Overlapping Community Detection Using Centrality Measure and Local Seed Information

Nyunt Nyunt Sein
University of Computer Studies(Kalay)
dynedyne098494@gmail.com

Abstract

Social Network Analysis (SNA)is the process of investigative social structures through the use of networks and graph theory. One of the most applicable features of graphs representing the real-world datasets is their community structure. Community detection is an important task in social network analysis. Communities are intuitively characterized as "unusually densely knit" subsets of a social network. When in real datasets, many vertexes belong to many communities, therefore overlapping community detection becomes a prominence topic in social analytic research fields. In this paper, we present a new overlapping community detection method based on local seed selection and local seed expansion approach. In our proposed method, we first partitioned the graph into disjoint clusters, and then select the node with highest eigenvector centrality value as seed node for each cluster and then expand these seed node to obtain the overlapping communities. We test our algorithm on small scale real-world datasets. Experimental results show that our presented algorithm outperforms in accuracy than other two popular overlapping community detection methods.

Keywords: seed expansion, overlapping community detection, community detection

1. Introduction

Various systems in the world can be characterized as networks in which network links represent the relationships between the interconnected parts (nodes) of the system. Examples of well-known networks include online social networking sites such as Facebook, Twitter, and collaboration networks. Network links represent different types of relationships in different domains, such as human friendship, organizational structure, physical proximity of animals, and interconnection of infrastructure.

In studies of complex networks, networks are said to have a community structure, where the nodes of the network are easily grouped into sets of nodes, where each set of nodes is tightly connected internally.

In recent years, researchers have developed numerous community detection methods. These methods have two types such as overlapping and disjoint community detection methods.

For disjoint community finding, the network naturally divides into groups of nodes with sparser links and internally dense links between groups. In online social network, a person can play in multiple communities such as family, friends and college. So, overlapping communities are also allowed. Overlapping communities is based on the principle that vertexes are more connected if they are members of the same general community (ies), and if they do not share communities, they are less likely to connect.

There are various methods for detecting non-overlapping community detection. Graph partitioning [1][2], clustering algorithm based in spectral analysis [3][4], clustering algorithm based on hierarchical [5] and density-based clustering algorithm [6][7] are disjoint community detection methods. There are many overlapping community detection methods such as clique percolation method [8], label propagation by Raghavan et. al [9] is some of the popular approaches.

In this research, the local seed expansion algorithm is proposed using Eigenvector centrality measure and Personalized PageRank algorithm. This seed set expansion method is one of the most popular approaches for overlapping community detection. The algorithm first partitioned the network into disjoint communities and then selects seed nodes for each community and expands these seed nodes to obtain overlapping communities.

This work is arranged as follows: Section 2 briefly describes the list of related works. Detailed stage for the presented algorithm is shown in Section 3. Section 4 displays a comparison of performance results with the Local Fitness Maximization (LFM) and Clique Percolation Method (CPM) algorithms. Finally, this paper ends in Section 5.

2. Related Works

In complex network, the study of cohesive groups, cliques, and communities is one of the most research topics and remarkable researches in this field have been described. There have been many overlapping community detection methods including CPM algorithm, LFM algorithm, COPRA algorithm, Eigenvector method, and Ego network analysis.

CPM proposed by Palla et. al is one of the most popular overlapping community detection method. In this approach, communities tend to form cliques due to the high density of internal edges. But CPM does not appropriate for large-scale complex network. This mechanism for defining cliques in k nodes is k cliques. According to the previous researchers' experiment, the value of k is appropriate between 3 to 6.

LFM [11] was proposed by Lancichinetti et al. 2009 is based on the local expansion approach. Firstly, LFM select random nodes as seed nodes. And then it expands these nodes to form a community. These expansion process stops if the fitness function is locally maximal.

Next, to create a new community, this method selects a node that does not included to any community. All of these communities' form overlapping social structures throughout the network.

It is based on the LPA algorithm represented by Cobra [12] Gregory. This extend LPA to copy social detection, allow each vertex to have up to u label. Where u is an input parameter. Label each vertex y with pair of pairs (a,b). C is the social ID and b is the link coefficient. This defined that the intensity range between y 0 to 1 and belong to the community. First each vertex is defined a label and the coefficient associated with the label is updated. Vertexes with the same label with eventually be divided into one community and vertexes with many labels will be copied vertexes linked to different communities.

The eigenvector method generalizes the spectral method and uses a soft clustering project put in to the eigenvectors of a normalized Laplacian or modularity matrix to estimate the community [13]. The ego network analysis method uses the theory of structural holes [14]. In this research, an overlapping seed expansion algorithm based on centrality measure is proposed to find out the best seeds and will be used these seeds to expand the communities for finding out the communities hidden in the network. According to the running results show that our new seed algorithm leads to improved coverage and quality of the generated community structure and competes with other popular overlapping community detection methods.

