Statistical significance of the rich-club phenomenon in complex networks

Zhi-Qiang Jiang$^{1,2}$ and Wei-Xing Zhou$^{1,2,3,4,5}$

$^1$ School of Business, East China University of Science and Technology, Shanghai 200237, People’s Republic of China
$^2$ School of Science, East China University of Science and Technology, Shanghai 200237, People’s Republic of China
$^3$ Research Center for Econophysics, East China University of Science and Technology, Shanghai 200237, People’s Republic of China
$^4$ Research Center of Systems Engineering, East China University of Science and Technology, Shanghai 200237, People’s Republic of China
E-mail: wxzhou@ecust.edu.cn

New Journal of Physics 10 (2008) 043002 (9pp)
Received 11 December 2007
Published 4 April 2008
Online at http://www.njp.org/
doi:10.1088/1367-2630/10/4/043002

Abstract. We propose that the rich-club phenomenon in complex networks should be defined in the spirit of bootstrapping, in which a null model is adopted to assess the statistical significance of the rich-club detected. Our method can serve as a definition of the rich-club phenomenon and is applied to analyze three real networks and three model networks. The results show significant improvement compared with previously reported results. We report a dilemma with an exceptional example, showing that there does not exist an omnipotent definition for the rich-club phenomenon.
1. Introduction

Almost all social and natural systems are composed of a huge number of interacting components. Many self-organized features that are absent at the microscopic level emerge in complex systems due to the dynamics. The topological properties of the underlying network of the interacting constituents have great impact on the dynamics of the system evolving on it [1]–[4]. Most complex networks exhibit small-world properties [5] and are scale-free in the sense that the distribution of degrees has a power-law tail [6]. In addition, many real networks have modular structures or communities expressing their underlying functional modules [7] and exhibit self-similar and scale invariant nature in the topology [8]–[14]. The modular and hierarchical structure of social networks may partly account for the log-periodic power-law patterns presented extensively in financial bubbles and antibubbles [15, 16]. A closely relevant feature is recently reported in some complex networks, termed the rich-club phenomenon; however, there is a lack of consensus on its definition [17]–[24].

The rich-club phenomenon in complex networks depicts the observation that the nodes with high degrees (called rich nodes) are inclined to intensely connect with each other. The average hop distance of the tight group is between one and two [18]. Intuitively, rich nodes are much more likely to organize into tight and highly interconnected groups (clubs) than low-degree nodes. Therefore, it is rational to accept that there is a rich-club phenomenon in the topology of the Internet [17]–[20]. This rationale can be characterized quantitatively by the rich-club coefficient \( \phi(k) \), which is expressed as follows [18]:

\[
\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)},
\]

where \( N_{>k} \) refers to the number of nodes with degrees higher than a given value \( k \) and \( E_{>k} \) stands for the number of edges among the \( N_{>k} \) nodes. The rich-club coefficient \( \phi(k) \) is the ratio of the real number to the maximally possible number of edges linking the \( N_{>k} \) nodes, which measures how well the rich nodes ‘know’ each other. For example, \( \phi = 1 \) means that the members within the club form a fully connected network.

Zhou and Mondragón [18] argue that an increasing function \( \phi(k) \) with respect to \( k \) provides evidence for the presence of the rich-club structure. However, Colizza et al [21] point out that a monotonic increase of \( \phi(k) \) is not enough to infer the presence of the rich-club phenomenon since even random networks generated from the Erdös–Rényi (ER) model, the Molloy–Reed (MR) model and the Barabási–Albert (BA) model have an increasing \( \phi(k) \) with respect to \( k \). Instead, the rich-club coefficient \( \phi(k) \) should be normalized by a reference and
the correct null model that can serve as a reference is the maximally random network with the same sequence of degrees as the network under investigation [21, 22]. The maximally random network can be generated with the chain-switching method [25, 26]. The normalized rich-club coefficient is defined by
\[ \rho(k) = \frac{\phi(k)}{\phi_{\text{ran}}(k)}, \] (2)
where $\phi_{\text{ran}}(k)$ is the average rich-club coefficient of the maximally random network [21]. The actual presence of the rich-club phenomenon in a network is confirmed if $\rho(k) > 1$ [21, 22]. In this framework, there is no rich-club ordering in the network of the Internet.

2. Statistical tests

We have repeated the analysis of Colizza et al [21] for three model networks, namely the ER model [27], the MR model [28] and the BA model [6], and three real-world networks, namely the protein interaction network [25] of the yeast Saccharomyces cerevisiae, the scientific collaboration network collected by Newman [29] and the Internet network at the autonomous system level collected by the Oregon Route Views project [30]–[32]. The rich-club coefficients $\phi$ of the six networks under investigation are presented in figure 1 with black circles as a function of the percentage $g$ of the richest nodes included in the rich club. The $\phi_{\text{ran}}$ functions are also shown for the corresponding maximally random networks. We note that when we plot $\phi(k)$ versus $k$, our results for the investigated networks are the same as shown in figure 1 obtained by Colizza et al [21].

