# *An Improved K-means Clustering Algorithm for Complex Networks*

Hao LI , Haoxiang WANG, Zengxian CHEN
School of Computer Science and Engineering, South China University of Technology,
Guangzhou, China
lihorecoco@qq.com

***Abstract*—The exploration about cluster structure in Complex Networks is crucial for analyzing and understanding Complex Networks. K-means algorithm is a widely used clustering algorithm. In this paper, a novel algorithm is proposed based on K-means. Considering, Complex Networks obeys Power-law Degree Distribution, this improved algorithm chooses nodes with high importance as the initial clustering centroids, and uses the distance to these key nodes as clustering measurement. The experiments prove that the new algorithm can conduct accurate clustering with acceptable performance.**

***Keywords—Complex Networks; K-means; Clustering; Node Importance***

## I. INTRODUCTION

Complex network consists of a large number of nodes and a great deal of connections between the nodes. Social network, Internet, power distribution network, scientific collaboration network can be regarded as real models of complex netw1ork. Complex network has the features of Power-law Degree Distributions [1], Network Cluster Structure [2] and other fundamental properties. Power-law Degree Distribution is that, for a random node in this network, the probability of its degree being k is equal to $k^{-C}$ (C is degree exponent). Network Cluster Structure shows us that there are dense connections between the nodes in the same clustering, while connections between the nodes from different clusters are much fewer.

Complex network is hard to be explored and analyzed efficiently in depth for the large amount of information hidden, but the study of clustering of complex network can help filter the information, furthermore, it also helps people understand the structure of complex network better, and even predict the change of complex network better. K-means algorithm has been widely accepted for its simpleness and fast convergence among lots of clustering algorithms. Based on the basic ideas of K-means algorithm, and the feature of Power-law Degree Distributions in Complex Networks, a novel clustering algorithm is proposed in this paper.

The structure of the paper is as follows: first we introduce the principle, limitation and related research of K-means algorithm. Then, we propose a novel algorithm overcoming the limitation of K-means algorithm. Finally we use this new algorithm to cluster some real network data.

This work is sponsored by Fundamental Research Funds for the Central Universities (10561201464), National Training Programs of Innovation and Entrepreneurship (201410561083).

## II. INTRODUCTION TO K-MEANS ALGORITHM

The clustering algorithms can be divided into three categories: optimization algorithms, heuristic algorithms and other algorithms. The first two algorithms are usually complex and time-consuming. K-means algorithm has been widely adopted because of its simpleness and fast convergence.

### *A. The limitations of K-means algorithm*

Although the K-means algorithm is simple and fast, yet there are also some limitations.

The value of the clusters k is implicit. People can only use subjective experience, or experimental results to determine the k value.

Random initialized cluster centroids may lead to inefficient clustering. The times of iterations may greatly increase if improper nodes are chosen as centroids. The choosing of different initial centroids may even cause different results.

Furthermore, K-means needs to use each node' associated vector to measure its Euclidean distance from the cluster centroid, however, not all nodes in real networks can easily find proper measuring scale.

### *B. Related work.*

Current improvements of K-means algorithm are mainly on two aspects: the prediction of the value k and the selection of cluster centroids.

Reference [10] proposes a method of using histogram to optimize the clustering, and determining the value of k by analyzing the distribution of node attribute vectors. Reference [11] explains an approach which can dynamically speculate the value of k based on the density of data. The approach first calculates the distance between sampled data, and then deduces the density of the data.

Reference [12] defines the correlation between nodes. The nodes with minimal correlation values are selected as cluster centroids. Nodes are clustered to the centroids with maximum correlation. Reference [13] uses density-sensitive similarity to measure the density of data, for generating the initial cluter centroids.

## III. IMPROVED K-MEANS ALGORITHM

The improved clustering algorithm presented in this paper leverages the characteristic that Complex networks follow

Power-law Degree Distribution to determine initial cluster centroids.

