*To appear in* Proceedings of the National Academy of Sciences, Volume 99, Issue 8, pp. 5207–5211, April, 2002. For more information, see http://modelingtheweb.com/.

# Winners don't take all: Characterizing the competition for links on the web

**David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, C. Lee Giles**[*]

NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
{dpennock,flake,lawrence,compuman,giles}@research.nj.nec.com
Phone: +1 (609) 951-2715 (Pennock) Fax: +1 (609) 951-2483

**As a whole, the World Wide Web displays a striking "rich get richer" behavior, with a relatively small number of sites receiving a disproportionately large share of hyperlink references and traffic. However, hidden in this skewed global distribution, we discover a qualitatively different and considerably less biased link distribution among subcategories of pages—for example, among all university homepages or all newspaper homepages. While the connectivity distribution over the entire web is close to a pure power law, we find that the distribution within specific categories is typically unimodal on a log scale, with the location of the mode, and thus the extent of the "rich get richer" phenomenon, varying across different categories. Similar distributions occur in many other naturally-occurring networks, including research paper citations, movie actor collaborations, and US power grid connections. A simple generative model, incorporating a mixture of preferential and uniform attachment, quantifies the degree to which the rich nodes grow richer, and how new (and poorly-connected) nodes can compete. The model accurately accounts for the true connectivity distributions of category-specific web pages, the web as a whole, and other social networks.**

The World Wide Web is a reflection of human culture—a massive social network encoding associative links among almost $10^9$ documents [1] authored by millions of people and organizations around the globe. The web's structure has emerged without central planning, the result of a bottom-up distributed process. Yet many aggregate web characteristics display a striking degree of regularity [2], including the distributions of traffic [3, 4], pages per site [5], file sizes [6, 7], and the depth to which a web user surfs [8]. Several independent investigations show that the distribution of the number of links to (and from) a web page obeys a power law over many orders of magnitude [9, 10, 11, 12]. Power law scaling arises from a variety of physical, biological, and social processes [10, 13, 14, 15]. The emergence of a power law tail seems to characterize the connectivity distribution of many networks in addition to the web, including the graph of movie actor collaborations, the pattern of research paper citations, the topology of the power grid in the western United States, and the metabolic networks of many microorganisms [10, 16, 17].

Barabási and Albert [10, 18] attribute power law scaling to a "rich get richer" mechanism called preferential attachment: as the network grows, the probability that a given vertex receives an edge is proportional to that vertex's current connectivity. Adamic and Huberman [19] give an alternative explanation for power law behavior by adapting their model of the growth of web sites [5] to the case of web links.

Obscured behind the nearly-pure power law distribution found for inbound links on the web as a whole, we uncover a richer structure among subsets of web pages in the same category. We find that these category-specific distributions exhibit very large deviations from power law scaling, with the magnitude of deviation varying from category to category. For these subsets of the web, we illustrate that the body of the distribution of incoming links is actually unimodal, rather than power law. Thus the "rich get richer" character of the web can be much less drastic among competing pages of the same type. In fact, pure power law scaling seems to be the exception rather than the rule. The distributions for outbound web links, and for a variety of other social and biological networks, also display significant de-

[*] C. Lee Giles is also with the School of Information Sciences and Technology and Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16801.

viations from power law, qualitatively similar in nature to those we find for web subsets [9, 10, 11, 12, 16].

We employ a generalized Barabási-Albert (BA) model (similar to recent models [20, 21, 22, 23] independently proposed elsewhere) that incorporates both preferential attachment and a uniform baseline probability of attachment. The model predicts the observed shape of both the body and tail of typical connectivity distributions, including those observed within specific categories of web pages where the divergence from power law is especially marked. In the model, larger modes arise from faster rates of growth of edges as compared to vertices, suggesting an explanation for the different modes observed within different categories of web pages.

