Growing and navigating the small world Web by local content

Filippo Menczer[†]

Department of Management Sciences, University of Iowa, Iowa City, IA 52242

Edited by Elwyn R. Berlekamp, University of California, Berkeley, CA, and approved August 26, 2002 (received for review June 10, 2002)

Can we model the scale-free distribution of Web hypertext degree under realistic assumptions about the behavior of page authors? Can a Web crawler efficiently locate an unknown relevant page? These questions are receiving much attention due to their potential impact for understanding the structure of the Web and for building better search engines. Here I investigate the connection between the linkage and content topology of Web pages. The relationship between a text-induced distance metric and a link-based neighborhood probability distribution displays a phase transition between a region where linkage is not determined by content and one where linkage decays according to a power law. This relationship is used to propose a Web growth model that is shown to accurately predict the distribution of Web page degree, based on textual content and assuming only local knowledge of degree for existing pages. A qualitatively similar phase transition is found between linkage and semantic distance, with an exponential decay tail. Both relationships suggest that efficient paths can be discovered by decentralized Web navigation algorithms based on textual and/or categorical cues.

The link structure of the Web is attracting considerable attention. Several studies have shown that the degree sequence of Web pages has a power-law distribution, $\Pr(k) \sim k^{-\gamma}$ where k is the degree of a page (number of inlinks or outlinks) and γ is a constant exponent (1–5). Two aspects of this scale-free link topology of the Web that may have important applications, for example, in the design of the next generation of Web search tools, are (*i*) growth models to explain the emergence of popular Web sites and (*ii*) navigation models to build effective crawlers.

The first goal of this article is to develop a generative model that predicts Web degree distributions based on realistic assumptions about what motivates authors to pick the sites to which they link their pages. I do not assume global knowledge of degree or other link information and I aim to tie the attachment process to a plausible content-based metric, modeling the author's intent to link a new page to existing sites that are both popular and related semantically. In the next section I report on a surprising relationship between the probability that two pages belong to a common link neighborhood and a content-based distance measure. A phase transition is observed between a region where linkage is not determined by content and one where linkage decays according to a power law with increasing content-based distance. Such a finding motivates a generative model based on this relationship and on the observed distribution of the content-based distance measure among pairs of Web pages. The model is described in the third section, where it is shown to accurately predict the distribution of degree in a large Web sample. There have been some other, more theoretical efforts recently to unify Web content and link generation based on latent semantic and link eigenvalue analysis (6, 7). My finding and model are motivated entirely by more empirical observations and actual data.

The second goal of this article is to tie the theoretical existence of optimal navigation algorithms for small world networks and the actual viability of efficient content-driven Web crawlers. Given the Web's small world and power-law topology (1, 2, 4, 5), its diameter scales as $\Theta(\log N/\log\log N)$ (8); therefore, if two pages belong to a connected component of the Web, some short path exists between them. Can a crawler navigate such a short path? In the fourth section I show that although purely link-based algorithms can be efficient in terms of path length (9, 10), the number of pages that must actually be crawled to determine the path is unreasonably large. I then draw a connection between the link-content analysis and a theoretical result regarding the existence of efficient decentralized algorithms to navigate small world networks in which links are generated based on a geographic topology (11, 12). An analogous connection is drawn by discovering an exponential relationship between link neighborhood probability and semantic distance between pages. My measurements reveal that the Web's linkage topology is consistent with network models where links are generated based on a topical hierarchy, for which the existence of efficient navigation algorithms is also known (13, 36). These findings suggest that realistic Web crawling algorithms based on textual and/or categorical cues can efficiently discover relevant pages and are consistent with data from state-of-the-art crawling algorithms.

Power-Law Relationship Between Web Content and Linkage

To gain insight into the Web's scale-free growth and mechanisms for efficient navigation, I want to study the connection between the two topologies induced over the Web by links and textual content. I start by introducing a distance measure based on the lexical similarity between the textual content of pages. Let us define such a lexical distance

$$r(p_1, p_2) = \frac{1}{s(p_1, p_2)} - 1, \qquad [1]$$

where (p_1, p_2) is a pair of Web pages and

$$s(p_1, p_2) = \frac{\sum_{k \in p_1 \cap p_2} w_{kp_1} w_{kp_2}}{\sqrt{\left(\sum_{k \in p_1} w_{kp_1}^2\right)\left(\sum_{k \in p_2} w_{kp_2}^2\right)}} \qquad [2]$$

is the cosine similarity function traditionally used in information retrieval (w_{kp} is some weight function for term k in page p, e.g., term frequency). The r distance measure is a natural local cue readily available in the Web, with the target content specified by a query, topic, or bookmark page of interest to the user. This measure also does not suffer from the dimensionality bias that makes L-norms inappropriate in the sparse word vector space.

