Colloquium

Extracting knowledge from the World Wide Web

Monika Henzinger* **and Steve Lawrence**

Google, Inc., 2400 Bayshore Parkway, Mountain View, CA 94043

The World Wide Web provides a unprecedented opportunity to automatically analyze a large sample of interests and activity in the world. We discuss methods for extracting knowledge from the web by randomly sampling and analyzing hosts and pages, and by analyzing the link structure of the web and how links accumulate over time. A variety of interesting and valuable information can be extracted, such as the distribution of web pages over domains, the distribution of interest in different areas, communities related to different topics, the nature of competition in different categories of sites, and the degree of communication between different communities or countries.

The World Wide Web has become an important knowledge and communication resource. As more people use the web for more tasks, it provides an increasingly representative and unprecedented in scale machine-readable sample of interests and activity in the world.

However, the distributed and heterogeneous nature of the web makes large-scale analysis difficult. We provide an overview of recent methods for analyzing and extracting knowledge from the web, along with samples of the knowledge that can be extracted.

Sampling the Web

The sheer size of the web has led to a situation where even simple statistics about it are unknown, for example, its size or the percentage of pages in a certain language. The ability to sample web pages or web servers uniformly at random is very useful for determining statistics. For example, we can use random URLs to estimate the distribution of the length of web pages, the fraction of documents in various Internet domains, or the fraction of documents written in various languages. We can also determine the fraction of web pages indexed by various search engines by testing the engines for the presence of pages chosen uniformly at random.

Random Walk. One approach to sample web pages approximately uniformly at random is based on the idea of a *random walk*, where we take successive steps in random directions. Henzinger *et al.* (1) have performed several such random walks on the web. Their main idea is to perform a random walk so that a page is visited by the walk with probability roughly proportional to its PageRank (2) value, and then to sample the visited pages with probability inversely proportional to their PageRank value. Thus, the probability that a page is sampled is a constant independent of the page.

One definition of the PageRank value of a web page uses a random walk: *The initial page of the walk is chosen uniformly at random from all pages. Assume the random walk is at page* p *at a given time step. With probability* d, *follow an outlink of paper* p, *chosen uniformly at random. With probability 1* − d, *select a random page out of all pages.* The PageRank of a page *p* is the fraction of steps that the walk spent at *p* in the limit, i.e., the PageRank is the stationary distribution of the random walk.

When trying to implement this random walk to generate random web pages, two problems arise: (*i*) The random walk assumes already that we can find a random page on the web, the very problem that we want to solve. (*ii*) Many hosts on the web have a large number of links within the same host and very few leaving them. If such a host is encountered early in the walk, then there is a good chance that most pages are from this host when the walk is stopped, i.e., the walk "never found its way out of the host." The main culprit is that any implementation can only take a finite number of steps, whereas the definition requires an infinite number.

To avoid these problems, Henzinger *et al.* (1) proposed and implemented the following modified random walk: *Given a set of initial pages, choose one page at random to be the start page. Assume the random walk is at page* p *at a given time step. With probability* d, *follow an outlink of page* p, *chosen uniformly at random. With probability 1* − d, *select a random host out of all hosts visited so far, and jump to a randomly selected page out of all pages visited on this host so far. In this definition*, *all pages in the initial set are also considered to be visited.*

The two problems are avoided as follows: (*i*) Instead of choosing a random page out of all pages, a random page from a subset of visited pages is chosen. (*ii*) Instead of jumping to a random page, the walk jumps to a random host and then to a random visited page on that host. In this way, even a host that has dominated the walk so far only has the same chance of being visited as any other visited host.

Because of the modification in the walk and because of the fact that the walk has to be finite in practice, the modified random walk visits a page with probability approximately proportional to its PageRank value, which is the stationary distribution of the PageRank random walk.

Afterward, the visited pages are sampled with probability inversely proportional to their PageRank value. If the PageRank value is not known, it can be approximated by computing PageRank on the graph of visited pages. Alternatively, the *visit ratio*, i.e., the ratio of the number of times the page was visited over the length of the random walk, can be used as an approximation of the PageRank value. The latter holds because the PageRank value of a page is defined to be the visit ratio of the PageRank random walk in the limit.

