Web and Social Networks

Ravi Kumar* Prabhakar Raghavan† Sridhar Rajagopalan* Andrew Tomkins*

Abstract

The study of the Web as a network has resulted in a better understanding of the sociology of Web content creation. This has paid off in higher precision search engines and more effective algorithms for data mining the Web. This paper reviews the research in this area in the broader context of social networks.

1 Introduction

The diverse authorship, style and distributed content creation on the Web are in sharp contrast to the more controlled and homogeneous domain of classical information retrieval. Link analysis has led to techniques that have dramatically improved the search experience on the Web. This in turn has spawned research into the Web's link structure in its own right, ranging from graph-theoretic studies (degree sequences, connectivity) to community mining and knowledge management.

Modern social network theory is built on the work of Stanley Milgram [19]. In 1967, Milgram conducted experiments in which each of several subjects in Omaha, Nebraska had to convey a letter to his associate in Boston. They could only send the letter to someone they knew on a first-name basis, who in turn had to forward to people they knew on a first-name basis with the objective of getting the letter to Milgram's associate with the smallest number of "hops". Milgram found that the median path length taken by successfully delivered letters was six, leading to the folklore that any two people in the United States are linked in a social network with "six degrees of separation."

In this paper, we review two link analysis algorithms and two structural discoveries about Web topology. There is a strong structural similarity between the Web as a network and social networks. It is our belief that these similarities will lead to progress in knowledge management. We present a number of research challenges that must be addressed in this arena.

Notation. We view the Web as a *directed graph*, with *nodes* (i.e., the pages) and directed *edges* (i.e., links) between certain pairs of the nodes. The notation $q \to p$ denotes that page q links to page p. We say p is an *out-link* of q and q is an *in-link* of p. The *adjacency matrix* A of a graph of n nodes is an $n \times n$ matrix with $A(p,q) = 1$ if and only if $p \to q$. The number of pages that point to p is called the *in-degree* of p and is denoted $\text{indeg}(p)$ and the number of pages that p points to is called its *out-degree*, denoted by $\text{outdeg}(p)$.

*IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA. {ravi, sridhar, tomkins}@almaden.ibm.com

†Verity, Inc., 892 Ross Drive, Sunnyvale, CA 94089, USA. pragh@verity.com

2 Link analysis of the Web

Text-based search engines often function rather poorly on the Web — the sheer volume of data and the low signal-to-noise ratio make them undesirable for locating high-quality pages for a given topic. Text-based engines do not exploit the annotative power of links. Specifically, when the author of a Web page links to another, it represents an implicit "endorsement" of the page being linked to. From the collective judgment in the set of such endorsements, a search system can distill highly relevant content from the Web. Kleinberg [16] and Brin and Page [6] pioneered the use of link information in devising search algorithms for the Web.

The HITS algorithm. HITS [16] identifies two kinds of pages on the Web — *authorities*, which are pages that are authoritative sources of information for the query and *hubs*, which are resource lists containing pointers to a list of resources on the topic. This relationship is mutually reinforcing — good hubs point to good authorities and vice versa. The HITS algorithm formalizes this into an iterative computation, using a *sampling* phase and a *weight-propagation* phase. The sampling phase uses the query terms to collect a *root set* of pages from a text-based search engine and expands this root set into a *base set* by including all pages that are linked to by pages in the root set, and all pages that link to a page in the root set. The idea is that even though the root set might not contain the best pages for the query, the base set will.

The weight-propagation phase works with the subgraph induced by the base set. The algorithm assigns a non-negative *authority weight* a_p and a non-negative *hub weight* h_p with each page p in the base set, both initialized to 1. The update rule for authority and hub weights is:

$$\begin{aligned} a_p &= \sum_{q:q\to p} h_q; \\ h_p &= \sum_{q:p\to q} a_q. \end{aligned} \tag{1}$$

The algorithm iteratively updates these weights by repeating the above computations. Following the iterations, the authorities (resp. hubs) are presented by the ordering of the authority (resp. hub) values.

The authority (resp. hub) values for all the pages form the vectors $\vec{a}$ (resp. $\vec{h}$). The update rules translate to $\vec{a} \leftarrow A^T\vec{h}$ and $\vec{h} \leftarrow A\vec{a}$. We thus have the linear system $\vec{a} \leftarrow (A^TA)\vec{a}$ and $\vec{h} \leftarrow (AA^T)\vec{h}$. The authority (resp. hub) vector is thus the *principal eigenvector* of the matrix A^TA (resp. AA^T). The update rules in Equation (1) turn out to be *power iterations* for computing these eigenvectors. (See the book by Golub and Van Loan [15] for background on eigenvectors and power iteration.) Two points are noteworthy here. Since the power iteration converges to the principal eigenvector for any "non-degenerate" choice of the initial vector, our initial choice for the authority and hub values is inconsequential. Secondly, although the convergence of eigenvector values is guaranteed, we are only interested in the ordering of these values and not their numerical values *per se*.