## Generic vs. Category-Specific Degree Distributions

Several studies find that the probability that a randomly selected web page has $k$ links is proportional to $k^{-\gamma}$ for large $k$ [9, 10, 11, 12], where $\gamma$ is a constant, empirically determined as roughly $2.1$ for inbound links and $2.72$ for outbound links [11]. When displayed on a log-log plot, this so-called *power law* distribution appears linear with slope $-\gamma$. A power law distribution has a heavy tail, which drops off much more slowly than the tail of a Gaussian distribution. As a result, although the vast majority of web pages have relatively small numbers of links, a few pages have enormous numbers of links—enough to skew the mean well above the median. If we interpret the number of inbound links to a web page as a measure of its popularity or impact, then power law scaling implies that a small fraction of web pages receive a disproportionately large share of such endorsements. As a result, these few popular pages typically benefit from a greater volume of traffic from web surfers, a higher probability of being indexed in search engine databases [1], and more prominent ranking within search engine results. Meanwhile, the majority of sites suffer from relatively poor visibility and new commercial sites may have a difficult time competing for consumer attention. This state of affairs on the web has been referred to (metaphorically, if somewhat inaccurately) as a “winners take all” phenomenon.

At small connectivities $k$, the distribution of links on the web fails to fit a power law, with the discrepancy larger for outbound links than for inbound links [11]. Systematic divergence from power law scaling at small $k$ is also seen in the connectivity distributions of graphs encoding actor collaborations, the western US power grid, scholarly citations, and outbound links from several subsets of the web [10].

Moreover, for some collections of web pages of the same type, we find that the distribution of inbound links departs drastically from a power law at small and moderate $k$. We examined the inbound link distributions for a set of public company homepages (obtained from `http://www.investorguide.com/StockListA.htm`, `StockListB.htm`, etc.), a set of American university homepages (from `http://www.clas.ufl.edu/CLAS/american-universities.html`), a set of US newspaper homepages (from `http://www.usnewspaperlinks.com/`), and a set of scientist homepages (from HPSearch at `http://hpsearch.uni-trier.de/hp/`). Diamond-shaped points in Figure 1 graph the connectivity distribution for company homepages as a log-linear histogram. Pages are placed into buckets according to the number of their inbound links. Buckets are of exponentially increasing width, or constant width on the log scale—the same histogram type used in characterizing web file sizes [6], although different from the histograms used in some previous studies [10, 5]. Specifically, in Figure 1, the $i$th bucket point marks the normalized number of pages with between $10^{i/6} - 1$ and $10^{(i+1)/6} - 1$ inbound links. Although the tail of the distribution continues to fit a power law, the body appears roughly lognormal, with a sharp and singular mode, indicating that a plurality of company homepages have between 99 and 146 inbound links.

Diamonds in Figure 2 display the connectivity distributions of company homepages, university homepages, scientist homepages, and newspaper homepages on log-log scales. All four display the same qualitative shape—unimodal body and power law tail—although the modes vary among the different categories of pages. Heavy tails indicate that a handful of popular pages still gain a disproportionate percentage of all inbound links. Nevertheless, among less popular web pages of the same type, the distribution of inbound links is more evenly balanced. Many web pages can fare well when compared against the mode of all competing pages within the same category. Relative to their community, winners don’t quite “take all”. Losing sites and mediocre sites attract a considerably higher proportion of links than would be the case under a pure power law distribution.

It is an open question exactly how peaked distributions for subsets of the web like those in Figures 1 and 2 sum together to produce the nearly pure power law for the web as a whole. We conjecture that the vast majority of subsets (or subsets containing the vast majority of pages) exhibit a nearly zero mode and dominate this sum, though more investigation is needed.

