Self-Similarity In the Web

STEPHEN DILL, RAVI KUMAR, KEVIN S. MCCURLEY, SRIDHAR RAJAGOPALAN, D. SIVAKUMAR, and ANDREW TOMKINS
IBM Almaden Research Center, San Jose

Algorithmic tools for searching and mining the Web are becoming increasingly sophisticated and vital. In this context, algorithms that use and exploit structural information about the Web perform better than generic methods in both efficiency and reliability.

We present an extensive characterization of the graph structure of the Web, with a view to enabling high-performance applications that make use of this structure. In particular, we show that the Web emerges as the outcome of a number of essentially independent stochastic processes that evolve at various scales. A striking consequence of this scale invariance is that the structure of the Web is "fractal"—cohesive subregions display the same characteristics as the Web at large. An understanding of this underlying fractal nature is therefore applicable to designing data services across multiple domains and scales.

We describe potential applications of this line of research to optimized algorithm design for Web-scale data analysis.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering*

General Terms: Experimentation, Measurement, Theory, Verification

Additional Key Words and Phrases: Fractal, graph structure, online information services, self-similarity, Web-based services, World-Wide-Web

1. INTRODUCTION

As the the size of the Web grows exponentially, data services on the Web are becoming increasingly complex and challenging tasks. These include both basic services such as searching and finding related pages, and advanced applications such as Web-scale data mining, community extraction, constructions of indices, taxonomies, and vertical portals. Applications are beginning to emerge that are required to operate at various points on the "petabyte curve"—billions of Web pages that each have megabytes of data, tens of millions of users in a peer-to-peer setting each with several gigabytes of data, etc. The upshot of the rate and diversity of this growth is that data service applications for collections of

Authors' address: IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120; email: {dill,ravi,mccurley,sridhar,siva,tomkins}@almaden.ibm.com.

hyperlinked documents need to be efficient and effective at several scales of operation. As we will show, a form of "scale invariance" exists on the Web that allows simplification of this multiscale data service design problem.

The first natural approach to the wide range of analysis problems emerging in this new domain is to develop a general query language to the Web. There have been a number of proposals along these lines [Mendelzon et al. 1997; Abiteboul et al. 1997; Spertus 1997]. Further, various advanced mining operations have been developed in this model using a Web-specific query language like those described above, or a traditional database encapsulating some domain knowledge into table layout and careful construction of SQL programs [Chakrabarti et al. 1999a; Spertus and Stein 1998; Arocena et al. 1997].

However, these applications are particularly successful precisely when they take advantage of the special structure of the document collections and the hyperlink references among them. An early example of this phenomenon in the marketplace is the paradigm shift witnessed in search applications—ranking schemes for Web pages were vastly improved when link-based analysis was added to more traditional text-based schemes [Kleinberg 2000; Brin and Page 1998].

The success of these specialized approaches naturally led researchers to seek a finer understanding of the hyperlinked structure of the Web. Broadly, there are two (very related) lines of research that have emerged. The first one is more theoretical, and is concerned with proposing stochastic models that explain the hyperlink structure of the Web [Kumar et al. 2000; Barabasi and Albert 1999; Aiello et al. 2000]. The second line of research [Broder et al. 2000; Barabasi and Albert 1999; Adamic and Huberman 1999; Kumar et al. 1999a] is more empirical; new experiments are conducted that either validate or refine existing models.

There are several driving applications that motivate (and are motivated by) a better understanding of the neighborhood structure on the Web. In particular, the "second generation" of data service applications on the Web—including advanced search applications [Chakrabarti et al. 1998a; Chakrabarti et al. 1998b; Bharat and Henzinger 1998], browsing and information foraging [Botafogo and Shneiderman 1991; Carriere and Kazman 1997; Chakrabarti et al. 1999b; Pirolli et al. 1996; Pitkow and Pirolli 1997], community extraction [Kumar et al. 1999a], taxonomy construction [Kumar et al. 1999b, 2001]—have all taken tremendous advantage of knowledge about the hyperlink structure of the Web. As just one example, let us mention the community extraction algorithm of [Kumar et al. 1999a]. In this algorithm, a characterization of degree sequences within Web-page neighborhoods allowed the development and analysis of efficient pruning algorithms for a subgraph enumeration problem that is in general intractable.

Even more recently, new algorithms have been developed to benefit from structural information about the Web. Arasu et al. [2001] have shown how to take advantage of the macroscopic "bow-tie" structure of the Web [Broder et al. 2000] to design an efficient algorithmic partitioning method for certain eigenvector computations; these are the key to the successful search algorithms of [Brin and Page 1998; Kleinberg 2000], and to popular database

indexing methods such as latent semantic indexing [Deerwester et al. 1990; Papadimitriou et al. 2000]. Adler and Mitzenmacher [2001] have shown how the random graph characterizations of the Web given in Kumar et al. [2000] can be used to compress the Web graph.

1.1 Our Results

In this article, we present a much more refined characterization of the structure of the Web. Specifically, we present evidence that the Web emerges as the outcome of a number of essentially independent stochastic processes that evolve at various scales, all roughly following the model of Kumar et al. [2000]. A striking consequence is that the Web exhibits *self-similarity,* i.e., each thematically unified region displays the same characteristics as the Web at large. In other words, the Web is a "fractal." This implies the following:

To design efficient algorithms for data services at various scales on the Web (vertical portals pertaining to a theme, corporate intranets, etc.), it is sufficient (and perhaps necessary) to understand the structure that emerges from one fairly simple stochastic process.

We believe that this is a significant step in Web algorithmics. For example, it shows that the sophisticated algorithms of Adler and Mitzenmacher [2001] and Arasu et al. [2001] are only the beginning, and the prospects are, in fact, much wider. We fully expect future data applications on the Web to leverage this understanding.

Our characterization is based on two findings we report in this article. Our first is an experimental result. We show that self-similarity holds for many different parameters, and also for many different approaches to defining varying scales of analysis. Our second finding is an interpretation of the experimental data. We show that, at various different scales, cohesive collections of Web pages (for instances, pages on a site or pages about a topic) mirror the structure of the Web at large. For example, consider the collection of Web pages that have at least one geographical reference to a location in the western half of the United States. We show that this cohesive collection of pages resembles the Web at large in terms of various graph-theoretic characteristics and parameters.

Furthermore, if the Web is decomposed into these cohesive collections, for a wide range of definitions of "cohesive," the resulting collections are tightly and robustly connected via a *navigational backbone* that affords strong connectivity between the collections. This backbone not only ties together the collections of pages, but also ties together the many different and overlapping *decompositions* into cohesive collections, suggesting that committing to a single taxonomic breakdown of the Web is neither necessary nor desirable. We now describe these two findings in more detail.