Figure 2 shows the $\rho$ functions, which are not the same as those in figure 2 presented by Colizza et al [21]. Specifically, we find that the normalized coefficients of the networks of protein interactions, scientific collaborations, the ER model and the MR model are qualitatively the same as those reported by Colizza et al [21], while the other two are not. We note that the AS-Internet data were created by Mark Newman from the data for 22 July 2006. Figure 2 shows that the normalized coefficient $\rho$ is not less than 1 for the Internet and the BA model. For the Internet case, we notice that its $\phi$ is close to 1 for the richest nodes. Intuitively, the corresponding $\phi_{\text{ran}}$ should be less than 1, which is observed in our analysis but not in that of Colizza et al [21].

The importance of the null model has been emphasized in the assessment of some properties claimed to be present in complex networks [22, 25, 33]. Other than the simple normalization of the rich-club coefficient, we argue that the correct way to assess the presence of the rich-club phenomenon is to perform a statistical test, which amounts to determining the probability that the identified rich-club phenomenon emerges by chance. The null hypothesis is the following: $H_0$: $\rho(g)$ is not larger than 1.

The alternative hypothesis is that $\rho(g) > 1$. We can compute the $p$-value, which is the probability that the null hypothesis is true. We adopt the chain-switching method on the original network to obtain a maximal random network. We can estimate $\hat{\rho}(g)$ through equations (1) and (2). This procedure is repeated $n$ times, which gives $n$ values of $\hat{\rho}(g)$ for each $g$. Finally, the $p$-value which is the probability of a false alarm for rich club (the so-called ‘false positive’ or error of type II), can be calculated as follows:
\[ p = \frac{\#[\hat{\rho}(g) \leq 1]}{n}, \] (3)
where $\#[\hat{\rho}(g) \leq 1]$ counts the number of $\hat{\rho}(g)$ whose values are not larger than 1. As $n \to \infty$, it is clear that the estimated bootstrap $p$-value will tend to the ideal bootstrap $p$-value. Note
that \( n = 1000 \) in our case. The smaller the \( p \)-value, the stronger the evidence against the null hypothesis and favoring the alternative hypothesis that the presence of rich-club ordering is statistically significant. The \( p \)-value is 100\% when \( g = 1 \). By adopting the conventional significance level of \( \alpha = 5\% \), the rich-club phenomenon is statistically significant if \( p < \alpha \).

Figure 3 shows the \( p \)-values as a function of the percentage \( g \) of rich nodes for the networks investigated. For the protein interaction network, the ER network and the MR network, the \( p \)-values are larger than \( \alpha \) when \( g < 10\% \). Therefore, there is no rich-club ordering in these three networks. For the Internet, except for the point at the smallest \( g \) and the point with \( g = 1 \), all \( p \)-values are well below \( \alpha = 5\% \), indicating significant rich-club ordering in the Internet. For the scientific collaboration network, the \( p \)-values are less than \( \alpha = 5\% \) for most values of \( g \). However, the most connected scientists corresponding to small \( g \) do not form a rich club. According to the top right panel of figure 2, the group of these most connected scientists has a relatively large normalized rich-club coefficient. What is most surprising is the fact that the BA network has a significant rich-club phenomenon.

Among these cases, the presence of a rich club in the Internet has stirred quite a few debates. In a recent work, Zhou and Mondragón [34] found that there is a clique of rich nodes that are completely connected, which we think is an undoubtable hallmark for the presence of a rich club. We can provide further evidence for this argument. As illustrated in figure 1, the rich-club coefficients \( \phi \) are close to 1 when \( g \) is small for the Internet and the BA model. This means
Figure 2. Normalized rich-club coefficients of the investigated networks. The ratio $\rho_{\text{ran}} = \phi/\phi_{\text{ran}}$ as a function of the percentage $g$ and compared with the baseline value equal to 1.

Figure 3. Statistical tests for the presence of rich-club phenomenon.

that the richest nodes in these networks are almost fully connected. This validates the intuitive definition that a rich club is a group of nodes with high degrees that are intensely linked. A statistical test puts further credit on the declaration of Zhou and Mondragón for the presence of a rich club in the Internet.
Figure 4. Broad distributions of the values of $\phi_{\text{ran}}$ for the six networks under investigation. Each curve corresponds to a given value of $g$ shown in the legend.

