Coarse Graining of Complex Networks: A k-means Clustering Approach

Shuang Xu[1], Pei Wang[1]

1. School of Mathematics and Statistical, Henan University, Kaifeng 475004, China.
E-mail: henuxs@foxmail.com
E-mail: wp0307@126.com

Abstract: Complex networks have been at the forefront of scientific research for more than a decade. A big challenge in complex networks is the share size of the considered systems, especially with the arriving of the era of big data. Coarse graining of complex networks is a possible way to overcome such difficulty. This paper tries to develop a new coarse graining method for complex networks, which is based on the well-known k-means clustering technique. Investigations on some artificial complex networks indicate that the proposed method can significantly reduce the network size and complication, meanwhile, some properties of the considered networks can be preserved to some extent. Moreover, the proposed algorithm allows people to freely choose the sizes of the reduced networks. The associated investigations have potential implications in the analysis and control of large-scale complex networks.

Key Words: Complex Network, Coarse Graining, k-means Clustering, Topological Structure

1 INTRODUCTION

With the arriving of the big data era, the sizes of more and more real-world complex networks become to be beyond our imagination. It's generally difficult to handle these big complex networks, since the storage of these big networks has strong impact on hard disk usage, and the computation of them will be also time-costing [1]. Some algorithms for large-scale networks are computationally prohibitive. For example, the network comprised of people using MSN all over the world is horribly large and could hardly be researched using traditional methods. Therefore, an interesting issue is how to reduce the size of a big complex network. A promising technique is the coarse graining, whose basic idea is as follows. Given a complex network with N nodes and E edges, which is considerably large and hard to be dealt with. The coarse graining technique merges the nodes or edges that share the similar characteristics by certain criteria, while keeping certain properties of the original networks roughly unchanged in the reduced one.

It is noted that, some well-known spectral coarse graining methods have been proposed in the last decade. For example, in the year 2007, Gfeller et al. [2] proposed a spectral coarse graining algorithm, which is based on the eigenvectors of the stochastic matrix that transformed from the adjacency matrix of a network. In the year 2008, Gfeller et al. [3] further used the same technique on the Laplace matrix of the adjacency matrix of a network, the authors merged with similar eigenvector components that corresponding to several of the smallest eigenvalues of the Laplace matrix. The authors reported that the spectral coarse graining technique can well preserve the synchronization ability of some networks, even the size of the original network is remarkably reduced. Following the spectral coarse graining scheme, Chen et al. [4] investigated the effects of the coarse graining process on synchronizability over complex clustered networks. They declared that there is a close correlation between the degree of clustering of the initial network and the ability of spectral coarse graining in preserving the network synchronizability. They found that synchronizability can be well preserved after applying the spectral coarse graining if the considered network has a clear cluster structure, whereas this is not true for networks with vague cluster structure.

The spectral coarse graining scheme can well preserve certain properties of the original network in coarse grained networks. However, the main hurdle of the spectral coarse graining scheme is the computation of eigenvalues and eigenvectors for large-scale complex networks, which limits its real-world applications. Moreover, it is generally difficult to exactly control the size of the reduced network. Motivated by the above problems, in this paper, based on the well-known k-means clustering algorithm, we introduce a new possible coarse graining technique. The algorithm is only based on the k-means clustering of the adjacency matrix, which is computationally simple, and more importantly, the size of the reduced network equals to k, which can be freely chosen. To validate whether the proposed scheme can preserve certain properties of the original network, numerical simulations on several properties during coarse graining process over three types of complex networks are considered, which include the Erdös-Rènyi (ER) network [5], the Newman-Watts (NW) small-world network [6] and the scale-free (SF) network [7]. The rest of the paper is organized as follows. The new algorithm will

This work is supported by National Nature Science Foundation under Grant 61304151, the National Science and Technology Major Project of China under Grants 2014ZX10004-001-014.