As an example of the statistics we can generate by using this approach, Table 1 shows the percentage of URLS in each top-level domain in fall 1999 generated with this method. Other approaches for sampling web pages based on a random walk methodology are presented in Bar-Yossef *et al.* (3) and Rusmevichientong *et al.* (4).

IP Address Sampling. An approach to obtaining a random sample of web servers is to randomly sample IP addresses, testing for a web server at the standard port (5). There are currently 256^4 (≈4.3 billion) possible IP addresses. If IPv6 was widely used on the web, this approach may not be possible; however, IPv6 has

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Mapping Knowledge Domains," held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

*To whom correspondence should be addressed. E-mail: monika@google.com.

Table 1. The top 10 top-level domains according to the percentage of sampled pages in each domain

Domain	Percentage of pages
com	46.93
edu	9.27
org	8.59
net	4.74
jp	3.51
de	3.17
gov	2.92
uk	2.75
ca	1.95
au	1.69
us	1.67
fr	0.81

jp, Japan; de, Germany; uk, United Kingdom; ca, Canada; au, Australia; us, United States; fr, France.

not been widely adopted and this approach is still practical today. Of the 4.3 billion possible IP addresses, some are unavailable and some are known to be unassigned. Many sites are temporarily unavailable due to Internet connectivity problems or web server downtime. To minimize this effect, all IP addresses can be checked multiple times.

This method finds many web servers that would not normally be considered part of the publicly indexable web. These include servers with authorization requirements (including firewalls), servers that respond with a default page, servers with no content (e.g., sites "coming soon"), web hosting firms that present their homepage on many IP addresses, printers, routers, proxies, mail servers, CD-ROM servers, and other hardware that provides a web interface. Many of these can be automatically identified, for example, by using regular expressions.

A number of issues lead to minor biases. The sample corresponds to the subset of servers that are active and respond to requests at the polling times. It is possible for one IP address to host several web sites, multiple IP addresses may serve identical content, and some web servers do not use the standard port. It is common for large sites to use multiple IP addresses that serve the same content (for load balancing and redundancy). This could potentially result in a higher probability of finding larger sites. To minimize the bias, we can use the domain name system to identify multiple IP addresses serving the same content, and consider only the lowest numbered address to be part of the publicly indexable web. Most major sites are not virtually hosted, and few public servers operate on a nonstandard port.

Fig. 1 shows a sample of the results of this approach, showing the distribution of server types found from sampling 3.6 million IP addresses in February 1999 (5). About 83% of servers were commercial, whereas $\approx$6% of web servers were found to have scientific/educational content (defined here as university, college, and research laboratory servers).

Also analyzed in the same study was metadata usage on the homepages of each server, where the results showed that only 34.2% of servers contained the common "keywords" or "description" metatags on their homepage. The low usage of the simple HTML metadata standard suggests that acceptance and widespread use of more complex standards, such as XML or Dublin Core, may be very slow (0.3% of sites contained metadata using the Dublin Core standard). High diversity was also noted in the HTML META tags found, with 123 distinct tags, suggesting a lack of standardization in usage.

Discussion. Unfortunately, current techniques for sampling web pages exhibit biases and do not achieve a uniform random sample. The main problem with the approaches based on random walks is that any implementation is limited to a finite random walk. The main challenge when using IP address sampling is how to subsample the pages that are accessible from a given IP address.

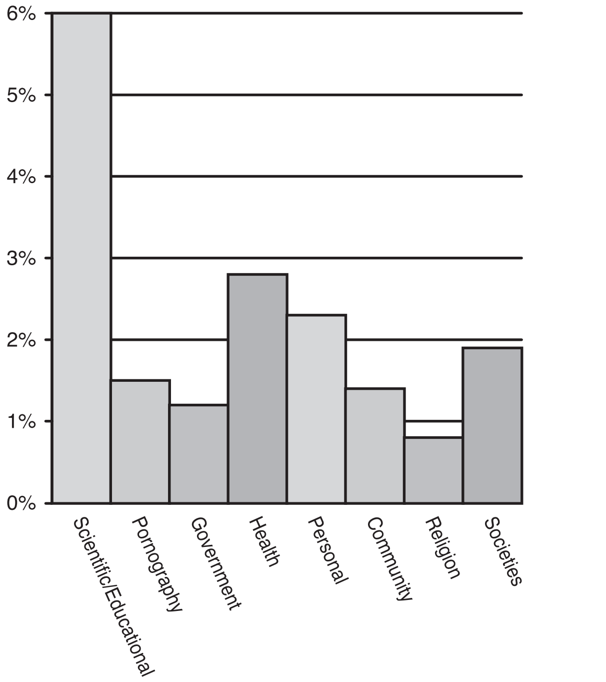

Fig. 1. The distribution of information on publicly indexable web servers. About 83% of servers contained commercial content (e.g., company homepages). The remaining classifications are shown above. Sites may have multiple classifications.

