Evolution of document networks

Filippo Menczer[†]

School of Informatics, Indiana University, Bloomington, IN 47408

How does a network of documents grow without centralized control? This question is becoming crucial as we try to explain the emergent scale-free topology of the World Wide Web and use link analysis to identify important information resources. Existing models of growing information networks have focused on the structure of links but neglected the content of nodes. Here I show that the current models fail to reproduce a critical characteristic of information networks, namely the distribution of textual similarity among linked documents. I propose a more realistic model that generates links by using both popularity and content. This model yields remarkably accurate predictions of both degree and similarity distributions in networks of web pages and scientific literature.

There are important social and economic implications in explaining how the topology evolves in information networks such as the World Wide Web. Although text analysis has been used for a long time to analyze documents, extract their meaning, retrieve information, and map knowledge domains (1–3), link analysis is increasingly used by search engines and digital libraries to estimate the importance or reputation of documents (4–6) and to map documents into topical clusters (7–10).

A number of models have been proposed to explain the growth of complex networks exhibiting characteristics such as the small-world property (high clustering and low diameter) and the scale-free property (power-law distribution of degree). A few representative models are reviewed in *Background*. Which of these competing theories is more plausible? In *Validating Prior Models*, I show that the answer is none: Although they all correctly predict degree distributions, they fail at predicting another observable feature of the information network, namely the distribution of textual similarity among linked documents. In *Degree-Similarity Mixture Model*, I propose a growth model to explicitly capture the trade-off between an author's desire to link related and popular documents. The model is validated against two data sets: a network of web pages sampled from the Open Directory Project (DMOZ; http://dmoz.org) and a collection of scientific articles published in PNAS. Web pages are connected by hyperlinks and articles by citations. Numerical simulations are used to generate predictions of degree and similarity distributions for both data sets.

Background

Since the discovery of scale-free and small-world phenomena in web pages (11–15) and bibliographic collections (16–18), physicists and computer scientists have developed growth theories that model the behavior of authors linking new pages to explain the emergence of these critical network properties (9, 19). Most growth models are based on some form of preferential attachment, whereby one node at a time is added to the network with new edges to existing nodes selected according to some probability distribution.

In the best known preferential attachment model, a node i receives a new edge with probability proportional to its current degree, $\Pr(i) \propto k(i)$ (13). This so-called BA model generates networks with power-law degree distributions, in which the oldest nodes are those with highest degree. The copying model and its extensions implement equivalent rich-get-richer processes based on local walks, without requiring explicit knowledge of degree (20–22).

To give newer nodes a chance to compete for links, an extension of the preferential attachment model is based on linking to a node dependent on its degree with some probability or to a uniformly chosen node with the remaining probability (23, 24). Such a mixture model generates networks that can fit the power-law degree distribution of the entire web as well as the different distributions observed in subsets of the web such as university and business homepages (25).

Some theories have explored similar mixture models in which links are created according to a trade-off between graph degree and metric distance measures, showing that certain trade-off regimes lead to power-law degree distributions (26, 27).

To study the decision process by which authors link documents, let us consider the relationship between the probability that two documents are linked and their content (text) similarity. Content similarity can be measured by the cosine metric traditionally used in information retrieval (1):

$$\sigma_c(d_1, d_2) = \frac{\|\vec{d}_1 \cdot \vec{d}_2\|}{\|\vec{d}_1\| \cdot \|\vec{d}_2\|}, \qquad [1]$$

where $\vec{d}$ is a vector space representation of the text in document d. Link probability can be approximated by a link similarity (or neighborhood) metric defined as the Jaccard coefficient:

$$\sigma_l(d_1, d_2) = \frac{|U_{d_1} \cap U_{d_2}|}{|U_{d_1} \cup U_{d_2}|}, \qquad [2]$$

where U_d is the set representing d's neighborhood, which consists of inlinks and outlinks for web pages and citations and references for articles (in which case σ_l is akin to cocitation and bibliographic coupling). To visualize the relationship between text and link similarity, one can map the joint distribution of σ_c and σ_l across pairs of documents. Fig. 1 shows two such maps, obtained from collections of web pages and PNAS articles. These maps demonstrate that for web pages as well as scientific papers, authors connect documents in a way that is significantly (although not strongly) correlated with the similarity of those documents to their own. The Pearson correlation coefficient between σ_c and σ_l is 0.10 for web pages (3.8×10^9 pairs) and 0.12 for PNAS articles (7.5×10^6 pairs).

