

A Stochastic Dynamics for the Popularity of Websites

Chang-Yong Lee

Department of Industrial Information

Kongju National University

Chungnam, 340-702, South Korea

clee@kongju.ac.kr

In this paper, we have studied a dynamic model to explain the observed power law distribution for the popularity of websites in the WWW. The dynamic model includes the self growth for each website and the external force acting on the website. With numerical simulations of the model, we can explain most of the important characteristics of websites, such as a power law distribution of the number of visitors to websites and fluctuation in the fractional growth of individual websites.

1 Introduction

As the Internet and the World Wide Web (the web, for short) plays an important role in our present society, research on these becomes more and more active. In particular, study of the characteristics of websites and their dynamical behavior has become recognized as a new field of research. Aside from the technical understandings of the Internet and the web, within this new field, the Internet can be regarded as an “artificial complex system” of which many interacting agents, or websites are composed. As is true for most complex systems, size and dynamic variations make it impractical to develop characteristics of the web deterministically.

Despite the fact that the web is a very complex system, seemingly an unstructured collection of electronic information, it is found that there exists a simple and comprehensible law: the power law distribution. According to the research [1], the number of visitors to websites exhibits a power law distribution. This finding implies that most of data traffic in the web is diverted to a few popular websites. This power law distribution of the popularity for websites is one of the characteristics of the Internet web market and contrasts with the traditional equal share markets in which the transaction cost and geological factors play important roles.

In addition to the power law distribution, the analysis of empirical data [1] shows a few additional characteristics for the number of visitors to websites: first, the number of visitors follows a power law with different exponents depending on the category of websites. More specifically, for the “.edu” domain sites the exponent $\beta = 1.45$, and for all websites, $\beta = 2.07$. This shows that the exponent of all categories is greater than that of a specific category. Second, the fluctuation of the growth rate in the number of visitors for each site is uncorrelated.

In this paper, we investigate the power law distribution of the number of visitors to websites and the dynamic properties among competing websites. In particular, we focus on the result of empirical data analysis in Ref. [1]. For this end, we first build up a stochastic model for the number of visitors to the websites, and then carry out both numerical and analytic calculations.

2 A Dynamic Model

In general, a dynamic system can be described schematically as

$$\frac{d X_i(t)}{dt} = f_i(\vec{X}), \quad (1)$$

where \vec{X} represents the state of the system and takes values in the state or phase space. In the present case, $X = \{X_i(t)\}, i = 1, 2, \dots, N(t)$, and $X_i(t)$ is the number of visitors to the website i at time t . Expanding $f_i(X)$ in the powers of X_i and keeping the lowest order term in X_i , Eq. (1) can be rewritten, after absorbing the constant term into X_i , as

$$\frac{d X_i(t)}{dt} = A_{ii} X_i + \sum_{j \neq i} A_{ij} X_j. \quad (2)$$

There is one more ingredient that should be taken into account: an exponential growth in the number of websites. It is known that the number of websites connected to the Internet is not constant but increases exponentially [2, 3]. Thus the number of websites $N(t)$ at time t satisfies, in the continuous time limit,

$$\frac{dN(t)}{N(t)} = \lambda dt, \quad (3)$$

where λ is the growth rate of the number of websites. To implement this exponential growth, we discretize time and take Δt as the time step such that within Δt a new website can be added into the Internet with the probability $N(t)\lambda\Delta t$. That is, in each time step Δt , on the average, the number of the websites will be increased at time t by an amount

$$\Delta N(t) = N(t + \Delta t) - N(t) = N(t)\lambda\Delta t. \quad (4)$$

We further assume that no two websites can be created within Δt .

Now, let us determine the coefficients A_{ii} and A_{ij} . The first term on the right hand side of Eq. (2) is the self growth term of the website i with the growth rate A_{ii} . The implication of this term is that in two successive time periods the increase in the number of visitors is proportional to the number of visitors to that site. This is reasonable because once a website is created, the website will be known to more users. In consequence, more users visit the website as time progresses. We also assume that each website would grow with an equal rate so that the coefficient A_{ii} could be set to the same irrespective of the website. This assumption is valid if there are no other factors affecting on the growth of a website. Furthermore, the coefficient A_{ii} can be absorbed with an appropriate re-scaling of X_i so that one can set $A_{ii} = 1$ for all i .

The second term on the right hand side of Eq. (2) can be regarded as an ‘‘external force’’ acting on the website i . The coefficient A_{ij} should satisfy the following. First, the force has to be global, that is, the website i experiences a force from all the others. Since websites distributed over the Internet can be accessed by a few clicks of a button [4], accessing a website does not depend on the geographical degree of freedom. Second, the force term should include environmental changes in the Internet, such as the bandwidth, Internet technologies, and topology. Since it is difficult to take these changes into account explicitly, we describe the influence of the environmental changes via a stochastic process. The environmental fluctuations in essence can be modeled as a random process, thus it is convenient to express these as a Gaussian white noise process.

