Self-Organization of the Web and Identification of Communities*

Gary William Flake, Steve Lawrence, C. Lee Giles, Frans M. Coetzee
{flake,lawrence,giles,coetzee}@research.nj.nec.com
NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
Phone: +1 609 951 2795 (Flake) Fax: +1 609 951 2488

Abstract

Despite the decentralized and unorganized nature of the web, we show that the web self-organizes such that communities of highly related pages can be efficiently identified based purely on connectivity. This discovery allows the identification of communities independent of, and unbiased by, the specific words used by authors. Applications include improved search engines, content filtering, and objective analysis of relationships within and between communities on the web.

1 Introduction

The existence of an increasing percentage of human knowledge and society in hyperlinked form on the web has advantages beyond the commonly stated improvements to information access. The potential for analysis of interests and relationships within science and society are great. However, analysis of content on the web is difficult due to the decentralized and unorganized nature of the web. Information on the web is authored and made available by millions of different individuals, operating independently, and having a variety of backgrounds, knowledge, goals, and cultures. We show that, despite its decentralized, unorganized, and heterogeneous nature, the web self-organizes such that the link structure allows efficient identification of communities.

Identification of communities on the web is significant for several reasons. Practical applications include automatic web portals and focused search engines, content filtering, and complementing text-based searches. More importantly, global community identification allows for analysis of the entire web and the objective study of relationships within and between communities (for example, scientific disciplines or countries). Such research could provide insight into the organization and interests of sectors of society, which individual members reflect by their linking practices. For example, links between scientific disciplines may allow more timely identification of emerging interdisciplinary connections.

The web can be modeled as a graph where vertices are web pages and hyperlinks are edges. We 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, these communities are "natural" in the sense that they are collectively organized by independently authored pages. We show that the web self-organizes such that these link-based communities identify highly related pages.

In comparison to previous methods of finding related pages on the web (see the sidebar), our method retains the transparency of methods such as co-citation and bibliographic coupling in explaining why pages are members of the community, yet can identify web communities of arbitrary diameter. Our algorithm

*Published as: G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee. Self-Organization and Identification of Web Communities. *IEEE Computer*, 35(3), 66–71, 2002

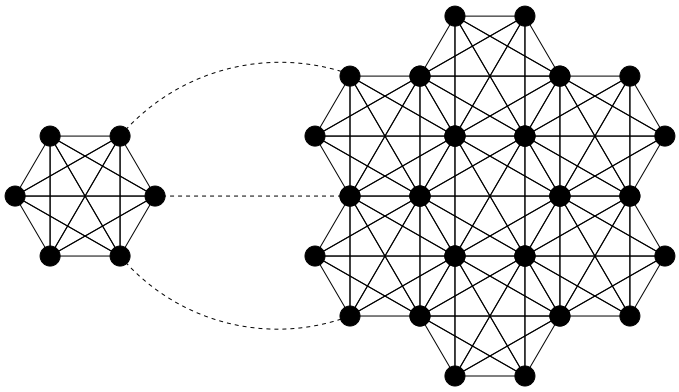

Figure 1: A simple community identification example. Maximum flow methods will separate the two subgraphs with any choice of source vertex s from the left subgraph and sink vertex t from the right subgraph, removing the three dashed links. As formulated with standard flow approaches, all community members must have at least 50% of their links inside of the community; however, additional artificial links can be used to change the threshold from 50% to any other desired threshold. Thus, communities of various sizes and with varying levels of cohesiveness can be identified and studied.

achieves this performance using only link information, without the text information used by algorithms such as HITS. In the absence of full natural language processing, the creation of an explicit link by a web author can be a stronger indication of relevance than implied links generated by the simple phrase and structure matching used by textual methods. In addition, this separation of link structure from content allows us to independently validate the performance of the link-based community estimation with content-based similarity measures.

Identifying a naturally formed community—according to our definition—is intractable in the general case because the basic task maps into a family of NP-complete graph partitioning problems [6]. However, if one assumes the existence of one or more *seed* web sites and exploits systematic regularities of the web graph [3, 8, 10], the problem can be recast into a framework that allows for efficient community identification by using a polynomial time algorithm that should scale well to studying the entire web graph.

2 Maximum Flow Communities

We recast the problem into a maximum flow framework which analyzes the flow between graph vertices. The s-t maximum flow problem [1] is defined as follows. Given a directed graph $G = (V, E)$, with edge capacities $c(u, v) \in \mathbb{Z}^+$, and two vertices, $s, t \in V$, find the maximum flow that can be routed from the source, s, to the sink, t, that obeys all capacity constraints. Intuitively, if edges are water pipes and vertices are pipe junctions, then the maximum flow problem tells you how much water you can move from one junction to another. The Max Flow-Min Cut theorem of Ford and Fulkerson [5] proves that the maximum flow of the network is identical to the minimum cut that separates s and t. Many polynomial time algorithms exist for solving the s-t maximum flow problem [7].

