W. A. Fowler, *Science* **143**, 465 (1964).
24. J. X. Mitrovica, *J. Geophys. Res.* **101**, 555 (1996).
25. B. H. Hager and M. A. Richards, *Philos. Trans. R. Soc. London Ser. A* **328**, 309 (1989).
26. This CMB response time is accelerated by a factor of 2, relative to a model without core heating.
27. The correspondence of the mantle response time and the time period for reliable reconstructions is not coincidental. Reconstructions are largely dependent on the magnetic isochrons of sea-floor spreading, which is limited to the characteristic maximum age for oceanic lithosphere. Simple boundary-layer convection theory (*22*) predicts that this characteristic boundary-layer age should be similar to the vertical transit time for convection. Based on the prediction that advection velocities scale roughly with the natural logarithm of the viscosity contrast (*18*), the mantle response time should be accurate within a factor of 2.
28. H.-P. Bunge and M. A. Richards, *Geophys. Res. Lett.* **23**, 2987 (1996).
29. H.-P. Bunge *et al.*, data not shown.
30. R. G. Gordon and D. M. Jurdy, *J. Geophys. Res.* **91**, 12389 (1986).
31. C. Lithgow-Bertelloni and M. A. Richards, *Rev. Geophys.* **36**, 72 (1998).
32. We thank G. Davies and R. van der Hilst for constructive reviews, J. Painter for supporting the 3D graphics, and the Los Alamos Branch of the Institute of Geophysics and Planetary Physics for continuing support. Computing resources were provided by the Advanced Computing Laboratory of Los Alamos National Laboratory.

27 October 1997; accepted 18 February 1998

Strong Regularities in World Wide Web Surfing

Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, Rajan M. Lukose

One of the most common modes of accessing information in the World Wide Web is surfing from one document to another along hyperlinks. Several large empirical studies have revealed common patterns of surfing behavior. A model that assumes that users make a sequence of decisions to proceed to another page, continuing as long as the value of the current page exceeds some threshold, yields the probability distribution for the number of pages that a user visits within a given Web site. This model was verified by comparing its predictions with detailed measurements of surfing patterns. The model also explains the observed Zipf-like distributions in page hits observed at Web sites.

The exponential growth of the World Wide Web is making it the standard information system for an increasing segment of the world's population. The Web allows inexpensive and fast access to unique and novel services, including electronic commerce, information resources, and entertainment, provided by individuals and institutions scattered throughout the world (*1*). Despite the advantages of this new medium, the Internet still fails to serve the needs of the user community in a number of ways. Surveys of Web users find that slow access and inability to find relevant information are the two most frequently reported problems (*2*). The slow access is at least in part a result of congestion (*3*), whereas the difficulty in finding useful information is related to the balkanization of the Web structure (*4*). Because it is difficult to solve this fragmentation problem by designing an effective and efficient classification scheme, an alternative approach is to seek regularities in user patterns that can then be used to develop technologies for increasing the density of relevant data for users.

A common way of finding information on the Web is through query-based search engines, which enable quick access to information that is often not the most relevant. This lack of relevance is partly attributable to the impossibility of cataloging an exponentially growing amount of information in ways that anticipate users' needs. But because the Web is structured as a hypermedia system, in which documents are linked to one another by authors, it also supports an alternative and effective mode of use in which users surf from one document to another along hypermedia links that appear relevant to their interests.

Here, we describe several strong regularities of Web user surfing patterns discovered through extensive empirical studies of different user communities. These regularities can be described by a law of surfing, derived below, that determines the probability distribution of the depth—that is, the number of pages a user visits within a Web site. In conjunction with a spreading activation algorithm, the law can be used to simulate the surfing patterns of users on a given Web topology. This leads to accurate predictions of page hits. Moreover, it explains the observed Zipf-like distributions of page hits to Web sites (*5*).

