Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 5, Issue 11: 79-84, DOI: 10.25236/AJCIS.2022.051112

Locally Expansion Directed Community Detection Method Considering Node Leadership

Fei Liu[1], Jun Gong[1,*], Laizong Huang[1], Shibin Li[1]

[1]*School of Software, Jiangxi Normal University, Nanchang, 330022, China*
[*]*Corresponding author: gongjun@jxnu.edu.cn*

Abstract: *In recent years, community structure has become a research hotspot in complex networks, and there are currently many excellent methods in the field of community detection, such as label propagation algorithms, LFM community detection algorithms, etc. These algorithms primarily investigate undirected networks, but the linking relationships between nodes in real complex networks are mostly directed and asymmetric, such as social networks. Therefore it is imperative to perform community detection on directed networks. We proposed a locally expansion community detection algorithm (DLE-NL) that can be applied to directed networks. There are four essential parts of DLE-NL algorithm: Firstly, the node cross-leadership index is proposed to select seed nodes considering local and global factors and directionality; secondly, seed node confidants are chosen to form the initial community using node following degree, and the core of the community is precisely mined; then node aggregation is carried out considering the node affiliation to the community with the initial community as the core. Finally, the initial reconstruction and optimization of the community are carried out by removing communities with less than three nodes, correcting the affiliation of nodes to the community based on the community directed-fitness function and the gain of the derivative, assigning free nodes and merging communities with high similarity. The experimental results show that the method performs well in the directed modularity and normalized mutual information metrics, while the community detection results are more stable.*

Keywords: *Directed network community detection; local expansion; node leadership*

1. Introduction

Complex networks can describe many complicated systems in the real world, and community structure is one of their essential structural characteristics, generally defined as "nodes in the same community are closely connected, while nodes between communities are sparsely connected" [1]. Studying the community structure of a network allows us to discover the functions of different types of complex networks more intuitively, helping us to grasp the evolutionary rules of their dynamics and to further investigate the dynamics of the topology of complex systems.

Due to the intricate relationships between individuals in real-world networks, link relationships usually show directionality and asymmetry; it is crucial to abstract actual networks into directed networks for community detection to explore the actual network ground community structure more effectively. Among the large-scale community detection methods for directed networks, community detection methods based on local extensions stand out in terms of efficiency and effectiveness. The algorithm mainly includes two parts: seed selection and community expansion, the main contributions of this paper are as follows:

1) A leadership-informed approach to high-density seed community generation. Considering both local and global importance, seed communities with high leadership and transmission capacity are used as a starting point for community expansion.

2) A community optimization strategy with approximately linear time complexity. Using the locally expansion method, directionality is incorporated into the process of selecting seed nodes and community expansion, community reconstruction is carried out through the movement of nodes, and isolated nodes in the network are reasonably allocated. Highly similar communities in the network are merged, improving the quality of community division, while the algorithm is easily parallelized and has good scalability.

The rest of the paper is structured as follows: Section 2 outlines the research work on directed network

ISSN 2616-5775 Vol. 5, Issue 11: 79-84, DOI: 10.25236/AJCIS.2022.051112

community detection algorithms; Section 3 defines the fundamental problem of the research; Section 4 describes the algorithms in detail; Section 5 shows and analyses the experimental results of the algorithms in this paper, and Section 6 concludes the whole article.

2. Related work

The literature [2] proposes a semi-supervised learning model that classifies directed networks. The literature [3] uses the idea of modularity optimization to transform a directed graph into an undirected network bipartite diagram for community detection. The literature [4] and others use spectral analysis and symmetry concepts to mine community structure by mapping directed networks to vector spaces. The literature [5] used a similarity matrix and spectral clustering algorithm to cluster nodes in a directed network to obtain community delineation results. The literature [6] transformed the directed network into an undirected network for community delineation using matrix symmetry methods and random walk symmetrization, etc. The literature [7] divides nodes based on feature vectors by extending the modularity metric to directed modularity and using spectral analysis. In the literature [8], the LinkRank modularity criterion was proposed to address the limitations of the directed modularity in Newman. The literature [9] studied directed network community detection using the idea of spectral equalization for the weighted cut problem. The literature [10] proposes algorithms for community detection based on local information and directed modularity, approaching linear complexity. The literature [11] proposes a parallelized spectral clustering algorithm with excellent clustering results.

In summary, scholars at home and abroad have achieved many results in studying complex networks and their association structures. Considering the directionality of connected edges is a critical direction in current association detection.

