International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
© Global Society of Scientific Research and Researchers
http://ijcjournal.org/

Overlapping Community Detection using Local Seed Expansion

Nyunt Nyunt Sein*

University of Computer Studies (Kalay), Myanmar
Email: dynedyne098494@gmail.com

Abstract

Communities are usually groups of vertices which have higher probability of being connected to each other than to members of other groups. Community detection in complex networks is one of the most popular topics in social network analysis. While in real networks, a person can be overlapped in multiple communities such as family, friends and colleagues, so overlapping community detection attracts more and more attention. Detecting communities from the local structural information of a small number of seed nodes is the successful methods for overlapping community detection. In this work, we propose an overlapping community detection algorithm using local seed expansion approach. Our local seed expansion algorithm selects the nodes with the highest degree as seed nodes and then locally expand these seeds with their entire vertex neighborhood into overlapping communities using Personalized PageRank algorithm. We use F1_score(node level detection) and NMI(community level detection) measures to assess the performances of the proposed algorithm by comparing the proposed algorithm's detected communities with ground_truth communities on many real_world networks. Experimental results show that our algorithm outperforms over other overlapping community detection methods in terms of accuracy and quality of overlapped communities.

Keywords: Community Detection; Overlapping Community Detection; Local Expansion .

1. Introduction

Finding of the cohesive groups, cliques and communities inside a complex network is one of the most studied topics in social networks analysis. It has attracted many researchers in sociology, biology, computer science, physics, criminology, and so on. Community detection aims at finding clusters as subgraphs within a given network. A community is a cluster where many edges link nodes within cluster but few edges link nodes between different clusters.

----

* Corresponding author.

In recent decades, community detection has found applications in many fields, such as complex network analysis [1], relation prediction [2], node identification [3], etc. In recent years, many community detection algorithms have been developed for various fields. In reality, there are many different kinds of communities in real-world networks : disjoint or nonoverlapping (e.g., students belonging to different disciplines in an institute) [4], overlapping (e.g., people in social network always belong to several group ,simultaneously, such as family, friends and colleagues) [5, 6], hierarchical (e.g., cells in the human body form tissues that in turn form organs and so on) [7], and local (e.g., a person having *uneven* interaction between certain members within a social group in Facebook) [8]. Disjoint community detection algorithms are partitional clustering, hierarchical clustering, spectral clustering, graph partitioning, genetic algorithms and random walks. Overlapping community detections are local expand and optimization, clique percolation method(CPM), link graph and link partitioning , agent based and dynamical algorithms, fuzzy detections, statistical inference based methods and non_negative matrix factorization(NMF) methods. In this paper, we propose an overlapped community detection algorithms based on see d expansion approach. The paper is organized as follows: Section 2, briefly outlines a list of related works. Detailed steps of the proposed method are presented in Section 3. Section 4, presents the experimental results, and comparison with state_of_art algorithms. Finally, this paper is concluded in Section 5.

2. Related Works

The problem of community detection in complex networks is an important research topic, and an impressive number of works in this field have been proposed. For overlapping community detection, there are many different approaches [6] including clique percolation, line graph partitioning, eigenvector methods, ego network analysis. The clique percolation method builds up the communities from *k*-cliques, which correspond to complete (fully connected) sub-graphs of *k* nodes. (E.g., a *k*-clique at *k* = 3 is equivalent to a triangle) [9]. According to our experiment, we can find out that the value of k appropriate for small scale networks (e.g. Karate) at k=3. Line graph partitioning is also known as link communities. Given a graph G= (V, E), the line graph of L (G) (also called the dual graph) has a vertex for each edge in G and an edge whenever two edges (in G) share a vertex. For instance, the line graph of a star is a clique. A partitioning of the line graph induces an overlapping clustering in the original graph [24]. Even though these clique percolation and line graph partitioning methods are known to be useful for finding meaningful overlapping structures, these methods often fail to scale to large networks like those we consider. Eigenvector methods generalize spectral methods and use a soft clustering scheme applied to eigenvectors of the normalized Laplacian or modularity matrix in order to estimate communities [11]. Ego network analysis methods use the theory of structural holes [12], and compute and combine many communities through manipulating ego networks [13,14]. There are many different community detection methods. However, some of these methods cannot have the ability to detect overlapping communities and to handle large real_world networks with reasonably computational cost. In addition, the community structure may be different to discover at the global level since the network organization at large scale would become very complex. The problem of community detection in large real networks is thus a challenging one, and the exploration of local view methods represents one of the alternatives in addressing this problem [6]. There have been a few local view methods for community detection [14,15,16]. Most of them apply one type of local community detection algorithm to expand the selected seeds. The local community detection algorithms usually start from the given seed and then expand by iteratively adding the neighbor node which contributes most to a specific score function (Modularity, Compactness-Isolation, etc…), until the score stops improving. The techniques based on approximate personalized PageRank are shown to be very effective for finding local communities [17,18]. In these techniques, an approximate personalized PageRank vector based on random walks from a seed is computed first, and the local community is then generated by performing a sweep over the PageRank vector using conductance or other community criteria. Gleich and Seshadhr propose using the nodes with local minimal conductance in their egonets as seeds [19], while Chen et.al. propose using nodes with local maximal degree in egonets as seeds[4]; Whang and his colleagues take the center of each cluster generated by Graclus [20] as a seed [16]. Moradi and his colleagues apply link prediction techniques to calculate the similarity of connected nodes and choose the nodes which are very similar to their neighbors as seeds, and then they further enhance this seed selection by applying a graph coloring algorithm [15]. In this paper, we propose an overlapping seed expansion algorithm to find out the best seeds and subsequently the best seeds from the proposed algorithm will be used to find out the communities hidden in the network. Experiments show that our seeding algorithm leads to an improvement on coverage and quality of the generated community structure, and is competitive in overlapping community detection.

