PHYSICAL REVIEW E, VOLUME 65, 036132

Dynamic model for the popularity of websites

Chang-Yong Lee[1],* and Seungwhan Kim[2]
[1]*Department of Industrial Information, Kongju National University, Chungnam 340-800, South Korea*
[2]*Basic Research Laboratory, Electronics and Telecommunications Research Institute, Taejon 305-350, South Korea*

(Received 9 October 2001; published 1 March 2002)

In this paper, we have studied a dynamic model to explain the observed characteristics of websites in the World Wide Web. The dynamic model consists of the self-growth term for each website and the external force term acting on the website. With simulations of the model, we can explain most of the important characteristics of websites. These characteristics include a power-law distribution of the number of visitors to websites, fluctuation in the fractional growth of individual websites, and the relationship between the age and the popularity of the websites. We also investigated a few variants of the model and showed that the ingredients included in the model adequately explain the behavior of the websites.

DOI: 10.1103/PhysRevE.65.036132 PACS number(s): 02.50.−r, 89.20.Hh, 89.75.Da

I. INTRODUCTION

The information era is emerging from the Internet and the World Wide Web. As this new era emerges, many aspects of social, economic, and academic life are rapidly changing. In particular, the World Wide Web (hereafter referred to as ‘‘the web’’), an important application on the Internet, is no longer confined to researchers in universities and research institutes. Its usage has been extended to many fields of interest, such as, education, advertisement, and especially electronic commerce. Due to the development of the network technologies and the reduced cost for information exchange via the Internet, it is expected that the usage of the Internet will be accelerated in the future. This expectation of accelerated usage has manifested itself through exponential increases in the number of computers as well as websites connected to the Internet [1].

As the Internet plays an important role in our present society, research on the Internet becomes more and more active. In particular, the study of the characteristics of websites and their dynamical phenomena has become recognized as a new field of research. Aside from the technical understandings of the Internet and the web, this new field can be regarded as an ‘‘artificial ecological system’’ of which many interacting agents, or websites, are composed. As is true for most complex systems, size and dynamical 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 [2] based on the web search engines, the probability distribution of the size of websites, that is, the number of pages per site, follows a power-law distribution. This power-law distribution implies that most of the websites have few pages, while only a few websites have many pages.

Another important finding is that the number of visitors to websites also exhibits a power-law distribution [3]. This finding suggests that most of the 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. The Internet web market contrasts with the traditional equal share markets in which the transaction cost and geological factors play important roles. Moreover, the analysis of empirical data shows a few additional characteristics for the number of visitors to websites: first, the distribution 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 websites is greater than that of websites in specific categories. Second, the fluctuation of the growth rate in the number of visitors for each site is uncorrelated. Third, the older websites do not necessarily have more visitors than younger ones. That is, site popularity and age are only slightly correlated.

Upon these observations, a study has been carried out by Adamic and Huberman [3] to explain these characteristics. In their study, they used a stochastic growth model and derived the power-law distribution. In the proposed model, however, the number of websites is not exponentially increasing but constant in time. Since the number of websites is in reality growing exponentially, this is an important factor in the dynamics of the websites. Thus, it is desirable to have a model that includes the exponential growth in the number of websites.

The web ecological system is subject to complex influence among agents and may belong to a dynamic system that can be modeled statistically. When the system is treated statistically, the predictable dynamics of the system, subject to all possible influence, can be handled via a stochastic process. The price we pay for introducing the stochastic factors into the system is the sacrifice of complete predictability. Unlike thermodynamics, which describes macroscopic system without the structure or dynamics of the system, the web ecological system is formed out of not only interacting, but also interdependent parts. In the dynamics of the system, the fine-scale details may influence large-scale behavior.

Complex systems, in general, are composed of interde-

*Email address: clee@kongju.ac.kr

pendent agents that have many degrees of freedom, whose time dependence can be very slow on a microscopic scale. While the microscopic dynamics of the system is rapidly changing and complex, the macroscopic behavior of the system can be described to be as simple or even static. The origin of this simplicity stems from an average over the fast microscopic variable on the time scale of the macroscopic observation.