As the web grows it has become impractical to retrieve all pages. Thus, it becomes more important to be able to uniformly sample pages to measure properties of the web. One pragmatic approach is to use two or more approaches that have different biases, for example, a random walk approach and an approach based on IP address sampling, and analyze the agreement between their results.

A fundamental question is what should be counted. For example, consider a web site that contains 10 million pages containing weather statistics for different points in time, compared to another containing the same statistics all on one page. Likewise, a research paper on the web may be on one page or split over multiple pages (6). Additionally, there can be many pages that do not contain original content, they may be transformations of content on other pages (extensions to methods for identify similar document such as ref. 7 can be valuable), or even randomly generated pages. This suggests that some measure of importance may be incorporated into the analysis; for example, we may consider creating a random sample of items that have at least *n* links to them from other sites, where an item may be a single web page or a collection of web pages (for example, the entire 10 million pages in the weather statistics example). Analysis of web sites as opposed to individual pages is also helpful here.

Analyzing and Modeling Web Growth

We can also extract valuable information by analyzing and modeling the growth of pages and links on the web. Several researchers (8–11) have observed that the link distribution of web pages follows a power law: the probability that a randomly selected web page has *k* inlinks is proportional to $k - \gamma$, where

$\gamma = 2.1$. The outlink distribution follows a power law with $\gamma = 2.72$. This observation led to the design of various models for the web graph. We describe two models, namely, the preferential attachment model by Barabási and Albert (8, 9) and the copy model by Kleinberg *et al.* (12). We also describe two extensions of these models to better account for deviations of the model from observations.

Preferential Attachment. Barabási and Albert (8, 9) attribute power law scaling to a "rich get richer" mechanism called preferential attachment. As the network grows, the probability that a given node receives an edge is proportional to that node's current connectivity. Specifically, Barabási and Albert propose the following (undirected) web graph model.

Growth. Starting with a small number m_0 of nodes, at every time step add a new node u with $m \leq m_0$ edges.

Preferential attachment. When choosing the nodes to which the new node connects, we assume that the probability p that a new node will be connected to node u depends on the degree k_u of node u, such that $p = k_u/\Sigma_{\text{node w}} k_w$.

An analysis based on mean-field theory shows that the probability for a randomly selected node to have k inlinks in this model is proportional to $k - 3$. More specifically, for a node u created at time step t_u, the expected degree is $m(t/t_u)^{0.5}$. Thus, older pages get rich faster than newer pages, leading to a "rich get richer" mechanism.

This model explains the observed power law inlink distribution. However, the model exponent is 3, whereas the observed exponent is 2.1. Additionally, it is not known that older web pages gain inlinks faster than new pages. Finally, different link distributions are observed among web pages of the same category, which we discuss below.

Competition Varies. The early models fail to account for significant deviations from power law scaling common in almost all studied networks. For example, among web pages of the same category, link distributions can diverge strongly from power law scaling, exhibiting a roughly log-normal distribution. In earlier models predicting a power law distribution, most members of a community fare poorly; they have none or very few links to them. However, for actual distributions, many community members can have a substantial number of inlinks, with the mode of the distribution varying up to ≈800 links for universities. Moreover, conclusions about the attack and failure tolerance of the Internet based on the early models may not fully hold within specific communities.

The distributions for outbound web links, and for a variety of other social and biological networks, also display significant deviations from power law (8, 10, 11, 13, 14).

Pennock *et al.* (15) introduced a new model of network growth, mixing uniform and preferential attachment, that accurately accounts for the true connectivity distributions found in web categories, the web as a whole, and other social and biological networks. Previous models imply a drastic "winners take all" scenario on the web, whereby highly referenced pages continue to grow richer in links, whereas new entrants languish in comparison. In fact, the situation is not so inequitable when examined at a local rather than a global level.

