A Community Identification Algorithm Based on Node Influence Gains

Yong Wang[a], Yunan Hou[b] and Jian Liu [c]

School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China.

[a]wangyong@hrbeu.edu.cn, [b]houyunan@hrbeu.edu.cn, [c]liujian@hrbeu.edu.cn

Abstract

With the rapid development of information technology, the social form of human society presents a trend of networking and gradually forms a social network in which people and relationships are interwoven. Community structure is an important topological feature of social network. It is of great practical significance to discover that the social structure of social network helps to understand the structural features of the network and reveal its inherent functional characteristics. In view of the fact that the previous community discovery algorithms mostly consider the global structure and neglect the attribute characteristics of the node social individuals and the interaction with the community nodes, a community discovery algorithm based on node influence gain is proposed. The algorithm firstly introduces the concept of topological potential of data field theory in physics and then uses it as a criterion to quantify the influence of nodes. Secondly, the random walk probability model based on node influence gain is constructed by using asynchronous thought, and the similarity measure of nodes is introduced. Then, using the hierarchical clustering method to complete the community discovery. Finally, the performance and efficiency of the proposed algorithm are verified through simulation experiments, which shows that the proposed algorithm has high efficiency and feasibility.

Keywords

Social network, topology potential, node influence, influence gains, community identification.

1. Introduction

With the rapid development of social economy and information technology and the popularization of the Internet, the social form of human beings has been gradually transformed into a network. With the social relations among people, organizations and organizations, different socialized network structures have gradually emerged. Network applications such as online shopping networks, social media networks and telecommunications networks have greatly affected people's lifestyles, which has also an important influence on the progress of society and the development of science and technology[1].

The analysis of social network was first proposed by scholar Barnes in 1954, which opened the door to social network research [2]. The purpose of social network analysis is to find out the characteristics of social network structure by studying the differences and changes of relationships in social networks in order to better understand, master and control the characteristics of the actual relationships represented by this social structure [3] . The study of social network structure stability is one of the important contents and main directions of social network analysis. Finding important nodes in the network and identifying the community structure in the network are the main ways to maintain the stability of the social network structure and crack down on cybercrime networks. The community

discovery is to divide the communities of social networks into the same nodes in a community and understand the overall structure of the network from a macro perspective so as to adjust the relationship between communities and grasp the trend of development and changes in communities, which can ensure the stability of the entire network structure [4].

To sum up, social network analysis has been widely applied to fields such as information, economy, military and security. It is of great practical significance and application value to conduct an in-depth study.

2. Related Work

The concept of community discovery in social networks was first proposed by Girvan and Newman. The classic GN algorithm finds out the effective community structure in social networks by removing the community with the largest number of intermediaries in the network. However, the algorithm does not specify the termination conditions, resulting in no specific termination criterion for the algorithm. Later, scholars also proposed a series of improved algorithms. For example, Newman et al. proposed the concept of modularity to measure the quality of network segmentation in social networks and improved GN algorithm to realize community discovery [5].Kerninghan and Lin propose K-L algorithm, which is suitable for dichotomous social networks with known size. The definition of network gain function P and then maximize the gain function P by constantly exchanging nodes between different communities to complete the community structure division. The disadvantage is the need for a priori knowledge and prior knowledge of the size and structure of the community[6]. Pothen et al. proposed a spectral dichotomy algorithm, which calculated its eigenvalues and eigenvectors according to the spectral characteristics of the matrix, and divided the network into two communities according to the sign of the eigenvalue objects [7].Vincent et al. proposed the hierarchical clustering algorithm using the hierarchy existing in the network, and defined the concept of similarity as the standard for node aggregation. Gradually, edges were added to the network according to the order of similarity, and finally the community was divided [8-9].Subsequent researchers also made a great deal of groundbreaking work on community discovery, many algorithms, such as Louvain algorithm, optimal modularity genetic algorithm, Potts-based algorithm, CkC algorithm, CxC algorithm and adaptive MIEN algorithm are proposed [10- 14].

Through the research and analysis, it is found that the above-mentioned community discovery algorithm is more about the division of community structure from the perspective of global topological structure, while neglecting the attribute characteristics of social individuals and the interaction with community nodes. Starting from the concept of community and the definition of topological potential, this paper makes use of the influence of nodes as the standard of community generation, which can objectively and truly describe the structure of social network and has strong practical significance and application value.