To investigate the relationship between the lexical topology induced by r and the link topology, it would be desirable to measure the probability that two pages at a certain lexical distance from each other have a direct link between them. Unfortunately, such a probability is extremely difficult to measure directly because the low ratio between the average degree of Web pages and the size of the Web makes the likelihood that

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: BA, Barabási-Albert.

[†]E-mail: filippo-menczer@uiowa.edu.

 www.pnas.org/cgi/doi/10.1073/pnas.212348399

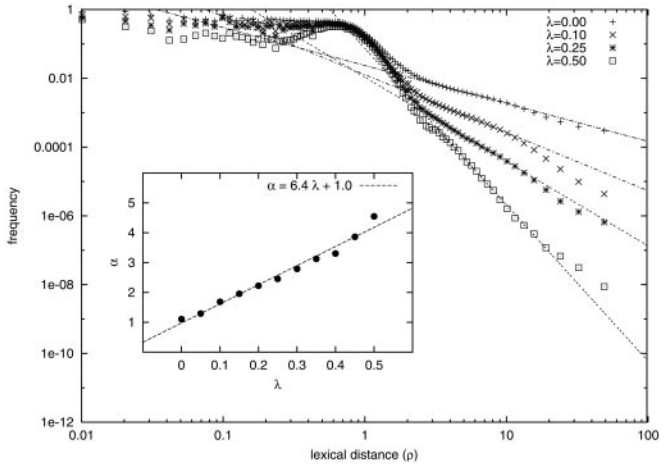

Fig. 1. Cluster probability $Pr(\lambda|\rho)$ as a function of lexical distance ρ, for various values of the linkage threshold λ. A nonlinear least-squares fit of the tail of each distribution to the power law model $Pr(\lambda|\rho) \sim \rho^{-\alpha}$ is also shown. (*Inset*) Plot of the relationship between the linkage threshold λ and the clustering exponent α of the power-law tail. The frequency data are based on $\approx 4 \times 10^9$ pairs of Web pages sampled from the Open Directory Project (http://dmoz.org).

two random pages link to each other negligibly small. Therefore let us instead focus on a neighborhood relation in link space, which approximates link probability but is easier to measure and is also used to identify Web communities (4, 14). I measured the frequency of neighbor pairs of pages as a function of the lexical distance:

$$\Pr(\lambda|\rho) = \frac{|(p_1, p_2) : r(p_1, p_2) = \rho \wedge l(p_1, p_2) > \lambda|}{|(p_1, p_2) : r(p_1, p_2) = \rho|}, \qquad [3]$$

where the neighborhood between two pages is expressed by the function

$$l(p_1, p_2) = \frac{|U_{p_1} \cap U_{p_2}|}{|U_{p_1} \cup U_{p_2}|} \qquad [4]$$

and U_p is the URL set representing p's neighborhood (inlinks, outlinks, and p itself). Note that l is akin to the well known cocitation and bibliographic coupling measures used in directed graphs for outlinks and inlinks, respectively. The neighborhood threshold λ models the ratio of local versus long-range links, where we think of local links as those characteristic of clustered nodes and long-range links as those characteristic of random connections. In small world networks such as the Web, local and random structures coexist in the form of node clusters (communities) connected by high-degree nodes (hubs).

Fig. 1 plots $Pr(\lambda|\rho)$ versus ρ for various values of λ.[‡] We observe an interesting phase transition between two distinct regions around a critical distance ρ^* that does not depend on λ ($\rho^* \approx 1$). For $\rho < \rho^*$ the probability that two pages are neighbors does not seem to be correlated with their lexical distance; for $\rho > \rho^*$ the probability decreases according to a power law

$$\Pr(\lambda|\rho) = c_1 \rho^{-\alpha(\lambda)} \qquad [5]$$

so that that the majority of clustered pages are lexically similar, but there is a long tail of pages that are clustered despite a very large lexical distance. These could be clusters around very popular hub pages or communities where textual content fails to capture the semantic relationship between pages. A fit of the tails to the power-law model reveals that the clustering exponent α grows roughly linearly with the linkage threshold λ for $\lambda \leq 0.5$ (see Fig. 1 *Inset*):

$$\alpha(\lambda) \approx 6.4171\lambda + 0.9735. \qquad [6]$$

For $\rho > \rho^*$ the relationship between $Pr(\lambda|\rho)$, ρ, and λ is qualitatively consistent with the intuitive correlation between the lexical and linkage topologies of the Web; pages that are semantically related tend to be both lexically similar and clustered, therefore pages that are more similar in content have a higher likelihood to be neighbors. However, for sufficiently small lexical distances ($\rho < \rho^*$) there is no additional information about link clustering to be gained from content.

