Local-structure-based community detection in firm network

Liang Zhang[1]*, Yuling Liao[2], Xing Cao[1], Guodong Hao[3]

[1]*Business School, Central South University, Changsha 410083, China*

[2]*Department of Legal and Public Administration, Hunan University of Finance and Economics, Changsha 410205, China*

[3]*School of Information Science and Engineering, Central South University, Changsha 410083, China*

Received 1 October 2014, www.cmnt.lv

Abstract

The community detection algorithm based on label propagation could discover latent community structure in a complex network by propagating the label of node between neighboring nodes. Due to the uncertainty and randomness involved in the propagation process, the output community structure is often unstable and lacks precision. The Local-Structure-Based Community detection algorithm propagates the label as an entity instead of propagates the labels of individual nodes, while defying the influence of label on the basis of the local structure in the network. Experiments shows that, applied on complex network data-bases, the algorithm can output high quality community structure, and the output was stable.

Keywords: firm network, community detection, label propagation, local structure

1 Introduction

Complex networks is a complex systems in the form of abstract expression in the real world [1]. Enterprise Network is an application example in the field of economics and management of complex networks. Enterprise network performs as a diagram form, composed of connecting nodes and edges. These nodes represent participants in economic activities like enterprises and organizations. Edges of the enterprise network represent a natural or man-made relationship in accordance with certain rules between participants. And in the perspective of inter-firm knowledge transfer and exchange research, Edges represent the knowledge transfer and exchanged contact between individual participants. Information as a key factor in business decisions, all activities between enterprises can be regarded as the behavior of access to information and knowledge processing. Enterprises learn interaction through the network to improve their ability to obtain information and to help them accelerate the development.

Community structure is an important topology feature of enterprise network. Several community structures compose of a network and there exists a more closely connection between nodes in the same network and relatively sparse one in the different network. Community structure in enterprise network, equivalent to functional units or organization groups which have common characteristics in reality. It is of great significance to study on corporates, such as public opinion analysis and early warning, accurately finding the target groups to create a more suitable business environment and so on.

Because community represents most activities of enterprise nodes in the network, in-depth study on the community can understand the knowledge information and community information of the community, as well as the development of the its organization structure, etc. Now many blog sites and online social networking sites have to rely on advertising to launch commercial revenues. However, not all are suitable for a wide commercials blanket deliver. Those for certain populations are more suitable for delivery within the relevant community, so as to make advertising audience more targeted, and to save resources, which also reflects the commercial value of a community study. The premise to realize these values is to find out the community in social networks. So, how to find out community structure of complex networks is the key issues of current research.

Although community discovery problem has been studied a lot of times, there still exist some fundamental issues which remain unresolved, like how to reduce the required amount of computation has been a research focus. Large number of algorithms have been developed and one of the most efficient and rapid calculation is the label propagation algorithm. The label propagation was first used by Zhu et al. [2] in 2002 as a semi-supervised learning method based on graph to solve the machine learning problem. It uses nodes marked with a label information to predict the label information of unlabelled nodes. It also uses the relationship between samples to build the complete graph model, in which there exists labelled and unlabelled data. Its edges represent similarity between two nodes, the label of the ones transmit to other nodes according to the similarity. As a fast and efficient algorithm, it is used in the community discovery of the literature [3, 4]. It also attracts scholars' wide attention and is widely applied to multimedia information classification, virtual communities mining and other fields.

The label propagation algorithm uses single node to transform its label. When there are several candidate labels

* ***Corresponding author's*** e-mail: zlang1982@sina.com

adjacent to a vertex point, it will randomly select one node. This approach greatly reduces the stability and precision of the algorithm, which makes community structure shock repeatedly or makes transmission happen between them. And it eventually leads to instability of the results. This paper proposes a new label propagation algorithm, which uses similarity of the local structure and makes all nodes with the same label spread as a whole. Actually, it effectively avoids the shock of local structures and to improve the accuracy of the community structure. However compared to this, the old algorithm propagated the most influential labels in a single node which prove to be low-efficiency.

2 Related work

Community structure.

Many scholars put forward the concept of communities in a complex network, but has not yet reached consensus on a uniform and an effective mathematical model of community has not been established either. Literature [7-9] thought that community structure is a structure characteristic in a complex networks between macro and micro. And it is an organization form of network nodes, which makes the connection density within the communities higher than that between communities. Feng [11] presents that community is a collection of a set of nodes with similar properties in the networks. And it also has a feature that a close relationship with the communities and a loose relationship between communities. However in Kang [12]'s eyes, community refers to the set of points which consisted of closely linked nodes. Gong [13] proposes that the community is divided by networks according to the nature of some kind, or certain kinds and it is composed of nodes and edges. Zhang [14] refers the community as the inner polyethylene sub-graph in the networks. There are more internal connections. By the above definition, we can infer that community discovery is a process to put the short distance or a higher degree of similarity of points together. The connection density is higher in the internal community which can be seen as a criterion whether there exists the close proximity of one another or a higher degree of similarity.

