The Self-Organized Web:
The Yin to the Semantic Web's Yang

Gary William Flake, David M. Pennock, and Daniel C. Fain

{gary.flake,david.pennock,dan.fain}@overture.com

Overture Services / 74 N. Pasadena Ave. 3rd floor / Pasadena, CA 91103 / USA

[*]Assigning superlatives to the Web is easy: it's massive, it's dynamic, it's decentralized—it's unlike anything else in the world. But one of the Web's most amazing attributes is that it is arguably the largest self-organized artifact in existence. Every day millions of Web publishers add, delete, move, and change their pages and links, yet what results is far from random or haphazard. Rather, from these millions of uncoordinated decisions emerge a startling number of regularities and patterns. The Web programmer's task—whether working on search, collaborative filtering, data mining, e-commerce, or scientific analysis—is to use these patterns to make the Web more digestible to users. This goal complements the Semantic Web's goal: to have humans help make the Web more digestible for computers. Exploiting the self-organized Web will improve tomorrow's algorithms; manually adding computer-friendly annotations to the Semantic Web will help today's less-sophisticated algorithms to cope.

Self-organization and the Web

Although Web authors' choices worldwide are largely uncoordinated, they are anything but uncorrelated. Unlike any (nondegenerate) random graph, the Web graph's large-scale structure has a "bow tie" organization that contains four distinct regions: a strongly connected core, an origination bow, a termination bow, and disconnected islands.[1] In the Web's core, hyperlinks are relatively sparse, yet they collectively possess small-world properties, forming many redundant and relatively short paths between most pages.[2] When sampled across the Web, inbound and outbound hyperlink distributions follow a clear power law distribution that resembles distributions in biology.[3] Viewed on a small scale, simple and small bipartite subgraphs are a signature of topical Web communities' formation.[4]

We also see clear structural self-organization at intermediate levels. For example, when aggregating only over a specific type of page (for example, movie, newspaper, photography, or university homepages), hyperlink distributions shift from a strict power law to a unimodal form, not unlike the change in the biomass distribution when aggregated over a single species instead of all species. A generative Web growth model with only one free parameter explains this unusual hyperlink distribution property with remarkable simplicity and accuracy.[5] Moreover, a simple definition—that a Web community is a collection in which each member is predominately hyperlinked to other community members—has yielded an efficient procedure for identifying self-organized Web communities. Empirically, these communities are topically and textually focused, even though the identification procedure uses only hyperlinks.[6] Figure 1 depicts a few of the Web's self-organizing properties.

Web data mining

Given the Web's size and decentralization, it seems almost a paradox that pure hyperlink data mining algorithms work. Nearly all current hyperlink methods strongly overlap with methods pioneered in graph clustering and where the problems are typically NP-hard.[7] The Web's enormous size suggests that it would be a more difficult domain, yet the reality is somewhat different.

Two popular Web data mining algorithms, HITS[8] (hypertext induced topic search) and PageRank,[9] have shown considerable success when coupled with text retrieval. Both methods attempt to capture Web pages' importance by recursively analyzing how pages hyperlink to each other. The PageRank algorithm can be roughly interpreted as simulating how a random walker would traverse the Web graph over forward links if also allowed to teleport to other pages with some small probability. After

[*] G. W. Flake, D. M. Pennock, and D. C. Fain, " The Self-Organized Web: The Yin to the Semantic Web' s Yang," ***IEEE Intelligent Systems***, 2003. To appear.

