GRAPH-THEORETIC ANALYSIS OF THE WORLD WIDE WEB: NEW DIRECTIONS AND CHALLENGES

Narsingh Deo **Pankaj Gupta**

Abstract

The World Wide Web is growing rapidly and revolutionizing the means of information access. It can be modeled as a directed graph in which a node represents a Web page and an edge represents a hyperlink. Currently, the number of nodes in this gigantic Web graph is estimated to be over ten billion, and is growing at more than seven million nodes a day—without any centralized control. Recent studies suggest that despite its chaotic appearance, the Web is a highly structured digraph, in a statistical sense. The study of this graph can provide insight into Web algorithms for crawling, searching, and ranking Web resources. Knowledge of the graph-theoretic structure of the Web graph can be exploited for attaining efficiency and comprehensiveness in Web navigation as well as enhancing Web tools, *e.g.*, better search engines and intelligent agents. In this paper, we discuss various problems to be explored for understanding the structure of the WWW. Many research directions are identified such as structural analysis of the Web, design of search engines, network security, *etc.*

1 Introduction

The World Wide Web (WWW or Web) has revolutionized the way we access information. The Web is currently estimated to have over ten billion pages and 70 billion hyperlinks [27]. It is growing at the rate of 7.3 million pages a day [19]. The Web can be viewed in two very divergent ways. From an internal standpoint, it is a distributed TCP-compliant application. We are interested in the other equally-important way of viewing the WWW, which is external and extensional. From this standpoint, the WWW is a vast and continuously growing repository of information: textual as well as audio and video. The

Keywords — Web graph, graph theory, random graph, Web model.

sheer mass of material and the apparently chaotic nature of the WWW can make locating, acquiring, and organizing information both time consuming and difficult. This runs exactly opposite to the promise of the WWW, namely that all information can be made readily available to the world's population all of the time. The first challenge to overcome in attaining this laudable goal would be to find methods for efficiently and comprehensively navigating the WWW. In WWW-intrinsic terms, efficiency means that one wants to minimize the number of links that must be followed to reach a desired piece of information from a designated starting point. Comprehensiveness means that the most relevant information is actually obtained during the navigation. At present, neither efficiency nor comprehensiveness can be reliably achieved.

Despite its chaotic appearance, the WWW is highly structured, but in a statistical sense. Models have been proposed which reproduce certain experimentally determined features of the WWW and these features could be exploited to attain efficiency and comprehensiveness in the WWW navigation. In this paper, we present new directions and challenges in the graph-theoretic analysis of the Web.

The paper is organized as follows: Section 2 surveys the random-graph models that explain the Web structure. Section 3 describes algorithms for efficient search and identification of Web communities. Section 4 identifies new directions and challenges in the graph-theoretic Web research. Section 5 presents the concluding remarks. The description of the graph-theoretic terms used in this paper can be found in standard graph-theory books such as ([8], [11]).

2 Random-Graph Models of the Web

The Internet is a constellation of data-communications technologies. Internet topology and traffic are defined by data-communications lines and resource allocation issues in entities like routers or operating systems. However, the Web topology is an aggregate sum of individual decisions and is not influenced by data communications or systems considerations. Web links reflect semantically motivated, intentional acts by human beings. The Web is different from a distributed database in the sense that there is no uniform structure, integrity constraints, transactions, standard query language, or a data model. The uncontrolled, decentralized, and rapid growth of the Web is absent in a distributed database. Moreover, the key features of a database such as reliability, recovery, *etc.*, are not present in the Web. Recent studies show that the Web structure resembles collective behavior reminiscent of complex physical systems and thermodynamics. The Web can be modeled as a directed graph, where each node represents a page and each edge, a hyperlink.

Empirical study is a useful technique for understanding the Web structure. Pirolli *et al.* [32] performed the earliest study of link structure of the Web in April, 1996. They studied three kinds of graphs to represent the strength of association among Web pages: hypertext-link topology, inter-page text similarity, and flows of users through a locality. Kumar *et al.* [22] showed that despite the chaotic nature of content creation on the Web, there exist well-defined communities. In August, 1999 Barabási and Albert [3] analyzed the induced graph of Notre Dame University (*www.nd.edu*) with 325,729 nodes and extrapolated the expected distance between any two nodes in the entire Web graph (with 800 million nodes at that time) to be about 19. They also found that the inverse power-law degree holds for the distribution of a Web page. Kleinberg *et al.* [19] report the distribution of bipartite cores on the Web from the result of *Alexa* crawl. In one of the most extensive empirical studies (conducted in May, 2000), Broder *et al.* [7] analyzed the link structure of 203 million pages and 1.5 billion links. Their study showed that the Web graph has four distinct regions. There exists a giant strongly connected component (SCC). There is a set of newly formed nodes called IN having only outgoing links and another set of nodes called OUT having only incoming links (*e.g.*, some corporate and e-commerce sites). There exists a directed path from each node in IN to SCC, and a directed path from SCC to each node in OUT. There exist directed paths from nodes in IN to a set of nodes, constituting the fourth region, called TENDRILS. There are also paths from TENDRILS to the OUT region. Molloy and Reed ([25], [26]) investigated the emergence of a giant component in a random graph with a given degree sequence as well as the structure of the graph after deleting the giant component. In September, 2000 Bar-Yossef *et al.* [2] performed random walks on the Web for discovering its properties. They built an application, named *WebWalker*, that performed random walks on a d-regular, undirected graph. *WebWalker* used resources such as HTML text analysis, search engines, and the random walk itself.