More specifically, during Δt , all factors for the environmental fluctuation are absorbed into a stochastic noise, which leads to a stochastic differential equation in time step Δt . That is, we lump all environmental influence on websites during Δt into a stochastic variable. Thus, we can write

$$A_{ij} \rightarrow \langle A \rangle + \kappa \eta_{ij}(t), \quad (5)$$

where κ is a time independent parameter representing the noise amplitude (or force strength) and $\eta_{ij}(t)$ is a Gaussian white noise characterized by

$$\langle \eta_{ij}(t) \rangle = 0 \text{ and } \langle \eta_{ij}(t) \eta_{kl}(t) \rangle = \delta(t-s) \delta_{ik} \delta_{jl}. \quad (6)$$

Note that we set $\langle A \rangle = 0$ for simplicity. We also assume that the external force strength acting on the website i depends on the number of websites influencing the website i . The physical implication of this assumption is that as the number of websites increases, the effective force strength from each website onto the website i decreases. Thus we take $\kappa \rightarrow \kappa / N(t)$ for the normalization.

With this stochastic nature of the external force term together with the exponential growth of the number of websites, the dynamics of the number of visitors to the website i can be expressed as

$$\frac{\Delta X_i}{\Delta t} = X_i + \frac{\kappa}{N(t)} \sum_{j \neq i}^{N(t)} \eta_{ij}(t) X_j, \quad (7)$$

where $\Delta X_i(t) = X_i(t + \Delta t) - X_i(t)$, and $N(t)$ satisfies Eq. (3). From the model, one finds that there are two parameters, κ and λ : κ being the noise strength and λ being the growth rate of the number of websites. Since the number of websites in the model is not constant but increases in time, it is not easy to solve the coupled dynamic equation analytically.

3 Simulation Results

With the stochastic dynamic equation of Eq. (7) and the exponential growth of the number of websites, we perform numerical simulations. In the simulation, we start with a small number of the websites (say, $N(0) = 10$) and at every time step Δt , a new website is added to the system with the probability .

Figure 1 shows cumulative distribution functions (CDF) of the number of visitors to websites with different number of total websites N_{total} at the end of each simulation. In the simulation, the growth rate and the force strength are held fixed as $\lambda = 0.5$ and $\kappa = 2.0$. From Fig. 1, one obtains a power law distribution as CDF, $C(x) \approx x^{-\alpha}$, and finds $\alpha \approx 0.5$. In terms of the probability density function (PDF), $P(x)$ of the number of visitors to the websites, we get $p(x) \propto x^{-\beta}$ with an exponent $\beta = 1 + \alpha \approx 1.5$. One can also see that the distribution of the number of visitors to websites follows a universal power law with the same exponent β irrespective of the total number of websites, N_{total} .

To see the effect of the force strength, we carried out simulations with different force strengths κ while keeping the other parameters fixed ($N_{total} = 2000$ and $\lambda = 0.5$). As can be seen in Fig. 2, the results for different κ fall into the same distribution, thus one can infer that the exponent does not depend on the force strength κ .

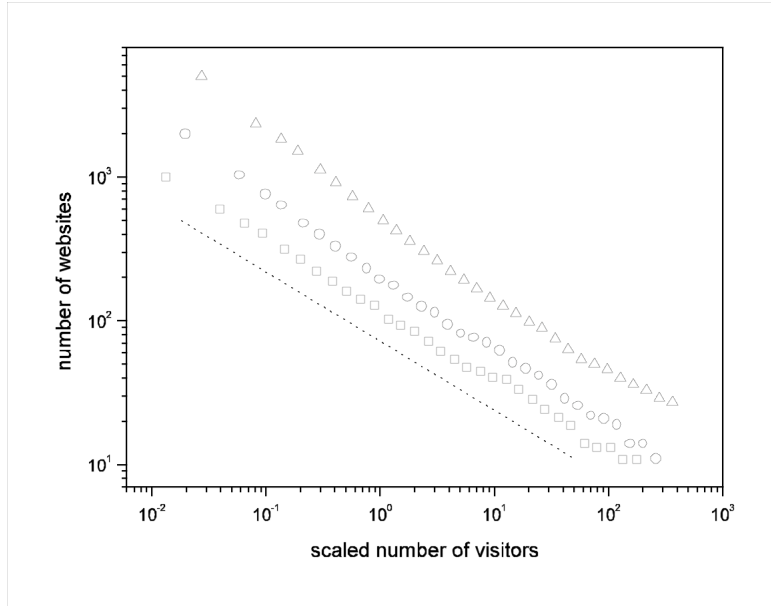


Figure 1 : Log-log scale plots of cumulative distribution functions of the number of visitors for the total number of websites at the end of simulations $N_{total} = 1000$ (\triangle), $N_{total} = 2000$ (\circ), and $N_{total} = 5000$ (\square). The dotted line has slope -0.5.

