Self-Organization and Identification of Web Communities

Despite its decentralized and unorganized nature, the Web self-organizes to allow identification of highly related pages based solely on connectivity, without the inherent bias of text-based approaches.

Gary William Flake
Steve Lawrence
NEC Research Institute

C. Lee Giles
Pennsylvania State University

Frans M. Coetzee
GenuOne

The vast improvement in information access is not the only advantage resulting from the increasing percentage of hyperlinked human knowledge available on the Web. Additionally, much potential exists for analyzing interests and relationships within science and society. However, the Web's decentralized and unorganized nature hampers content analysis. Millions of individuals operating independently and having a variety of backgrounds, knowledge, goals, and cultures author the information on the Web.

Despite the Web's decentralized, unorganized, and heterogeneous nature, our work shows that the Web self-organizes and its link structure allows efficient identification of communities. This self-organization is significant because no central authority or process governs the formation and structure of hyperlinks.

WEB COMMUNITIES

A Web *community* is a collection of Web pages in which each member page has more hyperlinks within the community than outside the community. We can generalize this definition to identify communities with varying sizes and levels of cohesiveness. Community membership is a function of both a Web page's outbound hyperlinks and all other hyperlinks on the Web because the rest of the Web collectively forms a page's inbound hyperlinks. Therefore, these communities are "natural" in that independently authored pages collectively organize them. The Web self-organizes such that these link-based communities identify highly related pages.

Compared to previous methods of finding related Web pages described in the "Finding Related Pages on the Web" sidebar, our approach retains the transparency of methods such as cocitation and bibliographic coupling in explaining why pages belong to a community, yet it can identify Web communities of arbitrary dimensions. Our algorithm achieves this performance using only link information, without the text information that algorithms such as Hyperlink-Induced Topic Search (HITS) use.

In the absence of full natural-language processing, a Web author's creation of an explicit link can be a stronger indication of relevance than the implied links that simple textual phrase and structure matching generate. In addition, separating link structure from content facilitates using content-based similarity measures to independently validate the performance of the link-based community estimation process.

We can model the Web as a graph in which Web pages are vertices and hyperlinks are edges. Identifying a naturally formed community—according to our definition—is generally intractable because the basic task maps into a family of NP-complete graph partitioning problems.[1] However, if we assume the existence of one or more *seed* Web sites—pages that are positive examples of community members—and exploit the Web graph's systematic regularities,[2-4] we can recast the problem. This approach provides a framework that permits efficient community identification via a polynomial time algorithm that should scale well to studying the entire Web graph.

MAXIMUM FLOW COMMUNITIES

We can recast the problem into a maximum flow framework to analyze the flow between graph ver-

0018-9162/02/$17.00 © 2002 IEEE

tices. If edges are water pipes and vertices are pipe junctions, the maximum flow problem tells us how much water we can move from one junction to another.

Lester Ford and Delbert Fulkerson's Max Flow-Min Cut theorem[5] proves that the maximum flow is identical to the minimum cut. Therefore, if you know the maximum flow between two points, you also know what edges you would have to remove to completely disconnect the same two points—the *cut set*.

Many polynomial time algorithms exist for solving the *s*-*t* maximum flow problem.[6] These algorithms formally define the problem with respect to a directed graph $G = (V, E)$, with edge capacities $c(u, v) \in Z^+$ and two vertices, $s, t \in V$, so that the result is the maximum flow that can be routed from the source *s* to the sink *t* that obeys all capacity constraints.[7]

Figure 1 shows the basic intuition of our approach. As formulated with standard flow ap-

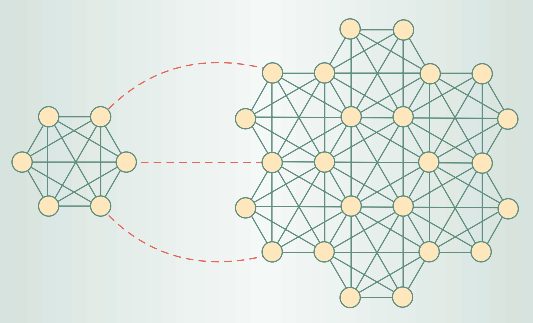

Figure 1. A simple community-identification example. Maximum flow methods separate the two subgraphs with any choice of source vertex *s* from the left subgraph and sink vertex *t* from the right subgraph, removing the three dashed links.