In this section, we present several models of random graph that have been proposed to describe the structure of the Web.

2.1 Erdős-Rényi Model

The earliest model of a random graph was proposed by Erdős and Rényi [13] (example in [12], p. 8) in which we start with n isolated nodes, and then each of the $\frac{n(n-1)}{2}$ pairs of nodes is connected by an edge with a specified uniform probability p. This static model does not represent the real-life WWW because it does not take into account: i) The increasing number of nodes in the Web, and ii) The non-uniform probability of edge formation between node pairs.

2.2 Small-World Models

In 1967, Milgram ([33], p. 23) articulated the basic properties of small-world networks, based on the well-founded folklore that each individual is indirectly linked through a short chain of acquaintances to practically anyone else in the world. Sparseness, small diameter, and cliquishness are the three properties that characterize small-world networks. Erdős-Rényi type random graphs have a small diameter, but they lack the cliquishness present in the small-world networks. The small-world effect has been identified in disparate contexts including neural network of the worm *C. elegans*, epidemiology, power-grid networks, collaboration graph of film actors, and the WWW [34]. Two variations of small-world model have been proposed. They are the edge-reassigning and edge-addition small-world network.

Edge-Reassigning Small-World Network: Watts and Strogatz [33] proposed the edge-reassigning small-world model. In this model, evolution starts with a ring lattice having each node connected to its d nearest neighbors. Each of the $\frac{n \cdot d}{2}$ edges is randomly removed and reassigned to distant nodes with a probability σ in a round-robin fashion (example in [12], p. 10). This model has two properties: characteristic-path length that measures the separation between two nodes (global property), and clustering coefficient that measures the cliquishness of neighborhood of a node (local property). The chief feature of this model is that a region of values for σ produces evolution, unlike the Erdős-Rényi type random-graph evolution. When $\sigma = 0$, the original regular-graph remains unchanged. As σ increases from 0 to 1, the characteristic path-length decreases rapidly. As edges are reassigned from a clustered neighborhood, a poorly clustered, small-world random network is formed. When $\sigma = 1$, we get an Erdős-Rényi type random graph.

Edge-Addition Small-World Network: The edge-addition small-world model was proposed by Newman *et al.* [30]. In this model, $\frac{p \cdot d \cdot n}{2}$ new edges are added randomly to an existing ring lattice (example in [12], p. 11). Here, p, d, and n denote the new-edge probability, degree of each node in the original ring-lattice, and the number of nodes, respectively.

Newman *et al.* ([28], [30]) considered the following situation in a small-world network. Some fraction f of nodes is populated by individuals who will contract a disease if exposed to it. The probability $P(j)$ that a randomly chosen node i belongs to a connected cluster of j nodes is determined. If any node i contains an infected individual, $P(j)$ is the probability that j people will be consequently infected because there are very short paths from the original infected individual to all others in the cluster. Newman *et al.* derive the threshold value of f at which the expected size j of infectious outbreak diverges.

The spread of epidemics in a population bears an interesting analogy to the spread of viruses on the WWW. The small-world characteristic of the Web aids in the spread of a virus epidemic. A file meant for general distribution on a popular site, if infected by a virus, can become the source of a major outbreak. The Web pages of a popular Web site are vulnerable, because these pages are part of a cluster. Once infected, they propagate the virus through their immediate neighbors and cause a global epidemic thereby causing enormous loss of time and money. Thus, the study of small-world networks is useful in predicting the nature of a virus outbreak and its effect on the Web.

2.3 The Preferential-Attachment Model

The small-world models cannot be applied directly to the WWW because the number of pages in the Web is not fixed. These models do not accommodate a birth/death process in which new pages are created, how new links are formed (possibly through editing old pages), and how both links and pages can be deleted. Two new models have been proposed recently to explain some of the empirical findings concerning the overall WWW structure, taking on board the reality of the birth/death process.

