Evolution of the Internet Map and Load Distribution

K.-I. Goh, B. Kahng*, and D. Kim

School of Physics and Center for Theoretical Physics, Seoul National University,
Seoul 151-747, Korea

Abstract. We track the evolutionary history of the Internet at the autonomous systems (ASes) level and provide the evidence that it can be described in the framework of the multiplicative stochastic process. It is found that the fluctuations arising in the process of diversifying connections of each node play an essential role in forming the *status quo* of the Internet. Extracting relevant parameters for the growth dynamics of the Internet topology, we are able to predict the connectivity (degree) exponent γ of the Internet AS map successfully. We also introduce a quantity called the load as the capacity of node needed for handling the communication traffic and study its distribution over the Internet across years. The load distribution follows a power law with the exponent $\delta \approx 2.0$ and the load at the hub scales with the network size as $\ell_h \sim N^{1.8}$.

1 Introduction

During recent years, the Internet has become one of the most influential media in our daily life, going beyond in its role as the basic infrastructure in the technological world. Explosive growth in the number of users and hence the amount of traffic poses a number of problems which are not only important in practice for, e.g., maintaining it free from any undesired congestion and malfunctioning, but also of theoretical interests as an interdisciplinary topic [1]. Such interests, also stimulated by other disciplines like biology, sociology, and statistical physics, have blossomed into a broader framework of network science [2,3,4]. The Internet is a primary example of complex networks. It consists of a large number of very heterogeneous units interconnected with various connection bandwidths, however, it is neither regular nor completely random. In their landmark paper, Faloutsos et al. [5] showed that the Internet at the autonomous systems (ASes) level is a scale-free (SF) network [6], meaning that it follows a power-law distribution

$$p_d(k) \sim k^{-\gamma} \tag{1}$$

in node degree k, the number of connections a node has. The degree exponent γ of the AS map is subsequently measured by a number of groups to be $\gamma \approx 2.1$.

Emergence of such power-law degree distribution calls for explanation and understanding of the basic mechanism underlying the growth of the Internet.

* Corresponding author (E-mail: kahng@phya.snu.ac.kr).

M. Bubak et al. (Eds.): ICCS 2004, LNCS 3038, pp. 1038–1045, 2004.

Once revealed, it can be used to predict what the Internet will be like in the future, as well as how it has evolved into the present shape. In the first part of the paper, we will address this issue, showing that it can be described by a simple physical model, the multiplicative stochastic process. By extracting relevant parameters for the stochastic process from the time history of the AS map deposited in the Oregon route views project, we can predict the degree exponent of the Internet accurately.

The Internet is not a quiet object. Data packets are sent and received over it constantly, causing momentary local congestion from time to time. To avoid such undesired congestion, the capacity, or the bandwidth, of the routers should be as large as it can handle the traffic. In the second part of the paper, we will introduce a rough measure of such capacity, called the load and denoted as ℓ. The distribution of the load reflects the high level of heterogeneity of the Internet: It also follows a power law,

$$p_l(\ell) \sim \ell^{-\delta}, \tag{2}$$

with the load exponent δ. We will discuss the implication of the power-law behavior of the load distribution.

2 Internet Evolution as a Multiplicative Stochastic Process

The mechanism of the emergence of SF network is mostly captured by the Barabási-Albert (BA) model [7] which assumes the linear growth in numbers of nodes and links in time and the preferential attachment (PA) in establishing links from a new node to other previously existing ones. The PA means that the probability $\Pi_i(t)$ that a node i will receive a link from the new node created at time t is linearly proportional to its present degree $k_i(t)$, i.e., $\Pi_i(t) = k_i(t)/\sum_j k_j(t)$. The empirical evidence of the PA in the Internet has been reported [8,9]. As we will see, however, the assumption that the numbers of nodes and links increase linearly in time does not apply to the real situation of the Internet. Rather, the numbers of nodes and links increase exponentially but with different rates. Furthermore, the interconnections between nodes are being updated continually in the Internet, which was not incorporated in the original BA model.