Web Growth

Background. Several models have been proposed to interpret the power-law distribution of Web page degree, for example, based on stochastic growth rates of Web sites (3). Here I focus on generative models, which attempt to explain the Web topology based on the behavior of individual authors linking new pages to existing ones. Most generative models are based on preferential attachment, whereby one node at a time is added to the network with m new edges to existing nodes selected according to some probability distribution, typically a function of some characteristic of the existing nodes. The best known preferential attachment model is the Barabási-Albert (BA) model, where a node i receives a new edge with probability proportional to its current degree, $\Pr(i) \propto k(i)$ (2, 15). The BA model generates network topologies with power-law degree distributions, but is based on the unrealistic assumption that Web authors have complete global knowledge of the Web degree. According to the BA model, links are generated according to popularity alone, and consequently the oldest nodes are those with highest degree.

To give newer nodes a chance to compete for links, some extensions of the BA model use $\Pr(i) \propto \eta(i)k(i)$, where $\eta(i)$ is the fitness of page i. These models still yield power-law degree distributions, but after enough time the winning pages are those with highest fitness (16, 17). Another class of variations that allow new pages to compete for links is based on linking to a node based on its degree with probability ψ or to a uniformly chosen node with probability $1 - \psi$ (18–20). Such a mixture model generates networks that can fit not only the power-law degree distribution of the entire Web, but also the unimodal degree distribution of subsets of Web pages such as university, company, or newspaper homepages (20). Unfortunately, all of these models still rely on global knowledge of degree. Furthermore, they fail to capture the cognitive processes that lead authors to pick pages to link. Neither the global fitness measure nor the uniform distribution match the heterogeneous nature of the Web and of the authors' interests.

The "copying" model (21, 22) implements a rich-get-richer process equivalent to the BA model, but without requiring explicit global knowledge of node degrees. For each new node, an existing prototype node i is first chosen at random. Then each link j from the new node is either attached to a uniformly chosen node with probability φ, or to the target of i's jth link with probability $1 - \varphi$. Higher-degree nodes are automatically favored by the copying mechanism, producing the same power-law degree distributions as the BA model. The prototype node could correspond to a related page known to the author, thus modeling some sort of local selection process.

A generative model that uses local information even more explicitly was recently proposed for trees (23). Nodes are given random coordinates in the unit square. Then new nodes are

[‡]The data were collected from a uniform random sample of 150,134 URLs extracted from 47,174 distinct categories in the Open Directory Project snapshot of February 14, 2002. Further details of the data collection procedure can be found in ref. 33.

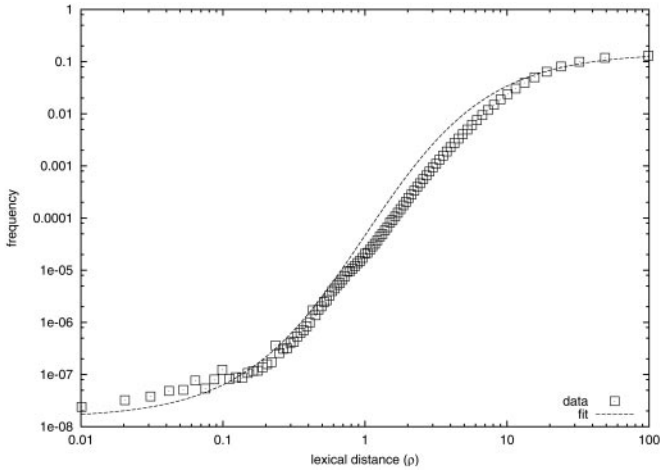

Fig. 2. Probability distribution of lexical distance, Pr(ρ). The frequency data are based on $\approx 4 \times 10^9$ pairs of Web pages sampled from the Open Directory Project (http://dmoz.org). A nonlinear least-squares fit of the distribution to the exponential model $\Pr(\rho) \sim \beta^{-\frac{1}{1+\rho}}$ is also shown.

attached to existing ones based on a linear combination of linkage and geographic bias. In this model a new node j is deterministically linked to the node arg $\min_i(\phi r_{ij} + g_i)$, where r_{ij} is the Euclidean distance and g_i is a "centrality" measure of i in the tree. It is shown that for critical values of ϕ ($\bar{\phi} \leq \phi = o(\sqrt{N})$, where $\bar{\phi}$ is a constant and N is the number of nodes), this model yields a power-law degree distribution. The Euclidean distance here is independent of link topology and thus could be used to model the factors that lead authors to link their pages to other pages based on, say, content rather than popularity alone. However, the linear combination model still requires global knowledge of the tree structure, and the uniform distribution of nodes in the unit square as a geographic model does not capture a realistic content topology.

