IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Instrumentation Mesure Metrologie
Vol. 18, No. 2, April, 2019, pp. 105-109
Journal homepage: http://iieta.org/journals/i2m

An Overlapping Community Detection Algorithm with Label Propagation Control for Complex Networks

Kun Deng*, Li Chen, Wenping Li

College of Mathematics Physics and Information Engineering, Jiaxing University, Jiaxing 314001, China

Corresponding Author Email: dengkun@hrbeu.edu.cn

https://doi.org/10.18280/i2m.180202

Received: 21 January 2019
Accepted: 10 March 2019

Keywords:
complex networks, community detection, label propagation, overlapping communities

ABSTRACT

Aiming at the problem that the accuracy of community detection is unstable and the labels appear vibration in the traditional overlapping community detection methods based on label propagation, this paper proposed OLPC (An Overlapping Community Detection Algorithm with Label Propagation Control for Complex Networks). The algorithm firstly initializes the labels and corresponding storage space for every node in networks. Then setting the number of reserved labels in the label storage space of nodes. And analyzing whether the node needs to continue the update operation in the way of judging whether the reserved labels in the storage space are same. Afterwards, every node receives the most appropriate community label by analyzing the neighbor nodes' conditions. Finally, if the newest community labels received by all nodes are consistent with all nodes' community labels received by previous generation's label propagation, the algorithm stops. Through the testing in benchmark networks, real-world networks and the analysis after comparing the algorithm with some typical algorithms, the experimental results verified the feasibility and validity of the algorithm proposed in this paper.

1. INTRODUCTION

Many complicated systems in real world can be expressed as complex networks, such as interpersonal relations and the cooperation between scientists. These complex networks have common statistical properties like the small-world effect [1], "power laws" in the link distribution [2] and community structure [3], etc. Among them, the community structure reflects an important feature of complex networks: the inner edges of a community are linked tightly but the communities are linked loosely with each other. The process of disclosing the community structure in complex networks is known as community detection. Once detected, the community structure helps to analyze the topology, examine the functions and predict the behavior of complex networks [3]. As a result, community detection has been extensively studied and widely applied in protein function prediction [4, 5], public opinion analysis and control [6] and design of search engine [7], and many other areas. It has become a hot issue in current research.

Recent years have witnessed the emergence of various community detection algorithms, namely, Girvan–Newman (GN) algorithm [8], Fast Newman (FN) algorithm [9], Blondel-Guillaume-Lambiotte-Lefebvre's (BGLL) algorithm [10], and label propagation algorithm (LPA) [11]. The GN is based on splitting, when the FN and BGLL are on modularity optimization, the LPA is on label propagation. Despite their excellent performance, these algorithms can only divide the complex networks into several disconnected communities, that is, each node only belongs to one community. In real-world complex networks, however, communities may overlap with each other, instead of being completely independent. In other words, some nodes in actual networks may belong to several communities (e.g. family, friends, occupation and hobbies) at the same time. This calls for the detection of the overlapping communities in complex networks.

To date, many algorithms have been developed to detect overlapping communities. For example, the clique percolation method (CPM) [12] suggests that only the edges inside a community can be connected to a large complete subgraph. References [13] and [14] propose several community structure detection algorithms similar to the CPM. The link clustering (LC) algorithm [15] assumes that an edge has only one role and belongs to only one community. In other words, the overlapping nodes after confirming the community of each edge must belong to multiple communities. References [16] and [17] provides algorithms similar to the LC. The LFM algorithm [18] determine the structure of all communities through local optimization: initializing several source nodes, optimizing the fitness function, expanding the local communities of each source node. Reference [19] contains several algorithms similar to the LFM. Cao [20] relied on nonnegative matrices to complete community detection. Specifically, the target network was decomposed by the normalized symmetric nonnegative matrix when the number of communities was known, and by Bayesian symmetric nonnegative matrix when the number was unknown. Similar algorithms were presented in References [21, 22]. In addition, the LPA-based algorithms have also been applied to overlapping community detection, thanks to their simplicity and efficiency. Typical examples include the community overlap propagation algorithm (COPRA) and speaker-listener label propagation algorithm (SLPA) [23, 24]. However, the LPA-based algorithms face instable detection accuracy in overlapping community detection, and their efficiency is dampened by the repeated updates of community labels.

