeds by finding the eigenvectors associated with the largest a few eigen-values. The similarity modularity matrices always possess both positive eigen-values and negative eigen-values, however, only those positive eigen-values are useful for partitioning similarity graphs. Before we use an eigenvector, there is a

need to check if the associated eigen-value is positive.

Algorithm 1 Spectral clustering based on SMM

Input: Dataset X and the clusters' number k.

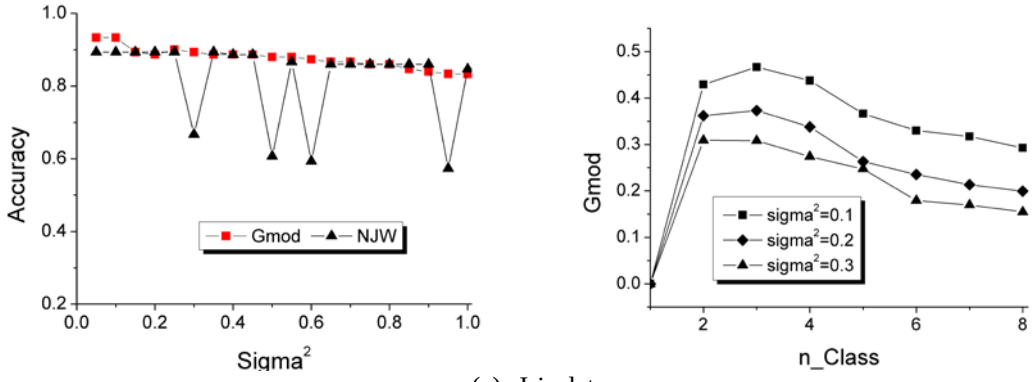
Output: The identification of clusters.

Step1: Obtain the similarity matrix W from X.

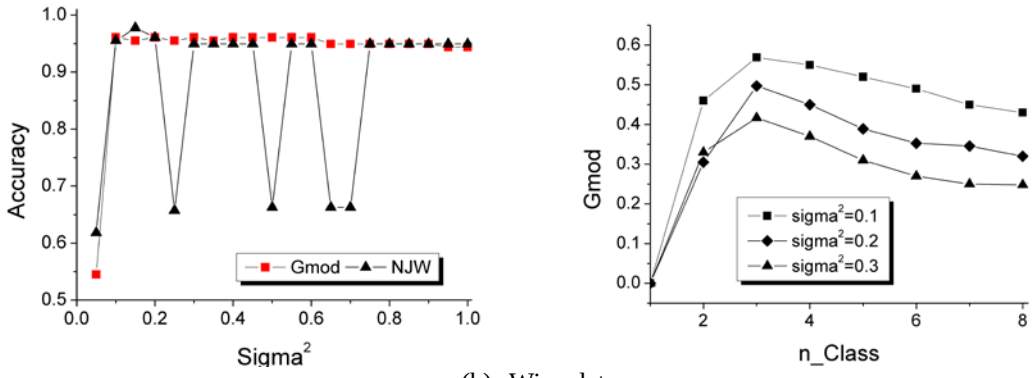
Step2: Form the normalized similarity matrix $W_n=D^{-1/2}W D^{-1/2}$

Step3: Build the *s-mod* matrix H from W_n.

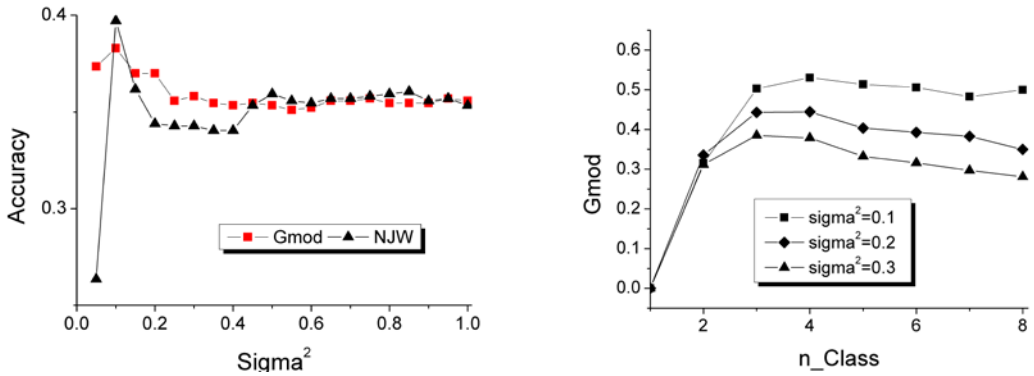
Step4: If $k = 2$, find the eigenvector associated with the largest eigenvalue of H and bi-partitioning the graph according to the signs of the entries, then terminated.



(a) Iris data



(b) Wine data



(c) Vehicle data

Figure 1. The experimental results of UCI datasets. The left are the variation of accuracies with sigma, and the right are the variation of generalized modularity with clusters' number.

Step5: Find $u_1,u_2,\ldots,u_{k-1}$ by solving $Hu = \lambda u$, the eigenvectors associated with the largest $k-1$ eigenvalues of H (if not all of the k-1 largest eigenvalues are positive, just use the eigenvectors associated with the positive ones); and then form the matrix $U = [u_1|u_2|\ldots|u_{k-1}] \in \mathbf{R}^{n\times(k-1)}$ by putting the eigenvectors in columns.

Step6: Form the matrix Y from U by re-normalizing each of U's rows to be unit length.

