Physica A 498 (2018) 50–65

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Overlapping communities detection based on spectral analysis of line graphs

Chun Gui, Ruisheng Zhang *, Rongjing Hu, Guoming Huang, Jiaxuan Wei

School of Information Science and Engineering, Lanzhou University, Lanzhou City, Gansu Province, 730000, China

HIGHLIGHTS

- Define the degree matrix and the Laplacian matrix of the line graph.
- Obtain appropriate number of communities by using micro-improvement partition density.
- Apply spectral analysis on line graph to reveal hierarchical organization and generate overlapping communities of networks.

ARTICLE INFO

Article history:
Received 28 August 2017
Received in revised form 16 November 2017

Keywords:
Overlap
Hierarchical structure
Spectral analysis
Line graphs
Community detection

ABSTRACT

Community in networks are often overlapping where one vertex belongs to several clusters. Meanwhile, many networks show hierarchical structure such that community is recursively grouped into hierarchical organization. In order to obtain overlapping communities from a global hierarchy of vertices, a new algorithm (named SAoLG) is proposed to build the hierarchical organization along with detecting the overlap of community structure. SAoLG applies the spectral analysis into line graphs to unify the overlap and hierarchical structure of the communities. In order to avoid the limitation of absolute distance such as Euclidean distance, SAoLG employs Angular distance to compute the similarity between vertices. Furthermore, we make a micro-improvement partition density to evaluate the quality of community structure and use it to obtain the more reasonable and sensible community numbers. The proposed SAoLG algorithm achieves a balance between overlap and hierarchy by applying spectral analysis to edge community detection. The experimental results on one standard network and six real-world networks show that the SAoLG algorithm achieves higher modularity and reasonable community number values than those generated by Ahn's algorithm, the classical CPM and GN ones.

1. Introduction

The community structure is one of prominent features of networks [1] with a topology structure of "external loose and inner tight" [2]. It consists of many subgraphs and appears with a high density of internal edges, yet the connection edges between the subgraphs are sparse. These subgraphs are called community, and emerge in all kinds of networked systems [3]. Communities represent the internal structure of the network and reveal the existence of hidden special correlations among vertices. These relations between vertices can be obtained by graph partitioning [4,5], hierarchical clustering [6,7], partition clustering [8–10], and spectral clustering and so on. Communities in the network are often overlapping, so one

* Corresponding author.
E-mail addresses: guich13@lzu.edu.cn (C. Gui), zhangrs@lzu.edu.cn (R. Zhang), hurj@lzu.edu.cn (R. Hu), huanggm12@lzu.edu.cn (G. Huang), weijx@lzu.edu.cn (J. Wei).

https://doi.org/10.1016/j.physa.2018.01.025
0378-4371/

of phenomenons is that a vertex belongs to more than one group synchronously. Meanwhile, the networks may show multi-layers of grouping of the nodes, with large clusters contain small clusters, and the large clusters are included in larger clusters, and so on, which is a hierarchical structure. Many real networks show that communities are both hierarchical and overlapping [11]. There are many achievements in the research of the overlapping community. Palla et al. [12] employed separate partition algorithms to split networks into significant numbers of cliques. Wu et al. [13] generalized the coupled Kuramoto oscillator to realize synchronization or divide into many clusters of networks, with the result that the overlapping vertices become the interspace of two communities. Reid et al. [14] analyzed the main statistical properties of overlapping communities which revealed the modular structure of the networks. In general, hierarchical clustering techniques are used to reveal the multilevel structure of the social networks. Gilpin and Davidson [15] formalized hierarchical clustering as an integer linear programming problem with an objective function and enforced the dendrogram attributes implemented as linear constraints. Depended on the similarity method in agglomerative hierarchal clustering, Lu et al. [16] proposed the travel-time method based on hierarchical clustering. The study [17] demonstrated that the two generic features of scale-free and a high degree of clustering are the fruits of a hierarchical organization. Although these traditional methods investigated the two properties of community separately, they are not enough to describe the complete structure of community accurately, due to the existence of overlap and hierarchy features of community structure [18].

