Locality, Hierarchy, and Bidirectionality in the Web[*]

Nadav Eiron and Kevin S. McCurley
IBM Almaden Research Center

Abstract

The World Wide Web has been previously observed to be a "small world network" in which nodes are clustered together. We provide evidence, based on a crawl of over a billion pages, that such a clustering effect corresponds very closely to the hierarchical nature of URLs. We also show that bidirectionality on the web graph is much more common than previous models predicted. We then propose a new paradigm for models of the Web that incorporates the hierarchical evolution and structure that is evident in the Web.

1 Introduction

The graph structure of the web contains within it an expression of relationships between information on the Web, and this collective set of relationships has led to several major innovations in information retrieval, including the HITS algorithm[8] and the PageRank algorithm [3]. For this reason, the graph induced by the hyperlink structure of the web represents an intriguing candidate for study and mathematical modeling.

One feature that seems to have been largely ignored in studies of the Web is the inherent hierarchical structure that is evident in the structure of URLs. It is a very common practice for information to be organized in a hierarchical tree structure in a system, with information at the upper levels of the tree being more general than the information at the bottom levels. In the case of paper documents, hierarchical organization of information dates back centuries, and in the case of computer file systems, it dates back to the time of Multics in 1965. This same hierarchical organization shows up on web sites, where for example we might find product information in one directory, support information in another, press releases in another, etc. Moreover, URLs reflect another layer of hierarchy from the domain name system (DNS), where domains are categorized by their top level domain of `edu`, `org`, etc.[1]

In his seminal work on complex systems, Simon [18] argued that all systems tend to organize themselves hierarchically. Moreover, he stated that "If we make a chart of social interactions, of who talks to whom, the clusters of dense interaction in the chart will identify a rather well-defined hierarchic structure." We believe that a similar phenomenon can be seen in the link structure of the World Wide Web, in which a large fraction of the hyperlinks between URLs tend to follow the hierarchical organization. In particular, we shall provide evidence that hyperlinks tend to exhibit a "locality" that is correlated to the hierarchical structure of URLs. The same is true for bidirectionality of hyperlinks. The hierarchical structure is at least as important as the hyperlink structure for understanding the information on the web. Moreover, the interaction between the hierarchical structure and the hyperlink structure reveals even more than the

[*] Extended abstract for WAW 2003

[1] The DNS is not very cleanly defined. For example, companies in the U.K. may have `.com` or `.co.uk` in the suffix of their domain name. For this reason we ignored this level of hierarchy.

individual structures. It is therefore important to understand how they relate to each other, and may be helpful in understanding the significance of hyperlinks and hierarchies themselves.

There has been a large literature on the subject of random graph models for the World Wide Web. The examination of these models has led to better understanding of several structural features, including the distribution of indegrees and outdegrees [4, 14, 15], the size and structure of the strongly connected components [4], and the existence of small communities [9]. Unfortunately, none of the previously described models are able to predict the pronounced locality that is present in hyperlinks.

The rest of the paper is structured as follows. In the next section we will describe the data set and methodology that is used for our observations. In section 3 we discuss locality measures for hyperlinks, and evidence for the fact that this locality follows the URL hierarchy. In section 4 we examine the situation in which links are bidirectional. In section 5 we briefly summarize previous models for the web and requirements that guide the formation of such models. We then describe a simple model that incorporates the hierarchical structure and locality of the Web as well as bidirectionality.

2 Experimental Methodology

Our observations are based on examination of a large subset of the Web as it existed in 2002. Most of our measurements have been performed from a crawl of over a billion pages from more than 18.5 million hosts, performed at IBM Almaden during the past year. Due to space limitations in our experimentation platform, we have limited ourselves to the first 616 million pages from this crawl, from 12.5 million sites. For some experiments we sampled from among these URLs in smaller proportion in order to keep the computations manageable. Our goal was to use as large a data set as possible in order to provide assurance that our observations are fairly comprehensive. Even with such a large data set, observations about the World Wide Web are complicated by the fact that the data set is constantly changing, and it is impossible to gather the entire web. The characteristics of the data set are also influenced by the crawl strategy used. The algorithm used by our crawler is fairly standard, by keeping a set of hosts active at one time, and crawling in round robin fashion from this set of hosts. After a time, these sites are evicted, to be replaced by other sites. The crawl order is approximated fairly well by a breadth first search.