The preferential-attachment model ([4], [5]) starts with a small, null graph having a finite number of nodes (n_0). At each successive time step, a new node with a bounded number of outgoing edges is added to the network. The probability that a link from the new node goes to an existing node i is $\frac{d_i}{\sum_j dj}$, where the denominator sum runs over all existing nodes (example in [12], p. 14). Thus, a newly-introduced node is more likely to be adjacent to a node with high degree.

This model reproduces one of the most significant empirical findings about the WWW, namely, the probability that a page or a node i has degree d_i is $\frac{A}{(d_i)^c}$, where A is proportional to the square of average degree of the network and c is a constant. The exponent c was empirically found to be about 2.9, and it is independent of the number of edges being added at each time step.

The preferential-attachment model does not allow for the reconnection of existing edges. Also, addition of new edges takes place only when new nodes are added in the system. However, in real-life, new links are added continuously between old nodes as well.

2.4 The Web-Site Growth Model

Huberman and Adamic [16] have proposed another model that exhibits an inverse, scale-free power law for the probability that a Web site s has N_s pages. The term Web site in their study is defined as a registered domain-name on the Internet. In the Web-site growth model, the number of pages added to a site s at any given time is considered proportional to those already existing on the site. The equation has the form

$$N_s(t+1) = N_s(t) + g(t+1) \cdot \quad N_s(t),$$

where, $N_s(t)$ = the number of pages at site s at time step t, and

$g(t)$ = universal growth rate, which is independent of a site.

Due to the unpredictable nature of growth of a site, $g(t)$ fluctuates about a positive mean g_0, and it can be expressed as

$$g(t) = g_0 + \xi(t),$$

where, g_0 = the basic, constant growth rate, and $\xi(t)$ = a Brownian motion variable.

The expected value of $N_s(t)$ is

$$N_s(t) = N_s(0) \cdot e^{(g_0 \cdot t \ + \ v_t)},$$

where, υ_t is a Wiener process such that $\upsilon_t^2 = e^{(var(g)\ \cdot\ t)}$ and $var(g)$ is variance of growth rate $g(t)$ of the Web site.

The probability of $N_s(t)$ pages at a site s is given by a weighted integral, which eliminates dependence on time t and yields a scale-free inverse power-law $\frac{c}{(N_s)^\gamma}$, where c is a constant and exponent γ is in the range $[1, \infty]$.

In a short note, Adamic and Huberman [1] criticized the preferential-attachment model for its prediction that older pages have larger in-degree than newer pages. Empirical data appears to show no correlation between the age and in-degree of a Web page. They argue that the growth rate of degree of a page depends on the current degree of page and not on its age.

3 Search and Page-Evaluation Algorithms

The enormous size and rapid growth of the Web makes it difficult for an individual to locate information and navigate by just employing Web addresses. A search engine is useful for locating information in the vast space of the Web. Ranking the pages returned by a search engine is done by ranking-function heuristics based on the frequency of occurrence of keywords, and sometimes on the position of keywords in the page [24]. However, such strategies may not deliver correct information. For example, some Web pages may have a keyword repeated many times to attract Web traffic or gain favorable ranking.

One of the factors in the effectiveness of a text-based search engine is the number of Web pages indexed in its database. As of January 2003, *Google* had the largest database with 3,083 million pages indexed. *Fast* had 2,112 million pages, *Altavista* had 550 million pages indexed, and *Inktomi* had 500 million pages (*http://searchenginewatch.com/reports/sizes.html*). The search engine with the largest index (*Google*) covers less than one-fourth of the present Web pages and this disparity is going to increase in future. Vast inconsistency in the number of hits for two similar queries on a search engine is another serious problem, *e.g.*, a query for “Oscar award” on *Altavista* search engine resulted in 1,480 hits while “award Oscar” yielded 1,208,710 hits. Additional examples of inefficient search results appear in ([12], [15]).

A level of abstraction higher than that of a Web page is helpful in understanding the Web topology. In this section, we explore how this abstraction can be used as a tool to identify order and hierarchy in the Web. This understanding is crucial for developing improved search algorithms and for enhancing the search-engine technology.

3.1 The WWW Communities

A WWW site has conventionally meant a collection of pages defined by design. In some cases, a site is coherent with respect to some semantical interpretation, *i.e.*, the pages focus on different aspects of one topic, but more often the home page is just the hub of the collection. A WWW site can be also defined through link counting. Here, we consider some notion of locality, *e.g.*, some subnet IP address range and then measure the ratio of the links among all pages inside the range to all the links going outside the range. Gibson *et al.* [14] identified two kinds of pages that together make possible a computational concept of a WWW community of pages. One kind of pages represents authorities focusing on a specified topic, while the other kind represents hubs, which point to many authorities on that topic. The abstract community of hubs and authorities arises due to the abundance of Web pages that are usually returned for any broad-based query. The goal of a search method is to return a smaller set of "authoritative" pages for the query. The hyperlinks can be exploited for understanding the inherent Web communities. A link from a page u to a page v can be viewed as conferral of authority on the page v. However, there are many links created which have no meaning with respect to conferral of authority. Hence, just counting the in-degree of any Web page is not the complete solution to the problem of identifying the authority.