To quantify the dependence of the web's link topology on content, I considered the conditional probability that the link neighborhood between two web pages is above some threshold λ, given that the two pages have some content similarity κ, as a function of κ:

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

[†]E-mail: fil@indiana.edu.

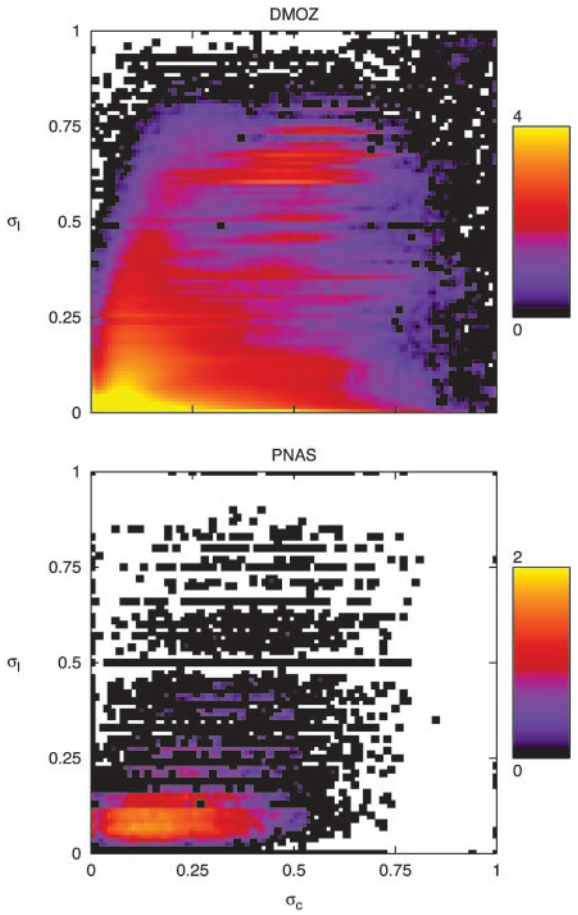

Fig. 1. Joint distribution maps for content and link similarity across pairs of web pages (*Upper*) and scientific articles (*Lower*). Colors code the $\log_{10}$ of the number of pairs with the corresponding similarity coordinates. The web data are based on a stratified sample of 109,648 pages from the DMOZ, with their inlinks and outlinks, crawled in 2002. The article data are based on the titles, abstracts, and references of 15,785 articles published in PNAS between 1997 and 2002.

$$\Pr(\lambda|\kappa) = \frac{|(p,q):\sigma_c(p,q) = \kappa \wedge \sigma_l(p,q) > \lambda|}{|(p,q):\sigma_c(p,q) = \kappa|}, \qquad [3]$$

where p, q are two web pages. An interesting phase transition was observed between two distinct regions around a critical distance κ^* independent of λ (28). For $\kappa > \kappa^*$, the probability that two pages are neighbors does not seem to depend on their content similarity; for $\kappa < \kappa^*$, the probability decreases according to a power-law $\Pr(\lambda|\kappa) \sim \kappa^\gamma$, with the decay exponent γ growing linearly with λ. This observation suggested a content-based growth model in which an author tends to link a new page to the most popular among related (similar) pages and with decreasing probability to less similar ones. As in the BA model, at each step t one new page t is added, and m new links are created from t to m existing pages, each selected from $\{i, i < t\}$ with probability:

$$\Pr(i,t) = \begin{cases} \dfrac{k(i)}{mt} & \text{if } \sigma_c(i,t) > \kappa^* \\ c\sigma_c^\gamma(i,t) & \text{otherwise} \end{cases}, \qquad [4]$$

where m, κ^*, and γ are constants derived from the data and c is a normalization factor. This degree-similarity phase model accurately predicted the degree distribution of the web pages in the sample from which the data were derived and was the first model to do so taking content into account (28).