Approximately 40% of the URLs in our crawl contain a `?` character in them, which proves to be a crucial consideration in our study. Such URLs are often used to fetch the results of a database query, with arguments following the `?` to indicate the data that is requested. An increasing number of web sites use such URLs to retrieve standard textual content (e.g., news sites), and it can be difficult to distinguish the two by an automated process. Statements about the aggregation of information on the web can be strongly influenced by whether such URLs are included in the study, and we shall be careful to indicate cases where we do not consider them. As the web continues to grow, we expect this to become increasingly important.

3 The Web Forest

The main purpose of this paper is to understand the interaction between the hyperlink structure of the web and the tree hierarchy present in URLs (i.e., the directory structure). At the top level, we think of the web as a collection of hosts. The distribution of the number of pages and directories per host in the dataset is shown in Figure 1(a). From the fact that the log-log plot for pages per host appears more quadratic than linear, we believe that this distribution is not a simple power law, but may instead be best described by a lognormal distribution or a double Pareto distrbution(see [13]). It should be noted however, that our crawling strategy, as well as the proliferation of aliases for hosts on the web, mean that this data is only approximate.

Moving down the hierarchy, to the directory structure within hosts, one might wonder how the shapes of directory trees of web servers are distributed, and how the URLs on a web server are distributed among the directories. For this purpose, we sampled the URLs and outlinks from approximately 100,000 web sites among the 12.5 million in our data set. For each directory on the server, we computed the number of URLs that correspond to the files in that directory, the number of subdirectories, and the depth of the directory. The distribution of the number of URLs and the number of subdirectories (the fanout) is shown in Figure 1(b). The shapes of the distributions suggest that both of these are distributed as a power law distribution, albeit with different exponents. Such a distribution arises from the class of plane-oriented recursive trees [19]. These trees are created recursively by a procedure in which new nodes are attached to nodes that are chosen with probability proportional to $1 + \alpha c(x)$, where α is a constant and $c(x)$ is the number of children of the node x. Thus they possess a "rich get richer" distribution of degrees. This agrees with the observations made in [6], in which the size of subtrees at a given depth was investigated. The directory trees of sites tend to be rather shallow and broad, in spite of the fact that URLs can theoretically be thousands of characters long.