### A. Choosing of initial clustering centroids

As the node degrees in the complex network follows Power-law Degree Distribution, there exist a number of nodes in the network, whose degrees are significantly higher than the average. Imaging the visualized topology of the network, these nodes with high-degree nodes are usually surrounded by many nodes . In a sense, these nodes can be defined as the relatively important nodes of the network [14].

Those nodes with high degree will naturally play more important role in clustering process. Firstly, high-degree nodes are connected tightly with other nodes around it, which is consistent with the criterion in complex network: the connections inside the cluster are intensive, whereas the connections from internal nodes to external ones are sparse. Secondly, the connection between remaining nodes and the important nodes can be used to cluster these nodes. This judgment is based on that people who share common friends in real social networks trend to be part of the same social circles. Or in the virtual social network, people who subscribe on common celebrities' account, are more likely to be classified in the same group because of common interests.

Based on the analysis above, following assumption are made: suppose there is a network $G(V,E)$ , where V is the set of nodes in the network , and E is the set of edges in the network. The vector can be defined as $D=\{d_0,d_1...d_n\}$ , where $d_n$ is the degree of the node n. Number of P nodes with highest degree are selected as important nodes set $I(i)$.

The value of P satisfies formula (1), where N is the number of nodes in the network, and K is the preset number of clusters.

$$Max(K, N\Big/(\sum_{i=0}^{N-1} d_i)) \tag{1}$$

For each node, the shortest path (defined as distance) to each node in I(i) are calculated, and the shortest distance matrix S can be represented as:

$$S=\begin{pmatrix} s_{00} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{n0} & \cdots & s_{np} \end{pmatrix} \tag{2}$$

In formula (2), $S_{ij}$ represents the shortest distance from node i to important node j, p = P - 1, n = N - 1. For selected important node i, its shortest distance to itself is $s=0$.

### B. Explanation of the algorithm

The basic idea of the algorithm is that P nodes are first found in the network as the set of the most important nodes, and then K highest-degree nodes are selected as the initial cluster centroids. The shortest distance from each node to important nodes is used as the measuring scale of clustering. formula (3) is used to calculate the distance from node i to initial cluster centroid j. Then, this algorithm also follows K-means' principle to iterate the processing until all the cluster stabilize.

$$dist_{ij}=\sqrt{\sum_{i=0}^{P-1}(s_{ix}-s_{jx})^2} \tag{3}$$

The steps of this algorithm can be summarized as follows:

```
Network G(V, E);
K;                              //number of cluster
For each v in V
    calDegree(v);
                    // compute degree for each node
P = keyNodeNum();
            //compute P value using (1)
I[P] = findKeyNodes();
S[N][P] = findShortestDist();
            //compute shortest distance matrix in (2)
KN[K] = initClusterCentroid();
            //initialize cluster centroids
While(!clusterStable()){
    //start iteration
    For each v in V{
            For each u in KN
                    Dist_uv = calDist(v, u);
            //compute distance between u&v using (3)
                    clusterNode(v);
            //add v into the cluster with least Dist_uv
    }
    reCalCentroid(KN);
    //re-calculate cluster centroids using (4)
}
```

In the algorithm above, method *reCalCentroid*() uses formula (4) to recalculate the centroid of the cluster. $C_i^j$ is the distance from the centroid of the i-th cluster to important node j. It is the average of each node's distance to the important node j.

$$C_i^j=(S_{i1j}+S_{i2j}\cdots+S_{inj})\Big/n \tag{4}$$

## IV. Algorithm Verification and Analysis

Using the famous Karate Club network dataset and Dolphin Social Network dataset to test the new algorithm that we have put forward in Section III, we find that the new algorithm not only clusters the two data sets correctly but also has advantage with less iterations. After the proof of correctness of the algorithm, we use Facebook5 concentrated Caltech social network data to analyze the performatnce of the algorithm.

### A. Karate Club Network Experiment

Karate Club dataset describes the relationship among the members of a karate club in a US university. The network includes 34 nodes and 78 edges, where nodes represent the club members and the edges represent the close relationship

between two club members. The club has been divided into two parts due to the charge problem. One is leaded by the coach, while the other's head is the director.

