ATLANTIS PRESS

Advances in Computer Science Research, volume 87

3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)

A Fast Community Detection Algorithm based on Clustering Coefficient

Run Zheng

College of Information and Engineering, Nanjing University of Finance & Economics, Jiangsu 210023, China.

zhengrun1101@163.com

Abstract. Community detecting has always been a hot topic in the complex network research area, the fast and accurate community detection can provide a good foundation for the research of complex network nature. With the number of network nodes increasing, the structure of network becomes complicated, the traditional community detection becomes more unpractical for they are all based on global information of network. We offered a local community detection algorithm which didn't need to know the whole complex network information, it just beginning with a node as initial community and computing the intensity between the initial node and its adjacent nodes. And add its adjacent nodes to community gradually; eventually get this node community structure. Above all, this method can achieve global network community detection. We applied this method to American College football network and dolphin social network, and experiment results show accuracy and feasibility.

Keywords: Community detection; Clustering coefficient.

1. Introduction

Real-world networks often show a high degree of organization, and we often describe this system as complex network, such as collaborating network, biological network [1]. Closely connected nodes have relatively dense interconnected edges, forming a relatively tight set of nodes. However, there are fewer connected edges and the relationship is loose between node groups. The research of community structure in complex networks can help to decompose large complex systems gradually and divide them into multiple more aggregated structures, which is easy to understand network composition and function, and discover the interaction between nodes in the network. And how to make correct community division in complex networks becomes a hot topic in current complex network research. Towards this problem, researchers proposed various community detection methods, for example, Girvan et al. proposed the GN algorithm by iterative removal of edges with high "betweenness" scores [3], Rovall et al. provided Infomap method [4], LPA method based on the idea of label propagation [6]. These methods detected the community based on the global network information. However, with the scale of network become larger, calculating the global information is more complicated.

In many circumstances, researchers only focused on the local community structure in complex network, such as just caring one person about which community he belongs to instead of exploring community structure in the whole network which is time-consuming. In order to addressing this problem, scholars proposed the local community detection algorithm, for instance, James P. Bagrow proposed the methods discovering the local community[7], Aaron Clauset proposed the concept of local modularity, the target is maximizing local modularity to complete the search of local community[8].

This paper proposed a local community detection algorithm based on clustering coefficient, which can not only be employed in global community detection, but discovery the local network community. In the circumstance of large scale of global network, we can first divide the whole network into several small network through calculating the clustering coefficient. Then we select the node with the highest degree of clustering coefficient as the initial node in the network, examine the tightness between the initial node and its neighboring nodes, add the adjacent nodes to the community, and run the above steps iteratively until all nodes are assigned to the community to complete the detection of community structure in global network. If to discover local community, firstly we initialize a node

and assess the tightness between this node and it's neighboring nodes, then iteratively add the neighboring nodes to existed community, algorithm terminated until we get the community structure of this node belongs to.

2. Definition

Given an undirected and unweighted graph $G(V, E)$, V is a set of n nodes and E is a set of m edges. Let e_{ij} be the edge between vertex i and vertex j, and $V = \{ i \mid i \in n \}$, $E = \{ e_{ij} \mid i, j \in n$ and $i \neq j \}$. Assuming δ is a community structure in G, $| \delta |$ denote the number of nodes in this community.

(1) Neighbor set

Neighbor set of a vertex i can be described as $N(i) = \{j \mid e_{ij} \in E\}$.

(2) Degree of vertex and community

The degree of vertex i in graph indicates the number of neighbors directly connecting to this vertex, denoted as k_i. We define the internal and external degree of vertex $i \in \delta$, k_i^{int} and k_i^{ext}, as the number of edges connecting i to other vertices of δ or to the rest of the graph, respectively. If $k_i^{\text{int}} = 0$, the vertex has neighbors only within δ, which is likely to be a good clustering for i; if $k_i^{ext} = 0$, instead, the vertex is disjoint from δ and it should better be assigned to a different cluster. The internal degree k_{int}^{δ} of δ is the sum of the internal degrees of its vertices. Likewise, the external degree k_{ext}^{δ} of δ is the sum of the external degrees of its vertices. The total degree k^{δ} is the sum of the degree of the vertices of δ. By definition, $k^{\delta} = k_i^{\text{int}} + k_i^{ext}$.

(3) Clustering coefficient

The clustering coefficient of vertex i is defined as the ratio of the number of actual edges between all adjacent vertices to the number of possible connected edges between adjacent vertices, described as follow:

$$coe(i) = \frac{2E(i)}{k_i(k_i - 1)} \tag{1}$$

where $coe(i) \in [0,1]$. It indicates that all neighbor vertices of node i are connected to each other when $coe(i) = 1$, and node i form complete mutuality with all directly connected neighbors.