978-1-4673-9714-8/16/$31.00 ©2016 IEEE

be proposed in Section 2. In Section 3, several properties of complex networks are investigated during the coarse graining processes. Discussion and concluding remarks will be in the last Section 4.

2 THE NEW COARSE GRAINING ALGORITHM

2.1 The new algorithm

For a complex network with N nodes, the $N \times N$ adjacency matrix $\mathbf{A}$ with elements a_{ij} describes the network wiring weight. For unweighted networks, $a_{ij} = 1$ if node i points to node j and 0 otherwise, and $\mathbf{A}$ is symmetric. Set $L_i = \{j | a_{ij} \neq 0\}$ contains all the nodes that have linkage with node i. We treat matrix $\mathbf{A}$ as a data set with N samples and N variables. In this sense, the i'th row of $\mathbf{A}$ is a N dimensional vector $\mathbf{a}_i = (a_{i1}, a_{i2}, ..., a_{iN})^T$. We treat it as the i'th sample or observation. Define distance function $d(\mathbf{a}_i, \mathbf{a}_j)$ between node i and node j, if $d(\mathbf{a}_i, \mathbf{a}_j)$ is small enough, intuitively, they shall be clustered into the same group. For an extreme example, in a large complex network, node 1 and node 2 are with $\mathbf{a}_1 = \mathbf{a}_2$, which indicates they own equal position, or they are topologically similar for the whole network. Therefore, the two nodes should be merged.

Hereinafter, the coarse graining problem transforms into the clustering problem. Traditionally, the hierarchical clustering method [8] is frequently used in various circumstances. However, the computational complexity of traditional clustering algorithm is $O(N^3)$ or $O(2^N)$, which makes them infeasible for large data sets. Whereas, k-means clustering algorithm [9] is computational simple, which is based on iteration and appropriate for large data situation.

The flow of the k-means clustering algorithm is as follows. Given a set of observations $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)^T$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into $k(\leq n)$ sets $C = \{C_1, C_2, ..., C_k\}$ so as to minimize the within-cluster sum of squares of deviations. In other words, its objective is to find:

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \mu_i) \tag{1}$$

where μ_i denotes the mean of set C_i, $d(\mathbf{x}, \mathbf{y})$ denotes the distance metric between vector $\mathbf{x}$ and $\mathbf{y}$. Since L_1 norm represents the exact amount of the differences between vectors $\mathbf{x}$ and $\mathbf{y}$, it is better to define the distance as L_1 norm, which is also called cityblock distance or Manhattan distance, that is, $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_i |x_i - y_i|$.

If we have divided nodes into k clusters and contracted each cluster into a supernode, the next step is to endow these k supernodes with appropriate topological structure to form the reduced network.

Let supernode S_m represents the m'th clustering set $C_m (1 \leq m \leq k)$. To keep the connectivity of the complex network, S_m should connect with S_n, if in the original graph a node in C_m connects with the other node in C_n. Mathematically, we can define the following set:

$$L_{S_m} = \{S_n | a_{ij} \neq 0, i \in C_m, j \in C_n\}, \tag{2}$$

which contains all the supernodes that the S_m should be connected with. According to the sets $L_{S_1}, \cdots, L_{S_k}$, the reduced complex network is capable of being reconstructed.

As a summary, the steps for the k-means clustering coarse graining algorithm are as follows:

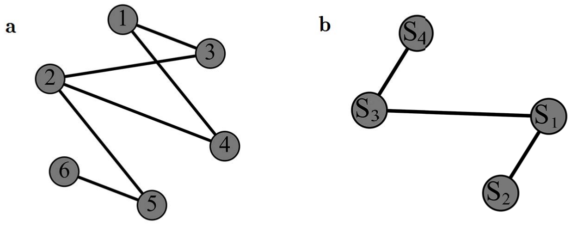
Step 1: Get the adjacency matrix A of the considered network.

Step 2: Set k, perform k-means clustering algorithm to the matrix A, and obtain k clusters $C_i (i = 1, 2, ..., k)$.