In this paper, we investigate the characteristics of the number of visitors to websites and the dynamical properties among competing websites by establishing a theoretical model and simulations with it. In particular, we focus on the result of empirical data analysis carried out in Ref. [3]. In doing so, we first build up a stochastic model for the number of visitors to the websites, and then carried out both numerical and analytic calculations.

This paper is organized as follows. In Sec. II, we establish a model from general principles of the websites dynamics. This is followed by the results of numerical simulation and approximated analytic calculations in Sec. III. A few possible variants of the model and their results are discussed in Sec. IV and the last section is devoted to the summary and conclusions.

II. MODEL

In general, a dynamical system can be described schematically as

$$\frac{dX_i(t)}{dt}=f_i(\vec{X}), \tag{1}$$

where $\vec{X}$ represents the state of the system and takes values in the state or phase space. In the present case, $\vec{X}=\{X_i(t)\}, i=1,2,\ldots,N(t)$, and $X_i(t)$ is the number of visitors to the website *i* at time *t*. The form of the $f_i(\vec{X})$ can be expanded, in the powers of X_i's, as

$$f_i(\vec{X})=a_i+\sum_j A_{ij}X_j+\sum_{lm} B_{ilm}X_lX_m+\cdots, \tag{2}$$

where a_i is some constant, and A_{ij} and B_{ilm} are appropriate coefficients. Assuming that the lowest-order term in X_i plays the most important role in the dynamics, we take the first nonconstant term in Eq. (2). After absorbing the constant term a_i into X_i, Eq. (1) can be rewritten as

$$\frac{dX_i}{dt}=A_{ii}X_i+\sum_{j\neq i} A_{ij}X_j. \tag{3}$$

Before trying to determine coefficients A_{ii} and A_{ij}, it is important to take the increase of the number of the websites into account, for it is known that the number of websites connected to the Internet is not constant but increases exponentially [4]. Since the number of websites $N(t)$ at time *t* grows exponentially in time, $N(t)$ satisfies, in the continuous time limit,

$$\frac{dN(t)}{N(t)}=\lambda\, dt, \tag{4}$$

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

$$\Delta N(t)=N(t+\Delta t)-N(t)=N(t)\lambda\Delta t. \tag{5}$$

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. (3) is the growth term of the website *i* with the growth rate A_{ii}. The implication of this term is that once a website is created the website will be known to more users, in consequence, more users visit the website as time progresses. As a result, in two successive time periods the increase in the number of visitors is proportional to the number of visitors to that site. 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 is no other factor affecting the growth of a website. Furthermore, the coefficient A_{ii} can be absorbed with an appropriate rescaling of X_i so that one can set $A_{ii}=1$ for all *i*.

The second term on the right-hand side of Eq. (3) can be regarded as an ‘‘external force’’ acting on the website *i*. The coefficient A_{ij} should include the following. First of all, the force has to be global, that is, the website *i* experiences a force from all the other websites. Since websites distributed over the Internet can be accessed by a few clicks of a button [5], accessing a website does not depend on the geographical degree of freedom. The result is that there is no spatial limitation. Second, the force 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

$$A_{ij}\rightarrow\langle A\rangle+\kappa\,\eta_{ij}(t), \tag{6}$$

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

$$\langle\,\eta_{ij}(t)\rangle=0 \quad\text{and}\quad \langle\,\eta_{ij}(t)\,\eta_{kl}(s)\rangle=\delta(t-s)\,\delta_{ik}\delta_{jl}. \tag{7}$$

Note that we set $\langle A\rangle$, time averaged strength, to zero for simplicity. One more ingredient that we take into consider-