Pennock's model generalizes the Barabási–Albert model to incorporate both preferential attachment and a uniform baseline probability of attachment. The model predicts the observed shape of both the body and tail of typical connectivity distributions, including those observed within specific categories of web pages where the divergence from power law is especially marked. In the model, larger modes arise from faster rates of growth of edges as compared to vertices, suggesting an explanation for the different modes observed within different categories of web pages.

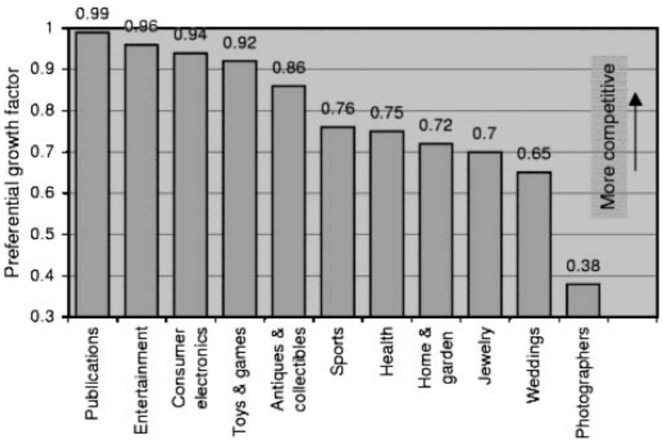

Fig. 2. Competition in different e-commerce categories in March 2002. "More competitive" refers to tougher competition, i.e., it is harder to compete with existing popular sites.

Pennock's model can be used to analyze competition in different categories on the web. Fig. 2 shows the degree of preferential growth for web sites in different e-commerce categories. The publications e-commerce category is the most competitive, where in this case we use competitive to mean that it is harder for a new site to compete with existing sites. The photographers category is the least competitive. There are multiple factors that can lead to the differences in competition that we see. For photographers, one likely factor is their local nature: photographers typically serve only a local community, and those serving different areas usually do not compete. Another factor may be that people looking for photographers use methods other than the web more often (e.g., referrals from friends). Perhaps because people typically use professional photographers rarely, they are also less likely to create and share information among related sites on the web.

A number of models related to Pennock's model have been proposed: Dorogovtsev *et al.* (16) and Levene *et al.* (17) independently propose similar generalizations of the Barabási–Albert model (the addition of a uniform component), motivating it in part as a natural way to parameterize the power-law exponent. Albert and Barabási (18) have proposed their own augmented model that involves a parameterized mixture of three processes: vertex additions, edge additions, and edge rewirings. The combination leads to a connectivity growth function that is roughly a sum of uniform and preferential terms. Even Simon (19) in 1955 invoked a similar process to explain Estoup–Zipf word frequency distributions.

Kleinberg *et al.* (12) explained the power-law inlink distributions with a *copy model* that constructs a directed graph. A slightly modified version as in ref. 20 works as follows: At each time step, one new node u is added with d outlinks. The destinations of these d links with source u are chosen as follows: First, an existing node v is chosen uniformly at random. Then, for $j = 1, 2, \ldots, d$, the jth link of u points to a random existing node with probability α, and to the destination of v's jth link with probability $1 - \alpha$.

Similarly to the Pennock *et al.* (15) model, this model is a mixture of uniform and preferential influences on network growth. A detailed analysis in ref. 20 shows that it leads to a power law inlink distribution as well as to a large number of bipartite cliques.

These models can be used to analyze the fault tolerance of the networks. Recently, Park *et al.* (21) analyzed the Internet for susceptibility to faults and attacks by using simulated data from models similar to those above and with actual data. They find