First, self-similarity in the Web is pervasive and robust—it applies to a number of essentially independent measurements and regardless of the particular method used to extract a slice of the Web. Second, we present a graph-theoretic interpretation of the first set of observations, which leads to a natural hierarchical characterization of the structure of the Web interpreted as a graph. In our characterization, collections of Web pages that share a common attribute

(for instance, all the pages on a site or all the pages about a particular topic) are structurally similar to the whole Web. Furthermore, there is a *navigational backbone* to the Web that provides tight and robust connections between these focused collections of pages.

(1) ***Experimental findings.*** Our first finding, that self-similarity in the Web is pervasive and appears in many unrelated contexts, is an experimental result. We explore a number of graph-theoretic and syntactic parameters. The set of parameters we consider is the following: indegree and outdegree distributions; strongly- and weakly-connected component sizes; bowtie structure and community structure on the Web graph; and population statistics for trees representing the URL namespace. We define these parameters formally below. We also consider a number of methods for decomposing the Web into interesting subgraphs. The set of subgraphs we consider is the following: a large Internet crawl; various subgraphs consisting of about 10% of the sites in the original crawl; 100 Websites from the crawl, each containing at least 10,000 pages; ten graphs, each consisting of every page containing a set of keywords (in which the ten keyword sets represent five broad topics and five subtopics of the broad topics); a set of pages containing geographical references (e.g., phone numbers, zip codes, city names, etc.) to locations in the western United States; a graph representing the connectivity of Web sites (rather than Web pages); and a crawl of the IBM intranet. More details about the crawl can be found in Section 3.3.

We then consider each of the parameters described above, first for the entire collection, and then for each decomposition of the Web into subcollections. Self-similarity is manifest in the resulting measurements in two flavors. First, when we fix a collection or subcollection and focus on the distribution of any parameter (such as the number of hyperlinks, number of connected components, etc.), we observe a Zipfian self-similarity within the pageset.[1] Namely, for any parameter x with distribution X, there is a constant c such that for all $t > 0$ and $a \geq 1$, $X(at) = a^c X(t)$. In many cases, even the constant c remains the same across different subcollection of pages—for example, our study suggests that for any cohesive collection of Web pages, the fraction of Web pages in this collection that have k hyper-inlinks is proportional to $k^{-2.1}$. Second, the phenomena (whether distributional or structural) that are manifest within a subcollection are also observed (with essentially the same constants) in the entire collection, and more generally, in all subcollections at all scales—from local Websites to the Web as a whole.

(2) ***Interpretations.*** Our second finding is an interpretation of the experimental data. As mentioned above, the subcollections we study are created to be cohesive clusters, rather than simply random sets of Web pages. We refer to them as *thematically unified clusters*, or simply TUCs. Each TUC has structure similar to the Web as a whole. In particular, it has a Zipfian distribution over the parameters we study, strong navigability properties, and

[1]For more about the connection between Zipfian distributions and self-similarity, see Section 2.2 and [32].

significant community and bowtie structure (in a sense to be made explicit below).

Furthermore, we observe unexpectedly that the central regions of different TUCs are tightly and robustly connected together. These tight and robust intercluster linking patterns provide a *navigational backbone* for the Web. By analogy, consider the problem of navigating from one physical address to another. A user might take a cab to the airport, take a flight to the appropriate destination city, and take a cab to the destination address. Analogously, navigation between TUCs is accomplished by traveling to the central core of a TUC, following the navigational backbone to the central core of the destination TUC, and finally navigating within the destination TUC to the correct page. We show that the self-similarity of the Web graph, and its local and global structure, are alternate and equivalent ways of viewing this phenomenon.

1.2 Related Prior Work

Zipf-Pareto-Yule and Power laws. Distributions with an inverse polynomial tail have been observed in a number of contexts. The earliest observations are due to Pareto [1897] in the context of economic models. Subsequently, these statistical behaviors have been observed in the context of literary vocabulary [Yule 1944], sociological models [Zipf 1949], and even oligonucleotide sequences [Martindale and Konopka 1996], among others. Our focus is on the closely related power law distributions, defined on the positive integers, with the probability of the value i being proportional to i^{-k} for a small positive number k. Perhaps the first rigorous effort to define and analyze a model for power law distributions is due to Simon [1955].

Recent work [Barabasi and Albert 1999; Broder et al. 2000; Kumar et al. 2001] suggests that both the in- and the outdegrees of nodes on the Web graph have power laws. The difference in scope in these three experiments is noteworthy. The first [Kumar et al. 2001] examines a Web crawl from 1997 due to Alexa, Inc., with a total of over 40 million nodes. The second [Barabasi and Albert 1999] examines Web pages from the University of Notre Dame domain `*.nd.edu` as well as a portion of the Web reachable from three other URLs. The third [Broder et al. 2000] examines a Web crawl from 1999 due to Altavista, Inc., with a total of over 270 million nodes. This collection of findings already leads us to suspect the fractal-like structure of the Web.

Graph-theoretic methods. Much recent work has addressed the Web as a graph and applied algorithmic methods from graph theory in addressing a slew of search, retrieval, and mining problems on the Web. The efficacy of these methods was already evident even in early local expansion techniques [Botafogo and Shneiderman 1991]. Since then, increasingly sophisticated techniques have been used; the incorporation of graph-theoretical methods with both classical and new methods that examine both context and content, and richer browsing paradigms have enhanced and validated the study and use of such methods. Following Botafogo and Shneiderman [1991], the view that

connected and strongly-connected components represent meaningful entities has become widely accepted.

Power laws and browsing behavior. The power law phenomenon is not restricted to the Web graph. For instance, Faloutsos et al. [1999] report very similar observations about the physical topology of the Internet. Moreover, the power law characterizes not only the structure and organization of information and resources on the Web, but also the way people use the Web. Two lines of work are of particular interest to us here. (1) Web page access statistics, which can be easily obtained from server logs (but for caching effects) [Adamic and Huberman 2000; Glassman 1994; Huberman et al. 1998]. (2) User behavior, as measured by the number of times users at a single site access particular pages also enjoy power laws, as verified by instrumenting and inspecting logs from Web caches, proxies, and clients [Barford et al. 1999; http://linkage.rockfeller.edu/wli/zipf/.].

There is no direct evidence that browsing behavior and linkage statistics on the Web graph are related in any fundamental way. However, making the assumption that linkage statistics directly determine the statistics of browsing has several interesting consequences. The Google search algorithm, for instance, is an example of this. Indeed, the view of PageRank put forth in Brin and Page [1998] is that it puts a probability value on how easy (or difficult) it is to find particular pages by a browsing-like activity. Moreover, it is generally true (for instance, in the case of random graphs) that this probability value is closely related to the indegree of the page. In addition, there is recent theoretical evidence [Kumar et al. 2000; Simon 1955] suggesting that this relationship is deeper. In particular, if one assumes that the ease of finding a page is proportional to its graph-theoretic indegree, and that otherwise the process of evolution of the Web as a graph is a random one, then power law distributions are a direct consequence. The resulting models, known as *copying* models for generating random graphs seem to correctly predict several other properties of the Web graph as well.

