Winners don't take all: A model of web link accumulation

David M. Pennock[1], **Gary W. Flake**[1], **Steve Lawrence**[1],
C. Lee Giles[1,2], **Eric J. Glover**[1,3]
{dpennock,flake,lawrence,giles,compuman}@research.nj.nec.com
Phone: +1 609 951 2715 (Pennock) Fax: +1 609 951 2488

[1] NEC Research Institute
4 Independence Way
Princeton, NJ 08540

[2] School of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16801

[3] Artificial Intelligence Laboratory
University of Michigan
Ann Arbor, MI 48109

As a whole, the World Wide Web displays a striking "winners take all" or "rich get richer" character, with a relatively small number of sites receiving a disproportionately (and increasingly) large share of hyperlink references *(1, 2, 3, 4)* and traffic *(5, 6, 7)*. Hidden in this skewed global distribution, however, we discover a qualitatively different, and considerably less biased, link distribution among pages of the same category—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, the distribution within specific categories is typically unimodal on a log scale, with the location of the mode, and therefore the extent of the "winners take all" 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 *(2, 8)*. We present a generative model, incorporating a mixture of preferential and uniform attachment, that quantifies the degree to which the rich nodes in a network 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. As commerce and communication move to the web, the dynamics of link accumulation—at both global and local granularities—can have a significant effect on competition and diversity throughout business and society. Our model may be used to study a variety of networks and communities, for example, by predicting growth patterns based on a static snapshot, or by inferring the degree of "winners take all" behavior within various communities.

The World Wide Web is a reflection of human culture—a massive social network encoding associative links among almost 10^9 documents *(9)* 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. Perhaps surprisingly, then, many aggregate web characteristics display a striking degree of regularity *(10)*, including the distributions of traffic *(5, 6)*, pages per site *(11)*, file sizes *(12, 13, 14)*, and depth to which a web user surfs *(7)*. 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 *(1, 2, 3, 4)*. Power law scaling arises from a variety of physical, biological, and social processes *(2, 15, 16, 17)*. 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 *(2, 8, 18, 19)*.

The probability that a web page has k links is proportional to $k^{-\gamma}$ for large k, where γ is a constant, empirically determined as roughly 2.1 for inbound links and 2.72 for outbound links *(3)*. 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 a "winners take all" scenario: a small fraction of web pages receive a disproportionately large share of the total number of inbound links. 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, and more prominent ranking within search engine results. The web's power law nature means that a majority of sites suffer from relatively poor visibility, and new commercial sites may have a difficult time competing for consumer attention.

Barabási and Albert *(2, 20)* 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. Albert and Barabási *(21)* generalize their original model to incorporate a mixture of network processes, including edge additions, edge rewirings, and vertex additions. Adamic and Huberman *(22)* give an alternative explanation for power law behavior by adapting their model of the growth of web sites *(11)* to the case of web links. Kleinberg et al. *(23)* propose a model where some edges are added at random and some are copied from existing vertices.

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. Thus the "winners take all" character of the web can actually 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 deviations from power law, qualitatively similar in nature to those we find for web subsets *(1, 2, 3, 4, 8, 18)*.

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 *(24)* 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, using exponentially increasing bucket widths, or constant widths on the log scale. Although the tail of the distribution continues to fit a power law, the body appears roughly lognormal, with a sharp and singular mode.

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 percent 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.

We propose a generative model of network growth to explain the observed connectivity distributions for the web, for web categories, and for other social networks. As in the Barabási-Albert (BA) model *(2)*, 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 from the new vertex to an old vertex according to preferential attachment: the probability $\Pi(k_i)$ that an edge connects to vertex i is $k_i / \sum_j k_j$, where k_i is the current number of edges incident on vertex i, and the summation is over all old vertices. Notice that, in the BA model, no vertex can have fewer than m edges except those in the initial seed set. BA networks tend to grow as a single connected component containing all vertices. In contrast, the

so-called "bow tie" characterization of the web's structure, based on a large scale empirical study, suggests that about one quarter of the web is disconnected from a dominant connected core *(3)*.

We introduce into the model the natural intuition that every vertex has at least some baseline probability of gaining an edge. To this end, both endpoints of edges are chosen according to a mixture of probability α for preferential attachment and $1 - \alpha$ for uniform attachment. The probability that an endpoint of a new edge connects to vertex i is

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

Note that $m_0 + t$ is the total number of vertices and $2mt$ the total connectivity at time t. In this augmented model, edge endpoints are chosen symmetrically, rather than pinned to the newest vertex. Solitary vertices are *not* destined to remain forever disconnected, and there is no discontinuity in the connectivity distribution at $k = m$. Simulations, with α set to roughly match web data, result in networks with about seventy percent of vertices in a dominant connected component, in reasonable agreement with the bow tie characterization *(3)*. Under preferential attachment alone, sites that are already rich in links tend to get richer, resulting in a 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, leading to a connectivity distribution more in line with empirical evidence.