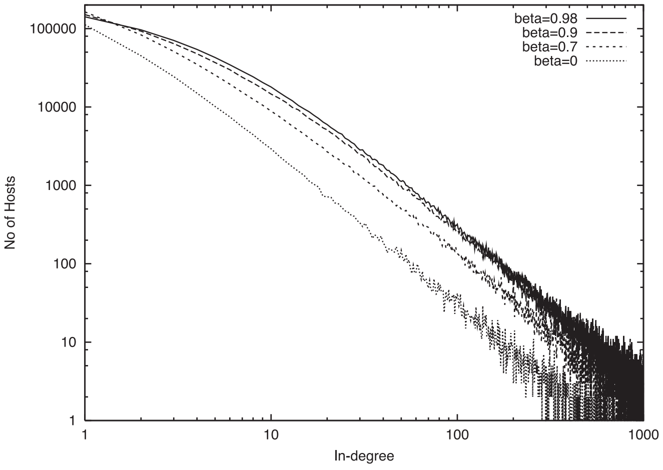

Fig. 3. Inlink distribution as predicted by the ''re-link model'' with varying β values.

that the Internet is becoming more preferential as it evolves: it is more robust to random failures but is also more vulnerable to attacks.

All of the current models of web growth are an approximation - the true nature of growth on the web is more complex. It is notable that relatively simple models can quite accurately reproduce the actual distributions and behavior of the networks. However, an open problem is refining the models to further improve their accuracy.

The Hostgraph Model. The web is a hierarchically nested graph, with domains, hosts, and pages introducing different levels of affiliation. Instead of modeling the web at the level of pages, one can also model it on the host or domain level. Using the host level leads to the following *hostgraph*: Each node represents a host, and each directed edge represents the hyperlinks from pages on the source host to pages on the target host. Bharat *et al.* (22) show that the weighted inlink and the weighted outlink distributions in the host graph have a power law distribution with $\gamma = 1.62$ and $\gamma = 1.67$, respectively. However, the number of small inlink hosts is considerably smaller than predicted by the model, i.e., there is "flattening" of the curve for low inlink hosts.

Bharat *et al.* (22) present the following modification to the copy graph model, called the *re-link model*, to explain this "flattening": At each time step, with probability β we select a random already existing node u, and with probability $1 - \beta$ we create a new node u. Then we add d new additional outlinks to it. The destinations of these d links with source i are chosen as follows: First, an existing node v is chosen uniformly at random. Second, one picks d random outgoing edges from v. Then, for $j = 1, 2, \ldots, d$, the jth link of u points to a random existing node with probability α, and to the destination of v's jth link with probability $1- \alpha$.

The difference to the copy model is that with probability $1 - \beta$ no new node is added. Because new nodes start without inlinks the number of low inlink nodes is reduced. Fig. 3 shows the resulting inlink distribution for a graph of 1 million nodes with $d = 7$ and $\alpha = 0.05$ for various β values.

In a recent paper, Cooper and Frieze (23) actually proved that an extension of a model very similar to the re-link model generates graphs whose link distributions follow a power law. Chakrabarti *et al.* (24) used a variant of the Bar-Yossef *et al.* random walk together with a topic classifier to analyze the link distributions of pages on the same topic.

Bharat *et al.* also analyzed affinity between top level country domains in June 2001. Table 2 shows the 20 source domains with the most outlinks together with the .com domain. For each source domain, it lists the percentage of outlinks into the same domain, into the .com domain, and into the four most highly linked country domains from that source domain.

