Vol 17 No 7, July 2008
1674-1056/2008/17(07)/2327-06

Chinese Physics B

© 2008 Chin. Phys. Soc.
and IOP Publishing Ltd

An estimation formula for the average path length of scale-free networks*

Li Ying(李 昊)[a)b)], Cao Hong-Duo(曹宏铎)[a)],
Shan Xiu-Ming(山秀明)[b)], and Ren Yong(任 勇)[b)]

[a)] *Department of Management Sciences, School of Business, SUN YAT-SEN University, Guangzhou* 510275, *China*

[b)] *Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*

(Received 12 September 2007; revised manuscript received 26 December 2007)

A universal estimation formula for the average path length of scale free networks is given in this paper. Different from other estimation formulas, most of which use the size of network, N, as the only parameter, two parameters including N and a second parameter α are included in our formula. The parameter α is the power-law exponent, which represents the local connectivity property of a network. Because of this, the formula captures an important property that the local connectivity property at a microscopic level can determine the global connectivity of the whole network. The use of this new parameter distinguishes this approach from the other estimation formulas, and makes it a universal estimation formula, which can be applied to all types of scale-free networks. The conclusion is made that the small world feature is a derivative feature of a scale free network. If a network follows the power-law degree distribution, it must be a small world network. The power-law degree distribution property, while making the network economical, preserves the efficiency through this small world property when the network is scaled up. In other words, a real scale-free network is scaled at a relatively small cost and a relatively high efficiency, and that is the desirable result of self-organization optimization.

Keywords: scale free network, power law, small world, power law exponent
PACC: 0250, 0520

1. Introduction

A large number of complex networks found in real-world applications have aroused more and more interest in the study of these networks. Many complex systems, artificial and natural, can be modelled as networks. Examples of these network-represented systems include the Internet, World Wide Web (WWW), metabolic networks, citation networks, cellular networks, etc.[1−3] The elements of these systems are represented as nodes of these networks, and the interactions between the elements are expressed as edges between corresponding nodes.[4] All these networks devised to study complex systems share some common properties. These networks, known as complex networks, all have large sizes, and the connectivity property cannot be easily expressed. Complex networks have been studied by using statistical methods, and they generally have three properties, i.e. scale-free, high clustering, and small world.

A scale-free network is a network in which degree distribution follows a power law,[1] that is, the probability distribution function of the degree of nodes follows $P(k) \sim k^{-\alpha}$, where k is the degree of a node, and α is usually a constant between 1 and 3. Small world property says that in these networks, despite the large size of the network, there is a relatively short path between any two nodes, where the path length l is defined as the number of edges along the shortest path connecting any two nodes. The average path length $< l >$ is a mean of all the lengths of paths in this network. The small world property can be written as $< l > \sim \ln(N)$, which says that the average path length $< l >$ increases logarithmically with the size of network, N, increasing. The high clustering property is the trend in the forming of charmed circle in the network. It is quantified by a clustering coefficient C_i.

In a complex network, the power-law degree distribution property and the small world property coexist in the same network. Is there any connection between the two properties? What is the relationship between the power-law degree distribution and the average path length?

Many studies have given different models for var-

*Project supported by the National Natural Science Foundation of China (Grant Nos 60672142, 60772053 and 90304005).

ious complex networks, for example, the Internet[5] and WWW networks,[6] but there is not any universal model that is applicable to all complex networks. Recently, there has been an increasing interest in formulas that calculate average path length of a scale-free network, or of real-world networks such as the email network and WWW. Zhang *et al*[7] gave an evolutionary Apollonian network and its topology property. They also studied the pseudofractal network, and indicated that the average path length was affected by the parameters of model and proportional to the logarithm of the size of network, N.[8] Their work was based on Dorogovtsev *et al*'s, in which the pseudofractal network has been first given. In Zhang *et al*'s work, the power-law exponent is 1+ln3/ln2, and the average path length approximately is $\frac{4}{9}\ln N_{\rm t}+\frac{5}{3}\ln\frac{3}{2}$, where $N_{\rm t}$ is the numbers of vertices.[9] Ebel *et al*[10] gave the average path length of the email network by using the generating function approach. Davidsen *et al*[11] studied an acquaintance network model, and gave its formula of average path length as $l \approx \dfrac{\log\left(\dfrac{N}{<k>}\right)}{\log\left(\dfrac{<k^2>-<k>}{<k>}\right)}+1$. Zou *et al*[12] investigated a fractal network with a mixed degree distribution and gave its topology property. Li and Cai[13] performed a statistical analysis of the airport network of China, and gave a formula of the diameter of the sub-cluster. Cohen and Havlin[14] studied the diameter of scale-free networks whose degree distribution $P(k)$ satisfies a power law $P(k)\sim k^{-\gamma}$, and by using analytical arguments they showed that $d(N)\sim\begin{cases}\ln(\ln N), & 2<\gamma<3,\\ \ln N/\ln(\ln N), & \gamma=3,\\ \ln N, & \gamma>3.\end{cases}$. Lee *et al*[15] measured the diameter of finite scale-free networks by random walks on finite SFNs, and their results were in agreement with Cohen and Havlin's conclusion.

