INFORMATION NETWORKS, LINK ANALYSIS, AND TEMPORAL DYNAMICS

(Summary of Invited Paper)

Jon Kleinberg
Cornell University

The Internet has given rise to two widespread communication media: the World Wide Web, and electronic mail. Both are sources of fearsome complexity, though in quite different ways.

Unlike other great networks of the past century — the electric power grid, the telephone system, or the highway and rail systems — the Web is not fundamentally an engineered artifact; its growth has been sudden, populist, and anarchic. The emergence of the Web has crystallized a view of large networks not just as technological creations, but as complex phenomena to be studied on their own terms. We are discovering that the Web and related information networks exhibit a characteristic 'geography'; they share a number of fundamental structural properties that presumably reflect the forces driving their growth and evolution [7, 19, 20, 27]. The study of these systems has led to methods for organizing the content of on-line document collections through analysis of their underlying link structures [6, 8, 16], and it has suggested research directions in models for large graphs [1, 3, 13, 21, 24], as well as computational perspectives on social network analysis [17, 30, 31].

E-mail has forced on us a different spectrum of problems — the personal complexity of managing a message stream that can reach a hundred pieces of mail per day, and organizing personal archives of correspondence that can easily grow to hundreds of megabytes in size. And at a still larger scale, e-mail has become the raw material for legal proceedings and historical investigation [22]. How can an algorithmic perspective suggest organizing principles for message streams of this magnitude? There has been research aimed at structuring e-mail archives by topic classification and keyword indexing [5, 9, 11, 12, 26]. A promising approach, complementary to these methods, is to make use of the tight relationship between topics and temporal dynamics — as time progresses, topics of interest are signaled by 'bursts of activity' in the stream. Using a concrete computational model for such 'bursts,' one can begin to structure the underlying content around them [18]. The resulting set of issues has interesting

R. Baeza-Yates et al. (eds.), *Foundations of Information Technology in the Era of Network and Mobile Computing*

connections to research in topic detection and tracking [2, 4, 28, 29], as well as to probabilistic models from queueing theory [14] and temporal data mining [10, 15, 23, 25].

References

[1] W. Aiello, F. Chung, L. Lu. "Random evolution of massive graphs," *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, 2001.

[2] J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, Y. Yang, "Topic Detection and Tracking Pilot Study: Final Report," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Feb. 1998.

[3] A.-L. Barabasi, R. Albert. "Emergence of scaling in random networks," *Science*, 286(509), 1999.

[4] D. Beeferman, A. Berger, J. Lafferty, "Statistical Models for Text Segmentation," *Machine Learning* 34(1999), pp. 177-210.

[5] A. Birrell, S. Perl, M. Schroeder, T. Wobber, *The Pachyderm E-mail System*, 1997, at http://www.research.compaq.com/SRC/pachyderm/.

[6] S. Brin, L. Page, "Anatomy of a Large-Scale Hypertextual Web Search Engine," *Proc. 7th International World Wide Web Conference*, 1998.

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener. "Graph structure in the Web," *Proc. 9th International World Wide Web Conference*, 2000.

[8] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, "Mining the link structure of the World Wide Web," *IEEE Computer*, August 1999.

[9] W. Cohen. "Learning rules that classify e-mail." *Proc. AAAI Spring Symp. Machine Learning and Information Access*, 1996.

[10] D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, MIT Press, 2001.

[11] J. Helfman, C. Isbell, "Ishmail: Immediate identification of important information," AT&T Labs Technical Report, 1995.

[12] E. Horvitz, "Principles of Mixed-Initiative User Interfaces," *Proc. ACM Conf. Human Factors in Computing Systems*, 1999.

[13] B. Huberman, L. Adamic, "Growth dynamics of the World Wide Web," *Nature* 401(1999).

[14] F.P. Kelly, "Notes on effective bandwidths," in *Stochastic Networks: Theory and Applications*, (F.P. Kelly, S. Zachary, I. Ziedins, eds.) Oxford Univ. Press, 1996.

[15] E. Keogh, P. Smyth, "A probabilistic approach to fast pattern matching in time series databases," *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, 1997.

[16] J. Kleinberg. "Authoritative sources in a hyperlinked environment." *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998. Extended version in *Journal of the ACM* 46(1999).

[17] J. Kleinberg. "Navigation in a Small World." *Nature* 406(2000).

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

[19] J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. "The Web as a graph: Measurements, models and methods." *Proc. Intl. Conf. on Combinatorics and Computing*, 1999.

[20] J. Kleinberg, S. Lawrence, "The Structure of the Web," *Science* 294(2001).

[21] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. "Stochastic models for the Web graph," *Proc. 41st IEEE Symposium on Foundations of Computer Science*, 2000.

[22] S.S. Lukesh, "E-mail and potential loss to future archives and scholarship, or, The dog that didn't bark," *First Monday* 4(9) (September 1999), at http://firstmonday.org.

[23] H. Mannila, M. Salmenkivi, "Finding simple intensity descriptions from event sequence data," *Proc. 7th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2001.

[24] D. Pennock, G. Flake, S. Lawrence, E. Glover, C.L. Giles, "Winners don't take all: Characterizing the competition for links on the Web," *Proc. Natl. Acad. Sci.* 99(2002).

[25] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE* 77(1989).

[26] R. Segal, J. Kephart. "Incremental Learning in SwiftFile," *Proc. Intl. Conf. on Machine Learning*, 2000.

[27] S. Strogatz, "Exploring complex networks," *Nature* 410(2001).

[28] R. Swan, J. Allan, "Automatic generation of overview timelines," *Proc. SIGIR Intl. Conf. on Research and Development in Information Retrieval*, 2000.

[29] R. Swan, D. Jensen, "TimeMines: Constructing Timelines with Statistical Models of Word Usage," *KDD-2000 Workshop on Text Mining*, 2000.

[30] D. Watts, *Small Worlds*, Princeton University Press, 1999.

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