The pioneering algorithm to detect both overlap and hierarchical structure of networks was proposed by Lancichinetti et al. in 2008 [19]. It used a local optimization fitness function to find the overlapping communities. The disadvantage is that the random selection of seed vertices may affect the detection result of the overlapping community and hierarchy in networks. The research of traditional algorithms focuses on the properties of the vertex, e.g. the degree distribution or the ranking computed by some measures. As is known, networks are composed of a set of vertices and edges, while edges reflect the relationships between vertices. So it may be helpful to research the network with the view point of edges rather than vertices. Evans and Lambiotte applied the popular k-cliques method to the line graph of the original graph to achieve the overlapping community detection [20]. In 2010, Ahn et al. used edge division to replace the node partition to realize the overlapping community detection [18]. The experimental results show that it successfully unified the overlap and hierarchy of networks by edge community detection. Both Evans and Ahn found that when the non-overlapping community detection algorithms were applied to the edge community, these algorithms can detect overlapping communities. In the present methods of community detection, the edge community reasonably explains the fact that the overlap and hierarchy coexist in social networks. Different from the traditional definition that the community is the partition of the vertex, we defined communities as the set of edges. A vertex may then have edges belonging to several communities and thus it may belong to two or more communities.

An algorithm combined both spectral analysis and hierarchical structure partition is proposed by Donetti and Munoz in 2004 for the first time [21]. The algorithm obtains the optimal community structure by finding the best splitting among all the possible partitions of the dendrogram based on the modularity function Q (Q is proposed by Newman). Spectral clustering is a partition method that separates the data set into clusters by employing the eigenvector of a matrix [22]. Eigenvectors make the clustering attributes of the original data set more clear, spectral clustering is superior to conventional clustering algorithms such as k-means [23]. Since the algorithm [21] cannot identify the overlapping community structure of networks, we used spectral analysis in line graphs and get overlapping vertices naturally. The most common matrix in spectral clustering is the Laplacian matrix by far. The first k eigenvectors of a Laplacian matrix must be calculated in spectral clustering. In 1973, Fiedler [24] proved that a bipartite network can be generated by using the eigenvector of the first non-trivial eigenvalue of Laplacian matrix. This method is based on a modular matrix, and has a good effect in the bipartite classification. In the following paper [25], there is a realization and explanation about the relation between the eigenvectors and the hierarchical community structure. However, the performance declines when the number of communities is more than two [26]. According to the Rayleigh–Ritz theory, the first non-trivial eigenvectors of the Laplacian matrix obtained by spectral analysis method are the relaxation solution of the optimal partition problem. For a network, components of the non-trivial eigenvector of the Laplacian matrix exactly correspond to each element of the network. So we can establish the eigenvector space of network elements for community detection. The performance is not ideal if only one-dimensional eigenvectors are used to construct eigenvector space [27]. This problem is resolved by enlarging the dimension of the eigenvectors space into two-dimensional in our algorithm. Community detection based on modularity is a popular method, however, the drawback of resolution limit [28,29] makes modularity cannot capture small communities although strongly clustered they are. So we use a micro-improvement partition density to obtain an appropriate number of communities. Jaccard index is a method to compute the similarity between vertices [18]. The drawback of Jaccard index is that it only focuses on the edge of a common vertex but ignore the two edges share no common vertex. Inspired by empirical observations: the vertices of a weakly connected within a community are "Euclidean distance" away, while are still in the same direction in the eigenvector space. We use angular distance to obtain the similarity between vertices [21].

In summary, the contributions of this paper are 1. defining the degree matrix and the Laplacian matrix of the line graph; 2. demonstrating Angular distance is the better one out of four similarity measures (four similarity measures are Manhattan distance, Euclidean distance, Jaccard index and Angular distance respectively); 3. getting appropriate number of communities by using a micro-improvement partition density; 4. applying spectral analysis on line graph to reveal hierarchical organization and generate overlapping communities of networks; 5. comparing the effectiveness of the proposed algorithm with three other state-of-the-art algorithms (the three other algorithms are Ahn's method [18], CPM [12] and a greedy modularity optimization method [30] and demonstrating that the proposed algorithm achieves higher modularity and reasonable community number values.