REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT DATE	3. DATES COVERED *(From - To)*
07/31/2002	Final Technical Report	May 1999 - April 2002

4. TITLE AND SUBTITLE
Algorithms for Networks and Link-Structured Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-99-1-0463

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kleinberg, Jon

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Cornell University
4130 Upson Hall
Ithaca, NY 14853

8. PERFORMING ORGANIZATION REPORT NUMBER
35498

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Ballston Centre Tower One
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; distribution is Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The research has focused on the design of algorithms for networks and the information resources that they support. Link analysis algorithms have been developed with applications to the design of World Wide Web search engines. A collection of models and algorithms, have been developed for decentralized search in so-called `small-world' networks; this has led to applications in both social network analysis and in the design of peer-to-peer computing systems. Techniques from these small-world algorithms have been extended to form the basis of distributed gossip algorithms for the robust dissemination of information through a network.
Algorithms have also been developed for network routing and design. The problem of provisioning a virtual private network has been approached through connections with multicommodity flow, leading to approximation algorithms with strong provable performance guarantees. The notion of fairness in resource allocation has been formalized, leading to algorithms that approximate the fairest allocations for a range of network problems. Algorithms have also been designed that deal with input continuously over time, in an event-driven fashion, for the problems of packet routing and load balancing. Finally, combinatorial algorithms have been developed for clustering and partitioning of networks; such algorithms are useful as primitives in a variety of network analysis tasks.

15. SUBJECT TERMS
Discrete optimization, approximation algorithms, networks, search, link analysis.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Esha Molette
U	U	U	UU	5	19b. TELEPONE NUMBER *(Include area code)* (607) 254-8948

ONR Young Investigator Final Annual Report

Title: Algorithms for Networks and Link-Structured Data

PI Name: Jon Kleinberg

Address: Dept. of Computer Science,
Cornell University,
Ithaca, NY 14853

Phone number: 607-255-7316

Fax Number: 607-255-4428

Email address: kleinber@cs.cornell.edu

Award Number: N00014-99-1-0463

Web site: http://www.cs.cornell.edu/home/kleinber

Long term goals:

The overall goal of this research has been to develop algorithmic techniques that are applicable to problems at the interface of networks and information resources. This includes problems related to communication networks such as the Internet, where network design and routing are central issues; it also includes problems related to information resources such as the WWW that are built on the Internet, where one can formulate problems of information discovery, data mining, and clustering in a combinatorial framework. In the latter case, the research has developed connections with social network analysis.

Objectives:

The goal of this work is the design of methods for optimization and resource allocation in networks; for disseminating information in large decentralized networks; and for identifying clusters and important resources in a networked information environment. Underlying this is the goal of new models for understanding the dynamics by which information is propagated in networks, and the processes by which networks such as the Web evolve over time.

Approach:

The research has considered algorithmic problems in a number of settings. Communication networks have been modeled using the tools of combinatorial optimization, and have incorporated notions of fairness in allocation and resilience against failures. On-line information resources have been modeled both at the level of their network topology -- as in the case of the Web graph – and their temporal properties, where characteristic patterns of burstiness can be used to identify events and episodes.

The algorithms developed as part of this research have been rooted in discrete optimization. Randomization has been extensively employed, as well as the design of approximation algorithms with provable performance guarantees. Algorithms have also been developed in event-driven models, where input arrives continuously over time and must be processed without knowledge of the future. Clustering and data mining algorithms for network data have been developed using a combination of linear algebraic and combinatorial techniques.

Results:

Since the most recent report on this research, I have extended and generalized my results on searching in small-world networks. In particular, I have shown that decentralized search is possible in network models significantly more general than previous results had suggested; the new models draw further connections both to social network analysis and peer-to-peer computing. The applications of small-world models to distributed gossip algorithms have been developed further as well. With my student David Kempe, I have shown how gossip algorithms based on small-world techniques can be used for distributed packet routing and the distributed construction of spanning trees. We have also shown that certain aspects of the small-world approach are in a sense necessary, by establishing a set of impossibility results for approaches based on more traditional gossip mechanisms.

With Elliot Anshelevich and David Kempe, I have developed algorithms for the distributed load balancing. In the dynamic load balancing problem, we seek to keep the job load roughly evenly distributed among the processors of a given network. The arrival and departure of jobs is modeled by an adversary restricted in its power. Work by Muthukrishnan and Rajaraman in 1998 gave a clean characterization of a restriction on the adversary that can be considered the natural analogue of a cut condition. They proved that a simple local balancing algorithm first proposed by Aiello et. al. in 1993 is stable against such an adversary if the insertion rate is restricted to be a fraction strictly less than 1 of the cut size. They left as an open question whether the algorithm is stable at rate 1.

