Link Structure of Hierarchical Information Networks

Nadav Eiron and Kevin S. McCurley
IBM Almaden Research Center

ABSTRACT

One feature that seems to have been largely ignored in previous models of the Web is the inherent hierarchy that is evident in the structure of URLs. We provide evidence that this hierarchical structure is closely related to the link structure of the web, and this relationship explains several important features of the web, including the locality and bidirectionality of hyperlinks, and the compressibility of the web graph. We describe how to construct data models of the web that capture both the hierarchical nature of the web as well as some crucial features of the link graph. Our analysis is based on observations from a crawl of over a billion URLs, as well as large-scale simulations of models. We also show how this interaction between hierarchical structure and link structure extends to other domains. In particular we describe some analysis on corporate instant messaging, in which there is similar correspondence between the corporate management structure and patterns of communication between individuals.

1. INTRODUCTION

The structure of hyperlinks on the World Wide Web has proved to be useful for the tasks of clustering, community extraction [17], classification [3], ranking of pages [24, 15], and identification of document structure [11]. The Web link structure therefore 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 hierarchy that is evident in the structure of URLs. For example, in the URL `http://ibm.com/products/server/` we might expect to find product information about servers, and we might further expect to find a layer below this containing information about the individual models. This hierarchical structure of the web reflects a very common practice for information in general, namely that it is often 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[1]. 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. The hierarchical structure on the web can be traced to the fact that most early web servers were built to retrieve files from their file system, and URLs were mapped directly to a subtree of their file system. To this day a large amount of web content still reflects this organization.

In addition to the hierarchical structure of file paths on a server, URLs also reflect another layer of hierarchy from the domain name system (DNS), where domains are categorized by their top level domain of `edu`, `org`, etc. By reversing the direction of the hostname in the URL structure, we can think of the web as being organized as a single tree, with the top levels of the tree being the top level domains, continuing down to the individual servers, and down into the information structure under a hostname. The hierarchical structure of DNS evolves under different social conventions from locally managed file systems, and the top levels of the hierarchy are managed centrally. At lower levels of the DNS, organization of the hostname to address mapping in DNS falls under local jurisdiction, and in some cases the local management of DNS resembles the pattern of organization found in local file systems. This dividing line where organization becomes local is the point at which the Web becomes a distributed information system, and this is the point at which social processes affect teh structure. Thus while it is possible to think of the Web as a single tree rooted at the top of the DNS hierarchy, it is perhaps more natural to model it as a forest rather than a tree, where the trees represent individual web sites or sub-domains.

In his seminal work on complex systems, Simon [29] 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 of information and social groups that administer the information. In particular, we shall provide evidence that hyperlinks tend to exhibit a "locality" that is correlated to the hierarchical structure of URLs, and that

[1]We follow the peculiar convention of computer scientists that trees have their leaves at the bottom.

many features of the organization of information in the web are predictable from knowledge of the hierarchical structure.

Our contributions are in three areas. First, we describe the results of some statistical analysis carried out on an extremely large sample from the World Wide Web. We then point out how previous models of the web fail to adequately predict some important characteristics of the web, including the correlation between hierarchical structure and link structure, a measure of entropy for links (as evidenced by compressibility of the web graph), the growth of individual web sites, and bidirectionality of links on the web. We then describe a new hierarchical approach to modeling the web that incorporates these characteristics. Our primary focus is on describing a class of models and how the observed structure is reflected in the model, rather than on precise mathematical analysis of a particular abstract model.

We believe that the hierarchical structure is at least as important as the hyperlink structure for understanding the organization of information on the web. Moreover, the interaction between the hierarchical structure and the hyperlink structure reveals even more than the individual structures by themselves. 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. In the case of the World Wide Web there is an obvious hierarchical structure, but it is clearly not the only taxonomy that can be constructed, and we expect that link information can also aid in automatic identification of hierarchical structure as well as improved algorithms for classification of hypertext [7].

While the motivation for this work arises from the information structure of the World Wide Web, we expect this feature to appear in other information networks for which there is an obvious hierarchical information organization, e.g., scientific citations, legal citations, patent citations, etc. Moreover, we expect that it will also apply to certain types of social networks, and in Section 9 we consider the specific case of social networks formed in corporate instant messaging, for which there is a natural hierarchical structure.

The rest of the paper is structured as follows. In the next section we review some previous work on models of the web and information systems. In section 3 we describe the data set and methodology that is used for our observations and experiments. In Section 4 we describe some previous work on modeling of random trees and forests, and its relationship to modeling of the web. In section 5 we discuss locality measures for hyperlinks, and evidence for the fact that this locality follows the URL hierarchy. In section 6 we examine the situation in which links are bidirectional. In section 7 we outline our requirements for a model of the Web, and describe our hierarchical paradigm for modeling of the web. In Section 8 we describe a measure of entropy for links under different models, and compare our model to others. In Section 9 we describe some analysis on corporate instant messaging logs and how the same phenomenon between hierarchy and links occurs in other realms. We conclude in Section 10 with a summary and some opportunities for future work.