Extensions to HITS. HITS sometimes has a tendency to generalize or drift to a nearby topic, especially when there are hubs that are quite diverse in the topics they cover. To address these and other issues, a number of researchers [4, 8, 9] introduced many variants to the basic HITS algorithm. Chakrabarti *et al.* [8] use the text surrounding a hyperlink (called the *anchortext*): this text is matched against the query term to obtain a weighted version of Equation (1). In further work, Chakrabarti *et al.* [9] use the tags on a large hub page to break it into smaller *hublets* so that the links within a hublet stay topically focused. Additionally, if several pages from a single domain

participate as hubs, their weights are scaled down so as to prevent a single site from becoming dominant. These heuristics, while retaining the clean mathematical properties of HITS (in terms of convergence, etc.), exploit the content of a page. Bharat and Henzinger [4] presented a number of different extensions to the basic HITS algorithm, substantiating the improvements via a user study. Some of their heuristic improvements include: weighting pages based on how similar they are to a given query topic and averaging the contribution of multiple links from any given site to a specific page.

Pagerank. A different way of utilizing link information was proposed by Brin and Page [6]; this has become the basis of the successful Web search engine `Google` (google.com). Here a query-independent ranking (called the *pagerank*) of all pages is obtained via link analysis. The pagerank of a page p is the limiting fraction of the time spent at p by the following process: at each step with probability ϵ the process jumps to a random page on the Web and with probability $(1-\epsilon)$ it follows a random out-link (if any present) from the current page. Typically, ϵ is chosen to be around 0.15. The pagerank of a page is given by its entry in the principal eigenvector of the matrix $(1-\epsilon)A^T + \epsilon\mathbf{1}$, where $\mathbf{1}$ is the matrix of all ones. The main advantage of pagerank comes from the fact that it is a static ordering and so, given a query term, the pages that contain the query term can be retrieved using a traditional text-based indexer and displayed in the pagerank order. While pagerank is reportedly a component in Google, it is not the only one; many other clever heuristics go into the making of a successful commercial search engine.

Salsa. Salsa[18] is a variant on HITS. Define two matrices, $W = [w_{ij}]$ where $w_{ij} = a_{ij}/\text{outdeg}(i)$, and $W' = [w'_{ij}]$ where $w'_{ij} = a_{ji}/\text{indeg}(j)$. Here $A = [a_{ij}]$ is the adjacency matrix used in HITS. It is easily verified, that both W and W' are stochastic, and thus represent Markov chains. Consequently, $H = WW'$ and $A = W'W$ too are stochastic (products of stochastic matrices remain stochastic). Salsa uses the principal (left) eigenvectors of H and A to rank pages as hubs and authorities respectively.

Borodin *et al.* [5] provide a comparative study of these algorithms and other variants.

3 Communities on the Web

A community on the Web is a collection of Web pages that deal with a common topic, presumably created by people with overlapping interests. Many communities are explicitly available on the Web — for example, newsgroups, email groups and mailing lists, Web rings, personal Web pages in portals, etc. On the other hand, many more are implicit. However, because of their evolving — and in many cases short-lived — nature, it is a formidable task to keep track of these communities manually. A method for extracting these implicit communities automatically was proposed by [17].

The success of HITS suggests that communities contain at their core a dense pattern of linkage from hubs to authorities. This motivates the identification of dense bipartite graphs as signatures of Web communities. By directed dense bipartite graph we mean a graph whose nodes can be partitioned into two sets A and B such that most potential links directed from a node in A to a node in B are in fact present. Given this characterization of communities, many of them can be expected to contain smaller bipartite subgraphs (called *cores*) that are in fact *complete bipartite graphs*: each node in A has a link to each node in B. The idea is to enumerate the cores and grow each core to the community it represents, using algorithms similar to those in Section 2.