3. Overview of the Proposed Method

There are three main stages in our proposed algorithm: (1) Finding non-overlapping communities. (2) Selecting seed nodes. (3) Expanding seed nodes. The architecture of our proposed algorithm is shown in figure (1).

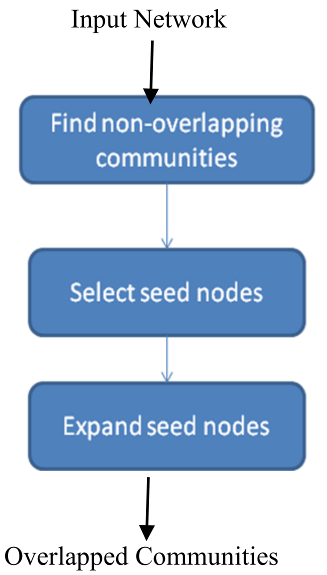

Figure 1. Architecture of the proposed algorithm

3.1. Finding Non-overlapping Communities

There are many graph partitioning methods. However, traditional graph partitioning methods need parameter to input the number of communities (number of seeds) in advance. This is impossible to input the number of seeds in advance for large scale complex networks.

So, we use the Louvain algorithm for graph partitioning. This algorithm does not need for parameter because of it automatically define for the number of communities. The Louvain Method of community detection first finds small communities by locally optimizing modularity on all nodes, then groups each small community into one node and repeats the first step. The optimized value is modular and is defined as a value between -1 and 1 that measures the density of links within a community compared to the links between communities.

3.2. Selecting Seed Nodes

For finding seed nodes, we use Eigenvector centrality measures. We find eigenvector value for each node in each disjoint community. And then we select the node with the highest eigenvector value as seed node. In graph theory, eigenvector centrality (also called eigen centrality or prestige score) is a measure of the influence of a node in a network. All vertexes in the networks are assigned relative scores based on the concept that relationships to high-scoring vertexes come up with more to the scores of problem nodes than same relations to low-scoring vertexes.

A high eigenvector score means that the vertex is connected to many nodes that themselves have a high score. The eigenvector centrality Eigen_u of vertex u is given by:

$$Eigenu = \frac{1}{\lambda}\sum_{k\in G} a_{u,k} x_k \tag{1}$$

Where, $A=(a_{k,u})$ be the adjacency matrix, i.e. $a_{k,u}=1$ if vertex **u** is connected to vertex k, $a_{k,u}=0$ otherwise. λ is a constant.

3.3. Expanding Seed Nodes

After finding seed nodes by using eigenvector centrality measure, we need to expand these seed nodes using Personalized PageRank algorithm. A personalized PageRank models the distribution of ranks when alpha determines the distance a random walker (called a random surfer in paper) from a source (often called the "source") can move. The main prominence of random walk-based approaches is that they can be calculated nearby and in equidistant, that the time and space requirements of such algorithms are independent of the size of the network [14], and they are identified by these types of algorithms. The community structurally close to the real community.

After expanding seed nodes, we get the set of nodes. And then we get the final community from these set of nodes with minimum conductance score. The most important aspect of cut-based measurement is conductance. The conductance of a cluster (set of vertices) begins with that cluster and measures the probability that a single step of randomness will leave that cluster.

4. Experimental Results and Analysis

To know the performance of the detected community by our proposed algorithm on many small-scale real-world networks, we use F1 score measure. And we compared the performance of our algorithm with other popular algorithms (LFM and CPM). The network datasets we tested are Zachary's karate club, American college football network, Risk Map Network, Dolphins' social network, Strike network.

F1-score is common score in binary classification, which is harmonic mean of precision (e.g., the proportion of positive identifications is correct) and recall (e.g., the proportion of actual positives is identified correctly). F1 score equation is shown as below:

F1(A,A′)=2∗precision(A,A′)∗recall(A,A′)/precision(A,A′) +recall(A,A′) (2)

where A is the ground-truth community and A' is the predicted community, precision (A, A′) is |A∩A′|/|A′| and recall (A, A′) is |A∩A′|/|A|. The higher F1 score, the higher community partition quality.

The detail information of the small-scale real-world network dataset are shown in Table1.

Table 1. Real-world network

Network	Node	Edge	Communities	Description
Karate	34	78	2	Zachary's karate club
Dolphins	62	160	4	Dolphin social network
Football	115	613	12	American football network
Strike	24	38	3	Strike social network
RiskMap	42	83	6	RiskMap network

Ground-truth communities and Predicted communities for Strike dataset are shown in Table 2.