Using the new algorithm to cluster the Karate Club dataset, by setting K=2, we only need two iterations to get a stable clustering, which is shown by Table 1. While the traditional k-means, which randomly selects cluster centroids, takes up to 7 iterations to produce the same result.

Table 1 The clustering result from Karate Club network

| Cluster | Nodes |
|---|---|
| Cluster 1 | 9 10 15 16 19 21 23 24 25 26 27 28 29 30 31 32 33 34 |
| Cluster 2 | 1 2 3 4 5 6 7 8 11 12 13 14 17 18 20 22 |

The clustering result on Karate Club Network topology is shown in Fig.1, which is exactly the same as the real groups in the club.

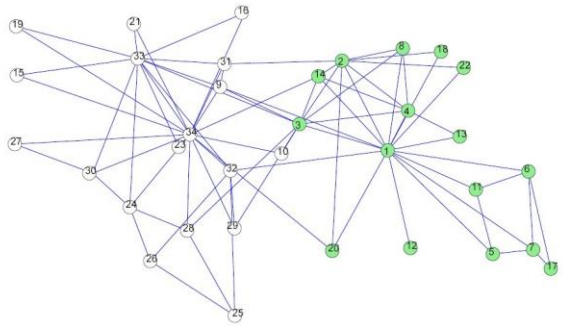

Fig.1 The clustering result on Karate Club Network topology

### B. *Dolphin Social Network Experiment*

Dolphin Social Network describes the relationship of Dolphin Social Network in the ocean near New Zealand. The dolphin network consists of two families, and the dataset includes 62 nodes,159 edges, where nodes represent dolphins and edges represent the close relationship of two certain dolphins.

Using the improved K-means algorithm to cluster the Dolphin Social Network dataset, it also only takes four iterations to get a stable clustering. The clustering result is shown as table 2.While traditional K-means takes up to 12 iterations to produce the same results.

Table 2 The clustering result from Dolphin Social network

| Cluster | Nodes |
|---|---|
| Cluster 1 | 1 3 4 5 8 9 11 12 13 15 16 17 19 21 22 24 25 29 30 31 34 35 36 37 38 39 41 43 44 45 46 47 48 50 51 52 53 54 56 59 60 62 |
| Cluster 2 | 2 6 7 10 14 18 20 23 26 27 28 32 33 40 42 49 55 57 58 61 |

The clustering result on Dolphin Social Network topology is shown in Fig.2, which is the same as real dolphin families division.

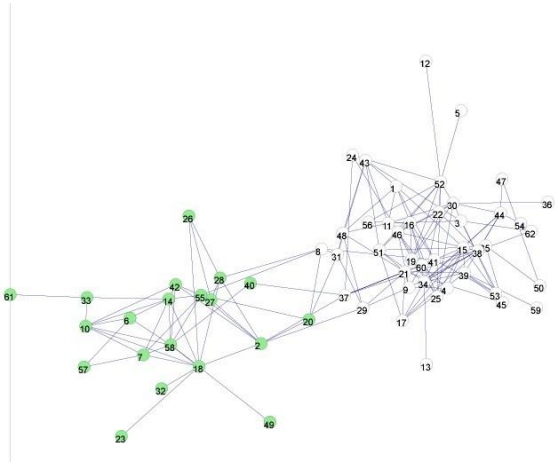

Fig.2 The clustering result on Dolphin Social Network topology

### C. *Facebook5's CalTech Data set experiment*

The environment of experiment is on a notebook with 2.67GHz I3 CPU and 2GB RAM, and the algorithm is completed by JAVA programming language.

We choose the Facebook5's Caltech dataset as the experimental subject. This data set describes the relationship of 769 users of the CalTech on the Facebook online community(including 24257 edges). After several experiments, we find that when we set K=10, the algorithm will not produce empty clustering (if K is set greater than 10,there will produce several clustering that include 0 node).