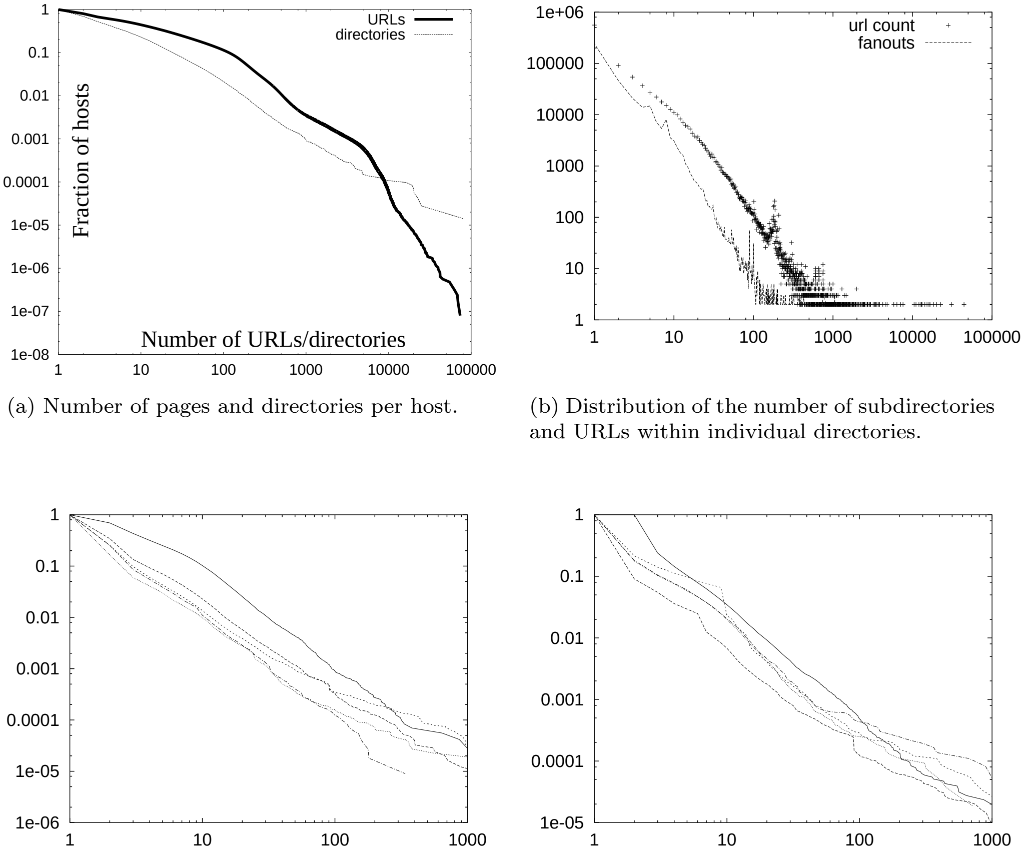
3.1 Link Locality

One feature of the web link graph that has not thus far been adequately studied the *locality* of links and its relation to the hierarchical structure of URLs. We loosely use the term "locality" to mean that links tend to be correlated to pages that are "nearby" in some measure. In practice there are various measures of locality that one might consider. Watts and Strogatz [20] examined the concept of "small world" graphs as measured by their *clustering coefficient*. This is defined as the probability that two neighbors of a page are also neighbors of each other (this version ignores the direction of hyperlinks). This measure was applied to a set of 100 million pages from the Web from 1998 by Adamic [1] to show that the World Wide Web is a "small world" graph. Another measure of clustering based on link distance in graphs was proposed by Newman [14]. He defined the *mutuality* of a graph as the ratio of the mean number of vertices two steps away from a vertex, divided by the mean number of paths of length two to those vertices. As far as we know, this measure has never been calculated for a significant portion of the web.

These measures provide evidence of a form of locality in the Web, but they do not shed much light on the process that creates the locality, and are therefore difficult to model. Davison [5] and Menczer [12] have also studied the "topical locality" of links, based on the observation that pages linked to or from a given page are usually on a similar topic. While this is clearly a form of locality that is present in the web, it does not easily lead to a model for the Web graph without a deeper understanding of the structure and distribution of topics that are present on the web.

3.2 Directory locality

We claim that much of the locality of links can be explained by a very strong correlation between the process of creating links and that of growing the hierarchy of a web site. We analyze the locality properties of links by dividing them into six distinct types:



(a) Number of pages and directories per host.

(b) Distribution of the number of subdirectories and URLs within individual directories.

(c) Distribution of number of subdirectories at directory depths 0 through 4.

(d) Distribution of URL counts at directory depths 0 through 4.

Figure 1: Distribution of URLs and subdirectories. In (c) and (d) we show the probability of the complementary cumulative distribution function (i.e., the probability that the count is greater than a given value). The nearly linear relationship for these distributions suggests the existence of a power-law distribution. The similarity for the distributions at different depths further suggests a scale-free phenomenon.

Self loops, Intra-directory links, Up and Down links (those that follow the directory hierarchy), Across links (all links within a host that are not of the other types), and External links that go outside of the site. The number of links of each type found in our corpus is shown in the second column of Table 1. This experiment clearly shows that external links make up for a relatively small fraction of the links, particularly when considering the fact that picking end points for links randomly by almost any strategy would result with almost all links being external. Note that when we limit ourselves to links for which we have crawled both ends, the fraction of external links is even smaller. This is partly because "broken" links are more common among external links, and partly because of our crawling strategy. The combination of our crawling strategy and the way authors typically limit crawling of their sites also help to explain the reduction in the numbers of Down links.