It would be helpful to provide evidence showing that a statistical test is necessary for rich-club identification. The main difference between our method and that proposed by Colizza et al [21] is that here one does not use a mean as a reference value of the null model, but actual values corresponding to specific realizations. Our method is more general in the sense that it allows us to include also the case in which the value of $\hat{\rho}(g)$ displays large fluctuations about the mean. This for instance happens for the Newman–Girvan modularity measure, which is also based on a comparison with the mean over maximal random networks, leading to unexpectedly large values of the measure for random networks [35]. Figure 4 shows that this is indeed the case for the six networks studied in this work. We observe that the values of $\phi_{\text{rnd}}$ are broadly distributed. It is interesting to note that there are fewer distinct values of $\phi_{\text{rnd}}$ with the decrease in $g$, which is consistent with the fact that the remaining network for smaller $g$ has a smaller size and fewer possible configurations.

3. Connectedness within rich clubs

A missing ingredient in the discussions of the rich-club phenomenon is the connectedness of the rich club. When we define rich nodes as those with for example $g > 1\%$ and start to investigate whether these nodes form a rich club, scrutiny should be carried out to see if this ‘club’ contains several disconnected subclubs. As illustrated in the upper panel of figure 5, the scientific collaboration network is not fully connected for small $g$. There are several separated clusters for small $g$. According to figure 3, all three subgraphs are rich clubs, which however contradicts the common intuition that the members are aware of each other forged by other members in the club. For $g = 0.141$, there are two rich clubs $(1, 4, 5, 9, 11, 12)$ and $(2, 3, 6, 10, 16, 20, 21, 14, 15, 17)$. With the increase of richness (smaller $g$ or larger $k$), the rich club $(1, 4, 5, 9, 11, 12)$ remains unchanged. The second rich club $(2, 3, 6, 10, 16, 20, 21, 14, 15, 17)$ splits into two
Figure 5. Reassessment of rich-club phenomenon for the scientific collaboration network. The upper panel shows the subgraphs with $g = 0.080\%$ ($k = 66$), $0.111\%$ ($k = 57$) and $0.141\%$ ($k = 54$). The lower panel shows the statistical analysis of the subclubs for different $g$. The blue markers in the left panel show the rich-club coefficient $\phi$ for all isolated subclubs with more than two nodes, while the red ones in the same panel are the associated $\phi_{\text{ran}}$. The middle panel presents the $\rho$ function and the right panel depicts the corresponding $p$-values. It is observed that $p < \alpha$ for all $g$.

clubs (2, 3, 6, 10, 16) and (14, 15, 17) when nodes 20 and 21 are removed for $g = 0.111$. When $g = 0.080$, the rich club (14, 15, 17) disappears and (2, 3, 6, 10, 16) degenerates to (2, 3, 6). Therefore, when there is more than one isolated cluster of nodes for a given $g$, we should investigate their statistical significance one by one except for the trivial cases of isolated nodes and pairs of nodes. The lower panel of figure 5 shows the results for the scientific collaboration network. One observes that $p < 5\%$ for all clusters.

4. An open problem

So far, we have shown that performing a statistical test is necessary, which does a good job in the detection of rich clubs in complex networks. However, a coin always has two sides. Consider a toy network shown in figure 6. The graph consists of two kinds of nodes identified with different colors: the degree of each white node is $k = 1$, while the red nodes are very ‘rich’ and fully connected. It is evident that the rich-club coefficient of the red nodes is $\phi(k = 1) = 1$ and one would say they are in a rich club without any doubt. Indeed, a qualitatively identical value was taken as an example of the presence of a rich club [21]. Surprisingly, this observation of $\phi(k = 1)$ does not ensure that the red nodes form a rich club in either the framework of $\rho > 1$ adopted by Colizza et al [21] or the statistical test proposed in this work since $\phi_t(k = 1) \equiv 1$ for

New Journal of Physics 10 (2008) 043002 (http://www.njp.org/)
all maximally random networks. Hence, we have \( \rho(k = 1) = 1 \), which means that there is no rich-club ordering when \( k = 1 \). This conclusion contradicts our intuitions.

We can generalize our discussion above by considering a network consisting of \( m \) rich nodes, which are linked to \( k_1, k_2, \ldots, k_m \) nodes of degree \( k = 1 \), respectively. Since each node with \( k = 1 \) has to be linked to a node with \( k > 1 \) to ensure the connectedness of the randomized network, the group of \( m \) rich nodes has \( \sum_{i=1}^{m} k_i \) out-edges and \( E_{>1} \) edges among them. The value of \( E_{>1} \) does not change for all randomized networks. In other words, \( \phi_{\text{ran}}(k = 1) = \phi(k = 1) \) and \( \rho(k = 1) = 1 \). This class of artificial networks invalidates the sophisticated approach based on statistical tests. Essentially, it shows that the null model used may not always be the natural choice. Rather than conserving the whole degree sequence, one could keep approximately constant only the first and second moments, or the average distance between nodes, or one or several eigenvalues of the adjacency matrix. There are many alternatives that make sense in certain contexts. Further investigations are called for on the choice of null models, which are, however, beyond the scope of the current work.