Finding Related Pages on the Web

Previous link-based research for identifying collections of related pages includes bibliometric methods such as cocitation and bibliographic coupling,[1] the PageRank algorithm,[2] the Hyperlink-Induced Topic Search (HITS) algorithm,[3] bipartite subgraph identification,[4] spreading activation energy (SAE),[5] and others.[6,7]

Localized approaches, such as cocitation, bibliographic coupling, and bipartite subgraph identification, seek to identify well-defined graph structures that exist inside a narrow region of the Web graph. More global approaches, such as PageRank, HITS, and SAE, work by iteratively propagating weights through a significant portion of the Web graph. The weights reflect an estimate of page importance (PageRank), how authoritative or hublike a Web page is (HITS), or how "close" a candidate page is to a starting region (SAE). PageRank and HITS relate to spectral graph partitioning[8] and therefore seek to find "eigen-Web-sites" of the Web graph's adjacency matrix or a simple transformation of it. Unlike SAE results, which show extreme sensitivity to the choice of parameters,[5] both HITS and PageRank are relatively insensitive to their choice of parameters.

Localized approaches are appealing because the identified structures unambiguously possess the properties that the algorithms seek by design. However, these approaches fail to find large related subsets of the Web graph because the localized structures are simply too small. At the other extreme, PageRank and HITS can operate on large subsets of the Web graph and can therefore identify large collections of related or valuable Web pages.

Because they are based on spectral graph partitioning, these methods often make it difficult to understand and defend the inclusion of a given page in the collections they produce. In practice, we can achieve meaningful results with HITS and PageRank only when we use textual content for either preprocessing (HITS) or postprocessing (PageRank). Without auxiliary text information, both PageRank and HITS have limited success in identifying collections of related pages.[9]

References

1. E. Garfield, *Citation Indexing: Its Theory and Application in Science*, John Wiley & Sons, New York, 1979.
2. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. 7th Int'l World Wide Web Conf.*, Elsevier Science, New York, 1998, pp. 107-117.
3. J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Proc. 9th Ann. ACM-SIAM Symp. Discrete Algorithms*, ACM Press, New York, 1998, pp. 668-677.
4. R. Kumar et al., "Trawling the Web for Emerging Cyber-Communities," *Proc. 8th Int'l World Wide Web Conf.*, Elsevier Science, New York, 1999, pp. 1481-1493.
5. P. Pirolli, J. Pitkow, and R. Rao, "Silk from a Sow's Ear: Extracting Usable Structures from the Web," *Proc. ACM Conf. Human Factors in Computing Systems*, ACM Press, New York, 1996, pp. 118-125.
6. D. Gibson, J. Kleinberg, and P. Raghavan, "Inferring Web Communities from Link Topology," *Proc. 9th ACM Conf. Hypertext and Hypermedia*, ACM Press, New York, 1998, pp. 225-234.
7. S. Chakrabarti, M. van der Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," *Proc. 8th Int'l World Wide Web Conf.*, Elsevier Science, New York, 1999, pp. 1623-1640.
8. F. Chung, *Spectral Graph Theory*, *CBMS Lecture Notes*, American Mathematical Soc., Providence, R.I., 1996.
9. K. Bharat and M. Henzinger, "Improved Algorithms for Topic Distillation in Hyperlinked Environments," *Proc. 21st Int'l ACM SIGIR Conf.*, ACM Press, New York, 1998, pp. 104-111.

Figure 2. Algorithms for identifying Web communities. (a) The exact-flow-community algorithm augments the Web graph in three steps. (b) The approximate-flow-community algorithm uses a subset of the Web graph found by a fixed-depth crawl that follows both inbound and outbound hyperlinks.

```
procedure EXACT-FLOW-COMMUNITY
    input : graph : $G = (V, E)$ ; set : $S \subset V$ ; integer : $k$ .
    Create artificial vertices, $s$ and $t$, and add to $V$ .
    for all $v \in S$ do
        Add $(s, v)$ to $E$ with $c(s, v) \equiv \infty$.
    end for
    for all $(u, v) \in E$ do
        Set $c(u, v) \equiv k$.
        if $(v, u) \notin E$ then add $(v, u)$ to $E$ with $c(v, u) \equiv k$.
    end for
    for all $v \in V, v \notin S \cup \{s, t\}$ do
        Add $(v, t)$ to $E$ with $c(v, t) \equiv 1$.
    end for
    call : MAX-FLOW $(G, s, t)$.
    output : all $v \in V$ still connected to $s$ .
end procedure
```

