Measuring microscopic evolution processes of complex networks based on empirical data

Liping Chi\textsuperscript{1,2}
\textsuperscript{1}Complexity Science Research Center, College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China
\textsuperscript{2}Center for Complex Network Research, Department of Physics, Northeastern University, Boston, MA 02115, USA
E-mail: chilp@mail.ccnu.edu.cn

Abstract. Aiming at understanding the microscopic mechanism of complex systems in real world, we perform the measurement that characterizes the evolution properties on two empirical data sets. In the Autonomous Systems Internet data, the network size keeps growing although the system suffers a high rate of node deletion ($r = 0.4$) and link deletion ($q = 0.81$). However, the average degree keeps almost unchanged during the whole time range. At each time step the external links attached to a new node are about $c = 1.1$ and the internal links added between existing nodes are approximately $m = 8$. For the Scientific Collaboration data, it is a cumulated result of all the authors from 1893 up to the considered year. There is no deletion of nodes and links, $r = q = 0$. The external and internal links at each time step are $c = 1.04$ and $m = 0$, correspondingly. The exponents of degree distribution $p(k) \sim k^{-\gamma}$ of these two empirical datasets $\gamma_{\text{data}}$ are in good agreement with that obtained theoretically $\gamma_{\text{theory}}$. The results indicate that these evolution quantities may provide an insight into capturing the microscopic dynamical processes that govern the network topology.

1. Introduction
There has been an increasing interest in characterizing the evolution properties of complex systems under the framework of network science. Models have been introduced to capture the basic evolution mechanism determining the system structure. The early model of \textit{growth} and \textit{preferential attachment} proposed by Barabasi and Albert in 1999 \cite{1} successful explained the heterogeneous degree distribution in real-world complex systems. Many other models have been built to obtain different dynamical properties observed in real networks, such as fitness \cite{2}, aging \cite{3}, copy model \cite{4}, temporal pattern \cite{5, 6}, to name a few. Recent research puts more emphasis on the time effect of complex systems with time-stamped connectivity, such as bursty in email data \cite{7}, time aggregation in mobile phone data \cite{8}. The other direction along this line is to perform the measurements directly on evolving systems to elucidate the formation of network structure. Newman measured the clustering coefficient and preferential attachment of scientific collaboration networks in physics and biology and provided a sublinear power-law degree distribution \cite{9, 10}. Jeong et al inspected the role of added external and internal links during the time evolution of real networks \cite{11}. Although similar in spirit of the latter, our approach emphasizes the measurement of evolution properties at a microscopic scale.

In Ref. \cite{12}, the model of four elementary processes in network evolution was analyzed...
theoretically to understand the individual roles of adding and deleting nodes or links, as well as the random and preferential attachment in link creation. In this paper we will employ the model to investigate two empirical data sets: the Autonomous Systems Internet data and Scientific Collaboration data. In addition to understanding their topology, the study is intended to uncover the microscopic dynamical properties of complex systems in real world and provide a way of probing the evolution processes that determine the network structure.

The rest of the paper is organized as follows. Section II briefly introduces the network evolution model in Ref. [12] and the empirical data sets we analyze. In Section III we present the methodology and corresponding results, especially the comparison of degree distribution between empirical data and theory. Conclusion and discussion are offered in the final section.

2. Model and data description
This paper attempts to characterize the evolution mechanism driving the network structure. The network evolution model will be performed on two empirical data sets to estimate the microscopic dynamical processes.

2.1. Model
In Ref. [12], to understand the individual roles of adding and deleting processes in forming network topology, four elementary processes are included at each time step:

(i) Add a new node along with \( c \) external links. Each link is attached to the existing node with probability \( \pi(k) = a + bk \);

(ii) Add \( m \) internal links among existing nodes. The probability of choosing one end of the internal link is given by \( f(k) = s + tk \);

(iii) Delete a node with probability \( r \);

(iv) Delete an internal link with probability \( q \).