In this research, we aim at finding a universal formula that can be used to estimate the average path lengths of all types of scale-free networks. This formula will give insights into the question of what the influence factors of the average path length are, and it will explain the relationship between the small world property and the power-law degree distribution. It will also help to better understand the advantages of scale-free networks.

2. The meaning of the power-law degree distribution and the small world

Consider the economical aspect of a scale-free network. If more edges are added to a network, the average path length is expected to become smaller. Thereby, consider a large network composed of a large number of nodes: only if it has a larger number of edges, will its average path length be shorter. A scale-free network has a relatively short average path length only because of a relatively small number of edges despite a large number of nodes. That the number of edges is not too large results from the power-law degree distribution. According to the power-law degree distribution $P(k)\sim k^{-\alpha}$, the larger the k is, the smaller the $P(k)$ will be. The result is the number of the nodes with a smaller degree is more than that with a large degree. So in a scale-free network, the total number of edges is not too large when considering the size of the network. If we use the total number of edges in a network to represent the cost of the network, the power-law degree distribution can make the network more economical.

On the other hand, the average distance between two nodes represents the efficiency of a network. Almost all networks, including the Internet and WWW, are expected to have a short average path length to achieve a better performance. So the average path length is often used as a measure of performance of a network. The small world property, which says that there is a relatively small distance between any two nodes, makes a complex network more efficient.

From an economical point of view, the power-law degree distribution and the small world together result in a complex network that is both economical and efficient. In other words, a scale-free network makes use of relatively few edges to obtain a relatively short average path length.

Our objective of this research is to find out how the goals in efficiency and economy are achieved simultaneously in complex networks, and how the power-law degree distribution affects the average path length.

3. Estimation of the average path length of complex network

In the power-law degree distribution $P(k) \sim k^{-\alpha}$, the exponent α shows the local connectivity property of the network. Apparently, the average path length $< l >$ is determined by this local connectivity property of the network. Considering that $< l >$ is also affected by the size of network N, it can be expressed as a function of α and N. In Ref.[5], we have analysed in detail the average path length of the Internet whose degree distribution follows a power law. One conclusion was that the path-length distribution function $P(l)$ must be a descending function of l, and also a function of α. In Ref.[5], we also gave a formula $P(l) = \left(1 - \frac{1}{\alpha}\right)^{1.7} \cdot l^{-\frac{3\ln(5.55-\alpha)}{5\ln\alpha}-1}$, which is suitable for the Internet but may not be applicable to other scale-free networks. By replacing the constant in the primary formula with a parameter, we give an empirical formula of $P(l)$ applicable to an arbitrary type of scale-free network as follows:

$$P(l) = \left(1 - \frac{1}{\alpha}\right)^{\frac{c}{3}} \cdot l^{-\frac{\log(c-\alpha)}{\log\alpha}-1}, \qquad (1)$$

where l is the path length, which is defined as the number of edges along the shortest path connecting any two nodes; $P(l)$ is the path-length distribution probability, which is determined by the connectivity property of network. In expression (1), besides α, the exponent of power law $k^{-\alpha}$, there is another parameter c, which varies with type of scale-free network. By adjusting this parameter c, expression (1) can be used for any type of complex network.

$< l >$ can be estimated based on $\sum_{l=1}^{l_{\max}} P(l)l$, so we have

$$< l >= \left(1 - \frac{1}{\alpha}\right)^{\frac{c}{3}} \sum_{l=1}^{l_{\max}} l^{-\frac{\log(c-\alpha)}{\log\alpha}}, \qquad (2)$$

where $l_{\max}$ is the longest possible path length, which is less than or equal to N. So expression (2) can be approximated as

$$< l > \approx \left(1 - \frac{1}{\alpha}\right)^{\frac{c}{3}} \sum_{l=1}^{N} l^{-\frac{\log(c-\alpha)}{\log\alpha}}. \qquad (3)$$

We compare $< l >_{\text{prediction}}$ from expression (3) and $< l >_{\text{real}}$ of many complex networks. The results are shown in Table 1 and Fig.1 indicating that the predictions are close to the observed values.