We start by deriving the probability $P(L)$ of the number of links L that a user follows in a Web site. This can be done by considering that there is value in each page a user visits, and that clicking on the next page assumes that it will be valuable as well. Because the value of the next page is not certain, we can assume that it is stochastically related to the previous one. In other words, the value of the current page V_L is the value of the previous one V_{L-1} plus or minus a random term. Thus, the page values can be written as

$$V_L = V_{L-1} + \varepsilon_L \qquad (1)$$

where the values ε_L are independent and identically distributed Gaussian random variables. A particular sequence of page valuations is a realization of a random process and thus is different for each user. Within this formulation, an individual will continue to surf until the expected cost of continuing is perceived to be larger than the discounted expected value of the information to be found in the future. This can be thought of as a real option in financial economics, for which it is well known that there is a threshold value for exercising the option to continue (*6*, *7*). Even if the value of the current page is negative, it may be worthwhile to proceed, because a collection of high-value pages may still be found. If the value is sufficiently negative, however, then it is no longer worth the risk to continue. That is, when V_L falls below some threshold value, it is optimal to stop.

The number of links a user follows before the page value first reaches the stopping threshold is a random variable L. For the random walk of Eq. 1, the probability distribution of first passage times to a threshold is given asymptotically by the two-parameter inverse Gaussian distribution

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} \exp\left[\frac{-\lambda(L-\mu)^2}{2\mu^2 L}\right] \qquad (2)$$

(*8*), with mean $E(L) = \mu$ and variance $\mathrm{Var}(L) = \mu^3/\lambda$, where λ is a scale parameter. This distribution has two characteristics worth stressing in the context of user surfing patterns. First, it has a very long tail, which extends much farther than that of a normal distribution with comparable mean and variance. This implies a finite probability for events that would be unlikely if described by a normal distribution. Consequently, large deviations from the average number of user clicks computed at a site will be observed. Second, because of the asymmetry of the distribution function, the typical behavior of users will not be the same as their average behavior. Thus, because the mode is lower than the mean, care must be exercised with available data on the average number of clicks, as this average overestimates the typical depth being surfed.

To test the validity of Eq. 2, we analyzed data collected from a representative sample

Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA.

of America Online (AOL) Web users. For each of 5 days (29 and 30 November and 1, 3, and 5 December 1997), the entire activity of one of AOL's caching proxies was instrumented to record an anonymous but unique user identifier, the time of each URL (uniform resource locator) request, and the requested URL. For comparison with the predicted distribution, a user who starts surfing at a particular site, such as www.sciencemag.org, is said to have stopped surfing after L links as soon as he or she requests a page from a different Web site. For this analysis, if the user later returned to that site, a new length count L was started. Requests for embedded media (such as images) were not counted.

On 5 December 1997, the 23,692 AOL users in our sample made 3,247,054 page requests from 1,090,168 Web sites. The measured cumulative distribution function (CDF) of the depth L for that day is shown in Fig. 1. Superimposed is the predicted function from the inverse Gaussian distribution fitted by the method of moments (8). To test the quality of the fit, we analyzed a quantile-quantile against the fitted distribution. Both techniques, along with a study of the regression residuals, confirmed the strong fit of the empirical data to the theoretical distribution. The fit was significant at the $P < 0.001$ level and accounted for 99% of the variance. Although the average number of pages surfed at a site was almost three, users typically requested only one page. Other AOL data from different dates showed the same strength of fit to the inverse Gaussian with nearly the same parameters.

For further confirmation of the model, we considered the simplest alternative hypothesis, in which a user at each page conducts an independent Bernoulli trial to make a stopping decision. This led to a geometric distribution of depths, which was found to be a poor fit to the data.