(a)

```
procedure APPROXIMATE-FLOW-COMMUNITY
    input : set : $S$ .
    while number of iterations is less than desired do
        Set $G = (V, E)$ to fixed depth crawl from $S$ .
        Set $k$ to $|S|$ .
        call : $C$ = EXACT-FLOW-COMMUNITY$(G, S, k)$.
        Rank all $v \in C$ by number of edges in $C$.
        Add highest ranked non-seed vertices to $S$.
    end while
    output : all $v \in V$ still connected to $s$ .
end procedure
```

(b)

proaches, all community members must have at least 50 percent of their links inside the community. However, maximum-flow methods use additional artificial links to change the threshold from 50 percent to any other desired threshold. Thus, we can identify communities of various sizes and with varying levels of cohesiveness.

One or more seed sites can play the role of the source vertex. For example, if the goal is to improve categories in a Web directory, we would use the existing pages in each category as seed sites. The sum total of the edges connected to the seed sites must be greater than the size of the cut set (the edges whose removal separates the source and the sink), represented by the dashed lines in Figure 1. If the seed sites do not meet this constraint, the procedure will only identify a subset of the community. In the worst case, we will only identify the seed sites as members of the community.

We could use an approximate centroid of the Web graph, such as Yahoo, as the sink. However, our method works without an explicit sink site via the graph augmentation steps shown in Figure 2, for which we have developed the corresponding theorem and proof.[8]

AUGMENTING THE WEB GRAPH

The exact-flow-community procedure augments the Web graph in three steps: It adds an artificial source *s* with infinite-capacity edges routed to all seed vertices in *S*, makes each preexisting edge bidirectional and rescales it to a constant value *k*, and routes all vertices except the source, sink, and seed vertices to the artificial sink with unit capacity. After augmenting the Web graph, we use a maximum-flow procedure to produce a residual-flow graph. All vertices accessible from *s* through nonzero positive edges form the desired result and satisfy our definition of a community.

The approximate-flow-community algorithm takes a set of seed Web sites as input, crawls to a fixed depth, including both inbound and outbound hyperlinks, and queries search engines to find inbound hyperlinks. The algorithm then applies the exact-flow-community procedure to the induced graph from the crawl, ranks the sites by the number of edges each has inside the community, adds the highest-ranked nonseed sites to the seed set, and iterates the procedure. The first iteration may only identify a very small community. However, adding new seeds identifies increasingly larger communities. Note that *k* is chosen heuristically.

With access to the entire Web graph, the exact-flow-community algorithm returns a set of Web pages that complies with our definition of a community because the maximum-flow procedure always finds a bottleneck from the source to the sink. Thus, any page that remains connected to the source must have more hyperlinks in the community than outside it; otherwise, a more efficient cut would have moved the Web site in question to the noncommunity.

In the exact-flow-community algorithm, the artificial sink is generic in that it is on the receiving end of an edge from every other vertex in the graph. Thus, separating the source from the sink finds a community that is strongly connected internally but is relatively disconnected externally from the rest of the graph.

We used the approximate-flow-community algorithm to find our experimental results. However, we could also exploit the Web's dynamic nature with an iterative approximate algorithm that tests for new candidate community members by count-