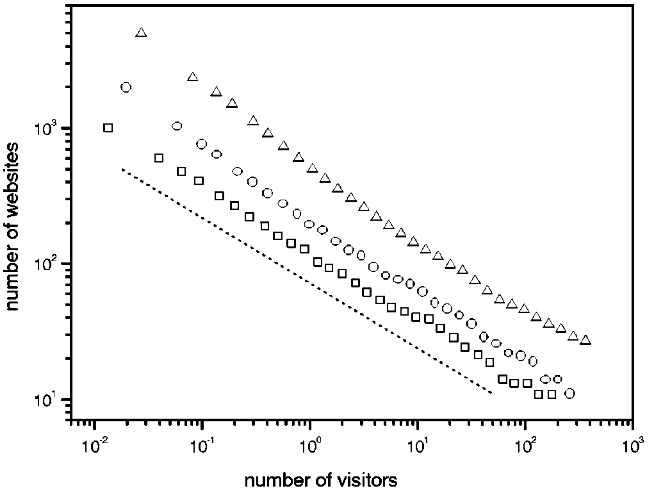

FIG. 1. Log-log scale plots of cumulative distribution functions of the number of visitors for the total number of websites at the end of simulations $N_{total}=1000$ (□), $N_{total}=2000$ (○), and $N_{total}=5000$ (△). The dotted line has slope -0.5, and all quantities are dimensionless.

ation is the normalization factor. We 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 $\kappa \rightarrow \kappa/N(t)$.

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

$$\frac{\Delta X_i}{\Delta t}=X_i+\frac{\kappa}{N(t)}\sum_{j\neq i}^{N(t)}\eta_{ij}(t)X_j, \tag{8}$$

where $\Delta X_i(t)=X_i(t+\Delta t)-X_i(t)$, and $N(t)$ satisfies Eq. (4). From this 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.

III. SIMULATION RESULTS

With the dynamic equation of Eq. (8), we perform numerical simulations. In the simulation, we start with a small number of websites [say, $N(0)=10$] and at every time step Δt, a new website is added to the system with the probability $N(t)\lambda\Delta t$, allowing interaction with websites already present in the system.

Figure 1 shows cumulative distribution functions (CDF) of the number of visitors to websites with different number of websites N_{total}, which is the total number of websites at the end of each simulation. That is, we carry out each simulation until $N(t)=N_{total}$ is satisfied. 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 can get a power-law distribution as CDF, $C(x)\propto x^{-\alpha}$, with $\alpha\approx 0.5$. Since the empirical exponents were obtained from probability density function (PDF), $P(x)$ of the number of visitors to the websites, we need to differentiate CDF with respect to *x* to get PDF. That is, $P(x)=dC(x)/dx$. Thus, we obtain a power-law probability distribution, $P(x)\propto x^{-\beta}$ with an exponent $\beta\approx 1.5$. One can 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}.

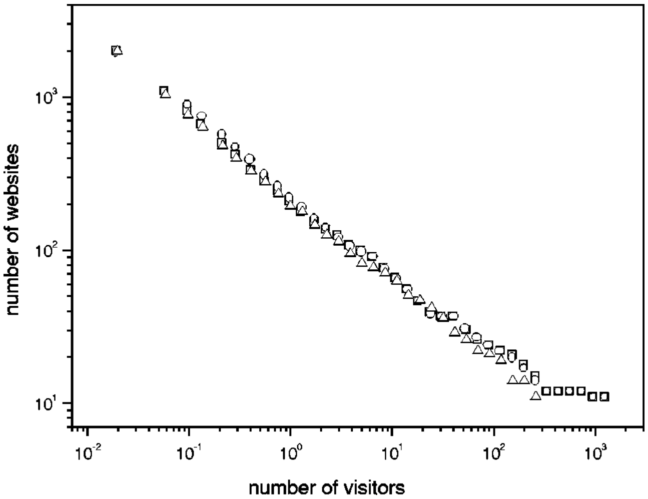

FIG. 2. Log-log scale plots of cumulative distribution functions of the number of visitors for $\kappa=0.5$ (□), $\kappa=1.0$ (○), and $\kappa=2.0$ (△). All quantities are dimensionless.

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 of the power law does not depend on the force strength κ.

The force term in the model is responsible for the fluctuation of the number of visitors to websites. It is found in Ref. [3] that the fractional fluctuation in the number of visitors for a given website are uncorrelated to each other. To show this, we plot in Fig. 3 the quantity

$$g(t)\equiv\frac{X(t+\Delta t)-X(t)}{X(t)}, \tag{9}$$

as a function of the time. This random fluctuation of the fractional growth can be verified in terms of the autocorrelation. The calculation of the autocorrelation 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.