Table 2. Communities of ground-truth and proposed algorithm for Strike dataset

Strike Dataset	Communities	Vertex in Community
Ground-Truth	Community-1	1,2,3,4,5,6,7,8,9
	Community-2	10,11,12,13
	Community-3	14,15,16,17,18,19,20,21,22,23,24
Proposed Algorithm	Community-1	2,5,6,7,8,9
	Community-2	10,11,12,13
	Community-3	14,17,18,19,20,21,22,23
	Community-4	22,23,24

Precision and Recall values calculated between the ground-truth Community-1 and the Commnity-1 detected by proposed algorithm are 1 and 0.6667. So, F1 score value for these two communities has 0.8000. For Strike dataset, average F1 score value for all detected communities that compared with ground-truth communities is 0.536250.

Table 3 Show the results compared the proposed method with CPM, LFM in terms of F1 score measure on the real-worlds networks.

According to the experimental results, for CPM algorithm, the K value (number of cliques) appropriate between 3 to 5 for small scale networks. For Karate network, the accuracy (F1 score) detected by our proposed method has only 0.265%. The proposed method is not satisfying the F1 score value in networks with highly sparse structure such as Karate dataset, however, the accuracy significantly high for dense networks such as Dolphins. The accuracy for Football dataset is 0.635, which is higher than other two methods. These football networks are dense network, which has average degree 10.6609. In real world, our proposed method appropriate since all real-world networks are large-scale dense networks.

Table 3. F1 score value for CPM, LFM and the proposed algorithm

	F1 score		
Network	CPM	LFM	Proposed Method
Karate	0.320 (k=3)	0.86	0.265
Dolphins	0.657(k=4)	0.467	0.707
Football	0.56(k=4)	0.450	0.635
Strike	0.261(k=3)	0.953	0.536
RiskMap	0.552(k=3)	0.384	0.624

5. Conclusion

In this paper, we presented the new overlapping community detection algorithm based on local seed selection and expansion strategy. The algorithm firstly partitioned the input network dataset into non-overlapping communities. For getting the seed nodes, the algorithm finds the eigenvector centrality value for each node in each disjoint cluster. Then the algorithm selects the nodes which have the highest eigenvector as seed nodes. Finally, it expands these seed nodes into overlapping communities using Personalized PageRank algorithm. We test our algorithm on many small-scale real-world networks and compared with two popular methods. The results of our new presented method that run on small- scale datasets show that our proposed method outperform than LFM, CPM methods.

References

[1]T. Heuer and S. Schlag, ``Improving coarsening schemes for hypergraph partitioning by exploiting community structure,'' in *Proc. Int. Symp. Exp. Algorithms*, 2017, pp. 21:1_21:19.

[2]Y. Wang, H. Huang, C. Feng, and Z. Liu, ``Community detection based on minimum-cut graph partitioning,'' in *Proc. Int. Conf. Web-Age Inf. Manage.*, 2015, pp. 57_69.

[3]J. Cheng, L. Li, M. Leng,W. Lu, Y. Yao, and X. Chen, ``A divisive spectral method for network community detection,'' *J. Stat. Mech. theory, Exp.*, vol. 2016, p. 033403, Mar. 2016.

[4]Z. Zhou and A. A. Amini, ``Analysis of spectral clustering algorithms for community detection: The general bipartite setting,'' to be published.

[5]B. Yang, J. Di, J. Liu, and D. Liu, ``Hierarchical community detection with applications to real-world network analysis,'' *Data, Knowl. Eng.*, vol. 83, pp. 20_38, Jan. 2013.

[6]T. Falkowski, A. Barth, and M. Spiliopoulou, ``DENGRAPH: A densitybased community detection algorithm,'' in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Nov. 2007, pp. 112_115.

[7]F. Meng, F. Zhang, M. Zhu, Y. Xing, Z. Wang, and J. Shi, ``Incremental density-based link clustering algorithm for community detection in dynamic networks,'' *Math. Problems Eng.*, vol. 2016, Dec. 2016, Art. no. 1873504.

[8]G. Palla, I. Derényi, I. Farkas, and T. Vicsek, ``Uncovering the overlapping community structure of complex networks in nature and society,'' *Nature*, vol. 435, no. 7043, pp. 814_818, Jun. 2005.

[9]U. N. Raghavan, R. Albert, and S. Kumara, ``Near linear time algorithm to detect community structures in large-scale networks,'' *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 3, p. 036106, 2007.

[10]G. Palla, I. Derényi, I. Farkas, et al., "Uncovering the overlapping community structure of complex networks in nature and society," Nature, Vol. 435, 2005, pp. 814-818.

[11]A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," New Journal of Physics, Vol. 11, 2009, No. 033015.

[12]S. Gregory, "Finding overlapping communities in networks by label propagation," New Journal of Physics, Vol. 12, 2010, No. 103018.

[13]S. Zhang, R.-S. Wang, and X.-S. Zhang, "Identification of overlapping community structure in complex networks using fuzzy C-means clustering," Physica A, vol. 374, no. 1, pp. 483–490, 2007.

[14]R. S. Burt, Structural Holes: The Social Structure of Competition. Cambridge, MA, US: Harvard Univ. Press, 1995.