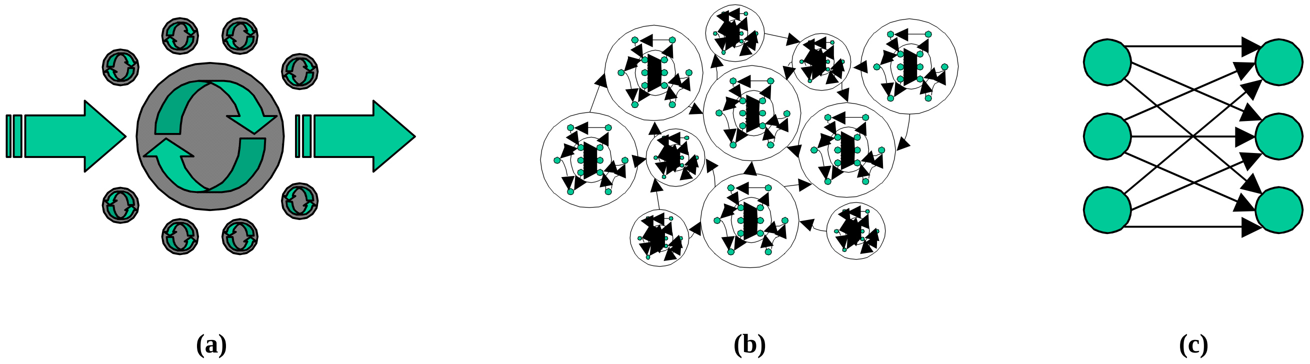

Figure 1: The Web viewed from three different scales: (a) On the whole, the Web has a bow-tie structure with a strongly connected core, pages that exclusively point into the core or are by pointed by the core (but not both), and disconnected islands; (b) At the intermediate level, the Web forms self-organized communities that are topically focused and have strong internal link structure; (c) At the lowest level, bipartite cores form the seeds of Web communities.

completion, PageRank assigns to a Web page a score that is approximately equal to the probability that the random walker will visit that page. Intuitively, the PageRank score is similar to the recursive definition that a Web page is important if other important pages link to it.

PageRank clearly improves text retrieval, as evidenced by the popularity of Google (the inventors' search engine[9]), which largely attributes its competitive edge to PageRank. What is truly interesting, however, is that PageRank is an existence proof that:

1. The Web is well behaved in that worst-case complexity results are far too pessimistic
2. The Web's self-organization can be used to improve Web search and data mining

Consider the first point. PageRank is technically equivalent to a power method estimation of the maximal eigenvector of a simple transformation of the Web graph's adjacency matrix. Linear algebra shows that the power method converges at a rate related to the ratio of the first two eigenvalues of the matrix being used. Simply put, the more similar the first two eigenvalues, the slower the procedure's convergence rate.

Many matrices are ill suited for power method procedures because of the convergence properties. However, in the Web's case, the power law distribution on inbound hyperlinks nearly guarantees that the Web will never possess such pathologies because Web pages with many other pages linking to them are exceedingly rare and in some sense very competitive with each other. In other words, the Web doesn't seem to like having two maximal eigenvalues of nearly equal value.[10]

As for the second point, think of PageRank as something of a collective voting scheme, where pages not only vote for each other but also vote to determine how many votes each should have. The votes are labeled, in a sense, with anchor text; a string match in a referring page's anchor text is weighed more heavily than a match in the target page itself. The links and labels from the Open Directory Project (ODP)—a voluntary effort to categorize Web sites—are the most important. In such a collective, distributed framework, it isn't at all obvious that the aggregate vote would add any value to the retrieval task. But, when coupled with text retrieval, PageRank is remarkably adept at producing the "correct" answer when correctness and popularity are highly correlated.

Top-down and bottom-up

Recently, many have championed the Semantic Web[11] as a means to improve information retrieval on the Web. Proponents argue that the Web is ill suited in its current form for automated processing because the information is unstructured to the point that semantics are nearly impossible for machines to infer. In the Semantic Web, authors will use a markup language to annotate data with semantic labels so that machines can identify content meaning and use rules for manipulating semantic information appropriately. In a best-case scenario, the markup language will be nearly complete and agreeable, and used consistently by Web authors. Authoring tools might generate the markups implicitly, but such markups will need to make sense alongside those the authors add manually. Implicit markup is easy to envision for product catalogs, but semantically marking up long passages of text—magazine articles, for example—could be daunting.