In our work, we have resolved this question positively, by proving stability of the local algorithm at rate 1. Our proof techniques are very different from the ones used by Muthukrishnan and Rajaraman, and yield a simpler proof and tighter bounds on the difference in loads. In addition, we introduce a multi-commodity version of this load-balancing model, and show how to extend the result to the case of balancing two different kinds of loads at once (obtaining as a corollary a new proof of the 2- commodity Max-Flow Min-Cut Theorem).

I have also developed models and algorithms for the analysis of temporal structure in on-line data. A particular application is that of mining document streams -- collections of text that arrive continuously over time. E-mail and news articles are two natural examples of such streams, each characterized by topics that appear, grow in intensity for a period of time, and then fade away. The published literature in a particular research field can be seen to exhibit similar phenomena over a much longer time scale. Underlying much of the text mining work in this area is the following intuitive premise -- that the appearance of a topic in a document stream is signaled by a ``burst of activity,'' with certain features rising sharply in frequency as the topic emerges. In recent work, I have proposed a formal approach for modeling such ``bursts,'' in such a way that they can be robustly and efficiently identified, and can provide an organizational framework for analyzing the underlying content. The approach is based on modeling the stream using an infinite-state automaton, in which bursts appear naturally as state transitions. The resulting algorithms are highly efficient, and yield a nested representation of the set of bursts that imposes a hierarchical structure on the overall stream. Experiments with e-mail and research paper archives suggest that the resulting structures have a natural meaning in terms of the content that gave rise to them. The work has interesting connections to topic detection and tracking, which has analyzed document streams comprised of news articles; to work in queueing theory, which provides some of the basic models used in the algorithms developed here; and to work in temporal data mining, time-series analysis, and traffic analysis, which seeks patterns in long sequences of time-indexed data.

Impact/Applications:

The research on Web structure and analysis has had impact on the design of search engines. The research on small-world networks has been applied to peer-to-peer search algorithms in decentralized settings. It has also served as part of the motivation for new research in social network analysis by Watts et al. (Columbia Dept. of Sociology), who are investigating the experimental methodology behind Milgram's original `six degrees of separation' study. The

research on gossip algorithms is proceeding via an interaction with distributed systems research groups at Cornell led by Ken Birman and Robbert van Rennesse. Finally, the research on temporal data mining of document streams is being applied to data from the physics e-Print arXiv housed
at Cornell.

Transitions:

Teoma (www.teoma.com), a search engine owned by by AskJeeves, is a recent example of a public search engine using algorithms based on the framework of hubs and authorities. CiteSeer, one of the largest scientific digital libraries, performs link analysis based on hubs and authorities as well. Through my student Amit Kumar, there has been close interaction between our work on network algorithms and a network management group at Lucent Bell Labs; Amit has graduate this spring, and will be spending the coming academic year at Bell Labs.

Related Projects:

In the area of Web search and structure analysis, work is being done by groups including Tomkins et al. and Broder et al. at IBM, Henzinger et al. at Google, Raghavan et al. at Verity, Lawrence et al. at NEC, Giles et al. at Penn State, and Chung et al. at UCSD. Algorithms for network optimization and routing are being investigated by groups at including several at Lucent Bell Labs and AT&T Research; Karger et al. at MIT; Charikar et al. at Princeton; Plotkin et al. at Stanford; Awerbuch et al. at Johns Hopkins; Upfal et al. at Brown; and Karp, Papadimitrou, and Shenker at Berkeley and ACIRI.

Publications (since most recent report):

J. Kleinberg,
``Bursty and Hierarchical Structure in Streams,"
Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining,
2002.

D. Kempe, J. Kleinberg,
``Protocols and Impossibility Results for Gossip-Based Communication Mechanisms,"
Proc. 43rd IEEE Symposium on Foundations of Computer Science, 2002.

E. Anshelevich, D. Kempe, J. Kleinberg,
``Load balancing in dynamic adversarial systems,"
Proc. 34th ACM Symposium on Theory of Computing, 2002.

J. Kleinberg,
``Small-World Phenomena and the Dynamics of Information,"
Advances in Neural Information Processing Systems (NIPS) 14, 2001.

J. Kleinberg, S. Lawrence,
``The Structure of the Web,"
Science 294(2001), 1849.

A. Blum, A. Kalai, J. Kleinberg,
``Admission Control to Minimize Rejections,"
Proc. 7th International Workshop on Algorithms and Data Structures, 2001.

D. Callaway, J. Hopcroft, J. Kleinberg, M. Newman, S. Strogatz,
``Are randomly grown graphs really random?"
Physical Review E 64, 041902 (2001).

J. Kleinberg, Y. Rabani, E. Tardos.
``Fairness in Routing and Load Balancing."
Journal of Computer and System Sciences 63(2001).

J. Kleinberg, A. Kumar,
``Wavelength Conversion in Optical Networks,"
Journal of Algorithms 38(2001).