The community detection algorithm based on label propagation.

Based on the label propagation algorithm, its main idea is as follows: A given set of vertices V, including K nodes {V1, V2,..., Vk}. Initially each node is given a label, followed by updating the neighboring node' labels which has most labels. After repeated iteration, those with the same label will eventually in the same community.

The algorithm can be summarized as follows:

1) Making the graph structure initialize. Assign a unique label (an integer usually) to each vertex as an identification of the community.

2) All labels are updated. In a single iteration update process, for each vertex V, its label is replaced by the label of all adjacent nodes in the largest number of the label, to update their own labels. If there are multiple labels for the same maximum amount, then randomly select a label as the vertex label. After several iterations, each vertex's neighbored label changes to stabilize.

3) If reaching the number of iterations, then add up the labels and classify the same label as a kind. If not, then turn to step 2.

The label propagation algorithm can form a better community structure in the ideal complex networks within the linear time. But it is worth noting that in the current label propagation algorithm, each vertex is initially assigned a unique label. So, every label has the equivalent propagating probability. However, there will still be two kinds of extreme situations: 1) Scattered and isolated small communities have emerged, forcing some communities in the true sense cannot be generated. The reason is each vertex has a unique and certain label, many vertexes with small influence will adversely affect those with great influence, which results in shocks. 2) The accuracy rate of community recognition is not high, because the nodes' label is easy to spread among different communities, which makes mode of transmission wrong among different nodes.

In conclusion, the reason for that above is there is no use of network-related characteristic, resulting in random spread. So, this paper proposes a new label propagation algorithm base on the local structure. The new one makes the label spread as a whole rather than a single node.

3 The label propagation algorithm based on local structure

The idea of the label propagation algorithm was inspired by the interpersonal phenomenon in our life: in real social networks, if a person's friends are all friends of another person, then the two of them and their friends will form a circle in general. The more mutual friends who appeared in two circles, the more possibility of becoming a community. Therefore, we can define "friends" number between two sets to reflect the impact of the set. In the following we will give several definitions of this set theory:

Definition 1: Influence domain.

Given an arbitrary set of A={V1, V2,... Vk}, its influence domain

SA={V1, V2,…,Vk}∪{N1, N2,...,Nk}

Among them:

Vi is one of the nodes in the network,

Ni to the neighbor node Vi.

This definition, for a community/set, its influence by itself and it consists of connected nodes.

Definition 2: Collection reachability.

Call set A and set B is reachable

$$S_A \cap S_B \neq \Phi$$

Can be written as

$$A \prec B \quad A \succ B$$

This definition implies that a community transmission of information must have mutual connection point.

Definition 3.

Set the influence domain of *A* contains a collection of *B*,

$$A \prec B \wedge S_A \cap S_B = S_A 。$$

Definition 3 points out that a person's friends is another man's friends with definitions in the online community.

Definition 4.

Effect of A on the B collection degree,

$$Infu(A \to B) = \frac{|S_A \cap S_B|}{|S_A|}$$

That, if the set *A* and set *B* more closely linked, the influence between the greater degree of; if a community is not reachable in another community, its influence is 0.

Definition 5.

A set of A for a community, that there is $B \subset A$ and $C \not\subset A$ have founded the Infu (*A*->*B*) < Infu (*A*->*B*).

This illustrates a community to any sub community in the degree of influence is the largest set.

By the above definition, we obtain the following label propagation algorithm based on local structure:

Input: G= (V, E)
Output: node community set C={Ci}.
1) Giving each node a community label Ci and then calculating the impact of each label domain.
2) For each cluster, looking for communities contained in impact within its reach, and attributing.
3) For each existing community labels, finding the most influential labels and attributing according to the label propagation algorithm.
4) Stop once arriving at the community number, or else turn to step 2.

In order to better illustrate the algorithm, we design an experiment to build a small network, as shown in Figure 1. When using the label propagation algorithm, each node's label is initialized and arranged in node number. Here we will combine the algorithm and then list the steps of community detection in this networks.

