An Improved Community Detection Algorithm Based on DCT and K-Means

Lin Li[1], Kefeng Fan[1,*], Jiezhong Gong[1] and Hao Peng[2]

[1]Research Center of Information Security, China Electronics Standardization Institute, Beijing, China 10007

[2]Department of Computer Science and Engineering, Zhejiang Normal University, Zhejiang, China 321004

*Corresponding author

***Abstract*—Detecting an overlapping and hierarchical community structure can give a significant insight into structural and functional properties in complex networks. In this paper, we propose an improved algorithm to detect communities in the complex network. The proposed algorithm use discrete cosine transform (DCT) to transfer the topology information into frequency domain, and reduce the dimension of frequency signal with a preliminary threshold, and at last cluster the nodes using K-means. Finally, we apply the proposed algorithm in the real and artificial networks. The simulation results show that the proposed algorithm can avoid curse of dimensionality and it is more accurate than some existing mechanisms.**

Keywords-community structure; DCT; k-means; curse of dimensionality

I. Introduction

Nowadays, more and more people paid attention to complex system in the real world, like online social network, biological system, World-Wide Web, power grids, collaboration networks and so on. In these complex systems, a wide variety of complex systems can be regarded as complex networks[1-3] which are composed of vertices connected together by edges in pairs. For example, the World-Wide Web can be regarded as web sites connected by hyperlink, modeled biological networks reveal interaction among individuals and scientific citation networks uncover collaboration relationship in scientists. The study of these complex networks has last for several centuries and many problems have been solved, but recently the network topological structure has draw a great attention. One of the typical topological properties of complex networks is community structure within which nodes are densely connected and edges connecting different communities are sparser, and many researcher on networks among mathematicians and physicists has focused on this classic topological structure. The Fig.1 is a toy complex network with three typical communities. These topology structure generally provides a significant insight to functional properties in complex network[4-6]. For example, communities in social network might indicate some reasonable organizations, nodes with densely edges in collaboration network might belong to the same team.

Although the definition of community structure is straightforward and easy understanding, the algorithms of detecting these groups of vertices are difficult to construct. In order to detect community structure, many algorithms are created to uncover the feature. There are two typical algorithms which are based on the analysis of eigenvector of Laplacian matrix[7] and the Kernighan–Lin method which iteratively splits a network into two sub-networks according to a greedy algorithm[8]. Recently, Newman and Girvan proposed a seminal algorithm to detect community structure based on the count of shortest-path in each pair of nodes[9]. The proposed algorithm calculated the edge betweenness which is the number of shortest-path passing through each edge, and then split the communities by removing the edge with the maximum betweenness step by step until all communities are detected. To evaluate community structure, Clauset and Newman defined a measurement named Modularity to detect the best partition of the complex network in Ref.[10]. Based on the concept of Modularity, some optimization algorithms are created. These methods use Modularity as the objective function, and proposed some constraint condition from topological feature of communities[11]. To detect community structure in a near linear time, Raghavan et al. proposed a notable algorithm named label propagation algorithm (LPA)[12]. After initializing each node with a unique label, LPA replaces every node's label repeatedly, and then updates the label with the greatest number of nearest neighbors. When all the labels cannot update, the nodes with the same label belong to the same community.

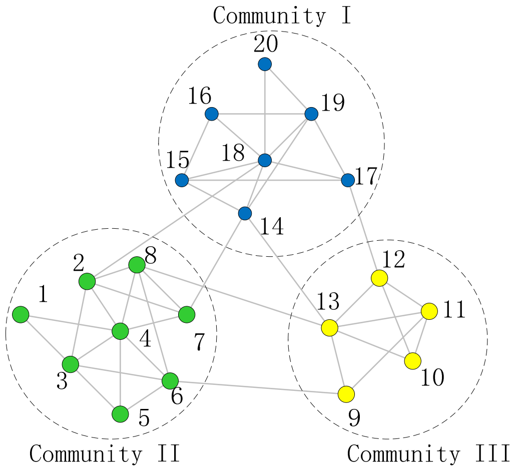

FIGURE I. A SMALL COMPLEX NETWORK WITH THREE TYPICAL COMMUNITIES

Since K-means algorithm is efficient to cluster nodes and fast convergence, some methods use K-means to detect community structure in the complex network. For Instance, Jiang et al. proposed an algorithm based on page rank and K-means to reveal communities[13]. However, many cluster algorithms based on K-means generally exist one fault called curse of dimensionality. To solve this problem, we introduce an improved algorithm to detect communities in the complex network. The proposed algorithm use DCT to transfer the topology information of complex network into frequency domain, and then reduce the dimensionality with a predefined threshold. At last, we use K-means to cluster nodes within the same community. We apply the improved algorithm in real-world and artificial networks, and the simulation results show the effectiveness of the proposed algorithm.