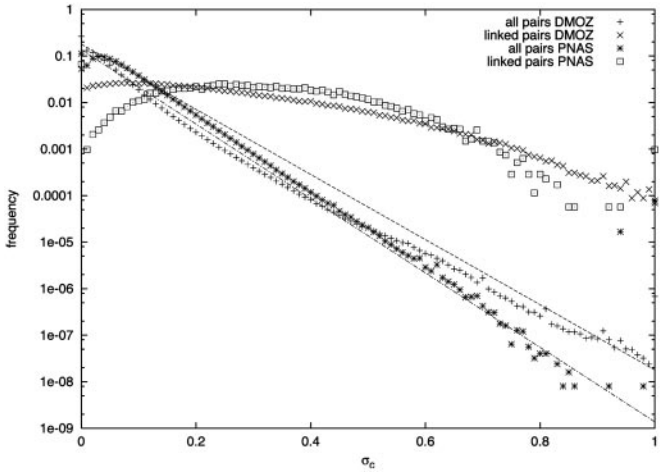

Fig. 2. Content similarity distributions for web pages (DMOZ) and scientific articles (PNAS). The distributions across linked documents are clearly different from the background distributions. The latter are exponential for both data sets (exponential fit curves appear as lines in the log-linear plot).

Validating Prior Models

Given that all of the models described in *Background* can predict the degree distribution of web pages and scientific articles, which is most plausible and/or powerful in explaining the emerging topology of information networks? To answer this question, we need an independent observation from the data, in addition to degree. I turned to the distribution of content similarity between linked (neighbor) pages. Fig. 2 demonstrates that this distribution is qualitatively different from the background similarity distribution for both web pages and scientific articles, clearly indicating that content must play a role in the evolution of information networks. I then looked for a model capable of predicting both the degree distributions and the similarity distributions among linked documents. Such a model would be more plausible than models predicting degree alone.

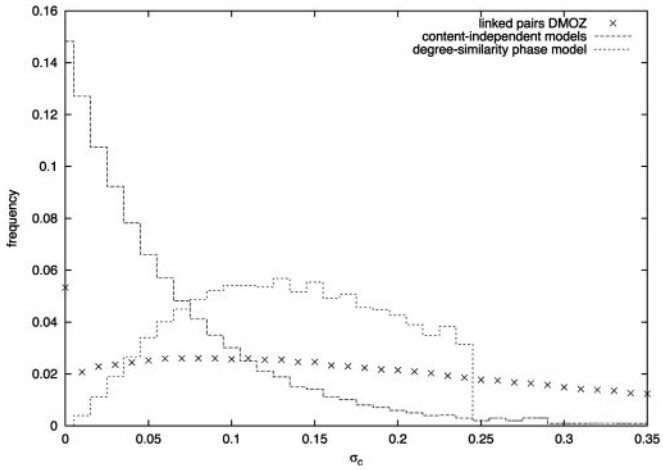

Fig. 3. Content similarity distributions across linked pages generated by simulating various growth models for web pages, compared with DMOZ data. In all simulations the parameters are set to match or fit the DMOZ data: n = 109,648 nodes, m = 15 links, and, in the degree-similarity phase simulation, κ^* = 0.25 and γ = 1.7.