One important question in the dynamics of the web, as well as other networks in general, is the correlation between age and popularity. That is, the correlation between the time at which a website is created and the number of visitors to the website. It is argued in Refs. [3,6] that age and site popularity are only slightly correlated, while there is an evidence that correlated tendency does exist once popularity is aver-

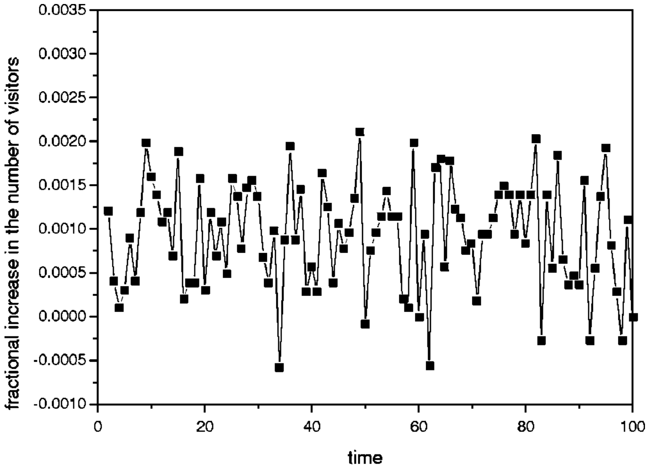

FIG. 3. Fractional fluctuations in the number of visitors as a function of time. All quantities are dimensionless.

aged over the same age [7]. To investigate the correlation between age and popularity in our model, we calculate Spearman's nonparametric statistics r_s [8]. The statistics is a rank-correlation method and is essentially the sum of the squared difference between each pair of ranks, which consists of ranks of age and its popularity. This statistics takes values $-1 \leqslant r_s \leqslant 1$: $r_s = 1$ being the total correlation, $r_s = -1$ being the total anticorrelation, and $r_s = 0$ being no rank correlation. We calculate r_s for various values of κ with $\lambda = 0.5$ and $N_{total} = 1000$. From the results shown in Fig. 4, it is easy to see that bigger the force strength κ, less the correlation will be. Following this and the above results, one can conclude that the force strength plays an important role not only in the fluctuation of the number of visitors, but also in the correlation between age and popularity.

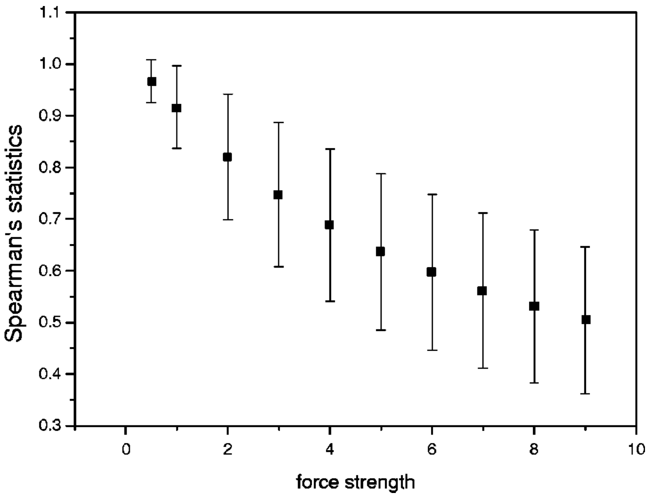

FIG. 4. Spearman's statistics r_s as a function of the force strength κ. Error bars are the standard deviation of ten independent trials. All quantities are dimensionless.

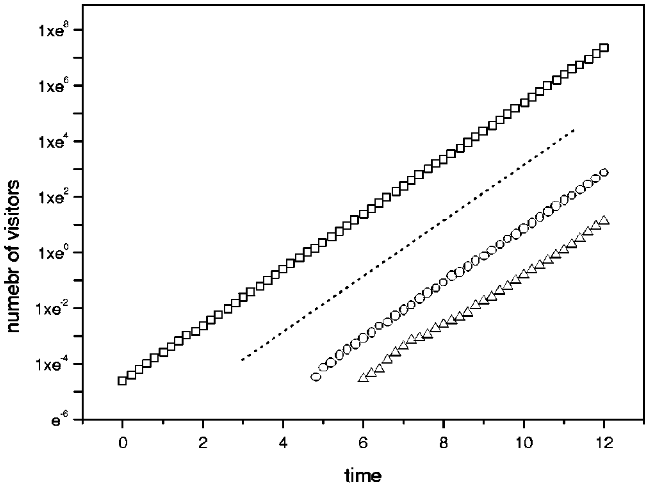

FIG. 5. Time evolution of the number of visitors $X_i(t)$ for websites $i=10$ (□), 50(○), and 100(△) added to the system. The ordinate is in the logarithmic scale with the natural base and the dotted line has slope 1.0. All quantities are dimensionless.