II. Detecting communities with DCT and K-means

A. Reduce Dimensionality with DCT

Some community detection methods transfer the topology feature of the complex network into the coordinate of the Euclidean space. And then use K-means to cluster the nodes within the same group. In general, the complex network contains hundreds of nodes, which indicates the coordinate of the Euclidean space is hundreds of dimensions. According to Ref.[14], a large number of dimension usually leads to the problem of curse of dimensionality, which means large number of dimensions cause all nodes disperse in the Euclidean space. Thus the cluster result is not as good as expected. In this paper, we use DCT into the classic community detection algorithm to avoid the problem of curse of dimensionality and improve the cluster results.

In signal processing, DCT[15] is a transform which can transfer the signal information from time domain to frequency domain, and in the frequency domain, the coefficient shows the energy in each frequency. We notice that the similar signals share the similar coefficient feature after DCT, and the low frequency coefficients own most of the energy of the signal. Thus if we trade the nearest neighbor nodes of each node as the signal, which means each row of the adjacency matrix is the sample of the signal, the DCT can show the coefficient of the signal in the frequency domain. The nodes with similar nearest neighbor nodes share the similar coefficient after DCT. And using this feature, we can cluster the nodes within the same community.

In this paper, we use formula (1) to transfer the topology information into frequency domain.

$$\begin{cases} \mathrm{F}(0)=\frac{1}{\sqrt{N}}\sum_{x=0}^{N-1} f(x) \\ F(\mu)=\sqrt{\frac{2}{N}}\sum_{x=0}^{N-1} f(x)\cos\frac{2(x+1)\mu\pi}{2N} \\ f(x)=\frac{1}{\sqrt{N}}F(0)+\sqrt{\frac{2}{N}}\sum_{\mu=1}^{N-1} F(\mu)\cos\frac{2(x+1)\mu\pi}{2N} \end{cases} \tag{1}$$

where *f(x)* is *N* sample in the time domain, x=0,1,2,...,*N*-1, $F(\mu)$ is the μ *th* coefficient of the DCT, $\mu(\mu=1,2,...,N-1)$ is the parameter of frequency domain, *N* is the number of the nodes.

As forementioned, Since the low frequency coefficients of the DCT cluster most energy of the signal, we can use the preliminary parameter δ to decide percentage of the energy we need, and then the dimension of the cluster can be reduced. Let matrix *A* is the adjacency matrix of the complex network, and *D* is the matrix of A after DCT, $d_{i\cdot}$ is the row vector of *D*, $d_{ij}(i,j=1,2,...N)$ is the element of D. Thus the dimension of *ith* (*i*=1,2,...,*N*) node in cluster algorithm k_i can be calculated with formula (2).

$$k_i=\left\{\mathrm{k_i} \mid \frac{\sum_{k_i=1}^{N} d_{ik_i}}{\sum_{j=1}^{N} d_{ij}} \geq \delta\right\} \tag{2}$$

where δ is the threshold which means the energy percentage of the DCT. For each row of D, we calculate the k_i (*i*=1,2,...,*N)* with formula (2) and then use formula (3) to confirm the dimension of cluster k_{max}.

$$k_{\max}=\{k_1,k_2,...,k_N\} \tag{3}$$

B. Cluster Nodes with K-Means

K-means is a classic cluster algorithm which comes from Expectation-maximization algorithm. This method can cluster nodes within the Euclidean space according to their coordinates into K groups. In this paper, we use K-means to cluster the nodes in the complex network according to the coordinate from DCT and threshold δ . First, the DCT can transfer each row of the adjacency matrix of the complex network, and then use the threshold δ to reduce the dimension in the frequency domain. The reduced dimension can be traded as the coordinate of the node which can be clustered by K-means.

The K-means is sensitive to the initial seeds. In general, K-means start from K initial seed nodes, and each node is put into the group whose center is nearest to the node. In this paper, we use the K-means++ method to choose the K initial seeds as far away from each other as possible[16], and this rule lead K initial seeds into K classes with great probability. Another problem is the value of K. In general, the possible value of K is from 2 to *N*-1 where *N* is the number of the network. In some situation, the number or the scope of K can be given by the user. For example, in the NCAA network[8], each node imply the football team from different state, so the scope of K can be estimated according the number of the state which has football team in the network.