2. PREVIOUS MODELS OF THE WEB

In recent years there has been an explosion of published literature on the subject of models for networked systems, including the World Wide Web, social networks, technological networks, and biological networks. For coverage of this we refer the reader to the survey by Newman [23]. Much of this work is in the spirit of Simon's work on complex systems; attempting to explain various features such as degree distribution that are ubiquitous across very different kinds of systems. Examples include:

Small world structure Most pairs of nodes are connected by relatively short paths.

Degree distributions In-degrees and out-degrees of nodes often appear to have a heavy-tailed distribution,

Transitivity the neighbors of a node are often neighbors of each other, producing graphs that have a relatively large number of "triangles".

Community The graph contains subsets with relatively high density of edges between nodes in the subset, but relatively low density of edges between nodes in different subsets.

Beyond these generic characteristics that show up across many different classes of networks, there are other features that may be unique to a particular type of network such as the World Wide Web. Some of these are due to the directed nature of the Web, but others are specific to the structure of information that the Web represents. A complete survey of previous evolutionary models for the World Wide Web is beyond the scope of this paper, and once again we refer the reader to the survey by Newman [23] to summarize the developments up until 2003.

Models of the web are generally defined as stochastic processes in which edges and nodes are added to the graph over time in order to simulate the evolution of the web (or any other network). Such models fall broadly into two categories. The first category is those that rely upon Price's concept of *cumulative advantage*, also sometimes referred to as *preferential attachment* or "the rich get richer". In this model, the probability of adding an edge with a given destination and/or source is dependent on the existing in or out degree of the node (usually in a linear fashion). The linear dependence on the existing degree can be varied to incorporate a mixture of two processes, in which cumulative advantage is mixed with some fraction of uniform assignments of edges [25]. The second class of models uses a notion of *evolving copying*, in which the destinations for edges from a node are copied as a set from an existing node chosen under some distribution [16].

In section 7 we will present a new paradigm for constructing models of information networks that incorporates their hierarchical structure. It is our hope that by breaking the web down into the component features of site size, hierarchical structure of information, and link structure, we will present a useful paradigm for future analysis that incorporates multiple features of the web. It should be noted that the hierarchical evolution of structure can be combined with previous techniques of cumulative advantage or copying.

A hierarchical model of the web was previously suggested by Laura et. al. [18]. 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 implied in the structure of URLs to establish the regions, which reflects a social division by organization.

Another recent model that incorporates hierarchical structure was proposed in [28]. Their model is generated in a very regular fashion, by starting with a small graph of five nodes, and replicating it five times, and joining these replicas together, and recursing this procedure. The resulting graph is shown to exhibit a clustering coefficient that resembles many real networks. Another recent model that results in a hierarchical organization of nodes was proposed in [6]. In both cases the models are fairly simple, and are designed to produce some specific generic properties such as clustering coefficient and degree distributions.

3. EXPERIMENTAL METHODOLOGY

Our observations are based on examination of a large subset of the Web that has been gathered at IBM Almaden since 2002. At the time of our experiments, the crawl had discovered at least 5.6 billion URLs on over 48 million hosts. For our analysis of tree structure we used the complete set of URLs, and for our analysis of link structure we used the first billion crawled URLs. For some of our experiments, we sampled from among the crawled 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 well approximated by a breadth first search.

More than 40% of the URLs discovered 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. Unfortunately, an increasing number of web sites use such URLs to retrieve standard textual content, encode session IDs, or indicate viewer preferences, and it is extremely difficult to distinguish these cases. Moreover, even if the URL lacks a `?` character, the content may still come from a relational database query that is encoded using a different convention. Because the purpose of this study is to investigate the relationship between hyperlinks and hierarchical organization of knowledge, we excluded URLs containing a `?` from our study altogether. As the web continues to grow, we expect this feature to become increasingly important.

We found that a significant fraction of the sites and pages from the crawl were pornographic in nature. The structure of these sites and the links between them is driven by a different social process from the rest of the web. In particular they are aggressive in their attempt to enhance their search engine rankings, and search engines are aggressive in their efforts to remove them. For these reasons, we used a simple classification scheme to remove these from our experimental set. We also chose to exclude URLs from our link analysis if they do not represent hypertext content (e.g., postscript, images, and other data types), as they represent leaf nodes without outlinks.