Content-Based Generative Model. The phase transition in Fig. 1 suggests a straightforward way to interpret the growth of the Web's link structure based on the content similarity between pages. I want to model an author's desire to link new pages to sites that are both similar (hence probably related) and popular (hence probably important). Let us assume in keeping with current models that page importance or popularity is correlated with degree. However, I do not assume that an author has global knowledge of degree; instead, an author has only local knowledge of degree, i.e., knowledge of importance only for pages with similar content. This is quite realistic as such pages are probably known to the author or else can be discovered simply by using a search engine.

I propose a generative model based on the above assumptions. At each step t one new page p_t is added, and m new links are created from p_t to m existing pages $\{p_i, i < t\}$, each selected with probability:

$$\Pr(p_i, t) = \begin{cases} \dfrac{k(i)}{mt} & \text{if } r(p_i, p_t) < \rho^* \\ c_1 r^{-\alpha}(p_i, p_t) & \text{otherwise} \end{cases}, \qquad [7]$$

where $k(i)$ is the indegree of p_i,[§] ρ^* is a lexical distance threshold, and c_1 and α are constants. This growth process is driven by local link decisions based on content. It is consistent with the phase transition of Fig. 1: lexical independence for close pages and an inverse power-law dependence for distant pages (where I model the link probability after the neighborhood probability expressed by Eq. **5**).

[§]This formulation is for directed graphs; in the indirected case the degree normalization factor is $2mt$.

Table 1. Parameter values for generative model simulation

Parameter	Value	Source (Open Directory data)
ρ^*	3	Fig. 1
c_1	0.01282	Fig. 1, Eq. **5**, $\lambda = 0.1$
α	1.68894	Fig. 1, Eq. **5**, $\lambda = 0.1$
c_2	0.14723	Fig. 2, Eq. **9**, $\int_0^1 \Pr(\rho)d\rho = 1$
β	10^7	Fig. 2, Eq. **9**
N	109,648	Fig. 3
m	15	Fig. 3

To test this model one needs a prior distribution for the values of r across Web pages. The frequency

$$\Pr(\rho) = \frac{|(p_1, p_2) : r(p_1, p_2) = \rho|}{|(p_1, p_2)|} \qquad [8]$$

measured from a large number of pairs of Web pages is shown in Fig. 2. I used a nonlinear fit of this frequency data to the exponential model

$$\Pr(\rho) = c_2 \beta^{-\frac{1}{1+\rho}} \qquad [9]$$

as a PDF to generate realistically distributed random $r(p_i, p_j)$ values for any pair (p_i, p_j), $i < j$. I then simulated the generative model of Eq. **7** to build a network with N nodes. The parameters of the simulation are shown in Table 1.

Fig. 3 shows the degree distribution obtained by the simulation of the local content-based generative model and compares it with the distribution measured from a sample of Web pages. Given the relatively small N and the fact that the sampled pages tend to be quite specific, the degree distribution of the data diverges significantly from the "pure" power laws reported for massive Web crawls (5). The general mixture model (20) matches the data very well for a preferential attachment coefficient $\psi \approx 0.3$. The novel result here is that the degree distribution generated by the local content-based

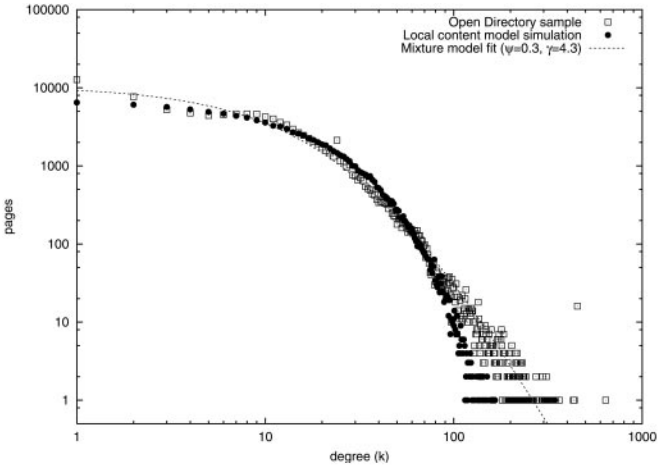

Fig. 3. Degree sequence distributions $N \Pr(k)$ generated by a simulation of the local content-based generative model and by a sample of $N = 109{,}648$ Web pages from the Open Directory Project (http://dmoz.org). A nonlinear least-squares fit of the Open Directory Project data to the mixture model $\Pr(k) = [m(1 - \psi)]^{\frac{1}{\psi}}/[\psi k + m(1 - \psi)]^{1+\frac{1}{\psi}}$ (adapted from ref. 20 for directed links) is also shown, yielding $\psi \approx 0.3 (\gamma = 1 + 1/\psi \approx 4.3)$.

