Multi-step locally expansion detection method using dispersed seeds for overlapping community

Simeng Wu, *Jun* Gong*, *Fei* Liu, and *Laizong* Huang

School of Software, Jiangxi Normal University, Nanchang 330022, China

Abstract. The local expansion method is a novel and promising community detection algorithm. Just based on part of network information, it can detect overlapping communities effectively, but some problems exist such as seed node aggregation, poor quality and inaccurate community coverage. Therefore, we propose a local expansion overlapping community detection algorithm based on dispersed seeds. There are four essential parts of this algorithm: 1) We firstly generate non-overlapping partitions of the network, and locate seed nodes with the largest influence in their own partition by using a new index of node influence, which combines the information centrality of nodes and the number of k-order neighbors. 2) Secondly, on the condition of the neighborhood overlap measure maximization, seed nodes merge unseeded nodes to generate a preliminary seed community; 3) Then based on the community conductance gain, the allocated nodes are screened and the free nodes are assigned to the seed community; 4) In the end, a node-community similarity based on common connection edge is proposed to re-allocate new free nodes and obtain the final community structure. This method can make the community distribution more proper and the coverage more reasonable. The experimental results on some artificial data and real network data show that the algorithm performs well on overlapping community indicators such as EQ and ONMI, while the community detection results are more stable.

Keywords: Overlapping community detection, Local expansion, Dispersed seeds, Node influence.

1 Introduction

The 21st century is the era of complex networks, as a significant feature, the community structure is generally defined as "nodes of the same community connect closely, while nodes between communities connect sparsely"[1]. Early research focused on disjoint community detection, and many representative algorithms have been put forward. However, the diversity in the real world makes objects not only belong to a certain category, thus the overlapping phenomenon of communities appears.

* Corresponding author: gongjun@jxnu.edu.cn

As the network size grows, it is difficult to obtain complete network information. Many methods have been developed to solve the problem, among these the local expansion algorithm has outstanding performance in terms of efficiency and effect. The algorithm mainly includes two parts: seed selection and community expansion, the main contributions of this paper are as follows:

We roughly divide several network partitions from a global perspective, and select seed nodes in each partition to make them evenly distributed in the entire network;

A node influence index, based on node information centrality and k-order neighbor number, is proposed to measure seed nodes in terms of "quantity" and "quality".

We consider the importance of common edges when measuring the strength of links between free nodes and the communities.

The remainder of this paper is structured as follows: Section 2 presents the related work of local expansion algorithms; Section 3 defines the basic problems of the research; Section 4 describes our algorithm in detail; Section 5 reports the experimental results; Section 6 summarizes the full text.

2 Related work

The LFM algorithm [2] detected network overlapping communities and hierarchical structure simultaneously for the first time. The EAGLE algorithm [3], different from the LFM algorithm, regarded the expanded modularity EQ as the target function. The GCE algorithm [4] selected the largest group as the seed, and expanded the community based on a greedy mechanism. In order to detect various types of networks, the OSLOM algorithm [5] took the direction and weight of edges into account, having advantages in directed graph and high overlapping networks. The DEMON algorithm [6] expanded communities by using the label information of neighbor nodes. The NISE framework [7] used the Graclus Centers algorithm to find high-quality seed nodes and proposed a new expansion strategy based on PageRank. The NDOCD algorithm [8] employed the greedy polynomial algorithm to find the maximum group. In order to reduce the effect of seed quality and built-in parameters on the algorithm, the LOCD algorithm [9] used structural centrality and weighted strategy to expand communities. The CFCD algorithm [10] introduced improved degree and fitness of k-core centrality. In order to further explore the overlapping phenomenon between communities, the LEBR algorithm [11] increased the processing of peripheral nodes of each community and reduced the dependence of algorithms on the target function.