4. THE WEB FOREST

At a coarse level of granularity, we can think of the web as a collection of hosts that grow more or less independently of each other. The distribution of the number of URLs per host is shown in Figure 1. The number of URLs per host was previously studied by Adamic and Huberman [14], who hypothesized that the growth of an individual web site is modeled as a multiplicative process, resulting in a lognormal distribution. This provides a model of the web as a mixture multiplicative processes, leading to a power law tail for the distribution. Indeed, our data in Figure 1 seems to exhibit a power law distribution in the tail (though there is some variation with a hump in the middle).

The use of multiplicative models for web site growth can be formulated as follows. Let $S(t)$ denote the size of a web site at time t, and let us hypothesize that $S(t) = g(t)S(t-1)$ for $t \geq 1$. In this case we have that

$$\log(S(t)) = \log(S(0)) + \sum_{i=1}^{t-1} \log(g(i)).$$

If the $g(i)$ are i.i.d., with finite mean and variance, then the central limit theorem suggests that $\log(S(t))$ would have an asymptotically normal distribution, which means that $S(t)$ would have a lognormal distribution. This is the model was suggested in [14], but there are several subtleties that underly this model (see [21]). For one thing, the values of $S(t)$ must be integers, which places somewhat unreasonable restrictions on the choice of $g(t)$. An alternative multiplicative model has been suggested by Reed (see also [21]), giving rise to a double Pareto distribution that combines two Pareto distributions at an inflection point. In either case the tail of the distribution ends up being a power law, and the only question is how to model the data at the head of the distribution.

Note however that statistics on web site size distribution are strongly affected by crawl strategy, and in particular many of the hosts that appear to have few pages are either protected from crawling by the existence of a `robots.txt` file, or else they are merely a redirect to another site with a different name. Moreover, resource restrictions dictate that our crawl truncates extremely large sites. These issues are of little concern to us because we only seek to model information networks as they appear to data mining applications.

One appealing aspect of the multiplicative model is that $S(t)$ is allowed to *shrink*, which is potentially important since web pages seem to disappear at a rapid rate. There are variations using this approach that can be proposed here, but the dynamics of the web seem to involve a number of complicated factors. For example, recent attempts to manipulate search engine rankings have led to a proliferation of many small cooperative sites that aggregate links to a single site in order to boost its rating, and these tend to distort observations of the size of sites.

Another complication of modeling the size of web sites arises from the growing presence of databases exposed through an HTTP interface. For example, if a relational database table is exposed through HTML content containing dynamic links to facilitate exploration of the table, then it may result in a number of URLs that is exponential in the number of rows and columns of the database table (reflecting the number of possible queries on the table). For sites constructed this way the rate of growth will be very bursty, and their web

pages would tend to have a very regular pattern of links, and therefore would not be well modeled by the work described here. Since we are interested in hierarchical organization of *information* rather than relational organization of *data*, we have specifically excluded URLs from our study if they contain a ? character. This minimizes the effect of relational data exposed through HTTP, but there remain many issues to be worked out in modeling the size of web sites accurately.

4.1 Tree shapes

Also shown in Figure 1 is the distribution of the number of directories per site. Based on observations from 60 million URLs on 1.3 million sites, it was previously observed in [10] that the size of directory subtrees at a given depth appears to follow a power law, which is consistent with our observations on a much larger data set. Note that in contrast with the distribution of URLs per site, the number of directories per site seems to behave more like a pure power law. We will return to this point in section 4.2.

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 sorted the static URLs in our set of 5.6 billion URLs by directory order. Then for each directory 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 2. Once again, the shapes of the distributions suggest that both of these are distributed as a power law distribution. Using the technique describe in [8], we estimate that the probability of finding more than n URLs in a subdirectory is approximately $c/n^{1.20}$ for large n, and the probability of finding more than d subdirectories of a directory is approximately $c/d^{1.43}$.

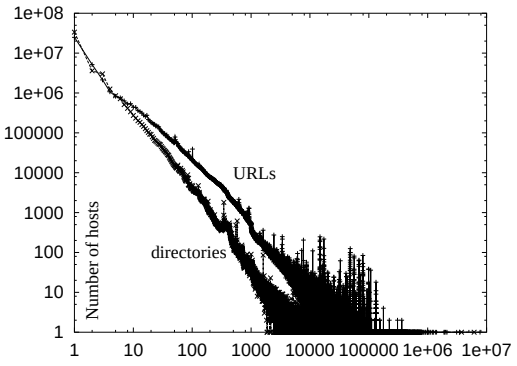

Figure 1: Number of pages and directories per host.

One might wonder whether the process of creating directories and creating URLs within the directory are correlated to each other, either positively or negatively, i.e., whether the existence of many URLs in a non-leaf directory is correlated to whether the directory has many (or few) subdirectories. In order to test this hypothesis, we computed a Goodman-Kruskal Gamma statistic [13] on fanouts and URL counts for a sample of non-leaf directories from our data set. Our computations suggest that they are only slightly concordant, so it is probably safe to model them as independent processes.