Figure 1 shows the basic intuition of our approach. We choose one or more seed sites to play the role of the source vertex. We require that the sum total of edges connected to the seed sites be greater than the size of the cut set (the dashed edges in Figure 1). If this constraint is not met, then our procedure will only identify a subset of the community with the worst case being that only seed sites will be discovered as being in the community.

One could imagine using an approximate centroid of the web graph (e.g., Yahoo!) as the sink; however,

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

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

Table 1: Algorithms for identifying web communities. EXACT-FLOW-COMMUNITY augments the web graph in three steps: an artificial source, s, is added with infinite capacity edges routed to all seed vertices in S; each preexisting edge is made bidirectional and rescaled to a constant value k; and all vertices except the source, sink, and seed vertices are routed to the artificial sink with unit capacity. After augmenting the web graph, a residual flow graph is produced by a maximum flow procedure. All vertices accessible from s through non-zero positive edges form the desired result and satisfy our definition of a community. APPROXIMATE-FLOW-COMMUNITY takes a set of seed web sites as input, crawls to a fixed depth including inbound hyperlinks as well as outbound hyperlinks (with inbound hyperlinks found by querying search engines), applies EXACT-FLOW-COMMUNITY to the induced graph from the crawl, ranks the sites in the community by the number of edges each has inside of the community, adds the highest ranked non-seed sites to the seed set, and iterates the procedure. The first iteration may only identify a very small community; however, when new seeds are added, increasingly larger communities can be identified. Note that k is heuristically chosen.

our method works without an explicit sink site via graph augmentation as described in Table 1. See [4] for the corresponding theorem and proof.

If one has access to the entire web graph, then EXACT-FLOW-COMMUNITY will return a set of web pages that obeys our definition of a community because the maximum flow procedure is guaranteed to always find 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 of the community; otherwise, a more efficient cut would have been to move the web site in question to the non-community.

In EXACT-FLOW-COMMUNITY, the artificial sink is generic in the sense 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 relatively disconnected externally to the rest of the graph.

Table 1 also shows an approximate version of the approach, APPROXIMATE-FLOW-COMMUNITY, which uses a subset of the web graph found by a fixed depth crawl that follows both inbound and outbound hyperlinks. Results are improved on each iteration by reseeding the algorithm with additional web sites found in the earlier steps. Our experimental results were found with the approximate version. However, we also note that the dynamic nature of the web can be exploited with a simpler iterative approximate algorithm that tests for new candidate community members by counting the number of candidate links that fall within the preexisting community.

Francis Crick Community	
Score	**Site Title or Description**
80	*Biography of Francis Harry Compton Crick* (Nobel Foundation)
79	*Biography of James Dewey Watson* (Nobel Foundation)
51	*The Nobel Prize in Physiology or Medicine 1962* (Nobel Foundation)
50	*Biographical Sketch of James Dewey Watson* (Cold Spring Harbor Lab.)
41	*A structure for Deoxyribose Nucleic Acid* (Nature, April 2, 1953)
⋮	
1	*Felix D'Herelle and the Origins of Molecular Biology* (Amazon.com)
1	Biography of Gregor Mendel
1	*Magazine: HMS Beagle Home*
1	*The Alfred Russel Wallace Page*
1	*U.S. Human Genome Project 5 Year Plan*

Stephen Hawking Community	
Score	**Site Title or Description**
85	*Professor Stephen W. Hawking's web pages*
46	*Stephen Hawking's Universe* at PBS
17	*The Stephen Hawking Pages*
15	*Stephen Hawking Builds Robotic Exoskeleton* (parody at *the Onion*)
14	*Stephen Hawking and Intel*
⋮	
1	*Did the cosmos arise from nothing?* MSNBC story
1	Spanish page for *Stephen Hawking's Universe*
1	*Relativity Group at DAMTP, Cambridge*
1	*Millennium Mathematics Project*
1	*Particle Physics Education and Information Sites*

Ronald Rivest Community	
Score	**Site Title or Description**
86	*Ronald L. Rivest : Home Page*
29	*Chaffing and Winnowing: Confidentiality without Encryption*
20	Thomas H. Cormen's home page at Dartmouth
9	*The Mathematical Guts of RSA Encryption*
8	German news story on Cryptography
⋮	
1	Phil Zimmermann's PGP web page
1	*A Very Brief History of Computer Science*
1	*Cormen / Leiserson / Rivest: Introduction to Algorithms*
1	*Security and Encryption Links*
1	*HotBot Directory: Computers & Internet, Computer Science, People: R*

Table 2: Sample results from community identification: The top five and bottom five pages (with ties) are shown for each community. The scores are the total number of inbound and outbound links that a web page has to other pages that are also in the community. Lower ranked pages often will not contain the name of the scientist used as the initial seed, yet they usually are highly topically related to the seed scientist.

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

Table 3: The fifteen most significant text features for each community, sorted in descending order of the Kullback-Leibler metric. A feature is either a word or consecutive word-pair. To extract features, all punctuation is removed, all uppercase letters are converted to lowercase, and extra white space is removed. Although only link information is used to identify the communities, the individual pages within each community are highly topically related.

