1. Chase, M. R., Moller, C., Kesseli, R. & Bawa, K. S. *Nature* **383,** 398–399 (1996).
2. Nason, J. D., Allen Herre, E. & Hamrick, J. L. *Nature* **391,** 685–687 (1998).
3. Crawford, T. J. *Heredity* **52,** 273–283 (1984).
4. Chambers, J. Q., Higuchi, N. & Schimel, J. P. *Nature* **391,** 135–136 (1998).
5. Reboud, X. & Zeyl, C. *Heredity* **72,** 132–140 (1994).
6. McCauley, D. E. *Trends Ecol .Evol.* **10,** 198–202 (1995).
7. Hamilton, M. B. *Mol. Ecol.* **8,** 521–522 (1999).
8. Taberlet, P. *et al. Plant Mol. Biol.* **17,** 1105–1109 (1991).
9. Weir, B. S. *Genetic Data Analysis II* (Sinauer, Sunderland, Massachusetts, 1996).
10. Rand, D. M. *Conserv. Biol.* **10,** 665–671 (1996).

Internet

Diameter of the World-Wide Web

Despite its increasing role in communication, the World-Wide Web remains uncontrolled: any individual or institution can create a website with any number of documents and links. This unregulated growth leads to a huge and complex web, which becomes a large directed graph whose vertices are documents and whose edges are links (URLs) that point from one document to another. The topology of this graph determines the web's connectivity and consequently how effectively we can locate information on it. But its enormous size (estimated to be at least 8×10^8 documents[1]) and the continual changing of documents and links make it impossible to catalogue all the vertices and edges.

The extent of the challenge in obtaining a complete topological map of the web is illustrated by the limitations of the commercial search engines: Northern Light, the search engine with the largest coverage, is estimated to index only 38% of the web[1]. Although much work has been done to map and characterize the Internet's infrastructure[2], little is known about what really matters in the search for information — the topology of the web. Here we take a step towards filling this gap: we have used local connectivity measurements to construct a topological model of the World-Wide Web, which has enabled us to explore and characterize its large-scale properties.

To determine the local connectivity of the web, we constructed a robot that adds to its database all URLs found on a document and recursively follows these to retrieve the related documents and URLs. We used the data collected to determine the probabilities $P_{out}(k)$ and $P_{in}(k)$ that a document has k outgoing and incoming links, respectively. We find that both $P_{out}(k)$ and $P_{in}(k)$ follow a power law over several orders of magnitude, remarkably different not only from the Poisson distribution predicted by the classical theory of random graphs[3,4], but also from the bounded distribution found in models of random networks[5].

The power-law tail indicates that the probability of finding documents with a large number of links is significant, as the network connectivity is dominated by highly connected web pages. Similarly, for incoming links, the probability of finding very popular addresses, to which a large number of other documents point, is non-negligible, an indication of the flocking nature of the web. Furthermore, while the owner of each web page has complete freedom in choosing the number of links on a document and the addresses to which they point, the overall system obeys scaling laws characteristic only of highly interactive self-organized systems and critical phenomena[6].

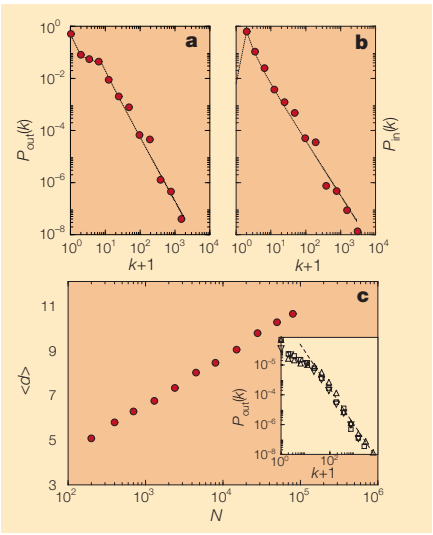

Figure 1 Distribution of links on the World-Wide Web. **a,** Outgoing links (URLs found on an HTML document); **b,** incoming links (URLs pointing to a certain HTML document). Data were obtained from the complete map of the nd.edu domain, which contains 325,729 documents and 1,469,680 links. Dotted lines represent analytical fits used as input distributions in constructing the topological model of the web; the tail of the distributions follows $P(k)\approx k^{-\gamma}$, with $\gamma_{out}=2.45$ and $\gamma_{in}=2.1$. **c,** Average of the shortest path between two documents as a function of system size, as predicted by the model. To check the validity of our predictions, we determined d for documents in the domain nd.edu. The measured $\langle d_{nd.edu}\rangle=11.2$ agrees well with the prediction $\langle d_{3\times10^5}\rangle=11.6$ obtained from our model. To show that the power-law tail of $P(k)$ is a universal feature of the web, the inset shows $P_{out}(k)$ obtained by starting from whitehouse.gov (squares), yahoo.com (triangles) and snu.ac.kr (inverted triangles). The slope of the dashed line is $\gamma_{out}=2.45$, as obtained from nd.edu in **a**.