2. PRELIMINARIES

In this section we formalize our view of the Web as a graph; here we ignore the text and other content in pages, focusing instead on the links between pages. In the terminology of graph theory [Harary 1975], we refer to pages as *nodes*, and to links as *arcs*. In this framework, the Web is a large graph containing over a billion nodes, and a few billion arcs.

2.1 Graphs and Terminology

A *directed graph* consists of a set of *nodes*, denoted V and a set of *arcs*, denoted E. Each arc is an ordered pair of nodes (u, v) representing a directed connection from u to v. The *outdegree* of a node u is the number of distinct arcs $(u, v_1), \ldots, (u, v_k)$ (i.e., the number of links from u), and the *indegree* is the number of distinct arcs $(v_1, u), \ldots, (v_k, u)$ (i.e., the number of links to u). A path from node u to node v is a sequence of arcs $(u, u_1), (u_1, u_2), \ldots, (u_k, v)$.

One can follow such a sequence of arcs to "walk" through the graph from u to v. Note that a path from u to v does not imply a path from v to u. The *distance* from u to v is one more than the smallest k for which such a path exists. If no path exists, the distance from u to v is defined to be infinity. If (u, v) is an arc, then the distance from u to v is 1. Given a graph (V, E) and a subset V' of the node set v, the *node-induced subgraph* (V', E') of (V, E) is defined by taking E' to be $\{(u, v) \in E \mid u, v \in V'\}$, i.e., the node-induced subgraph corresponding to some subset V' of the nodes contains only arcs that lie entirely within V'.

Given a directed graph, a *strongly connected component* of this graph is a set of nodes such that for any pair of nodes u and v in the set there is a path from u to v. In general, a directed graph may have one or many strong components. Any graph can be partitioned into a disjoint union of strong components. Given two strongly connected components, C_1 and C_2, either there is a path from C_1 to C_2 or a path from C_2 to C_1 or neither, but not both. Let us denote the largest strongly component by *SCC*. Then, all other components can be classified with respect to the SCC in terms of whether they can reach, be reached from, or are independent of, the SCC. Following Broder et al. [2000], we denote these components *IN*, *OUT*, and *OTHER* respectively. The SCC, flanked by the IN and OUT, figuratively forms a "bowtie."

A *weakly connected component* of a graph is a set of nodes such that for any pair of nodes u and v in the set, there is a path from u to v if we disregard the directions of the arcs. Similar to strongly connected components, the graph can be partitioned into a disjoint union of weakly connected components. We denote the largest weakly connected component by (*WCC*).

2.2 Zipf Distributions and Power Laws

The power law distribution with parameter $a > 1$ is a distribution over the positive integers. Let X be a power law distributed random variable with parameter a. Then, the probability that $X = i$ is proportional to i^{-a}. The Zipf distribution is an interesting variant on the power law. The Zipf distribution is a defined over any categorical-valued attribute (for instance, words of the English language). In the Zipf distribution, the probability of the i-th most likely attribute value is proportional to i^{-a}. Thus, the main distinction between these is in the nature of the domain from which the random variable takes its values. A classic general technique for computing the parameter a characterizing the power law is due to Hill [1975]. We use Hill's estimator as the quantitative measure of self-similarity.

While a variety of socio-economic phenomena have been observed to obey Zipf's law, there is only a handful of stochastic models for these phenomena of which satisfying Zipf's law is a consequence. Simon [1955] was perhaps the first to propose a class of stochastic processes whose distribution functions follow the Zipf law (http://linkage.rockfeller.edu/wli/zipf/). Recently, new models have been proposed for modeling the evolution of the Web graph [Kumar et al. 2000]. These models predict that several interesting parameters of the Web graph obey the Zipf law.

3. EXPERIMENTAL SETUP

3.1 Random Subsets and TUCs

Since the average degree of the Web graph is small, one should expect subgraphs induced by (even fairly large) random subsets of the nodes to be almost empty. Consider for instance a random sample of 1 million Web pages (say out of a possible 1 billion). Consider now an arbitrary arc, say (a, b). The probability that both endpoints of the arc are chosen in the random sample is about 1 in a million (1/1000 * 1/1000). Thus, the total expected number of arcs in the induced subgraph of these million nodes is about 8000, assuming an average degree of 8 for the Web as a whole. Thus, it would be unreasonable to expect random subgraphs of the Web to contain any graph-theoretic structure. However, if the subgraphs chosen are not random, the situation could be (and is) different. In order to highlight this dichotomy, we introduce the notion of a *thematically unified cluster* (*TUC*). A TUC is a cluster of Webpages that share a common trait. In all instances we consider, these thematically unified clusters share a fairly syntactic trait. However, we do not wish to restrict our definition only to such instances. For instance, one could consider linkage-based concepts [Pirolli et al. 1996; Spertus 1997] as well. We now detail several instances of TUCs.

(1) ***By content.*** The premise that Web content on any particular topic is also "local" in a graph-theoretic context has motivated some interesting earlier work [Kleinberg 2000; Kumar et al. 2001]. Thus, one should expect Web pages that share subject matter to be more densely linked than random subsets of the Web. If so, these graphs should display interesting morphological structure. Moreover, it is reasonable to expect this structure to represent interesting ways of further segmenting the topic.

The most naive method for judging content correlation is to simply look at a collection of Webpages that share a small set of common keywords. To this end, we have generated 10 slices of the Web, denoted henceforth as KEYWORD1, ..., KEYWORD10. To determine whether a page belongs to a keyword set, we simply look for at least one occurrence of the keyword in the body of the document after simple preprocessing (removing tags, javascript, transform to lowercase, etc.). The particular keyword sets we consider are shown in Tables III and IV. The terms in the first table correspond to mesoscopic subsets and the corresponding terms in the second table are microscopic subsets of the earlier ones.

(2) ***By location.*** Websites and intranets are logically consistent ways of partitioning the Web, hence they are obvious candidates for TUCs. We look at intranets and particular Websites to see what structures are represented at this level. We are interested in what features, if any, distinguish these two cases from each other and, indeed, from the Web at large. Our observations here would help determine what special processing, if any, would be relevant in the context of an intranet. To this end, we have created TUCs consisting of 100 Websites (of the form `www.*.*`) denoted SUBDOMAIN1, ..., SUBDOMAIN100, each containing at least 10 K pages and the IBM intranet, denoted INTRANET.

(3) ***By geographic location.*** Geography is becoming increasingly evident in the Web, with the growth in the number of local and small businesses represented on the Web (restaurants, shows, housing information, and other local services), as well as local information Websites such as `sidewalk.com`. We expect the recurrence of similar information structures at this level. We hope to understand more detail about overlaying geospatial information on top of the Web. We have created a subset of the Web based on geographic cues, denoted GEO henceforth. The subset contains pages that have geographical references (addresses, telephone numbers, and ZIP codes) to locations in the western United States. This was constructed through the use of databases for latitude–longitude information for telephone number area codes, prefixes, and postal zipcodes. Any page that contained a zipcode or telephone number was included if the reference was within a region bounded by Denver (Colorado) on the east and Nilolski (Alaska) on the west, Vancouver (British Columbia) on the north, and Brownsville (Texas) on the south.