The algorithm costs 6736ms, and after 20 iterations it can produce a stable clustering.

### D. *Analysis of experimental results*

Karate Club Experiment and Dolphin Social Network Experiment has proved that this improved K-means algorithm can correctly divide the clustering of network with Power-law Degree Distributions. For setting nodes with higher node importance as the centers of clustering, comparing to traditional method which chooses the center of clustering randomly, this improved algorithm avoids incorrect clustering result and improve the performance by reducing iterations.

Facebook5 CalTech dataset experiment improves that this new algorithm has reasonable performance and can be suitable for processing large-scale complex network data.

However, we still find some disadvantages of this algorithm. For example, the algorithm can't reduce the dependence of the value K. But we can predict K with the outcomes of the reference document [10][11] in future work.

On the other hand, it is difficult to verify whether the clustering is correct for complex network data. During the Facebook5 experiment, because the average degree of the nodes of the data set is 63.087 and the data set includes 24257 edges, it is nearly impossible verify the clustering from the network topology. In the future work, the clustering algorithms can be integrated with visualization algorithms (ex. Force-directed Placement algorithm) to achieve better display and analyze of the clustering structure of huge complex network.

## V. CONCLUSION

Here we proposed an assumption that selecting important node as the centroids of clusters and using distance vector between normal nodes and important nodes as the measuring scale of the clustering division. Experimental results show that this improved K-means algorithm can correctly produce the clustering of networks. By reducing the iterations, this improved algorithm also has descent performance.

## *Acknowledgment*

This work was financially supported by Fundamental Research Funds for the Central Universities (10561201464), National Training Programs of Innovation and Entrepreneurship (201410561083).

## *References*

[1] Newman M E J. The structure and function of complex networks [J]. SIAM Review, 2003, 45(2): 167~ 256

[2] Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structures of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814~818.

[3] MacQueen J. Some methods for classification and analysis of multivariate observations [C]. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967:281~297.

[4] Bo YANG, Dayou LIU, Di JIN, Haibing MA. Complex Network Clustering Algorithms[J]. Journal of software, 2009, 20(1):54~66.

[5] Newman M E J. Modularity and communities structure in networks [J]. Proc. of the National Academy of Science, 2006, 103(23):8577~8582.

[6] Shiga M, Takigawa I, Mamitsuka H. A spectral clustering approach to optimally combining numerical vectors with a modular network [C]. In: Berkhin P, Caruana R, Wu X, eds. Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 647~656.

[7] Guimera R, Amaral LAN. Functional cartography of complex metabolic networks [J]. Nature, 2005, 433(7028):895~900.

[8] Girvan M, Newman MEJ. Community structure in social and biological networks [J]. Proc. of the National Academy of Science, 2002, 9(12): 7821~7826.

[9] Wu F, Huberman BA. Finding communities in linear time: A physics approach. European Physical Journal B, 2004, 38(2): 331~338.

[10] Jiangpei ZHANG,Yue YANG,Jing YANG. Algorithm of Initialization of K-Means Clustering Center Based on Optimal-Division [J].Journal of System Simulation, 2009, 21(9): 2586~2590.

[11] TIAN Sen-ping WU Wen-liang, Algorithm of automatic gained parameter value k based on dynamic k-means[J]. Computer engineering and design , 2011, 32(1):274~277

[12] ZHAO Feng-xia, XIE Fu-ding. Detecting community in complex networks using K-means cluster algorithm[J]. APPLICATION RESEARCH OF COMPUTERS, 2009, 26(6): 2041~2044

[13] WANG Zhong, LIU Gui-Quan, CHEN En-Hong. A K-means Algorithm Based on Optimized Initial Center Points[J].PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE, 2009, 22(2): 299~304

[14] Callaway, D.; Newman, J.; Strogatz, S H.; Watts, D J. Network robustness and fragility: percolation on random graphs [J]. Physical Review Letters, 2000, 85(25), 5468~5471