Editor's Preface
A brief overview of network algorithms

In the rapidly growing world of Internet infrastructures, we face many challenging new algorithmic problems. These arise in part because the usual assumptions made in problems of this general type may no longer hold. For example, many typical questions deal with massive data sets and huge networks of prohibitively large sizes so that the (exact) number of nodes or edges is no longer a useful parameter. Instead, only partial information can be obtained. In addition, the network and the data are evolving dynamically. For example, the number of Internet hosts as of July 2005 topped 353 million and the number of Web pages indexed by large search engines now exceeds 8 billion [24].

Driven by the need of current technology and guided by the myriad of existing examples of large scale networks, the area of *network algorithms* has been rapidly expanding and evolving into a new and rich field. This field straddles both theoretical computer science and networking through the strong interplay between theory and practice. Furthermore, this field is closely connected to a variety of areas including graph theory, game theory, probability, statistics, physics, bioinformatics and the social sciences. Through this connection, a wealth of knowledge and methodology from diverse areas have been brought into play. As a result, new tools and insights have been developing and enriching all these relevant areas as well.

To help foster research in this area, we include here a list of topics in network algorithms, together with brief discussions.

1. **Network structures**
   What are the 'laws' that govern discrete structures such as the Internet? For a massive network, is it possible to extract relatively few parameters to predict its shape and structure ? How are the structural properties related to various behaviors of the network? Why are most existing networks so 'similar' to each other? Are there simple *local* growth rules that dictate the *global* behavior of the network?

   In spite of the complexity of these problems, a great deal of progress has been made. It has been observed that numerous networks, including the Web graph, the telephone graph, the collaboration graph, among others [1, 2, 3, 6, 17], all satisfy a *power law*, which asserts that the number of nodes of degree k is proportional to $k^{-\beta}$ for some positive β. To model power law graphs, various versions of random graphs come into play — including the on-line and the off-line models. (In an on-line model, a random graph grows discretely at each tick of the clock, so to speak, which is different from what happens for an off-line model.) The analysis for on-line models is of course considerably harder than that for off-line models. Nevertheless, the study of the off-line models is helpful for understanding on-line models [13].

   The on-line models are also called the generative models. The most popular model is perhaps the preference attachment scheme that generates new vertices and edges under the principle that "the rich get richer". There is a comprehensive survey by Mitzenmacher [34] of this model and the power law degree distributions that are satisfied by such randomly generated graphs. Another on-line model is the duplication model [14] that also generates random graphs with a power law distribution and is particularly suited for analyzing biological networks. There is also an interesting game-theoretical model of how the network evolves — subject to optimizing certain costs [18].

   The off-line models can be partitioned into two types. The configuration model uses random matchings and then groups the vertices of the matchings into nodes of the required degrees. In spite of the dependencies of the edges, there are quite a few results known for the connectivity [35, 36] and other properties [21] of such graphs. Another off-line random graph model is a natural generalization of the classical Erdős-Rényi model. Here, each edge is randomly and independently chosen according to some probability depending on the two endpoints. Consequently, useful tools from the classical theory can be used and extended to derive results on the diameter, average distance, eigenvalues, etc. as summarized in [12].

Another approach is to find an efficient approximation of a given graph. For dense graphs, the celebrated regularity lemma gives almost 'finite' approximations (see [4, 27]). For sparse graphs, the regularity lemma does not quite work so more research is needed. The recent breakthrough of Green and Tao [20] of finding arbitrarily long arithmetic progressions in primes (or some quasi-random sequences) rests in part on a hypergraph version of the regularity lemma.

One major direction is the sparsification of graphs, i.e., approximating a given graph by a sparse graph of linear size. Spielman and Teng [41] obtain remarkable sparsification results by using a result of Lovász and Simonovits [32]. Although these results are for undirected graphs, some can be extended to directed graphs, as described in [10, 11].

2. **Caching**

   In the historical setting, caching means managing the transfer of data files or virtual memory between RAM and hard disks. The classical issues concern load balancing and traffic redirection for reducing congestion and increasing efficiency [30].