The clustering coefficient of the edge in the network is defined as the ratio between the triangle formed by the two ends of the edge and the common neighbors and the number of triangles may contain [9], described as follow:

$$coe(e_{ij}) = \frac{z_{ij}}{\min(k_i - 1, k_j - 1)} \tag{2}$$

Where $coe(e_{ij}) \in [0, 1]$, which reflects the tightness of the connection between the nodes at both ends of a certain edge. z_{ij} denotes the real number of triangles in the network that contain this edge.

(4) Clustering centrality degree

The clustering centrality degree of a vertex i is defined as the sum of the edge clustering coefficient of the edges directly connected to the vertex, computed as follows:

$$C_{ccd}(i) = \sum_{j \in N(i)} coe(e_{ij}) \tag{3}$$

The $C_{ccd}(i)$ reflects the number of neighbors of a node and the closeness between the node and its neighbors. The node with a larger centrality degree is more similar to the community centrality node.

3. A Local Community Detection Method based on Clustering Coefficient

3.1 Basic Ideas

This paper proposed an algorithm begins with an initial vertex, and merges the vertex into the result community to get a local community. Then we calculate the tightness of all neighbors of the community and add the vertex with largest tightness value to community to conform preliminary community division. Eventually we incorporate communities based on tightness between different communities. Algorithm terminated until the community is no longer vital to all its neighbors to reach result of community detection.

3.2 Algorithm Details

3.2.1 Initial Node

Consider the select of initial node in community detection algorithm based on the clustering coefficient, the general methods start the node with the largest clustering centrality degree in the network as the initial node for community detection, which is simple and convenient. However, it only takes into account the characteristics of the core nodes in the network. It does not take into account that the core node has a vital feature that its connection with neighbors is relatively close, and its neighbors also contact closely [11]. Therefore, we initialized the vertex considering the tightness between the node and its neighbors, and the closeness between adjacent node. So, the concept of node clustering degree is defined as formula (4):

$$C_{ncd}(i) = \sum_{j \in N(i)} C(i) + \frac{coe(e_{ij})}{k_i} \tag{4}$$

Where $C_{ncd}(i)$ express the sum of a node clustering coefficient and the means of edge clustering coefficient directly connected with this node.

3.2.2 Algorithm Description

a) Initial partition

Our method divides the initial community according to the following four criteria.

If a community with clustering coefficient $C_\delta = 1$, we add all its adjacent nodes to community.

If the node clustering coefficient of any adjacent node in community is 1($coe(i) = 1$), then add the adjacent node and other nodes directly connected to it to the community.

If any adjacent node of the community satisfies: $k^{in}(\delta) > k^{out}(\delta)$, we add this adjacent node to community.

If edge eij clustering coefficient meet: $coe(e_{ij}) > coe(e_{jk})$, $j \in N(i)$, $k \in N(j)$. we add the vertex j to community.

If all the nodes in the network were divided into their belonging communities, we complete the initial division and detect many sub-communities.

b) Incorporation refinement

We inspect the result communities detected from initial partition, and the eligible communities will be merged with the following three criteria.

Firstly, we calculate the D-value of internal degree and external degree of the existed communities. If D-value > 0 or D-value < 0, we do not need to incorporate any communities.

If D-value is 0, we call this community as non-connected community, and merge the community into weakly connected community or non-connected community.

If the tightness of newly incorporated community is improved, we keep the result reserved; and the tightness is not improved, we still incorporate the non-connected community with other directly connected community, algorithm terminated until we get the newly incorporated community with improved tightness or no communities can be merged.

3.2.3 Algorithm Implementation

Algorithm implementation steps can be described in detail as algorithm 1:

Algorithm 1: community detection algorithm based on clustering coefficient

Step 1: if (local community detection):
node *i* as initial node, go Step 3;
endif
Step 2: if (global community detection):
calculate the edge clustering coefficient $coe(e_{ij})$, and remove the edge of $coe(e_{ij}) = 0$, get small-scaled set of sub-communities *S*;
calculate node clustering degree $C_{ncd}(i)$, the vertex with max $C_{ncd}(i)$ is the initial node, go Step 3;
endif
Step 3: regard node *i* as initial community;
do ($j \in N(i)$ and *j* satisfy the four criteria):
add node *j* to initial community, removing node *j* from previous community;
end
if (global community detection):
select the initial node iteratively in sub-communities *S* and Step 3, until all nodes were added to corresponding communities, go Step 4;
endif
Step 4: calculate the D-value of n communities, expressed as $D(i)$;
foreach ($D(i) == 0$):
incorporate the community with adjacent community, calculate the D-value again;
if ($D(i) > 0$):
keep the community reserved, and refresh the community;
endif
if ($D(i) == 0$):
incorporate the community with other directly connected community until get the improved connectivity relevance or no existed communities to incorporated, output n communities;
endif
end