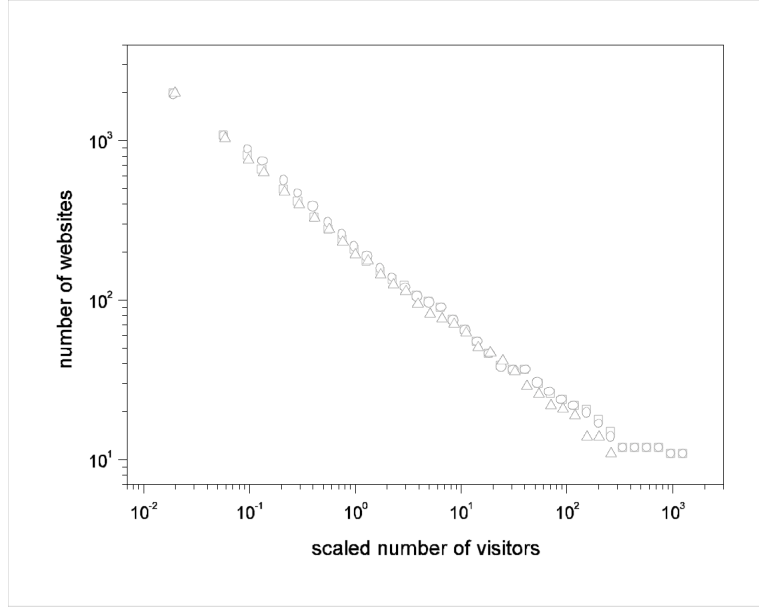


Figure 2 : Log-log scale plots of cumulative distribution functions of the number of visitors for $\kappa = 0.5$ (\circ), $\kappa = 1.0$ (\triangle), and $\kappa = 2.0$ (\square).

The force term in the model is responsible for the fluctuation of the number of visitors to websites. It is found in Ref. [1] that the fractional fluctuations in the number of visitors for a given website are uncorrelated to each other. To show this, we calculate the quantity

$$g(t) \equiv \frac{X(t + \Delta t) - X(t)}{X(t)}, \quad (8)$$

as a function of time. This random fluctuation of the fractional growth can be verified in terms of the auto-correlation. The calculation of the auto-correlation function shows that the fractional fluctuation is linearly uncorrelated. It should be also stressed that this uncorrelated fluctuation is independent of the force strength κ as well as the growth rate λ .

The power law distribution observed in Fig. 1 and Fig. 2 can be derived analytically within an appropriate approximation. Following the procedure similar to Ref. [5], we plot in Fig.3 the number of visitors $X_i(t)$ to various websites as a function of time with parameters $N_{total} = 2000$, $\kappa = 1.0$, and $\lambda = 0.5$. From Fig. 3 one can obtain an approximate differential equation for X_i as

$$\frac{\partial \ln X_i(t)}{\partial t} \approx \alpha, \quad (9)$$

where α is estimated from Fig. 3 as $\alpha \approx 1$.

The solution to Eq. (9) is given as

$$X_i(t) = m_0 e^{(t-t_i)}, \quad (10)$$

where $m_0 = X_i(0)$, and t_i is the time at which the website i is added to the system. Equation (10) implies that older websites (smaller t_i) increase their visitors at the expense of younger ones (larger t_i); « rich-get-richer » phenomenon that was observed in the dynamics of the various networks [5].

With the above result, we can get the probability distribution analytically. The probability that a website i has visitors smaller than x , $P(X_i(t) \leq x)$, can be written as $P(t_i \geq \tau)$, where $\tau = t - \ln(x/m_0)$. Note that $P(t_i \geq \tau)$ is the probability that the website i can be found in the system up to time τ . Therefore, the desired probability is just a fraction of the number of added websites up to time τ to the total number of websites up to time t . Thus we have

$$P(t_i \geq \tau) = 1 - P(t_i \leq \tau) = 1 - e^{-\lambda(t-\tau)}, \quad (11)$$

where λ is the growth rate of the number of websites. With the above, we get,

$$P(X_i(t) \leq x) = 1 - (m_0/x)^\lambda, \quad (12)$$

which yields

$$P(x) = \frac{\partial P(X_i(t) < x)}{\partial x} \propto x^{-(1+\lambda)}, \quad (13)$$

from which the exponent of the power law distribution can be obtained as $\beta = 1 + \lambda$. This result is consistent with the simulation results that are shown in Fig. 1 in which we obtained $\beta \approx 1.5$ with $\lambda = 0.5$.

