International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.8 (2015), pp.207-214
http://dx.doi.org/10.14257/ijmue.2015.10.8.21

Detecting the Hierarchical Community Structure Based on Advanced LPA

Pengtao Jia, Sha Chao and Shuang Qiu

Institute of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China
pengtao.jia@gmail.com

Abstract

LPA is a classical community detection algorithm with linear time complexity. It can be applied to networks whose community structure is known. However, the randomness of it's initial nodes causes a poor stability. In this paper, we proposed a new method to improve the algorithm by reordering the initial nodes. The advanced algorithm uses ascending order instead of descending order based on degree. It can improve the stability of the original algorithm and detect the hierarchical community structure of the complex network. We tested the algorithm on Zachary's karate network, Dolphins social network and American football network. The experiments confirmed that compared to the classical algorithm the improved algorithm has better performance in terms of rationality and accuracy.

Keywords: *Community Detection; Hierarchical Community; Label Propagation; Degree*

1. Introduction

Community detection aims to research the topology structure of the network. The network is an abstract concept, and it can be abstracted from any real-world networks. For example, the nodes represent everyone, and the link between nodes represent the relationship between people, then the network is a social network. The community structure [1-2] has a characteristic: the community internal connection is relatively tight, while the connection between the community is relatively sparse. We usually detect communities according to this characteristic. In recent years, a lot of community detection algorithms have been developed in complex networks. They respectively adopted physics, mathematics, computer science and other fields' theories and technologies [3]. On account of this, they can detect communities respectively based on partition [2], modular optimization [1,4], label propagation [5-8], dynamics [9], bionic calculation method [10] *etc.*. In addition, Fortunato [11] make a detailed introduction about the various community detection algorithms research in his article.

However, the big data information has a high requirement that the time complexity of community detection algorithm should be as low as possible at present in the social network. In 2002, Zhu [12] proposed label propagation algorithm (LPA), which is a semi-supervised learning method based on graph. In 2007, LPA was applied to community detection by Raghava [5] for the first time. LPA has low time complexity, and it is well adapt in large-scale community. The algorithm begins convergent after five iterations. The process of community detection completely relies on the network topology structure, and it neither need to optimize the predefined objective function, nor require any prior knowledge, for example the number and size of the community. These advantages of the algorithm greatly enhance the efficiency of the community detection. However, the randomness of traditional LPA algorithm is its serious drawback, which results in

ISSN: 1975-0080 IJMUE

different partitions at every runs. Such randomness is not worthy to be used. After that, Xie [13] optimized the running speed of LPA. Leung [6] extended LPA by incorporating heuristics like hop attenuation score. Gregory [14] proposed Copra and Xie [15] proposed SLPA which extended LPA to detect the overlapping communities. Unfortunately, none of these extension algorithms can settle the matter of randomness of traditional LPA algorithm. The traditional algorithm randomly sorts each node which has independent label at the initial time, then sequentially update the label of nodes, that is to say: update the label in random order, which results in the increased randomness and more worse stability. In order to solve the problem mentioned above, this paper advanced the traditional LPA algorithm by reordering the initial nodes. In the end, we got more stable and reasonable result for divided communities.

2. The Algorithm Based on Advanced LPA

2.1. The Thought of the Advanced Algorithm

Because of the randomness of the initial nodes, the traditional LPA algorithm division effect is not ideal. In order to improve the performance, we proposed the advanced algorithm of LPA. In this paper, we will start with the degree of the initial nodes in ascending order, and update the label values of these nodes in turn, then continue the traditional LPA algorithm.

We consider simple graphs only, that is to say, the graphs without self-loops or multi-edges.

Definition 1: Assuming that there is a graph $G \overset{def}{=} \{V, E\}$, where $V \overset{def}{=} \{v_1, v_2, \cdots, v_i, \cdots, v_n\}$, $i = 1, 2, 3, \cdots, n$, denotes the sets of the vertices, and $V \neq \phi$. In which, the v_i is the ith vertex, the n is the amount of vertices. $E \overset{def}{=} \{e_1, e_2, \cdots, e_k, \cdots, e_n\}$, $i = 1, 2, 3, \cdots, n$, denotes the sets of the edges. In which, $e_k = (v_i, v_j)$, $i, j = 1, 2, 3, \cdots, n$, is the edge which connects the vertex i and j.