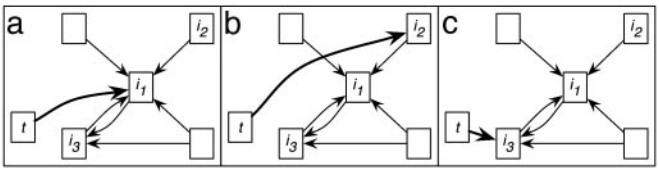

Fig. 4. (*a*) In all preferential attachment models, a new page *t* is likely to be linked to a page with high degrees such as i_1. (*b*) In the degree-uniform mixture model, there is also some probability to link to any random page such as i_2. (*c*) In the degree-similarity mixture model, *t* is more likely to be linked to a page that has high degree but is close in content space, i.e., similar to *t*, such as i_3.

The models in *Background* were validated by numerical simulations, with content similarity drawn from the exponential distributions obtained by fitting the background distributions in Fig. 2: $\Pr(\sigma_c) \sim 10^{-\mu\sigma_c}$ where $\mu = 7$ for web pages and $\mu = 8$ for PNAS articles. The degree distribution of the data are well matched by those generated by the mixture model (25) and the degree-similarity phase model (28). On the other hand, as shown in Fig. 3, for web pages the growth models that do not consider content (13, 20, 25) predictably generate similarity distributions across linked pages that mirror the background exponential distribution. The degree-similarity phase model does somewhat better, but it generates a distribution that goes to zero too rapidly for small and large σ_c. Results are analogous for PNAS articles. Thus, none of the growth models outlined above generates distributions of textual similarity across linked documents in qualitative agreement with the data.

Degree-Similarity Mixture Model

The class of mixture models has a free parameter that can be tuned to fit the data. At each step, one new document is added, and *m* new links or references are created from it to existing documents. At time *t* the probability that the *i*th document is selected and linked from the *t*th document is

$$\Pr(i) = \alpha \frac{k(i)}{mt} + (1 - \alpha)\overline{\Pr}(i)\,, \qquad [5]$$

where $i < t$ and $\alpha \in [0,1]$ is a preferential attachment parameter (Fig. 4*a*). In the degree-uniform mixture model (25) $\overline{\Pr}(i) = 1/t$, the uniform distribution (Fig. 4*b*). Let us now introduce an alternative degree-similarity mixture model in which

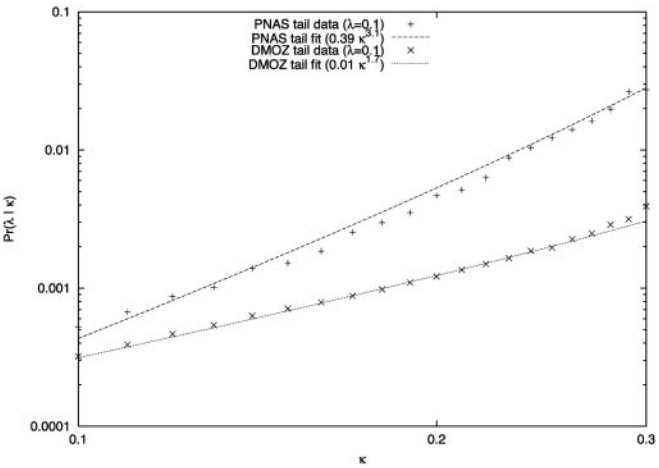

Fig. 5. Tails of conditional link probability $\Pr(\lambda|\kappa)$ as a function of κ for pairs of web pages (DMOZ) and PNAS articles, with power-law fit exponents $\gamma = 1.7$ and $\gamma = 3.1$ respectively.