Table 2. Asymptotic power law exponents from least-squares fit of degree distribution tails, with standard errors

Distribution	$\bar{\gamma}$	$\sigma_{\bar{\gamma}}$
Open Directory sample	4.28	0.14
Mixture model (20)	4.31	0.22
Local content model	4.26	0.07

model also yields a very accurate fit of the data. The tails of all three distributions are consistent with a single power law with exponent $\gamma \approx 4.3$, as shown in Table 2.

Web Navigation

Background. The relationships between Web link topology and notions of semantic similarity stemming from page content or topic classification have important applications for the design of more effective search tools (24, 25). Topic-driven crawlers (26–30) are increasingly seen as a way to address the scalability limitations of current universal search engines, by distributing the crawling process across users, queries, or even client computers. The context available to such crawlers can guide the navigation of links with the goal of efficiently locating highly relevant target pages. Given the need to find unknown target pages, we are interested only in decentralized crawling algorithms, which can use only information available locally about a page and its neighborhood. Starting from some source Web page we aim to visit a target page by visiting $\ell \ll N$ pages, where N is the size of the Web, several billion pages.

Since the Web is a small world network (1, 2, 4, 5) we know that its diameter [at least for the largest connected component (5, 31)] scales logarithmically with N, or more precisely a short path of length

$$\ell' \sim \frac{\log N}{\log \log N} \qquad [10]$$

is likely to exist (8) between some source (e.g., a bookmarked page or a hit returned by a search engine) and some unknown relevant target page. Can a crawler navigate such a short path?

If the only local information available is about the hypertext link degree of each node and its neighbors, then simple greedy algorithms that always pick the neighbor with highest degree lead to paths whose length ℓ' scales logarithmically ($\ell' \sim \log N$) (10) or sublinearly ($\ell' \sim N^{\varepsilon}$, $\varepsilon \approx 0.7 < 1$) (9). However, a real Web crawler would have to visit all of the neighbors of a page to determine their degree. Due to the power-law degree distribution, moving to high-degree nodes leads to covering most of the nodes in a short number of steps. For example, simulations suggest that the number of steps to cover $N/2$ nodes scales as $N^{0.24} < \ell'$ (9). Therefore a small ℓ' implies a large ℓ ($\ell \sim N$), making the degree-seeking strategy too costly for actual Web navigation.

Kleinberg (11–13) and Watts *et al*. (36) characterized certain classes of networks whose link topology depends on external geographic or hierarchical structures, showing that in these special cases navigation algorithms are guaranteed to exist with polylogarithmic bounds $\ell = O(\log^{\varepsilon} N)$, $\varepsilon \geq 1$. I now outline these models and show that they can be applied to the Web.

Content-Based Crawling Algorithms. If the link topology of a network follows a D-dimensional lattice, with local links to lattice neighbors plus some long-range connections, and information about the geographic location of nodes is available, Kleinberg (11, 12) proved that a greedy algorithm, which always picks the neighbor geographically closest to the target, yields a bound $\ell = O(\log^2 N)$. In this geographic model the optimal path length is achieved if, and only if, long-range links are chosen with probability $\Pr(r) \sim r^{-\alpha}$, where r is the lattice (L_1) distance between the two nodes, and α is a critical constant clustering exponent dependent on the dimensionality of the lattice ($\alpha = D$).

While Kleinberg's geographic model is inspired by social small world networks where location knowledge exists, the lattice model cannot be applied directly to the Web where such a notion of L_1 norm-based geography is unrealistic. However, Kleinberg's result would be applicable to the Web if one could use a realistic topological distance measure for r. In fact, the result of the second section, namely the power law in Eq. **5** obtained from the tail of the conditional probability distribution in Fig. 1, makes the distance induced by the lexical similarity between the textual content of pages an obvious candidate for r. In the lattice model, local links are equiprobable whereas long-range links are governed by the power-law distribution. Analogously, in the Web lexical similarity defines the geographic topology of the pages and the critical distance ρ^* marks the border between local links (whose probability is independent of r) and long-range links. Therefore my finding suggests that Kleinberg's analysis based on geographic networks can be applied to the Web to infer the existence of efficient crawling algorithms. There is one caveat. Kleinberg's result requires a critical clustering exponent $\alpha = D$, but what is the meaning of D in the Web? One could imagine defining an analogous "Web dimensionality" based, say, on the communities induced by the Web's link structure (14). Fortunately, it is not necessary to quantify such a dimensionality to extend Kleinberg's analysis to the Web; it is sufficient to assume that the dimensionality can be defined by a "critical neighborhood threshold" λ^*. Then Eq. **6** can be used to obtain a corresponding critical clustering exponent $\alpha^* = \alpha(\lambda^*)$. Since the power-law relationship of Eq. **5** holds over a wide range of λ values (compare Fig. 1), it is reasonable to conjecture the existence of a decentralized Web navigation algorithm that can be proven to be optimally efficient under the linkage topology induced by λ^* and the geographic topology induced by r.