3. Community Discovery Algorithm Based on Node Influence Gains

3.1 Community Concept

Definition 1 Radicchi et al. introduced two concepts of community structure: strong community and weak community, the specific idea is as follows [15-16]:

Strong community, in graph G, the following condition is satisfied for any node i in all the subgraphs.

$$K_i^{in}(C) > K_i^{out}(C), \forall i \in C$$

That is, the neighbor nodes of any mode contained in the same community are more than the neighbor nodes belonging to different communities.

Weak community, in Figure G, the following condition is satisfied for any sub-network C.

$$\sum_{i \in C} K_i^{in}(C) > \sum_{i \in C} K_i^{out}(C)$$

Compared with the definition of strong community, a weak community can consider a community as a "node" in the network which meets the definition of strong community.

3.2 Topology Potential and Node Influence

The concept of topology potential is proposed based on the data field theory in cognitive physics. A network can be considered as a physical system that contains several nodes, each representing a field source, and each of them interact with each other. Nodes in the network structure are not only affected by itself but also by the neighboring nodes. This combined influence produces the potential of the node, which can be described by the topology of the node [17].

Definition 2 Topological Potential of Nodes

Assuming the network topology as G (V, E), using the Gaussian function to represent the topological potential φ (vi) at node vi as:

$$\varphi(V_i) = \sum_{j=1}^{n} (m_j \times e^{-\frac{d_{ij}^2}{\sigma}})$$

In the above formula, mj is the quality of node j, which usually reflects some attributes of nodes; dij is the distance from node i to node j, and the shortest path between two nodes is generally chosen as the value of this parameter; σ is the impact factor, which is mainly used to restrain the extent of the topology potential to a certain extent. In this paper, we directly introduce the calculation method of node topological potential in literature [18] and take the topological potential of the node as the node's influence.

Definition 3 Random Walk Model Based on Influence of Node

In random walk model, for any node, the walker will walk to the neighbor node with probability and jump to any other node with the hop number of any with probabilities, and then walk from any node to The walking probability is

In random walk model, for any node u , the walker will walk to the neighbor node of u with probability p , and jump to any other node whose hoop count is t with p with probability p^t ,then the walking probability from any node u_i to u_j can be described as:

$$p_{u_i,u_j} = \frac{E_{u_i,u_j}}{d_{u_i}}$$

Where d_{u_i} is the degree of node, E is the link between nodes. The formula shows that the probability of random walk walking is actually related to the node degree of start node and transit node. At the same time, the influential nodes are more likely to attract each other and increase their probability of arrival due to their higher importance and topological potential. Based on the above assumptions, the random walk model is adjusted, and the node importance factor is added, taking the node influence as the measure.

Definition 4 Node Similarity Measurement Based on Random Walk Model

First of all, we define the following conditions for satisfaction of similarity: the higher the probability of mutual visits between two nodes, the greater the similarity between two nodes. This condition mainly explains two problems: (1) The reachability between nodes cannot be a one-way behavior, and only the two-way walk can satisfy the internal structure of the community ;(2) Only nodes with high walking probability may appear in the same community. Based on the above ideas, a heuristic node similarity measurement function is proposed, which is specifically as follows:

$$sim_{i,j} = \frac{\max(P_{ij}, P_{ji})}{|P_{ij} - P_{ji}|}$$

The numerator of the formula ensures that the nodes in the community have higher accessibility to travel, and the denominator part acts as a penalty factor to reduce the walking probability between nodes with lower mutual reachability.

3.3 Algorithm Ideas

In this paper, we use the method in Section 3.2 to calculate the influence of each node in the network and rank the influence of nodes. We proposes a new node discovery algorithm called NIGCI (node Influence Gains based Community Identification), using node influence as a community generation criterion not only objectively and truly reflect the attribute characteristics of social individuals, but also the identified community structure is considered to have a high degree of cohesion.

The basic idea of the algorithm is that the information flow between nodes is attracted by the nodes with more influence, so as to change the probability of information transmission.At present, a common view of scholars on community structure analysis is that there is a stronger correlation between nodes in the community.Therefore, the higher the efficiency of information dissemination between two nodes, the greater the probability that two nodes are in the same community.Based on the above ideas, this paper proposes a method for node similarity measurement based on random walk of node influence gain, and considers that the higher the mutual reachability of random walk is, the higher the similarity between nodes is.Then a greedy strategy of similar node integration is proposed. Finally, the identification process is completed by selecting the appropriate community structure at the appropriate level.

3.4 Algorithm Description

The community discovery algorithm (NIGCI) based on node influence gain is described as follows:

Input: node topology potential φ (vi), network topology;

Output: Community findings.