Another point one may consider when examining the distribution of links of the

Type of link	All static links	Both ends crawled	Bidirectional
Intra-directory	32.3%	41.1%	80.3%
Up	9.0%	11.2%	4.5%
Down	5.7%	3.9%	4.5%
Across directories	18.4%	18.7%	10.0%
External to host	33.6%	25.0%	0.7%
Total	5.1 billion	534893	156859

Table 1: Distribution of links by type. Shown are the distribution of links for the complete corpus, a sample among links where both source and destination pages were crawled, and a sample among bidirectional links. Self loops (which were not included in the sample) account for roughly 0.9% of the links.

various types is the influence of normalizing the distribution by the number of possible targets of the various types. For example, in a random sample of approximately 100,000 web sites, we found that approximately 83% of the URLs appear at the leaves of the directory tree. Clearly, leaves cannot have outgoing "down" links. How much does the tree structure dictate the distribution we see? To answer this question we picked a random sample of roughly 100,000 sites, and for each page, generated outlinks to other pages from the same site uniformly at random. We generated the same *number* of outlinks as the pages originally had. We compare this to the distribution of types of outlinks in general, normalized to exclude self-loops and external links, in Table 2. The data clearly shows a significantly higher number of links that follow the hierarchy (intra-directory, up and down links) in the real data, compared to what a random selection of targets will generate. This shows that the creation of links is highly correlated with the hierarchical structure of a web site.

Type of link	Crawled links	Random links
Intra-directory	48.6%	32%
Up	13.6%	6%
Down	8.6%	5%
Across directories	22.7%	57%

Table 2: Distribution of intra-host links in our test corpus and in a randomly generated graph on a sample of sites.

Another measure of locality that bears some relationship to the hierarchical structure is the measure of directory distance. We consider a distance measure between URLs known as the "tree distance". This distance is calculated by considering the directory structure implicitly exposed in a URL as a tree, and measuring the tree traversal distance between the directories (e.g., the number of directories between slashes that must be removed and appended to get from one URL to the other).

We hypothesized that links tend to span a short distance in this measure, and in order to test this we calculated the distances for all links in the data set for which both the source and destination URL do not contain a `?` character (a total of 5.1 billion links). Figure 2 shows the results of the distribution of tree distance from this data set. From this data it appears that links have a great deal of locality when distinguished by the tree distance. For across, up, and down links, the probability of a link covering a distance d appears to decrease exponentially in d. For down links this preference is pronounced, but for up links this is misleading because up links tend to be *more* distant than if the targets were selected randomly, since many up links are to the top of a site, bypassing all intermediate destinations.

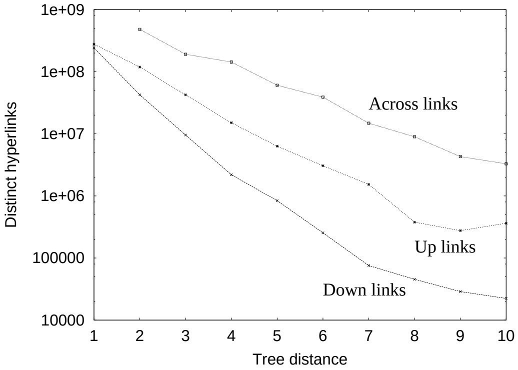


Figure 2: Distribution of tree distance for hyperlinks. As distance increases, the probability of a hyperlink decreases.