C. Workflow of the Improved Algorithm

According to the forementioned concepts and formulas, we proposed the improved community detection algorithm based on DCT. The detailed description of the introduced algorithm is as follows:

- **Step1**: According to the adjacency matrix *A* of the complex network, remove the isolated nodes and change the *A* into *A′*.
- **Step2**: Let $a'_{i\bullet}$ (i =1,2,...,N) is the *i*th of row the *A′*. For each $a'_{i\bullet}$, use DCT to transfer $a'_{i\bullet}$ into the $d_{i\bullet}$, and $d_{i\bullet}$ (i =1,2,...,N) compose the matrix *D* whose row vectors are $d_{i\bullet}$ (i =1,2,...,N).
- **Step3**: Calculate the k_{max} as the dimension of cluster with formula (2) and (3).
- **Step4**: For each row $d_{i\bullet}$ (i =1,2,...,N) of *D*，keep k_{max} items and let the kept k_{max} items as the elements of row vector $d'_{i\bullet}$, and construct the matrix $D'_{N\times K_{\max}}$ whose row vectors is $d'_{i\bullet}$ (i =1,2,...,N).
- **Step5**: For a given *K*, Trade $d'_{i\bullet}$ (i =1,2,...,N) as *N* coordinates in the Euclidean space and cluster the *N* nodes with *K*-means mentioned in section *B*.
- **Step6**: using the output of K-means $\{C'_1, C'_2, ..., C'_K\}$ to calculate the Modularity value.
- **Step7**: For $K \in [2, N-1]$, repeat **Setp5** and **Step6**, and calculate the maximum Modularity and the corresponding $\{C_1, C_2, ..., C_K\}$
- **Step8**: Output the $\{C_1, C_2, ..., C_K\}$ and let each isolated nodes as a single community.

In this paper, we use Modularity function proposed by Mark Newman et al. to estimate the proper community structure in the complex network.

III. Applications

To validate the performance of the proposed algorithm, we test and verify it to a number of real-world networks with known community structures and artificial networks. These networks include the Karate club network[17], the National Collegiate Athletic Association(NCAA)[8] College-Football network, the classic Girvan-Newman artificial network(GN artificial network)[8].

A. Karate Club Network

The Zachary's karate club network is a typical real network data and it is taken from one of the classic studies of social network analysis. In the early 1970s, Zachary observed social interactions among members of a karate club in an American university for about two years, and described the relationship by a graph. The graph consists of 34 nodes and 78 edges shown in Figure II. In real world, members of the club are separated into two communities because of the dispute between club administrator (node 1) and principal karate teacher (node 33). The two split communities are shown in Figure II by a dotted line.

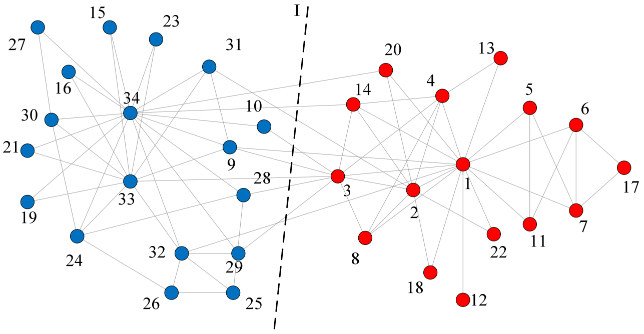

FIGURE II. THE RESULT OF THE ZACHARY'S KARATE NETWORK USING THE PROPOSED METHOD

As is shown in Figure II, the proposed algorithm divides the network into two communities. The blue community represent administrator group while the red one is instructor group. This partition exactly corresponds to the situation in the real-world and many works support this partition. Comparing the partition result with the modularity-based algorithm[8], the proposed algorithm identifies red and blue communities successfully. Modularity-based algorithm also finds community {24,25,26,28,29,32} which is clearly unreasonable, because node 24 more densely connects with the community of instructor.

B. NCAA Football Network

The second real-world network is NCAA football network. In Ref.[8], Girvan and Newman proposed the football network with 115 nodes and 613 edges. All nodes represent college football teams and each edge represents a regular season game between the two teams. The cluster result is shown in Figure III.

