The Mathematical Work of Jon Kleinberg

Gert-Martin Greuel, John E. Hopcroft, and Margaret H. Wright

The *Notices* solicited the following article describing the work of Jon Kleinberg, recipient of the 2006 Nevanlinna Prize. The International Mathematical Union also issued a news release, which appeared in the October 2006 issue of the *Notices*.

The Rolf Nevanlinna Prize is awarded by the International Mathematical Union for "outstanding contributions in mathematical aspects of information sciences". Jon Kleinberg, the 2006 recipient of the Nevanlinna Prize, was cited by the prize committee for his "deep, creative, and insightful contributions to the mathematical theory of the global information environment".

Our purpose is to present an overall perspective on Kleinberg's work as well as a brief summary of three of his results:

- the "hubs and authorities" algorithm, based on structural analysis of link topology, for locating high-quality information on the World Wide Web;
- methods for discovering short chains in large social networks; and
- techniques for modeling, identifying, and analyzing bursts in data streams.

Kleinberg's Nevanlinna citation includes two other areas in which he has made important contributions: theoretical models of community growth in social networks and the mathematical theory of clustering. Readers interested in learning more about his work on these topics should consult the list of papers given on his home page [8].

Gert-Martin Greuel is professor of mathematics at the Universität Kaiserslautern, Germany, and director of the Mathematical Research Institute Oberwolfach. His email address is greuel@mathematik.uni-kl.de.
John E. Hopcroft is IBM Professor of Engineering and Applied Mathematics in Computer Science at Cornell University. His email address is jeh@cs.cornell.edu.
Margaret H. Wright is professor of computer science and mathematics and chair of the computer science department at the Courant Institute of Mathematical Sciences, New York University. Her email address is mhw@cs.nyu.edu.

Find the Mathematics; Then Solve the Problem

Kleinberg broadly describes his research [8] as centering "around algorithmic issues at the interface of networks and information, with emphasis on the social and information networks that underpin the Web and other on-line media". The phenomena that Kleinberg has investigated arise from neither the physical laws of science and engineering nor the abstractions of mathematics. Furthermore, the networks motivating much of his work lack both deliberate design and central control, as distinct from (for example) telephone networks, which were built and directed to achieve specific goals. He has focused instead on networks that feature very large numbers of unregulated interactions between decentralized, human-initiated actions and structures.

A striking motif in Kleinberg's research is his ability to discern and formulate plausible mathematical structures to describe problems that represent vague, even elusive, human goals. Some of his most brilliant work has begun by asking questions that might seem initially to have no clear answers—"What do people really want from a Web query?", "How can individuals find short paths in a social network using only local information?"—and then coming up with mathematical insights that illuminate the important features of reality. Once he has the mathematical definitions in hand, Kleinberg goes on to create powerful solution techniques that are both mathematically elegant and successful in practice.

It is difficult to overstate the impact of Kleinberg's work on several major real-world problems, and for this reason alone his work is well known to the broad scientific and technological community.

Two additional reasons for Kleinberg's high visibility are that his two best known results can be grasped intuitively without invoking details of the underlying mathematics and that he is a master expositor, with a much-admired ability to motivate and explain his work so that readers can follow his logic every step of the way.

Hubs and Authorities

There is no better way to appreciate Kleinberg's signature style than reading the journal version of his most famous paper, "Authoritative sources in a hyperlinked environment" [2]. The paper opens with a discussion of an important but imprecisely defined problem—finding the "most relevant" webpages in response to a given broad query. Beyond the quandary of how to define "most relevant" in a meaningful way, Kleinberg notes that the difficulty with broad queries is the vast overabundance of possibly relevant hits, so that what is needed is an automated way to filter out the most "definitive" pages.

Starting with what seem at first to be the obvious solutions, he carefully explains the impossibility of using purely internal features of a page to rate its authority, as well as the flaw in relying on the query words themselves. The next part of the paper motivates and proposes the nonobvious concept of using a link-based model of the graph representing the Web to create an algorithm for deciding which pages are "authoritative".