We also examined the navigational patterns of the Web user population at Georgia Institute of Technology for a period of 3 weeks, starting on 3 August 1994. The data were collected from an instrumented version of the National Center for Supercomputing Applications' Xmosaic that was deployed across the students, faculty, and staff of the College of Computing (9). One hundred and seven users (67% of those invited) chose to participate in the experiment. The instrumentation of Xmosaic recorded all user interface events. Of all the collected events, 73% were navigational, resulting in 31,134 page requests. As with the AOL experiment, the surfing depth of users was calculated across all visits to each site for the duration of the study. For the combined data, the mean number of clicks was 8.32 and the variance was 2.77. Comparison of the quantile-quantile, the CDF, and a regression analysis of the observed data against an inverse Gaussian distribution of same mean and variance confirmed the ability of the law of surfing to fit the data ($R^2 = 0.95$, $P < 0.001$). Hence, the model was able to fit surfing behavior with data sets from diverse communities of users, several years apart, who used different browsers and connection speeds.

An interesting implication of the law of surfing can be obtained by taking logarithms on both sides of Eq. 2, which yields

$$\log P(L) = -\frac{3}{2}\log L - \frac{\lambda(L-\mu)^2}{2\mu^2 L} + \log\left(\sqrt{\frac{\lambda}{2\pi}}\right) \quad (3)$$

That is, a log-log plot shows a straight line whose slope approximates 3/2 for small values of L and large values of the variance. As L gets larger, the second term provides a downward correction. Thus, Eq. 3 implies that, up to a constant given by the third term, the probability of finding a group surfing at a given level scales inversely in proportion to its depth, $P(L) \propto L^{-3/2}$. This Pareto scaling relation was verified by plotting the available data on a logarithmic scale. Figure 2 shows that the inverse proportionality holds well over a range of depths.

The previous data validated the law of surfing for a population of users who had no constraints on the Web sites they visited. We also considered the case of surfing within a single large Web site, which is important from the point of view of site design. The site used was the Xerox Corporation's external Web site (www.xerox.com). During the period 23 to 30 August 1997, the Xerox site consisted of 8432 HTML documents and received an average of 165,922 requests per day. The paths of individual users were reconstructed by a set of heuristics that used unique identifiers ("cookies"), when present, or otherwise used the topology of the site along with other information to disambiguate users behind proxies. Automatic programs that request the entire contents of the site ("spiders") were removed from the analysis. Additionally, a stack-based history mechanism was used to infer pages cached either by the client or by intermediary caches. This resulted in a data set consisting of the full path of users and the number of clicks performed at the Xerox Web site.

Figure 3 shows the CDF plot of the Xerox Web site for 26 August 1997 against the fitted inverse Gaussian defined by Eq. 2. The mean number of clicks was 3.86, with a variance of 6.08 and a maximum of 95 clicks. As with the client path distributions, both the quantile-quantile and the CDF plots of the site data showed a strong fit to Eq. 2. Moreover, these results were very consistent across all the days in the study.

We next describe how Eq. 2 (in conjunction with a spreading activation algorithm) can predict the number of hits for each page in a Web site, a quantity of interest in electronic commerce. Spreading activation refers

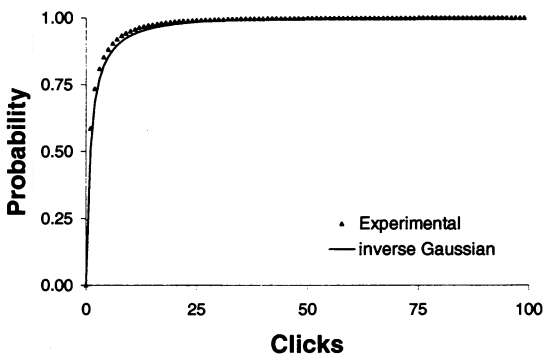

Fig. 1. The CDF of AOL users as a function of the number of surfing clicks. The observed data were collected on 5 December 1997 from a representative sample of 23,692 AOL users who made 3,247,054 clicks. The fitted inverse Gaussian distribution has a mean of $\mu = 2.98$ and $\lambda = 6.24$.