Huberman and Adamic (HA) [10] proposed another scenario for SF networks. They argued that the fluctuation effect arising in the process of connecting and disconnecting links between nodes is an essential feature to describe the dynamics of the Internet topology correctly. In their model, the total number of nodes $N(t)$ increases exponentially with time as

$$N(t) = N(0)\exp(\alpha t). \tag{3}$$

Next, they assumed that the degree k_i at a node i evolves through the multiplicative stochastic process,

$$k_i(t+1) = k_i(t)[1 + \zeta_i(t+1)], \tag{4}$$

where $\zeta_i(t)$ is the growth rate of the degree k_i at time t, which fluctuates from time to time. Thus, one may divide the growth rate $\zeta_i(t)$ into two parts, $\zeta_i(t) = g_{0,i} + \xi_i(t)$, where $g_{0,i}$ is the mean value over time, and $\xi_i(t)$ the rest part, representing fluctuations over time. $\xi_i(t)$ is assumed to be a white noise satisfying $\langle \xi_i(t) \rangle = 0$ and $\langle \xi_i(t)\xi_j(t') \rangle = \sigma_{0,i}^2 \delta_{t,t'}\delta_{i,j}$, where $\sigma_{0,i}^2$ is the variance. Here $\langle \cdots \rangle$ is the sample average and $\delta_{a,b}$ is the Kronecker delta symbol. For later convenience, we denote the logarithm of the growth factor as $G_i(t) \equiv \ln[1+\zeta_i(t)]$. Then a simple application of the central limit theorem ensures that the probability distribution of $k_i(t)/k_i(t_0)$, t_0 being a reference time, follows the log-normal distribution for sufficiently large t. To get the degree distribution, one needs to collect all contributions from different ages τ_i, growth rates $g_{0,i}$, standard deviations $\sigma_{0,i}$ and initial degree $k_i(t_0)$. HA further assumed that ζ_i are identically distributed so that $g_{0,i} = g_0$ and $\sigma_{0,i} = \sigma_0$ for all i. Then the conditional probability for degree, $P_d(k, \tau |\, k_0)$, that $k_i(t_0 + \tau) = k$, given $k_i(t_0) = k_0$ is given by

$$P_d(k, \tau |\, k_0) = \frac{1}{k\sqrt{2\pi\sigma_{\text{eff}}^2 \tau}} \exp\left\{ -\frac{\left(\ln(k/k_0) - g_{\text{eff}}\tau\right)^2}{2\sigma_{\text{eff}}^2 \tau} \right\}, \qquad (5)$$

where $g_{\text{eff}} \equiv \langle G_i(t) \rangle$ and $\sigma_{\text{eff}}^2 \equiv \left\langle (G_i(t) - \langle G_i(t) \rangle)^2 \right\rangle$. g_{eff} and σ_{eff}^2 are related to g_0 and σ_0^2 as $g_{\text{eff}} \approx g_0 - \sigma_0^2/2$ and $\sigma_{\text{eff}}^2 \approx \sigma_0^2$, respectively [11]. Since the density of nodes with age τ is proportional to $\rho(\tau) \sim \exp(-\alpha\tau)$, the degree distribution collected over all ages becomes $p_d(k) = \int d\tau \rho(\tau) P_d(k, \tau |\, k_0) \sim k^{-\gamma}$, where the degree exponent γ is given in terms of the growth parameters as

$$\gamma = 1 - \frac{g_{\text{eff}}}{\sigma_{\text{eff}}^2} + \frac{\sqrt{g_{\text{eff}}^2 + 2\alpha\sigma_{\text{eff}}^2}}{\sigma_{\text{eff}}^2}. \qquad (6)$$

In the next section, we will measure such parameters from the real evolutionary history of the Internet AS map and check if the HA scenario holds.