Then, G is an undirected graph, if $(v_j, v_i) \in E$, for $\forall (v_i, v_j) \in E$;

or, G is a directed graph, if $(v_j, v_i) \notin E$, for $\forall (v_i, v_j) \in E$.

Definition 2: There is a matrix $A \overset{def}{=} [a_{ij}]_{n \times n}$, $i, j = 1, 2, 3, \cdots, n$, if A satisfies the following conditions: $\begin{cases} i = j, \ a_{ii} = 0 \\ i \neq j, \ a_{ij} = \begin{cases} 1 & \text{if } v_i \text{ and } v_j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \end{cases}$

Then, A is called the adjacency matrix of graph G.

Definition 3: If $D \overset{def}{=} \{d_1, d_2, \cdots, d_i, \cdots, d_n\}$, where $d_i = d(v_i) = \sum_{j=1}^{n} a_{ij}$.

Then, d is the degree of the nodes.

Theorem 1: If D_{ASC} denotes the degree of the nodes reordered with ascending order,

Then, $D_{ASC} = \left\{ \min\left\{d_{remain(1)}\right\}, \min\left\{d_{remian(2)}\right\}, \cdots, \min\left\{d_{remian(\mathrm{n})}\right\} \right\}$

Where, $\min\left\{d_{remian(1)}\right\} = \min\{d_1, d_2, \cdots, d_n\}$

$$\min\left\{d_{remian(2)}\right\} = \{d_1, d_2, \cdots, d_n\} - \min\left\{d_{remian(1)}\right\}$$

$$\min\left\{d_{remian(3)}\right\} = \{d_1, d_2, \cdots, d_n\} - \min\left\{d_{remian(1)}\right\} - \min\left\{d_{remian(2)}\right\}$$

$$\min\left\{d_{remian(n)}\right\}=\left\{d_1,d_2,\cdots,d_n\right\}-\min\left\{d_{remian(1)}\right\}-\min\left\{d_{remian(2)}\right\}-\min\left\{d_{remian(n-1)}\right\}$$

Theorem 2: For any vertices $v_i, v_j \in G$, if the label $C_i = C_j$, then they belong to the same community.

2.2. The Description of the Advanced Algorithm

Using the idea of label propagation, the advanced algorithm can detect community structure quickly. At the initial moment, each node with an independent label is regarded as a community. Next, the nodes are arranged by ascending order based on their degrees, then the label values of the nodes are iteratively updated in turn. After that, each node changes its label to the one carried by the largest number of its neighbors in every step. Finally, nodes with the same label are classified together after convergence.

The advanced algorithm in this paper can be described as follows:

Input: network $G(V,E)$, which is an undirected graph; initial label array C_V; the sequence p by reordering the label of nodes according to the D_{ASC}.

Output: the updated label array C_V'.

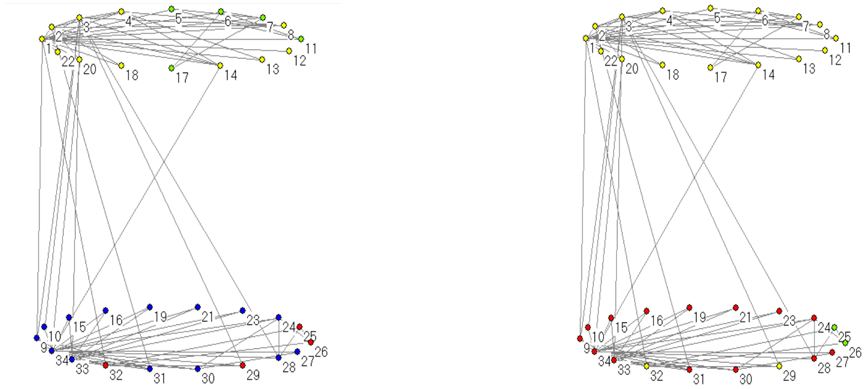
The cycle process:

```
for  i=1: n                  /* n means the number of nodes in the community*/
  (1) Command $v = p(i)$; find the neighbors $v\_Conn$ of node v;
  (2) Obtain the label set $label$ of the neighbors $v\_Conn$.
      Command $label = C_V(G(v,:) == 1)$;
  (3) Select label set $max\_label$ whose frequency of occurrence is the most in $label$;
  (4) If $\mathrm{size}(max\_label)>1$
        Select one of the $max\_label$ randomly;
      else
        Update the label value directly $C_V(v) = max\_label$;
      end if
end for
```