Step 3: Merge nodes in each cluster and get k supernodes $S_i (i = 1, 2, ..., k)$.

Step 4: Wire supernodes S_m and S_n if there exists connection between nodes of cluster C_m and C_n in the original network. The reduced network consists of the k supernodes and the connections among them.

2.2 A toy example



$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad \widetilde{\mathbf{A}} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 1: A toy example for k-means coarse graining. (a)A 6-node simple network with adjacency matrix $\mathbf{A}$. (b)The reduced network by k-means coarse graining with adjacency matrix $\widetilde{\mathbf{A}}$.

A 6-node toy example is shown in Fig.1 (a). Here, $L_1 = \{3, 4\}$, $L_2 = \{3, 4, 5\}$, $L_3 = \{1, 2\}$, $L_4 = \{1, 2\}$, $L_5 = \{2, 6\}$,$L_6 = \{5\}$. The distance matrix based on L_1 norm is shown in Tab.1. Since $L_3 = L_4$, the two nodes have totally equal topological roles. Intuitively, to reduce the size of the network, node 3 and 4 should be merged. As to node 1 and 2, both of them connect with node 3 and 4, while node 2 also connect with node 5. They are very similar in topological roles. Hence they can be merged as well to further reduce the size of the network. Now we apply the k-means clustering coarse graining algorithm to the network. The result agrees with our expectation, $C_1 = \{1, 2\}$, $C_2 = \{3, 4\}$, $C_3 = \{5\}$,$C_4 = \{6\}$. According to $L_1 - L_6$, we can calculate $L_{S_1} = \{S_2, S_3\}$, $L_{S_2} = \{S_1\}$, $L_{S_3} = \{S_1, S_4\}$, $L_{S_4} = \{S_3\}$. Based on C_i and $L_{S_i} (i = 1, 2, 3, 4)$, we can create the coarse grained network with adjacency matrix $\widetilde{\mathbf{A}}$, as shown in Fig.1 (b).

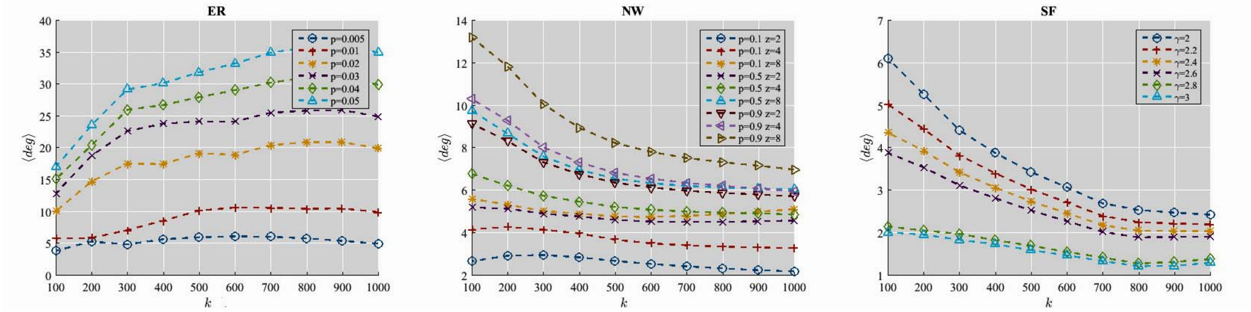


Figure 2: Evolutions of average degrees with the k-means clustering coarse graining processes. The three panels correspond to the three types of artificial networks. Similarly hereinafter.

Table 1: Distance matrix based on the L_1 norm for the 6-node toy network.

Node number	1	2	3	4	5	6
1	0	1	4	4	4	3
2		0	5	5	5	2
3			0	0	2	3
4				0	2	3
5					0	3
6						0

3 PROPERTIES OF THE NEW COARSE GRAINING ALGORITHM

Hereinafter, we investigate several properties of the k-means coarse grained complex networks, including the degree, heterogeneity, assortativity, small-worldness and controllability. For simplicity, we mainly consider the ER random [5] networks with connecting probability p, the NW small-world [6] networks with rewiring probability p and coordinator number z, as well as the SF [7] networks with power-law exponent γ.