His method for finding these pages begins by constructing a small, focused subgraph G_σ of the Web, where σ is the query string, using link structure to identify its strong authorities. Along the way, Kleinberg explains how, in addition to finding highly authoritative pages, we would expect to find *hub pages*, i.e., those that contain links to many relevant authoritative pages and thereby allow unrelated pages to be discarded. Based on a natural equilibrium between hubs and authorities, a novel iterative algorithm is defined that updates numerical weights for each page until a fixed point is reached. Happily, Kleinberg also shows that a similar process and algorithm can be adapted when seeking "similar" webpages.

A feature of Kleinberg's algorithm that brought joy to fans of linear algebra is that the desired authority-weight and hub-weight vectors form, respectively, the principal eigenvectors of A^TA and AA^T, where A is the adjacency matrix of the graph of a collection of linked pages. As well, the nonprincipal eigenvectors of these matrices can be used to extract additional densely linked collections of hubs and authorities.

Always a careful scholar, Kleinberg provides a summary of previous approaches to a variety of related problems: measuring "standing" in social networks, "impact" in scientific citations, ranking of Web pages, hypertext document retrieval, clustering of explicitly linked moderate-size structures, and spectral graph partitioning. A fascinating historical note is his discussion of the then recently published Brin–Page page-rank algorithm [1], soon thereafter to become the basis of Google.

Near the end of the paper (which was published before the rise of Google), Kleinberg surveys three user studies designed to evaluate the effectiveness of his algorithm. These reported favorable results concerning Web users' perception of improved quality in their query results, but he cautions that such an evaluation is a challenging task because individual judgments of relevance are inherently subjective.

How Can It Be a Small World?

Stanley Milgram's social psychology experiments in the 1960s (see, for example, [6]) reported on and popularized the idea of a "small world"—that any two individuals who are apparently far apart in a social network can find "short paths" to reach one another. To model the mathematical properties required to ensure the existence of short paths, Watts and Strogatz [7] proposed a "superposition" structure in which a relatively small number of random long-range links are added to a high-diameter network with edges at each node representing local social links. The long-range links provide the opportunity for a short chain through the entire network. Informally, a "small-world" network is an n-node graph such that almost all pairs of nodes are connected by chains whose length is a polynomial in $\log n$, i.e., the number of links traversed to reach one node from another is likely to be exponentially smaller than the number of nodes.

In his fundamental paper "The small-world phenomenon: an algorithmic perspective" [3], Kleinberg noted that Milgram's findings not only indicated the surprising existence of short chains, but also revealed an equally surprising result about the existence of *algorithms* that would enable an arbitrary person, knowing only information about the locations of his/her individual acquaintances, to construct a short communication path to a target stranger.

To begin his analysis of small-world models, Kleinberg first proved a negative result—that, using the Watts-Strogatz model, no *decentralized* algorithm, meaning one whose decisions are based solely on local information, can produce paths of small expected length relative to the diameter of the network.

Next, he generalized the Watts-Strogatz model into an infinite family of random network models. Starting with two parameters characterizing a

node's local and long-range contacts, the model can be simplified to a one-parameter family with an associated clustering exponent α that represents the probability of a long-range connection between two nodes as a function of their lattice distance.

When there is a uniform distribution over long-range contacts (corresponding to $\alpha = 0$), Kleinberg showed that, although short paths exist with high probability, a decentralized algorithm cannot find them efficiently, since the expected time is exponential in the expected minimum path length. In effect, the long-range links are "too random" to be useful to a decentralized algorithm. Although larger values of α allow a decentralized algorithm to take advantage of the structure of the long-range contacts, they become less useful in transmitting the message to an arbitrary far-away node.

The central result is that there is a unique value of α ($\alpha = 2$) for which a decentralized algorithm (using a greedy heuristic) will find a target node in a number of steps bounded by a polynomial in $\log n$. Furthermore, no efficient decentralized algorithm exists for any other value of α. These results generalize from two-dimensional to d-dimensional lattices, with the critical value of α equal to the dimension d. A later paper by Kleinberg [4] includes an extension of these results to other network models, including hierarchical models or models based on set systems.