In the field of community detection, when the initial node is the center of community or the center of community was updated first, the results of the algorithm usually tend to be more stable [16]. The center node of community is a node with big degree, which means it has more neighbors, so the node has an important position in the community. Therefore we can obtain a better effect of community detection by giving priority to the node that has the characteristic. However, the situation is opposite when the traditional LPA algorithm improved. The initial nodes are arranged according to the increase of the degree, then update the labels of nodes in turn. This method achieves good results. The reason is that each node has an independent label in the beginning, if the neighbor labels of the current node are same and the number of same label are maximum, then randomly select one of the labels to update the label of the current node. In that way, when we update the node with big degree at first, random selectivity was increased because of the neighbors with different label value. For example, in a data set with 34 nodes, we update the label value of the node 34 that has the maximum degree 17, because the label values of its neighbors are independent of each other (not yet updated), so its occurrences is 1 and the maximum. Therefore we need to choose a label value from 17 neighbors randomly to update the label value of the node 34.

3. Experiments

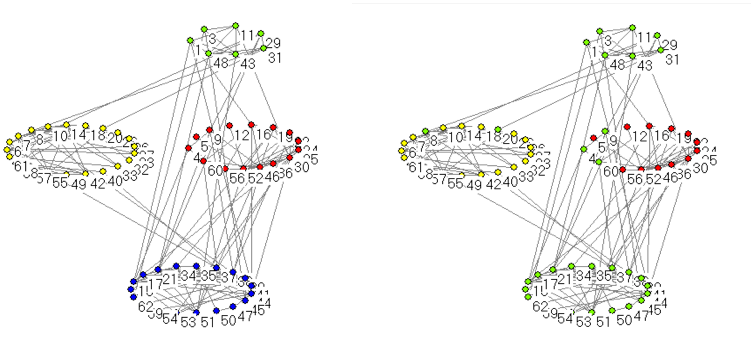
Firstly, the advanced algorithm (LPA-Degree) was tested on Zachary's karate network [17] with 34 nodes and 78 edges. The result is shown in Figure1. In (a), we can see that LPA-Degree found four communities with different hierarchy. Obviously the nodes 5-7, 11, 17 and 25-26, 29, 32 are formed the smaller sub community structure, and they should be integrated as a whole. In (b), the traditional LPA algorithm found three communities. It distinguished nodes 25, 26 as a small community only and meanwhile nodes 29 and 32, which was supposed to belong to the green bottom community, are identified as the member of top yellow community. So our algorithm based on degree achieves the more accurate and reasonable division result.



(a) LPA-Degree (b) LPA

Figure 1. It Shows Community Structure Identified by LPA-Degree and LPA on Zachary's Karate Network. The Community is Identified by Different Colors

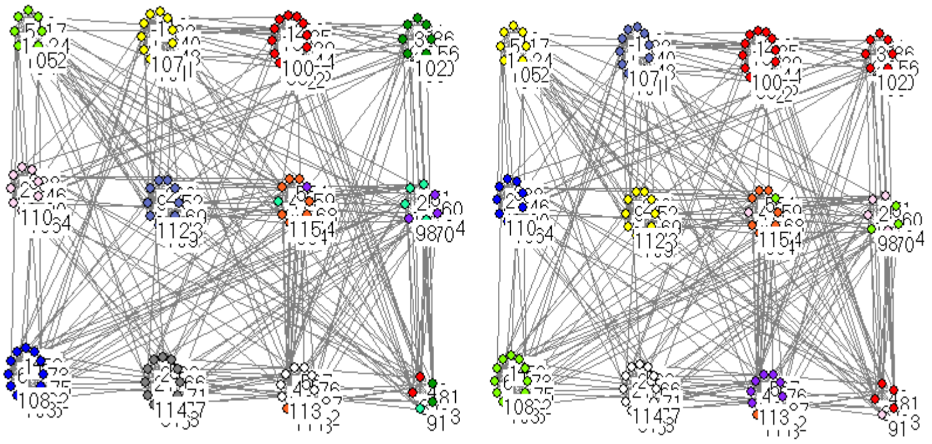
Secondly, the advanced algorithm was tested on Dolphins social network[18] with 62 nodes and 159 edges. The result is shown in Figure2. From this Figure, we can see that LPA-Degree divided the nodes into four communities accurately. The traditional LPA algorithm identified three communities. And it has a grievous mistake that the upper and the lower communities are identified as a community. In addition, there are five nodes were wrongly identified.