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, α, is optimized using a non-linear regression. Circles in Figure 1 plot the resulting connectivity histogram, which corresponds very well with the true distribution.

We derive in closed form the distribution implied by (1), by applying a mean-field approximation technique similar to that employed by Barabási and Albert *(20)*. We assume that k is continuous and that $\Pi(k_i)$ is the growth rate of k_i. Then,

$$\frac{\partial k_i}{\partial t} = A\Pi(k_i) = A\alpha \frac{k_i}{2mt} + A(1 - \alpha) \frac{1}{m_0 + t}. \tag{2}$$

Since $\Pi(k_i)$ sums to one, and the total connectivity increase per time step is $2m$, A must equal $2m$. For the remainder of the derivation we assume that $t \gg m_0$. Substituting in for A, we find that,

$$\frac{\partial k_i}{\partial t} = \alpha \frac{k_i}{t} + (1 - \alpha) \frac{2m}{t}.$$

Using the initial condition that vertex i begins at time t_i with no incident edges, or $k_i(t_i) = 0$,

$$k_i(t) = 2m \left(\frac{1-\alpha}{\alpha} \right) \left(\frac{t^\alpha - t_i^\alpha}{t_i^\alpha} \right).$$

Then the probability that vertex i's connectivity $k_i(t)$ is less than k is

$$\Pr(k_i(t) < k) = \Pr\left(t_i > t \left[\frac{2m(1-\alpha)}{\alpha k + 2m(1-\alpha)} \right]^{\frac{1}{\alpha}} \right) = 1 - \left[\frac{2m(1-\alpha)}{\alpha k + 2m(1-\alpha)} \right]^{\frac{1}{\alpha}}, \tag{3}$$

where the last step assumes that vertices are added at equal time intervals, and thus that the probability density of t_i is uniform, or $\Pr(t_i) = 1/t$. The probability density of k is the partial derivative of the cumulative distribution (3) with respect to k.

$$\Pr(k) = \frac{\partial \Pr(k_i(t) < k)}{\partial k} = [2m(1-\alpha)]^{\frac{1}{\alpha}} \left[\alpha k + 2m(1-\alpha) \right]^{-1-\frac{1}{\alpha}} \tag{4}$$

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 α 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 *(3)*.

Our log-scale histograms in Figures 1 and 2 employed exponentially increasing bucket sizes. We can perform an analogous transformation of the probability density (4), in order to facilitate comparison on log-scale plots. We substitute $k = 10^{k'/6}$ into the cumulative distribution (3), 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

$$\widetilde{\Pr}(k) = \frac{\ln 10}{6} \cdot [2m(1-\alpha)]^{\frac{1}{\alpha}} \cdot k \cdot \left[\alpha k + 2m(1-\alpha) \right]^{-1-\frac{1}{\alpha}}. \tag{5}$$

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 as compared to vertices, the mode shifts toward higher connectivities k. As the mixture parameter α approaches one, or as m approaches zero,

the distribution approaches a pure power law, and the mode appears at much lower connectivities.

Albert and Barabási *(21)* have independently proposed a different extension of the BA model. Their augmented model involves a parameterized mixture of three processes: vertex additions, edge additions, and edge rewirings. As in the original BA model, all new vertices begin with m edges. The new model also allows for internal edge additions. These additions are asymmetric: one edge endpoint is chosen uniformly and the other preferentially. Note that this assumption does not seem to hold for the web, since both inbound and outbound link distributions follow a power law. Edge endpoint *subtractions*—due to rewiring—are also chosen uniformly. The combination of these three processes leads to a connectivity growth function similar in form to (2)—roughly a sum of uniform and preferential terms.

The dashed line in Figure 1 graphs (5), where t, m_0, m, and α are set to the same values as in the company homepages simulation. The closeness of fit between the analytic solution and the simulation reflects the accuracy of the mean-field approximation. Figure 2 illustrates the fit between our 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 (5) for the four sets of web pages on log-log scales. 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, α, affects both the mode and the slope of the tail, yet a single best-fit α 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. Our model fits these distributions closely as well; moreover, the mixture parameters α imply power law slopes $\gamma = 1 + 1/\alpha$ precisely in line with previous measurements *(3)*. Table 1 reports the modes, best-fit parameters α, and power law exponents γ for the four data sets and for the web as a whole.

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 *(3, 5, 7, 10, 11, 12)* some of which are shared with other social and biological networks *(2, 18, 19)*. Improved tools for measuring, characterizing, and modeling the web will have significant scientific, social, 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.