To solve the defects of the existing LPAs, this paper puts forward an **o**verlapping community detection algorithm with **l**abel **p**ropagation **c**ontrol for complex networks (OLPC). The

algorithm firstly initializes the labels and their storage space in each network node; Then, the number of labels reserved in the storage space of each node was set up, followed by judging if the reserved labels are the same; if yes, the labels of the node will not be updated; otherwise, the labels will be further updated; next, each node will receive the most suitable label by analyzing the adjacent nodes; finally, the iteration will be terminated if the latest labels of all nodes are consistent with those in the previous generation.

2. ALGORITHM ANALYSIS

2.1 Problem overview

As mentioned before, the traditional LPAs face the problem of instable detection accuracy and their efficiency is dampened because the label of a node can be affected easily by the labels of its adjacent nodes. Taking the SLPA for example, the labels and their storage space are initialized in each network node; in the iterative process, each node sends a label to its adjacent nodes, then receives a label from each adjacent node, and saves the received label randomly in its label storage space; if the labels take up more than *r* percent of the storage space, the labels will be preserved; then, the node will be identified as an overlapping node, if it contains more than one label in its label storage space. Figure 1 shows the possible scenarios of the SLPA in the label propagation. It can be seen that the labels in the storage spaces of nodes b, c, e and f remain unchanged in the last three iterations, and the nodes need to continue to update the labels in the next iteration. Thus, the algorithm becomes less efficient. Meanwhile, the algorithm stability is affected as the storage spaces are updated irregularly for the randomly received labels.

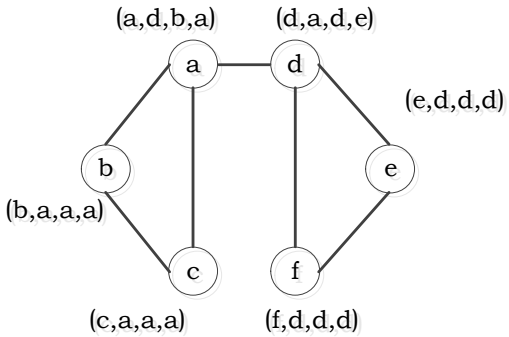


Figure 1. The sketch map of SLPA community detection algorithm

2.2 OLPC algorithm

Considering the defects of the traditional LPAs, this paper sets the control mark of label propagation as a basis that stop updating the node labels which have not changed after multiple iterations, thus enhancing the iteration efficiency. Compared with most LPA-based algorithms of overlapping community detection, the OLPC selects the most suitable label and adds it into each node's storage space by analyzing all labels received from the adjacent nodes rather than received labels randomly. The elimination of randomness makes the operation stable. The detailed steps of the OLPC are as follows:

Step 1: Initialize the label storage space S_v of each node v in the network as $S_v = \{(l_0), cl\}$, where l_0 is the initial label of node v; cl is the control mark of label propagation, i.e. the judgment of the necessity for further label update.

Step 2: Let r be the number of labels reserved in each storage space, which only saves the labels received in the latest r iterations.

Step 3: If there are r identical labels in the storage space S_v of node v, terminate the label propagation; otherwise, go to Step 4.

Step 4: In the t-th iteration, each node v receives the labels from all adjacent nodes, and stores the label l_t with the largest value by formula (1) into the storage space S_v, producing $S_v = \{(l_0, l_1, \ldots, l_t), cl\}$.

$$b_t = \frac{count(x_t)}{|L_t|} \tag{1}$$

where L_t is the set of repetitive elements received in the t-th iteration; x_t is the non-repetitive elements in L_t; t=1, 2,…, N, with N being the number of different values of elements; $count(x_t)$ is the number of elements x_t in L_t.