(a) LPA-Degree (b) LPA

Figure 2. It Shows Community Structure Identified by LPA-Degree and LPA on Dolphins Social Network. The Community is Identified by Different Colors

Thirdly, the advanced algorithm was tested on American football network [2] with 115 nodes and 616 edges. The result is shown in Figure3. LPA-Degree has a good overall performance, and it can correctly detect all the communities and distributes nodes into the right communities (they actually belong to). Figure4 shows several nodes formed three new communities which have weak links between each other, such as nodes 47, 50, 54, 68, 74, 84, 89, 115, nodes 12, 25, 51, 70, 91, 29, and nodes 59, 60, 98, 64. This phenomenon conformed the characteristic of community structure by analyzing the network topological structure. Besides, Figure5 shows LPA-Degree can detect three communities with smaller sub communities structure, which should be integrated into a whole according to the further network topology structure analysis, and they still satisfy the characteristics of community structure, such as the third community of the first line, the first and the second community of the third line in Figure3(a). However, the traditional LPA algorithm has two mistakes. In Figure3(b), the first yellow community of the first column and the second yellow community of the second column are identified as a community, the third and the four red community of the first line are identified as a community too.



(a) LPA-Degree

(b) LPA

Figure 3. It Shows Community Structure Identified by LPA-Degree and LPA on American Football Network. The Community can be Identified by different Colors

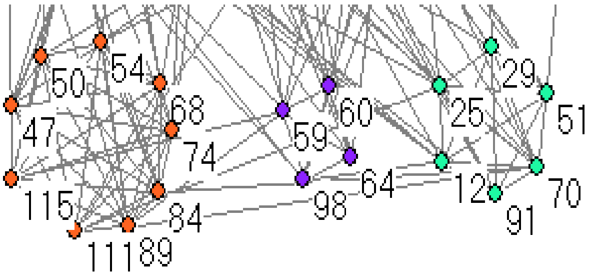
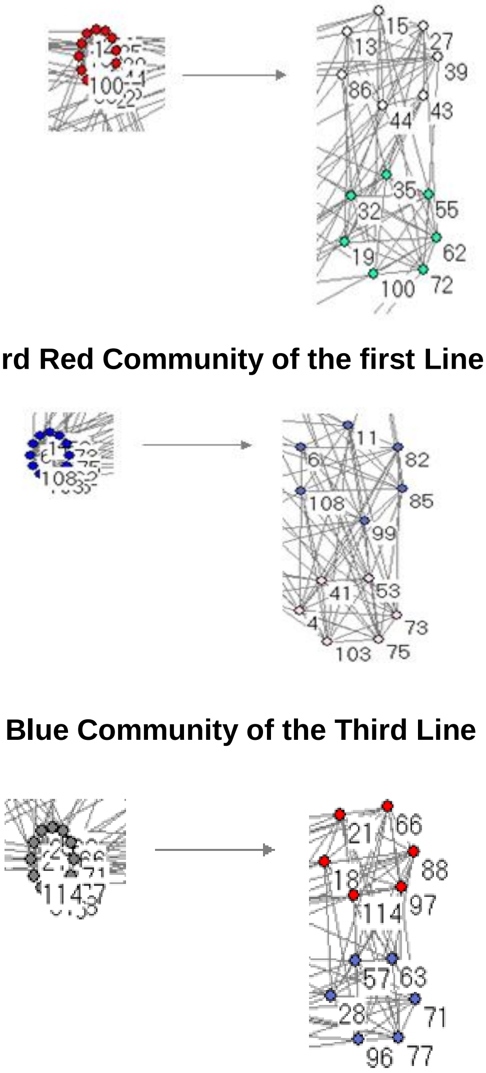


Figure 4. The Figure Shows a More Reasonable Division Results through the Further Analysis of Network Topology Structure by LPA-Degree



(a) The Third Red Community of the first Line in Figure3 (a)

(b) The First Blue Community of the Third Line in Figure3 (a)

(c) The Second Gray Community of the Third Line in Figure3 (a)

Figure 5. The LPA-Degree can Detect Three Communities that have Smaller Sub Communities Structure. It Also Shows the Characteristics of Community Structure that Community Internal Connection is Relatively Tight, and the Connection Between the Communities is Relatively Sparse

Finally, the advanced algorithm was tested on Zachary's karate network, Dolphins social network and American football network in algorithms' running time and accuracy. Because of the certain randomness of both algorithms, each algorithm obtained average value after running the algorithm for 10 times as shown in Table 1 and Table 2. It is well worth sacrificing an average running time of 0.0337s for an average accuracy of 3.87%. The dividing result is more reasonable because of the networks' detected hierarchical community structure. It turned out that LPA-Degree's division performance is more stable and effective.