3. Overview of the Proposed Method

We introduce our overlapping community detection algorithm , which consists of four phases: (1) find disjoint community (2) find core seed from each disjoint community (3) expand core seed nodes (4) calculate minimum conductance. The flow-graph of the proposed algorithm is shown in **figure 1** based on the seed expansion strategy.

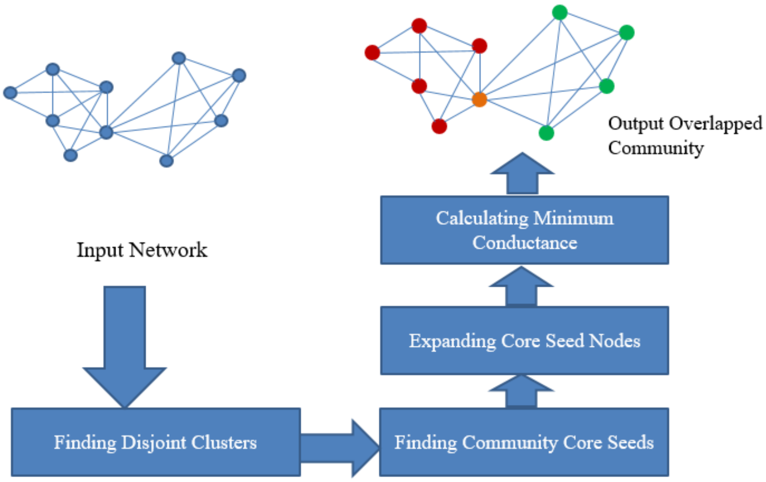
Output Overlapped Community
Input Network
Calculating Minimum Conductance
Expanding Core Seed Nodes
Finding Disjoint Clusters
Finding Community Core Seeds

Figure 1: The flow-graph of the proposed algorithm

3.1. Find disjoint community

One way to achieve these goals is to apply a high quality and fast graph partitioning scheme (disjoint clustering of nodes in a graph) in order to compute a collection of sets with fairly small conductance. In this phase, we use the Louvain method to find disjoint cluster. It processes a network edge by edge in the order that the network is fed to the algorithm. If a new edge is added, it just updates the existing community structure in constant time, and does not need to re-compute the whole network. Therefore, it can efficiently process large networks in real time. Our algorithm optimizes expected modularity instead of modularity at each step to avoid poor performance [21].

3.2. Find community core seeds

The selection of a seed node or a seed set is really important in the algorithms which use seeds to find out communities. Centrality analysis is one of the major research direction in social network analysis. The intuition behind centrality analysis is to discover group of central nodes from where maximum influence can be propagated in the network. Computation in detecting community can be minimized if central nodes can be identified at initial phase. Central node plays an important role in social relationships and community evolution. The person who has higher number of relationships seems to be an important person. Communities can be identified based on the set of central node in the network. Degree centrality, betweenness centrality, eigenvalue centrality and closeness centrality are some of the measures used to find out the central nodes in social network. Degree centrality is the most common feature from where community evolution is carried out. In this phase, we use degree centrality to find out community core nodes. Firstly, we find degree centralities on all nodes in each disjoint community. And then we sort nodes in descending order for each disjoint community. From each disjoint cluster, we pick the highest degree node and then select these nodes with its neighbor as seed nodes. Degree centrality of a node is defined as the number of edges incident upon it. In a directed network, there are two measures of degree centrality namely in-degree and out-degree. In-degree of vertex ‘v’ corresponds to the number of edges directed into a vertex ‘v’. Out-degree of vertex ‘v’ corresponds to the number of edges that are outward from the vertex ‘v’. Mathematically, degree centrality of a vertex ‘v’ of a graph G = (V, E) is denoted as DC (v), which is equal to degree of node ‘v’. The degree centrality of the network is defined by the following equation: [22]

$$DC_N = \frac{\sum_{i=1}^{n} D_{max} - D_i}{(n-1)(n-2)} \qquad (1)$$

where DC_N and D_{max} are the degree centrality and maximum degree of the network respectively.