(1) Initialize community results $C = C_1, C_2, \ldots, C_n$;

(2) while $|C| \neq 1$ do //| C | represent the number of C in the collection

(3) Select the most similar community $C_a, C_b \in C$

(4) $C_k \leftarrow C_a \bigcup C_b \quad C_k \leftarrow C_a \bigcup C_b$;

(5) $C = C \setminus \{C_a, C_b\} \bigcup C_k$; //$C \setminus \{C_a, C_b\}$represents remaining set except for C_a、 C_b

(6) Use Equation (3-12) to calibrate this variable, and calibrate matrix ranks

(7) end while

(8) $C_{tree} = \{C^1, C^2, \ldots, C^H\}$; //Hierarchical cluster in any layer H set division

(9) $C = \arg\max_{C^i \in C^{tree}} Q(C^i)$; //A certain level set division when getting a maximum

(10) return C

(11) $Q(C^i)$ in the above algorithm is a module function, and the module function proposed by Newman et al. in literature [19] is used as the evaluation criterion. That is, the layer with the largest module function is the final community identification result. The function is defined as:

$$Q = \sum_i (e_{ii} - a_i^2)$$

Where e_{ii} represents the proportion of the inner side of the community, a_i^2 represents the expectations of connecting between the community and the random side.

This section presents a random walk community discovery algorithm based on nodes' influence gain. The algorithm adopts a greedy strategy to merge the communities with the highest similarity in each merge and select the appropriate level for the hierarchical tree generated by node merging , In order

to obtain the optimal community result, so the controllability of the algorithm is stronger, at the same time, it has higher efficiency.

4. Experimental Analysis

This section will demonstrate the performance and efficiency of the algorithm through simulation experiments. The experimental environment is Window 7, CPU: Intel I5 6402P, memory: 8G, programming language: Matlab, Java. The experimental dataset adopts synthetic network, It is more practical to use the LFR Benchmark datum network which has a wider coverage of the LFR synthesis network in line with the power law distribution of the real world. The LFR synthesis network is embedded with the ground-truth community, thus the performance of the proposed algorithm can be verified by comparing the similarity of two community sets. Synthetic networks mainly use NMI (normalized mutual information) normalized mutual information statistics to identify the match between community results and ground-truth communities, where NMI index equals 1, which means that two community sets match exactly, NMI equals 0 and vice versa .

In the aspect of result comparison and analysis, this paper selects Cfinder, GCE and COPRA algorithms, which are considered to have high community recognition performance at present. This section focuses on the performance of synthetic network to verify algorithms. In order to avoid the deviation of experimental results, LFR Benchmark parameters set with the same three algorithms above settings. During the experiment, the degree of ambiguity of the edge of ground-truth community and the impact of network size on each algorithm are determined by adjusting the mixing parameters / mu and the number of nodes N, and the specific results are shown in Figure 1 and Figure 2. (The NIGCI algorithm uses a square Said, Cfinder triangle, COPRA diamond, GCE indicated by the lower triangle).