3.2 HITS Algorithm

Kleinberg [18] proposed an iterative algorithm, called HITS (Hyperlink-Induced Topic Search), for identifying authorities and hubs using the adjacency matrix of a subgraph of the WWW. For a broad-topic search, the algorithm starts with a root set S of pages returned by a text-based search engine. The set S induces a small subgraph focused on the query topic. This induced subgraph is then expanded to include all the nodes that are successors of each node in set S. In addition, a fixed number of predecessors of each node in the set S are also included. Let G be the graph induced by the nodes in this expanded node set. It should be noted that the links that are used purely for navigation within a Web site are not included in the graph G. For each node, x, in the graph G a non-negative authority weight $a(x)$ and a non-negative hub weight $h(x)$ are computed. The authority and hub weights of all the nodes may be expressed as vectors $a()$ and $h()$, respectively. The elements of vectors $a()$ and $h()$ are initialized to one. In each iteration, $a(x)$ is replaced by the sum of $h(x_i)$'s of all the predecessors of node x, and $h(x)$ is replaced by the sum of the $a(x_j)$'s of all the successors of node x. The iterations may be expressed as

$$a(x) = \sum_{v \to x} h(v), \text{ and } h(x) = \sum_{x \to w} a(w).$$

The authority and hub scores are normalized in each iteration so that $\sum (a(x))^2 = 1$, and $\sum (h(x))^2 = 1$. This iterative process converges to yield the authority and hub vector for the initial query. If M is the adjacency matrix of the graph G, then the iterations of vector a when normalized, converge to the principal eigenvector of $M^T M$. Similarly, multiple iterations of the normalized vector h converge to the principal eigenvector of MM^T. Thus, HITS applies a link-based computation for identifying the hubs and authorities on a query topic.

3.3 Web-Page Evaluation

A common problem with search engines that evaluate Web page relevance based on the frequency of keywords, is that they can be easily biased by deliberate inflation of keywords in the contents of a Web page. Another problem is that many Web pages do not contain the keywords that best describe what the Web page is known for or the services it provides. Evaluating the importance of a Web page, using the graph structure of the Web, can solve both these problems. We now present a graph-theoretic algorithm for measuring the importance of a Web page.

PageRank Algorithm

Conventional search-engines have relied on matching keywords and strings in the Web pages to index and search the Web for information. *Google* uses the graph structure of the Web to produce better search results. It uses PageRank algorithm [31] that attempts to give a ranking to a Web page, regardless of its content, based solely on its location in the Web graph.

Here,
i = a node (Web page) in the Web graph,
d_i^+ = out-degree of node i,

$w_1, w_2, \ldots, w_k$ = predecessors of node i,
η = normalization constant ($\eta < 1$), and
$PR(u)$ = PageRank of a Web page u.

The PageRank of a page i is given as

$$PR(i) = (1 - \eta) + \eta \cdot (\frac{PR(w_1)}{d_1{}^+} + \frac{PR(w_2)}{d_2{}^+} + \cdots + \frac{PR(w_k)}{d_k{}^+}).$$

The PageRank algorithm starts with assigning a rank of one to all pages and recursively computes the PageRank value for each page. The rank of a page is divided equally among its outgoing links. A page has high PageRank if pages having high PageRank point to it. The PageRank vector $PR()$ corresponds to the principal eigenvector of normalized adjacency-matrix of the Web graph. Normalization constant, η, is the probability that a random surfer does not follow an outgoing link of a page i, and selects another page randomly.

3.4 Small-World Algorithmics

Kleinberg ([20], [21])proposed an algorithm for finding the shortest or near-shortest path from one node to another node in a graph with small-expected diameter. The algorithm considers a two-dimensional grid with directed edges. The probability of an edge between nodes u and v is proportional to $[L(u, v)]^{-r}$, $(r \geq 0)$, where $L(u, v)$ is the distance between nodes u and v. Using an extension of the Watts-Strogatz model, Kleinberg devised a decentralized algorithm that finds a near-shortest path in expected time which is polylogarithmic in the number of nodes. The algorithm considers the problem of passing a message from a node u to another node v. In each step, an intermediate node i chooses a contact node that is as close to the target node v as possible. It assumes that every node knows the location of the target node in the network as well as the location and long-range contact of all nodes that have met the message. Kleinberg proved that at $r = 2$, the decentralized algorithm takes best advantage of the geographic structure of the network and generates paths of length O(log n), where n is the number of nodes in the grid.