To complete our study, we also define some additional graphs derived from the Web. Strictly speaking, these are not TUCs. However, they can be derived from the Web in a fairly straightforward manner. As it turns out, some of our most interesting observations about the Web relates to the interplay between structure at the level of the TUCs and structure at the following levels. We define them now:

(4) ***Random collections of Websites.*** We look at all the nodes that belong in a random collection of Websites. We do this in order to understand the fine-grained structure of the SCC, which is the navigational backbone of the Web. Unlike random subgraphs of the Web, random collections of Websites exhibit interesting behaviors. First, the local arcs within a Website ensure that there is fairly tight connectivity within each Website. This allows the small number of additional intersite arcs to be far more useful than would be the case in a random subgraph. We have generated seven such disjoint subsets. We denote these STREAM1, ..., STREAM7.

(5) ***Hostgraph.*** The hostgraph contains a single node corresponding to each Website (for instance `www.ibm.com` is represented by a single node), and has an arc between two nodes, whenever there is a page in the first Website that points to a page in the second. The hostgraph is not a subgraph of the Web graph, but it can be derived from it in a fairly straightforward manner, and more importantly, is relevant to understanding the structure of linkage at levels higher than that of a Web page. In the following discussion, this graph is denoted HOSTGRAPH.

3.2 Parameters

We study the following parameters:

(1) ***Indegree distributions.*** Recall that the indegree of a node is the number of arcs whose destination is that node. We consider the distribution of indegree over all nodes in a particular graph, and consider properties of that

distribution. A sequence of papers [Adamic and Huberman 1999; Barabasi and Albert 1999; Broder et al. 2000; Kumar et al. 1999a] has provided convincing evidence that indegree distributions follow the power law, and that the parameter a (called *indegree exponent*) is reliably around 2.1 (with little variation). We study the indegree distributions for the TUCs and the random collections.

(2) ***Outdegree distributions.*** Outdegree distributions seem to not follow the power law at small values. However, larger values do seem to follow such a distribution, resulting in a "drooping head" of the log-log plot, as observed in earlier work. A good characterization of outdegrees for the Web graph has not yet been offered, especially one that would satisfactorily explain the drooping head.

(3) ***Connected component sizes.*** (cf., Section 2) We consider the size of the largest strongly-connected component, the second-largest, third-largest, and so forth, as a distribution for each graph of interest. We consider similar statistics for the sizes of weakly-connected components. Specifically, we show that they obey power laws at all scales, and study the exponents of the power law (called *SCC/WCC exponent*). We also report the ratio of the size of the largest strongly-connected component to the size of the largest weakly-connected component. For the significance of these parameters, we refer the reader to Broder et al. [2000], and note that the location of a Web page in the connected component decomposition crucially determines the reachability of this page (often related to its popularity).

(4) ***Bipartite cores.*** Bipartite cores are graph-theoretic signatures of community structure on the Web. A $K_{i,j}$ *bipartite core* is a set of $i + j$ pages such that each of i pages contains a hyperlink to all of the remaining j pages. We pick representative values of i and j, and focus on $K_{5,7}$'s, which are sets of 5 "fan" nodes, each of which points to the same set of 7 "center" nodes. Since computing the exact number of $K_{5,7}$'s is a complex subgraph enumeration problem that is intractable using known techniques, we instead estimate the number of node-disjoint $K_{5,7}$'s for each graph of interest. To perform this estimation, we use the techniques of Kumar et al. [1999a, 1999b]. The number of communities (cores) is an estimate of community structure with the TUC. The $K_{5,7}$ *factor* of a TUC is the ratio of the number of the nodes in the TUC to the number of nodes that participate in $K_{5,7}$'s in the TUC. According to this definition, it is easy to see that the higher the factor, the less one can view the TUC as a single well-defined community.

(5) ***URL compressibility and namespace utilization.*** The URL namespace can be viewed as a tree, with the root node represented by the null string. Each node of the tree corresponds to a URL prefix (say, `www.foo.com`) with all URLs that share that prefix, (e.g., `www.foo.com/bar` and `www.foo.com/rab`) being in the subtree rooted at that node. For each subgraph and each value d of the depth, we study the following distribution: for each s, the number of depth-d nodes whose subtrees have s nodes. We will see that these follow the power law. Following conventional source coding theory, it follows that this skew in the population distributions of the URL namespace can be used to design improved

Table I. Results for STREAM1 through STREAM7

Nodes $\times 10^6$	Arcs $\times 10^6$	Expansion factor	Indeg. exp.	Outdeg. exp.	SCC exp.	WCC exp.	WCC $\times 10^6$	SCC/ WCC	IN/ WCC	OUT/ WCC	$K_{5,7}$ factor
6.55	46.8	2.06	−2.07	−2.12	−2.16	−2.32	4.69	0.24	0.23	0.23	47.2
6.47	45.7	2.06	−2.08	−2.24	−2.14	−2.28	4.60	0.23	0.19	0.24	50.1
6.38	48.1	2.05	−2.06	−2.15	−2.15	−2.24	4.47	0.24	0.20	0.23	49.5
6.84	50.0	2.04	−2.12	−2.30	−2.14	−2.27	4.86	0.23	0.21	0.23	43.5
6.83	48.2	2.06	−2.08	−2.27	−2.11	−2.29	4.90	0.24	0.20	0.23	45.4
6.77	49.3	2.01	−2.10	−2.32	−2.11	−2.25	4.78	0.23	0.20	0.24	45.3
6.23	43.5	2.03	−2.13	−2.19	−2.15	−2.27	4.31	0.22	0.19	0.23	46.9

compression algorithms for URLs. The details of this analysis are beyond the scope of the present article.

3.3 Experimental Infrastructure

We performed these experiments on a small cluster of Linux machines with about 1TB of disk space. We created a number of data sets from two original sets of pages. The first set consists of about 500 K pages from the IBM intranet. We treat this data as a single entity, mainly for purposes of comparison with the external Web. The second set consists of 60 M pages from the Web at large, crawled in Oct. 2000. These 60 M pages represent the pages that were actually crawled and amount to approximately 750 GB of content. The crawl was seeded with a set of external IBM sites and commercial sites, and the crawling algorithm obeyed all politeness rules, crawling no site more often than once per second. Therefore, while we had collected 750 GB of content (crawling about 1.3 M sites) no more than 12 K pages had been crawled from any one site. More details of the crawling algorithm can be found in Edwards et al. [2001].

4. RESULTS AND INTERPRETATION

Our results are shown in the following tables and figures. Though we have an enormous amount of data, we try to present as little as possible, while conveying the main thoughts. All the graphs here refer to node-induced subgraphs, and the arcs refer to the arcs in the induced subgraph. Our tables show the parameters in terms of the graphs, while our figures show the consistency of the parameters across different graphs, indicating a fractal nature.