3 Growth Dynamics of Internet

A number of projects exist aiming to map the world-wide topology of the Internet. One such is the Route Views project initiated at the university of Oregon [12], the data of which are also archived at the National Laboratory of Applied Network Research (NLANR) [13]. Among the daily data from November 1997 to January 2000, we sample one AS map a month, with the total period of 26 months, and analyze them for various quantities. First we measure the growth rate of the number of ASes α. We also measure directly the growth rate of the number of links β, which can be crosschecked for consistency later. In Fig. 1, we show the total number of ASes $N(t)$ and the total number of links $L(t)$ as a function of time t. The straight line in log-linear plot means $N(t)$ and $L(t)$ indeed grows exponentially. The growth rates are determined to be $\alpha \approx 0.029$ and $\beta \approx 0.034$. We also find that the newly appeared AS would connect to only one or two existing ASes so that the average number of links the new AS establishes

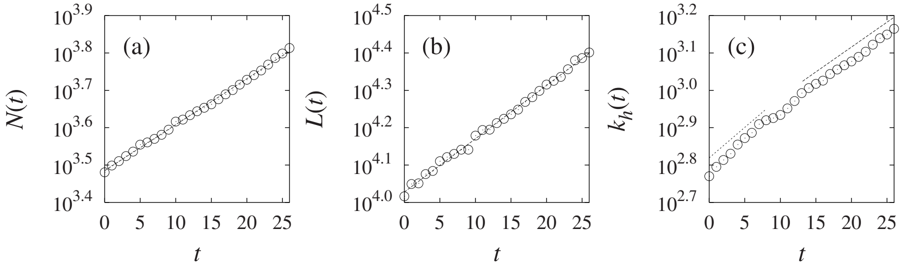

Fig. 1. The time evolution of the number of ASes $N(t)$ (a), the number of links between ASes $L(t)$ (b), and the degree of the hub $k_h(t)$ (c). Note that the ordinates in the figures are in logarithmic scale, indicating the exponential increase of corresponding quantities. The fitted line has a slope 0.029 in (a), 0.034 in (b), and 0.043 (0.030) for the dotted (dashed) one in (c).

is $k_{\text{new}} \approx 1.35$. Fig. 1(c) shows the growth of the degree of the hub, the node with the largest degree. It shows a change of growth rate around $t \approx 14$.

The measurement of g_0 and σ_0 is nontrivial due to the presence of large fluctuations. To this end, we measure the degree growth rate of a node i, $G_i(t)$, defined earlier as $G_i(t) \equiv \ln[1 + \zeta_i(t)] = \ln[k_i(t)/k_i(t-1)]$. To keep $G_i(t)$ well-defined for all t, we consider only the nodes existing for the entire time range $0 \leq t \leq 26$, the set composed of which is denoted by S hereafter. By the existence of a node we mean that its degree is nonzero, since we cannot identify an AS with no connection. For each i ($i \in S$), let $g_i = \langle G_i(t) \rangle_t$ and $\sigma_i^2 = \left\langle (G_i(t) - g_i)^2 \right\rangle_t$, where $\langle \cdots \rangle_t$ means the temporal average over the period $16 < t \leq 26$ ($T = 10$). If the HA scenario holds, the histogram of $\{g_i\}$ for all nodes would follow the Gaussian distribution with the mean g_{eff} and the variance σ_{eff}^2/T. We show such histogram in Fig. 2, the fit of which to the Gaussian gives the mean $\overline{g}$ as 0.016 and the standard deviation σ_g as 0.04. The measured values of $\{\sigma_i^2\}$ give the mean value $\overline{\sigma^2} \approx 0.017$.