The technique used for enumerating such cores is called *trawling*. The main challenge is the efficient enumeration of cores. Naive enumeration is infeasible: consider the example of examining every set of six Web pages to see whether three of them all point to the other three (3×3 cores). Even on a subset of the Web with 100 million pages, this would require the examination of over 10^{40} subsets. The key then is to efficiently prune away most of these subsets from contention. The paper [17] describes a family of such pruning techniques and show that all cores with up to twenty Web pages can be enumerated exhaustively on a standard desktop PC in about 3 days of running time. They used a crawl from Alexa (`www.alexa.com`) circa 1997.

The experiment yielded about 130,000 3×3 cores. Were these linkage patterns coincidental? Manual inspection of a random sample of about 400 communities suggested otherwise: fewer than 5% of the communities discovered lacked a unifying topic. Moreover, about 25% of the communities were not represented in Yahoo!, even in 1999. Of those that do appear in Yahoo!, many appear at as deep as the sixth level in the Yahoo! topic tree. Some sample communities identified by the study include: the community of people interested in *Hekiru Shiina*, a Japanese pop singer; the community of people concerned with oil spills off the coast of Japan; and the community of Turkish student organizations in the U.S. These studies lead to believe that trawling a current copy of the Web will result in the discovery of many more communities that will become explicitly recognized in the future.

In a more recent work [14], a slightly different notion of communities was defined. In this work, a community is a collection of Web pages that have more links to the members of the community than to non-members. Members of a community can be found using maximum flow between a source (consisting of known members of the community) and a sink (consisting of known non-members of the community). Unfortunately, this approach is not fully automatic since it requires specifying the source and sink explicitly. Moreover, unlike trawling, it is unclear how to scale these algorithms for the entire Web.

4 Connectivity and the diameter of the Web

Broder *et al.* [7] aimed to understand the connectivity properties of the Web — is the Web well-connected or does the Web break into small pieces? Is it possible to reach any page from any other page by just following hyperlinks? These questions received impetus from work of Barabasi *et al.* [2, 3] suggesting that the diameter of the Web digraph is 19.

Broder *et al.* first studied a crawl of the Web from Altavista consisting of over 200 million pages and 1.5 billion links, subsequently validating their findings on larger crawls. The results from this paper can be classified into three categories — degree distributions, the bowtie structure, and distance/diameter studies of the Web. Several earlier studies on small portions of the Web demonstrated a power-law behavior for in-degree distributions [17, 3]. The experiments in [7] confirm this phenomenon on a much larger scale. The power-law exponent of the in-degree distribution was determined to be 2.1 and has remained consistent for over three years. The out-degree distribution also conforms to a power-law, albeit in a less striking manner.

Connectivity analysis of the Web graph breaks it into strongly and weakly connected components. Recall that a set of Web pages forms a *strongly connected* component if there is a path following hyperlinks from any page in the set to any other. A set of Web pages is *weakly connected* under a similar definition, except that hyperlinks can be followed in the forward or backward direction.

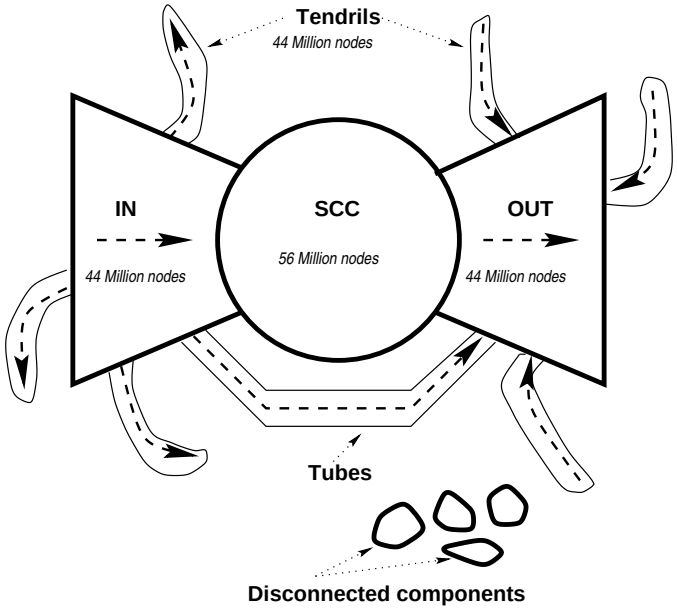

Figure 1: The bowtie structure of the Web.