Table I shows all the parameters for the STREAM1 through STREAM7. The additional parameter, *expansion factor*, refers to the fraction of hyperlinks that point to nodes in the same collection to the total number of hyperlinks. As we can see, the numbers are quite consistent with earlier work. For instance, the indegree exponent is −2.1, the SCC exponent is around −2.15, and the WCC exponent is around −2.3. As we can see, the ratios of IN, OUT, SCC with respect to WCC are also consistent with earlier work.

Table II shows the results for the three special graphs: INTRANET, HOSTGRAPH, and GEO. The expansion factor for the INTRANET is 2.158, while the indegree exponent is very different from that of other graphs. The WCC exponent for HOSTGRAPH is not meaningful, since there is a single component that is 99.4% of the entire graph.

Table II. Results for Graphs: INTRANET, HOSTGRAPH, and GEO

Subgraph	Nodes $\times 10^3$	Arcs $\times 10^3$	Indeg. exp.	SCC exp.	WCC exp.	WCC $\times 10^3$	SCC/ WCC	IN/ WCC	OUT/ WCC	$K_{5,7}$ factor
INTRANET	285.5	1910.7	−2.31	−2.53	−2.83	207.7	0.20	0.48	0.17	56.13
HOSTGRAPH	663.7	1127.9	−2.34	−2.81		659.9	0.82	0.04	0.13	72.64
GEO	410.7	1477.9	−2.51	−2.69	−2.27	2.1	0.87	0.03	0.10	139.9

Table III. Results for Single Keyword Query Graphs KEYWORD1 through KEYWORD5

Subgraph	Nodes $\times 10^3$	Arcs $\times 10^3$	Indeg. exp.	SCC exp.	WCC exp.	WCC $\times 10^3$	SCC/ WCC	$K_{5,7}$ factor
BASEBALL	336.5	3444.4	-2.09	-2.16	-2.30	33.2	0.12	55.85
GOLF	696.8	8512.8	-2.06	-2.06	-2.18	47.3	0.15	44.48
MATH	831.7	3787.8	-2.85	-2.66	-2.73	50.2	0.28	148.7
MP3	497.3	7233.2	-2.20	-2.39	-2.20	47.6	0.28	57.18
RESTAURANT	623.0	3592.5	-2.33	-2.47	-2.28	7.96	0.31	115.2

Table IV. Results for Double Keyword Query Graphs KEYWORD6 through KEYWORD10

Subgraph	Nodes $\times 10^3$	Arcs $\times 10^3$	Indeg. exp.	SCC exp.	WCC exp.	WCC $\times 10^3$	SCC/ WCC	$K_{5,7}$ factor
BASEBALL YANKEES	24.0	320.0	−2.11	−2.35	−2.27	3.81	0.73	45.82
GOLF TIGER WOODS	14.9	62.8	−2.07	−2.10	−2.15	1.50	0.20	83.02
MATH GEOMETRY	44.0	86.9	−2.58	−2.65	−2.78	1.90	0.27	407.52
MP3 NAPSTER	27.1	321.4	−2.20	−2.35	−2.20	1.76	0.36	35.19
RESTAURANT SUSHI	7.4	23.7	−2.19	−2.40	−2.20	0.17	0.72	132.14

Table V. Averaged Results for SUBDOMAIN1 through SUBDOMAIN100

Nodes $\times 10^3$	Arcs $\times 10^3$	Indeg. exp.	SCC exp.	WCC exp.	WCC $\times 10^3$	SCC/ WCC	$K_{5,7}$ factor
7.17	108.42	−2.11	−2.20	−2.30	7.08	0.42	22.97

Table III shows the results for single keyword queries. The graphs in the category are only in few hundreds of thousands. Table IV shows the results for double keyword graphs. The graphs in this category are in few tens of thousands. (Since the graphs in this category are relatively small and more fragmented, many parameters that were presented for larger graphs do not make statistical sense, and therefore we drop them from our tables.) Note that the ratio of WCC to the total number of nodes for these graphs is much smaller than the corresponding value for the other graphs, suggesting that these graphs might have a different overall connectivity structure. Another specific interesting case is the large $K_{5,7}$ factor for the keyword MATH, which probably arises because pages containing the term MATH is probably not a TUC since it is far too general.

Table V shows the averaged results for the 100 sites SUBDOMAIN1, . . . , SUBDOMAIN100.

Next, we point out the consistency of the parameters across various graphs. For ease of presentation, we picked a small set of TUCs and plotted the distribution of indegree, outdegree, SCC, WCC on a log-log scale (see the figures in Section 4.1). Figure 1 shows the indegree and outdegree distributions for five of the TUCs. As we see, the shape of plots are strikingly alike. As observed in earlier studies, a drooping initial segment is observed in the case of