3.2.4 Algorithm Analysis

Consider local community detection, this method starts with the target node, the time of search an adjacent node is $O(e \times d)$, where e, d denote the number of edges in community and average degree of neighbors respectively. The method is convenient and simple compared to traditional global algorithm. As to global community detection, the time of node clustering degree is $O(m \times d)$, m expresses the the number of whole networks; the time of initial partition is $O(a \times d^2)$, a is the average neighbors of communities; the time of incorporating initial partition is $O(n)$, n is the number of vertices. It can be clearly observed that time-consuming step is in initial partition, especially calculation of $coe(i)$ and $coe(e_{ij})$, and less time complexity compared to classic algorithms.

4. Experimental Setup and Results Analysis

In this section, we verified our algorithm on two real-world networks including college football network [12] and dolphin social network [13,14]. Also, we compared this method with two other algorithms proposed by Kong [2] and Zhang [11], we called them *NCC* algorithm and *LCC* algorithm respectively. Experiment results demonstrate that our algorithm can detect the communities more accurate and effective, and thereby verify the quality of our algorithm accordingly.

4.1 Real-world Network

Dolphin network: Dolphin dataset was also a real-world network, and often used to testify the results of community detection methods. The network was observed by D. Lusseau reflecting living

situation of 62 dolphins in New Zealand, which consisting 62 nodes and 159 edges, the nodes represent dolphins and the edges represent the contact frequency between the two dolphins. This dolphin network included two dolphin families with 42 members in comparatively larger family and 20 members in the smaller. its real information is shown as Figure 1, where difform nodes indicates different community.

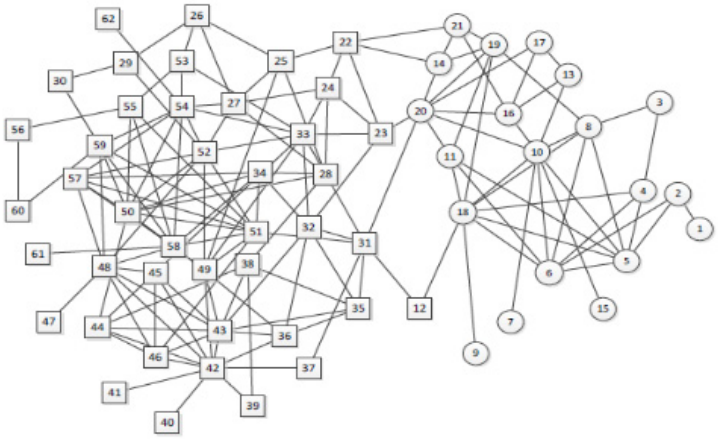

Figure 1. dolphin network

College Football Network: College football network is a complex social network created by Newman based on the American College Football League. The network consists of 115 nodes and 616 edges, where the nodes in the network represent the football team and the edges between the two nodes represent one match between two teams. The 115 college students participated in the competition were divided into 12 leagues. The regulation of the game is that the teams within the league play the group match first, then the team between the leagues. This expresses the information that the number of matches between teams within the league is greater than the number of matches between teams in the league. The alliance can be expressed as the real community structure of the network. As is shown in Figure 2.