3.3 Link Compressibility

It has been observed by several authors that the link graph is highly compressible [17, 16]. In [17] they report that it takes only 6 bits on average to store the outlinks from a set of 350 million pages (6 billion links). If the links were *random* then of course this would not be possible, as an easy probabilistic argument says that we would require at least 28 bits to store a single link from each page. In fact the hierarchical locality for links that we have observed is closely related to why they are able to achieve such a small data structure for the link database. The primary method used in [17] is to sort the URLs lexicographically, and encode a link from one URL to another by the difference between their positions in the list. This delta encoding is small precisely because the URLs of source and destination often agree on a long prefix of the strings. Thus the compressibility of the link graph is closely related to the locality of links in the hierarchical structure.

4 Hyperlink Bidirectionality

In order for two web pages to share links to each other, the authors must at least know of their existence. Thus if the pages are created at different times, the page created first must either be created with a “broken” link, or else it is later modified to include a link to the page created later. In the case when pages are created by different authors, either they must cooperate to create their shared links, or else one page must be modified after creation of the other. This perhaps explains why many bidirectional links appear between page that are authored by the same person at the same time.

In order to examine the existence of bidirectional links in our corpus, we randomly sampled 1/64th of the URLs, recording the links between pairs of pages that had been crawled, and noting when there were links going in each direction between the pair of pages. The results, broken down by link type, are shown in Table 1. From this data we can draw several conclusions. First, bidirectional links are far more frequent than previous models would have predicted. Second, it is evident that the vast majority of bidirectional links occur in the same directory, and probably arise from simultaneous creation by the same author. Bidirectional links between pages on dissimilar sites are extremely rare and probably indicates a high degree of cooperation or at least recognition between the authors of the two pages.

5 Hierarchy in Models for the Web

There have been many papers presenting variations on evolving models for the web graph, with somewhat different goals. Albert et al. [2] presented an early evolutionary model of small world networks in which vertices and edges are appended over time. This model sought to explain the observed distribution of degrees of in and out links using *preferential attachment*. In addition, their models provided some explanation for the "small world" or "scale-free" features of the web. More recent models (see [15, 11]) have mixed preferential attachment with a uniform attachment process in order to better explain the observed statistics for the low-connectivity region, and to parameterize the power law exponent. Kumar et. al. [9] presented a class of *evolving copying* models in order to explain the existence of small thematic communities in the web.

A hierarchical model of the web was previously suggested in by Laura et. al. in [10]. In their model, every page that enters the graph is assigned with a constant number of abstract "regions" it belongs to, and is allowed to link only to vertices in the same region. This forces a degree of locality among the vertices of the graph, though the definition of regions is unspecified, and the model artificially controls connections between these regions. In our model, we use the explicit hierarchy in the structure of URLs to establish the regions, which reflects a social division by organization.

We believe that the approach to modeling the web should be based on a model of the social process of authorship, and the nature of social relationships within increasingly larger groups. These social processes are not always present, and can have somewhat vague definitions, but they are very instructive for understanding a more formal model. Consider the the social process by which a web site of a large company or university is built. At the lowest level we might start with an individual who authors a single page consisting of a personal page, a news release, privacy statement, or other small bit of information. This then fits into a larger group of pages that might be authored by the same author (often on the same or closely related topics). The author of these pages may be a member of a small group, department, or family, in which there are other authors who contribute material.

Continuing up the chain, a department or group might be part of a division, college, or physical location within a larger organization consisting of a university, company, or ISP. This larger organization can be groups with other organizations of the same type, such as other universities under the edu domain, or other companies, or other domains in the same geographic region. As we move up the hierarchy of social structure, there is generally less social coordination between authors of pages, and the probability of links between documents decreases as we move up the chain.

The hierarchical structure of the social groups of authors of web information follows very closely the development of other social phenomenon as described by Simon [18]. In addition to this social hierarchy, web information has a topical hierarchy associated with it that is often recognizable from the URL hierarchy. Both of these provide a good basis for a model of relationships between different information items on the World Wide Web.