A qualitatively similar result (the existence of polylogarithmic navigation algorithms, $\ell = O(\log^{\varepsilon} N)$ was given by Kleinberg for the case where information about a semantic hierarchy is available, where nodes are classified into the leaves of a topical tree and links are drawn from a critical probability distribution based on the semantic distance between nodes induced by the tree (13). The link probability distribution must have an exponential tail $\Pr(h) \sim \xi^{-h}$ where ξ is a constant and h is the semantic distance between two nodes, expressed by the height of the lowest common ancestor in the tree. The hierarchical model is more plausible than the geographic model for the Web because directories such as Yahoo and the Open Directory Project can play the role of semantic hierarchy trees. An almost identical hierarchical model was proposed by Watts *et al*. (36) with an even stronger (constant) bound on navigation time.

I have previously studied the relationship between link topology and semantic similarity by analyzing the probability of finding links to a set of pages on a certain topic as one crawls away from that set. This probability remains remarkably constant in exhaustive breadth-first crawls of depth up to three links starting from pages classified by Yahoo (34). This finding has been confirmed and extended by recent experiments based on crawls from the Open Directory Project, showing that the linkage probability is significantly higher between pages on the same topic than between pages on different topics (25). To investigate whether the hierarchical model analysis can be applied to the Web, let us define a semantic distance between topics:

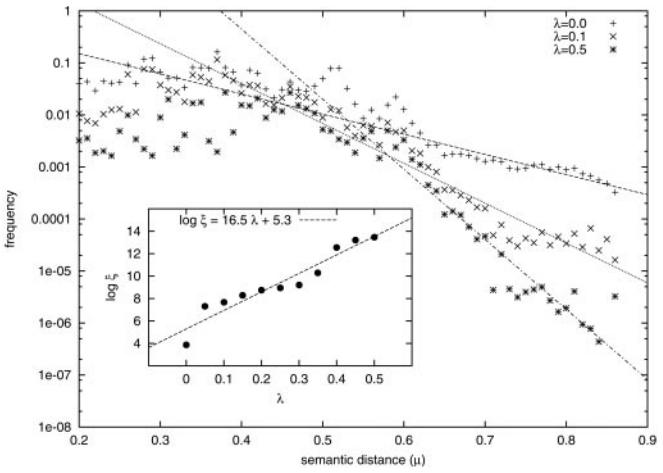

Fig. 4. Tail of cluster probability $\Pr(\lambda|\mu)$ as a function of semantic distance μ, for various values of the linkage threshold λ. A nonlinear least-squares fit of each tail to the exponential model $\Pr(\lambda|\mu) \sim \xi^{-\mu}$ is also shown. (*Inset*) Plot of the relationship between the linkage threshold λ and the base ξ of the exponential tail. The data are based on the same Open Directory Project sample used for Fig. 1.

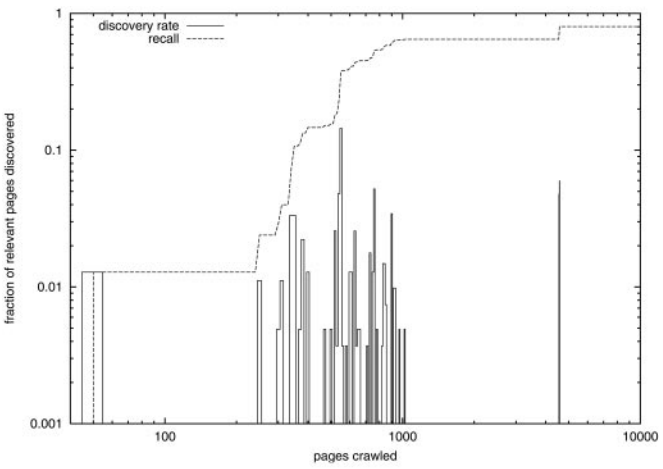

Fig. 5. Distribution of discovery times for unknown relevant pages obtained by a crawler using a content-driven (and slightly less greedy) version of the algorithm described in ref. 12. The crawler is given a short topic query and a limited store for unvisited URLs and starts at least three links away from any relevant pages.

$$h(p_1, p_2) = \frac{\log \Pr(p_1) + \log \Pr(p_2) - 2 \log \Pr(p_0)}{\log \Pr(p_1) + \log \Pr(p_2)}, \quad [11]$$

where p_0 is the lowest common ancestor of p_1 and p_2, and $\Pr(p)$ represents the fraction of pages classified at node p. This is a straightforward extension of the information-theoretic semantic similarity measure (32), and it generalizes the hierarchical models to account for the fact that pages are classified by actual Web directories into internal nodes as well as leaves.