Table 1. The Average Running Time

	LPA-degree	LPA
Zachary's karate network	0.1833s	0.1465s
Dolphins social network	0.2113s	0.1837s
American football network	0.2017s	0.1651s

Table 2. The Average Accuracy

	LPA-degree	LPA
Zachary's karate network	96.76%	88.82%
Dolphins social network	89.52%	86.29%
American football network	94.26%	93.83%

4. Conclusions

In this paper, we have proposed an advanced LPA algorithm. It breaks the traditional way of thinking and adopts ascending order rather than descending order for the initiate nodes according to the degree. LPA-Degree not only improves the stability of the algorithm but also detect the hierarchical structure of networks. The experimental results show that it is superior to the traditional method for applying in real data sets, and it is more accurate and reliable. Of course, LPA-Degree is not perfect. It has a disadvantage that the discovery of the hierarchical structure is random but not controlled. We look forward to improve the inadequacy by adding control factor in subsequent research.

Acknowledgements

This work is supported by the Scientific Research Program Funded by Shaanxi Provincial Education Department (No. 2013JK0870).

References

[1] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks,” Physical Review E, vol.69 no.22, **(2004)**, pp.1-15.

[2] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences, vol.99 no.12, **(2002)**, pp.7821-7826.

[3] D. Y. Liu, D. Jin, D. X. He, J. Huang, J. N. Yang and B. Yang, “Community mining in complex Networks,” Journal of Computer Research and Development, vol.50 no.10, **(2013)**, pp.2140-2154.

[4] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” Physical Review E, vol.69, no.62, **(2004)**, pp.1-5.

[5] U. N. Raghavan, R. Albert and S. Kumara, “Near linear time algorithm to detect community structures in large scale networks,” Physical Review E, vol.76 no.3, **(2007)**, pp. 106-115.

[6] I. X. Y. Leung, P. Hui, P. Lio and J. Crowcroft, “Towards real-time community detection in large networks,” Physical Review E, vol.79, no.6, **(2009)**, pp.1-10.

[7] M. J. Barber and J. W. Clark, “Detecting network communities by propagating labels under constrains,” Physical Review E, vol.80 no.2, **(2009)**, pp.1-16.

[8] X. Liu and T. Murata, “Advanced modularity-specialized label propagation algorithm for detecting communities in networks,” Physic A, vol.389 no.7, **(2010)**, pp.1493-1500.

[9] V. Dongen and S. Marinus, “Graph clustering by flow simulation,” D. Utrecht, Netherlands: University of Utrecht, **(2000)**.

[10] D. Jin, B. Yang, J. Liu, D. Y. Liu and D. X. He, “Ant colony optimization based on random walk for community detection in complex networks,” Journal of Software, vol.23 no.3, **(2012)**, pp.451-464.

[11] S. Fortunato, “Community detection in graphs,” Physics Reports, vol.486 no.3, **(2010)**, pp.75-174.

[12] X. J. Zhu and Z. Ghanramani, "Learning from labeled and unlabeled data with label propagation," Technical Report CMU-CALD-02-107, Carnegie Mellon University, **(2002)**.

[13] J. Xie and B. K. Szymanski, "Community detection using a neighborhood strength driven label propagation algorithm," In IEEE Network Science Workshop, **(2011)**, pp.188-195.

[14] S. Gregory, "Finding overlapping communities in networks by label propagation," New Journal of Physics, vol.12 no.10, **(2010)**, pp.103-118.

[15] J. Xie and B. K. Szymanski, "Towards linear time overlapping community detection in social networks," In PAKDD, **(2012)**, pp.25-36.

[16] N. Du, B. Wang and B. Wu, "Community detection in complex networks," Journal of Computer Science and Technology, vol.4 no.23, **(2008)**, pp.671-683.

[17] W. W. Zachary, "An information flow model for conflict and fission in small groups," Journal of Anthropological Research, vol.33 no.4, **(1977)**, pp.452-473.

[18] D. Lussesu, "The emergent properties of a dolphin social network," Proceeding of the Royal Society of London. Series B: Biological Sciences, 270. Supply 2, **(2003)**, S186-S188.