References and Notes

1. R. Albert, H. Jeong, A.-L. Barabási, *Nature* **401**, 130 (1999).
2. A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
3. A. Broder *et al.*, In *Proceedings of the Ninth International World Wide Web Conference* (2000).
4. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, In *Proceedings of the Eighth International World Wide Web Conference* (1999).
5. L. A. Adamic and B. A. Huberman, *Quarterly Journal of Electronic Commerce* **1**(1), 5 (2000).
6. S. Glassman, *Computer Networks and ISDN Systems* **27**, 165 (1994).
7. B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, R. M. Lukose, *Science* **280**, 95 (1998).
8. A.-L. Barabási, R. Albert, H. Jeong, Technical report University of Notre-Dame (1999).
9. S. Lawrence and C. L. Giles, *Nature* **400**, 107 (1999).
10. J. E. Pitkow, *Computer Networks and ISDN Systems* **30**, 551 (1998).
11. B. A. Huberman and L. A. Adamic, *Nature* **401**, 131 (1999).
12. P. Barford, A. Bestavros, A. Bradley, M. Crovella, *World Wide Web, Special Issue on Characterization and Performance Evaluation* **2**, 15 (1999).
13. C. R. Cunha, A. Bestavros, M. Crovella, Technical Report TR-95-010 Department of Computer Science, Boston University (1995).
14. A. Woodruff, P. M. Aoki, E. Brewer, P. Gauthier, L. A. Rowe, *Computer Networks and ISDN Systems* **28**, 963 (1996).
15. J. L. Casti, *Complexity* **1**(1), 12 (1995).
16. R. M. May, *Science* **214**, 1441 (1988).
17. S. Wasserman and K. Faust, *Social Network Analysis : Methods and Applications* (Cambridge University Press, Cambridge, 1994).
18. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabási, *Nature* **407**, 651 (2000).
19. D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).
20. A.-L. Barabási, R. Albert, H. Jeong, *Physica A* **272**, 173 (1999).
21. R. Albert and A.-L. Barabási, Technical report LANL ArXiv (2000).
22. L. A. Adamic and B. A. Huberman, *Science* **287**, 2115a (2000).
23. J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. S. Tomkins, In *Proceedings of the the International Conference on Combinatorics and Computing* (1999).
24. G. Hoff and M. Mundhenk, Technical Report D-54286 Department of Mathematics/Computer Science, University Trier, Germany (1999).

data set	mode	α	γ
universities	839	0.612	2.63
companies	136	0.950	2.05
newspapers	87	0.948	2.05
web outlinks	8	0.581	2.72
scientists	7	0.602	2.66
web inlinks	0	0.909	2.10

Table 1: Mixture parameters α, 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. The web inlink and outlink exponents γ match precisely with Broder et al.'s *(3)* measurements. 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 α 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 and to newspaper homepages is most dominated by the "rich get richer" process of preferential attachment, while link growth on scientist homepages and university homepages suggests a more balanced mixture of preferential and uniform terms.

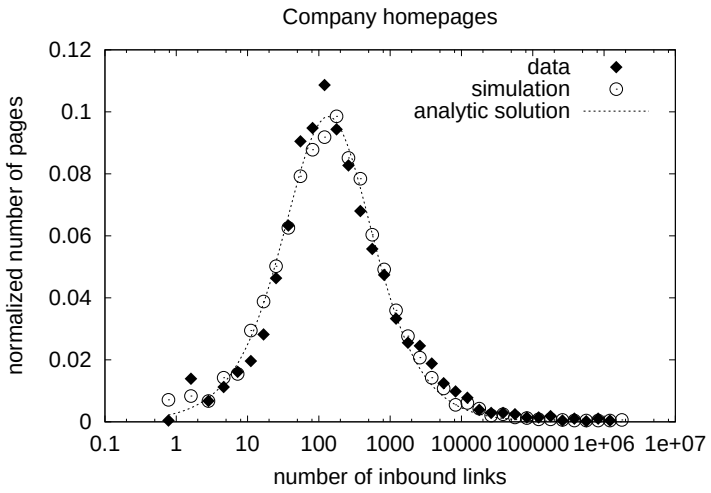

Figure 1: Diamonds indicate a histogram of the number of inbound links to company homepages. Pages are placed into buckets according to the number of their inbound links. Buckets are of exponentially increasing width, or constant width on the log scale—the same histogram type used in characterizing web file sizes *(12, 13)*, though different than employed in some previous studies *(2, 11)*. Specifically, the ith bucket point marks the probability that the number of links to a page falls between $10^{i/6} - 1$ and $10^{(i+1)/6} - 1$. The distribution has a sharp and singular mode, indicating that a plurality of company homepages have between 99 and 146 inbound links. Circles display the histogram resulting from a simulation of our model, with parameters set to match the company data: $t = 4923$ is set to the actual number of pages, m_0 is set to zero, $2m = 2712$ is set to the average number of inbound links per homepage, and $\alpha = 0.950$ is set according to a non-linear least-squares fit of the analytic solution (5) to the data. Multiple edges between two vertices are allowed, though self-edges are not. The dashed line marks the analytic solution (5) instantiated with the same parameters.