Table 2. Most frequently linked-to domains from country domains

	% of weighted outdegree					
	com	self	1	2	3	4
com	82.9		**net** 6.5	**org** 2.6	**jp** 0.8	**uk** 0.7
au	27.0	58.8	**uk** 1.0	**ch** 0.5	**ca** 0.4	**de** 0.3
br	17.8	69.1	**uk** 0.4	**pt** 0.4	**de** 0.4	**ar** 0.2
ca	19.4	65.2	**uk** 0.6	**fr** 0.4	**se** 0.3	**de** 0.3
cn	15.8	74.1	**tw** 0.4	**jp** 0.2	**de** 0.2	**hk** 0.1
cz	8.1	82.4	**sk** 1.0	**de** 0.7	**uk** 0.4	**ch** 0.1
de	16.0	71.2	**uk** 0.8	**ch** 0.6	**at** 0.5	**nl** 0.2
dk	13.8	73.0	**uk** 1.1	**de** 1.0	**int** 0.7	**no** 0.7
es	38.9	42.3	**de** 1.3	**uk** 1.0	**fr** 0.5	**int** 0.3
fr	20.9	61.9	**ch** 0.9	**de** 0.8	**uk** 0.7	**ca** 0.5
it	19.3	64.6	**de** 1.0	**uk** 0.7	**fr** 0.4	**ch** 0.3
jp	17.4	74.5	**to** 0.8	**cn** 0.6	**uk** 0.2	**de** 0.1
kr	26.5	57.1	**jp** 0.6	**uk** 0.5	**de** 0.3	**to** 0.3
nl	21.2	61.7	**de** 1.3	**uk** 1.1	**be** 0.6	**to** 0.5
no	16.1	65.6	**de** 1.2	**se** 0.9	**uk** 0.7	**dk** 0.6
pl	4.2	92.2	**de** 0.2	**uk** 0.1	**ch** 0.1	**nl** 0.1
ru	10.0	84.9	**ua** 0.4	**su** 0.2	**uk** 0.2	**de** 0.2
se	22.6	60.0	**nu** 1.6	**uk** 0.9	**de** 0.7	**to** 0.6
tw	22.0	66.0	**to** 1.3	**au** 0.6	**jp** 0.6	**ch** 0.4
uk	34.2	45.9	**de** 0.7	**ca** 0.5	**jp** 0.3	**se** 0.3
us	34.4	33.1	**ca** 0.6	**uk** 0.5	**au** 0.2	**de** 0.2

Domains are listed in boldface. au, Australia; br, Brazil; ca, Canada; cn, China; cz, Czech Republic; de, Germany; dk, Denmark; es, Spain; fr, France; it, Italy; jp, Japan; kr, Korea; nl, The Netherlands; no, Norway; pl, Poland; ru, Russia; se, Sweden; tw, Taiwan; uk, United Kingdom; us, United States.

Communities on the Web

The web allows communities to rapidly form with members spread out around the world. Identification of communities on the web is valuable for several reasons. Practical applications include automatic web portals and focused search engines, content filtering, and complementing text-based searches. Community identification also allows for analysis of the entire web and the objective study of relationships within and between communities.

Flake *et al.* (25–27) define a web community as a collection of web pages such that each member page has more hyperlinks (in either direction) within the community than outside of the community (this definition may be generalized to identify communities with varying sizes and levels of cohesiveness). Community membership is a function of both a web page's outbound hyperlinks as well as all other hyperlinks on the web; therefore, the communities are "natural" in the sense that they are collectively organized by independently authored pages. They show that the web self-organizes such that these link-based communities identify highly related pages (Fig. 4).

Identifying a naturally formed community, according to Flake's definition, is intractable in the general case because the basic task maps into a family of nonparametric–complete graph partitioning problems (28). However, if one assumes the existence of one or more *seed* web sites and exploits systematic regularities of the web graph (8, 30, 31), the problem can be recast into a framework that allows for efficient community identification using a polynomial time algorithm.

This is just one of many link-based approaches proposed for identifying collections of related pages. Kumar *et al.* (11) con-

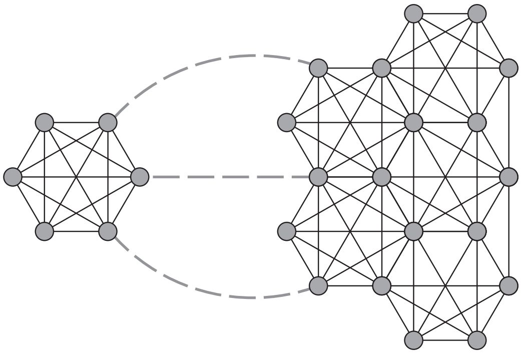

Fig. 4. A community identification example. Maximum flow methods will separate the two subgraphs using any choice of source vertex *s* from the left subgraph and sink vertex *t* from the right subgraph, removing the three dashed links. As formulated with standard flow approaches, all community members must have at least 50% of their links inside the community; however, artificial links can be added to change the threshold from 50% to any other desired threshold. Thus, communities various sizes and with varying levels of cohesiveness can be identified and studied.

sider dense bipartite subgraphs as indications of communities, and expand such graphs into larger sets with HITS (32). Reddy and Kitsuregawa (33) propose a related approach that can be used to identify a hierarchy of communities. Other approaches include bibliometric methods such as cocitation and bibliographic coupling (34–36), the PageRank algorithm (2), the HITS algorithm (32), bipartite subgraph identification (11), spreading activation energy (37), and others (33, 38, 39).