5.1 Requirements for a Model

Mathematical modeling and computational science has a rich history of advancing the state of numerous scientific disciplines, including material science, chemistry, economics, biology, astronomy, physics, and computational fluid dynamics. The practice of computational science consists of a cycle of analysis, model construction, computational simulation, prediction, and validation. Past experience in other fields suggests that once we capture the *essential elements* of a problem in a model, the model can lead to a deeper understanding of the complexity of a problem and a useful complement for experimental work in the field. In seeking to model the Web, we consider the following axioms to be important, though the list is not exhaustive.

Evolutionary The model should allow for an *evolution* of the graph. The simplest form of evolution recognizes that the web is growing over time, and that pages are constantly being created. More complicated models would reflect the fact that many pages are modified over time, or may be removed (creating broken links).

Social realism The model should reflect the social and authorship processes that influence the World Wide Web. The evolving copying models of [9] provides an example of this property.

Indegree/outdegree The indegree distributions should be parameterized to achieve a power-law distribution in the tail, and preferably also conform to the discrepancy in the tail observed in [15]. Outdegree distributions should also conform to a heavy tail distribution, though the distribution is different than indegrees and it may be truncated.

Hierarchical The model should reflect the hierarchical organization that is evident in structure of URLs on the Web. In particular, the fanout of the trees that result from creating directories should be distributed as a power law, and the distribution of URLs within directories should be distributed as a power law.

Locality The model should exhibit a degree of locality in the link structure, in conformity with that seen in the hierarchical nature of the web. It should also exhibit the "small world" nature of the web.

Communities The existence of small communities of thematically related pages should not be precluded by the model [9].

Bidirectionality The probability of a link being bidirectional should be strongly correlated to the locality of links.

Simplicity To the extent possible, the model should be simple enough to analyze or at least simulate, and yet capture the features described above.

5.2 A Hierarchical Model of the Web

We propose a model in which the web grows in two different (but related) ways. First, new hostnames get added to the web, and second, new URLs get added to existing hosts. We treat these processes separately, by evolving two graph structures for the hierarchical directory structure and the hyperlinks. The model evolves over time, and at discrete time steps we perform one of two steps: either we add a site to the web (with a single URL at the top, and potentially some outlinks to other sites), or else we add a URL to a site along with some outlinks to that URL. Sites themselves grow in a hierarchical fashion, with a site starting as a single URL, and growing into a tree. Our model therefore treats the web as a collection of trees (a forest), with hyperlinks that are separate from the tree structure (but, as we have demonstrated, not independent of it).

At each step in time a new URL is added to the Web. With probability ϵ, this URL is added as a new tree (i.e., a new domain), containing a single URL. With probability $1 - \epsilon$ we pick an existing directory to add a URL to. The probability that we pick a particular directory D is proportional to $s_D + f_D$ where s_D is the number of URLs in D, and f_D is the number of sub-directories of D. After picking the directory, we use the power-law distribution with parameter δ for the number of URLs in a directory to decide whether to add a new URL to the directory, or create a sub-directory with a single new URL in it. If the directory that the new URL is created in has s_D URLs after adding the new one, then the probability of not splitting the directory is proportional to $s_D^{-\delta}$. In order to connect this page to the rest of the site, we create a link to this URL by randomly selecting another URL on the site and creating a link from that page to the newly added page.

We now have to say how to create links from this URL. We hypothesize the existence of five parameters that are preserved as the graph grows, namely the probabilities of a link being internal, up, down, across, or external. For each type of link t we have a fixed probability p_t that remains constant as the graph grows, and $\sum_t p_t = 1$. For each type of link we also have a fixed probability b_t that the link will be bidirectional. In assigning links to a page, we first decide the number of links in accordance with a hypothesized power law distribution on the outdegrees from pages. For each created link we assign it a type t with probability p_t. We pick the target for the link from among the eligible URLs with a probability that is proportional to 1 plus its existing indegree. If there are no eligible URLs to create a link to, then we simply omit the link. If we create a link, then we create a backlink from that link with probability b_t.