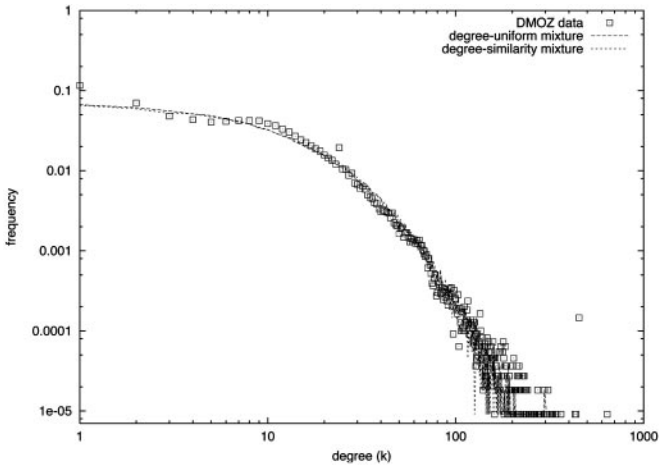

Fig. 6. Degree distributions of web pages predicted by simulating the two mixture models. In the degree-uniform mixture simulation, $\alpha = 0.3$; in the degree-similarity simulation, $\alpha = 0.2$ and $\gamma = 1.7$. All parameters were set by fitting the DMOZ data.

$$\overline{\Pr}(i) \propto \left(\frac{1}{\sigma_c(i,t)} - 1\right)^{-\gamma}, \qquad [6]$$

where γ is a constant (Fig. 4*c*). Like the degree-similarity phase model, this model is inspired by the idea that authors tend to link new documents to popular and related ones and by the observation that link probability between two web pages decays with decreasing similarity as a power-law $\Pr(\lambda|\kappa) \sim \kappa^\gamma$ with $\gamma = 1.7$ for $\lambda = 0.1$ (Fig. 5). However, the free parameter α in the degree-similarity mixture allows us to explicitly model the trade-off between linking to related (similar) versus popular (high-degree) documents.

To validate the degree-similarity mixture model, the networks of web pages and PNAS articles were built by simulation and compared with those obtained by simulating the degree-uniform mixture model. Figs. 6 and 7 show the predictions generated for web pages. Although both models accurately predict the degree distribution, only the degree-similarity mixture model reasonably approximates the similarity distribution.

The PNAS article data were analyzed analogously to the DMOZ data, yielding a conditional citation probability with a tail that scales as a power-law $\Pr(\lambda = 0.1|\kappa) \sim \kappa^\gamma$ just as for web

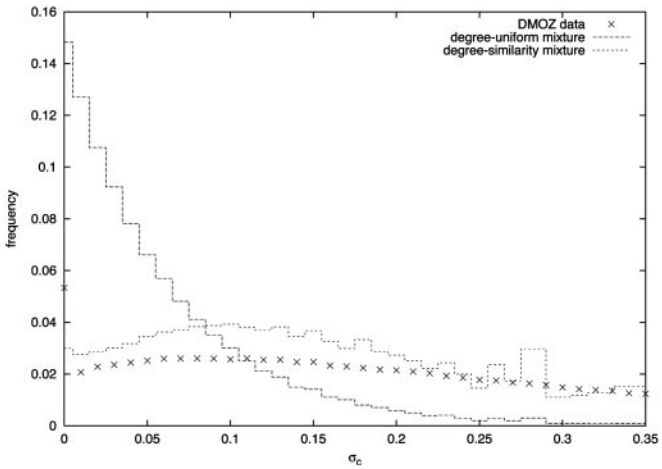

Fig. 7. Distribution of content similarity among linked web pages predicted by simulating the two mixture models.