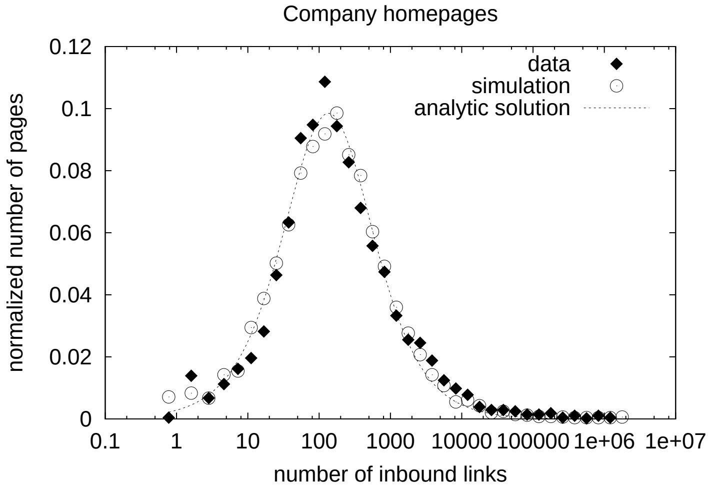

Figure 1: Diamonds plot the empirically observed connectivity distribution for company homepages. Circles display the histogram resulting from a simulation of the model, with parameters $t = 4923$, $m_0 = 0$, $m = 1356$, and $\alpha = 0.950$ set to match the company data. The dashed line marks the analytic solution (3) instantiated with the same parameters.

## Network Growth Model

**Generative process description.** We employ a generative model of network growth to explain the observed connectivity distributions for the web, for web categories, and for other social networks. The model is similar to other generalized BA models recently developed independently by other authors [20, 21, 22, 23]. The network begins with $m_0$ vertices. At each time step $t$, one vertex and $m$ edges are added to the network. In the BA model, all $m$ edges connect the new vertex with an old vertex according to preferential attachment: the probability $\Pi(k_i)$ that an edge connects to vertex $i$ is proportional to $k_i$, where $k_i$ is the current number of edges incident on vertex $i$.

We presume instead that every vertex has at least some baseline probability of gaining an edge. Both endpoints of edges are chosen according to a mixture of probability $\alpha$ for preferential attachment and $1 - \alpha$ for uniform attachment. The probability that an endpoint of a new edge connects to vertex $i$ is

$$\Pi(k_i) = \alpha \frac{k_i}{2mt} + (1 - \alpha) \frac{1}{m_0 + t}, \qquad (1)$$

where $m_0 + t$ is the total number of vertices and $2mt$ the total connectivity at time $t$. Edge endpoints are chosen symmetrically, rather than pinned to the newest vertex. Solitary vertices are not destined to remain forever disconnected. Under preferential attachment alone, sites that are already rich in links tend to get richer, resulting in a nearly pure power law distribution over connectivities. On the other hand, with the addition of a component for uniform attachment, the poorer sites (with some luck) can get rich too. Intuitively, the two growth components can be viewed as capturing two common behaviors of web page authors: (a) creating links to pages that the author is aware of because they are popular, and (b) creating links to pages that the authors is aware of because they are personally interesting or relevant, largely idependent of popularity.

We generated a simulated network using (1), with parameters set to model the company homepages data: $t$ and $2m$ are set to the actual number of web pages (4923) and the average number of inbound links per page (2712), respectively. The seed set size $m_0$ is set to zero. The only tuning parameter, $\alpha$, is set according to a non-linear least-squares fit of the analytic solution (3) to the data (resulting in $\alpha = 0.950$). Multiple edges between two vertices are allowed, though self-edges are not. Circles in Figure 1 plot the resulting connectivity histogram, which corresponds very well with the true distribution. Circles in Figure 2 display simulation results for all four data sets on log-log scales, again showing good agreement with empirical measurements.

Notice that the simulation builds a graph among subset members only, while empirical data includes inbound links originating from the entire web. There are two ways to interpret the model to reconcile this difference. First, one can think of the model as a prescription not for graph