Bipartite subgraph identification, cocitation, and bibliographic coupling are localized approaches that aim to identify well defined graph structures existing in a narrow region of the web graph. PageRank, HITS, and spreading activation energy (SAE) are more global and iteratively propagate weights through a significant portion of the graph. The weights reflect an estimate of page importance (PageRank), how authoritative or hub-like a page is (HITS) or how "close" a candidate page is to a starting region (SAE). PageRank and HITS are related to spectral graph partitioning (40), seeking to find "eigen-web-sites" of the web graph's adjacency matrix or a simple transformation of it. Both HITS and PageRank are relatively insensitive to their choice of parameters, unlike SAE, where results are extremely sensitive to the choice of parameters (37).

Localized approaches are appealing because the structures they identify unambiguously have the properties that the algorithms were designed to find. However, one limitation of these approaches is that they cannot find large related subsets of the web graph because the localized structures are too small. At the other extreme, PageRank and HITS operate on large subsets of the web graph and can identify large collections of web pages that are related or valuable. One limitation of these methods is that it may be hard to understand and defend the inclusion of a given page in the collections that are produced. In practice, HITS and PageRank are combined with textual content either for preprocessing (HITS) or postprocessing (PageRank) (41).

The current approaches to finding communities work well in many, but not all, cases, and have not yet moved from research to widely used products. The approaches often produce some communities with unexpected or missing members. One difficulty is the definition of a community; different people often have different opinions on how a set of pages should be grouped into clusters or communities (29). This is an open area of research.

Summary

The web offers both great opportunity and great challenge in the quest for improving our understanding of the world. The combined efforts of many researchers has resulted in several valuable methods for analysis, and the extraction of a wide variety of valuable knowledge.

However, there are still many open problems and areas for future research. Many of the web analysis studies as presented in this paper provide valuable results for a particular point in time; however, few of these provide directly comparable results at different points in time. It would be very interesting to repeat many of the studies to provide updated analysis, and to provide additional insight into the evolution of the web. The problem of uniformly sampling the web is still open in practice: which pages should be counted, and how can we reduce biases? Web growth models approximate the true nature of how the web grows: how can the current models be refined to improve accuracy, while keeping the models relatively simple and easy to understand and analyze? Finally, community identification remains an open area: how can the accuracy of community identification be improved, and how can communities be best structured or presented to account for differences of opinion in what is considered a community?

1. Henzinger, M., Heydon, A., Mitzenmacher, M. & Najork, M. (2000) *Comput. Networks* **33,** 295–308.
2. Brin, S. & Page, L. (1998) in *Proceedings of the 7th International World Wide Web Conference* (Elsevier, Amsterdam), pp. 107–117.
3. Bar-Yossef, Z., Berg, A., Chien, S., Fakcharoenphol, J. & Weltz, D. (2000) in *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB)* (Morgan Kaufman, San Francisco), pp. 535–544.
4. Rusmevichientong, P., Pennock, D. M., Lawrence, S. & Giles, C. L. (2001) in *American Association for Artificial Intelligence Fall Symposium on Using Uncertainty Within Computation* (Am. Assoc. Artificial Intelligence, Menlo Park, CA), pp. 121–128.
5. Lawrence, S. & Giles, C. L. (1999) *Nature* **400,** 107–109.
6. Eiron, N. & McCurley, K. S. (2003) in *Proceedings of Hypertext 2003* (Assoc. Comput. Machinery Press, New York), pp. 85–94.
7. Broder, A., Glassman, S., Manasse, M. & Zweig, G. (1997) in *Sixth International World Wide Web Conference* (Assoc. Comput. Machinery Press, New York), pp. 391–404.
8. Barabási, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
9. Barabási, A.-L., Albert, R. & Jeong, H. (1999) *Physica A* **272,** 173–187.
10. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000) *Comput. Networks* **33,** 309–320.
11. Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999) *Comput. Networks* **31,** 1481–1493.
12. Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. S. (1999) in *Proceedings of the 5th International Conference on Computing and Combinatorics* (Springer, Berlin), pp. 1–18.
13. Albert, R., Jeong, H. & Barabási, A.-L. (1999) *Nature* **401,** 130–131.
14. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. (2000) *Nature* **407,** 651–654.
15. Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J. & Giles, C. L. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 5207–5211.
16. Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. (2000) *Phys. Rev. Lett.* **85,** 4633–4636.
17. Levene, M., Fenner, T., Loizou, Y. & Wheeldon, R. (2002) *Comput. Networks* **29,** 277–287.
18. Albert, R. & Barabási, A.-L. (2000) *Phys. Rev. Lett.* **85,** 5234–5237.
19. Simon, H. A. (1955) *Biometrika* **42,** 425–440.
20. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. & Upfal, E. (2000) in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE Press, Piscataway, NJ), pp. 57–65.
21. Park, S.-T., Khrabrov, A., Pennock, D. M., Lawrence, S., Gilles, C. L. & Ungar, L. H. (2003) *IEEE Infocom 2003, San Francisco, CA, April 1–3 2003* (IEEE Press, Piscataway, NJ), CD-ROM.
22. Bharat, K., Chang, B.-W., Henzinger, M. & Ruhl, M. (2001) in *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)* (IEEE Press, Piscataway, NJ), pp. 51–58.