5. Summary

The analysis presented here provides a more rigorous methodology for detecting rich clubs in complex networks. This allows us to understand the rich-club phenomenon on a solid basis. However, there exists a class of artificial networks with rich clubs invalidating the methods based on null models taking maximally random networks. In this sense, the definition of the rich-club phenomenon remains an open problem.

Acknowledgments

We are grateful to Dr Newman for providing autonomous systems network data and scientific collaboration network data and two anonymous referees for insightful suggestions. This work
was partially supported by the National Natural Science Foundation of China (grant no 70501011), the Fok Ying Tong Education Foundation (grant no 101086), the Shanghai Rising-Star Program (grant no 06QA14015), and the Program for New Century Excellent Talents in University (grant no NCET-07-0288).

References

[1] Albert R and Barabási A L 2002 Rev. Mod. Phys. 74 47–97
[2] Newman M E J 2003 SIAM Rev. 45 167–256
[3] Dorogovtsev S N and Mendes J F F 2003 Evolution of Networks: From Biological Nets to the Internet and the WWW (Oxford: Oxford University Press)
[4] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D U 2006 Phys. Rep. 424 175–308
[5] Watts D J and Strogatz S H 1998 Nature 393 440–2
[6] Barabási A L and Albert R 1999 Science 286 509–12
[7] Newman M E J 2004 Eur. Phys. J. B 38 321–30
[8] Kim B J 2004 Phys. Rev. Lett. 93 168701
[9] Song C M, Havlin S and Makse H A 2005 Nature 433 392–95
[10] Strogatz S H 2005 Nature 433 365–66
[11] Song C M, Havlin S and Makse H A 2006 Nat. Phys. 2 275–81
[12] Goh K I, Salvi G, Kahng B and Kim D 2006 Phys. Rev. Lett. 96 018701
[13] Kim J S, Goh K I, Salvi G, Oh E, Kahng B and Kim D 2007 Phys. Rev. E 75 016110
[14] Zhou W X, Jiang Z Q and Sonnerte D 2007 Physica A 375 741–52
[15] Sonnerte D 2003 Phys. Rep. 378 1–98
[16] Zhou W X, Sonnerte D, Hill R A and Dunbar R I M 2005 Proc. R. Soc. London B 272 439–44
[17] Zhou S and Mondragón R J 2004 Phys. Rev. E 70 066108
[18] Zhou S and Mondragón R J 2004 IEEE Commun. Lett. 8 180–82
[19] Zhou S and Mondragón R J 2005 Lect. Notes Comput. Sci. 3391 207–14
[20] Zhou S, Zhang G Q and Zhang G Q 2007 IET Commun. 1 209–14
[21] Colizza V, Flammini A, Serrano M A and Vespignani A 2006 Nat. Phys. 2 110–15
[22] Amaral L A N and Guimera R 2006 Nat. Phys. 2 75–6
[23] McAuley J J, Costa L d F and Caetano T S 2007 Appl. Phys. Lett. 91 084103
[24] Xu X P, Hu J H and Liu F 2007 Chaos 17 023111
[25] Maslov S and Sneppen K 2002 Science 296 910–13
[26] Milo R, Kashtan N, Itzkovitz S, Newman M E J and Alon U 2004 Uniform generation of random graphs with arbitrary degree sequences Preprint cond-mat/0312028
[27] Erdős P and Rényi A 1959 Publ. Math. 6 290–7
[28] Molloy M and Reed B 1995 Random Struct. Algorithms 6 161–79
[29] Newman M E J 2001 Proc. Natl Acad. Sci. USA 98 404–9
[30] Faloutsos M, Faloutsos P and Faloutsos C 1999 Comput. Commun. Rev. 29 251–62
[31] Pastor-Satorras R, Vazquez A and Vespignani A 2001 Phys. Rev. Lett. 87 258701
[32] Vazquez A, Pastor-Satorras R and Vespignani A 2002 Phys. Rev. E 65 066130
[33] Guimerà R and Sales-Pardo M 2006 Mol. Syst. Biol. 2 42
[34] Zhou S and Mondragón R J 2007 New J. Phys. 9 173
[35] Newman M E J and Girvan M 2004 Phys. Rev. E 69 026113

New Journal of Physics 10 (2008) 043002 (http://www.njp.org/)