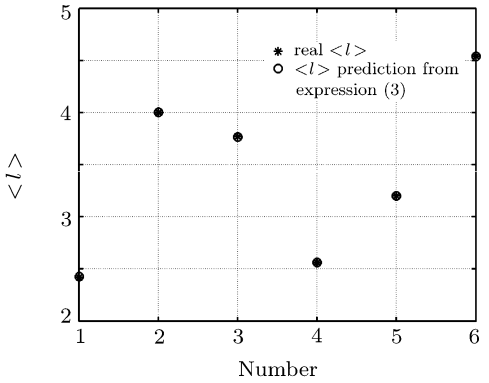


Fig.1. The average path lengths of real networks, $< l >_{\text{real}}$, from the reference and the predictions, $< l >_{\text{prediction}}$, from expression (3), showing that $< l >_{\text{prediction}}$ is almost equal to $< l >_{\text{real}}$.

Table 1. Comparison between the average path lengths of real networks, $< l >_{\text{real}}$, from the references, and the predictions $< l >'_{\text{prediction}}$, $< l >''_{\text{prediction}}$, and $< l >'''_{\text{prediction}}$, respectively from expressions (3), (4) and (6).

network	size N	α	c	$\alpha \times 2$	$<l>_{\text{real}}$	$<l>'_{\text{prediction}}$	$<l>''_{\text{prediction}}$	$<l>'''_{\text{prediction}}$	Ref.	Number
software	1439	1.5	3.013	3.00	2.42	2.4271	2.6164	2.6162	[3]	1
package Internet	3015	2.15	4.248	4.30	4	4.0020	3.5030	3.5028	[16]	2
router Internet	3530	2.16	4.299	4.32	3.77	3.7681	3.5730	3.5729	[16]	3
router metabolic network	765	2.2	4.478	4.40	2.56	2.5635	2.9670	2.9666	[17]	4
Merabolic, E. coli	778	2.2	4.363	4.40	3.2	3.2028	2.9739	2.9735	[17]	5
Movie actors	212250	2.3	4.653	4.60	4.54	4.5399	5.3545	5.3544	[1]	6

To study the potential correlation between c and α in Table 1, we plot c against α as shown in Fig.2. From this figure, we can see that c grows linearly with α, and c is approximately twice as big as α, i.e. $c \approx 2\alpha$. By substituting this into expression (3), we obtain

$$< l >= \left(1 - \frac{1}{\alpha}\right)^{\frac{2\alpha}{3}} \sum_{l=1}^{N} l^{-1}. \qquad (4)$$

We compare the values of $< l >$ obtained from expression (4) and the values of $< l >_{\text{real}}$ observed from the above complex networks. The results are shown in Table 1 and Fig.3.

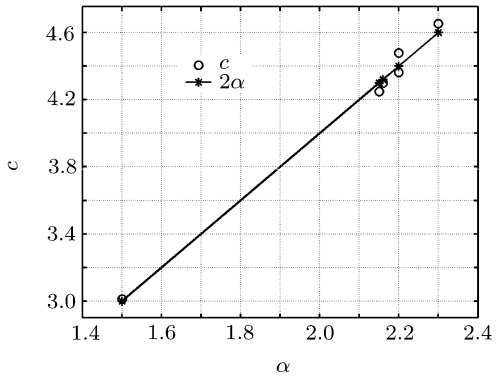


Fig.2. The relation between α and c, showing they have a good linear relation, where $c \approx 2\alpha$.

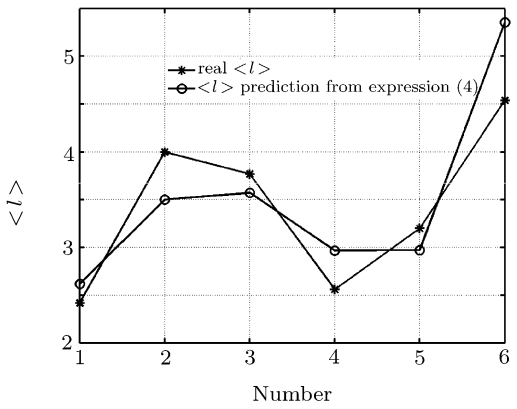


Fig.3. The average path lengths of real networks and the predictions from expression (4).