For simplicity, we assume \( s = a \) and \( t = b = 1 \), which leads to \( \pi(k) = f(k) = a + k \). The parameter \( a \) is called the random attachment, or initial attractiveness by Dorogovtsev et al [13].

At time \( t \), the number of nodes in the network is \( n_t = (1 - r)t \) and the number of edges is \( m_t = [c - r\langle k \rangle + m(1 - q)]t \). According to \( \langle k \rangle = \frac{2m_t}{m} \), we obtain the average degree

\[
\langle k \rangle = \frac{2}{1 + r}[c + m(1 - q)].
\]  

(1)

Degree distribution is an important metric to describe the network properties. We assume the degree distribution has the form

\[
p(k) \sim k^{-\gamma} \Omega^k,
\]  

(2)
in which \( \Omega \) is a controllable parameter. If \( \Omega \leq 1 \), the degree distribution is a power-law distribution with an exponential cutoff. The exponent \( \gamma \) is given by [12]

\[
\gamma = 2 + \frac{(1 + r)[c(c + m(1 - q)) + a(c + m)(1 + r)]}{c^2(1 - r) + c[m(1 - q)(3 - r) - ar(1 + r)] + m[2m(1 - q)^2 - a(1 + r)(q + r)]}.
\]  

(3)

When \( 2 < \gamma < 3 \), the degree distribution is typically a heterogeneous distribution. As \( \gamma \) gets much larger, the degree distribution decays rapidly, indicating that the hubs are few and the distribution tends toward homogeneous.
2.2. Data description

Two empirical data sets are chosen for analysis.

Autonomous System (AS) Internet data: The data contains 685 daily instances which span an interval from November 8, 1997 to January 2, 2000 [14].

Scientific Collaboration data: The scientific collaboration data consists of all the papers published in APS journals with DOI prefix 10.1103, providing a range of 117 years from July 1893 to Dec 2009 [15, 16]. The data in the considered year is a cumulated result of all the papers and authors from 1893 up to the given year.

3. Method and Results

3.1. Autonomous System Internet data

In the Autonomous System network (ASN), each node is an Autonomous System number and the link represents the communication between Autonomous Systems. Figure 1 presents the number of nodes \( N \) and the average degree \( \langle k \rangle \) as a function of time. In spite of frequently adding or deleting nodes and links, the network size \( N \) keeps growing. However, it is interesting to notice that the average degree \( \langle k \rangle \) is almost unchanged.

We plot the degree distribution in Fig. 2 for ASN. The method of logarithmic binning is used in Fig. 2(b) to reduce the noise in the tail. We see that \( p(k) \) approximately fits to a power-law degree distribution \( p(k) \sim k^{-\gamma} \) with exponent \( \gamma_{\text{data}} = 2.21 \). The subscript \( \text{data} \) denotes the exponent is from the empirical data.

Next we will employ the model to measure the following quantities that characterize the microscopic dynamical processes in the data. First, we will discuss the quantities that describe the topological changes at each time step, which include external links attached to a new node \( c \), added internal links \( m \), probability of node deletion \( r \) and probability of link deletion \( q \). These parameters are crucial to understand the evolution processes. Figure 3 reports the time evolution of these parameters. The peak values (ie, most probable numbers) are chosen to approximate the values of these parameters. For ASN, \( c = 1.1 \), \( m = 8 \), \( r = 0.4 \) and \( q = 0.81 \), as shown in Table 1.

Second, we will focus on the measurement of random attachment \( a \) in \( \pi(k) = f(k) = a + k \). As stated in Ref. [11], the random attachment is very small and available statistics is insufficient for us to measure it. Hence in this paper we take the random attachment as \( a = 0 \). It implies that the attachment \( \pi(k) \) linearly depends on degree \( k \).