3.3. Expand seed nodes

Once we have a set of seed vertices, we wish to expand the clusters around those seeds. An effective technique for this task is using a personalized PageRank vector , also known as a random-walk with restart. A personalized PageRank vector is the stationary distribution of a random walk that, with probability α follows a step of a random walk and with probability (1-α) jumps back to a seed node. If there are multiple seed nodes, then the choice is usually uniformly random. Thus, nodes close by the seed are more likely to be visited. The main advantages of random walk-based techniques are that they can be computed locally and in parallel, the time and space requirements of such algorithms do not depend on the size of the network [23], and the communities identified by these types of algorithms are structurally close to real-world communities. In personalized

PageRank algorithm, there are two parameters (ε,α). We set α =0.99 for an undirected graph. This value yields results that are similar to those without damping. The parameter ε is an accuracy parameter. After expanding seed nodes, we get the set of nods. And then we get the final community from these set of nodes with minimum conductance score. Conductance is one of the most important cut-based measure. The conductance of a cluster (a set of vertices) measures the probability that a one-step random walk starting in that cluster leaves that cluster. The conductance of a cluster is defined to be the cut divided by the least number of edges incident on either set Ci or V\Ci:

$$cond(C_i) = \frac{cut(C_i)}{min(links(C_i,V),links(V\backslash C_i,V))} \qquad (2)$$

4. Experimental Results

In order to quantitatively test our algorithm, we test it on real-world networks. We compare our algorithm with other state-of-the-art overlapping community detection methods: LFM [24] and CPM[9]. We demonstrate that our proposed algorithm' ability to detect meaningful overlapping communities in the real-world networks. The network datasets include: Zachary's karate club, American college football network, RiskMap Network, Dolphins' social network, Strike network. Some descriptions are given in Table I. There are many methods to measure the quality of detected communities. But only a few measures are suitable for overlapping communities. In our study, we introduce well-known evaluation criteria to measure the performance of the applied algorithms, Average (F1-score)[25].

$$F1 = \frac{2*precision*recall}{precision+recall} \qquad (3)$$

The statistics of the performance of the proposed algorithm, CPM, LFM algorithms for these real-world networks is collected in Table 2, Table 3, Table 4.

Table 1: Real World Networks

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

Table 2: F1-score for the proposed algorithm on Real-world Networks

Networks	Ground_Truth communities	Identified Communities	Ground_Truth Matched	Node Coverage	Proposed method
Karate	2	4	1	1	0.265
Dolphins	4	5	1	0.667	0.512
Football	12	9	0.750	0.870	0.569
Strike	3	4	1	0.791	0.446
RiskMap	6	7	1	0.881	0.686

Table 3: F1-score for the CPM algorithm on Real-world Networks**.**

Networks	Ground_Truth communities	Identified Communities	Ground_Truth Matched	Node Coverage	CPM method
Karate	2	3	1	0.941	0.320
Dolphins	4	4	0.750	0.730	0.322
Football	12	4	0.333	0.870	0.162
Strike	3	6	1	0.791	0.261
RiskMap	6	4	0.667	0.905	0.552

Table 4: F1-score for the LFM algorithm on Real-world Networks**.**

Networks	Ground_Truth communities	Identified Communities	Ground_Truth Matched	Node Coverage	LFM method
Karate	2	2	1	1	0.85
Dolphins	4	5	1	0.667	0.186
Football	12	9	0.750	0.870	0.605
Strike	3	3	1	1	0.983
RiskMap	6	5	1	0.857	0.832

5. Conclusion and Recommendations

The local seed expansion algorithm is proposed in this paper. The algorithm firstly partition the input network graph into disjoint communities by using Louvain algorithm and then from each disjoint community, select seed nodes which have the highest degree centrality , then expand these seed nodes into overlapping communities using Personalized PageRank. The proposed method is compared with other overlapping community mining algorithms in this paper. Experimental results compared with the existing prominent algorithms over different types of large real networks show that the proposed seed expansion algorithm is relatively insensitive to the keeping rate of support and is competitive in overlapping community detection. In this paper, we test our proposed algorithm on only small scale real world networks to measure the performance of the detected community and we can compare only two another algorithms.

References

[1]. Y. Liu, L. Nie, L. Han, L. Zhang, A. D. S. Rosenblum. Action2activity: recognizing complex activities from sensor data, in Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, 2015, pp. 1617–1623.