Step 5: Perform Steps 3 and 4 repeatedly. If the latest label generated from the storage space S_v of any node v is consistent with that in the previous generation, terminate the label propagation.

Step 6: Allocate the nodes with the same label into the same community. If the storage space contains multiple labels, then the node must be an overlapping node.

The above description shows that the OLPC mainly introduces a control mark *cl* to label propagation, and uses it to judge the label variation after multiple iterations. If the label of a node does not change after multiple iterations, the label is very unlikely to change in the subsequent iterations. Thus, if the *cl* indicates no need to go on with the update, the label will not be updated in the subsequent iterations, thus eliminating the label variation in traditional LPAs and improving the propagation efficiency. Moreover, the OLPC ensures that all nodes receive the most suitable nodes, which greatly suppresses the randomness.

2.3 Time complexity analysis

Let G be a network of n nodes with the mean node degree of k, t be the number of iterations of OLPC's label update, and O(n) be the time complexity of the OLPC's initialization of the label of each node. During the label update, the time complexity of label propagation will not exceed O(tkn), because each node needs to receive the labels of its adjacent nodes through the t iterations. Finally, the OLPC's time complexity can be expressed as $O(n+tkn)$. Since k is far smaller than n which is the number of network nodes, the time complexity of our algorithm can also be expressed as $O(jn)$, with j being a constant.

3. EXPERIMENTAL ANALYSIS

To verify its performance, the OLPC was tested with the datasets of benchmark networks and real-world networks, and compared with traditional algorithms like CFinder [12], LFM, COPRA and SLPA. The number of labels r reserved in the storage space of the OLPC was set to 4.

3.1 Evaluation indices

The performance of each algorithm was evaluated against by two classical indices: the community detection accuracy and community closeness.

The community detection accuracy is denoted as normalized mutual information (NMI) [18]. The NMI ranges from 0 to 1. If NMI=1, the detected community structure is exactly the same with the actual structure; If NMI=0, the detected community structure is completely different from the actual structure. The NMI value is positively correlated with an algorithm's accuracy in detecting overlapping communities.

The community compactness is evaluated by the extend Q (*EQ*) [13]. The value of EQ is positively correlated with the closeness between intra-community nodes.

3.2 The datasets of benchmark networks

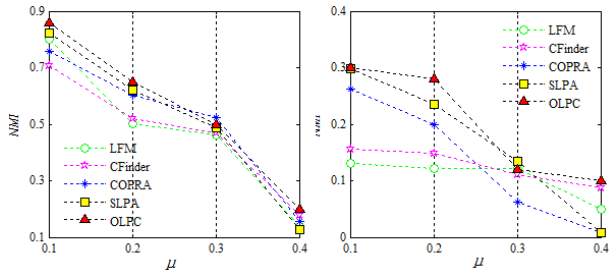
In the LFR benchmark network [25], the node degree and community size obeys power rate distribution, which is similar to the distribution in real complex networks. Thus, this dataset was adopted for our testing and comparison.

The LFR benchmark network contains the following parameters: the total number of network nodes N, the mean node degree k, the maximum node degree k_{max}, the number of nodes in the largest community C_{max}, the number of nodes in the smallest community C_{min}, the mixture proportion μ (the value of μ is positively correlated with the clarity of the community structure), the number of overlapping nodes O_n, and the greatest number of communities than an overlapping node can belong to O_m. Here, the parameters are initialized as: N=200, k=10, k_{max}=30, C_{min}=20 and C_{max}=50. The setting of other parameters is listed in Table 1 below.