The relationship between semantic distance h and link topology can be analyzed in the same way as was done for lexical distance r in the second section, by measuring the frequency of neighbor pairs of pages as a function of the semantic distance:

$$\Pr(\lambda|\mu) = \frac{|(p_1, p_2) : h(p_1, p_2) = \mu \wedge l(p_1, p_2) > \lambda|}{|(p_1, p_2) : h(p_1, p_2) = \mu|}, \quad [12]$$

where h was computed by using the Open Directory Project tree.[¶] Fig. 4 plots the tail of $\Pr(\lambda|\mu)$ versus μ for various values of λ. We observe a good fit between the data and the exponential model

$$\Pr(\lambda|\mu) = c_3 \xi(\lambda)^{-\mu} \quad [13]$$

for $\mu > \mu^* \approx 0.5$. The fit also reveals that the base $\xi(\lambda)$ is roughly exponential in λ (see Fig. 4 *Inset*). As for the geographic model, this finding suggests that the analyses based on hierarchical networks can be applied to the Web to conclude that efficient crawling algorithms exist. In fact, the crawler proposed in ref. 27 attempts to prioritize links based on h estimates obtained from a classifier and thus is an implementation of the optimal greedy algorithm described in ref. 13.

The greedy algorithm proposed by Kleinberg (12) would locate a target page by visiting $\ell \leq c_4 \log^2 N$ pages with a constant $c_4 \approx 128$. For $N \approx 10^{10}$, $\ell \leq 10^4$ pages. This finding is consistent with our experimental data, as illustrated by the discovery time distribution in Fig. 5; the majority of relevant pages are located based on local content before 10^4 pages have been crawled.[‖]

Conclusion

In this article I addressed two questions of considerable interest for potential Web applications: whether one can effectively model the emergence of the scale-free topology of the Web from realistic assumptions about authors' local knowledge and behavior and whether one can design crawling algorithm to efficiently locate unknown pages of interest to a user or search engine.

I found that the answer to both of these questions is yes, and in doing so I uncovered two interesting relationships between the Web's link topology and distance metrics based on lexical and semantic similarity. The two relationships (compare Figs. 1 and 4) are qualitatively very similar: in both cases a critical distance (μ^* or ρ^*) marks a phase transition between a relatively flat section at small distances and a decay at large distances. The decay has a long (power law) tail in the case of lexical distance and a short (exponential) tail for semantic distance. These are probably two manifestations of the same behavior. Authors tend to link their pages to semantically related clusters, identified via page content. The link probability decreases rapidly with increasing semantic/lexical distance. But among the most closely related pages, the choice of which pages to link is largely driven by other factors such as authority or popularity.

This analysis led to a generative model for the Web graph based on local content cues, which accurately matches the distribution of degree in a representative sample of Web pages. This generative model yielded accurate predictions of degree sequence based on page content data. My model may help us gain a better understanding of the evolving structure of the Web and its cognitive and social underpinnings and may lead to more effective authoring guidelines as well as to improved ranking, classification, and clustering algorithms.

Finally, the link/lexical power law relationship and the related link/semantic exponential relationship led to analogies between the Web and special classes of geographic and hierarchical graphs for which the existence of optimally efficient (polylogarithmic) navigation algorithms was proven.

[¶]The full Open Directory Project snapshot of February 14, 2002 was used to compute the probabilities associated with each of 97,614 distinct categories containing a total of 896,233 URLs.

[‖]These data are based on crawls for six topics, from experiments described in ref. 35.

These results strongly suggest that short paths can be discovered by decentralized Web crawlers based on textual and/or categorical cues. These data-supported results yielded efficient bounds on the number of pages visited by Web crawlers to reach unknown targets. The field of focused crawlers is gaining much empirical attention owing to its potential to cope with the scalability limitations of current search engine technology. The present findings are consistent with, and give some theoretical grounding to, data from a growing body of work on crawling algorithm design; they may have a large impact toward the construction of effective decentralized search tools.

I am grateful to Jon Kleinberg, Mark Newman, Soumen Chakrabarti, Albert-László Barabási, Lada Adamic, Christos Papadimitriou, Rik Belew, Padmini Srinivasan, Dave Eichmann, Nick Street, Alberto Segre, and Gautam Pant for helpful comments; the paper was greatly improved by the suggestions of an anonymous referee. Thanks to the Open Directory Project for making their data publicly available. This work is funded in part by National Science Foundation Career Grant IIS-0133124.