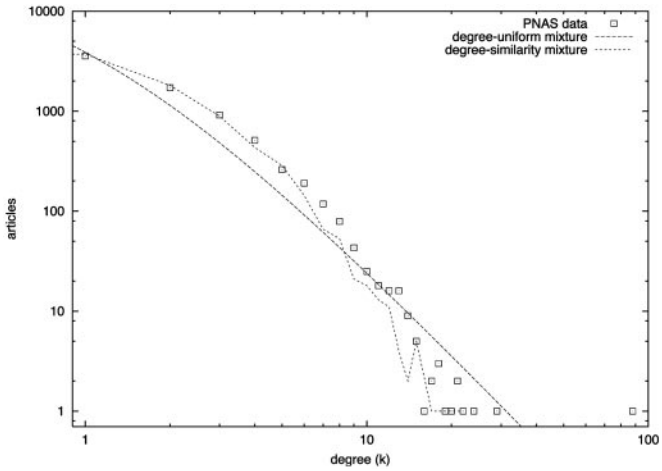

Fig. 8. Distributions of degree (citation count) of PNAS articles, as predicted by the two mixture models. In all of the simulations, $n = 15{,}785$ nodes and $m = 1$ reference. In the degree-uniform mixture simulation, $\alpha = 0.5$; in the degree-similarity simulation, $\alpha = 0.1$ and $\gamma = 3.1$. All parameters were set by matching or fitting the PNAS data (only references to other papers in the PNAS collection were considered).

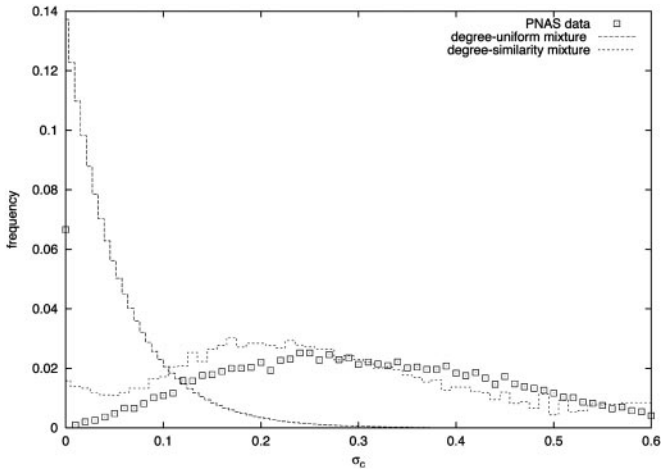

Fig. 9. Distribution of content similarity among titles and abstracts of articles that cite one another, as predicted by the two mixture models.

pages, but with a larger exponent $\gamma = 3.1$ (Fig. 5). Figs. 8 and 9 show the predictions generated by simulating the growth of the PNAS article network according to the two mixture models. Both models accurately predict the distribution of citation counts, although the degree-similarity model fits the data better. Again, the degree-similarity mixture model generates a similarity distribution in remarkable agreement with the data.

Conclusion

In this paper I have shown that existing growth models for document networks generate the wrong predictions for the distribution of content similarity across linked documents. Models that do not take content into account yield distributions that are heavily skewed toward low similarities because those are exponentially more frequent in the data. On the contrary, if authors tend to link and cite related documents, one would expect a similarity distribution with a fatter tail and a peak for $\sigma_c > 0$, as displayed by both web pages and scientific articles. This behavior is captured by the degree-similarity mixture model.

The similarity measures and document representations used in the presented analysis and model are quite crude. Cosine similarity here is based on simple term frequencies. The Jaccard coefficient for link similarity not only is a rough approximation of link probability for web pages but is further limited by incomplete knowledge owing to the necessary reliance on search engines for inlink information. One natural direction for future work is to repeat the analysis and validate the degree-similarity mixture model by using more sophisticated document representations and similarity measures (2, 3, 29, 30). Another is to extend the validation of the model to see whether it can predict additional properties of the networks, such as clustering coefficient and degree correlation (18, 22). Finally, further insight must be gained by studying the relationship between the mechanism studied here (linking similar documents) and other processes likely to play a role in the evolution of link/citation networks, such as copying (22) and coauthorship (17, 18).

The results presented here strongly suggest that page content cannot be neglected when we try to understand the evolution of document networks. The tension between referring to popular versus related documents provides us with a plausible and unified model of how authors link nodes in such different networks as the web and the scientific literature. This model generates remarkably accurate predictions of how such a process can lead to the emergent link and content structure of document spaces.