3.5 Related-URL and Topic-Distillation Algorithms

The graph topology of the Web can be exploited for discovering novel search-techniques. Dean and Henzinger [10] proposed a search algorithm for finding

Web pages related to a URL. A related Web page is one that addresses the same topic as the original page, but is semantically different. The Dean-Henzinger algorithm consists of the following steps:

1. Build a vicinity graph for a given URL, *i.e.*, node U.

The vicinity graph is an induced, edge-weighted digraph that includes the URL node U, up to B randomly selected predecessor nodes of U, and for each predecessor node up to B_F successor nodes different from U. In addition, the graph includes F successor nodes of U, and for each successor node up to F_B of its predecessor nodes different from the node U. There is an edge in the vicinity graph if a hyperlink exists from a node v to node w, provided both the nodes v and w do not belong to the same Web site.

2. Eliminate duplicate and near-duplicate nodes.

Duplicate nodes are mirror sites or different aliases of the same Web page. Two nodes are defined as near-duplicate nodes if they have more than 95% of the links in common and each node has more than 10 links. The near-duplicate nodes are replaced by a node with links that are union of links of all the near-duplicate nodes.

3. Compute edge weights based on connections between Web sites.

An edge between nodes on the same Web site is assigned a weight 0. If there are m_1 edges directed from a set of nodes on one Web site to a single node on another Web site, then each edge is given an authority weight $1/m_1$. If there are m_2 edges directed from a single node on one Web site to a set of nodes on another Web site, each edge is assigned a hub weight $1/m_2$. This prevents the influence of a single Web site on the computation.

4. Compute a hub and an authority score for each node in the graph.

The ten top-ranked authority nodes are returned as the pages that are most related to the start page U (modified version of HITS algorithm, Section 3.2).

Bharat and Henzinger [6] proposed a search-engine enhancement algorithm, based on the Web topology, called *Topic distillation*. *Topic distillation* is defined as the process of finding quality Web pages related to a query topic. The topic-distillation algorithm solves three problems associated with the HITS algorithm. The first problem is the mutually reinforced relationship between Web sites where hub and authority scores of nodes on each Web site increase. The other two problems are automatically generated links and presence of non-relevant nodes. If there are m_1 edges directed from a set of nodes on one Web site to a single node on another Web site, then each edge is assigned an authority weight (*edge_auth_wt*) of $1/m_1$. Similarly, if there are m_2 edges directed from a single node on one Web site to a set of nodes on another Web site, then each edge is assigned a hub weight (*edge_hub_wt*) of $1/m_2$. In addition, isolated nodes are eliminated from the graph. The hub weight and authority weight of each node is calculated iteratively as:

$\forall u \in V$,

$a(u) = \sum_{(v,u)\in E} h(v) \times edge\_auth\_wt(v,\ u)$ and

$h(u) = \sum_{(u,v)\in E} a(v) \times edge\_hub\_wt(u,\ v)$.

The similarity between the query and the node returned by a search engine is defined as the relevance weight of the node. The relevance weight of each node is computed and the nodes whose relevance weights fall below a threshold level are eliminated.

4 New Directions and Challenges in Web Graph Research

In this section, we discuss some new directions for further research in the graph-theoretic analysis of the Web.

4.1 Structural Analysis of Web Connectivity

The Web now constitutes one of the largest directed graphs ever encountered in a real-life situation. Our challenge is to understand, model, and exploit this gigantic structure. One promising direction of research is to investigate the various meaningful and useful graph-theoretic properties of the Web. For example, locating various subgraphs on the Web such as bi-directional stars, cliques, trees, k-bipartite components, *etc.* Enumeration of these structures can help us to understand the communities existing on the Web as well as invent new search methods. Another approach is to identify frequently occuring subgraphs and understand their significance in the overall structure of the Web. From an algorithmic aspect, the problem involves finding a subgraph H in a large directed G. We may consider graphs H for which the solution is feasible (*i.e.*, in polynomial time). To identify these structures quickly in the dynamically changing structure of the Web graph is a challenging problem.

The random-walk method for sampling the Web graph is a promising tool for analyzing the Web structure. An accurate sampling of the Web can help us understand the dynamic properties of the Web. This can be accomplished by comparing samples of the Web at different intervals and analyzing changes in the static properties. This is particularly important for designing scalable search-techniques. In addition, the stability and relative fractions of each component of the Web is yet to be investigated.

Near-Clique Characterization: Closely related topics on the Web can be explored using the link structure of the Web. These topics tend to form a near-clique, *i.e.*, almost all the nodes (corresponding to the topics) are mutually adjacent to each other. Characterizations of such near-cliques can help us to analyze and extract such structures from the vast Web graph. There is a need to develop efficient heuristics for automatically locating such *near-cliques* structures on the Web. Identifying the near-cliques structures in the Web graph will make the ranking algorithm of a search engine more effective.