It is most likely that $\overline{g}$ and $\overline{\sigma^2}$ would have a distribution over nodes. As HA assumed, however, we try to approximate the growth process by a single process whose effective mean growth rate and standard deviation are g_{eff} and σ_{eff}, respectively. Then the Eq. (5) should hold for all i and all t with a suitable choice of those parameters. For this purpose, we consider the distribution $P[k_i(t)/k_i(t_0)]$ in terms of the scaled variables x and y defined as

$$x \equiv \frac{\ln[k_i(t)/k_i(t_0)] - g_d(t - t_0)}{\sqrt{2\sigma_d^2(t - t_0)}}, \tag{7}$$

and

$$y \equiv P[k_i(t)/k_i(t_0)][k_i(t)/k_i(t_0)]\sqrt{2\pi\sigma_d^2(t - t_0)}, \tag{8}$$

where we set $t_0 = 0$ and g_d and σ_d are parameters to be chosen. From Eq. (5), with suitably chosen parameters g_d and σ_d, the distribution for different time

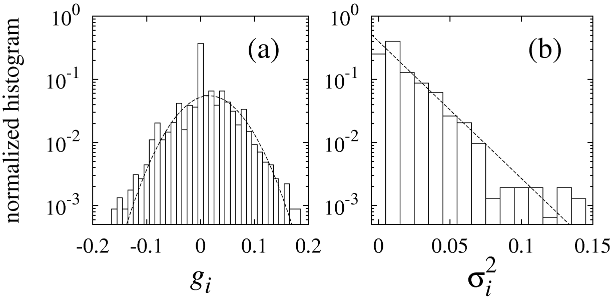

Fig. 2. The normalized histogram of g_i (a) and σ_i^2 (b). In (a), the data is fitted with a Gaussian with the mean 0.016 and the standard deviation 0.04. In (b), the data is fitted with an exponential decay $\exp(-x/x_c)$ with the characteristic scale $x_c \approx 0.02$. The measured value of the average is $\overline{\sigma^2} \approx 0.017$.

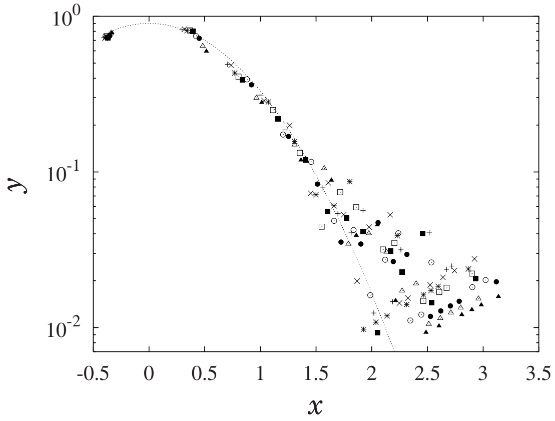

Fig. 3. Plot of $P[k_i(t)/k_i(t_0)]$ versus $k_i(t)/k_i(t_0)$ for different times $t > 16$ in terms of the scaled variables y and x defined in Eqs. (7) and (8). Larger deviations for large x are due to t being finite and are caused by the rare statistics.

t would collapse onto a single curve, $\ln y = -x^2$. We show such data in Fig. 3. The best collapse can be accomplished by choosing the parameters $g_d = 0.016$ and $\sigma_d = 0.14$, which should be identified with g_{eff} and σ_{eff}, respectively. The effective growth parameters found in this way are in accordance with the ones estimated before as $\overline{g} = 0.016$ and $\overline{\sigma^2} = 0.017$. As noted earlier, the consistency of estimated parameters can be checked as, for example, it should satisfy $\beta = \max(\alpha, g_{\text{eff}} + \sigma_{\text{eff}}^2)$, for which we have $\beta = 0.034$ and $g_{\text{eff}} + \sigma_{\text{eff}}^2 = 0.035$, being reasonably consistent with each other. Thus we conclude the parameters $g_{\text{eff}} = 0.016$ and $\sigma_{\text{eff}} = 0.14$ can be regarded as the effective parameters of the degree growth dynamics of the Internet AS map as a single process. Applying those values together with $\alpha = 0.029$ found earlier into Eq. (6), we found $\gamma \approx 2.1$, which is in excellent agreement with the directly measured ones.