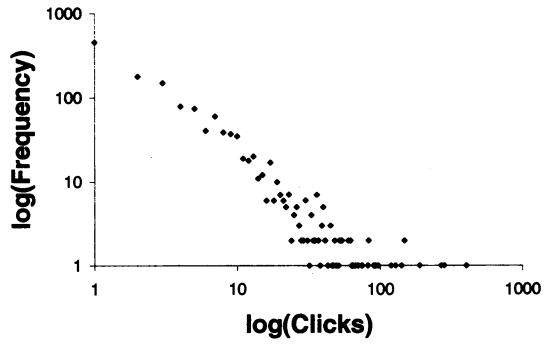

Fig. 2. The frequency distribution of surfing clicks on log-log scales. Data collected from the Georgia Institute of Technology, August 1994.

to a class of algorithms that propagate numerical values (or activation levels) among the connected nodes of a graph (*10*). Their application ranges from models of human memory (*11*) and semantics (*12*) to information retrieval (*13*). In the context of the Web, the nodes correspond to pages and the arcs to the hyperlinks among them, so that spreading activation simulates the flow of users through a Web site.

Consider a collection of n Web pages, each indexed by $i = 1, 2, \ldots, n$, and connected by hyperlink edges to form a graph. The surfing activity of users can be simulated by assigning a weight, $S_{j,i}$, between the ith and jth node. This weight can be interpreted as the fraction of continuing users at node i who proceed to node j if $\Sigma_j S_{j,i} = 1$, where the sum is over all the nodes to node i by an edge. Let f_L be the fraction of users who, having surfed along $L - 1$ links, continue to surf to depth L. If the activation value $N_{i,L}$ is defined as the number of users who are at node i after surfing through L clicks, the resulting expression is

$$N_{i,L+1} = f_L \sum_j S_{i,j} N_{j,l} \qquad (4)$$

The fraction f_L is derived from Eq. 2; that is, in a group of users f_L is equal to the ratio of the number of users who surf for L or more links to the number who surf for $L - 1$ or more links. In terms of $F(L,\mu,\lambda)$, the CDF of the inverse Gaussian, it is given by

$$f_L = \frac{1 - F(L,\mu,\lambda)}{1 - F(L-1,\mu,\lambda)} \qquad (5)$$

With this definition, Eq. 4 can be iterated from initial conditions $N_{i,1}$. After most of the surfers have stopped, the predicted aggregate number of hits at each page is simply the sum over all iterations for each page.

Figure 4 shows the observed and predicted daily average number of hits per page for the Xerox corporate Web site data described above. Equation 4 was initialized with estimates from the data of the number of users who started surfing at each page i, $N_{i,1}$, and the proportion of users who surfed from each page i to connected pages j, $S_{i,j}$. We used the inverse Gaussian estimated in Fig. 3 for the Xerox site to compute f_L and iterated Eq. 4 for $L = 15$ levels of surfing. Figure 4 shows that the hits predicted by spreading activation are highly correlated with the observed hits ($r = 0.995$).

This algorithm can also be used for a number of interesting Web applications. For example, if a Web site were to be reorganized, spreading activation plus the law of surfing could give an indication of the expected usage. Alternatively, it may be possible to automatically reorganize the Web site structure in order to obtain a desired hit pattern.

We also used a spreading activation model to address another universal finding in studies of Web activity, that of a Zipf-like distribution (*14*) in the number of hits per page. We ran spreading activation simulations on random graphs of 100 nodes each, with an average of five links per node, using various initial conditions. The resulting probability distribution of the number of hits received over the collection of pages followed a Zipf's law, in agreement with observed data (*5*).

These results show that surfing patterns on the Web display strong statistical regularities that can be described by a universal law. In addition, the success of the model points to the existence of utility-maximizing behavior underlying surfing. Because of the Web's digital nature and great use, it is relatively easy to obtain online data that could reveal more novel patterns of information foraging. For example, these studies could be extended to determine the relation between the characteristics of different user communities and the law of surfing parameters.