Francis Crick Community		Stephen Hawking Community		Ronald Rivest Community	
Score	**Site title or description**	**Score**	**Site title or description**	**Score**	**Site title or description**
80	Biography of Francis Harry Compton Crick (Nobel Foundation)	85	Professor Stephen W. Hawking's Web pages	86	Ronald L. Rivest home page
79	Biography of James Dewey Watson (Nobel Foundation)	46	*Stephen Hawking's Universe* (PBS)	29	"Chaffing and Winnowing: Confidentiality without Encryption"
51	The Nobel Prize in Physiology or Medicine 1962 (Nobel Foundation)	17	The Stephen Hawking pages	20	Thomas H. Cormen's home page at Dartmouth
50	"Biographical Sketch of James Dewey Watson" (Cold Spring Harbor Lab.)	15	"Stephen Hawking Builds Robotic Exoskeleton" (parody in *The Onion*)	9	"The Mathematical Guts of RSA Encryption"
41	"A Structure for Deoxyribose Nucleic Acid" (*Nature*, 2 Apr. 1953)	14	Stephen Hawking and Intel	8	German news story on Cryptography
…		…		…	
1	*Felix D'Herelle and the Origins of Molecular Biology* (Amazon.com)	1	*"Did the Cosmos Arise from Nothing?"* (MSNBC)	1	Phil Zimmermann's PGP Web page
1	Biography of Gregor Mendel	1	Spanish page for *Stephen Hawking's Universe*	1	"A Very Brief History of Computer Science"
1	Magazine: *HMS Beagle Home*	1	Relativity Group at DAMTP, Cambridge	1	Cormen/Leiserson/Rivest: Introduction to Algorithms
1	The Alfred Russel Wallace Page	1	Millennium Mathematics Project	1	Security and encryption links
1	US Human Genome Project 5 Year Plan	1	Particle physics education and information sites	1	HotBot Directory: Computers & Internet, Computer Science, People: R

Figure 3. Sample results generated using the approximate-flow-community algorithm, showing the top five and bottom five pages for each community.

ing the number of candidate links that fall within the preexisting community.

EXPERIMENTAL RESULTS

To test the approximate-flow-community identification algorithm, we used the personal home pages of three prominent scientists—Francis Crick, Stephen Hawking, and Ronald Rivest—as a single seed in three separate runs. Each trial of the approximate algorithm produced communities consisting of approximately 200 Web pages. At the later stages of the runs, the induced graphs often contained tens of thousands of vertices. Thus, the algorithm pruned many pages to produce these communities.

Figure 3 shows the top five and bottom five pages for each community, with the scores indicating the total number of inbound and outbound links that a Web page has to other pages in its community. The majority of Web pages found were highly topically related, often in nontrivial ways. For example, the Crick community contained many references to Darwin, the Human Genome Project, and Rosalind Franklin. Likewise, the Hawking community contained many sites dealing with cosmology, relativity, and Cambridge University, while the Rivest community contained numerous encryption Web sites along with sites focusing on his coauthors. Lower-ranked pages were usually topically related to the seed scientist, although they might not include that scientist's name.

Table 1 shows how we more completely characterized the three communities by extracting all text *features*—a word or consecutive word pair—from the pages within a community and from 10,000 randomly chosen Web pages. We then sorted all features in the community according to their ability to separate community pages from noncommunity pages, as measured by the Kullback-Leibler metric. Thus, Table 1 lists the features that are most useful for separating community pages from noncommu-

Table 1. The 15 most significant text features for each community, sorted in descending order using the Kullback-Leibler metric.

Community	Most significant text features
Crick	crick, nobel, dna, “francis crick,” “the nobel,” “of dna,” watson, “james watson,” francis, molecular, biology, genetics, “watson and,” “structure of,” “crick and”
Hawking	hawking, “stephen hawking,” stephen, “hawking s,” “s universe,” physics, “black holes,” “the universe,” cambridge, cosmology, einstein, relativity, damtp, “universe the”
Rivest	rivest, “l rivest,” “ronald l,” ronald, cryptography, rsa, “ron rivest,” lcs, “theory lcs,” encryption, “lcs mit,” theory, chaffing, winnowing, crypto

nity pages. The extracted features support our hypothesis that link-based communities are topically related.

To obtain more precise characterizations of the communities, we exhaustively searched for all three-term binary classifiers that disambiguate community from noncommunity pages. Simple disjunctive expressions of community-related keywords matched a large fraction of the communities with low false-alarm rates. For example, *crick* or *nobel* or *darwin* matched 54 percent of the Francis Crick community but only 0.5 percent of random Web pages. Similarly, *hawking* or *relativity* or *for mathematical* matched 84 percent of the Stephen Hawking community, but only 0.2 percent of random pages. Finally, *rivest* or *cormen* or *to encrypt* matched 85 percent of the Ronald Rivest community versus 1.3 percent of random pages. The pages in the communities are highly related topically in that they have simple and compact descriptions in the form of binary classifiers.