3 Problem definition

Definition 1. Node influence I_v, measuring the influence of a node in the community, is defined as follows:

$$\mathrm{I_v} = |\ \mathrm{K\_Neighbor(v)}\ | \cdot \mathrm{Info\_Centrality(v)} \tag{1}$$

Where *Info_Centrality*(v) is the information centrality of node v, and q_{vj} is the sum of information propagated in all paths between the node pairs (v, j).

$$\mathrm{Info\_Centrality} = \frac{\mathrm{n}}{\sum_{\mathrm{j}} \mathrm{q_{vj}}} \tag{2}$$

Definition 2. Extended neighborhood overlap O_{ic}, based on neighborhood overlap degree [12], is extended to measure the closeness degree between non-seed node i and seed community c, defined as follows:

$$O_{ic} = \frac{|\ N(i) \cap N(c)\ |}{d_i + d_c - 2 - |\ N(i) \cap N(c)\ |} \tag{3}$$

Definition 3. Node-community similarity S_{ic}, measuring the degree of similarity between free node i and seed community c, is defined as follows:

$$S_{ic} = e_{ic} \cdot t_{ic} \tag{4}$$

Where, e_{ic} measures the importance of common edge to community c. t_{ic} measures the importance of common edge to the node i.

Definition 4. Conductivity $Cond(c)$ [13], measuring the closeness of the community c external connection, is defined as follows:

$$\mathrm{Cond(c)} = \frac{\mathrm{Edge(c, \bar{c})}}{\min\{\mathrm{degree(c)}, 2m - \mathrm{degree(c)}\}} \tag{5}$$

4 Multi-step locally expansion detection method using dispersed seeds

4.1 Algorithm framework

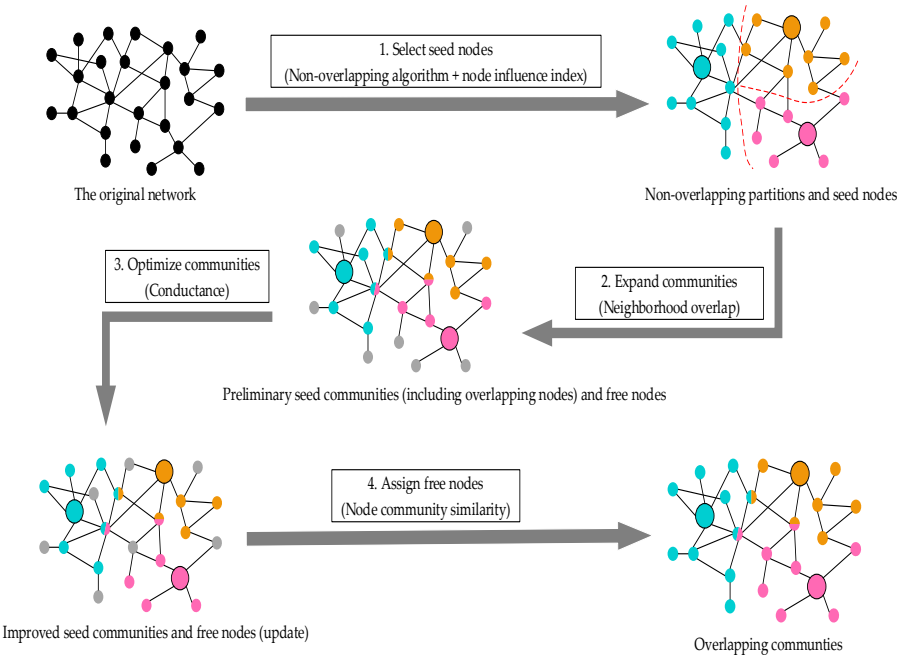


Fig. 1. Algorithm framework.

4.2 Algorithm complexity

While n represents the number of nodes, m represents the number of edges, and i represents the number of seed nodes. The time complexity of stage 1 seed nodes selection is $O(n)$; The time complexity cost of stage 2 community extension process is $O(snm)$. The number of free nodes is f_1, so the time complexity of stage 3 is $O(sf_1 + n - f_1)$. The number of free nodes is f_2, so the time complexity of stage 4 is $O(sf_2)$. In summary, since the values of s

and f in each step are small enough, so the total complexity of the algorithm in this paper is $O(mn)$.

