Hubs, Authorities, and Communities

Jon M. Kleinberg

Cornell University Web: http://www.cornell.edy/
Department of Computer Science Web: http://www.cs.cornell.edu/
Ithaca, NY 14853
Web: http://www.cs.cornell.edu/home/kleinber/

Categories and Subject Descriptors: H.5.4 **[Information Interfaces and Presentation]**: Hypertext / Hypermedia ; F.2.2 **[Analysis of Algorithms and Problem Complexity]**: Non-numerical algorithms and problems - *Computations on discrete structures*

General Terms: Algorithms, Human Factors.

Additional Key Words and Phrases: Hypertext structure, World Wide Web, Link analysis, Graph algorithms.

This has been a computer scientist's revolution; but we all share in its results. Only a few years ago, the World Wide Web was known just to a small research community; it is hard to remember that the voluminous content we see on the WWW, expanding by an estimated million pages each day, has grown up around us in so short a time. Researchers now release their results to the Web before they appear in print; corporations list their URLs alongside their toll-free numbers; news media and entertainment companies vie for the attention of a browsing audience. The Web has become the most visible manifestation of a new medium: a global, populist hypertext.

The speed with which this medium has emerged is a testament to the universality of the computational models on which it is built; much of the software and network infrastructure supporting the Web was developed long before there was a Web to support. In much the same way, when we investigate the structural properties of the WWW, we may make use of well-studied models of discrete mathematics -- the combinatorial and algebraic properties of graphs. The Web can be naturally modeled as a directed graph, consisting of a set of abstract nodes (the pages) joined by directional edges (the hyperlinks). Hyperlinks encode a considerable amount of latent information about the the underlying collection of pages; thus, the structure of this directed graph can provide us with significant insight into its content. Within this framework, we can search for signs of meaningful graph-theoretic structure; we can ask: What are the recurring patterns of linkage that occur across the Web as a whole?

The profound complexity of the WWW is a crucial challenge in this search for structure. Content on the Web is being created by millions of autonomous participants, a group that includes multinational corporations, government agencies, academic researchers and students, and individual AOL subscribers. The variability in the content and in its quality is infinitely greater than what we encounter in more traditional media. Faced with these problems, we can adopt a perspective that has emerged in the study of citation analysis and social networks: A valuable way to gain understanding of a complex network is through the identification of its important, or prominent, nodes.

In the field of citation analysis, a number of methods have been proposed for measuring the importance of scientific journals [Egghe 1990]. Perhaps the most widely used is Garfield's *impact factor*, which provides a quantitative

---

"score" for each journal proportional to the average number of citations per paper published in the previous two years [Garfield 1972]. This measure encodes the fundamental intuition that more heavily-cited journals have more overall impact on a field, and it has been applied to rank journals in the Journal Citation Reports of the Institute for Scientific Information.

Beginning with this measure, we could picture enhancing our estimate of the important journals as follows. Suppose we have concluded, by counting citations, that the journals *Science* and *Nature* are highly prominent. Then if we are comparing two more obscure journals which have received roughly the same number of citations as one another, and we discover that one of these journals has received many citations from *Science* and *Nature*, we may wish to elevate its ranking. In other words, it is better to receive citations from an important journal than from an unimportant one. We can see this phenomenon on the WWW as well: counting the number of links to a page can give us a general estimate of its prominence on the Web, but a page with very few incoming links may also be prominent, if two of these links come from the home pages of *Yahoo!* and *Netscape.* Defining such a richer notion of importance, or prominence, contains an intrinsic element of circularity: it arises from the fragile intuition that a node is important if it receives links from other important nodes. Several measures incorporate this basic circular notion, and each contains a method for capturing the implicit equilibrium that this circularity encodes.

Two early approaches to embrace this theme in the study of social networks are the measures of Katz [Katz 1953] and Hubbell [Hubbell 1965]. (See also the discussion in Wasserman and Faust [Wasserman 1994].) In Hubbell's formulation, each node has an internal, *a priori* weight that is given at the outset. We are also given a specified *connection strength* between each pair of nodes. We seek to assign a global weight, or prominence value, to each node in such a way that a node's global weight is equal to the sum of its internal weight and the global weights of all nodes that link to it, scaled by their connection strengths. This can be represented as a collection of linear equations; its solution captures a version of the equilibrium discussed above. The solution has some of the key features we were seeking; if a node has large weight, the nodes it links to will tend to have large weights as well. In the field of citation analysis, Pinski and Narin [Pinski 1976] developed a similar notion of *influence weights*, using a somewhat different mathematical model. First, they define the strength of the connection from one journal to another to be the percentage of the citations in the first journal that refer to the second. They then seek a set of weights that obey the following equilibrium: the weight of each journal *J* should be equal to the sum of the weights of all journals citing *J*, scaled by the strengths of their connections to it. Again, we can see desirable features of this definition; if a journal receives regular citations from other journals of large weight, it too will acquire large weight.