The SF networks' degree distributions follow $p(deg) \propto deg^{-\gamma}$, where γ is called the power-law exponent, which typically lies between 2 and 3 for many real-world systems [10]. In this paper, we consider the cases with $\gamma = 2, 2.2, 2.4, 2.6, 2.8, 3$. The NW small-world network algorithm is proposed by Newman and Watts in 1999, which has a rewiring probability p and a coordinator number z. In this paper, we consider the cases with $p = 0.1, 0.5, 0.9$ and $z = 2, 4, 8$. In terms of ER networks, we consider the networks with connecting probabilities $p = 0.005, 0.01, 0.02, 0.03, 0.04, 0.05$, respectively. Additionally, for each type of the artificial complex networks, we fix the size of these networks as $N = 1000$, and for each network, we consider $k = 1000, 900, 800, 700, 600, 500, 400, 300, 200, 100$.

3.1 Average degree

Degree is a fundamental metric that describes the importance of a node in certain extent, and average degree $\langle deg \rangle$ can weigh the relatively connectedness of the whole network.

Fig. 2 shows the average curves of $\langle deg \rangle$ versus network sizes for the three types of networks. Each curve is averaged over 10 independent simulation running. For the ER networks, with the connecting probability p varying from 0.005 to 0.05, the average degrees have great differences. We find that $\langle deg \rangle$ roughly decrease with the decreasing of network sizes (k becomes smaller and smaller), but the decrement is very low, especially for $k \geq 500$. This result indicates the average degree for the ER networks can be well preserved even the size of the original network is reduced to half.

For the NW networks, the curves of $\langle deg \rangle$ roughly slowly increase with the decreasing of network sizes. However, for very low k, the average degrees of the coarse grained networks can roughly keep the same the original ones. This also indicates the average degrees for the NW small-world networks can be well preserved for certain k.

For the SF networks under each γ, the curves for $\langle deg \rangle$ versus k increase with the decreasing of k. Whereas, for $k \geq 800$, there are no much changes on the average degrees between the coarse grained ones and the original ones, and implies the k-means clustering coarse graining algorithms can roughly preserve the average degrees of SF networks for certain k.

3.2 Degree heterogeneity

Degree heterogeneity [11] measures the degree irregularity of a network. All nodes in highly irregular networks have distinct degrees. The more irregular of the degrees values for a network, the more heterogeneous. Heterogeneity can be viewed as a kind of measurement of degree distribution, which is defined as

$$H = \frac{\langle deg^2 \rangle}{\langle deg \rangle^2}. \tag{3}$$

Generally speaking, $H_{ER} < H_{NW} < H_{SF}$. The degree distribution for the ER network is Possion. For the ER networks with high connecting probabilities, its degree distributions are very narrow, and therefore such ER networks with very low H. For the NW networks, they have very low H with very high rewiring probabilities, and they have low H under high coordinator numbers. For the SF networks, with the increasing of γ, the degree irregularity is also increased. Therefore, the SF networks have lower H under higher power-law exponent γ.

The evolutions of network heterogeneity with the k-means clustering coarse graining processes for the three types of networks are shown in Fig.3. In this figure, the normal-