To investigate the connectivity and the large-scale topological properties of the web, we constructed a directed random graph consisting of N vertices, assigning to each vertex k outgoing (or incoming) links, such that k is drawn from the power-law distribution of Fig. 1a,b. To achieve this, we randomly selected a vertex i and increased its outgoing (or incoming) connectivity to k_i+1 if the total number of vertices with k_i+1 outgoing (or incoming) links is less than $NP_{out}(k_i+1)$ (or $NP_{in}(k_i+1)$).

A particularly important quantity in a search process is the shortest path between two documents, d, defined as the smallest number of URL links that must be followed to navigate from one document to the other. We find that the average of d over all pairs of vertices is $\langle d\rangle=0.35+2.06\log(N)$ (Fig. 1c), indicating that the web forms a small-world network[5,7], which characterizes social or biological systems. For $N=8\times10^8$, $\langle d_{web}\rangle=18.59$; that is, two randomly chosen documents on the web are on average 19 clicks away from each other.

For a given N, d follows a gaussian distribution so $\langle d\rangle$ can be interpreted as the diameter of the web, a measure of the shortest distance between any two points in the system. Despite its huge size, our results indicate that the web is a highly connected graph with an average diameter of only 19 links. The logarithmic dependence of $\langle d\rangle$ on N is important to the future potential of the web: we find that the expected 1,000% increase in the size of the web over the next few years will change $\langle d\rangle$ very little, from 19 to only 21.

The relatively small value of $\langle d\rangle$ indicates that an intelligent agent, who can interpret the links and follow only the relevant one, can find the desired information quickly by navigating the web. But this is not the case for a robot that locates the information based on matching strings. We find that such a robot, aiming to identify a document at distance $\langle d\rangle$, needs to search $M(\langle d\rangle)\approx0.53\times N^{0.92}$ documents, which, with $N=8\times10^8$, leads to $M=8\times10^7$, or 10% of the whole web. This indicates that robots cannot benefit from the highly connected nature of the web, their only successful strategy being to index as much of the web as possible.

The scale-free nature of the link distributions indicates that collective phenomena play a previously unsuspected role in the development of the web[8], forcing us to look beyond the traditional random graph models[3–5,7]. A better understanding of the web's topology, aided by modelling efforts, is crucial in developing search algorithms or designing strategies for making information widely accessible on the World-Wide Web. Fortunately, the surprisingly small diameter of the web means that all that information is just a few clicks away.

Réka Albert, Hawoong Jeong, Albert-László Barabási
Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA
e-mail:alb@nd.edu

1. Lawrence, S. & Giles, C. L. *Nature* **400,** 107–109 (1999).
2. Claffy, K., Monk, T. E. & McRobb, D. Internet tomography. *Nature* [online] <http://helix.nature.com/webmatters/tomog/tomog.html> (1999).
3. Erdös, P. & Rényi, A. *Publ. Math. Inst. Hung. Acad. Sci.* **5,** 17–61 (1960).
4. Bollobás, B. *Random Graphs* (Academic, London, 1985).

5. Watts, D. J. & Strogatz, S. H. *Nature* **393,** 440–442 (1998).
6. Bunde, A. & Havlin, S. *Fractals in Science* (Springer, Berlin, 1994).
7. Barthélémy, M. & Amaral, L. A. N. *Phys. Rev. Lett.* **82,** 3180–3183 (1999).
8. Barabási, A.-L., Albert, R. & Jeong, H. <http://www.nd.edu/~networks>.

Internet

Growth dynamics of the World-Wide Web

The exponential growth of the World-Wide Web has transformed it into an ecology of knowledge in which highly diverse information is linked in an extremely complex and arbitrary manner. But even so, as we show here, there is order hidden in the web. We find that web pages are distributed among sites according to a universal power law: many sites have only a few pages, whereas very few sites have hundreds of thousands of pages. This universal distribution can be explained by using a simple stochastic dynamical growth model.

The existence of a power law in the growth of the web not only implies the lack of any length scale for the web, but also allows the expected number of sites of any given size to be determined without exhaustively crawling the web. The distribution of site sizes for crawls by Alexa and Infoseek is shown in Fig. 1. Both data sets display a power law over several orders of magnitude, so on a log–log scale the distribution of the number of pages per site appears as a straight line. This distribution should not be confused with Zipf's like distributions[1,2], where a power law arises from rank ordering the variables[3].

In order to describe the growth process underlying this distribution[4], we assume that the day-to-day fluctuations in site size are proportional to the size of the site. One would not be surprised to find that a site with a million pages has lost or gained a few hundred pages on any given day. On the other hand, finding an additional hundred pages on a site with just ten pages within a day would be unusual. So we assume that the number of pages on the site, n, on a given day, is equal to the number of pages on that site on the previous day plus or minus a random fraction of n.

If a set of sites is allowed to grow with the same average growth rate but with individual random daily fluctuations in the number of pages added, their sizes will be distributed log-normally after a sufficiently long period of time[5]. A log-normal distribution gives high probability to small sizes and small, but significant, probability to very large sizes. But although it is skewed and has a long tail, the log-normal distribution is not a power-law one.