What are the strongly and weakly connected components of the Web graph? The analysis reveals the following *bowtie* structure (Figure 1), showing that the Web breaks into four natural pieces. The first is called the *SCC*, whose every page is reachable from every other page in the SCC by following links. This is the *largest* strongly connected component of the Web graph. The SCC is a collection of the most valuable resources on the Web. Most portals, university home pages, corporations, and companies can be presumed to be present in the SCC. The second piece is called *IN* and represents those pages that can reach the SCC, but not vice versa. This component may consist of sites that are fairly new to the Web and point to pages in the SCC, but their own identity is yet unknown to the rest of the Web. The third piece is called OUT and includes those pages that can be reached from the SCC, but not vice versa. It could consist of pages in corporate Web sites that do not point back to any page in the SCC. The fourth piece consists of those pages that do not fall into the above classification. Some of them are a consequence of dead links.

In the crawl examined by Broder *et al.* these four pieces are of roughly the same size; this could very well be a strange coincidence as the sizes are in many ways artifacts of crawling policies. The exact relative sizes of the components is not the most interesting aspect of this finding; rather the important message out of this study is that the structure of the Web graph is *not* one of the following two:

(1) A well-connected graph where given any two pages, one could click from one page and get to the other.

(2) A fragmented graph where portions of the graph are well-connected but these well-connected portions are disconnected from each other.

We now argue why the Web cannot be one of the above. By the way in which the Web was decomposed into four pieces, given a pair (p, q) of pages, the only situation where q can be reached from p is when both p is in IN or SCC and q is in SCC or OUT. Since the sizes of IN and OUT are non-trivial, this shows that for roughly $3/4$ of pairs p, q, page q is not reachable from p. This dispels possibility (1). Moreover, this is in contrast with earlier studies [2] which predicted that the Web is well-connected by interpolating connectivity results from a small set of pages collected

from a single site.

On the other hand a large fraction of pages (1/4 in the study) are in the SCC. Moreover, the second largest strongly connected component turns out to be two orders of magnitude smaller than the SCC. This suggests that the Web does not break in regions of well-connected components. Rather, there is a central SCC that holds most of the Web together. This dispels possibility (2).

An off-shoot of this study was to analyze the diameter of the Web. The diameter of the Web, in strict graph-theoretic terms, is infinite as there are (in fact, many) pairs of pages in which one cannot be reached from the other. We need a modified notion of diameter, called the *average connected distance*, which is the average length of the path from page p to page q, *conditioned* upon q being reachable from p. From the study, the average connected distance of the Web is roughly 16, which means that *if* there is a path from p to q, then on average 16 clicks are needed to go from p to q. If we ignore the directions of the links (i.e., if one has the ability to surf to those pages that point to a given page – as apparently in the calculation of [2]), then the average connected distance is only seven. These findings suggest that (under some dramatic simplifications) the Web exhibits a "small-world" behavior.

5 Fractal nature of the Web

Several earlier studies of the Web graph at different scales [2, 1, 17, 7] showed remarkable similarities in various measurements of the Web graph. These observations lead to the natural question: to what extent is the Web a fractal? In other words, do subgraphs of the Web look like "mini Webs"? These and related questions were addressed in a recent paper [11]. The subgraphs studied include a large internet crawl; various subgraphs consisting of about 10% of the sites in the original crawl; 100 Web sites from the crawl each containing at least 10,000 pages; ten graphs, each consisting of every page containing a set of keywords (in which the ten keyword sets represent five broad topics and five sub-topics of the broad topics); a set of pages containing geographical references (e.g., phone numbers, zip codes, city names, etc.) to locations in the western United States; a graph representing the connectivity of Web sites (rather than Web pages); and a crawl of the IBM intranet. The graph properties studied include the in- and out-degree distributions, the bowtie structure (Section 4), distribution of connected components, and the number of communities (Section 3).

The main finding is that self-similarity in the Web is both *pervasive* and *robust*. It is pervasive in that so long as the slice of the Web considered is meaningful, the slice can be thought of as a "mini Web" — its graph-theoretic properties are very similar to that of the entire Web. It is robust in that the parameters corresponding to various properties do not change significantly with the slice considered. For instance, for many of the subgraphs, the power-law exponent of the in-degree turned out to be close to 2.1 (see Figure 2 for a log-log plot of the in-degree distribution for five of the "mini Webs").

Based on this experimental finding, one can derive a graph-theoretic interpretation leading to a natural hierarchical characterization of the graph structure of the Web. According to this, collections of Web pages that share a common trait (for example, all the Web pages that deal with golf) appear similar to the Web as a whole. These "mini Webs" are connected by a *navigational backbone* which not only ties together the collections of pages, but also ties together the many different and overlapping "mini Webs". The user navigates through the Web by going from one "mini Web" to the other uses the navigational backbone.