   For large networks such as the Internet, caching strategies are carried out by deploying content distribution networks (CND). One of the primary considerations for CND is the geographic location of caching. What are optimal geographic locations for the cache placement? Such problems are closely related to the k-median server problem, the k-center problem and facility location problems. Although the general versions of these problems have been shown to be NP-complete, efficient approximation algorithms have been found [9, 22, 42]. Variations of these problems restricted to topologies related to the Web and other information networks are considered in [26, 33].

   How should cache be managed? Which, where and how can decisions be made about file replication and eviction? There are different modes of cache management — each node (e.g., a server) can be treated as an isolated object or an entire set of servers acting in a coordinated manner. There are also a variety of cache algorithms that such as the least recently used (LRU), the least frequently used (LFU) and a myriad of schemes that keep track of additional parameters such as size, time and various types of correlations. Several excellent surveys give overviews of cache placement/replacement strategies [28, 38].

   Since Web servers have access to extensive statistics of Web accesses, the cache management algorithms can take full advantage of the access logs. Irani [25] examined page cache algorithms by using competitive analysis for on-line algorithms. Several refinements and improvements that further reflect the cache costs or incorporate the correlations can be found in [8, 44]. Until now, on-line analysis has mainly focused on worst case bounds, usually over all possible input sequences. Thus, the results are often overly pessimistic.

3. **Games and economics in networks**

   A large network, such as the Internet, is built and operated by a large number of entities typically with diverse economic interests in varying relationships of competition and collaboration. Very much like in a game-theoretical setting, there is a set of commonly agreed-upon rules (such as various routing or transmission protocols) without a central authority. How much is the loss of efficiency due to the lack of a central authority? In the other direction, is a given strategy, protocol or algorithm stable in spite of selfish users? Numerous problems can be formulated in a game-theoretical fashion.

In a classical game-theoretical setting, each of the players can choose among a set of strategies and there are functions which assign to each combined choice a *payoff* for each player. How should a player make his/her choice which may be countered by the choices that all other players can make. A classical concept for predicting the collective behavior of the game is the *Nash equilibrium*. In a Nash equilibrium, no player can choose a better strategy given the choices of the other players, and therefore no player has an incentive to deviate from the chosen strategy.

To route traffic in a network, there are often many alternative routes for sending messages from one node to another. Can a selfish choice of routes cause congestions in the network? How does the selfish optimization of various metrics at each node affect the overall performance of the network? How can economic incentives (such as the pricing or taxing) influence selfish behavior in networks? There have been a number of results using game theory to address various problems arising in selfish routing [16, 39] as well as a survey in [19].

In the Internet, the congestion control is governed by the Transmission Control Protocol TCP, under a simple greedy principle, 'increase the batch size by one if the previous batch of packets go through, otherwise decrease by one'. How effective is such a congestion control algorithm? Will greedy behaviors at end-points result in instability of the entire network? The TCP game of certain restricted models are analyzed in [5].

Wireless ad hoc networks consist of both wireless mobile and stationary nodes. The edges may occur when nodes are within radio range (one-hop) but not necessarily so. The edges change dynamically and sometimes arbitrarily. The de facto access standard, DCF (short for distributed coordination function), can be interpreted by various game models [43]. The game-theoretical approach leads to transmission strategies relying on the consideration of the Nash equilibrium.

For web caching, servers have obvious incentives to maximize the benefit in their own domain, possibly at the expense of the global optimum. The selfish caching problem is the game version of this problem. In [15], various versions of the selfish caching games are analyzed again using the concept of Nash equilibrium.

Although the existence of the Nash equilibrium can be assured under certain general convexity conditions, the computational complexity of finding the Nash equilibrium is wide open. It belongs to a class of problems 'between P and NP'. As stated by Papadimitriou [5]: "Together with factoring, the complexity of finding a Nash equilibrium is in my opinion the most important concrete open question on the boundary of P today." Recently, Papadimitriou [37] gave a polynomial time algorithm for computing the Nash equilibrium for mixed strategies under the additional assumption that the input of the game has a *succint* representation. The class of succinct games includes numerous versions of games for facility location, network design, scheduling and congestion games. Still open is the problem of computing a mixed Nash equilibrium even in a two person game [31, 37, 40].