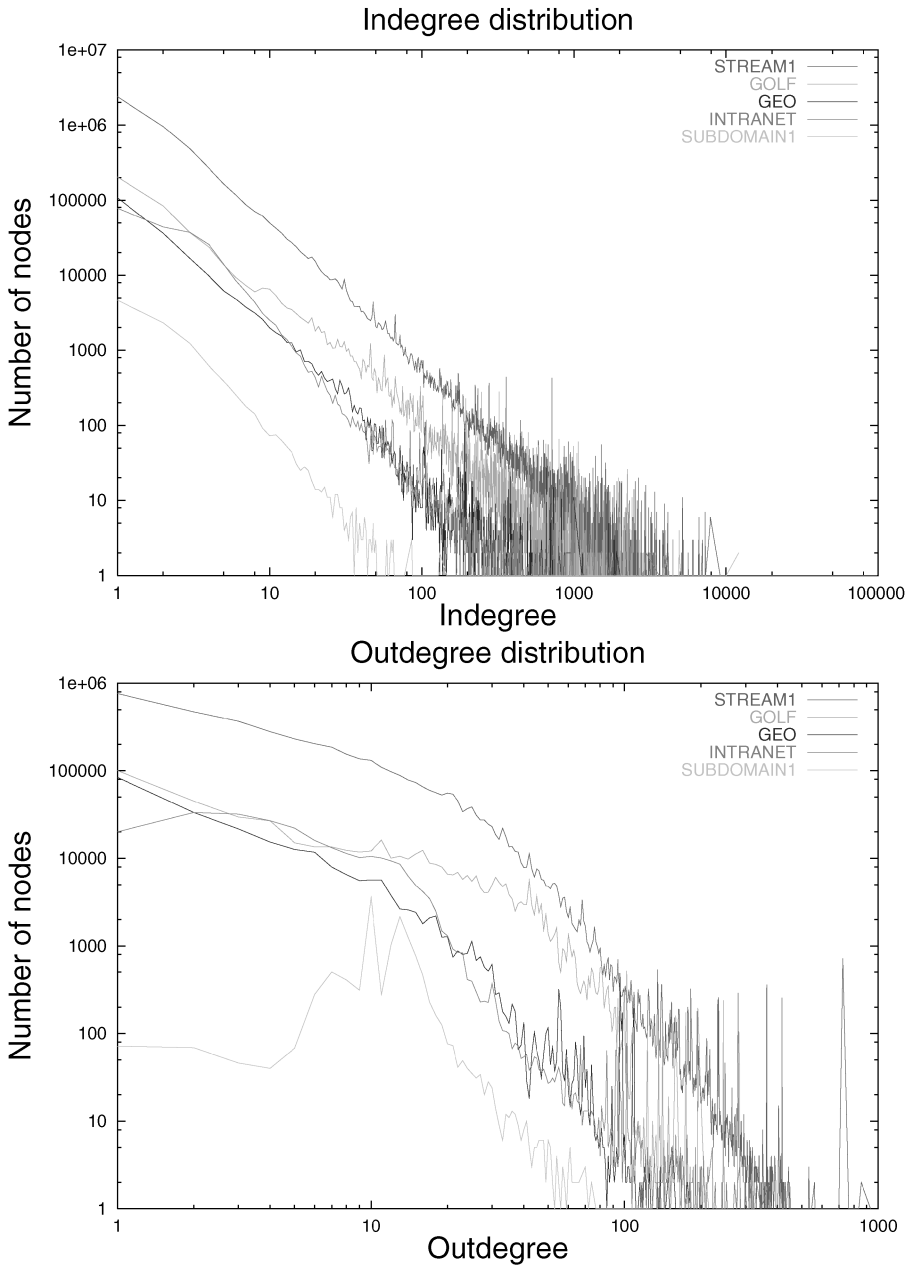

Fig. 1. Indegree and outdegree distributions for STREAM1, GOLF, GEO, INTRANET, SUBDOMAIN1. X-axis denotes the degree d and Y-axis denotes the number of nodes with degree d.

outdegree. Figure 2 shows the component distributions for the graphs. Again, the similarity of shapes is striking. The URL tree sizes also show remarkable self-similarity, which exists both across graphs and within each graph across different depths (see Figure 3).

4.1 Discussion

We now mention four interesting observations based on the experimental results. Following Broder et al. [2000] (see also Section 2), we say that a slice of the Web graph *has the bowtie structure* if the SCC, IN, and OUT, each accounts for a large constant fraction of the nodes in the slice.

(1) Almost all nodes (82%) of the HOSTGRAPH are contained in a giant SCC (Table II). This is not surprising, since one would expect most Websites to have at least one page that belongs to the SCC.

(2) The (microscopic) local graphs of SUBDOMAIN1, . . . , SUBDOMAIN100, look surprisingly like the Web graph (see Table V). Each has an SCC flanked by IN and OUT sets that, for the most part, have sizes proportional to their size on the Web as a whole, about 40% for the SCC, for instance. Large Websites seem to have a more clearly defined bowtie structure than the smaller, less developed ones.

(3) Keyword based TUCs corresponding to KEYWORD1, . . . , KEYWORD10 (see Tables III and IV) seem to have a different overall connectivity structure, since the WCC is only a small portion of the whole TUC. When restricted to WCC, however, these TUCs exhibit similar phenomena as other TUCs; the differences are often due to the extent to which a community has a well-established presence on the Web. For example, it appears from our results that the GOLF is a well-established Web community, while RESTAURANT is a newer developing community on the Web. While the mathematics community had a clearly defined bowtie structure, the less developed geometry community lacked one.

(4) Considering STREAM1, . . . , STREAM7, we find that (Table I) the union of a random collection of TUCs contains a large SCC. This shows that the SCC of the Web is very resilient to node deletion and does not depend on the existence of large taxonomies (such as `yahoo.com`) for its connectivity. Indeed, as we remarked earlier, each of these streams contain very few arcs that are not entirely local to the Website. However, the bowtie structure of each Website allows the few intersite arcs to be far more valuable than one would expect.

4.2 Analysis and Summary

The foregoing observation about the SCC of the streams, while surprising, is actually a direct consequence of the following theorem about random arcs in graphs with large strongly-connected components.

THEOREM 1. *Consider the union of n/k graphs on k nodes each, where each graph has a strongly-connected component of size αk. Suppose we add dn arcs whose heads and tails are uniformly distributed among the n nodes, then, provided that d is at least of the order $1/(\alpha k)$, with high probability we will have a strongly-connected component of size of the order of αn on the n-node union of the n/k graphs.*