Further more, we also know

$$\begin{aligned}\sum_{l=1}^{N} l^{-1} =& \frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\cdots\cdots+\frac{1}{N}\\ \approx& \ln(N)+C\\ =& \ln(N)+0.577,\end{aligned} \tag{5}$$

where C is Euler's constant and approximately equal to 0.577.

By substituting expression (5) into expression (4), we obtain

$$< l >= \left(1-\frac{1}{\alpha}\right)^{\frac{2\alpha}{3}} (\ln(N)+0.577). \tag{6}$$

The above expression is very interesting and important, and it will be further discussed in the next section. It is consistent with the findings of many previous studies, such as those done by Albert and Barabási etc. We use this expression to re-estimate the predicted average path length of aforementioned real-world networks, and the results are shown in Table 1 and Fig.4.

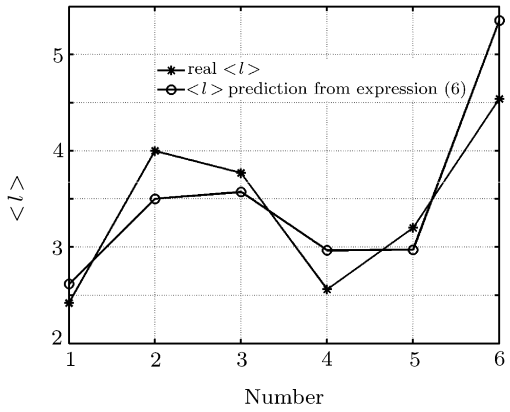


Fig.4. The average path lengths of real networks and the predictions from expression (6).

Using expression (6), besides the size of the network, N, the remaining parameter is only α, the value of which is required to be estimated in order to obtain the average path length of a scale-free network. This makes this estimation method easy to use, and the accuracy is still satisfactory. For example, for co-authorship network of neuroscience,[18] N=70975, α=2.1, and the real average path length $< l >$=6, while the average path length is predicted from expression (6) to be 5.2.

4. Analysis of the formula

Two variables: N, which is the number of nodes in the network, and α, which is the power-law exponent, are included in expression (6). From this expression, N has a positive correlation with $< l >$, and α has a more complex relation with $< l >$, which is shown in Fig.5. In Fig.5, when $\alpha < 1$ the value of $\left(1-\frac{1}{\alpha}\right)^{\frac{2\alpha}{3}}$ decreases, and when $\alpha > 1$ it increases. For most real-world networks, α falls in a range approximately from 2 to 3, so in these networks $< l >$ increases with α increasing. Thereby, $< l >$ will increase if either N or α or both grow larger.

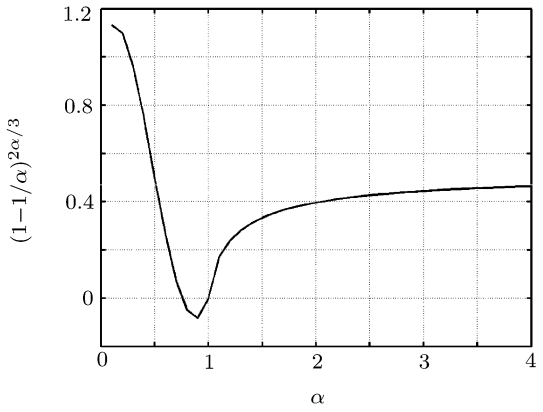


Fig.5. The relation between α and $\left(1-\frac{1}{\alpha}\right)^{\frac{2\alpha}{3}}$.

Because $< k >$, the mean of the degree of node k, satisfies that $< k > \sim \sum k^{-\alpha+1}$, a larger α will result in a smaller average degree $< k >$ and therefore fewer total edges in the network. Because of this decrease in number of edges in the network, the average path length $< l >$ will increase. This is consistent with the correlation between $< l >$ and α we conclude above based on our formula in expression (6).

Expression (6) also indicates that the average path length $< l >$ increases logarithmically with the size of the network, N, i.e. $< l > \sim \ln(N)$, which is the definition of small world. This is also consistent with the results of many previous studies, such as those performed by Albert and Barabási,[1] Newman,[3] and Barabási and Albert.[4] But these previous studies are different from ours in that they each give a different formula for a different network, which is often derived from statistic data for that specific network, while our formula is obtained from studying a general network at a microscopic level and is not restricted to any specific network. For example, Albert gave a formula for WWW, in which he stated that the average path length $< l >$ and the size of the network N satisfied $< l >= 0.35 + 2.06\ln(N)$.[6] This research, along with many other previous studies, was performed at a macroscopic level by using a statistical approach, but our study starts at a microscopic level of a general network, without considering the macroscopic properties such as the size of the network. It finally reaches conclusions consistent with those drawn from all previous studies. Therefore, a big advantage of our formula is its universal applicability. Given the values of α and N for any scale-free network, the average path length can be estimated from our formula.