4 Load Distribution of Internet

To a large extent, the Internet is the medium of communication. The continuous communication between hosts generates certain amount of data traffic. To make the best use of it, we have to avoid congestions, from which we suffer possible delays and the loss of information. What's worse, one doesn't know when and how much a host will generate the traffic. Absent is the central regulation in the Internet, hence each node should do its best for its own ends. To give a measure of such activity, we define the load ℓ_i as the amount of capacity or bandwidth that a node i can handle in unit time [14]. Not knowing the level of traffic, one assumes that every node sends a unit packet to everyone else in unit time. One further assumes that the packets are transfered from the source to the target only along the shortest paths between them, and divided evenly upon encountering any branching point. To be precise, let $\ell_i^{s\to t}$ be the amount of packet sent from s (source) to t (target) that passes through the node i. Then the load of a node i, ℓ_i, is the accumulated sum of $\ell_i^{s\to t}$ for all s and t,

$$\ell_i = \sum_{s\neq t} \ell_i^{s\to t}. \tag{9}$$

In other words, the load of a node i gives us the information how much the capacity of the node should be in order to maintain the whole network in a free-flow state. To calculate load on each node, we use the modified breath-first search algorithm introduced by Newman [15] and independently by Brandes [16], which can evaluate $\{\ell_i\}$ in time of order $\mathcal{O}(N^2)$ for sparse binary graphs.

For a number of SF networks in nature and society, the load distribution is also found to follow a power law, Eq. (2) [17]. The Internet AS map is no exception and the load exponent δ of the power law is estimated to be approximately $\delta \approx 2.0$ [17,18]. The power-law load distribution means that a few ASes should handle an extraordinarily large amount of load while most others should do only a little.

The load of a node is highly correlated with its degree. The Pearson correlation coefficient between the two quantities is as high as 0.98. This suggests a scaling relation between the load and the degree of a node as

$$\ell \sim k^{\eta} \tag{10}$$

and the scaling exponent η is estimated as $\eta = 1.06 \pm 0.03$ for January 2000 AS map (Fig. 4a). In fact, the exponent η depends on γ and δ as $\eta = (\gamma - 1)/(\delta - 1) \approx 1.1$, which is consistent with the direct measurement.

The time evolution of the load at each AS is also of interest. Practically, how the load scales with the total number of AS (the size of the AS map) is an important information for the network management. In Fig. 4b, we show $\ell_i(t)$ versus $N(t)$ for 5 ASes with the highest rank in degree, i.e., 5 ASes that have largest degrees at $t = 0$. The data of $\{\ell_i(t)\}$ shows large fluctuations in time. Interestingly, the fluctuation is moderate for the hub, implying that the connections of the hub is rather stable. The load at the hub is found to scale

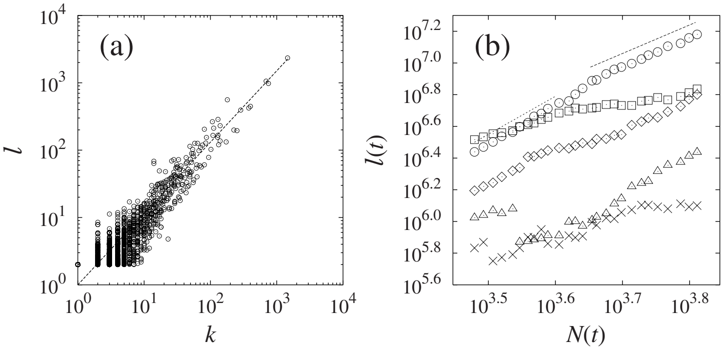

Fig. 4. (a) The scatter plot of the load versus the degree of each node for the AS map as of January 2000. The slope of the dashed line is 1.06, drawn for the eye. (b) Time evolution of the load versus $N(t)$ at the ASes of degree-rank 1 (○), 2 (□), 3 (◇), 4 (△), and 5 (×). The dashed line for larger N has slope 1.8, drawn for the eye.

with $N(t)$, $\ell_h(t) \sim N(t)^\mu$, but the scaling shows a crossover from $\mu \approx 2.4$ to $\mu \approx 1.8$ around $t \approx 14$, as it did for the degree.