Step7: Treating each row of Y as a data point in $\mathbf{R}^{k-1}$, cluster them into k groups via K-means algorithm.

5. Experiments on UCI Data

The datasets we chosen are of availability from UCI machine leaning repository [23], including Iris, Wine

and Vehicle, which have been extensively applied in testing the performances of new clustering algorithms.

For the details of the data introduction, see the web site in Ref. [23]. For each dataset, we simply use all the

features when construct the similarity graph.

As the results shown in Fig.1, for each dataset, the accuracies of similarity modularity based spectral clustering tends to be significantly robust to the variation of σ. However, the well-known NJW algorithm tends to be sensitive to the variation of σ, particularly on the datasets Iris and Wine, the oscillations exceeding 25%. On the other hand, we can see that the s-mod values directly aid the estimation of intrinsic clusters' number. For right choices of σ, the intrinsic clusters' numbers are often associated with the largest values of s-mod. Particularly on Wine data, for all the three σ^2s, the largest s-mod values always correspond to three, the intrinsic clusters' number.

6. Conclusion

This paper has shown that the generalizing of modularity function to data clustering is feasible at the pure mathematical angle, and has also shown that the derived k-way spectral clustering algorithm can perform more robust than the well-known NJW algorithm on a set of benchmark datasets. The s-mod based spectral algorithm is capable of estimating the intrinsic clusters' number by maximizing the s-mod optimization objectives.

Acknowledgment

The author would like to thank the subsidized fund of no. 16YKS011 and no. 17JMR39.

References

[1] Y.C. Wei and C.K. Cheng. Toward efficient hierarchical designs by ratio cut partitioning. *In Proceedings of the IEEE International Conference on Computer Aided Design*, pp. 298-301, 1989.

[2] B. Mohar. Some applications of Laplace eigenvalues of graphs. *Graph Symmetry: Algebraic Methods and Applications*, Vol. NATO ASI Ser. C 497, pp. 225-275,1997.

[3] C. Alpert, A. Kahng, and S. Yao. Spectral partitioning: The more eigenvectors, the better. *Discrete Applied Math*, Vol. 90, pp. 3-26, 1999.

[4] J.B. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 888-905, 2000.

[5] S. Sarkar and P. Soundararajan. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 5, pp. 504-525, 2000.

[6] R. Kannan, S. Vempala, and A. Vetta. On clustering: good, bad, and spectral. *In Proceedings of the 41st Annual Symposium on the Foundation of Computer Science*, pp. 367-380, 2000.

[7] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *In Proceedings of Neural Information Processing Systems*, pp. 849-856, 2001.

[8] J. Mcqueen. Some methods for classification and analysis of multivariate observations. *In*

Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297,1967.

[9] J. Malik, S. Belongie, T. Leung, and J.B. Shi. Contour and texture analysis for image segmentation. *Perceptual Organization for Artificial Vision Systems*, Kluwer, 2000.

[10] M. Meila and J. Shi. Learning segmentation by random walk. *In Proceedings of Neural Information Processing Systems* 13, 2002.

[11] C. J. Alpert and A. B. Kahng. Multiway partitioning via geometric embeddings, orderings and dynamic programming. *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, Vol. 14, No. 11, pp. 1342-58, 1995.

[12] P. K. Chan, M. D. F. Schlag, and J. Y. Zien. Spectral k-way ratio-cut partitioning and clustering. *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, Vol. 13, No. 9, pp. 1088-96, 1994.

[13] B. Hendrickson and R. Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. Sci. Comput.*, Vol. 16, No. 2, pp. 452-459, 1995.

[14] R. Van Driessche and D. Roose. An improved spectral bisection algorithm and its application to dynamic load balancing. *Parallel Comput*, Vol. 21, No. 1, pp. 29-48, 1995.

[15] Y. Kluger, R. Basri, J.T. Chang, and M. Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*. Vol. 13, No. 4, pp. 703-716, 2003.

[16] A. Paccanaro, C. Chennubhotla, J.A. Casbon, and M.A.S. Saqi. Spectral clustering of protein sequences. *International Joint Conference on Neural Networks*, Vol. 4, pp. 3083-3088, 2003.

[17] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, Vol. 17, No. 4, pp. 395-416, 2007.

[18] D. Verma and M. Meila. A comparison of spectral clustering algorithms. *Technical Report, Department of CSE University of Washington Seattle, WA, 98195-2350*, 2005.

[19] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta. A survey of kernel and spectral methods for clustering. *Pattern Recognition*, Vol. 41, No. 1, pp. 176-190, 2008.

[20] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, Vol. 103, pp. 8577-8582, 2006.

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

[22] U. Luxburg, M. Belkin, and O. Bousquet. Consistency of Spectral Clustering. *Annals of Statistics*, Vol. 36, No. 2, pp. 555-586, 2008.

[23] A. Asuncion and D.J. Newman. UCI Machine Learning Repository [http://www.ics.uci.edu/mlearn/ MLRepo- sitory.html], Irvine, CA: University of California, School of Information and Computer Science, 2007.

[24] Alzate C and Suykens J A K. Multiway Spectral Clustering with Out-of-Sample Extensions through Weighted Kernel PCA[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, Vol. 32, No, 2, pp. 335-347, 2010.

[25] Yan D H, Huang L and Jordan M I. Fast approximate spectral clustering[C]. ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, ACM Press, New York, 2009: 907-915.