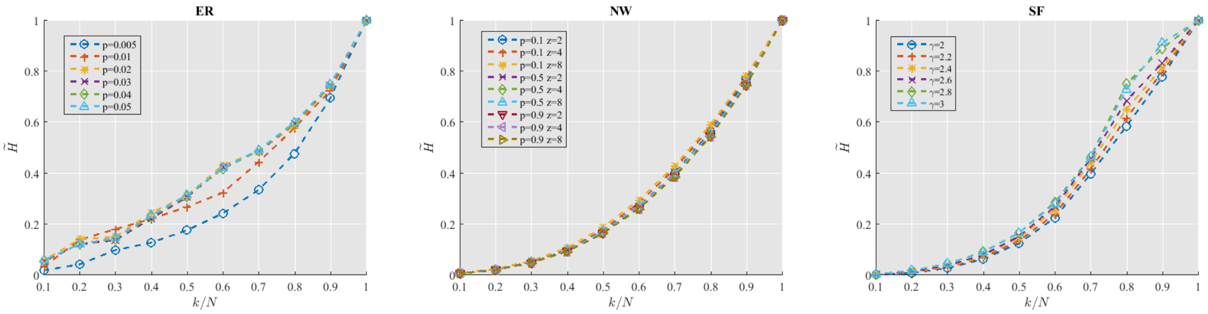


Figure 3: Evolutions of network heterogeneity with the k-means clustering coarse graining processes.

ized heterogeneity $\widetilde{H}$ index is used to unify scales, which is normalized through being divided by its first elements. For all types of networks, $\widetilde{H}$ firstly sharply decreases and then slowly decreases with the decreasing of k/N, which indicates the degree heterogeneity can not be preserved according to the k-means clustering coarse graining algorithm. Another interesting finding is beyond our expectation, it seems that $\widetilde{H}$ is weakly correlated with network parameters, such as the connecting probability, the coordinator number, the rewiring probability and the power-law exponent. This phenomenon seems remarkably true for the NW networks, the curves under different parameters almost coincide with each other. For the ER networks, the curves under $p = 0.005, 0.01$ have subtle differences from the other curves, we suspect that this regular pattern emerges when $p \geq 0.02$ for the ER networks.

3.3 Assortativity

Assortativity [12, 13] describes the phenomenon that nodes with certain degree tend to be connected to other nodes with similar degree. Assortativity coefficient is defined as

$$r = \frac{M^{-1}\sum_i j_i l_i - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - [M^{-1}\sum_i \frac{1}{2}(j_i + k_i)]^2}. \quad (4)$$

where j_i and l_i are the degree of two terminal nodes of the i'th edge, and M denotes the edge number. Generally, the assortativity coefficient r ranges between -1 and 1. If $r = 1$, the network is completely assortative, if $r = -1$, the network is completely disassortative, and if $r = 0$, the network is non-assortative. The theoretical values of r is 0 for both the ER and SF networks [13]. However, $r > 0$ for the NW networks.

For the ER and the NW small-world networks, the curves decrease with the decreasing of k, and different curves almost coincide with each other. But for the curves with parameters $p = 0.005, 0.01$ for the ER networks and $p = 0.1, z = 2\&p = 0.1, z = 4$ for the NW networks, there are a little differences. Moreover, curves for the ER networks are approximately linear, however, curves for the NW networks decrease sharply at first, and then slowly. For the SF networks, the curves tend to increase firstly and then decrease with the decreasing of k. The observations indicate that the k-means coarse graining tends to reduce an assortative or a non-assortative complex network into a disassortative one, where nodes with higher degrees tend to connect with nodes that are with lower degrees. Furthermore, the results in Fig.4 also indicate the assortative mixing property can not be preserved during the k-means clustering coarse graining processes.

3.4 Clustering Coefficient and Average Path Length

The clustering coefficient [14] measures the density of triangles in a network, defined as the ratio of the triangle number to the connected triples number. The average path length L is another useful network metric, which is defined as the mean shortest distances among all pairs of nodes. That is,

$$L = \frac{2}{N(N+1)} \sum_{i \geq j} d_{ij}. \quad (5)$$

where N is network size, d_{ij} denotes the length of the shortest path between nodes i and j.

The ER networks are with small clustering coefficients and small average path lengths. However, the small-world networks are typically with high clustering coefficients and short average path lengths. Fig.5 shows the evolutions of average path lengths and clustering coefficients with the k-means clustering coarse graining processes. From Fig.5, the clustering coefficients for almost all networks increase significantly, and L decrease with the descending of k, except for the SF networks with $\gamma \geq 2.8$. The results indicate that k-means clustering coarse graining can make the reduced complex networks more small-world; and also indicate the small-world property can be preserved. Specially, the clustering coefficients for the ER networks tend to be independent of the connecting probability p. Interestingly, L converges at a certain point with the decreasing of k, which holds regardless of the types or parameters of the complex networks.