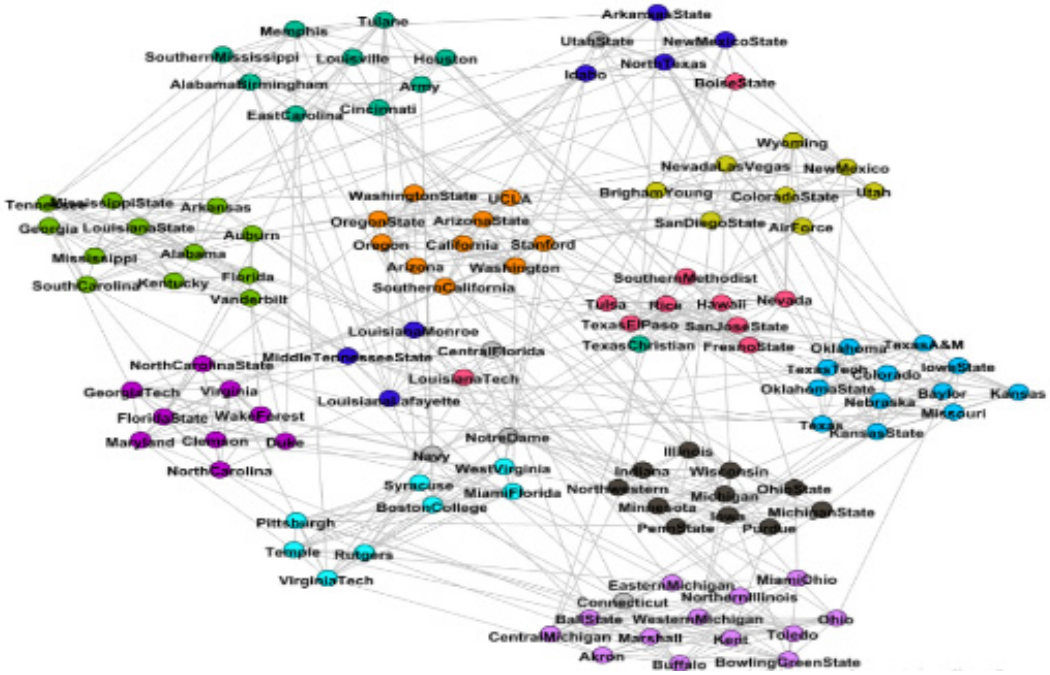

Figure 2. College Football Network

Polbooks network: This is a network of books on U.S. politics, founded by V.Krebs and is used for demonstration by Newman. The nodes represent books on American politics bought from Amazon.com, while edges between books represent frequent co-purchasing of books by the same

buyers. By reading the descriptions and reviews of the books posted on Amazon, the books are eventually classified into three categories.

NetScience network: This is a network of co-authorship of scientists working on network theory and experiment. The network is weighted, with weights assigned as described by Newman.

Table 1. network parameters

Network	Vertex	Edge	Weighted	Directed
Dolphin	62	159	No	No
Football	115	616	No	No
Polbooks	105	441	Yes	No
NetScience	1589	2742	Yes	No

4.2 Results Analysis

We applied our algorithm in two ground-truth networks, and corresponding division results are demonstrated as follow.

Dolphin network: This dolphin dataset included two clusters consisting 42 nodes and 20 nodes in corresponding cluster. LCC algorithm and NCC algorithm include 34 sub-communities and 18 sub-communities in corresponding division in initial partition stage, and we eventually detect 3 communities and 2 communities through incorporation stage. Table 2 demonstrate specific division result, we can observe the untrustworthy division result of LCC algorithm for dividing the network into 3 communities, and after the NCC algorithm, it detected same number of real networks with node 22 and node 12 separated wrongly. Our algorithm detected the trusted division result, and meanwhile node 12 and node 22 were divided correctly.

Table 2. Division result comparison

	LCC		*NCC*		Our algorithm	
	communities	nodes	communities	nodes	communities	nodes
division result	3	(1,2,3,4,5,6,7,8,9,10,11,13,14,15,16,17,18,19,20,21,23,32)	2	(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,21,22)	2	(1,2,3,4,5,6,7,8,9,10,11,13,14,15,16,17,18,19,20,21)
		(35,37,38,39,40,41,42,43,44,45,46,47, 51,52		(20,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62)		(12,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62)
		(12,22,24,25,26,27,28,29,30,31,33,34,36,48,49,50 ,53,54,55,56,57,58,59,60,61,62)				

College Football Network/ Polbooks network/ NetScience: We use two metrics to evaluate the quality of the different algorithms. The first metric is the *F1Score*, following the approach taken by the authors of the community benchmark [15]. The second metric is the Normalized Mutual Information (*NMI*) [16], which is based on information theory concepts. The *NMI* provides a real number between zero and one that gives the similarity between two sets of sets of objects. An exact match between the two inputs obtains a value of one.