3. Problem definition

Definition 1. Directed cross-leadership L_v, which measures the influence of a node on other nodes in the neighborhood, consists of the node's cross-over degree and communication efficiency capability, is defined as follows :

$$L_v = Cross\_degree(v) * Node\_efficiency(v) \tag{1}$$

Where $Cross\_degree(v)$ is an extension of the degree metric for use in directed networks, and it is a local measure of node leadership. $Node\_efficiency(v)$ is the communication efficiency of node v. It reflects the node's global characteristics, the overall ease of communication between a node and other nodes in the network. $d_{v_i v_j}$ represents the distance from node v_i to node v_j.

$$Cross\_degree(v) = (k_{in}(v)\_ + 1)^{\alpha} * (k_{out}(v) + 1)^{1-\alpha} - 1 \tag{2}$$

$$Node\_efficiency(v_i) = \frac{1}{n-1} \sum_{j=1, j\neq 1}^{n} \frac{1}{d_{v_i v_j}} \tag{3}$$

Definition 2. Node following degree $F(v_i, v_j)$, depended on the proportion of the number of common neighbors of the two nodes to the number of neighboring nodes of the central node. $\delta(v_i, v_j)$ is the number of common neighboring nodes of node v_i and node v_j. It is defined as follows :

$$F(v_i, v_j) = \frac{\delta(v_i, v_j)}{d(v_i)},\ L(v_i) > L(v_j) \tag{4}$$

Definition 3. Node-to-community affiliation, measuring the affiliation degree of node to community. When there exists an edge with node i pointing to j, eij =1, and the opposite is 0. It is defined as follows :

$$Affiliate(i, C) = \frac{\sum_{j\in C} e_{ij} + e_{ji}}{\sum_{j\notin C} e_{ij} + e_{ji}} \tag{5}$$

Definition 4. Community similarity $S_{C_1 C_2}$. The ratio of the number of contiguous edges to the total

ISSN 2616-5775 Vol. 5, Issue 11: 79-84, DOI: 10.25236/AJCIS.2022.051112

number of nodes in the smaller community is calculated based on the similarity of the adjacent edges between the two communities, is defined as follows :

$$S_{C_1C_2} = \frac{|\{(i,j)\,|i \in C_1, j \in C_2\}|}{\min\{|C_1|,|C_2|\}} \tag{6}$$

Definition 5. Directed optimization fitness $f(c)$, used to determine the degree of influence a node has on a community c when it joins the community c. When f(c) reaches a peak, it forms an optimal score for community c, defined as follows :

$$f(c) = \frac{k_c^{in\_out} + k_c^{in/out}}{k_c^{in\_out} + k_c^{in/out} + k_{c\_in}^{in\_out} + k_{c\_in} + k_{c\_out}} \tag{7}$$

Where $k_c^{in\_out}$ denotes the number of all edges in community c that point to each other; $k_c^{in/out}$ represents the number of edges in community c that are unidirectionally connected.

4. Locally expansion directed community detection method considering node leadership

4.1. Algorithm description and framework

In response to the fact that most current association detection algorithms are limited to undirected networks, this paper adopts the idea of the local extension, taking into account the importance of both local and global aspects, and uses a high-density, high-total-degree and non-overlapping seed community as the starting point for community extension, incorporating directionality into the process of selecting seed nodes and community extension for community detection. It consists of four main stages: seed node selection, seed community formation, community neighborhood expansion, and community optimization. The algorithm framework is shown in Figure 1.