His work on small-world phenomena has had a direct effect on the design of peer-to-peer systems and focused Web crawling techniques, and it has also inspired numerous papers by other authors. Using one of Kleinberg's contributions to illustrate the influence of the other, submission of the combination of "small world" and "Kleinberg" to Google on February 11, 2007, produced nearly 52,000 hits!

Word Bursts and Temporal Analysis

In his 2002 paper [5], Kleinberg considers the problem of extracting meaningful information from document streams (such as email messages or news articles) that arrive continuously over time—in particular, spotting the "burst of activity" that signals the first appearance of a new topic. His stated scientific goal in this work was to devise a mathematical model that allows such bursts to be identified efficiently, with a further aim of analyzing the content via the associated organizational framework. But, as he amusingly relates in the paper's introduction, his personal motivation (one we can all share) was a wish to find an organizing principle based on *time* rather than topic for his ever-increasing volume of accumulated email.

The model proposed by Kleinberg—an infinite-state automaton in which bursts are state transitions—is conceptually related to queueing theory models of bursty traffic. In the most basic form of his model, the gap in time between two consecutive events (messages) is given by an exponential density function such that the expected gap between messages is $1/\alpha$ for some $\alpha > 0$, where α can be interpreted as the rate of message arrivals. Bursts can be added by allowing the model to include interleaved periods with lower and higher rates; these correspond to different states for which the rate depends on the state. In the ultimate model analyzed by Kleinberg in detail, there are an infinite number of states, each denoted by q_i. The sequence $q_0, q_1, \ldots$ models inter-arrival times that decrease geometrically, and there is a cost $\tau(i, j)$ corresponding to the transition from state i to state j. Given this model, Kleinberg shows that an optimal (cost-minimizing) state sequence can be found efficiently by adapting a standard forward dynamic programming algorithm for hidden Markov models. From the results of running this algorithm, one can then define the hierarchical structure that is implicit in a sequence of bursts.

Several fascinating conclusions emerge from this work, including the fact that use of a state-transition model means that bursts are characterized by unambiguous beginnings and endings. Kleinberg notes that, when the document stream consists of email messages, the initial message at which the state transition took place can be seen as a "landmark" in subsequent extended message sequences. Later related work by Kleinberg and others has addressed traffic-based feedback on the Web and the temporal dynamics of online information streams (see [8]).

Summary

Jon Kleinberg's work perfectly fits the Nevanlinna Prize specification since, as we have seen, his mathematical insights have had wide application to multiple elements of information science—the effectiveness of advanced Web search engines, Internet routing, data mining, and the sociology of the World Wide Web. We refer the interested reader to his website [8] for further pointers to papers on these and other topics, including network analysis and management, gossip algorithms, clustering, data mining, comparative genomics, and geometric pattern matching.

Returning to our opening theme, Kleinberg's much-lauded work on information networks is characterized by (i) identifying and formulating fundamental mathematical structures in questions about the real world, (ii) defining meaningful mathematical models that represent crucial features of real-world phenomena, and finally (iii) creating effective algorithms that solve the resulting mathematical problems.

References

[1] S. Brin and L. Page, Anatomy of a large-scale hypertextual Web search engine, *Proceedings of the 7th World Wide Web Conference*, Elsevier Science B. V., Amsterdam, 1998.

[2] J. Kleinberg, Authoritative sources in a hyperlinked environment, *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, SIAM, Philadelphia, PA, 1998, 668–677 (a longer version appears in the *Journal of the ACM* **46** (1999)).

[3] _____, The small-world phenomenon: An algorithmic perspective, *Proceedings of the 32nd ACM Symposium on the Theory of Computing*, 2000.

[4] _____, Small-world phenomena and the dynamics of information, *Advances in Neural Information Processing Systems* **14** (2001).

[5] _____, Bursty and hierarchical structure in streams, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, New York, NY, 2002, 91–101.

[6] S. Milgram, The small-world problem, *Psychology Today* **1** (1967) .

[7] D. Watts and S. Strogatz, Collective dynamics of small-world networks, *Nature* **393** (1998).

[8] `http://www.cs.cornell.edu/home/kleinber`.