I thank Jon Kleinberg, Rob Axtell, David Aldous, László Barabási, Reka Albert, Mark Newman, Lada Adamic, Alessandro Vespignani, and Katy Börner for reviewing an earlier draft of this manuscript; Rich Shiffrin and two anonymous reviewers for providing helpful suggestions; the Open Directory Project for the DMOZ data; and the National Academy of Sciences for the PNAS data. This work was supported by National Science Foundation Career Award IIS-0133124/0348940.

1. Salton, G. & McGill, M. (1983) *An Introduction to Modern Information Retrieval* (McGraw–Hill, New York).
2. Belew, R. (2000) *Finding Out About: A Cognitive Perspective on Search Engines and the WWW* (Cambridge Univ. Press, Cambridge, U.K.).
3. Börner, K., Chen, C. & Boyack, K. (2003) *Annu. Rev. Inf. Sci. Technol.* **37,** 179–255.
4. Brin, S. & Page, L. (1998) *Comput. Networks ISDN Syst.* **30,** 107–117.
5. Kleinberg, J. (1999) *J. Assoc. Comput. Mach.* **46,** 604–632.
6. Mendelzon, A. & Rafiei, D. (2000) *IEEE Data Eng. Bull.* **23,** 9–16.
7. Kleinberg, J. & Lawrence, S. (2001) *Science* **294,** 1849–1850.
8. Girvan, M. & Newman, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8271–8276.
9. Henzinger, M. & Lawrence, S. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5186–5191.
10. Hopcroft, J., Khan, O., Kulis, B. & Selman, B. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5249–5253.
11. Albert, R., Jeong, H. & Barabási, A.-L. (1999) *Nature* **401,** 130–131.
12. Huberman, B. & Adamic, L. (1999) *Nature* **401,** 131.
13. Barabási, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
14. Broder, A., Kumar, S., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000) *Comput. Networks ISDN Syst.* **33,** 309–320.
15. Adamic, L. & Huberman, B. (2000) *Science* **287,** 2115.
16. de Solla Price, D. (1965) *Science* **149,** 510–515.
17. Newman, M. E. J. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5200–5205.
18. Börner, K., Maru, J. T. & Goldstone, R. L. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5266–5273.
19. Dorogovtsev, S. & Mendes, J. (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ. Press, Oxford).
20. Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999) *Lect. Notes Comput. Sci.* **1627,** 1–18.
21. Kumar, S., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. & Upfal, E. (2000) in *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science* (IEEE Comput. Soc., Silver Spring, MD), pp. 57–65.
22. Vazquez, A. (2003) *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* **67,** 056104.

23. Dorogovtsev, S., Mendes, J. & Samukhin, A. (2000) *Phys. Rev. Lett.* **85,** 4633–4636.
24. Cooper, C. & Frieze, A. (2001) in *Proceedings of the 9th Annual European Symposium on Algorithms*, Lecture Notes in Computer Science, ed. Meyer auf der Heide, F. (Springer, Berlin), Vol. 2161, pp. 500–511.
25. Pennock, D., Flake, G., Lawrence, S., Glover, E. & Giles, C. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 5207–5211.
26. Fabrikant, A., Koutsoupias, E. & Papadimitriou, C. (2002) in *Proceedings of the 29th International Colloquium on Automata, Languages, and Programming*, Lecture Notes in Computer Science, eds. Widmayer, P., Triguero, F., Morales, R., Hennessy, M., Eidenbenz, S. & Conejo, R. (Springer, Berlin), Vol. 2380, pp. 110–122.
27. Aldous, D. (2003) arXiv.:cond-mat/0304701
28. Menczer, F. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14014–14019.
29. Ganesan, P., Garcia-Molina, H. & Widom, J. (2003) *Assoc. Comput. Mach. Trans. Inf. Syst.* **21,** 64–93.
30. Landauer, T. K., Laham, D. & Derr, M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5214–5219.