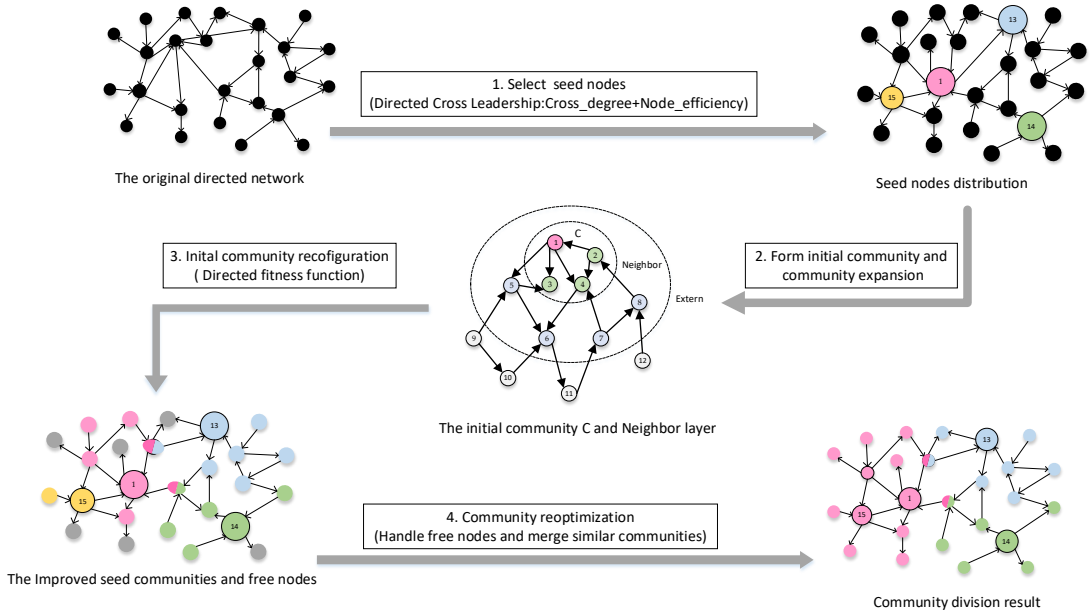

Figure 1: Algorithm framework

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 $O(smn)$. 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

ISSN 2616-5775 Vol. 5, Issue 11: 79-84, DOI: 10.25236/AJCIS.2022.051112

of the algorithm in this paper is $O(mn)$.

5. Analysis of experimental results

5.1. Experimental data

The paper conducts experiments on three real-world network datasets: the Poblogs political blog network, the Cora citation network and the BlogCatalog social network. We compared the DLE-NL algorithm with the Newman directed algorithm.

Also, we used computer-generated directed Benchmark artificial networks [12]. The mixing coefficient mu is a parameter that affects the clarity of the network community structure. The smaller the mu value, the more pronounced the community structure is. The community structure becomes increasingly blurred when mu is more significant than 0.5, so the experiments in this paper set the mu value in the range of 0.1 to 0.7. The DLE-NL algorithm was compared with the FTA algorithm [13] directed algorithm.

5.2. Evaluation indicators and parameter settings

Directed modularity [7]. Considering the directional nature of directed networks, based on the modularity degree Q [14], we use the extended directed modularity degree DQ. The closer the value is to 1, the better the result of community segmentation is implied. It is defined as follows:

$$DQ = \frac{1}{m}\sum_{i,j}\left[A_{ij} - \frac{k_i^{in}k_i^{out}}{m}\right]\delta(c_i, c_j) \tag{8}$$

Where A_{ij} denotes the adjacency matrix of the directed network, and k_i^{in} and k_i^{out} represent the indegree and outdegree of the node.

Standard mutual information NMI[15]. It measures the similarity of the two partitioning results, such as algorithms divide results and real community distribution. It is defined as follows:

$$NMI(C_1, C_2) = \frac{2\sum_{i\in C_1}\sum_{j\in C_2} P(i,j)\log\frac{P(i,j)}{P(i)P(j)}}{\sum_{i\in C_1} P(i)\log P(i) + \sum_{j\in C_2} P(j)\log P(j)} \tag{9}$$

Where C_A, and C_B are the number of associations in the division results of A and B, respectively, M_{ij} denotes the number of nodes in the actual community i that are identical to those in the experimentally found community j, $M_{i.}$ is the sum of row i of the M matrix, $M_{.j}$ is the sum of column j of the M matrix. N is the total number of nodes in the network. The closer the value is to 1, the more similar the experimental community division result is to the actual community structure.

5.3. Result presentation and analysis

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

Table 1: The result of DLE-NL algorithm and Newman directed algorithm

Dataset	DLE-NL algorithm			Newman directed algorithm		
	DQ	NMI	Number of communities	DQ	NMI	Number of communities
Dirnet	0.682	0.97	4	0.659	0.96	4
Poblogs	0.406	0.73	2	0.448	0.68	3
Cora	0.422	0.54	8	0.363	0.48	10

Table 2 presents information on the structure of the directed artificial network for three different mu values (total number of nodes N=1000, average entry k=3, maximum entry maxk=10, minimum community size minc=50, maximum community size maxc=200).