In comparison, simple breadth-first crawl strategies quickly lose topical relevance. For the three scientists we investigated, only about 10 percent of pages at a depth of two from the seed site matched our classification rules. In contrast, the communities that we identify have pages up to a depth of five links from the seed site. Breadth-first crawling to this depth would yield an enormous number of pages.[9]

Based only on the self-organization of the Web’s link structure, we can efficiently identify highly topically related communities whose individual members can be spread over a large area of the Web graph. Because our method is completely divorced from text-based approaches, we can use the communities we identify to infer meaningful text rules and augment text-based methods.

Global community identification permits analysis of the entire Web and the objective study of relationships within and between communities such as scientific disciplines or countries. Such research could provide insight into the organization and interests of sectors of society. For example, links between scientific disciplines could facilitate more timely identification of emerging interdisciplinary connections.

Applications of our method include creating improved search engines, content filtering, and objective analysis of Web content and the relationships between Web communities. Specialized search engines could identify the community of pages within their domains, individuals could identify communities of others with similar interests, and Web-filtering software could identify the community of pages to be filtered. Finally, objective and rigorous analysis of the entire Web, taking into account issues such as the “digital divide,”[10] may help improve our understanding of the world.

Acknowledgment

We thank Inktomi for the random URL data.

References

1. M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, New York, 1979.
2. A-L. Barabási and R. Albert, “Emergence of Scaling in Random Networks,” *Science*, 15 Oct. 1999, pp. 509-512.
3. B.A. Huberman et al., “Strong Regularities in World Wide Web Surfing,” *Science,* 3 Apr. 1998, pp. 95-97.
4. D. Watts and S. Strogatz, “Collective Dynamics of “Small-World” Networks,” *Nature*, 4 June 1998, pp. 440-442.
5. L.R. Ford Jr. and D.R. Fulkerson, “Maximal Flow through a Network,” *Canadian J. Math*, vol. 8, no. 3, 1956, pp. 399-404.
6. A.V. Goldberg and R.E. Tarjan, “A New Approach to the Maximum Flow Problem,” *J. ACM*, vol. 35, no. 4, 1988, pp. 921-940.
7. R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall, Englewood Cliffs, N.J., 1993.
8. G.W. Flake, S. Lawrence, and C.L. Giles, “Efficient Identification of Web Communities,” *Proc. 6th Int’l Conf. Knowledge Discovery and Data Mining*, ACM Press, New York, 2000, pp. 150-160.
9. R. Albert, H. Jeong, and A-L. Barabási, “Diameter of the World Wide Web,” *Nature*, 9 Sept. 1999, pp. 130-131.
10. T.P. Novak and D.L. Hoffman, “Bridging the Digi-

tal Divide: The Impact of Race on Computer Access and Internet Use," *Science*, 17 Apr. 1998, p. 919.

Gary William Flake *is a research scientist at NEC Research Institute. His research interests include machine learning, data mining, and complex systems. He received a PhD in computer science from the University of Maryland, College Park. Flake is a member of the IEEE, the ACM, and DIMACS. Contact him at flake@research.nj.nec.com.*

Steve Lawrence *is a research scientist at NEC Research Institute. His research interests include information retrieval and machine learning. Lawrence received a PhD in computer science from the University of Queensland, Australia. Contact him at lawrence@necmail.com.*

C. Lee Giles *is the David Reese Professor of Information Sciences and Technology, professor of computer science and engineering, and associate director of research at the eBusiness Research Center at Pennsylvania State University. Currently, he is a consulting research scientist at the NEC Research Institute and an adjunct professor in computer and information science at the University of Pennsylvania. His research interests are Web and Internet computing, computational models of e-business, intelligent information processing and agents, and fundamental models of intelligent systems. He received a PhD in optical science from the University of Arizona. Giles is a Fellow of the IEEE and a member of the ACM, AAAI, and AAAS. Contact him at giles@research.nj.nec.com.*

Frans M. Coetzee *is currently the chief technical officer of GenuOne Inc. and a visiting scientist at NEC Research Institute. His research interest is signal processing, principally in the area of pattern recognition and sensor fusion. Coetzee received a PhD in electrical and computer engineering from Carnegie Mellon University. Contact him at coetzee@research.nj.nec.com.*