3.5 Structural Controllability

A system is controllable if it can be driven from any initial state to any desired final state in finite time. In the years 2011 and 2013, Liu et al.[15] and Yuan et al. [16] investigated the structural controllability for directed complex networks and undirected complex networks, respectively. Liu et al. developed some analytical tools to study the controllability of an arbitrary complex directed network. The method can well identify the set of driver nodes with time-dependent control that can guide the system's entire dynamics [15]. In the above two works, the following canoni-

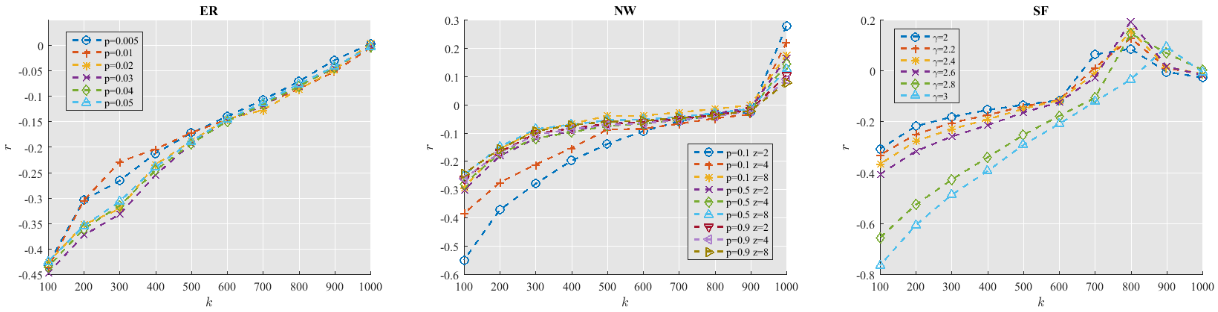


Figure 4: Evolutions of network assortativity with the k-means clustering coarse graining processes.