ISSN 2616-5775 Vol. 5, Issue 11: 79-84, DOI: 10.25236/AJCIS.2022.051112

Table 2: Information about the LFM artificial network community for different mu values

mu	Number of edges	Number of communities	Minimum community nodes	Maximum community nodes
0.1	3522	7	83	196
0.6	3494	9	57	210
0.7	3495	7	113	197

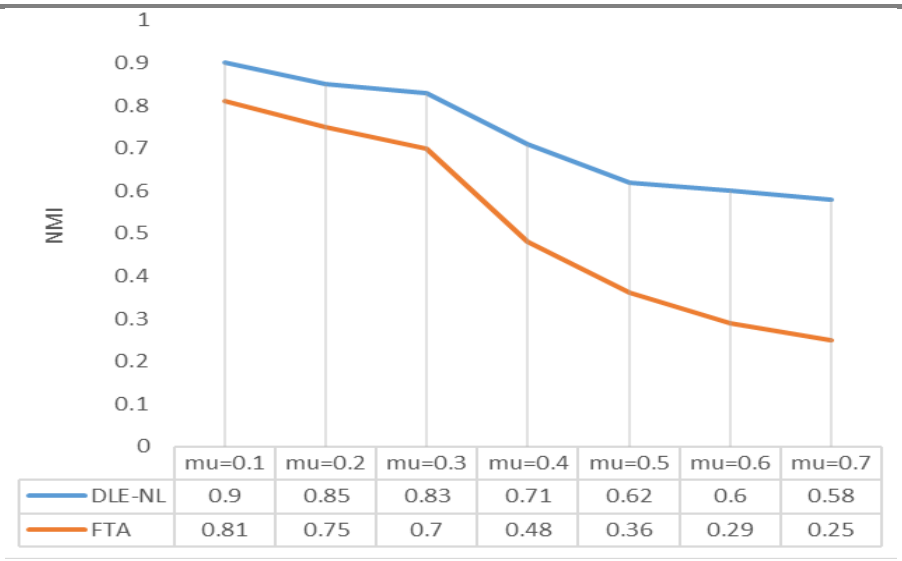

Figure 2: Experimental results of different mu value

Figure 2 line graph shows the NMI values for community detection in the directed network for different values of the mixing parameter mu for the DLE-NL algorithm and the FTA algorithm. The DLE-NL algorithm performs better for smaller values of mu, and for larger values of mu, the DLE-NL algorithm NMI values drop but are still higher than the FTA algorithm.

6. Discussion

This paper proposes a local extension-based community detection algorithm for directed networks, which integrates local and global importance, incorporates directionality into the directed cross-leadership index and reduces the randomness of seed selection.

It also adopts a community expansion method that considers leadership and parallelization to distribute the nodes that stray in the network. It better solves the problem of low local community coverage and obtains a high-quality community distribution through multi-step optimization. The experimental results show that the algorithm proposed in this paper not only can fully take into account the topological structure information of the directed network but also has higher stability, efficiency and accuracy and is suitable for multi-class structural networks. In the next step, we will study real networks with richer attribute structures and apply the algorithm to larger-scale network data by combining it with parallel computing.

References

[1] AL Barabá si. The New Science of Networks [J]. Physics Today, 2003, 6(5): 444.
[2] Schölkopf B, Hofmann T, Zhou D. Semi-supervised learning on directed graphs [J]. Nips, 2005, 17(2005):1633-1640.
[3] Guimerà R, Sales-Pardo M, Amaral L A N. Module identification in bipartite and directed networks [J]. Physical Review E, 2007, 76(3): 036102.
[4] Lai D, Lu H, Nardini C. Finding communities in directed networks by PageRank random walk induced network embedding [J]. Physica A Statistical Mechanics & Its Applications, 2010, 389(12):2443-2454.
[5] Cooper K, Barahona M. Role-based similarity in directed networks [J]. Physics, 2010.1012(2726): 1-4.
[6] Satuluri V, Parthasarathy S. Symmetrizations for clustering directed graphs[C]. Proceedings of the

ISSN 2616-5775 Vol. 5, Issue 11: 79-84, DOI: 10.25236/AJCIS.2022.051112

14th International Conference on Extending Database Technology. ACM, 2011: 343-354.
[7] Leicht E A, Newman M E J. Community structure in directed networks[J]. Physical Review Letters, 2008, 100(11): 2339-2340.
[8] Kim Y, Son S W, Jeong H. Finding communities in directed networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 81(1):016103.
[9] Meilă M, Pentney W. Clustering by weighted cuts in directed graphs[C]. Proceedings of the 2007 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2007: 135-144.
[10] Liu S G. Detecting communities of directed networks via a local algorithm[J]. J. Math. Inform, 2013, 1: 43-51.
[11] Chen W Y, Song Y, Bai H, et al. Parallel spectral clustering in distributed systems.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33(3):568-586.
[12] Lancichinetti A, Fortunato S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities [J]. Physical Review E, 2009, 80(1):016118.
[13] Lidong Fu, Dan Li, Zhanli Li. The membership tree algorithm detects overlapping communities in complex networks [J]. Computer Science, 2019, 46 (12): 330-334.
[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.