Table 1. Parameter setting in LFR benchmark network

Network	O_n	O_m	μ
R1	20	2	0.1~0.4
R2	100	2	0.1~0.4
R3	20	2~6	0.1
R4	20	2~6	0.3

3.3 Comparison of detection accuracy

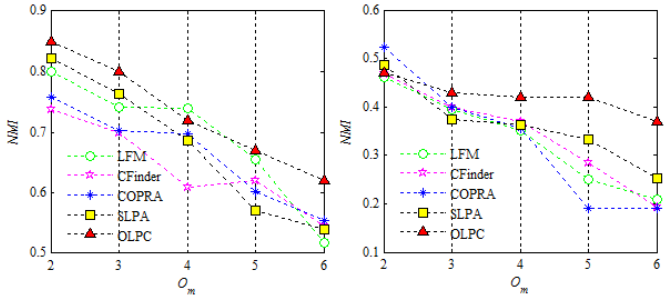


(a)R1 (b)R2

Figure 2. Comparison of detection accuracy at μ=01~0.4

The contrastive algorithms were compared in terms of detection accuracy based on their NMIs in a network with low overlap degree (R1) and in a network with a high overlap degree (R2). As shown in Figure 2, the detection accuracies of CFinder and LFM were not greatly affected by the gradually increasing μ, but were lower than the detection accuracy of the OLPC. Despite their relatively high detection accuracies at the beginning, SLPA and COPRA witnessed a rapid decline in detection accuracy, as the community structure became blurrier, indicating that the two algorithms are not stable. As for the OLPC, its detection accuracy was slightly affected by the increase of μ, but decreased slower than that of SLPA and COPRA. Hence, the OLPC outperformed the other algorithms in detection accuracy.

Next, the detection accuracies of the contrastive algorithms were compared based on their NMIs in a network with clear community structure (R3) and in a network with blurry community structure (R4). As shown in Figure 3, the detection accuracies of CFinder, LFM, COPRA and SLPA all decreased faster than the detection accuracy of the OLPC, with the growing number of communities O_m which the overlapping nodes belong to, although these algorithms had a high detection accuracy at the beginning. The detection accuracy of the OLPC was not greatly affected by the increase of O_m.



(a)R3 (b)R4

Figure 3. Comparison of detection accuracy at O_m=2~6

Overall, the testing on benchmark network datasets shows that the OLPC's detection accuracy was not greatly affected by the increase of O_m, but significantly influenced by the growth of μ. This is because the LPAs update the label of each node based on the labels of the adjacent nodes. In terms of detection accuracy, the LPA are sensitive to the closeness between intra-community nodes. As a result, the SLPA, COPRA and OLPC are greatly affected by the increase of μ, i.e. the fuzzification of community structure. Of course, the comparison also reveals that the OLPC had better detection stability than traditional algorithms like SLPA and COPRA. Despite the vibration of detection accuracy, the OLPC outperformed the other algorithms in detection accuracy with the increase of μ.

3.4 The datasets of real-world networks

Considering the topological difference of real-world networks from benchmark networks, the performance of OLPC was further verified with the datasets of real-world networks. Table 2 shows the datasets of six real-world networks used for the verification, including small networks with dozens of nodes and large network with 10,000 nodes. The community detection quality of each contrastive algorithm was evaluated by the index *EQ*. The comparison results are recorded in Table 3, where "\" means the algorithm failed to detect communities or the *EQ* is lower than 0.001.

It can be seen from Table 3 that the OLPC acquired the largest *EQ* in four networks (Karate, Dolphins, Polbooks and PGP), the second largest *EQ* in Email network (only behind COPRA), the third largest *EQ* in Lesmis network (only after LFM and COPRA algorithms). In general, the OLPC have better community detection ability than the other algorithms in

real-world networks, which promises a great application potential.