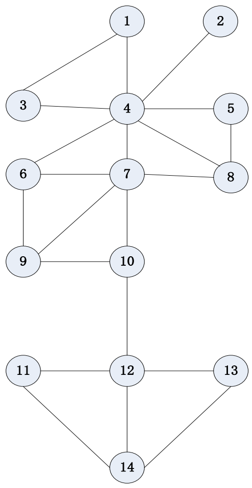

FIGURE 1 simple network examples

First, label every node and the value of the label is the node number. Each influence domain and the reachable node is stored in Table 1.

TABLE 1 The initial node of the domain of influence, reach

Node	Effect of domain S	Reachable node
V1	V1, V3, V4	V1, V2, V3, V4, V5, V6, V7, V8
V2	V2, V4	V1, V2, V3, V4, V5, V6, V7, V8
V3	V1, V3, V4	V1, V2, V3, V4, V5, V6, V7, V8
V4	V1, V2, V3, V4, V5, V6, V7, V8	V1, V2, V3, V4, V5, V6, V7, V8, V9, V10
V5	V2, V4	V1, V2, V3, V4, V5, V6, V7, V8
V6	V4, V6, V7, V9	V1, V2, V3, V4, V5, V6, V7, V8, V9, V10
V7	V2, V4, V6, V7, V9, V10	V1, V2, V3, V4, V5, V6, V7, V8, V9, V10, V12
V8	V4, V5, V7, V8	V1, V2, V3, V4, V5, V6, V7, V8, V9, V10
V9	V6, V7, V9, V10	V4, V6, V7, V8, V9, V10, V12
V10	V7, V9, V10	V4, V6, V7, V8, V9, V10, V11, V12, V13, V14
V11	V11, V12, V14	V10, V11, V12, V13, V14
V12	V10, V11, V12, V13, V14	V7, V9, V10, V11, V12, V13, V14
V13	V12, V13, V14	V10, V11, V12, V13, V14
V14	V11, V12, V13, V14	V10, V11, V12, V13, V14

For example, we can conclude that there exists $S_1 \prec S_3$, $S_1 \prec S_4$ and Infu(1->3)=1, Infu(1->4)=1. Label 1 is contained by label 3 and 4, then we change label 1 to any one of label 3 or 4 according to the algorithm. We change label 1 to label 3 and update label 3's influence domain. Then label 3 's nodes have label 1 and 3 at the same time. However, the label 3's influence domain contain label 4, all nodes label 3 will be updated to label 4. This update is according to labels rather than the nodes, which eliminates the random nature of the original algorithm. After eliminating all the labels contained, we can get the final label in the network diagram (Figure 2).

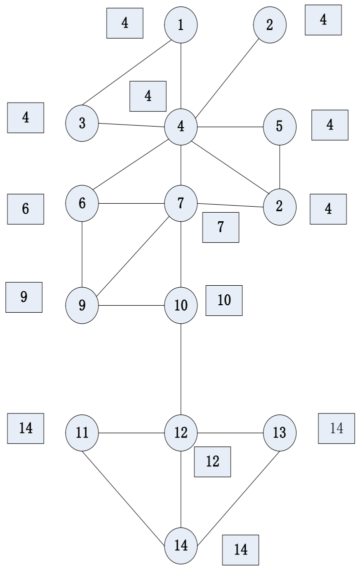

FIGURE 2 The label propagation intermediate results (circles represent the node enterprises; box represents its label)

After a resolution, the rest of the tag value 4, 14, 6, 7, 9, 10, 12. We found the label 7 for tab 4 impact for 0.4.

The 7 is changed to label 4. In this order, until all categories have good (Table 2).

TABLE 2 A label and the influence attributed domain

Label	Up to label	Effect of domain S
4	6 7	1,2,3,4,5,6,7,8
14	12	11 12 13 14
6	4 7	4,6,7，8，9,10
7	4 6 10	1,2,3,4,5,6,7,8
9	6 10	4,8
10	7 9 12	4,7,9
12	10 14	4,6,8,10

4 Experiment

To test the performance of the algorithm proposed in this paper, we use the data samples in the fields of complex networks and conduct a comparing experiment. This experiment includes Karate, Dolphins, Football, Jazz, American Airlines network and Netscience [10]. In addition, we apply the Modular Measurement proposed by Newman and Girvan [14] to compare algorithm. They presented an indicator to evaluate the quality of community.

The Modular Measurement was defined as follows:

$$NMI(C_0,C_e)=\frac{H(C_0)+H(C_e)-H(C_0,C_e)}{\sqrt{H(C_0)\cdot H(C_e)}}$$

C_0 represents the original classified information, $H(C)$ is Shannon entropy. When C_e is exactly equal to C_0, $NMI(C_e,C_0)=1$; when C_e and $C_0$are totally different, $NMI(C_e,C_0)=0$.