5 Summary

We have studied the temporal evolution of the Internet AS map and showed that it can be described in the framework of multiplicative stochastic process for the degree growth dynamics. We measured the values of relevant parameters from the history of the AS map. With those values, the AS number growth rate $\alpha = 0.029$, the effective degree growth rate $g_{\text{eff}} = 0.019$, and its effective standard deviation $\sigma_{\text{eff}} = 0.14$, we were able to predict the degree exponent γ as $\gamma \approx 2.1$, which is in excellent agreement with the previously reported empirical values $\gamma_{\text{measured}} = 2.1 \sim 2.2$. Although it successfully accounts for the emergence of the scale-free characteristics of the Internet, the present description is by no means complete. More elaborated modeling [19,20,21] would improve our understanding of the evolutionary and organizational principle of the Internet and the research in this direction is highly called for.

In the second part of the paper, we introduced a quantity called the load. It can be thought of as the amount of traffic that a node should handle to keep the whole network away from the unwelcome congestion and maintain free-flow state, giving a measure of desired capacity of the nodes. The load distribution also follows a power law. The load and the degree of an AS are highly correlated with each other. The analysis of the temporal change of the load reveals that the load at the hub scales with the system size as $N^{1.8}$. Finally, we note that the contents of this article is in part overlapped with our previous studies published in [14,17,19].

References

1. Pastor-Satorras, R., Vespignani, A.: Evolution and Structure of the Internet. Cambridge University Press, Cambridge (2004)
2. Albert, R., Barabási, A.-L: Statistical mechanics of complex networks. Rev. Mod. Phys. **74** (2002) 47–97
3. Dorogovtsev, S. N., Mendes, J. F. F.: Evolution of Networks. Oxford University Press, Oxford (2003)
4. Newman, M. E. J.: The structure and function of complex networks. SIAM Rev. **45** (2003) 167–256
5. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships in the Internet topology. Comput. Commun. Rev. **29** (1999) 251–262
6. Barabási, A.-L., Albert, R., Jeong, H.: Mean-field theory for scale-free random networks. Physica A **272** (1999) 173–187
7. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286** (1999) 509–512
8. Pastor-Satorras, R., Vázquez, A., Vespignani, A.: Dynamical and correlation properties of the Internet. Phys. Rev. Lett. **87** (2001) 258701
9. Jeong, H., Néda, Z., Barabási, A.-L.: Measuring preferential attachment for evolving networks. Europhys. Lett. **61** (2003) 567–572
10. Huberman, B. A., Adamic, L. A.: Evolutionary dynamics of the World Wide Web. e-print (http://arxiv.org/abs/cond-mat/9901071) (1999)
11. Gardiner, C. W.: Handbook of Stochastic Methods. Springer-Verlag, Berlin (1983)
12. Meyer, D.: University of Oregon Route Views Archive Project (http://archive.routeviews.org)
13. The NLANR project sponsored by the National Science Foundation (http://moat.nlanr.net)
14. Goh, K.-I., Kahng, B., Kim, D.: Universal behavior of load distribution in scale-free networks. Phys. Rev. Lett. **87** (2001) 278701
15. Newman, M. E. J.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys. Rev. E **64** (2001) 016132
16. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25** (2001) 163–177
17. Goh, K.-I., Oh, E., Jeong, H., Kahng, B., Kim, D.: Classification of scale-free networks. Proc. Natl. Acad. Sci. USA **99** (2002) 12583–12588
18. Vázquez, A., Pastor-Satorras, R., Vespignani, A.: Large-scale topological and dynamical properties of the Internet. Phys. Rev. E **65** (2002) 066130
19. Goh, K.-I., Kahng, B., Kim, D.: Fluctuation-driven dynamics of the Internet topology. Phys. Rev. Lett. **88** (2002) 108701
20. Yook, S.-H., Jeong, H., Barabási, A.-L.: Modeling the Internet's large-scale topology. Proc. Natl. Acad. Sci. U.S.A. **99** (2002) 13382–13386
21. Rosvall, M., Sneppen, K.: Modeling dynamics of information networks. Phys. Rev. Lett. **91** (2003) 178701