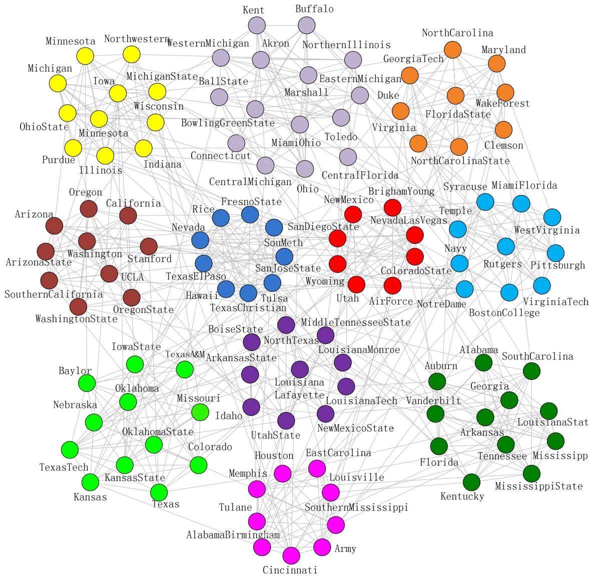

FIGURE III. CLUSTERING RESULT OF NCAA FOOTBALL NETWORK WITH COMMUNITIES ARE SHOWN IN DIFFERENT COLORS

As is shown in Figure III, the NCAA College-football network is divided into 11 communities by the proposed algorithm, and the nodes within the same community are expressed by the same color. Seven of the communities are correctly clustered and they are: Atlantic Coast, Big10, Big12, Pac10, SEC, Conference USA and Mountain West. Total only 11 nodes are included into the incorrect conferences. Further analysis shows that some misclassified nodes are also reasonable for the community definition. For instance, in the real-world, Texas Christian is not the member of the dark blue community. However, we notice that the dark blue community is a complete subgraph which means Texas Christian is a reasonable member of the dark blue community. This eleven-community partition is better than the results got from Newman in Ref.[8] and MSLM, and the cluster results of the three algorithms are shown in Table I.

TABLE I. COMPARISON OF PRECISION OF SEVERAL ALGORITHMS IN NCAA FOOTBALL NETWORK

Algorithm	Community Num	Accuracy(%)
GN[8]	11	78%
MSLM[18]	11	70%
ER[19]	12	88%
Our method	11	90%

C. GN Artificial Network

To further test the performance of the proposed algorithm, in this section, we use artificial network to evaluate the performance of the algorithm proposed here. In [8], Girvan and Newman proposed an artificial network named GN artificial network. The GN network consist of 128 nodes which belong to 4 equal sized communities for each, and each community owns 32 nodes. The average degree <k> of each node is equal to 16 and the edges are generated independently at random between pairs of nodes with probabilities depending on whether the two nodes belong to the same community or not. Each node has k_{in} edges on average connecting with nodes in the same community and kout edges between communities, and k_{in} + k_{out} =16. We use NMI proposed by Danon et al. to evaluate the quality of the two different community partitions[20].

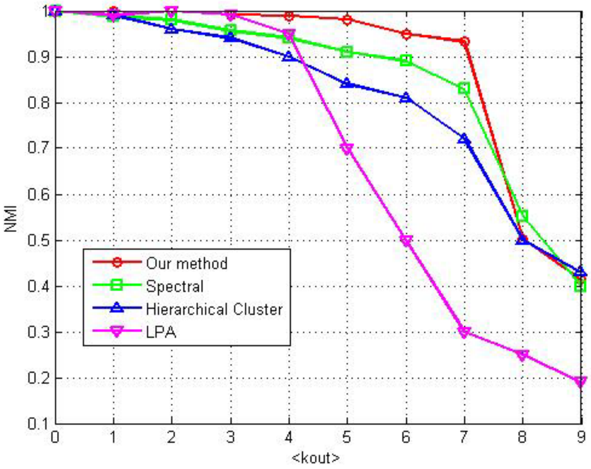

FIGURE IV. THE NMI RESULTS OF FOUR METHOD IN GN ARTIFICIAL NETWORKS

Figure IV shows the community detecting results of the four algorithms: Spectral Partitioning, LPA, and Hierarchical Clustering and the improved method proposed in this paper. According to Figure IV, we notice that the LPA cannot discover reasonable communities when k_{out}>4. When 4<k_{out}<8, the proposed algorithm perform best in the four methods. When k_{out}=8, all the other methods perform similar results. When k_{out} >8, each node has approximate intra-community edges and inter-community edges which leads to the unreasonable community structure, so all the algorithms fail to uncover community structure in the GN artificial network.