Substituting these quantities into Eq. (3), we obtain the theoretical exponent of degree distribution \( \gamma_{\text{theory}} = 2.43 \). Compared with \( \gamma_{\text{data}} = 2.21 \), the result demonstrates that these quantities can possibly capture the microscopic evolution mechanism of empirical data.
Figure 2. (a) Plot of degree distribution $p(k) \sim k^{-\gamma}$ for ASN with exponent $\gamma_{data} = 2.21$. (b) The method of logarithmic binning is used to reduce the noise in the tail.

Figure 3. Plot of empirically measured quantities that govern the evolution processes for ASN. (a) added external links attached to a new node $c$; (b) added internal links $m$; (c) relative probability of node deletion $r$ and; (d) relative probability of link deletion $q$. The peak values (ie, most probable numbers) are chosen to approximate the values of these parameters. For ASN, $c = 1.1$, $m = 8$, $r = 0.4$ and $q = 0.81$. 
Table 1. Quantities that describing the evolution processes at each time step, including the external links $c$, internal links $m$, probability of node deletion $r$ and link deletion $q$, as well as the power-law exponents obtained theoretically $\gamma_{\text{theory}}$ and empirically $\gamma_{\text{data}}$.

|    | $c$ | $m$ | $r$ | $q$ | $\gamma_{\text{theory}}$ | $\gamma_{\text{data}}$ |
|----|-----|-----|-----|-----|--------------------------|---------------------|
| ASN| 1.1 | 8   | 0.4 | 0.81| 2.43                     | 2.21                |
| SCN| 1.04| 0   | 0   | 0   | 3                        | 2.13 - 4.01         |

Figure 4. (a) Cumulative number of authors $N$ and (b) the average degree $\langle k \rangle$ in the given year. The inset exhibits that the average degree $\langle k \rangle$ increases linearly with $N$.

3.2. Scientific Collaboration data

In the Scientific Collaboration network (SCN), the nodes are the authors and the links represent that two authors wrote a paper together. The number of nodes $N$ and the average degree $\langle k \rangle$ as a function of time are plotted in Fig. 4. We see that the network size $N$ grows exponentially with time due to the addition of new authors. The average degree $\langle k \rangle$ is approximately a constant before 1960s and then increases sharply afterwards.

Figure 5 plots the cumulative degree distribution of the given year, assuming the degree distribution is of the form $p(k) \sim (\gamma - 1)k_0^{\gamma - 1}\gamma^\gamma(k + k_0)^{-\gamma}$ [18]. The exponents $\gamma_{\text{data}}$ vary for different years, from 4 in 1950 to 2.13 in 2009. The inset exhibits the changes of $\gamma_{\text{data}}$ with time.

Now we calculate the evolution quantities for SCN. It is noticed that the Scientific Collaboration data is a cumulated data. There is no deletion of nodes and links, indicating that $r = 0$ and $q = 0$. The corresponding theoretical exponent of degree distribution is

$$\gamma = 2 + \frac{a + c}{c + 2m}. \quad (4)$$

We plot the external links $c$ and internal links $m$ as a function of time in Fig. 6. The external links $c$ give the average collaborators a new author has. The internal links $m$ represent the number of collaborations between old authors who did not collaborate before. The peak values are $c = 1.04$ and $m = 0$ (see Table 1), which shows that each new author has about one existing old collaborator and there is no collaboration between old authors from the view point of over a century’s time.

Providing the random attachment $a = 0$, the theoretical exponent of degree distribution for
Figure 5. Cumulative degree distribution for SCN, assuming the degree distribution is $p(k) \sim (\gamma - 1)k_0^{\gamma - 1}(k + k_0)^{-\gamma}$. The exponents $\gamma_{\text{data}}$ vary for different years, falling into the range of 2.13 to 4.

Figure 6. Plot of empirically measured quantities that govern the evolution processes for SCN. (a) added external links attached to a new node $c$ and; (b) added internal links $m$. The peak values of $c$ and $m$ are 1.04 and 0, respectively.

SCN is $\gamma_{\text{theory}} = 3$. The result also exhibits a consistency between the theory and empirical data.