23. Cooper, C. & Frieze, A. (2002) *Random Struct. Algorhythms* **22,** 311–335.
24. Chakrabarti, S., Joshi, M. M., Punera, K. & Pennock, D. (2002) in *Proceedings of the 11th International World Wide Web Conference* (Assoc. Comput. Machinery Press, New York), pp. 517–526.
25. Flake, G. W., Lawrence, S. & Giles, C. L. (2000) in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining* (Assoc. Comput. Machinery Press, New York), pp. 150–160.
26. Flake, G. W., Lawrence, S., Giles, C. L. & Coetzee, F. (2002) *IEEE Comput.* **35,** 66–71.
27. Flake, G., Tsioutsiouliklis, K. & Tarjan, R. (2002) *Graph Clustering Techniques Based on Minimum Cut Trees* (NEC, New York), Technical Report TR 2002-06.
28. Garey, M. R. & Johnson, D. S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, New York).
29. Macskassy, S., Banerjee, A., Davison, B. & Hirsh, H (1998) in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)* (AAAI Press, New York), pp. 264–268.
30. Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E. & Lukose, R. M. (1998) *Science* **280,** 95–97.
31. Watts, D. & Strogatz, S. (1998) *Nature* **393,** 440–442.
32. Kleinberg, J. M. (1998) in *Proceedings of the Ninth Annual Association for Computing Machinery/Society for Industrial and Applied Mathematics Symposium on Discrete Algorithms* (Assoc. Comput. Machinery/SIAM Press, New York), pp. 668–677.
33. Reddy, P. K. & Kitsuregawa, M. (2002) in *Workshop on Web Analytics, April 13 2002* (Arlington, VA), pp. 11–13.
34. Garfield, E. (1979) *Citation Indexing: Its Theory and Application in Science* (Wiley, New York).
35. Larson, R. (1996) in *Proceedings of the Annual Meeting of the American Society for Information Science* (Assoc. Comput. Machinery Press, New York), pp. 71–78.
36. White, H. D. & McCain, K. W. (1989) *Annu. Rev. Info. Sci. Technol.* **24,** 119–186.
37. Pirolli, P., Pitkow, J. & Rao, R. (1996) in *Proceedings of the Association for Computing Machinery Conference on Human Factors in Computing Systems, Chicago, IL* (Assoc. Comput. Machinery Press, New York), pp. 118–125.
38. Gibson, D., Kleinberg, J. & Raghavan, P. (1998) in *Proceedings of the 9th Association for Computing Machinery Conference on Hypertext and Hypermedia* (Assoc. Comput. Machinery Press, New York), pp. 225–234.
39. Chakrabarti, S., van der Berg, M. & Dom, B. (1999) in *Proceedings of the 8th International World Wide Web Conference* (Elsevier, Amsterdam), pp. 545–562.
40. Chung, F. (1996) *Spectral Graph Theory*, CBMS Lecture Notes (Am. Math. Soc., Providence, RI).
41. Bharat, K. & Henzinger, M. (1998) in *Proceedings of the 21st International ACM SIGR Conference on Research and Development in Information Retrieval* (Assoc. Comput. Machinery Press, New York), pp. 104–111.