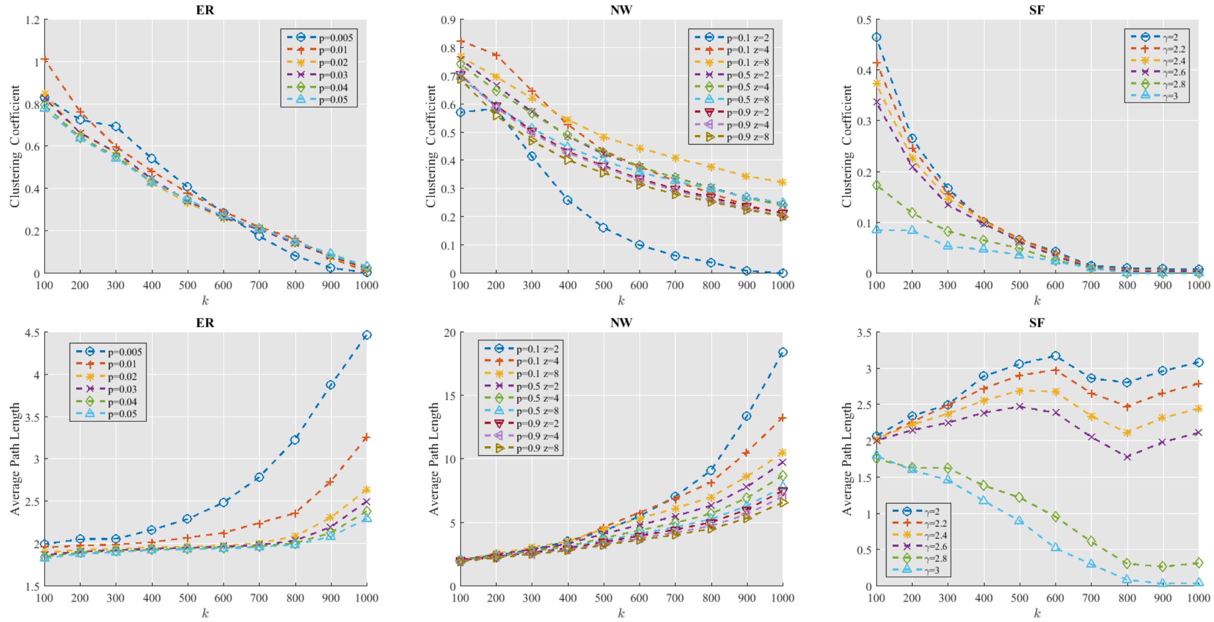


Figure 5: Evolutions of the average path lengths and the clustering coefficients with the k-means clustering coarse graining processes.

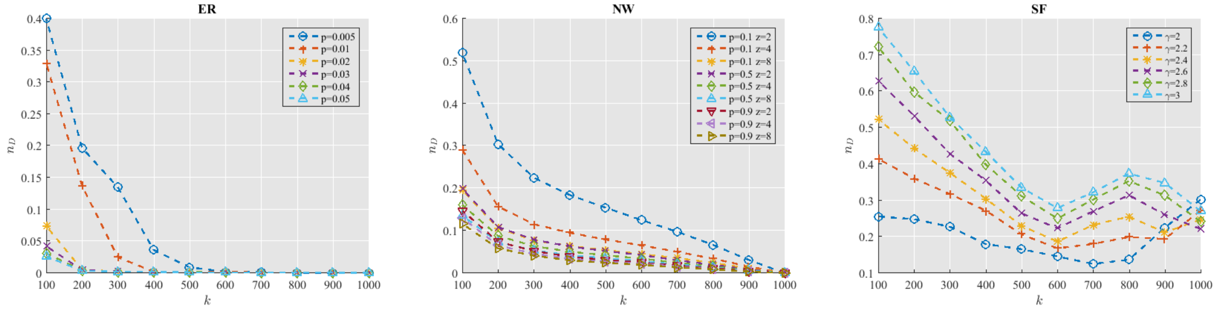


Figure 6: Evolutions of the structural controllability with the k-means clustering coarse graining processes.

cal linear, time-invariant dynamics are considered in a complex network.

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad (6)$$

where $\mathbf{x}(t) = (x_1(t), \cdots, x_N(t))^T$ captures the state of a system with N nodes at time t. **A** denotes the adjacency matrix of the system, and **B** identifies the nodes controlled by outside controllers. The system is controlled using the time-dependent input vector $\mathbf{u}(t) = (u_1(t), \cdots, u_M(t))^T$ imposed by the controller. The system described by formula 6 is called controllable if and only if the $N \times NM$ controllability matrix $C = (B, AB, A^2B, \cdots, A^{N-1}B)$ has full rank $rank(C) = N$. Actually, the rank method is computationally feasible. Fortunately, Liu et al.[15] introduced the maximum matching algorithm and vastly simplified the structural controllability problem. Yuan et al. built the controllability framework for undirected complex networks. It's proved that the minimum number N_D of controllers or drivers is determined by the maximum algebraic multiplicity for undirected networks with arbitrary link weights[16]. Hereinafter, the used controllability metric n_D is defined as the ratio of N_D to the network size N. That is, $n_D = N_D/N$.

The evolutions of the structural controllability are shown in Fig.6. The coarse grained controllability property varies greatly among different complex network types. For the ER networks, there is a threshold value of k for each network, when k is larger than the threshold value, the structural controllability property can be preserved. Whereas, when k is smaller than the threshold value, one needs to control more and more nodes with the decreasing of k. For most of the k values, the n_D roughly slowly linearly increases with k for all the NW small-world networks, and indicates the controllability property for the NW small-world networks can also roughly be preserved under some k values. For the SF networks, the curves firstly increase, then descend during k decreasing and finally strongly linearly increase under small k. For $k \geq 500$, the n_D values for different networks fluctuate around 0.3, and also indicates the controllability property of the coarse grained SF networks can be preserved under some circumstances.