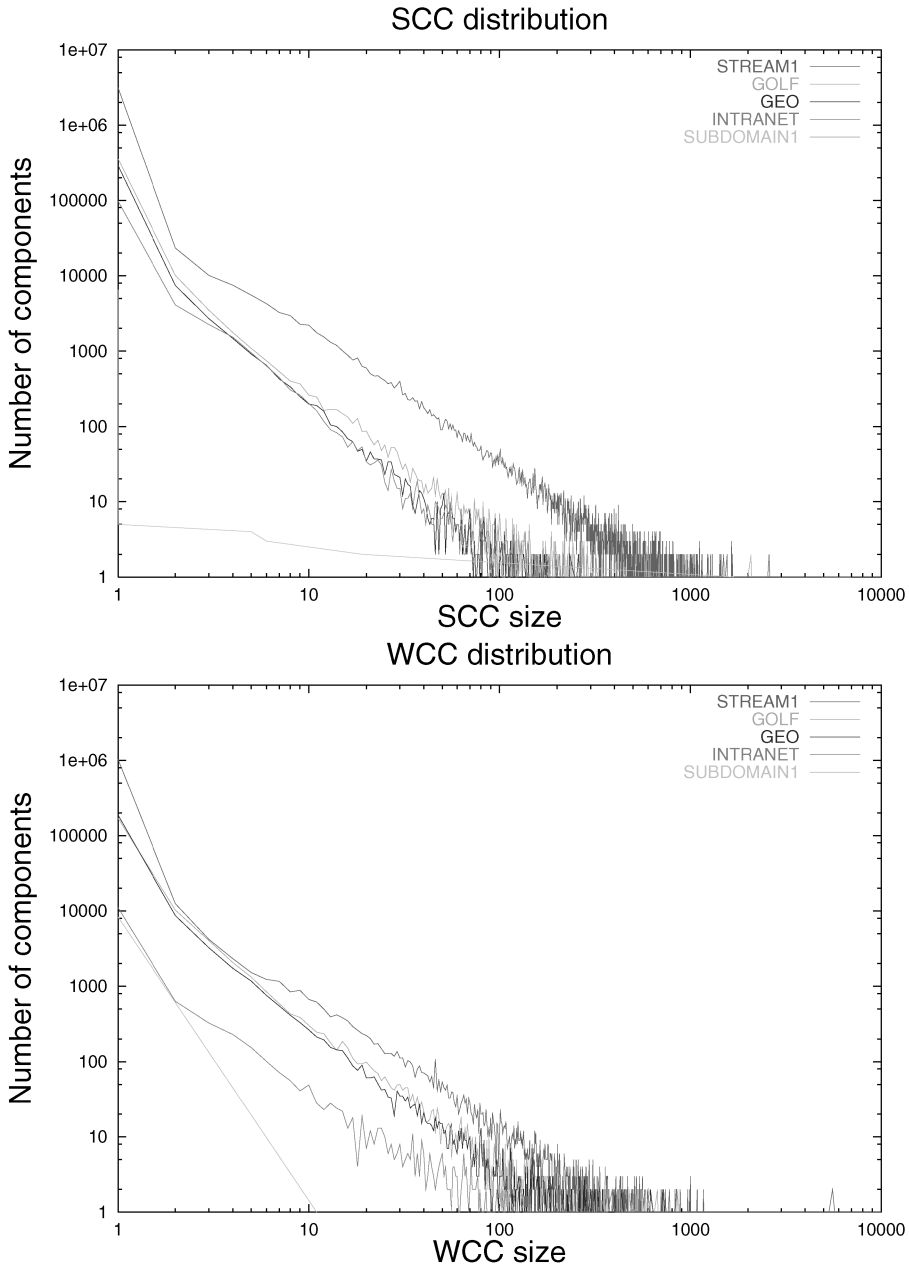

Fig. 2. SCC and WCC distributions for STREAM1, GOLF, GEO, INTRANET, SUBDOMAIN1. X-axis denotes the size s and Y-axis denotes the number of components with size s.

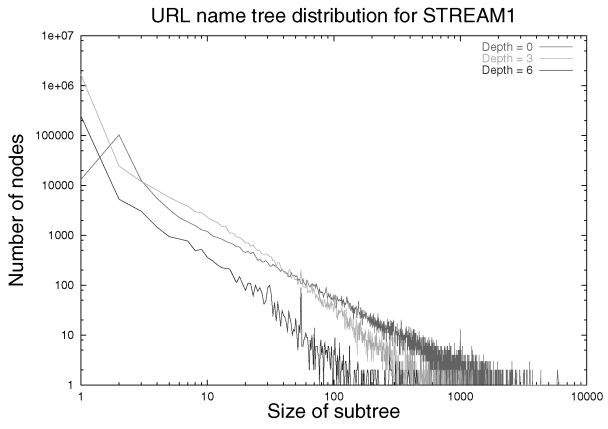

Fig. 3. (a) and (b): Self-similarity in URL name trees for STREAM1 and GEO at depths 0, 3, and 6. (c) Self-similarity in URL name trees between STREAM1, GOLF, GEO, and the INTRANET at depth 3. X-axis denotes the size s and Y-axis denotes the number of depth d-nodes whose subtrees have s nodes.

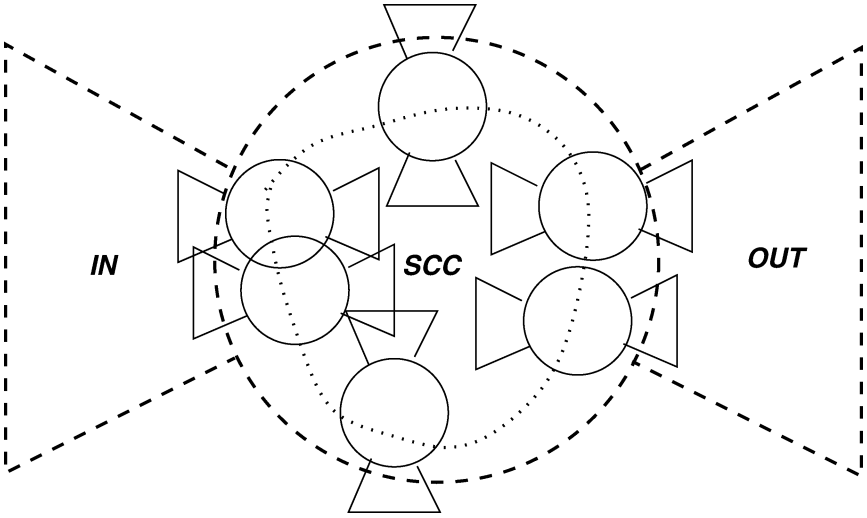

Fig. 4. TUCs connected by the navigational backbone inside the SCC of the Web graph.

The proof of Theorem 1 is fairly straightforward. On the Web, n is about 1 billion, k, the size of each TUC, is about 1 million (in reality, there are more than 1 K TUCs that overlap, which only makes the connectivity stronger), and α is about 1/4. Theorem 1 suggests that the addition of a mere few thousand arcs scattered uniformly throughout the billion nodes will result in very strong connectivity properties of the Web graph!

Indeed, the evolving copying models for the Web graph proposed in Kumar et al. [2000] incorporates a uniformly random component together with a copying stochastic process. Our observation above, in fact, lends considerable support to the legitimacy of this model. These observations, together with Theorem 1, imply a very interesting detailed structure for the SCC of the Web graph.

The Web comprises several thematically unified clusters *(TUCs). The common theme within a TUC is one of many diverse possibilities. Each TUC has a* bowtie structure *that consists of a large strongly-connected component (SCC). The SCCs corresponding to the TUCs are integrated, via the* navigational backbone, *into a global SCC for the entire Web. The extent to which each TUC exhibits the bowtie structure and the extent to which its SCC is integrated into the Web as a whole indicate how well-established the corresponding community is.*

An illustration of this characterization of the Web is shown in Figure 4.

5. CONCLUSIONS

In this article we have examined the structure of the Web in greater detail than earlier efforts. The primary contribution is twofold. First, the Web exhibits self-similarity in several senses, at several scales. The self-similarity is pervasive, in that it holds for a number of parameters. It is also robust, in that it holds irrespective of which particular method is used to carve out small subgraphs of the Web. Second, these smaller thematically unified subgraphs are organized

into the Web graph in an interesting manner. In particular, the local strongly-connected components are integrated into the global SCC. The connectivity of the global SCC is very resilient to random and large-scale deletion of Websites. This indicates a great degree of fault tolerance on the Web, in that there are several alternate paths between nodes in the SCC.

While our understanding of the Web as a graph is greater now than ever before, there are many holes in our current understanding of the graph-theoretic structure of the Web. One of the principal holes deals with developing stochastic models for the evolution of the Web graph (extending Kumar et al. [2000]) that are rich enough to explain the fractal behavior of the Web in such amazingly diverse ways and contexts.

ACKNOWLEDGMENTS

Thanks to Raymie Stata and Janet Wiener (Compaq SRC) for some of the code. The second author thanks Xin Guo for her encouragement for this project.

REFERENCES

ABITEBOUL, S., QUASS, D., MCHUGH, J., WIDOM, J., , AND WIENER, J. 1997. The Lorel query language for semistructured data. *Int. J. Digital Libr. 1*, 1, 68–88.

ADAMIC, L. AND HUBERMAN, B. 2000. The nature of markets on the world wide web. *Q. J. Econ. Commerce 1*, 1, 5–12.