4.2 Accurate Web-Models

Most of the models that have been proposed for explaining the growth of the Web do not consider deletion of existing edges and nodes, nor realigning of existing edges. What is needed is a random-graph model that can fully represent the characteristics of the Web ([17],[23]). Such a model could be enhancement of an existing model that overcomes the specific drawbacks such as realigning of existing hyperlinks. The model also needs to be resilient and should be able to take into account dynamic changes in the properties as the Web grows.

An interesting model for the Web could be one that handles intra-site links differently from inter-site links. Another direction for further research will be in exploiting results of such studies to improve access to information on the Web. As described earlier, community topologies can be used to cluster and improve search results, and mirror listings can be used for reducing duplication in search results and for Web page request redirection. An interesting direction is exploiting the temporal information about Web pages. We can use timestamp information to validate some hypotheses about what roles different nodes play in the link structure of a community.

4.3 Design of Better Search Engines

The current search engines depend on the creation of a keyword index that is refreshed at a particular frequency. This means that the pages returned by a search engine query may not necessarily be available, as they may have been be deleted after their inclusion in the search-engine index. In addition, the index size cannot be scaled with the rapid increase in the size of the Web. Thus, there is a need to develop graph-theoretic search techniques that can search the Web in real-time. However, such methods may be computationally intensive. The contents of a Web page cannot be the only measure of its importance. An interesting research direction could be defining an authoritative set of metrics that represents the page quality.

Next Generation Searching: In the future, there will be an increasing demand for search engines that are increasingly specialized and subject-specific. For instance, search engines specializing in legal-case histories, academic literature, patents, recipes, *etc.* There is a need for search engines that exploit the organization and specialized semantics of this data [29]. What are the common principles that apply to such specialized search engines and the models of data they exploit? What sort of algorithmic toolkit is needed? How can the data be organized to facilitate such searching?

4.4 Security Issues

A "denial-of-service attack" on the Web occurs when a site (or a subnetwork) is flooded with packets whose sole purpose is to overload the local system, thus hampering or preventing access by legitimate users. An interesting question is how to exploit the graph-theoretic structure of the Web in conjunction with the knowledge of network contents to predict and prevent a denial-of-service attack. The key may be the development of intelligent router-protocols that can detect performance degradation. Can router protocols be developed to prevent or disarm such attacks before substantial performance degradation occurs? The problem is especially complicated because of the possibility of distributed attacks, where the malicious packets come from corrupted machines at many different locations.

4.5 Flow Analysis

Some edges in the Web graph are traversed more frequently than others. The data transfer due to the traversal of different edges varies very widely. Such patterns of edge traversal hold important information about the navigation preferences of the users. The structural information about the Web graph and the paths traversed by a specific user as he visits a sequence of pages can be combined to understand his navigational behavior. We need to develop tools to exploit the information gained from such traffic patterns.

4.6 Design of Adaptive Web-Sites

The hypertext structure of a Web site is relatively static, and is designed to accommodate all possible users. However, after observing usage patterns of the site, it may be desirable to restructure a site to be more effective. Furthermore, we may want to restructure the site for certain classes of users, or even for individuals. Thus, there is a need to develop methods for automatically restructuring Web sites for different users over time. How can we employ graph-theoretic techniques for designing an adaptive Web site? The challenge involves two parts: a) Detecting patterns of usage that suggest interesting restructuring, and b) Automatically restructuring the site in a consistent manner.

4.7 The Web of Queries

Various services keep users informed of new developments in areas of interest by tracking the changes in relevant Web pages. As such services grow, one can envision the growth of an augmented version of the graph associated with the Web. In this new graph, the nodes are not only Web pages, but also queries that depend on Web pages (and possibly on the results of other queries) [17]. There is an edge from a query to a Web page or another query, if the result of the first query depends on the contents of the Web page or the result of the second query. Should this "Web of Queries" develop, we will need to confront such questions and issues as detecting changes in queries and answers, as well as implicit dependencies among them. Efficient distributed processing of large numbers of queries in response to changes in Web pages is also important. We need to maintain consistency between answers as the contents of the Web change dynamically. How can we extract useful information from the graph structure of the query Web?