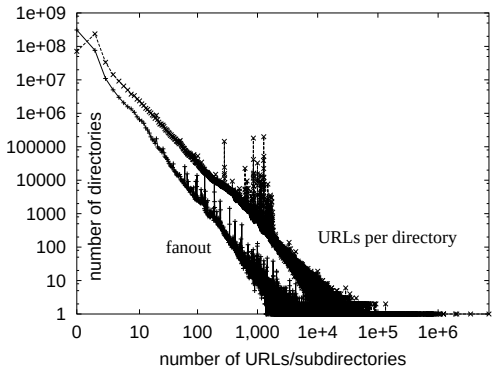

Figure 2: Number of subdirectories of a directory and number of URLs within individual directories. Note that some internal node directories have no URLs in them.

4.2 Growth of Trees

In order to understand the distribution of links in the hierarchy, we first need to understand the structure of directory trees for web sites. At least two models for the shape of random trees have been studied extensively, namely the class of random recursive trees and the class of plane-oriented random recursive trees (see [30]). Random recursive trees are built up by selecting a node uniformly at random and attaching a new child to it. This class of trees results in a fanout distribution where $\Pr(\text{degree} = k) \sim 2^{-k}$, and are thus unsuitable for describing the type of trees seen here. By contrast, the construction of plane oriented trees chooses a node with probability proportional to 1+degree, resulting in a fanout distribution where $\Pr(\text{degree} = k) \approx \frac{4}{(k+1)(k+2)(k+3)}$, or $\approx 4/k^3$ for large k. This therefore gives a power law for the fanout distribution, but with the wrong exponent. It is however a simple matter to modify the plane-oriented model to incorporate a mixture between cumulative advantage and uniform attachment in order to achieve a more realistic degree distribution for trees.

The web does not grow as a single tree however; it grows as a forest of more or less independent trees. Rather little has been written about models for random forests, but one model was studied in [4]. In their model, forests with node set $\{1, \ldots, n\}$ are grown as follows. Designate 1 as a root, and in order to determine the vertex k for $k = 2, 3, \ldots, n$, we select y from $\{0, 1, \ldots, k-1\}$ with probability $p_k(y)$. If $y \in \{1, \ldots, n\}$ then we join k to y, and if $y = 0$ then we designate k as a new root. Most of their results are however concerned with the special case of $p_k(y) = 1/k$, in analogy to the uniform random recursive tree model. Among other things, they proved that in this model the number of trees in a forest of size n nodes is asymptotically $\log n$. In our data set from the web, we found approximately 48 million trees (websites) in a forest of 417 million nodes (directories). For this reason alone, the model of [4] does not seem appropriate as a model of the web forest, as it would have predicted a much smaller number of websites.

Random recursive forests were also considered by Mitzenmacher [22], where his goal was to construct a model of file sizes in a file system. In his model, files are either created from scratch according to a fixed distribution, or else they

are created by copying an existing file and modifying it. The trees are then used to model the evolution of files, and he used a constant probability of creating a new root at each step, resulting in many more trees in the forest than the uniform recursive model of [4]. We adopt a similar strategy in our model of web forest growth, by maintaining a constant probability of creating a new web site at each time step. It may be the case that this probability will vary over time, but we leave this to future studies of web evolution.

Another potential problem is related to the fact that web sites tend to grow largely independent of each other, whereas in [4] the placement of new leaves in the forest is dependent on the structure of the entire existing forest. In reality, the particular size of one web site usually has no bearing on the size of another web site (excluding mirrors and hosts within a single domain). For this reason we believe it is natural to model the growth of the forest as a collection of independent trees.

There are a number of interesting statistics that might be investigated concerning the growth of the web forest. One difference between the two classes of random trees is found in the number of leaves. For random recursive trees the expected number of leaves is asymptotically 1/2 the number of nodes, whereas for plane-oriented random recursive trees the expected number of leaves is asymptotically 2/3 of the number of nodes. In the case of our web sample, we found that for hosts with more than 10 directories, the average number of leaves was 60%, and across the entire web the number of leaves is 74%. Hence the presence of many small sites on the web with few directories contributes many of the leaves. Note also that the 60% figure agrees with our suggestion to interpolate between plane-oriented trees and uniform recursive trees.

5. LINK LOCALITY

It has been observed in various contexts that links on the web seem to exhibit various forms of "locality". We loosely use the term locality to mean that links tend to be correlated to pages that are "nearby" in some measure. The locality of links in the web is important for various applications, including the extraction of knowledge on unified topics and the construction of efficient data structures to represent the links. We shall consider the latter issue in Section 8.