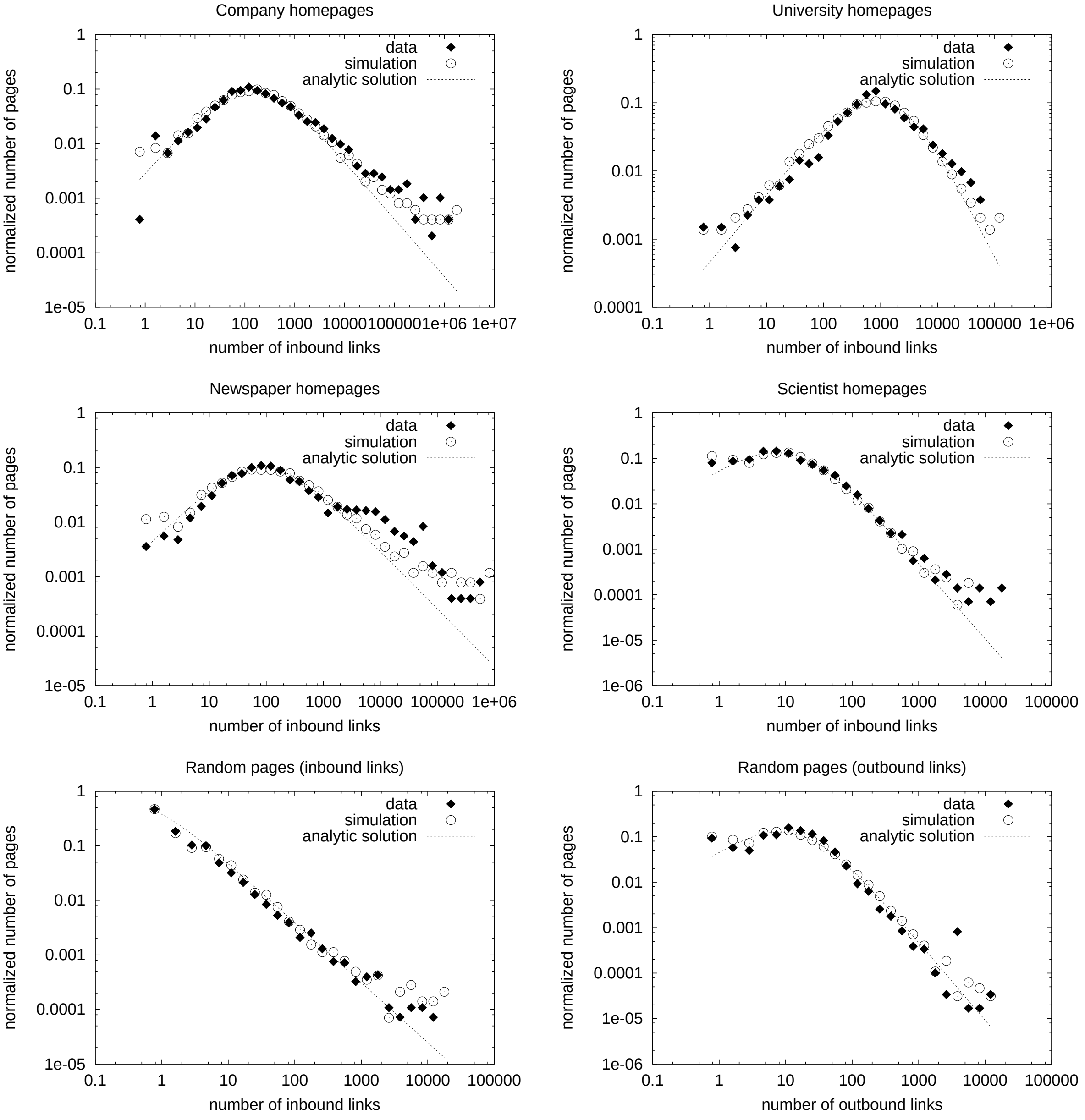

Figure 2: Diamonds display log-log histograms of inbound connectivities for category-specific homepages, and inbound and outbound connectivities for random web pages. Circles mark the connectivity distributions for corresponding simulations, with $m_0 = 0$, $t$ set equal to the number of web pages, $2m$ set equal to the average number of inbound links per page, and $\alpha$ chosen according to a non-linear least-squares fit. Dashed lines indicate the analytic solutions (3).

generation, but simply for connectivity growth, where each vertex increments its connectivity independently according to (1); this is the same abstraction adopted in other models [20, 22]. Second, one can interpret an edge between vertex $i$ and vertex $j$ as a path between $i$ and $j$, possibly traversing outside subset boundaries (with both endpoints being inbound links).