[2]. Y. Liu, Y. Zheng, Y. Liang, S. Liu, A. D. S. Rosenblum. Urban water quality prediction based on multi-task multi-view learning. in Proceedings of the International Joint Conference on Artificial Intelligence, 2016.

[3]. X. Wang, L. Nie, X. Song, D. Zhang, T.S. Chua. Unifying virtual and physical worlds: Learning towards local and global consistency. ACM Transactions on Information Systems, 36 (1) (2017) 1-26. https://doi.org/10.1145/3052774

[4]. S. Fortunato. 2010. Community detection in graphs. Physics Reports 486, 3–5 (2010), 75–174.

[5]. Chakraborty, S. Ghosh, and N. Ganguly. 2012. Detecting overlapping communities in folksonomies. In 23rd ACM Conference on Hypertext and Social Media. 213–218.

[6]. J. Xie, S. Kelley, and B. K. Szymanski. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Computing Surveys 45, 4, Article 43 (Aug. 2013), 35 pages.

[7]. M. F. Balcan and Y. Liang. 2013. Modeling and detecting community hierarchies. In Proceedings of the 2nd International Conference on Similarity-Based Pattern Recognition (SIMBAD'13). Springer-Verlag, Berlin,160–175.

[8]. X.Wang et al.: "Overlapping Community Detection Based on Structural Centrality in Complex Networks", in IEEE Access · November 2017, DOI: 10.1109/ACCESS.2017.2769484

[9]. G. Palla, I. Der_enyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," Nature, vol. 435, pp. 814–818, 2005.

[10]. Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," Nature, vol. 466, pp. 761–764, 2010.

[11]. 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.

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

[13]. B. S. Rees and K. B. Gallagher, "Overlapping community detection by collective friendship group inference," in Proc. Int. Conf. Adv. Social Netw. Anal. Mining, 2010, pp. 375–379.

[14]. M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon: A Local-first discovery method for overlapping communities," in Proc. 1 8th ACM Int. Conf. Knowl. Discovery Data Mining, 2012, pp. 615–623.

[15]. F. Moradi, T. Olovsson and P. Tsigas, " A local seed selection algorithm for overlapping community detection,", Proc, 2014 IEEE/ACM I nternational Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014),Beijing, China, 2014, pp. 1-8, doi:10. 1109/ASONAM.2014.6921552.

[16]. J.J. Whang. D.F. Gleich and I.S. Dhillon, "Overlapping community detection using seed set

expansion,' Proc. 22nd ACM International Conference on Information and Knowledge Management (CIKM'13) ,San Francisco, CA, USA, 2013, pp. 2099-2108, doi:10.1145/2505515.2505535.

[17]. R. Andersen, F. Chuang and K. Lang, "Local graph partitioning using PageRank vectors," Proc. 47th Annual IEEE Symposium on Foundations of Computer Science(FOCS' 06), Berkeley, CA, USA, 2006, pp. 475-486, doi: 10.1109/ FOCS. 2006.44.

[18]. J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground truth," Proc. 12th IEEE International Conference on Data Mining (ICDM 2012), Brussels, Belgium, 2012, pp. 745-754, doi:10.1109/ICDM. 2012.138.

[19]. D.F. Gleich, C.Seshadhri and Vertex neighborhoods, "low conductance cuts, and good seeds for local community methods, "Proc, 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD' 12), Beijing, China, 2012, pp. 597-605, doi:10.1145/2339530.2339628.

[20]. L.S. Dhillon, Y. Guan and B. Kulis, "Weighted graph cuts without eigenvectors: A multilevel approach," IEEE Transactions on Pattern Analaysis and Machine Intelligence (PAMI), vol, 29(11), 2007, pp. 1944-1957, doi:10.1109/TPAMI. 2007.1115.

[21]. G.Pan, W.Zhang, Z.Wu and S.Li , " Online Community Detection for Large Complex Networks", July 2014 | Volume 9 | Issue 7 | e102799.

[22]. R.K.Behera, D.Naik, "Centrality Approach for Community Detection in Large Scale Network", ACM COMPUTE '16, October 21-23, 2016.

[23]. R. Andersen and K. Lang. Communities from seed sets. In Proceedings of the 15th conference on World Wide Web, 2006.

[24]. A.Lancichinetti, S.Fortunato, J. Kertesz, "Detecting the overlapping and hierarchical community structure in complex networks," New Journal of Physics, 2009, 11(3):033015

[25]. Q.Cai, L. Ma, and M. Gong, "A survey on network community detection based on evolutionary computation," Int. J. Bio-Inspired Comput.,l vol.8, no.2, pp. 84-98,2014.