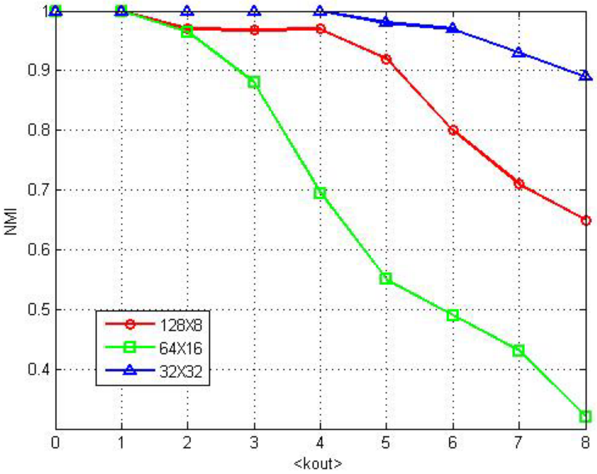

FIGURE V. THE NMI RESULTS IN THREE GN ARTIFICIAL NETWORKS WITH 1024 NODES

Figure V shows the proposed method works on three different GN networks. By changing size of community structure, we can see the performance of the algorithm proposed here in different topologies. First we construct an artificial network with 8 communities, and each community owns 128 nodes. This network contains large communities and connections between communities are densely. The second GN artificial network consists of 16 communities and each community obtain 64 nodes. The third network are divided into 32 small communities and each community consists of 32 nodes. As is shown in Figure V, we notice that the proposed algorithm can detect communities with different size and is more sensitive to edge number between pairs of communities instead of total number of inter-community edge.

IV. CONCLUSION

In this paper, we introduce an improved community detection algorithm based DCT and K-means. First, we use DCT to transfer the topology structure into frequency domain. Since the low frequency coefficients own most of energy, we use a preliminary threshold to reduce the dimension to avoid the curse of dimensionality. At last, K-means is used to identify the communities in the complex network. Applying the proposed algorithm in the real-world and artificial networks, we notice that the proposed algorithm is efficient and robust to detect community structure. In the future, we will focus on reducing the time complexity of the proposed algorithm.

ACKNOWLEDGMENT

Thanks Mark Newman's great help in finishing this paper. This work is supported by 2015 Intelligent Manufacturing Special Project: The Comprehensive Standardized Test, National Key Basic Research Program of China (2013CB329603), National Natural Science Foundation of China under Grant 61602418, Social development project of Zhejiang provincial public technology research under Grant 2016C33168, MOE (Ministry of Education in China) Project of Humanity and Social Science under Grant 15YJCZH125, Key Lab of Information Network Security, Ministry of Public Security under Grant C15610, Zhejiang Provincial Natural Science Foundation of China under Grant LQ16F020002.

REFERENCES

[1] Strogatz S H, Exploring complex networks, Nature 410 (2001) 268.

[2] Watts D J, Strogatz S H. Collective dynamics of "small-world" networks [J]. Nature , 1998, 393(6684): 440-442.

[3] M. E. J. Newman and Tiago P. Peixoto, Generalized communities in networks, Phys. Rev. Lett. 115, 088701 (2015).

[4] Jianshe Wu, Rui Lu, Licheng Jiao, Fang Liu, Xin Yu, Da Wang, and Bo Sun,

[5] Phase transition model for community detection, Physica A 392 (2013) 1287–1301.

[6] M. E. J. Newman, Prediction of highly cited papers, Europhys. Lett. 105, 28002 (2014).

[7] Fiedler M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(2): 298-305.

[8] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs[J]. Bell system technical journal, 1970, 49(2): 291-307.

[9] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113 (2004).

[10] Aaron Clauset, M. E. J. Newman, and Cristopher Moore, Finding community structure in very large networks, Phys. Rev. E 70, 066111 (2004).

[11] Shi C, Yan Z, Cai Y, et al. Multi-objective community detection in complex networks[J]. Applied Soft Computing, 2012, 12(2): 850-859.

[12] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3): 036106.

[13] Jiang Y W, Jia C, Yu J. An efficient community detection method based on rank centrality[J]. Physica A: Statistical Mechanics and its Applications, 2013, 392(9): 2182-2194

[14] Richard Ernest Bellman; Rand Corporation (1957). Dynamic programming. Princeton University Press.

[15] Narasimha, M.; Peterson, A. IEEE Transactions on Communications. 26 (6): 934–936. (June 1978).

[16] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[C]//Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007: 1027-1035.

[17] W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452-473 (1977)

[18] Xie F, Ji M, Zhang Y, Huang D. The detection of community structure in network via an improved spread algorithm [J]. Physica A: Statistical Mechanics and its Applications, 2009, 388: 3268-3272.

[19] Cafieri S, Hansen P, Liberti L. Edge ratio and community structure in networks[J]. Physical Review E, 2010, 81(2): 026105.

[20] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification[J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(09): P09008.