In practice there are various measures of locality that one might consider. Watts and Strogatz defined the concept of a *clustering coefficient*, and similar measures have been studied by others (see [23, § IIIB]). The clustering coefficient is a local measure of how often a page will link to two pages that link to each other, although it is usually studied in the context of undirected graphs. Experiments by Adamic [1] on a set of 100 million Web pages in 1998 showed that the clustering coefficient for the Web is relatively large, and this provides at least one form of evidence for locality in links.

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 explain directly by a model. Davison [9] and Menczer [20] have studied a more natural measure of locality in the form of "topical locality", based on the observation that pages linked to or from a given page are usually on a similar topic. A similar point of view can be found in work toward identifying community structure in the web [16].

5.1 Locality and Hierarchy

None of these measures take into account the purposes for which links are created. We believe 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. Specifically, links can be of two types: navigational links within a cohesive set of documents, and informational links that extend outside the corpus developed by the author. Navigational links can further be broken down into templatized links designed to give web pages a common look and feel, and informational links that facilitate exploration of a body of knowledge.

We can categorize links to be one of several types based on the relative placement of the source and destination within the hierarchies, and divide them into six distinct types: 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 second column of Table 1 shows the distribution of links into the various types, based on a sample of links from our entire corpus. This data clearly shows that external links are relatively rare, 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 discrepancy between the number of down and up links is perhaps surprising at first, but reflects several factors. First, many sites have every page equipped with a link to the top of the site (i.e., the starting point of the site), but downward links often target a single "entry page" in a directory [11]. Second, resource limitations on crawling and author-imposed restrictions on crawling via a `robots.txt` file will result in some down links being discovered without crawling the lower level pages to discover up links.

Another point one may consider when examining the distribution of links of the 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 92% 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.

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

Type of link	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 by type for a sample of links. Shown are a sample of links where both source and destination are static URLs, and the subset where both ends were crawled. In the final column we tabulate the number of bidirectional links. Self loops (which were not included in the sample) account for roughly 0.9% of the links.

Type of link	Crawled links	Random links
Internal	48.6%	32%
Up	13.6%	6%
Down	8.6%	5%
Across	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. Random assignment produces a distinctly different distribution of link types.

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). For external links we add 1 for a change of hostname.

We hypothesized that links tend to span a short distance in this measure, and in order to test this we calculated the distances for a sample of links for which both the source and destination URL do not contain a `?` character. Figure 3 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. In each case it appears that the probability of a link covering a distance d appears to decrease exponentially in d, in spite of the fact that the number of eligible targets for a link initially *increases* with the distance.

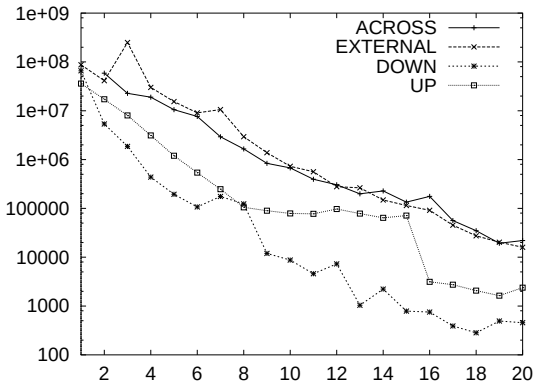

Figure 3: Distribution of tree distance for hyperlinks. As distance increases, the probability of a hyperlink decreases.

6. 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 may explain why many bidirectional links appear between pages that are authored by the same person at the same time.

For these experiments, we used roughly the first 600 million pages from the crawl. 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.

7. HIERARCHY IN MODELS FOR THE WEB

We believe that the approach to modeling the web should incorporate the social process of authorship, and the nature of social relationships within increasingly larger groups. Consider the social process by which a web site of a large company or university is built. At the lowest level we start with an individual who authors a few pages such as a personal page or a news release. 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 grouped 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.

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 [29]. In addition to this social hierarchy, web information often has a topical hierarchy associated with it that is often recognizable from the URL hierarchy. For example, and individual author will typically organize their files into directories, grouping them by topic. Users are often given access to a directory on a server, using file permissions to grant access to everything under that node of the tree. This combination of social hierarchy and filesystem hierarchy encapsulates a great deal of structure that we shall incorporate into our model of the World Wide Web.

7.1 Requirements for a Model

Efforts to model complex systems usually have to make tradeoffs between accuracy and simplicity. Simple models generally lend themselves to direct mathematical analysis and extrapolation, but they often fail to adequately describe the many features that are present in a complex system. Complex models tend to defy direct analysis, but are better able to describe the system. For complex models, simulation is a viable alternative to direct analysis.