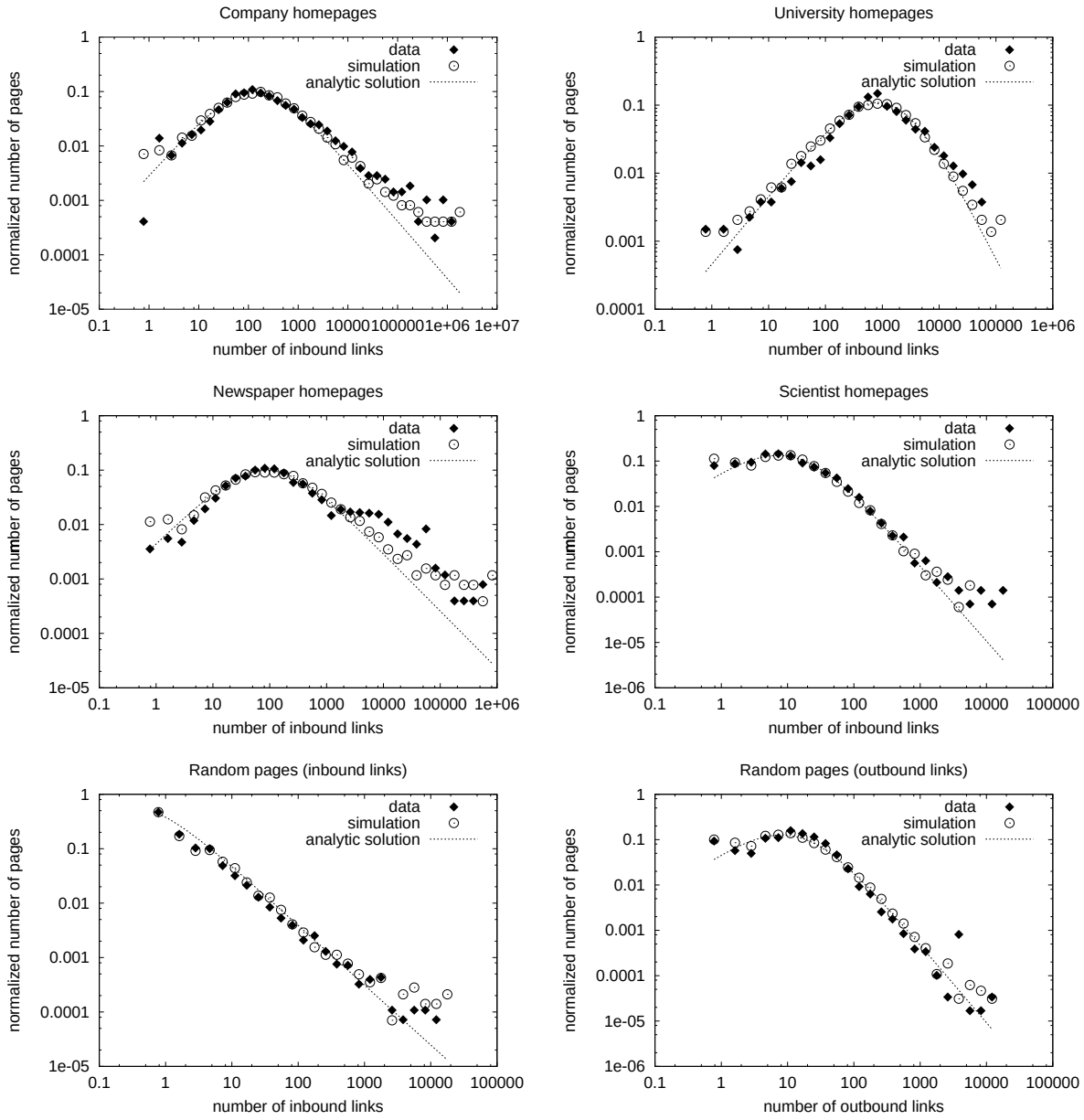

Figure 2: Diamonds display log-log histograms of inbound connectivities for category-specific homepages, and inbound and outbound connectivities for random web pages. Circles mark the connectivity distributions for corresponding simulations, with $m_0 = 0$, t set equal to the number of web pages, $2m$ set equal to the average number of inbound links per page, and α chosen according to a non-linear least-squares fit. Dashed lines indicate the analytic solutions (5).