Another major difference between our formula and those proposed in the previous studies is that our formula includes the parameter α besides the size of network, N, which is not included in other estimation formulas. Since α represents a local connectivity property and $< l >$ represents a global connectivity property, our formula captures the fact that the local connectivity property at a microscopic level of a network determines the global connectivity property of the whole network. In other words, our formula reflects the joint effect of the local connectivity property and the size of network on the global property of the whole network. This distinguishes expression (6) from other estimation formulas.

A side conclusion from our formula is that if a network is a scale-free network, then it must also be a small world network. This is because any scale-free network will follow the rule of power-law degree distribution, and our formula states that such a network always has a relatively short average path length despite the large size, or in other words, such a network has the small world property. The power-law degree distribution makes the scale-free network economical, and the small world makes the network efficient. Thereby, the scale network is intrinsically both economical and efficient, which makes it scaled up efficiently with a relatively small growth in cost. This desirable result is because of self-organization optimization.

5. Conclusion

In this paper, we present a new formula to estimate the average path length based on both the size and the local connectivity property of the network. We argue that the local connectivity property and the size of network both affect the global connectivity property of networks, and that the previous efforts to estimate the average path length solely from the size of the network, N, is not sufficient. Our new estimation formula as shown in expression (6), takes into consideration both the size of the network, N, and a parameter α, which effectively captures the local connectivity property at a microscopic level, in estimating the average path length. The use of this parameter α marks a major difference between our formula and those proposed in the previous studies, and makes our formula applicable to all types of scale-free networks: for any scale-free network, given the values of α and N, the average path length $< l >$ can be easily estimated by inserting these values into expression (6). We have also shown that our formula, as a generalization of

all network-specific estimation methods, draws conclusions consistent with those from the previous studies focusing on specific types of networks. In summary, our formula provides a universal and easy-to-compute estimation for all types of scale-free networks.

Through our new estimation formula we have also shown the intrinsic relationship between the average path length and the size of the scale-free network. Expression (6) discloses that in a scale-free network, the average path length will grow logarithmically with the size of the network, i.e. $< l > \sim \ln(N)$, which is the definition of the small world property. This shows that the small world property is only a side result of the power-law degree distribution. In other words, the power-law degree distribution determines the small world property. Therefore, the power-law degree distribution, while governing the economical growth of the network on its own, preserves the efficiency of the network with a slowly increasing average path length when the network is scaled up. In conclusion, a scale-free network is both economical and efficient, and that is a desirable result of self-organization optimization.

References

[1] Albert R and Barabási A L 2002 *Rev. Mod. Phys.* **74** 47

[2] Li Y, Liu Y, Shan X M, Ren Y, Jiao J and Qiu B 2005 *Chin. Phys.* **14** 2153

[3] Newman M E J 2003 cond-mat/0303516

[4] Barabási A L and Albert R 1999 *Science* **286** 509

[5] Li Y, Shan X M and Ren Y 2004 *Acta Phys. Sin.* **53** 69 (in Chinese)

[6] Albert R, Jeong H and Barabási A L 1999 *Nature* **401** 130

[7] Zhang Z Z, Rong L L and Zhou S G 2005 cond-mat/0511609

[8] Zhang Z Z, Zhou S G and Chen L C 2007 *Eur. Phys. J.* B **58** 337

[9] Dorogovtsev S N, Goltsev A V and Mendes J F 2002 *Phys. Rev.* E **65** 066122

[10] Ebel H, Mielsch L I and Bornholdt S 2002 *Phys. Rev.* E **66** 035103

[11] Davidsen J, Ebel H and Bornholdt S 2002 *Phys. Rev. Lett.* **88** 128701

[12] Zou L H, Pei W J, Li T and He Z Y 2006 physics/0612056

[13] Li W and Cai X 2004 *Phys. Rev.* E **69** 046106

[14] Cohen R and Havlin S 2003 *Phys. Rev. Lett.* **90** 058701

[15] Lee S M, Yook S-H Y and Kim Y 2007 cond-mat/0701070

[16] Faloutsos M, Faloutsos P and Faloutsos C 1999 *Comput. Commun. Rev.* **29** 251

[17] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabási A L 2000 *Nature* **407** 651

[18] Barabási A L, Jeong H, Néda Z, Ravasz E, Schubert A and Vicsek T 2001 cond-mat/ 010416