Finally, we note that the model can be easily generalized into a directed graph model. Two parameters $\alpha_{in}$ and $\alpha_{out}$ encode mixture probabilities for inbound and outbound links. The source of each new edge is chosen according to (1) using $\alpha_{out}$ and the destination according to (1) using $\alpha_{in}$. Simulation and analytic results describing inbound (or outbound) connectivity growth are unaffected by this modification, modulo a factor of two.

**Analytic solution.** Using a continuous mean-field approximation similar to that employed by Barabási and Albert [18], we can derive the connectivity distribution for the model in closed form. Assuming that $t \gg m_0$, the probability that a vertex has connectivity $k$ is

$$\Pr(k) = [2m(1-\alpha)]^{\frac{1}{\alpha}} \, [\alpha k + 2m(1-\alpha)]^{-1-\frac{1}{\alpha}} \qquad (2)$$

In the limit as $k \to \infty$, the density $\Pr(k)$ is proportional to $k^{-(1+1/\alpha)}$, or a power law with exponent $\gamma = 1 + 1/\alpha$. For example, if $\alpha = 1/2$, then $\gamma = 3$, the same as predicted in the BA model. Mixture parameters $\alpha$ of 0.909 and 0.581 yield exponents of 2.1 and 2.72, respectively, the empirically observed exponents for inbound and outbound web links [11]. Other values for $\alpha$ yield alternative power law exponents.

Our log-scale histograms in Figures 1 and 2 employed exponentially increasing bucket sizes. We can perform an analogous transformation of the probability density (2), in order to facilitate comparison on log-scale plots. We substitute $k = 10^{k'/6}$ into the cumulative distribution, take the derivative with respect to $k'$, and substitute back using $k' = 6\log_{10} k$. The resulting function displays the instantaneous probability mass at each $k$, where the widths of the infinitesimal "buckets" $dk$ are constant on the logarithmic scale. This transformed density $\widetilde{\Pr}(k)$, suitable for log-scale visualization, is

$$\widetilde{\Pr}(k) = \frac{\ln 10}{6} \cdot [2m(1-\alpha)]^{\frac{1}{\alpha}} \cdot k \cdot [\alpha k + 2m(1-\alpha)]^{-1-\frac{1}{\alpha}} \, . \qquad (3)$$

The maximum of this function, corresponding to the mode of the distribution on a log scale, occurs at $k = 2m(1-\alpha)$. The location of the mode is directly proportional to $m$, the rate of edge additions per time step. If $\alpha = 1/2$, for example, then the mode is simply $m$, or the number of edges added per vertex. As the growth rate of edges increases compared to that of vertices, the mode shifts toward higher connectivities $k$. As the mixture parameter $\alpha$ approaches one, or as $m$ approaches zero, the distribution approaches a pure power law, and the mode appears at much lower connectivities.

**Related models.** Dorogovtsev et al. [20] and Levene et al. [22] independently propose similar generalizations of the BA model (the addition of a uniform component), motivating it in part as a natural way to parameterize the power-law exponent. Dorogovtsev et al. [20] solve for the exact degree distribution, showing that BA's mean-field approximation is correct in determining the asymptotic power-law exponent; the authors go on to study the connectedness properties of the network as it grows [21]. Levene et al. [22] reformulate the growth process in terms of an urn transfer model, enabling them to obtain exact solutions in certain cases. Albert and Barabási [24] have proposed their own extension of their original model. Their augmented model involves a parameterized mixture of three processes: vertex additions, edge additions, and edge rewirings. The combination of these three processes leads to a connectivity growth function that is roughly a sum of uniform and preferential terms. Kleinberg et al. [23] propose a model where some edges are added at random and some are copied from existing vertices, again leading to a mixture of uniform and preferential influences on network growth. Even Simon [25] in 1955 invokes a similar process to explain Estoup-Zipf word frequency distributions. While most authors point out that the generalized form is flexible enough to admit any asymptotic power law exponent in the range $(2, \infty)$, we focus on the fact that the same single additional degree of freedom is also sufficient to explain the often large deviation from power law behavior observed in the low connectivity region.