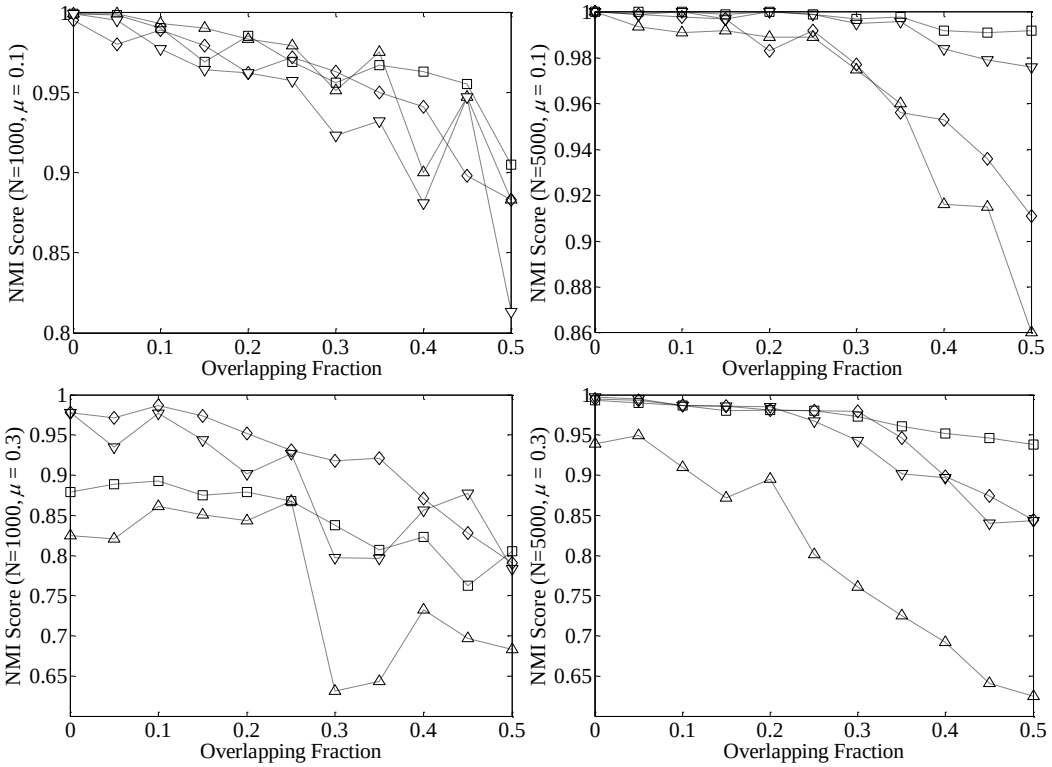

Fig.1 NMI indicator anaysis

It can be seen from the experimental results that the proposed NIGCI algorithm based on node influence gain has the highest NMI index and therefore is closest to the real community. For other algorithms, Cfinder uses community density as a community generation criterion, so the controllability is poor. Although COPRA and GCE also use localization, they ignore the importance of nodes in the network (that is, they ignore the node attraction to neighbor nodes), the resulting

community results are less compact and do not conform to the LFR artificial network generation criteria.

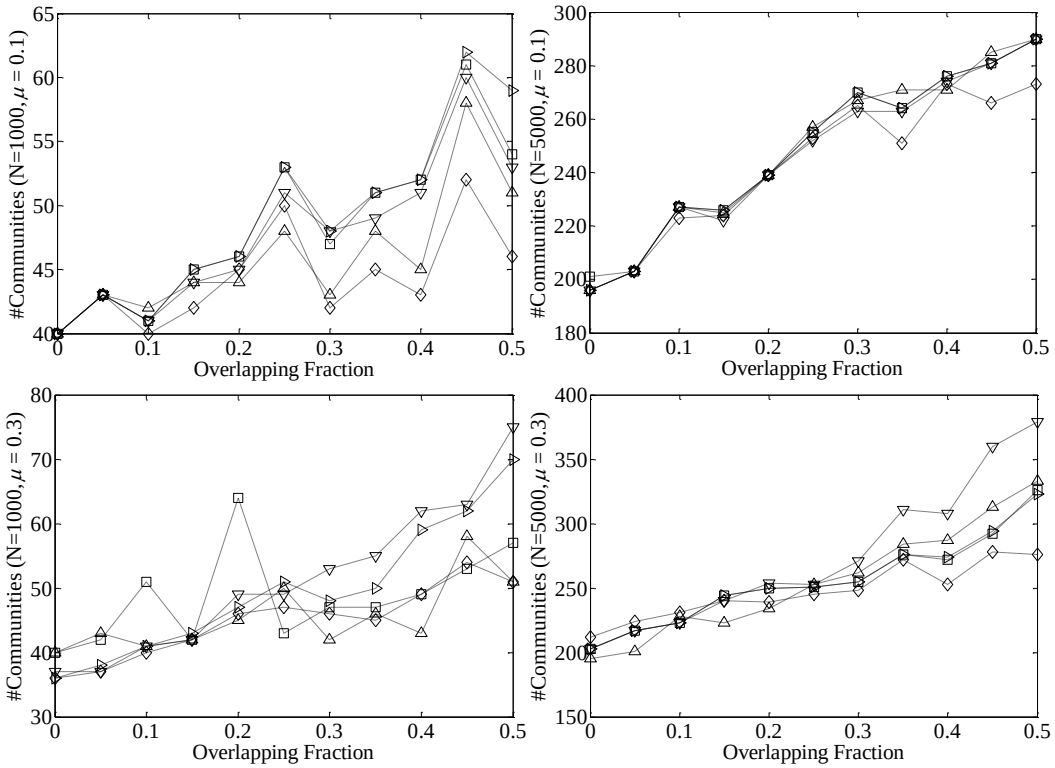

Fig.2 Analysis of the number of network communities

From the above figure 2, the proposed NIGCI algorithm can identify a more reasonable number of communities, with reasonable performance under all parameters. Although the localization algorithm can improve the efficiency of the algorithm, but most can only identify smaller communities. The gain-random walk model used by NIGCI can overcome this problem, mainly due to the increasing role of node importance on the walk model, so that it can abandon some noise nodes on the walk, as shown in the figure experimental results can be clearly reflected.