Table 2. The datasets of real-world networks

Network	Nodes	Edges	Average degree	Description
Karate	34	78	4.59	Karate club network [26]
Dolphins	62	159	5.13	Dolphin social network [27]
Lesmis	77	254	6.6	The tragic world relations network [28]
Polbooks	105	441	8.4	American political book network [29]
Email	1 133	5 451	9.62	Email communication network [30]
PGP	10 680	24 316	4.55	Trust network [31]

Table 3. The comparison results of each algorithm's value of EQ

EQ	OLPC	CFinder	LFM	COPRA	SLPA
Karate	**0.3543**	0.1072	0.2146	0.3239	0.3472
Dolphins	**0.5041**	0.2885	0.2374	0.4206	0.3879
Lesmis	0.4215	0.1855	**0.4812**	0.4779	0.3209
Polbooks	**0. 4642**	0.4304	0.3476	0.4586	0.4568
Email	0.3055	0.2641	0.1822	**0.3523**	0.1837
PGP	**0.6959**	\	\	0.4335	0.6928

4. CONCLUSION

Aiming at the problem that the results of detection are unstable and labels appear the phenomenon of vibration in the overlapping community detection methods based on traditional label propagation, this paper proposed the OLPC algorithm. This method sets the control mark of label propagation in the process of label propagation. When the number of reserved labels in the label storage space is *r* and these labels are same, meaning that the labels don't need to change in past *r* times iterations. So sets the control mark of label propagation as the mark which forbids updating. In the process of the later updates of labels, this node label will not change, which raising the operating efficiency. Meanwhile, choosing the most appropriate label from the received labels and storing it into the label storage space, which replaces the method of traditional algorithms that randomly choosing the label. Through the testing in benchmark networks and real-world networks, and the analysis after comparing OLPC with multiple comparison algorithms, it can be seen that the OLPC algorithm is viable and efficient.

ACKNOWLEDGMENT

This work is sponsored by the Humanity and Social Science Youth Foundation of Ministry of Education of China (No.17YJCZH033 and No.15YJCZH088), Zhejiang Provincial Natural Science Foundation of China (No.LY15F020040).

REFERENCES

[1] Watts, D.J., Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. Nature, 393(84): 440-442. https://doi.org/10.1038/30918

[2] Barabási, A.L., Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439): 509-512. https://doi.org/10.1126/science.286.5439.509

[3] Jin, D., Liu, D.Y., Yang, B., Liu, J., He, D.X. (2011). Fast complex network clustering algorithm using local detention. Acta Electronica Sinica, 39(11): 2540-2546.

[4] Wang, Z., Zhang, J. (2007). In search of the biological markificance of modular structures in protein networks. Plos Computational Biology, 3(6): 1011-1021. https://doi.org/10.1371/journal.pcbi.0030107

[5] Farutin, V., Robison, K., Lightcap, E. (2006). Edge-count probabilities for the detection of local protein communities and their organization. Proteins Structure Function and Bioinformatics, 62(3): 800-818. https://doi.org/10.1002/prot.20799

[6] Qian, C., Cao, J., Lu, J., Kurths, J. (2011). Adaptive bridge control strategy for opinion evolution on social networks. Chaos: An Interdisciplinary Journal of Nonlinear Science, 21(2): 025116. https://doi.org/10.1063/1.3602220

[7] Sidiropoulos, A., Pallis, G., Katsaros, D., Stamos, K., Vakali, A., Manolopoulos, Y. (2008). Prefetching in content distribution networks via web communities detection and outsourcing. World Wide Web, 11(1): 39-70. https://doi.org/10.1007/s11280-007-0027-8

[8] Newman, M.E.J., Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2): 026113. https://doi.org/10.1103/PhysRevE.69.026113

[9] Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. Physical Review E, 69(6): 066133. https://doi.org/10.1103/PhysRevE.69.066133

[10] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics Theory and Experiment, 10: 10008.