5 Analysis of experimental results

5.1 Experimental data

The paper conducts experiment on 6 real-world networks datasets, including Karate, Dolphin, Les Miserables, Polbooks, Football and Power. Our algorithm is compared with 5 classical overlapping community detection algorithms, that is LFM, CPM, SLPA, GCE and LINK.

In this paper, LFR artificial data generation program [2] is used to generate 4 groups of network data with different structures by adjusting different parameters, including network mixed parameter μ, number of nodes n, number of overlapping nodes on, and maximum degree of nodes $maxk$.

5.2 Evaluation indicators and parameter settings

Overlapping module degree EQ [13]. Based on the modular degree Q [14], we adopt the extended modular degree EQ. The closer the value is to 1, meaning the better the result of community division is.

Overlapping standard mutual information $ONMI$ [15]. It measures the similarity of the two partitioning results, such as algorithms divide results and real community distribution.

Neighborhood overlap threshold h. In stage 2 of our algorithm, in order to avoid nodes with low-tightness joining the community, we set the overlap threshold h reasonably, the experiment results show that h is supposed to be set within the range [0, 0.2].

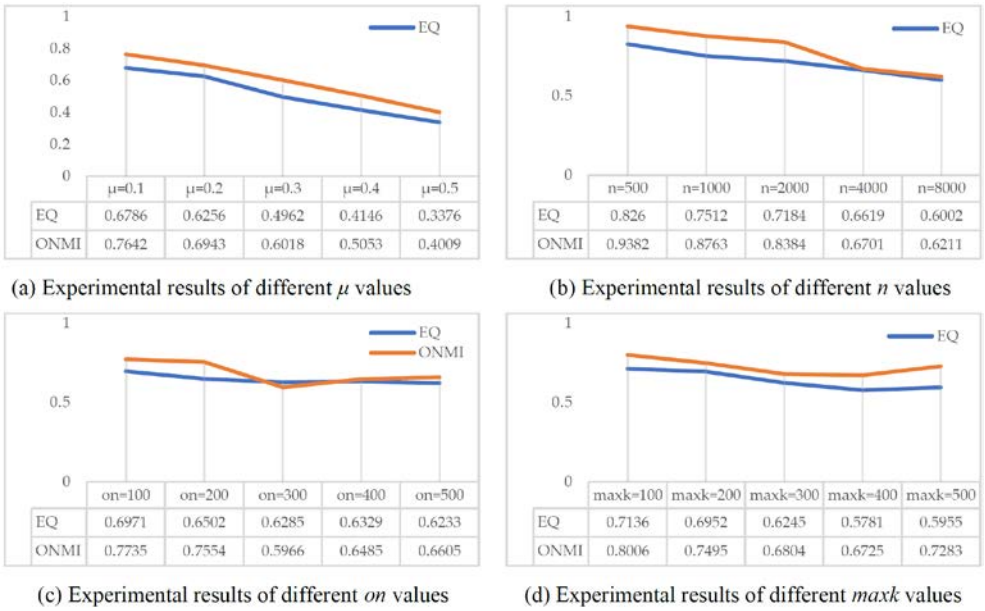
5.3 Result presentation and analysis

As shown in Table 1, the EQ and $ONMI$ values corresponding to the experimental results are superior to other comparison algorithms, and the index values all perform well.

Table 2. Formatting sections, subsections and subsubsections.