The results are shown in Table 3. The *NMI* of label propagation algorithm is improved when compared to the original algorithm, especially the American Airlines network. The *NMI* of the original algorithm is 0, which means the original one has a large rate of errors. And our algorithm can detect the network structure and prove to be effective. As for the community numbers, we can detect more small networks. However, for large-scale scientific papers network, label propagation algorithm based on the local structure detects less. It explains that our algorithm can effectively avoid meaningless online community structure.

TABLE 3

Network	The original NMI algorithm	Based on the local structure	The label algorithm NMI	Community number
Karate club network	0.3684	3	0.38711	4
Dolphin network	0.51345	4	0.64877	5
American University Football League Network	0.58432	12	0.60124	10
Jazz musicians network (Jazz)	0.28221	2	0.34512	4
America aviation network	0.0000	3	0.17323	10

5 Conclusions

Community detection is a hot issue in the complex networks field and it can reveal important structural features of the corporate network. In this paper, the new label propagation algorithm is based on the full study on the local structure of the network. The label spread as a whole rather than a single node, which avoid the instability and improve the quality of the community detection. Experiments show that our algorithm is easier to find out the community with higher quality, when c

6 Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant no. 71371071) and the fund of Hunan Province philosophy social science major project (Grant no. 12WTA45).

References

[1] Fortunato S 2010 Community detection in graphs *Physics Reports* 486(3) 75-174

[2] Zhu X, Ghanramani Z 2002 Learning From Labeled and Unlabeled Data With Label Propagation *Technical Report CMU-CALD-02-107 Carnegie Mellon University*

[3] Raghavan UN, Albert R, Kumara S 2007 Near Linear Time Algorithm To Detect Community Structures In Large-Scale Networks [J] *Physical Review* 76(3)

[4] Leung I X Y, Hui P, Lio P2009 Towards Real-Time Community Detection In Large Networks *Physical Review* E **79**(6) 066107

[5] Zhang J, Chang Y, Wen S 2013 The label propagation algorithm theory and its application research overview of *Application Research of computers* (01) 14

[6] Barber M J, Clark J W 2009 Detecting network communities by propagating labels under constraints *Physical Review E* **80**(2) 026129

[7] Newman M 2003 The Structure And Function Of Complex Net-Works *Siam Review* **45**(2) 167-256

[8] Scheer M 2010 Complex Systems: Foreseeing Tipping Points *Nature* 467(7314) 411-2

[9] Newman M 2010 Networks: An Introduction *New York Ox-Ford University Press*

[10] Newman M 2001 Scientic Collaboration Networks: Net-Work Construction And Fundamental Results *Physical Review* **64**(1) 016131

[11] Feng H, Xue A 2013 The overlapping modularity community outlier detection computer applications and software based on (5) 7-10

[12] Kang X, Jia C 2013 An improved label propagation fast community discovery method *Journal of HeFei University of Technology (Natural Science Edition)* (1) 43-7

[13] Gong S, Chen W, Jia P 2013 Hierarchical clustering community discovery algorithm *Application Research of computers* (11) 3216-20

[14] Newman M, Girvan M 20047 Finding And Evaluating Community Structure In Networks *Physical Review E* **69**(2) 026113

[15] MacDiarmid A G, Epstein A J 1991 in ed. Salaneck W R, Clark D T, Samuelson E J Science and Applications of Conducting Polymers *Adam Hilger Bristol* 117

[16] Eaton D I 1975 Porous glass support material *US Patent No* 3 904

Authors

Liang Zhang, Sep 23th 1982, Shanxi, China.

Current position, grades: PhD candidate.
University studies: Central South University.
Scientific interest: technology innovation, computer science.
Publications: 10 papers.

Xing Cao, Jan 28th 1964, Sichuan, China.

Current position, grades: PhD, professor.
University studies: Central South University.
Scientific interest: technology innovation, technology management, enterprise knowledge management, investment and financing analysis and decision-making areas.
Publications: 60 papers.

Yuling Liao, Sep 15th 1983, Jiangxi, China.

Current position, grades: PhD candidate.
University studies: Central South University.
Scientific interest: management science, behavioral economics, data analysis.
Publications: 10 papers.

Goudong Hao, Oct 14th 1983, Shanxi, China.

Current position, grades: PhD candidate.
University studies: Central South University.
Scientific interest: computer science, machine learning, pattern recognition, image processing.