4.8 Determining the Geography of the Web

Web addresses are typically of two forms. First, there is the URL (Universal Resource Locater) such as *www.ucf.edu* that provides a linguistically meaningful address. Corresponding to this URL is the 32-bit IP address. There is, however, a third address that is of interest: the network location of the Web site (not necessarily where it is geographically, but where it is in the graph whose nodes are computers and other computational devices, and whose links are communications lines). The actual geographic location corresponding to an IP address constitutes a fourth address. This network location is implicitly encoded in the IP address, which originally at least was supposed to present a hierarchical description of the networks or domains that contain the location. Today, however, a correct interpretation of the encoding may involve information that is distributed in routers all over the world and cannot be directly queried. Locations with almost identical IP addresses can be very far apart, due to changes in the network and other idiosyncrasies. There is much potential value in knowing precise information about the network and geographic location of an IP address ([9], [17]). In particular, content distribution systems, such as the one *Akamai* provides, would like to be able to direct a Web site visitor to a cache that is close to the user (in the network sense), so as to eliminate or reduce the delays in content delivery that are caused when the content has to traverse many network links. Knowing geographic locations can be valuable for marketing too. The problem is to design algorithms that efficiently construct and update the current physical map of the Internet (the graph structure in which nodes are IP addresses and edges are communications links), using the protocols and mechanisms currently available to help pin down geography.

4.9 Network Reliability Analysis

Measuring the reliability of a large, unstructured network can be difficult: one may not have full knowledge of the network topology, and detailed global measurements may be infeasible. A valuable approach to such a problem is to take measurements from selected locations within the network and then aggregate them to infer large-scale properties. One sees this notion applied in settings that range from Internet topology discovery tools to remote software agents that estimate the download times of popular Web pages. Many questions arise related to these techniques: How reliable are the results? How much does the choice of measurement locations affect the aggregate information one infers about the network? Suppose we want to detect events of the following form: an adversary destroys up to k nodes or edges, after which two subsets of the nodes, each at least a fraction of the network, are disconnected from one another ([9],[17]). Scale-free networks, describing a number of systems, such as the WWW, Internet, social networks or a cell, display an unexpected degree of robustness, the ability of their nodes to communicate being unaffected by even unrealistically high failure rates. However, this error tolerance comes at a high price: these networks are extremely vulnerable to attacks, *i.e.*, to the selection and removal of a few nodes that play the most important role in assuring the network's connectivity.

5 Conclusion

The topology of the World Wide Web exhibits the characteristics of a new type of random graph, which at present, is only dimly understood. In this paper, we have considered several models that help describe the growth of the Web, and we have pointed out some of the features of the Web graph. Recent studies have uncovered only a few fundamental properties of the Web graph. We believe there are still more subtle, but important graph-theoretic properties yet to be discovered about the Web.

Structural analysis of Web connectivity can help us understand its complex regions. For example, the near-clique structures represent a rich source for any topic search. Hence, heuristics for identifying such structures can lead to improvement in the existing search engines. This understanding coupled with the information about user traffic traversing the hyperlinks can be applied for network-reliability analysis and design of adaptive Web sites. This can also help us prevent security threats such as denial of services.

The rapid growth of the Web poses a challenge to the present search-engine technology. The solution for improving search quality involves more than just scaling the size of the search-engine index database. Graph-theoretic algorithms, that take into account the link structure of the Web graph, will lead to development of better search engines and smart agents for providing relevant information to the end user, with efficiency and comprehensiveness. New graph-theoretic research directions identified in this paper will lead to the development of tools and techniques for harnessing the vast potential of the Web.

References

[1] Adamic, L.; Huberman, B., *Technical comment to 'Emergence of scaling in random networks'*, Science, (2000), 287- 2115.

[2] Bar-Yossef, Z.; Berg, A.; Chien, S.; Fakcharoenphol, J.; Weitz, D., *Approximating aggregate queries about Web pages via random walks*, In: Proceedings of the 26 th International Conference on Very Large Databases, Cairo, Egypt, Sept. 10-14, (2000), 535-544.

[3] Barabási, A.-L.; Albert, R.; Jeong, H., *Scale-free characteristics of random networks: The topology of the World Wide Web*, Physica A, 281, (2000), 69-77.

[4] Barabási, A.-L.; Albert, R., *Emergence of scaling in random networks*, Science, 286, (1999), 509-512.

[5] Barabási, A.-L., *Linked: The New Science of Networks*, Perseus publication, Cambridge, MA, (2002).

[6] Bharat, K.; Henzinger, M., *Improved algorithms for topic distillation in a hyperlinked environment*, In: Proceedings of the 21 st ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, Aug. 24-28, (1998), 104-111.

[7] Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; Wiener, J., *Graph structure in the Web: Experiments and models*, Computer networks, 33(1), (2000), 309-320.

[8] Bollobás, B., *Random Graphs*, 2nd ed., Cambridge University Press, (2001).

[9] Buyukkokten, O.; Cho, J.; Garcia-Molina, H.; Gravano, L.; Shivakumar, N., *Exploiting geographical location information of Web pages*, Working Paper SIDL-WP-2000-0136, Stanford Digital Library Project, Stanford University, (2000).

[10] Dean, J.; Henzinger, M., *Finding related pages in the World Wide Web*, In: Proceedings of the 8^{th} International World Wide Web Conference, Toronto, Canada, May 11-14, (1999).