The power-law distribution observed in Figs. 1 and 2 can be derived analytically with an appropriate approximation. Following the procedure similar to Ref. [9], we plot in Fig. 5 the number of visitors $X_i(t)$ to various websites as a function of the time with parameters $N_{total} = 2000$, $\kappa = 1.0$, and $\lambda = 0.5$. From Fig. 5 one can obtain an approximate differential equation for X_i as

$$\frac{\partial \ln X_i(t)}{\partial t} \approx \alpha, \quad \text{or,} \quad \frac{\partial X_i(t)}{\partial t} \approx X_i, \tag{10}$$

where α is estimated from Fig. 5 as $\alpha \approx 1$.

Comparing Eq. (10) with Eq. (8), one can see that the force term in Eq. (8) just introduces fluctuation of the number of visitors and plays little role in the growth dynamics of the number of visitors to the websites. The solution of Eq. (10) is given as

$$X_i(t) = m_0 e^{(t-t_i)}, \tag{11}$$

where $m_0 = X_i(0)$, and t_i is the time at which the website i is added to the system. Equation (11) 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 [9].

The probability that a website i has visitors smaller than x, $P(X_i(t) < x)$, can be written as $P(t_i > \tau)$, where $\tau = t - \ln(x/m_0)$. Note that $P(t_i \leqslant \tau)$ is the probability that the website i can be found in the system up to time τ, and the number of websites increases exponentially with the rate λ. 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

$$P(t_i > \tau) = 1 - P(t_i \leqslant \tau) = 1 - e^{\lambda(\tau - t)}, \tag{12}$$

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

$$P(X_i(t) < x) = 1 - (m_0/x)^{\lambda}, \tag{13}$$

which yields

$$P(x)=\frac{\partial P(X_i(t)<x)}{\partial x}\propto x^{-(1+\lambda)}, \tag{14}$$

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

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 [3]. In Ref. [3] 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.

IV. VARIANTS OF THE MODEL

The characteristics of the dynamics of the websites we explored, indicate that the exponential growth of the number of websites and the stochastic nature of the external force play an important role in the websites dynamics. To show that these ingredients are necessary, we investigated a few variants of the model.

The first variant replaces the stochastic noise $\eta(t)$ with quenched one. With this variant, even though we still obtained the power law, no fractional fluctuation in the growth rate is seen in the dynamics of the websites. In addition, with the quenched noise, one finds a strong correlation between age and popularity. That is, older websites necessarily get more visitors.

The second variant of the model includes a linear growth of the number of websites in time. To implement this, we start with a small number of the websites (say, $m_0=10$), and at every time step Δt we add a new website allowing interaction with the websites already present in the system. Thus after $n\Delta t$ time step, we have m_0+n websites in the system. Figure 6 shows the distribution of the number of visitors to websites with $\kappa=2.0$ and $N_{total}=2000$. As can be seen from Fig. 6, we have a power-law distribution with exponent $\beta \approx 1$. Note also that this result is independent of κ as we expect.