[11] Raghavan, U.N., Albert, R., Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 76(3): 036106. https://doi.org/10.1103/PhysRevE.76.036106

[12] Palla, G., Derényi, I., Farkas, I. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435(7043): 814-818. https://doi.org/10.1038/nature03607

[13] Shen, H.W., Cheng, X.Q., Guo, J.F. (2009). Quantifying and identifying the overlapping community structure in networks. Journal of Statistical Mechanics-Theory and Experiment, 53(7): 07042. https://doi.org/10.1088/1742-5468/2009/07/P07042

[14] Zhang, Z.W., Wang, Z.Y. (2015). Mining overlapping and hierarchical communities in complex networks. Physica A: Statistical Mechanics and its Applications, 421: 25-33. https://doi.org/10.1016/j.physa.2014.11.023

[15] Ahn, Y.Y., Bagrow, J.P., Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. Nature, 466(7307): 761-764. https://doi.org/10.1038/nature09182

[16] Meng, F., Zhang, F., Zhu, M. (2016). Incremental density-based link clustering algorithm for community detection in dynamic networks. Mathematical Problems in Engineering, 2016(6): 1-11. https://doi.org/10.1155/2016/1873504

[17] Kim, P., Kim, S. (2015). Detecting overlapping and hierarchical communities in complex network using interaction-based edge clustering. Physica A: Statistical Mechanics and its Applications, 417: 46-56. https://doi.org/10.1016/j.physa.2014.09.035

[18] Lancichinetti, A., Fortunato, S., Kertesz, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 11(3): 033015. https://doi.org/10.1088/1367-2630/11/3/033015/meta

[19] Wang, M., Yang, S., Wu, L. (2016). Improved community mining method based on LFM and EAGLE. Computer Sliena and Infornaton Systems, 13(2): 515-530. https://doi.org/10.2298/CSIS160217012W

[20] Cao, X., Wang, X., Jin, D. (2014). The (un)supervised detection of overlapping communities as well as hubs and outliers via (Bayesian) NMF. International Conference on World Wide Web Companion, 2014: 233-234. https://doi.org/10.1145/2567948.2577307

[21] Chang, Z.C., Chen, H.C., Huang, R.Y. (2016). Semi-supervised dgnamic community detection bused on non-negative matrix foutorizalion. Journal on Communications, 2(37): 132-142.

[22] He, D., Wang, H., Jin, D. (2016). A model framework for the enhancement of community detection in complex networks. Physica A Statistical Mechanics & Its Applications, 461: 602-612. https://doi.org/10.1016/j.physa.2016.06.033

[23] Gregory, S. (2010). Finding overlapping communities in networks by label propagation. New Journal of Physics, 12(10): 103018. https://doi.org/10.1088/1367-2630/12/10/103018/meta

[24] Xie, J.R., Szymanski, B.K., Liu, X. (2011). Slpa: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. IEEE ICDM Workshop on DMCCI. IEEE, Vancouver, Canada, 2011: 344-349. https://doi.org/10.1109/ICDMW.2011.154

[25] Lancichinetti, A., Fortunato, S., Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. Physical Review E, 78(4): 046110. https://doi.org/10.1103/PhysRevE.78.046110

[26] Zachary, W.W. (1977). An information flow model for conflict and fission in small communitys. Journal of Anthropological Research, 33(4): 452-473. https://doi.org/10.1086/jar.33.4.3629752

[27] Lusseau, D. (2003). The emergent properties of a dolphin social network. Proceedings of the Royal Society B: Biological Sciences, 270(S2): 186-188. https://doi.org/10.1098/rsbl.2003.0057

[28] Knuth, D.E. (1993). The Stanford graphbase: a platform for combinatorial computing [EB/OL].

[29] Newman, M.E.J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Science, 103(23): 8577-8582.

[30] Guimera, R., Danon, L., Diaz-guilera, A. (2003). Self-similar community structure in a network of human interactions. Physical Review E, 68(6): 065103. https://doi.org/10.1103/PhysRevE.68.065103

[31] Boguñá, M., Pastor-satorras, R., Díaz-guilera, A. (2004). Models of social networks based on social distance attachment. Physical Review E, 70(5): 056122. https://doi.org/10.1103/PhysRevE.70.056122