[11] Deo, N., *Graph Theory with Applications to Engineering and Computer Science*, Prentice-Hall, Englewood Cliffs, NJ, (1974).

[12] Deo, N.; Gupta, P., *World Wide Web: A graph-theoretic perspective*, Technical Report CS-TR-01-001, School of Computer Science, University of Central Florida, Orlando, FL, (2001).(http://www.cs.ucf.edu/∼pgupta/publication.html)

[13] Erdős, P; Rényi, A., *On the evolution of random graphs*, Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5 (1960), 17-61.

[14] Gibson, D.; Kleinberg, J.; Raghavan, P., *Inferring Web communities from link topology*, In: Proceedings of the 9^{th} ACM Conference on Hypertext and Hypermedia, Pittsburg, PA, June 20-24, (1998), 225-234.

[15] Greenlaw, R.; Hepp, E., *In-line/On-line: Fundamentals of the Internet and the World Wide Web*, McGraw-Hill, New York, NY, (1998), 177-181.

[16] Huberman, B.; Adamic, L., *Growth dynamics of the World-Wide Web*, Nature, 401, (1999), 131.

[17] Johnson, D., *Challenges for theoretical computer science*, White paper, AT&T Research, (2000). (http://www.research.att.com/∼dsj/nsflist.html)

[18] Kleinberg, J., *Authoritative sources in a hyperlinked environment*, Journal of the ACM, 46(5), (1999), 604-632.

[19] Kleinberg, J.; Kumar, R.; Raghavan, P.; Rajagopalan, S.; Tomkins, A., *The Web as a graph: Measurements, models and methods*, Lecture Notes in Computer Science, 1627, (1999), 1-17.

[20] Kleinberg, J., *The small-world phenomenon: An algorithmic perspective*, Technical Report 99-1776, Dept. of Computer Science, Cornell University, Ithaca, NY, (1999).

[21] Kleinberg, J., *Navigation in a small world*, Nature, 406, (2000), 845.

[22] Kumar, R.; Raghavan, P.; Rajagopalan, S.; Tomkins, S., *Trawling the Web for emerging cyber-communities*, In: Proceedings of the 8^{th} International World Wide Web Conference, Toronto, Canada, May 11-14, (1999).

[23] Kumar, R.; Raghavan, P.; Rajagopalan, S.; Sivakumar, D.; Tomkins, A.; Upfal, E., *Stochastic graph models for the Web graph*, In: Proceedings of the 41^{st} Symposium on Foundations of Computer Science, Redondo Beach, CA, Nov. 12-14, (2000).

[24] Lawrence, S.; Lee Giles, C., *Accessibility of information on the Web*, Nature, 107, (1999), 107-109.

[25] Molloy, M.; Reed, B., *A critical point for random graphs with a given degree sequence*, Random Structures and Algorithms, 6, (1995), 161-180.

[26] Molloy, M.; Reed, B., *The size of the giant component of a random graph with a given degree sequence*, Combinatorics, Probability and Computing, 7, (1998), 295-305.

[27] Moore, A,; Murray, B. H., *Sizing the Web*, Cyveilliance, Inc. White Paper, July 10, (2000).
(http://www.cyveillance.com/resources/7921S_Sizing_the_Internet.pdf)

[28] Moore, C.; Newman, M. E. J., *Epidemics and percolation in small-world networks*, Physics Review E, 61, (2000), 5678-5682.

[29] Munson, K. I., *Internet search engines: Understanding their design to improve information retrieval*, Journal of Internet Cataloging, 2(3-4), (1998), 47-60.

[30] Newman, M. E. J.; Moore, C.; Watts, D., *Meanfield solution of the small-world network model*, Working paper 99-09-066, Santa Fe Institute, Santa Fe, NM, (1999).

[31] Page, L.; Brin, S.; Motwani, R.; Winograd, T., *The PageRank citation ranking: Bringing order to the Web*, Working Paper SIDL-WP-1999-0120, Stanford University, CA, (1999). (http://www-diglib.stanford.edu/WP/WWW/WPTitles.html)

[32] Pirolli, P.; Pitkow, P.; Rao, R., *Silk from a sow's ear: Extracting usable structures from the Web*, In: Proceedings of the ACM Conference on Human factors in computing, Vancouver, Canada, Apr. 13-18, (1996), 118-125.

[33] Watts, D. J., *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, NJ, (1999).

[34] Watts, D. J.; Strogatz, S., *Collective dynamics of 'smallworld' networks*, Nature, 393, (1998), 440-442.

N. Deo and P. Gupta
School of Electrical Engineering and Computer Science
University of Central Florida, Orlando, FL 32816-2362
USA
E-mail: deo, pgupta@cs.ucf.edu