There are endless variations on this model, including the incorporation of copying, a preference for linking to URLs that are a short distance away, preferences for linking to URLs that are at a given level of the directory tree, etc. The purpose of our exposition here is to propose a simple model that satisfies the hierarchical requirement mentioned previously. We plan to expand upon this model in future work, as well as analyze the growth processes.

6 Conclusions

In this work we concentrated on the properties of the web graph that are the result of the interaction between two evolutionary processes that shape the web: the growth of hierarchical structures as reflected in URLs, and the creation of hyperlinks on the web. We have shown that the hyperlink structure is highly correlated with the hierarchical structure underlying URLs. This correlation is particularly strong for bidirectional links. We therefore conclude that an evolutionary model of the web cannot accuractly model locality and bidirectionality properties of hyperlinks without acounting for the underlying growth process of the hierarchical structure.

In this work we have proposed a model that incorporates an evolutionary process that acts on both the hierarchical structure and the hyperlink graph. The model is further motivated by how web sites evolve, from the general to the specific. Ours is certainly not the final word in models of the web, and it is natural to expect that more complicated models will arise in the future that incorporate other features. Natural candidates for examination include topical locality [5] and similarity [7], author relationships, and institutional missions. It is our hope that the study of the features of the web that we examine, and the model we propose to explain them, will lead to a better understanding of the web and more effective algorithms for information retrieval tasks.

References

[1] L. A. Adamic. The small world web. In *Proceedings of ECDL '99*, volume 1696 of *Lecture Notes in Computer Science*, pages 443–454, 1999.

[2] R. Albert, H. Jeong, and A.-L. Barabási. The diameter of the world wide web. *Nature*, 401:130–131, 2000.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th World Wide Web Conference*, pages 107–117, 1998.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. In *Proc. 9th WWW*, pages 309–320, 2000.

[5] B. Davison. Topical locality in the web. In *Proceedings of the 23rd Annual International Conference on Information Retrieval*, pages 272–279, Athens, 2000.

[6] S. Dill, R. Kumar, K. S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205–223, 2002.

[7] P. Ganesan and H. G.-M. J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1):64–93, January 2003.

[8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46(5):604–632, 1999.

[9] R. Kumar, P. Raghavan, S. Rajagopalan, and D. Sivakumar. Stochastic models for the Web graph. In *Proc. of the 41st IEEE Symposium on Foundations of Comp. Sci.*, pages 57–65, 2000.

[10] L. Laura, S. Leonardi, G. Caldarelli, and P. D. L. Rios. A multi-layer model for the web graph. In *2nd International Workshop on Web Dynamics*, Honolulu, 2002. Also presented in 33rd Annual Conference of the Operational Research Society of Italy, September, 2002.

[11] M. Levene, T. Fenner, G. Loizou, and R. Wheeldon. A stochastic model for the evolution of the web. *Computer Networks*, 39:277–287, 2002.

[12] F. Menczer. Growing and navigating the small world web by local content. *Proc. Natl. Acad. Sci. USA*, 99(22):14014–14019, 2002.

[13] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1, 2003. to appear.

[14] M. E. J. Newman. Random graphs as models of networks. http://www.santafe.edu/sfi/publications/wpabstract/200202005.

[15] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *PNAS*, pages 5207–5211, 2002.

[16] S. Raghavan and H. Garcia-Molina. Representing web graphs. In *IEEE International Conference on Data Engineering (ICDE03)*, 2003.

[17] K. H. Randall, R. Stata, R. G. Wickremesinghe, and J. L. Wiener. The link database: Fast access to graphs of the Web. In *Proceedings of the 2002 Data Compression Conference (DCC)*, pages 122–131, 2002.

[18] H. A. Simon. *The Sciences of the Artifical.* MIT Press, Cambridge, MA, 3rd edition, 1981.

[19] R. T. Smythe and H. M. Mahmoud. A survey of recursive trees. *Theoretical Probability and Mathematical Statistics*, 51:1–27, 1995. Translation from *Theorya Imovirnosty ta Matemika Statystika*, volume 51, pp. 1–29, 1994.

[20] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, 393:440–442, 1998.