A methodology closely related to Pinski-Narin influence weights has been applied to the WWW in the work of Brin and Page [Brin 1998], which forms the basis of the search engine *Google.* The many dead-ends in the link structure of the Web -- the fact that many prominent sites have no links to the outside world -- causes the pure weight-balancing model of Pinski and Narin to exhibit counter-intuitive behavior in this domain. Brin and Page apply a ``smoothing'' operation to the model, giving every page a small but positive connection strength to every other page, and compute equilibrium weights with respect to these modified connection strengths.

The types of equilibrium we have been discussing so far are characterized by the following notion: important pages are those that receive links from other important pages. In many cases, however, the relationships one sees among Web pages suggests that a different model is at work. Consider, for example, the home pages of the main WWW search engines. Although these are all important pages on a common topic, they do not link to each other; this is for the simple reason that they are in competition with one another, and the goal of each is to keep users on its site for as long as possible. And yet there is a way to recognize them as a unified set of prominent pages, and this is through the large collection of pages that act as resource lists and guides to search engines, linking to all of them from a single point. Such pages need not themselves receive many links; they are recognizable because they link to many prominent sites in this focused manner.

In our recent work, we have identified a form of equilibrium among WWW sources on a common topic in which we explicitly build into the model this diversity of roles among different types of pages [Kleinberg 1998]. Some pages, the most prominent sources of primary content, are the *authorities* on the topic; other pages, equally intrinsic to the

structure, assemble high-quality guides and resource lists that act as focused *hubs*, directing users to recommended authorities. The nature of the linkage in this framework is highly asymmetric. Hubs link heavily to authorities, but hubs may themselves have very few incoming links, and authorities may well not link to other authorities. This, as we suggested above, is completely natural; many good hubs on the Web are being created by relatively anonymous individuals, and the main authorities on a topic are often in competition with one another, either explicitly or implicitly. A formal type of equilibrium consistent with this model can be defined as follows: we seek to assign two numbers -- a *hub weight* and an *authority weight* -- to each page in such a way that a page's authority weight is proportional to the sum of the hub weights of pages that link to it; and a page's hub weight is proportional to the sum of the authority weights of pages that it links to.

A common premise in all these approaches is that a large ``community'' of thematically related information sources can be characterized by the way in which it links and refers to its most central, prominent members. These prominent sources serve as a form of broad-topic summary of a much larger underlying ensemble; one can thereby condense an enormous volume of information down to a more tractable representation. We believe that this framework has considerable potential to provide further understanding of the structure and content of the World Wide Web.

Acknowledgments

The author is supported in part by an Alfred P. Sloan Research Fellowship, a David and Lucile Packard Foundation Fellowship, an ONR Young Investigator Award, and NSF Faculty Early Career Development Award CCR-9701399.

References

[Brin 1998] Sergey Brin and Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine" in Proceedings of World-Wide Web '98 (WWW7), [Online: ttp://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm], April 1998.

[Egghe 1990] Leo Egghe and Ronald Rousseau. Introduction to Informetrics, Elsevier, 1990.

[Garfield 1972] Eugene Garfield. Citation analysis as a tool in journal evaluation. Science 178, 1972.

[Hubbell 1965] Charles H. Hubbell. "An Input-Output Approach to Clique Identification" in Sociolmetry, 28, 377-399, 1965.

[Katz 1953] L. Katz. "A new status index derived from sociometric analysis" in Psychometrika 18(1), 39-43, March 1953.

[Kleinberg 1998] Jon M. Kleinberg. "Authoritative Sources in a Hyperlinked Environment" in Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 668-677, [Online: http://www.cs.cornell.edu/home/kleinber/auth.ps], January 1998.

[Pinski 1976] G. Pinski and Francis Narin. "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics" in Information Processing and Management. 12, 1976.

[Wasserman 1994] Stanley Wasserman and Katherine Faust. Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.