5. Conclusion

In this paper, for the community discovery problem of social network, we introduces the concept of topological potential of data field theory in physics and proposes NIGCI algorithm based on node influence gains. The algorithm takes the characteristics of the social nodes and their interaction with the community nodes into account, which can describe the structure of the social network more objectively and truthfully. Finally, the simulation results show that the proposed algorithm is effective and feasible, and has certain practical significance for further research on community discovery methods.

Acknowledgements

The research work was supported by The Fundamental Research Funds for the Central Universities under Grant No. HEUCF170604, The Youth Foundation of Heilongjiang Province of China under Grant No. QC2016083, and Innovative Talents Research Special Funds of Harbin Science and Technology Bureau under Grant No. 2016RQQXJ128.

References

[1] Qi Li, Jiancheng Li, Songyu Li.The Logical Structure of Governance Reform in Network Society[J].Chinese Public Administration,2017(7), p.49-54.

[2] Dekker A H. Applying Social Network Analysis Concepts to Military C4ISR Architectures. Connections, the official journal of the International Network for Social Network Analysis, 2001, 24(3), p.93-103.
[3] Freeman L C. Centrality in social networks. Conceptual clarification. Social Networks, 1979, 16(1), p. 215-239.
[4] Haibo Hu, Ke Wang, Ling Xu, Xiaofan Wang.Analysis of Online Social Networks Based on Complex Network Theory.COMPLEX SYSTEMS AND COMPLEXITY SCIENCE, 2008, 5(2), p. 5-6.
[5] Wenyan Gan.Community Discovery Method in Networks Based on Topological Potential.JOURNAL OF SOFTWARE, 2009,44(16), p.236-239.
[6] Youfang Lin, Tianyu Wang, Rui Tang.An Effective Model and Algorithm for Community Detection in Social Networks.Journal of Computer Research and Development, 2012,49(2), p. 337-345.
[7] Jianwei Niu, Bin Dai, Chao Tong, et al.Complex network clustering algorithm based on Jordan-form of Laplace-matrix[J].Journal on Communications, 2014(3), p.11-21.
[8] Liuqiang LI, Xiaolin Gui, Jian An, et al.Overlapping Community Detection Algorithm Based on Fuzzy Hierarchical Clustering in Social Network[J]. Journal of Xi'an Jiaotong University, 2015, 49(2), p.6-13.
[9] Vincent DB, Guillaume JL, Lambiotte R, et al. Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics: Theory and Experiment. 2008(10), p.1-12.
[10] Jinshi Guo, Hongbo Tang, Xiaolei Wang. A Dynamic Community Structure Detection Scheme Based on Social Network Incremental[J]. Journal of Electronics & Information Technology, 2013(9), p.2240-2246.
[11] Xueqi Cheng, Huawei Shen.Community Structure of Complex Networks.COMPLEX SYSTEMS AND COMPLEXITY SCIENCE, 2011,8(1), p.57-70.
[12] Newman M E J. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 2006,74(3), p.36-104.
[13] Dinh T N, Xuan Y, Thai M T. Towards Social-aware Routing in Dynamic Communication Networks[C]. Performance Computing and Communications Conference (IPCCC), 2009 IEEE 28th International, 2009, p.161-168.
[14] Haifeng Du, Shuzhuo LI, Marcus W. Feldman,et al. Detecting Algorithm Based on Prior Knowledge and Modularity for Networked Community Structure[J]. JOURNAL OF XI'AN JIAOTONG UNIVERSITY, 2007, 41(6), p.750-754.
[15] S. Fortunato. Community detection in graphs. Phys. Rep. 2010, 486(3-5), p.75-174.
[16] M.E.J. Newman, M. Girvan. Finding and evaluating community structure in networks. Phys. Rev. E, 2004, 69(2), p.26-113.
[17] Deyi Li, Yi Du.Artificial Intelligence with Uncertainty[M].National Defend Industry Press. 2005, p.193-216.
[18] Wenyan Gan.Community Discovery Method in Networks Based on Topological Potential.JOURNAL OF SOFTWARE.2009,44(16), p.236-239.
[19] Newman M E J. Finding community structure in networks using the eigenvectors of matrices. Physical Review E.2006,74(3), p.36-104.