As the world becomes increasingly connected by the Internet, the discovery of new patterns in the use of the Web can throw a timely light on the growth and development of this new medium. This is important because the sheer reach and structural complexity of the Web makes it an ecology of knowledge, with relationships, information "food chains," and dynamic interactions that could soon become as rich as, if not richer than, many natural ecosystems.

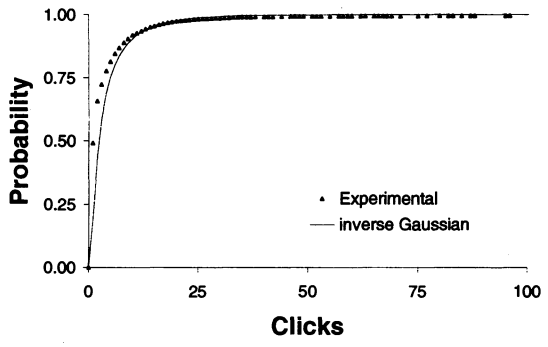

Fig. 3. CDF for the number of users surfing through the Xerox Web site (www.xerox.com) on 26 August 1997 as a function of the number of clicks.

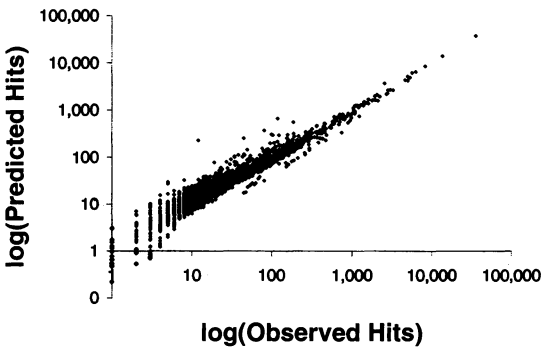

Fig. 4. Histogram of the predicted number of visits per page (hits) to the Xerox Web site versus the observed number of visits generated by spreading activation simulations on a log-log scale.

REFERENCES AND NOTES

1. Special Issue on the Internet, *Sci. Am.* **276** (March 1997).
2. J. E. Pitkow and C. M. Kehoe, *GVU's WWW User Surveys* [online] (1997). Available at www.gvu.gatech.edu/user_surveys.
3. B. A. Huberman and R. M. Lukose, *Science* **277**, 535 (1997).
4. M. Van Alstyne and E. Brynjolfsson, *ibid.* **274**, 1479 (1996).
5. S. Glassman, *Comput. Networks ISDN Syst.* **27**, 165 (1994).
6. A. K. Dixit and R. S. Pindyck, *Investment Under Uncertainty* (Princeton Univ. Press, Princeton, NJ, 1994).
7. R. M. Lukose and B. A. Huberman, paper to be presented at the Fourth International Conference on Computational Economics, Cambridge, UK (1998).
8. V. Seshardri, *The Inverse Gaussian Distribution* (Clarendon, Oxford, 1993).
9. L. Catledge and J. Pitkow, *Comput. Networks ISDN Syst.* **27**, 1065 (1995).
10. J. Shrager, T. Hogg, B. A. Huberman, *Science* **236**, 1092 (1987).
11. J. R. Anderson and P. L. Pirolli, *J. Exp. Psychol. Learn. Mem. Cogn.* **10**, 791 (1984).
12. M. R. Quillan, *Semantic Memory* (Bolt, Beranek, and Newman, Cambridge, MA, 1966).
13. P. Pirolli, J. Pitkow, R. Rao, *Silk from a Sow's Ear: Extracting Usable Structures from the Web*, paper presented at the Conference on Human Factors in Computing Systems, CHI '96, Vancouver, BC, Canada, April 1996.
14. G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA, 1949).

15. Parts of this research were supported by NSF grant IRI-961511 (B.A.H.), Office of Naval Research grant N00014-96-C-0097 (P.L.T.P. and S. Card), and a Citibank fellowship at Stanford University (R.M.L.). We also thank M. Crovella for his comments on a draft of this paper and D. Aronson for providing the AOL data.

14 November 1997; accepted 12 February 1998