## Web Data and Model Comparisons

The model's ability to fit both the body and tail of typical degree distributions is especially evident for category-specific web data. Figure 2 illustrates the fit between the model and the actual connectivity distributions for company, university, newspaper, and scientist homepages. The figure overlays web data, simulation data, and the mean-field solution (3) for the four sets of web pages on log-log scales; Figure 1 displays the same information on a log-linear scale for the company homepages. Any discrepancy between the analytic solution and the simulation is a result of the mean-field approximation. For the simulation and the analytic solution, the model parameters $t$ and $2m$ are set to the number of web pages and the average number

| data set | $\alpha$ | mode | $\gamma$ |
|---|---|---|---|
| companies | 0.950 | 136 | 2.05 |
| newspapers | 0.948 | 87 | 2.05 |
| web inlinks | 0.909 | 0 | 2.10 |
| universities | 0.612 | 839 | 2.63 |
| scientists | 0.602 | 7 | 2.66 |
| web outlinks | 0.581 | 8 | 2.72 |

Table 1: Mixture parameters $\alpha$, modes $2m(1-\alpha)$, and power law tail exponents $\gamma = 1 + 1/\alpha$ for inbound links to category-specific homepages, and for inbound and outbound links on the web as a whole.

of inbound links per page, respectively. The seed set size $m_0$ is set to zero and $\alpha$ is optimized using a non-linear regression. In all four cases, the model distributions fit very closely to the true distributions, capturing the same unimodal body and power law tail observed in the data.

Note that the only tuning parameter, $\alpha$, affects both the mode and the slope of the tail, yet a single best-fit $\alpha$ captures both dimensions well. We also computed distributions for inbound and outbound links for the web as a whole, using a collection of 100,000 pseudo-random web pages, sampled from roughly one billion URLs in Inktomi Corporation's webmap. The model fits these distributions closely as well; moreover, the mixture parameters $\alpha$ imply power law slopes $\gamma = 1 + 1/\alpha$ precisely in line with previous measurements [11]. Table 1 reports the best-fit parameters $\alpha$, modes, and power law exponents $\gamma$ for the four data sets and for the web as a whole.

The distribution of links to university homepages exhibits the largest deviation from a power law; on the other end of the spectrum, the distribution of inbound links on the web as a whole is closest to a pure power law. In all cases studied, mixture parameters $\alpha$ are greater than 1/2. Thus preferential attachment appears to play a larger role in web link growth than does uniform attachment. The growth of links to company homepages ($\alpha = 0.950$) and to newspaper homepages ($\alpha = 0.948$) is most dominated by the "rich get richer" process of preferential attachment, while link growth on scientist homepages ($\alpha = 0.602$) and university homepages ($\alpha = 0.612$) suggest a more balanced mixture of preferential and uniform terms. The model also appears consistent with the shape of connectivity distributions reported for the graph of actor collaborations, the networks of western US power stations, the citation pattern among publications, and outbound links from subsets of the web [10]. The distribution of file sizes on the web is also qualitatively similar to our category-specific link distributions, with a lognormal body and a power law tail [6]. Previous studies characterized the body and tail separately [6]; Equation 3 (when interpreted purely as a growth model rather than a graph generation model) serves as a single-function alternative for describing the full distribution.

# Conclusions

The addition of pages and links to the web is a distributed, asynchronous, complex and continual process: to an outside observer, fine-grained changes must appear almost haphazard. Yet, when examined on the large, discernible patterns emerge [11, 3, 8, 2, 5, 6] some of which are shared with other social and biological networks [10, 16, 17]. For one, the distribution of the number of links to (and from) a page has been shown to follow a power law over many orders of magnitude [10, 11, 12]. We demonstrate that, among web pages of the same type, the body of the distribution of inbound links deviates strongly from a power law, exhibiting a roughly lognormal shape. A generative model incorporating uniform as well as preferential attachment explains data from the web as a whole, as well as category-specific data from company, university, newspaper, and scientist homepages.