Two additional factors that determine the growth of the web need to be considered: sites appear at different times and grow at different rates. The number of web sites has been growing exponentially since its inception, which means that there are many more young sites than old ones. Once the age of the site is factored in to the multiplicative growth process, $P(n)$, the probability of finding a site of size n, is a power law, that is, it is proportional to $n^{-\beta}$. Similarly, considering sites with a wide range of distributions in growth rates yields the same result: a power-law distribution in site size. The simple assumption of stochastic multiplicative growth, combined with the fact that sites appear at different times and/or grow at different rates, therefore leads to an explanation of the observed power-law behaviour.

The existence of this universal power law, which is yet another example of the strong regularities[6,7] revealed by studies of the web, also has practical consequences. The expected number of sites of any arbitrary size can be estimated, even if a site of that size has not yet been observed. This can be achieved by extrapolating the power law to any large n; for example, $P(n_2) = P(n_1) \times (n_2/n_1)^{-\beta}$. The expected number of sites of size n_2 in a crawl of N sites would be $NP(n_2)$. For instance, from the Alexa data we can infer that, if data were collected from 250,000 sites, the probability of finding a site with a million pages would be 10^{-4}. This information is not readily available from the crawl alone, as it stops at 10^5 pages per site.

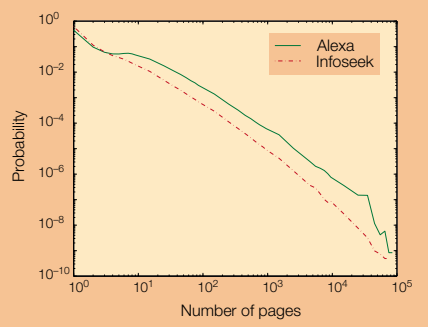

Figure 1 Log–log plot of the distribution of pages in sites for Alexa and Infoseek crawls, which covered 259,794 and 525,882 sites, respectively. There is a drop-off at approximately 10^5 pages because server limitations mean that search engines do not systematically collect more pages per site than this. A linear regression on the variables log(number of sites) and log(number of pages) yielded [1.647, 1.853] as the 95% confidence interval for the exponent β in the Alexa crawl, and [1.775, 1.909] for the Infoseek crawl. These estimates for the power-law slope are consistent across the two data sets and with the model, which predicts that β is greater than 1.

Bernardo A. Huberman, Lada A. Adamic
Xerox Palo Alto Research Center,
3333 Coyote Hill,
Palo Alto, California 94304, USA
e-mail: ladamic@parc.xerox.com

1. Zipf, G. K. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, Massachusetts, 1949).
2. Mantegna, R. N. *et al. Phys. Rev. E* **52,** 2939–2950 (1995).
3. Gunther, R. *et al. Int. J. Theor. Phys.* **35,** 395–417 (1996).
4. http://www.parc.xerox.com/iea/www/growth.html
5. Crow, E. L & Shimizu, K. *Lognormal Distributions: Theory and Applications* (Dekker, New York, 1988).
6. Huberman, B. A., Pirolli, P., Pitkow, J. & Lukose, R. M. *Science* **280,** 95–97 (1998).
7. Huberman, B. A. & Lukose, R. M. *Science* **277,** 535–538 (1997).

Genome evolution

Global methylation in eutherian hybrids

O'Neill *et al.* propose that epigenetic processes help to drive karyotypic evolution in marsupials[1]. Here we present evidence that global methylation patterns do not undergo dramatic changes in interspecific hybrids among three orders of placental mammals, indicating that the mechanisms underlying genome evolution may be different in placental mammals and marsupials.

Interspecific hybridization in mammals frequently results in male sterility[2], abnormal growth[2] and placental dysplasia[3–5], which together may cause post-meiotic reproductive isolation. It has been proposed that incompatibility between rapidly evolving genes that interact normally in the intraspecific context[6] and genomic rearrangements[7] may explain interspecific hybrid defects.

O'Neill *et al.* have given a striking example for the latter mechanism in an interspecific hybrid of the marsupials *Macropus eugenii* and *M. bicolor*[1]. This first-generation (F_1) hybrid exhibited genome-wide demethylation, retrotransposon amplification and centromere expansion on the autosomes derived from *M. eugenii*. Undermethylation of F_1 genomes compared with those of the parental species was also seen in two hybrids of other species within the genus *Petrogale*[1]. These findings were taken to indicate that retrotransposon amplification and chromosome expansion secondary to genome-wide undermethylation could be a frequent phenomenon in mammalian hybrids, leading to rapid karyotypic evolution and finally to reproductive isolation[1].

We have analysed genome-wide methylation in interspecific hybrids in the placental mammalian families of three orders, Equidae (Perissodactyla), Muridae (Rodentia) and Camelidae (Artiodactyla), by following the digestion of genomic DNA with the methylation-sensitive and methylation-insensitive enzymes *Hpa*II and *Msp*I, respectively, and Southern blotting the digest. This analysis included hybrids between horse and donkey, three species of mouse (*Mus musculus*, *Mus spretus* and *Mus macedonicus*), and llama (*Lama glama*) and dromedary (*Camelus dromedarius*). This analysis gave no indication for any changes in genome-wide methylation in any of the F_1 hybrids when compared with parental animals (Fig. 1).