ADAMIC, L. AND HUBERMAN, B. 1999. Scaling behavior on the world wide web. Technical comment on Barabasi and Albert [1999].

ADLER, M. AND MITZENMACHER, M. 2001. Towards compressing web graphs. In *Proceedings of the IEEE Data Compression Conference*. To appear.

AIELLO, W., CHUNG, F., AND LU, L. 2000. A random graph model for massive graphs. In *Proceedings of the 32nd STOC*, 171–180.

ARASU, A., TOMKINS, A., AND TOMLIN, J. 2001. Pagerank computation and the structure of the web: Experiments and algorithms. Manuscript, 2001.

AROCENA, G., MENDELZON, A., AND MIHAILA, G. 1997. Applications of a web query language. In *Proceedings of the 6th WWW/Comput. Netw. 29*, 8–13, 1305–1315.

BARABASI, A. AND ALBERT, R. 1999. Emergence of scaling in random networks. *Science 286*, 509.

BARFORD, P., BESTAVRO, A., BRADLEY, A., AND CROVELLA, M. E. 1999. Changes in web client access patterns: Characteristics and caching implications. *World Wide Web, Special Issue on Characterization and Performance Evaluation 2*, 15–28.

BHARAT, K. AND HENZINGER, M. 1998. Improved algorithms for topic distillation in hyperlinked environments. In *Proceedings of the 21st SIGIR*, 104–111.

BOLLOBAS, B. 1985. *Random Graphs.* Academic Press.

BOTAFOGO, R. A. AND SHNEIDERMAN, B. 1991. Identifying aggregates in hypertext structures. In *Proceedings of the 3rd Hypertext Conference* (1991), 63–74.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large scale hypertextual web search engine. In *Proceedings of the 7th WWW/Comput. Netw. 30*, 1–7, 107–117.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., AND WIENER, J. 2000. Graph structure in the web. In *Proceedings of the 9th WWW/Comput. Netw. 33*, 1–6, 309–320.

CARRIERE, J. AND KAZMAN, R. 1997. WebQuery: Searching and visualizing the web through connectivity. In *Proceedings of the 6th WWW 29*, 8–13, 1257–1267.

CHAKRABARTI, S., DOM, B., GIBSON, D., KLEINBERG, J., RAGHAVAN, P., AND RAJAGOPALAN, S. 1998a. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th WWW/Comput. Netw. 30*, 1–7, 65–74.

CHAKRABARTI, S., DOM, B., GIBSON, D., RAVI KUMAR, S., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1998b. Experiments in topic distillation. In *SIGIR Workshop on Hypertext Information Retrieval on the Web*.

CHAKRABARTI, S., VAN DEN BERG, M., AND DOM, B. 1999a. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings of the WWW/Comput. Netw. 31*, 11–16, 1623–1640.

CHAKRABARTI, S., GIBSON, D., AND MCCURLEY, K. 1999b. Surfing the web backwards. In *Proceedings of the 8th WWW/Comput. Netw. 31*, 11–16, 1679–1693.

DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T., AND HARSHMAN, R. 1990. Indexing by latent semantic analysis. *J. ASIS 41*, 6, 391–407.

EDWARDS, J., MCCURLEY, K., AND TOMLIN, J. 2001. An adaptive model for optimizing performance on an incremental web crawler. In *Proceedings of the 10th WWW*, 106–113.

FALOUTSOS, M., FALOUTSOS, P., AND FALOUTSOS, C. 1999. On power law relationships of the internet topology. In *Proceedings of the ACM SIGCOMM Conference*, 251–262.

GLASSMAN, S. 1994. A caching relay for the world wide Web. In *Proceedings of the 1st WWW/Comput. Netw. 27*, 2, 165–173.

HARARY, F. 1975. *Graph Theory*. Addison Wesley.

HILL, B. 1975. A simple approach to inference about the tail of a distribution. *Ann. Stat. 3*, 5, 1163–1174.

HUBERMAN, B., PIROLLI, P., PITKOW, J., AND LUKOSE, R. 1998. Strong regularities in world wide web surfing. *Science 280*, 95–97.

KLEINBERG, J. 2000. Authoritative sources in a hyperlinked environment. *J. ACM 46*, 5, 604–632.

KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., SIVAKUMAR, D., TOMKINS, A., AND UPFAL, E. 2000. Stochastic models for the web graph. In *Proceedings of the 41st FOCS Conference*, 57–65.

KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1999a. Trawling the web for cyber communities. In *Proceedings of the 8th WWW/Comput. Netw. 31*, 11-16, 1481–1493.

KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1999b. Extracting large scale knowledge bases from the web. In *Proceedings of the Conference on Very Large Data Bases*, 639–650.

KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 2000. On semi-automated taxonomy construction. In *Proceedings of the 4th Web Data Base Conference*.

LUKOSE, R. M. AND HUBERMAN, B. 1998. Surfing as a real option. In *Proceedings of the 1st International Conference on Information and Computation Economies*.

MARTINDALE, C. AND KONOPKA, A. K. 1996. Oligonucleotide frequencies in DNA follow a Yule distribution. *Comput. Chem. 20*, 1, 35–38.

MENDELZON, A., MIHAILA, G., AND MILO, T. 1997. Querying the world wide web. *J. Digital Libr. 1*, 1, 68–88.

MENDELZON, A. AND WOOD, P. 1995. Finding regular simple paths in graph databases. *SIAM J. Comput. 24*, 6, 1235–1258.

PALMER, E. M. 1985. *Graphical Evolution*. Wiley.

PAPADIMITRIOU, C., RAGHAVAN, P., TAMAKI, H., AND VEMPALA, S. 2000. Latent semantic indexing: A probabilistic analysis. *JCSS 61*, 2, 217–235.

PARETO, V. 1897. *Cours d'economie politique*. Rouge, Lausanne et Paris.

PIROLLI, P., PITKOW, J., AND RAO, R. 1996. Silk from a sow's ear: Extracting usable structures from the web. In *Proceedings of the ACM SIGCHI Conference*, 118–125.

PITKOW, J. AND PIROLLI, P. 1997. Life, death, and lawfulness on the electronic frontier. In *Proceedings of the ACM SIGCHI Conference*, 383–390.

SIMON, H. A. 1955. On a class of skew distribution functions. *Biometrika 42*, 425–440.

SPERTUS, E. AND STEIN, L. 1998. A hyperlink-based recommender system written in Squeal. In *Proceedings of the CIKM Workshop on Web Information and Data Management*.

SPERTUS, E. 1997. ParaSite: Mining structural information on the web. In *Proceedings of the 6th WWW/Comput. Netw. 29*, 8–13, 1205–1215.

WHITE H. D. AND MCCAIN, K. W. 1989. Bibliometrics. *Ann. Rev. Inf. Sci. Technol.*, 119–186.

YULE, G. U. 1944. *Statistical Study of Literary Vocabulary*. Cambridge University Press.

ZIPF, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Received June 2001; revised April 2002; accepted May 2002