To derive the above exponent analytically, we carry out a procedure similar to the one in the preceding section. We plot the number of visitors $X_i(t)$ to different websites as a function of the time and the result is shown in Fig. 7. From Fig. 7, the logarithmic rate at which a website i acquires visitors can be expressed approximately as

$$\frac{\partial \ln X_i}{\partial t}\approx \alpha, \tag{15}$$

where α is estimated as $\alpha\approx 1$. The solution of the equation again becomes

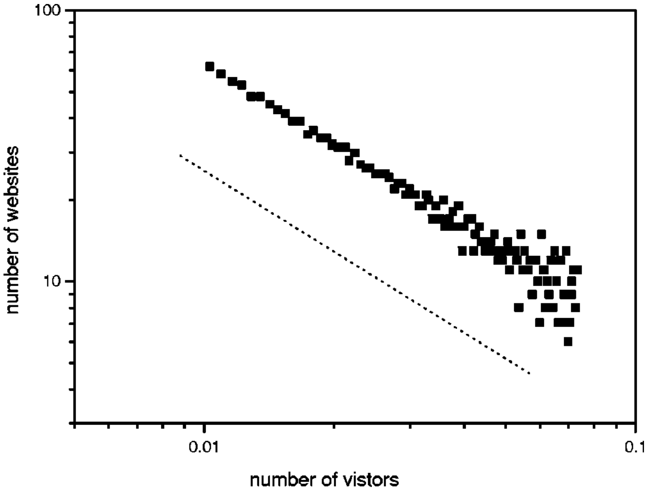

FIG. 6. Log-log scale plots of the distributions of the number of visitors. The dotted line has slope -1. All quantities are dimensionless.

$$X_i(t)=m_0 e^{\alpha(t-t_i)}, \tag{16}$$

where $m_0=X_i(0)$ and t_i is the time at which the website i is added to the system. Following the same argument as the one in the preceding section and taking into account the fact that the number of websites added to the system is proportional to the number of time step, we have

$$P(t_i>\tau)=1-\tau/t=\ln(x/m). \tag{17}$$

This yields

$$P(x)=\frac{\partial P(X_i(t)<x)}{\partial x}\propto x^{-1}. \tag{18}$$

From the above analysis, one finds that even though the linear growth model give rise to a power-law distribution, the

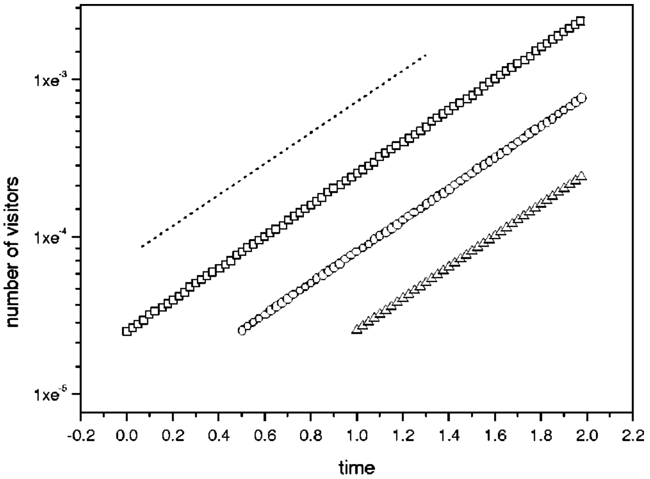

FIG. 7. Time evolution of the number of visitors $X_i(t)$ for websites $i=10$ (□), 500 (○), and 1000 (△) added to the system. The ordinate is in the logarithmic scale of the natural base and the dotted line has slope 1.0. All quantities are dimensionless.

exponent is fixed with $\beta=1$ and cannot account for the different values of the exponent for different categories of the websites.

V. 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 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. Using this 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. It was also found that although some variants of the model are quite possible, these did not exhibit any observed characteristics of the websites. 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 nonequilibrium 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, that 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 possible directions of research in this field.

ACKNOWLEDGMENT

This work was supported by Grant No. R02-2000-00292 from the Korea Science & Engineering Foundation (KOSEF).

[1] The source for the exponential growth of the websites is 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

[2] B. Huberman and L. Adamic, Nature (London) **401**, 131 (1999).

[3] L. Adamic and B. Huberman, QJEC **1**, 5 (2000).

[4] It is known in Ref. [1] that between August of 1992 and August 1995, the number of web servers increased 100 times for every 18 months, and between August 1995 and February 1998, 10 times every 30 months.

[5] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).

[6] L. Adamic and B. Huberman, Science **287**, 2115 (2000).

[7] A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi, Science **287**, 2115 (2000).

[8] See, for example, L. Chao, *Statistics: Methods and Analyses* (McGraw-Hill, New York, 1969), Chap. 17.

[9] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).