As commerce and communication move to the web, the dynamics of link accumulation—at both global and local granularities—can strongly influence competition and diversity in business and society. Improved tools for characterizing and modeling these dynamics will have significant scientific and commercial value [23]. Beyond the web, understanding commonalities among diverse network types promises to enrich our understanding of the evolution of social and ecological structures.

### Acknowledgements

We thank A.-L. Barab´asi, D. Watts, F. Coetzee, P. Rusmevichientong, A. Khrabrov, S.-T. Park, S. Chakrabarti, and the anonymous reviewers for insights, advice, and pointers to related work. We thank Inktomi Corporation for access to their webmap.

# References

[1] Lawrence, S. & Giles, C. L. *Nature* **400**, 107–109 (1999).

[2] Pitkow, J. E. *Computer Networks and ISDN Systems* **30**, 551–558 (1998).

[3] Adamic, L. A. & Huberman, B. A. *Quarterly Journal of Electronic Commerce* **1**(1), 5–12 (2000).

[4] Glassman, S. *Computer Networks and ISDN Systems* **27**, 165–173 (1994).

[5] Huberman, B. A. & Adamic, L. A. *Nature* **401**, 131 (1999).

[6] Barford, P., Bestavros, A., Bradley, A. & Crovella, M. *World Wide Web, Special Issue on Characterization and Performance Evaluation* **2**, 15–28 (1999).

[7] Woodruff, A., Aoki, P. M., Brewer, E., Gauthier, P. & Rowe, L. A. *Computer Networks and ISDN Systems* **28**, 963–980 (1996).

[8] Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E. & Lukose, R. M. *Science* **280**, 95–97 (1998).

[9] Albert, R., Jeong, H. & Barab´asi, A.-L. *Nature* **401**, 130–131 (1999).

[10] Barab´asi, A.-L. & Albert, R. *Science* **286**, 509–512 (1999).

[11] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. *Computer Networks (Proceedings of WWW9)* **33**(1–6), 309–320 (2000). Available online at `http://www.www9.org/w9cdrom/160/160.html`.

[12] Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. *Computer Networks (Proceedings of WWW8)* **31**(11–16), 1481–1493 (1999). Available online at `http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html`.

[13] Casti, J. L. *Complexity* **1**(1), 12–15 (1995).

[14] May, R. M. *Science* **214**, 1441–1449 (1988).

[15] Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, (1994).

[16] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barab´asi, A.-L. *Nature* **407**, 651–654 (2000).

[17] Watts, D. J. & Strogatz, S. H. *Nature* **393**, 440–442 (1998).

[18] Barab´asi, A.-L., Albert, R. & Jeong, H. *Physica A* **272**, 173–187 (1999).

[19] Adamic, L. A. & Huberman, B. A. *Science* **287**, 2115a (2000).

[20] Dorogovtsev, S., Mendes, J. & Samukhin, A. *Physical Review Letters* **85**(21), 4633–4636 (2000).

[21] Dorogovtsev, S., Mendes, J. & Samukhin, A. *Phys. Rev. E* **64**, 066110 (2001).

[22] Levene, M., Fenner, T., Loizou, G. & Wheeldon, R. Technical Report cond-mat/0110016, LANL ArXiv, (2001). Available online at `http://www.arXiv.org/abs/cond-mat/0110016/`.

[23] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. S. in *Proceedings of the 5th International Conference on Computing and Combinatorics*, 1–18 (Springer-Verlag, Berlin, 1999).

[24] Albert, R. & Barab´asi, A.-L. *Physical Review Letters* **85**(24), 5234–5237 (2000).

[25] Simon, H. *Biometrika* **42**, 425–440 (1955).