3 Experimental Results

To test the approximate community identification algorithm, we used the personal home pages of three prominent scientists as a single seed in three separate runs: Francis Crick, Stephen Hawking, and Ronald Rivest. 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; hence, a considerable number of web pages were pruned to produce the communities.

Table 2 shows sample web pages within the communities. On visual inspection the majority of web pages found were highly topically related and in non-trivial ways. For example, the Crick community contained many references to Darwin, the Human Genome Project, and Rosalind Franklin; the Hawking community contained many sites dealing with cosmology, relativity, and Cambridge University; and the Rivest community contained numerous encryption web sites along with sites focused on his co-authors.

Table 3 gives a more complete characterization of the three communities. We extracted all text features from the pages within a community and for ten thousand randomly chosen web pages. We then sorted all features in the community according to their ability to separate community pages from non-community pages, as measured by the Kullback-Leibler metric. Thus, the features shown in Table 3 can be interpreted as the most useful features for separating community pages from non-community pages. As can be seen, the extracted features support our hypothesis that linked-based communities are topically related.

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

In comparison, simple breadth-first crawl strategies lose topical relevance very quickly. For the three scientists we investigated, only about 10% of pages at a depth of two from the seed site match the classification rules given above. 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 [2].

4 Conclusion

Based only on the self-organization of the link structure of the web, we are able to efficiently identify highly topically related communities, individual members of which may be spread over a very large area of the web graph. Since our method is completely divorced from text-based approaches, identified communities can be used to infer meaningful text rules and to augment text-based methods.

Applications of our method include the creation of improved search engines, content filtering, and objective analysis of the content of the web and relationships between communities represented on the web. Such analysis, taking into account issues such as the "digital divide" [9], may help improve our understanding of the world.

Acknowledgments

We thank Inktomi for the random URL data.

References

[1] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows : Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] R. Albert, H. Jeong, and A.-L. Barabasi. Diameter of the World Wide Web. *Nature*, 401:130–131, 1999.

[3] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[4] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proc. 6th Int. Conf. on Knowledge Discovery and Data Mining*, pages 150–160, 2000.

[5] L. R. Ford Jr. and D. R. Fulkerson. Maximal flow through a network. *Canadian J. Math.*, 8:399–404, 1956.

[6] M. R. Garey and D. S. Johnson. *Computers and intractability: A guide to the theory of NP-completeness*. W. H. Freeman, New York, 1979.

[7] Andrew V. Goldberg and Robert E. Tarjan. A new approach to the maximum flow problem. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, pages 136–146, Berkeley, California, 28–30 May 1986.

[8] Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong regularities in World Wide Web surfing. *Science*, 280(5360):95–97, 1998.

[9] T.P. Novak and D.L.Hoffman. Bridging the digital divide: The impact of race on computer access and Internet use. *Science*, 281:919, 1998.

[10] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

SIDEBAR: Finding related pages on the web

Previous link-based research for identifying collections of related pages includes bibliometric methods such as co-citation and biliographic coupling [5], the PageRank algorithm [2], the HITS algorithm [7], bipartite subgraph identification [8], Spreading Activation Energy (SAE) [9], and others [6, 3].

Co-citation, bibliographic coupling, and bipartite subgraph identification are localized approaches in the sense that they seek to identify well-defined graph structures that exist inside of a narrow region of the web graph. PageRank, HITS, and SAE, are more global since they 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 hub-like a web page is (HITS), or how "close" a candidate page is to a

starting region (SAE). PageRank and HITS are related to spectral graph partitioning [4] and therefore seek 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 spreading activation energy, which yields results that are extremely sensitive to the choice of parameters [9].

Localized approaches are appealing in that the identified structures unambiguously have the properties that the algorithms were designed to find. 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 operate on large subsets of the web graph and, therefore, can identify large collections of web pages that are related or valuable. However, because these methods are based on spectral graph partitioning, it is often difficult to understand and defend the inclusion of a given page in the collections that these algorithms produce. In practice, meaningful results are only achieved by HITS and PageRank when textual content is used for either preprocessing (HITS) or postprocessing (PageRank); without auxiliary text information, both PageRank and HITS have limited success in identifying collections of related pages [1].

References

[1] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in hyperlinked environments. In *Proc. 21st Int. ACM SIGIR Conf.*, 1998.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. 7th Int. World Wide Web Conf.*, 1998.

[3] S. Chakrabarti, M. van der Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Proc. 8th Int. World Wide Web Conf.*, 1999.

[4] F. Chung. *Spectral Graph Theory*. CBMS Lecture Notes. Amer. Math. Soc., 1996.

[5] E. Garfield. *Citation Indexing: Its Theory and Application in Science*. Wiley, New York, 1979.

[6] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proc. 9th ACM Conf. on Hypertext and Hypermedia*, 1998.

[7] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, 1998.

[8] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. In *Proc. 8th Int. World Wide Web Conf.*, 1999.

[9] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*. ACM Press, 1996.