In seeking to model the Web, we consider the following axioms to be important, though the list is not exhaustive. First, it should be evolutionary, ideally including both birth and death processes. Next, the model should reflect the social and authorship processes that influence the World Wide Web. Third, the tail of the indegree distributions should exhibit a power-law. Outdegree distributions should also display a power-law tail, though they arise from a different process. Fourth, the model should reflect any inherent hierarchical organization of the system. Fifth, the model should exhibit a degree of locality in the link structure. Other desirable features include the existence of small communities of thematically related pages [16], and the probability of a link being bidirectional being strongly correlated to the locality of links.

7.2 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 forest directory structure and the hyperlinks. Sites themselves grow in a hierarchical fashion, with a site starting as a single URL, and growing into a tree. There are many variations on the procedure that we describe, and we defer discussion of these until after we describe the basic model.

We first describe how the forest structure is built. 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 site), containing a single URL. With probability η we create a new directory on an existing site to put the URL into. With probability γ we pick an existing leaf directory (a directory that has no sub-directories) and add the new URL to it. Finally, with probability $1-\gamma-\epsilon-\eta$, we pick an existing non-leaf directory and add the new URL to it. In the case where a new directory is to be created, we pick the parent directory uniformly at random with probability c_f, and with probability $1-c_f$, in proportion to the current fanout of the directory. When adding a URL to an existing directory, we pick a directory uniformly at random with probability c_s, and with probability $1-c_s$ with proportion to the number of URLs in the directory.

We now describe how the links are created. At the time that we create a URL, we create a single inlink to the newly created page (this makes the resulting graph connected). If the URL is created on a new site, the source for the inlink is chosen uniformly at random from all URLs in the graph. If it is created in an existing site, we pick a URL uniformly at random from the set of URLs in the directory where the new URL is attached and the directory immediately above it.

We now have to say how to create links from each newly created 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 from a page, we first decide the number of links in accordance with a hypothesized distribution on the outdegrees from pages. We expect that this distribution has a power law tail, but the small values are unimodal and dictated by the common conventions on page design (in our simulation we use the observed outdegree distribution averaged over all sites). 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 mix of uniform and preferential attachment, namely with probably δ we choose one uniformly at random, and with probably $1-\delta$ we pick one with probability that is proportional to its existing indegree. If there are no eligible URLs to create a link to, then we simply omit the link (for example, in the case of attempting to create a down link from a URL at a leaf directory). If we create a link, then we create a backlink from that link with probability b_t.

The mix of uniform and preferential attachment for inlinks is designed to guarantee the power law distribution for indegree. 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.

8. LINK COMPRESSION AND ENTROPY

It has been observed by several authors that the link graph is highly compressible [5, 27, 26]. Randall et al. [27] report that it takes only 6 bits on average to store the outlinks from a set of 350 million pages (6 billion links), and more recently Boldi and Vigna [5] have found encodings that use only 3 bits per link. If the links were *random* then of course this would not be possible, as an easy probabilistic argument says that at least 28 bits would be required to store a single link from each page, and this number would grow as $\log(N)$ for a graph with N nodes. One possible source of redundancy in the link structure may be attributed to the power law distribution of indegrees. However, it was observed by Adler and Mitzenmacher [2] that a simple Huffman encoding scheme that exploits only this redundancy for compression of the web graph would still require $\Omega(\log(N))$ bits to represent an edge in an N-node link graph. This suggests that there are other sources of redundancy in the link graph that allows for such high levels of compression.

In fact the hierarchical locality for links that we have observed is closely related to why such good compression schemes for the web graph are achievable. The primary method used in [27] 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, and are therefore close together in a lexicographic sort order. Since lexicographic order of URLs is a good approximation of directory order, the compressibility of the link graph is closely related to the locality of links in the hierarchical structure. This observation that locality is the source of compressibility of the web graph was also made in [2].

To further explore the link between hierarchical structure and compression, we wish to examine how well various web models explain link compressibility. Rather than considering the various compression schemes devised (which are mostly based on the textual structure of the URLs, and are designed to facilitate efficient implementation), we concentrate on the information-theoretical measure of the *entropy* of the link graph. We choose to focus on the following entropy measure, which we call *isolated destination entropy*. We define the probability distribution whose entropy we measure as follows: First, the evolutionary model is used to grow a graph to a given number of nodes N. Then, we consider the distribution of the destination URLs that are linked to from each URL (where the distribution is over the set of all nodes in the graph). In other words, for each source URL v, we consider the distribution of the random variable D_v whose values are the destinations of links originating at v.

Our motivation in picking the entropy measurement to be based on a static snapshot of the graph, rather than considering the entropy of the selection process employed by the various evolutionary model, is to mimic the conditions faced by a compression algorithm for a web. When compressing the web, a compression algorithm typically has no knowledge of the order in which URLs and links were created. Furthermore, we would like a measure of compressibility that is independent of the evolutionary model used, to allow for an apples-to-apples comparison of the various models. The isolated destination entropy is a lower bound on the compression that may be achieved by compression schemes that encode each destination independently of other destinations. It obviously also dictates a lower bound for more sophisticated methods.