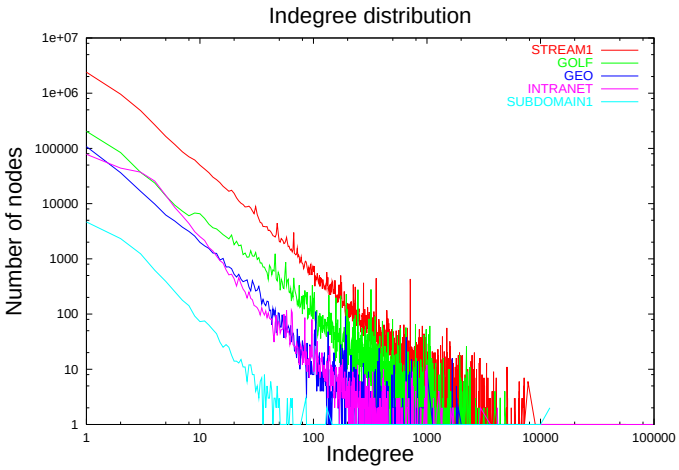

Figure 2: Indegree distribution for subgraph of a crawl (Stream1), golf-related pages (Golf), geographically related pages (Geo.Western), IBM intranet (IBM.intranet), and pages corresponding to a small corporation (Subdomain1).

Self-similarity is pervasive in social networks. While self-similarity on the Web has been observed in other contexts like Web traffic [10] and physical topology of the internet [13], the fractal nature of the Web in a graph-theoretic setting adds further evidence to its small-world nature.

References

[1] Adamic, L., Huberman, B. (2000): Scaling behavior on the world wide Web. A comment in Science **287**, 2115

[2] Albert, R., Jeong, H., Barabasi, A. L. (1999): Diameter of the world wide Web. Nature **401**, 130–131

[3] Barabasi, A., Albert, R. (1999): Emergence of scaling in random networks. Science **286**, 509–512

[4] Bharat, K., Henzinger, M. (1998): Improved algorithms for topic distillation in hyperlinked environments. Proc. 21st Annual Intl. ACM SIGIR Conf., 104–111

[5] Borodin, A., Roberts, G. O., Rosenthal, J. S., Tsaparas, P. (2001): Finding authorities and hubs from link structures on the world wide Web. Proc. 10th World-Wide Web Conf., 415–429

[6] Brin, S., Page, L. (1998): The anatomy of a large scale hypertextual Web search engine. Proc. 7th World-Wide Web Conf./Comp. Networks **30**(1-7), 107–117

[7] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J. (2000): Graph structure in the Web. Proc. 9th World-Wide Web Conf./Comp. Networks **33**(1-6), 309–320

[8] Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., Rajagopalan, S. (1998): Automatic resource compilation by analyzing hyperlink structure and associated text. Proc. 7th World-Wide Web Conf./Comp. Networks **30**(1-7), 65–74

[9] Chakrabarti, S., Dom, B., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. (1998): Experiments in topic distillation. Proc. SIGIR Workshop on Hypertext Inf. Retrieval, 13–21

[10] Crovella, M. E., Bestavros, A. (1997): Self-similarity in world wide Web traffic: Evidence and possible causes. IEEE/ACM Trans. on Networking **5**(6), 835–846

[11] Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D., Tomkins, A. (2001): Self-similarity in the Web. Proc. 27th Intl. Conf. on Very Large Databases, 69–78

[12] Egghe, L., Rousseau, R. (1990): *Introduction to Informetrics.* Elsevier, Amsterdam

[13] Faloutsos, M., Faloutsos, P., Faloutsos, C. (1999): On power law relationships of the internet topology. Proc. ACM SIGCOMM, 251–262

[14] Flake, G. W., Lawrence, S., Giles, C. L. (2000): Efficient identification of Web communities. Proc. 6th ACM SIGKDD Intl. Conf. on Knowledge Disc. and Data Mining, 150–160

[15] Golub, G., Van Loan, C. F. (1999): Matrix Computations. Johns Hopkins University Press, London

[16] Kleinberg, J. M. (2000): Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632

[17] Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. (1999): Trawling the Web for cyber communities. Proc. 8th World-Wide Web Conf./Comp. Networks **31**(11-16), 1481–1493

[18] Lempel, R., Moran, S. (2000): The stochastic approach for link-structured analysis (SALSA) and the TKC effect. Proc. 9th World-Wide Web Conf./Comp. Networks **33**(1-6), 387–401

[19] Wasserman, S., Faust, K. (1994): Social Network Analysis. Cambridge University Press, New York