1. Albert, R., Jeong, H. & Barabási, A.-L. (1999) *Nature* **401,** 130–131.
2. Barabási, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
3. Huberman, B. & Adamic, L. (1999) *Nature* **401,** 131.
4. Kumar, S., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999) *Comput. Networks* **31,** 1481–1493.
5. Broder, A., Kumar, S., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000) *Comput. Networks* **33,** 309–320.
6. Achlioptas, D., Fiat, A., Karlin, A. & McSherry, F. (2001) in *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science* (IEEE Computer Society, Silver Spring, MD), pp. 500–509.
7. Cohn, D. & Hofmann, T. (2001) in *Advances in Neural Information Processing Systems 13*, eds. Leen, T. K., Dietterich, T. G. & Tresp, V. (MIT Press, Cambridge, MA), pp. 430–436.
8. Bollobas, B. (2001) *Random Graphs* (Cambridge Univ. Press, Cambridge, U.K.), 2nd Ed.
9. Adamic, L., Lukose, R., Puniyani, A. & Huberman, B. (2001) *Phys. Rev. E* **64**, 046135, http://link.aps.org/abstract/PRE/v64/e046135.
10. Kim, B., Yoon, C., Han, S. & Jeong, H. (2002) *Phys. Rev. E* **65**, 027103, http://link.aps.org/abstract/PRE/v65/e027103.
11. Kleinberg, J. (2000) *Nature* **406,** 845.
12. Kleinberg, J. (2000) in *Proceedings of the 32nd Association for Computing Machinery Symposium on Theory of Computing*, eds. Yao, F. & Luks, E. (Association for Computing Machinery, New York), pp. 163–170.
13. Kleinberg, J. (2002) in *Advances in Neural Information Processing Systems 14*, eds. Dietterich, T. G., Becker, S. & Ghahramani, Z. (MIT Press, Cambridge, MA), in press.
14. Flake, G., Lawrence, S., Giles, C. & Frans Coetzee, F. (2002) *IEEE Comput.* **35,** 66–71.
15. Barabási, A.-L., Albert, R. & Jeong, H. (1999) *Physica A* **272,** 173–187.
16. Adamic, L. & Huberman, B. (2000) *Science* **287,** 2115.
17. Bianconi, G. & Barabási, A.-L. (2001) *Phys. Rev. Lett.* **86,** 5632–5635.
18. Dorogovtsev, S., Mendes, J. & Samukhin, A. (2000) *Phys. Rev. Lett.* **85,** 4633–4636.
19. 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.
20. Pennock, D., Flake, G., Lawrence, S., Glover, E. & Giles, C. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 5207–5211.
21. Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999) in *Proceedings of the 5th Annual International Conference on Combinatorics and Computing*, *Lecture Notes in Computer Science*, eds. Asano, T., Imai, H., Lee, D., Nakano, S. & Tokuyama, T. (Springer, Berlin), No. 1627, pp. 1–18.
22. 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 Computer Society, Silver Spring, MD), pp. 57–65.
23. 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.
24. Chakrabarti, S., Punera, K. & Subramanyam, M. (2002) in *Proceedings of the 11th International World Wide Web Conference*, eds. Lassner, D., De Roure, D. & Iyengar, A. (Association for Computing Machinery, New York), pp. 148–159.
25. Chakrabarti, S., Joshi, M., Punera, K. & Pennock, D. (2002) in *Proceedings of the 11th International World Wide Web Conference*, eds. Lassner, D., De Roure, D. & Iyengar, A. (Association for Computing Machinery, New York), pp. 251–262.
26. Cho, J., Garcia-Molina, H. & Page, L. (1998) *Comput. Networks* **30,** 161–172.
27. Chakrabarti, S., van den Berg, M. & Dom, B. (1999) *Comput. Networks* **31,** 1623–1640.
28. Menczer, F. & Belew, R. (2000) *Machine Learning* **39,** 203–242.
29. Menczer, F., Pant, G., Ruiz, M. & Srinivasan, P. (2001) in *Proceedings of the 24th Annual International Conference on Research and Development in Information Retrieval*, eds. Kraft, D. H., Croft, W. B., Harper, D. J. & Zobel, J. (Association for Computing Machinery, New York), pp. 241–249.
30. Pant, G. & Menczer, F. (2002) *Autonomous Agents Multi-Agent Systems* **5,** 221–229.
31. Kleinberg, J. & Lawrence, S. (2001) *Science* **294,** 1849–1850.
32. Lin, D. (1998) in *Proceedings of the 15th International Conference on Machine Learning*, ed. Shavlik, J. (Morgan Kaufmann, San Francisco), pp. 296–304.
33. Menczer, F. (2002) preprint, http://dollar.biz.uiowa.edu/~fil/Papers/maps.pdf.
34. Menczer, F. (2002) *Computing Research Repository Technical Report*, http://arxiv.org/cs.IR/0108004.
35. Menczer, F., Pant, G. & Srinivasan, P. (2002) preprint, http://dollar.biz.uiowa.edu/~fil/Papers/TOIT.pdf.
36. Watts, D. J., Dodds, P. S. & Newman, M. E. J. (2002) *Science* **296,** 1302–1305.