This measure captures the redundancy in information that is present because outlinks from a given page are typically made to pages in the close proximity to the source page. However, this does not capture the more global phenomenon that makes pages that are close to each other in the hierarchy have links similar to each other. In fact, it does not even directly exploit the dependency between different pages linked to from the same page. The effects of this phenomenon are part of the explanation for the improvements over the Link2 scheme in [27] achieved by schemes, such as the Link3 scheme in [27], that use delta encodings between outlink lists of multiple pages.

Unfortunately, because of the complexity of the models and the fact that we are measuring entropy on snapshots of the graph, we are unable to analytically compute the isolated destination entropy. Instead, we provide empirical measurements for various models. To measure the isolated destination entropy we use each model to generate 225 random graphs, each containing a million nodes. Where applicable, we use the same arbitrary set of URLs (and hierarchical structure) for all graphs generated by a model, and only allow the link generation process to be driven by a pseudo-random number generator. We then sample a fraction of the nodes in all the graphs, and empirically estimate their average isolated destination entropy by calculating the entropy of the empirical distribution of outlinks from a node. We express our entropy measurement in bits per link, as is customary in works that describe compression schemes for the web graph [27, 2]. When comparing these results to the theoretical maximum entropy, one must note that because of the relatively small sample that we use relative to the domain of the random variable D_v, the upper bound on the entropy is much lower than the usual $\log N$ for a graph with N nodes. Instead, if the average outdegree is d, and we generate m distinct graph, $\log(md)$ is an upper bound on the empirical isolated destination entropy we can expect This is because, on average, only md outlinks from any given node will be encountered in the sample.

The models we compare are the following:

PAModel A preferential attachment model based on [19, 25]. In this model, the destination for outlinks is chosen by a mixture of preferential attachment and a uniform distribution over all previously cerated URLs.

AMModel A copying model, similar to that of Kumar et al. [17]. Following Adler and Mitzenmacher's choice of parameters for this model labeled G_4 [2], we set the parameters for this model to copy links from zero to four previous nodes, where each link is copied with probability 0.5, and either one or two additional links are then added with the destination chosen uniformly at random.

Hierarchical Model Our hierarchical web model as described in Section 7.2.

Model	Empirical Entropy	Max. Entropy
PAModel	11.72	12.56
AMModel	9.6	12.05
Hierarchical	8.08	12.49

Table 3: Empirical measurements of isolated destination entropy on graphs generated by three models. Measurements are based on sampling from 225 graphs for each model, of size one million nodes each.

Both PAModel and the hierarchical model require outdegrees to be drawn from a power law distribution. Rather than using a pure power law distribution, we use a sample of outdegrees from our web crawl to determine the "head" of the outdegree distribution in these models, with the tail being determined by a power law. This distribution has a mean of about 27 outlinks. The Hierarchical model exhibits a slightly lower average outdegree in practice, because some outlinks may not be feasible (e.g., uplinks from top level URLs, etc.). The results of our experiments are summarized in Table 3. The results clearly point out that the

destinations for outlinks in our hierarchical web model are far less random than those generated by the previous models we compare against. The results also demonstrate that graphs generated by an evolutionary copying model tend to have a less random structure than graphs where link destinations are chosen through a preferential attachment process. This suggests that incorporating copying into the hierarchical model may reduce the uncertainty in link creation even further, and yield an even more realistic model, as far as the measure of compressibility of the link graph is concerned.

9. CORPORATE INSTANT MESSAGING

In this section we describe a totally different example of the relationship between hierarchical structure and link structure. We note that corporations tend to be organized hierarchically, and we might therefore suspect that Simon's hypothesis on communication following a hierarchy to be exhibited within a corporation. In this section we examine the communication patterns exhibited in text instant messaging within a corporation.

Text instant messaging has gained wide acceptance inside IBM for global communication among the approximately 325,000 employees around the world. By arrangement with the corporate IT department, we were able to obtain[2] a log of the email addresses for 175,288 individual sessions (involving 75,587 different individuals). For each session we used the online database of the management structure to retrieve the management chain of the two people involved in the session. Simon's hypothesis suggests that most of the conversation will take place between people that are close to each other in the management hierarchy.

One fact that may seem surprising is that the IBM management structure is a forest and not a tree. In particular, the general managers of the divisions in all the different countries do not report directly to the CEO. Thus the management structure resembles the hierarchical structure of the web in the sense that it can be traced to a few top level nodes.