4 CONCLUSION

The giant sizes of large-scale complex networks hinder our understanding of them. The coarse graining techniques are promising ways to reduce the complicity of large-scale complex networks. Although many coarse graining methods have been proposed, seldom algorithms allow freely choosing the final sizes of the reduced networks. Moreover, many algorithms are computational difficult at the present stage. In this paper, we have developed a new algorithm to reduce the sizes of complex networks, which is based on the k-means clustering. The size of the reduced complex network is exactly k and the method is computationally feasible. Several properties of the networks are considered during the coarse graining processes, and we find some properties can be preserved during the processes, while many other properties can not be preserved. For example, the average degrees for the ER and NW small-world networks can be well preserved for certain k. The related investigations have potential implications in future analyze and control of large-scale complex networks.

In this paper, only undirected and unweighted complex networks are considered, our future works will consider the cases for directed and weighted complex networks. Furthermore, except the network metrics considered in the above section, one can consider many more other metrics, such as the motif centrality [10]. Another interesting question that deserves to be further investigated is to design specific coarse graining algorithms to preserve certain network property.

REFERENCES

[1] J. Fan, F. Han and H. Liu, Challenges of big data analysis, Nat. Sci. Rev., Vol.1, 293-314, 2014.

[2] D. Gfeller and P.D.L. Rios, Spectral coarse graining of complex networks, Phys. Rev. Lett., Vol.99, No.3, 038701, 2007.

[3] D. Gfeller and P.D.L. Rios, Spectral coarse graining and synchronization in oscillators networks, Phy. Rev. Lett., Vol.100, No.17, 174104, 2008.

[4] J. Chen, J.A. Lu, X.F. Lu, X.Q. Wu and G.R. Chen, Spectral coarse graining of complex clustered networks, Commun. Nonlinear Sci. Numer. Simulat., Vol.18, No.11, 3036-3045, 2013.

[5] P. Erdös and A. Rènyi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci., Vol.5, 17-61, 1960.

[6] M.E.J. Newman and D.J. Watts, Renormalization group analysis of the small-world network model, Phys. Lett. A, Vol.263, No.4, 341-346, 1999.

[7] A.L. Barabási and R. Albert, Emergence of scaling in random networks, Science, Vol.286, No.5439, 509-512, 1999.

[8] T. Hastie,R. Tibshirani,J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Berlin: Springer, Chap. 4, 2009.

[9] J. MacQueen, Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp. Math. Statistics Prob., Vol.1, No.14, 281-297, 1967.

[10] P. Wang, X. Yu and J. Lü, Identification and evolution of structurally dominant nodes in protein-protein interaction networks, IEEE Trans. Biomed. Circuits Syst., Vol.8, No.1, 87-97, 2014.

[11] P. Wang, C.G. Tian and J.A. Lu, Identifying influential spreaders in artificial complex networks, J. Syst. Sci. Complex., Vol.27, No.4, 650-665, 2014.

[12] M.E.J. Newman, Mixing patterns in networks, Phys. Rev. E, Vol.67, No.2, art.no. 026126, 2003.

[13] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett., Vol.89, No.20, art. no. 208701, 2002.

[14] M.E.J. Newman, Structure and function of complex networks, SIAM Rev., Vol.45, No.2, 167-256, 2003.

[15] Y.Y. Liu, J.J. Slotine and A.L. Barabási, Controllability of complex networks, Nature, Vol.473, No.7346, 167-173, 2011.

[16] Z.Z. Yuan, C. Zhao, Z.R. Di, W.X. Wang and Y.C. Lai, Exact controllability of complex networks, Nat. Commun., Vol.4, No.2447, art.no. 2447, 2013.