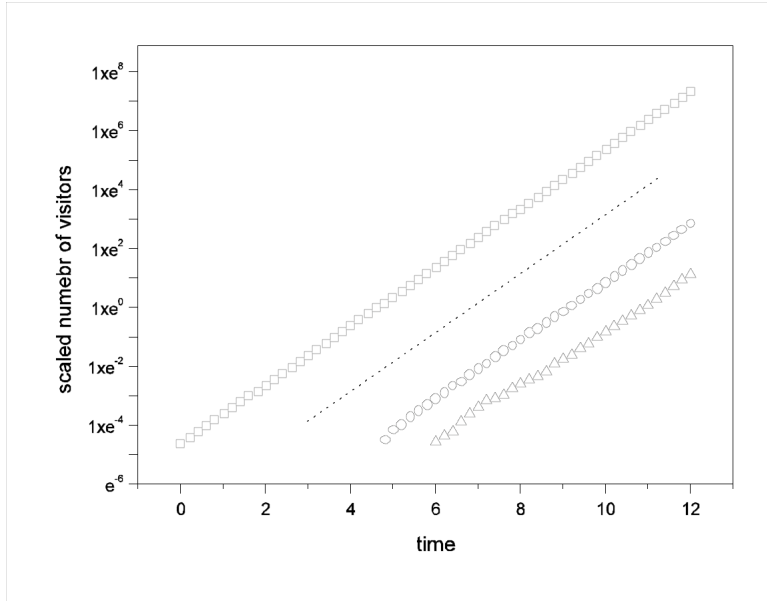


Figure 3 : Time evolution of the number of visitors $X_i(t)$ for websites $i = 10$ (\square), 50 (\circ), and 100 (\triangle) added to the system. The ordinate is in the logarithmic scale with the natural base and the dotted line has slope 1.0

From the above result, we infer that the exponent in the power law distribution depends only on the growth rate λ : the higher is the growth rate, the greater the exponent. This relationship between β and λ also explains dependence of the exponent on the category of the websites that are observed in the empirical study [1]. In Ref. [1] it was found that for the .edu category the power law exponent $\beta = 1.45$, while for all categories the exponent $\beta = 2.07$. Since the growth rate λ of all categories is greater than that of one specific category (.edu category for instance), the exponent for overall websites should be greater. This explains why the exponent for overall websites is bigger than that for .edu sites.

4 Summary and Conclusion

In this paper we investigated the origin of the empirically observed power law distribution of the number of visitors to websites. In order to explain the characteristics of the websites, we established a stochastic dynamic model, which includes the following: the growth of an individual website, the external forces acting on each website, and the exponential growth of the number of websites. With the model, we were able to show most of the characteristics of the dynamics of the websites, such as power law distributions of the number of visitors to websites and the fluctuation in the individual website's growth. Moreover, we found that the

exponential growth rate λ of the number of websites determines the exponent β in the power law distribution: the higher the growth rate, the bigger the exponent. We also performed an analytic calculation and compared the result with that of the numerical simulations. Within the approximation we formulated the exponent in terms of the growth rate λ and confirmed the simulation results.

Thus the key ingredients in the dynamics of the websites are the following. First, there is a global interaction in terms of the stochastic force strength among websites with which one can view the web ecology as a competitive complex system. Second, the web ecological system stays in non-equilibrium in the sense that the number of the websites in the system is not fixed but exponentially increased. These two ingredients in the web ecological system lead to the characteristics of the system.

Needless to say, this approach is not the unique way to explain the power law nature of the dynamics of the websites. Other approaches that lead to the same characteristics of the dynamics of the websites are possible and one candidate model might be the one in which the interaction among websites are included. This could be one of the further directions of research in this field. This work was supported by Grant No. R02-2000-00292 from the Korea Science & Engineering Foundation (KOSEF).

Bibliography

- [1] L. Adamic and B. Huberman, 2000, Quarterly Journal of Electronic Commerce, 1, 5.
- [2] Source for the exponential growth of the websites are from the World Wide Web Consortium, Mark Gray, Netcraft Server Survey and can be obtained at <http://www.w3.org/Talks/1998/10/WAP-NG-Overview/slide10-3.html>.
- [3] It is known in Ref. [2] that between August of 1992 and August 1995, the number of web servers increases 100 times for every 18 months, and between August 1995 and February 1998, 10 times every 30 months.
- [4] R. Albert, H. Jeong, and A.-L. Barabasi, 1999, Nature, 401, 130.
- [5] A.-L. Barabasi and R. Albert, 1999, Science, 286, 509.