Realistically, Semantic Web advocates understand that some holes and inconsistencies in the markup language are inevitable and that author adoption rates and proficiencies will be heterogeneous. Witness the failure of Xanadu,[12] a hypertext framework arguably superior to HTML/HTTP. It failed partly because its rules and guarantees imposed too great a burden on authors. A simple example of Web authors' natural laziness is the prevalence of rasterized text unmarked by ALT text tags. Some of Xanadu's ideas were revived in less-demanding forms. For example, instead of keeping an up-to-date set of backlinks at the target page, Web backlinks are retained in search engines. But relaxing enforcement in the Semantic Web will lead to another form of self-organized entity, even if more structured than today's Web. To add to the confusion, as long as there is search spam, a contingent supplying false information through metadata will exist. This underscores the value of objective third-party annotation, even when minimal, such as ODP.

A complementary best-case scenario envisions Web algorithms intelligent enough to infer semantics from the current, non-annotated, self-organized Web, without the aid of semantic markups. Information extraction tools are progressing, making inroads on problems such as parsing resumes and finding contact information, but today's best algorithms still fall woefully short of people's capability to extract meaning from the Web. We have only scratched the surface of the self-organized Web's potential. Future data mining methods will use efficient algorithms optimized for the expected case of Web data (rather than the worst case or random case), use predictable Web properties to reduce problem sizes and input dimensionality, and have performance bounds consistent with generative Web growth models. With these algorithms, search engines will be able to cluster and classify the entire Web along multiple dimensions (topic, type, and genre, for example). The most advanced search engines and autonomous Web agents will be able to use richer forms of metadata if and when a new markup structure is adopted.

But for the foreseeable future, efforts to leverage the self-organized Web will complement efforts to build the Semantic Web. Both open up opportunities for innovative new algorithms—data mining on one hand, symbolic inference on the other. Where these efforts meet, tools will arise for vastly improved search, filtering, personalization, economic efficiency, and scientific understanding of the social forces and trends reflected in the Web. We believe that the Web's properties—structure, content, and explicit or inferred metadata—will continue to evolve in a decentralized and self-organized way. Users will benefit most if work on creating the Semantic Web coevolves with work on tools for data-driven analysis of the self-organized Web.

References

1. A. Broder et al., "Graph Structure in the Web: Experiments and Models," *Proc. 9th World Wide Web Conf.* (WWW9), Elsevier, 2000.
2. D.J. Watts and S.H. Strogatz, "Collective Dynamics of 'Small World' Networks," *Nature*, vol. 393, no. 6684, June 1998, pp. 440–442.
3. AL Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, Oct. 15, 1999, pp. 509–512.
4. R. Kumar et al., "Trawling the Web for Emerging Cyber-communities," *Proc. 8th World Wide Web Conf.* (WWW8), Elsevier, 1999.
5. D.M. Pennock et al., "Winners Don't Take All: Characterizing the Competition for Links on the Web," *Proc. Nat'l Academy of Sciences*, vol. 99, no. 8, April 2002, pp. 5207–5211.
6. G.W. Flake et al., "Self-organization of the Web and Identification of Communities," *IEEE Computer*, vol. 35, no. 3, Mar. 2002, pp. 66–71.
7. M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, 1979.
8. J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *J. ACM*, vol. 46, no. 5, Sept. 1999, pp. 604–632.
9. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. 7th World Wide Web Conf.* (WWW7), Elsevier, 1998.
10. F. Chung and L. Lu, "The Average Distances in Random Graphs with Given Expected Degrees," *Proc. Nat'l Academy of Sciences*, vol. 99, no. 25, Dec. 2002, pp. 15,879–15,882.
11. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 284, no. 5, May 2001, pp. 34–43.
12. T.H. Nelson, "Xanalogical Structure, Needed Now More Than Ever: Parallel Documents, Deep Links to Content, Deep Versioning, and Deep Re-use," *ACM Computing Surveys*, vol. 31, no. 4, Dec. 1999, pp. 194–225.