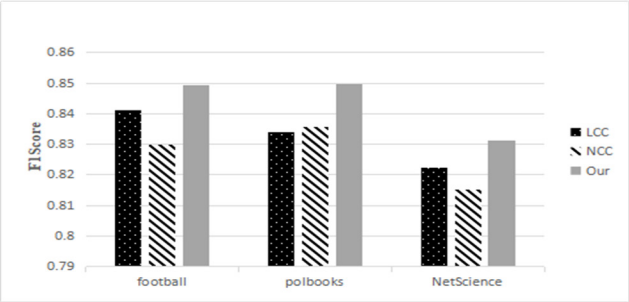

Figure 3. F1Score using ground truth

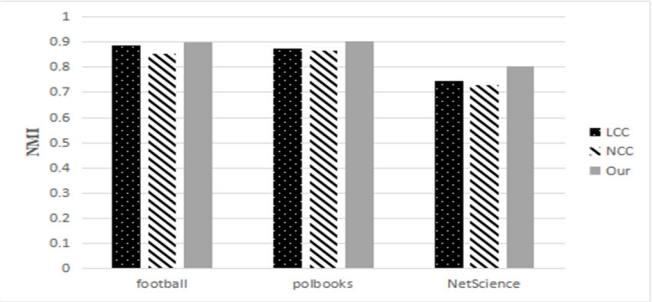

Figure 4. NMI using ground truth

Figures 3 and 4 show the *F1Score* and *NMI* quality scores of the tested algorithms, respectively. We observe that both metrics are correlated, though some small deviations exist among them.From these results, we conclude that our algorithm obtains the best quality. In terms of *F1Score*, our algorithm obtains the best quality in all the tested graphs, and performance fluctuation is not clearly with the different scales of network. In case of *NMI*, our algorithm quality is also the best though the performance of the tested algorithms is little different. Broadly speaking, the proposed algorithm improved the quality of the community detection algorithms.

5. Conclusion

This paper proposes a local community detection algorithm based on clustering coefficient, which can not only be applied in both community detection of global networks, but in local networks. In local detection of network, we initialize the pending node and calculate the tightness this node and it's neighbors, then add the adjacent node to community gradually, and the structure of communities. In the circumstance of global detection, we firstly remove the edges with 0 of clustering coefficient, then select the nodes with largest clustering degree as initial centrality and add these nodes to communities. Subsequently, the tightness between initial nodes and the neighbors are calculated to incorporate nodes to existed communities gradually, eventually we obtained the division of whole network. This algorithm obtains the less number of scattered division in initial partition, which incorporate nodes more conveniently in subsequent stage. Especially in local detection, the method still can detect the community without whole network information, improving the performance of traditional algorithm with the scale of network becoming larger.

Acknowledgments

The authors wish to thank the anonymous editors and reviewers for their valuable comments and helpful suggestions that greatly improved this paper's quality. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1]. Fortunato, Santo. Community detection in graphs[J]. Physics Reports, 2009, 486(3):75-174.

[2]. Kong-Wen L I. Local Community Detecting Method Based on the Clustering Coefficient[J]. Computer Science, 2010.

[3]. Newman MEJ, Girvan M. Finding and evaluating community structure in networks[J]. J, Phys Rev E,2004,69:026113.

[4]. Rosvall M, Bergstrom C T. Maps of Information Flow Reveal Community Structure in Complex Networks[J]. Proceedings of the National Academy of Sciences Usa, 2007:1118--1123.

[5]. Newman M E J. Fast algorithm for detecting community structure in networks. [J]. Phys Rev E Stat Nonlin Soft Matter Phys, 2004, 69(6 Pt 2):066133.

[6]. Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2007, 76(3 Pt 2):036106.

[7]. Bagrow J P, Bollt E M. Local method for detecting communities. [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2005, 72(2):046108.

[8]. Clauset A. Finding local community structure in networks. [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2005, 72(2):026132.

[9]. Nascimento M C V. Community detection in networks via a spectral heuristic based on the clustering coefficient[J]. Discrete Applied Mathematics, 2014, 176(3):89-99.

[10]. Zhang R, Li L, Bao C, et al. The community detection algorithm based on the node clustering coefficient and the edge clustering coefficient[C]// Intelligent Control & Automation. 2015.

[11]. Dongxiao H E, Liu J, Yang B O, et al. An Ant-Based Algorithm with Local Optimization for Community Detection in Large-Scale Networks[J]. Advances in Complex Systems, 2012, 15(08):1250036.

[12]. M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA 99, 7821-7826 (2002).

[13]. V. Lusseau, K. Schneider, OJ. Boisseau et al. The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations. Behavioral Ecology and Socio-biology, 2003, 54(4):392-405.

[14]. Lusseau D, Newman MEJ. Identifying the role that animals play in their social networks. Proc. of the Royal Society B: Biological Sciences, 2004,271(Suppl_6): S477-S481.

[15]. J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In WSDM, pages 587–596, 2013.

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