	Karate	**Dolphins**	**Les Miserables**	**Polbooks**	**Football**	**Power**
LFM	0.3320/0.3725	0.4078/0.4661	0.3115/0.6988	0.4310/0.4649	0.3799/0.5478	0.6210
CPM	0.1858/0.1653	0.3612/0.2752	0.2998/0.5462	0.4409/0.4219	0.3570/0.8822	0.5580
SLPA	0.3012/0.1321	**0.5040**/0.1206	0.4548/0.6213	0.3934/0.2258	0.5830/0.3019	0.6310
GCE	0.2930/0.5220	0.4550/0.4150	0.4370/0.6634	0.5020/0.4730	0.5430/**0.9250**	0.6980
LINK	0.1330/0.2810	0.1110/0.1540	0.2131/0.4380	0.0740/0.0600	0.2390/0.0670	0.4312
Mine	**0.3924/0.5995**	0.4983/**0.4902**	**0.5488/0.8869**	**0.5172/0.4936**	**0.6025**/0.8967	**0.7012**

Experimental results of mixed parameter μ, number of nodes n, number of overlapping nodes on, and maximum degree of nodes $maxk$ are as follows. With the increase of different parameter values, EQ value and $ONMI$ have slight fluctuation but have overall good performance, reflecting that the algorithm has high detection ability for complex, different scale networks even with fuzzy community structures. The specific information is shown in Figure 2.



(a) Experimental results of different μ values

(b) Experimental results of different n values

(c) Experimental results of different *on* values

(d) Experimental results of different *maxk* values

Fig. 2. Experimental results of each parameter value.

6 Discussion

This paper proposes multi-step locally expansion detection method using dispersed seeds for overlapping community, and the nodes influence index ensures the core position of seed node from two aspects of "quality" and "quantity". It properly solves the problem of low local community coverage, and gets overlapping communities with high quality by the multi-step optimization. Experimental results show that our algorithm has higher stability and efficiency and accuracy, compared with the 5 overlapping community detection algorithms (LFM, CPM, SLPA, GCE and LINK), and is suitable for multi-class structure networks. In the next step, we will study real networks with richer attribute structure, and apply the algorithm to a larger scale of network data by combining with parallel computing.

References

1. AL Barabási. The New Science of Networks[J]. Physics Today, 2003, 6(5): 444.
2. Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E, 2008, 78(4 Pt 2):046110.
3. Shen H W, Cheng X, et al. Detect overlapping and hierarchical community structure in networks[J]. Phys. A Stat. Mech. Appl. 2009, 388, 1706–1712.
4. Lee C, Reid F, Mcdaid A, et al[1]. Detecting highly overlapping community structure by greedy clique expansion[J]. 2010.
5. Lancichinetti A, Radicchi F, Ramasco J J, et al. Finding statistically significant communities in networks[J]. PloS one, 2011, 6(4): e18961.
6. Coscia M, Rossetti G, Giannotti F, et al. DEMON: a Local-First Discovery Method for Overlapping Communities[C]// Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

7. Whang J J, David F, Gleich, Inderjit S. Dhillon. Overlapping community detection using seed set expansion[P].[2,3] Information & Knowledge Management,2013: 2099-2108.
8. Ding Z, Zhang X, D Sun, et al. Overlapping Community Detection based on Network Decomposition[J]. Scientific Reports, 2016, 6:24115.
9. Wang X, Liu G, Li J. Overlapping Community Detection Based on Structural Centrality in Complex Networks[J]. IEEE Access, 2017:1-1.
10. Zhang J, Ding X, J Yang. Revealing the role of node similarity and community merging in community detection[J]. Knowledge-Based Systems, 2019, 165(FEB.1):407-419.
11. Ding X, Zhang J, Yang J. Node-community membership diversifies community structures: An overlapping community detection algorithm based on local expansion and boundary re-checking[J]. Knowledge-Based Systems, 2020, 198:105935.
12. Easley D A, Kleinberg J M. Networks, Crowds, and Markets: Reasoning About A Highly Connected World[M]. 2010.
13. Kannan R, Vempala S, Vetta A. On clusterings: Good, bad and spectral[J]. Journal of the Association for Computing Machinery, 2004, 51(3): p. 497-515.
14. Newman M, Girvan M. Finding and Evaluating Community Structure in Networks[J]. Physical Review E, 2004, 69(2 Pt 2):026113.
15. Newman M, Clauset A. Structure and inference in annotated networks[J]. Nature Communications, 2015, 7(2-3):11863.