4. Conclusion and discussion

The analysis of microscopic evolution processes of complex networks is performed on two empirical data sets: the Autonomous Systems Internet data and Scientific Collaboration data. A detailed description of the evolution properties has been investigated, including the external links $c$ that a newly added node has, added internal links between existing nodes $m$, the probability of node deletion $r$ and link deletion $q$. In the Autonomous Systems Internet data, the network size keeps growing although the system suffers a high rate of node deletion ($r = 0.4$) and link deletion ($q = 0.81$). The average degree $\langle k \rangle$ is approximately a constant as the network size $N$ expands. The external and internal links at each time step for ASN are $c = 1.1$ and $m = 8$. It implies that the internal link plays an important role in network growth. The theoretical and empirical exponents of degree distribution are $\gamma_{\text{theory}} = 2.43$ and $\gamma_{\text{data}} = 2.21$, showing a good agreement. For the Scientific Collaboration data, it is a cumulated result with $r = q = 0$. The external and internal links are $c = 1.04$ and $m = 0$, correspondingly. A similar consistency in the exponent of degree distribution has been observed for SCN.
The agreement between the empirical data and theory indicates that these quantities can capture the dynamical processes determining the network topology. However, two key questions still remain uncertain. First, in the theoretical analysis of dynamical processes, time $t$ is a continuous variable and being strictly defined. It is unclear how to set time $t$ when analyzing empirical data. In this paper we take one added node as a time step. Second, the time interval is an important parameter in describing the network evolution. If the time interval is too small, there are only several nodes and links in each interval. It is hard to observe the formation of clusters. On the other hand, if the time interval is too large, it contains approximately all the nodes in each interval. It is difficult to inspect the emergence of critical links. It is still an open question on how to choose an optimal time interval to analyze the time-stamped network. In spite of all these challenges, this work provides a way of probing the microscopic evolution mechanism driving the network structure.

Acknowledgement
The author thanks Gourab Ghoshal for valuable discussions and suggestions. This work was partially supported by the Programme of Introducing Talents of Discipline to Universities under Grant No. B08033, and the Fundamental Research Funds for the Central Universities of Central China Norma University (CCNU).

References
1. Barabási A-L and Albert R 1999 Science 286 509-512
2. Bianconi G and Barabási A L 2001 Europhys. Letts. 54 436-442
3. Dorogovtsev S N and Mendes J F F 2000 Phys. Rev. E 62 1842-1845
4. Krapivsky P L and Redner S 2005 Phys. Rev. E 71 036118
5. Holme P and Saramäki J 2012 Physics Reports 519 97
6. Nicosia V, Tang J, Mascolo C, Musolesi M, Russo G and Latora V 2013 Temporal networks: understanding complex systems, (Springer Berlin Heidelberg) pp 15-40
7. Barabási A 2005 Nature 435 207-211
8. Kriou G, Karsai M, Bernhardsson S, Blondel V D and Saramäki 2012 EPJ Data Science 1(4) 1-16
9. Newman M E J 2003 Phys. Rev. E 64 025102
10. Newman M E J 2010 Networks: An Introduction (Oxford: Oxford University Press)
11. Jeong H, Neda Z and Barabási A L 2003 Europhys. Letts. 61 567-572
12. Ghoshal G, Chi L P and Barabási A L 2013 Scientific Reports 3 2920
13. Dorogovtsev S N, Mendes J F F and Samukhin A 2000 Phys. Rev. Lett. 85 4633
14. http://snap.stanford.edu/data/us.html
15. Barabási A L, Jeong H, Néda Z, Ravasz E, Schubert A and Vicsek T 2002 Physica A 311 590-614
16. Deville P, Wang D, Sinatra R, Song C, Blondel V D and Barabási A L 2014 Scientific Reports 4 4770
17. Martin T, Ball B, Karrer B and Newman M E J 2013 Phys. Rev. E 88 012814
18. Ghoshal G and Barabási A L 2011 Nature Commun. 2 394