Table 4 shows the breakdown of sessions according to the relationship between the two parties in the management chain. The most striking thing about this breakdown is the fact that they bear some resemblance to the statistics of hyperlinks given in Table 1. In particular, the largest fraction of communications takes place between people in the same department, and a much smaller percentage takes place between people who are in distinct parts of the IBM management forest.

One can form a single tree from the IBM management structure by placing a virtual node at the top of the company, and linking all of the individuals who report to themselves up to this virtual node. In this tree we computed the tree traversal distance between all pairs of people that communicated using instant messaging, and the results are shown in Figure 4. From this plot it is evident that the further people are from each other in the management tree, the less likely they are to communicate with each other. The situation is however slightly more complicated than this, and the hierarchical structure influences the patterns of social interactions in other ways. For example, people seem somewhat more likely to talk to their peers, i.e., those at their same level in the management tree. This explains in part why distance 3 is somewhat less likely than distance 4, because odd distances correspond to communicating with the manager of a nearby peer.

Relationship	percentage
same department	34.5%
across departments	43.7%
up to manager	7.0%
down to employee	5.4%
outside organization	8.9%

Table 4: Management relationships between parties engaged in instant message communications inside IBM

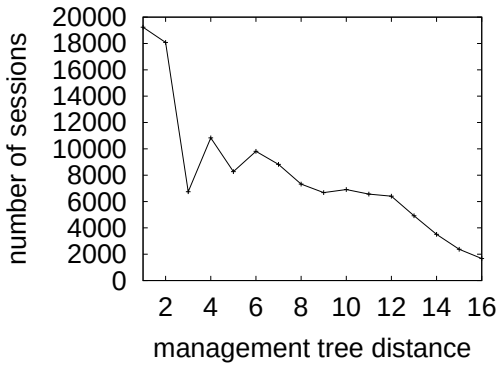

Figure 4: Distribution of tree distance in the virtual IBM management tree for users of instant messenger communication. As distance in the management chain increases, the probability of communication declines. There is also slight tendency for people to be more likely to communicate with each other if they are at the same depth of the management tree, which is why the odd distances are slightly less likely than the next larger distance.

The data clearly shows a strong tendency for locality in communication within a corporation, and provides a further domain in which to apply hierarchical models of random graphs.

10. 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 accurately model locality and bidirectionality properties of hyperlinks without accounting for the underlying growth process of the hierarchical structure.

We have proposed a framework for models that incorporates an evolutionary process for both the hierarchical

[2]We thank Savitha Srinivasan for obtaining access to this data.

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 [9] and similarity [12], 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.

We have demonstrated that at least one feature of the actual web graph, namely the compressibility of the link structure, is directly related to the hierarchical structure. We believe that many other features of the web graph may be more accurately explained once the hierarchical structure of web sites is incorporated into the model.

11. 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] M. Adler and M. Mitzenmacher. Towards compressing web graphs. Technical report, Harvard University Computer Science Dept., 2001. Short version in Data Compression Conference, 2001.

[3] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The connectivity sonar: detecting site functionality by structural patterns. In *ACM Hypertext '03*, pages 38–47, Nottingham, UK, 2003.

[4] K. T. Balińska, L. V. Quintas, and J. Szymański. Random recursive forests. *Random Structures and Algorithms*, 5(1):3–12, 1994.

[5] P. Boldi and S. Vigna. The webgraph framework I: Compression techniques. In *Proc. Int. WWW Conf.*, New York, 2004.

[6] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-MAT: A recursive model for graph mining. In *Proc. SIAM Int. Conf. on Data Mining*, 2004.

[7] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proc. ACM SIGMOD*, pages 307–318, 1998.

[8] M. E. Crovella and M. S. Taqqu. Estimating the heavy tail index from scaling properties. *Methodology and Computing in Applied Probability*, 1(1), 1999.

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

[10] 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.

[11] N. Eiron and K. S. McCurley. Untangling compound documents in the web, 2003. Proc. ACM Conf. on Hypertext and Hypermedia.

[12] 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.

[13] L. A. Goodman and W. H. Kruskal. Measures of association for cross classifications. *J. of the American Statistical Assoc.*, pages 723–763, 1954.

[14] B. A. Huberman and L. A. Adamic. Evolutionary dynamics of the world wide web. Technical report, XEROX PARC, 1999.

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

[16] 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.

[17] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the Web. In M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie, editors, *Proc. 25th VLDB*, pages 639–650, Edinburgh, Scotland, 1999. Morgan Kaufmann.

[18] 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.

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

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

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

[22] M. Mitzenmacher. Dynamic models for file sizes and double pareto distributions. *Internet Mathematics*, 2004.

[23] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. Paper SIDL-WP-1999-0120 (version of 11/11/1999).

[25] 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.

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

[27] 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.

[28] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67(026112), 2003.

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

[30] 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.