4. **Information retrieval**

To find a needle in the haystack of massive data remains as a major challenge in this information age. The tasks are both theoretical and algorithmic. A first step is to take advantage of the relationships among the data and to form an information network. The Web crawler views the Web as a graph with Web pages as nodes and hyperlinks as edges. The crawler retrieves Web pages off-line and the indexer builds the inverse index that allows a query handler to answer user queries on-line. There has been a great deal of progress in this area although numerous challenges still remain. There are three excellent surveys on this subject.

In the survey by Henzinger [23], six algorithmic problems are posted and are only partially solved:

(1) uniformly sampling of web pages,
(2) modeling the web graph,
(3) finding duplicate hosts,
(4) finding top gainers and losers in data streams;
(5) finding large dense bipartite subgraphs;
(6) understanding how eigenvectors partition the Web.

A second survey by Langville and Meyer [29] is a comprehensive survey of many issues associated with PageRank. There has been a great deal of work since Kleinberg first used spectral analysis to distinguish *hubs* from *authorities*. Since then, the spectral analysis of the Web graph has become an active subfield. There is a great deal of interplay between spectral methods and probabilistic techniques.

A third survey is by Berkhin [7] on pagerank computing. It covers acceleration of convergence, the relevance models and personalized page rank sets among others.

Most of the analyses are based on Web graphs which represent pairwise relations. In many realistic scenarios, the relations often involve more than two objects. The combinatorial platform for such relations are hypergraphs. However, the spectral analysis for hypergraphs barely exists. Most spectral results for hypergraphs still only involve pairwise relations arising from the projections of the hypergraphs. The difficulty seems to arise from the lack of 3-dimensional matrix multiplication. Some fundamental breakthrough is needed here.

The above list of topics is by no means complete. Research in network algorithms is rapidly evolving. Here we include a number of problems and directions to function as Socratic “gadflies”. The list of references is far from exhaustive. The selection of references emphasizes recent survey papers.

Fan Chung Graham
Guest Editor for the special issue on network algorithms

References

[1] J. Abello, A. Buchsbaum, and J. Westbrook, A functional approach to external graph algorithm, *Proc. 6th European Symposium on Algorithms,* (1998), 332–343.

[2] W. Aiello, F. Chung and L. Lu, A random graph model for massive graphs, *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, (2000) 171-180.

[3] R. Albert and A.-L. Barabási, statistical mechanics of complex networks, *Review of Modern Physics*, **74**, (2002), 47-97.

[4] N. Alon, R. A. Duke, H. Lefmann, V. Rödl and R. Yuster, The algorithmic aspects of the regularity lemmas, *J. of Algorithms* **16**, (1994), 80-109.

[5] A. Akella, S. Seshan, R. Karp, S. Shenker and C. Papadimitriou, Selfish behavior and stability of the Internet: A game-theoretic analysis of TCP, *Proceedings of SIGCOMM'02*, (2002),

[6] Albert-László Barabási and Réka Albert, Emergence of scaling in random networks, *Science* **286** (1999) 509-512.

[7] P. Berkhin, A survey on pagerank computing, *Internet Mathematics*, to appear.

[8] P. Cao and S. Irani, Cost-Aware WWW Proxy Caching Algorithms, *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, (1997), 193-206.

[9] M. Charikar and S. Guha, Improved combinatorial algorithms for the facility location and k-median problems, *FOCS*, (1999), 378-388.

[10] F. Chung, Laplacians and the Cheeger inequality for directed graphs, *Annals of Combinatorics*, **9** (2005), 1-19.

[11] F. Chung, Random walks and cuts in directed graphs, preprint.

[12] F. Chung and L. Lu, *Massive Graphs*, *CBMS Lecture Notes*, AMS, to appear.

[13] F. Chung and L. Lu, Coupling online and offline analysis for random power law graphs, *Internet Mathematics*, **1** (2004), 409-461.

[14] F. Chung, L. Lu, G. Dewey and D. J. Galas, Duplication models for biological networks, *J. Computational Biology*, **10**, no. 5, (2003), 677-688.

[15] B.-G. Chun, M. Barreno, K. Chaudhuri, C. H. Papadimitriou, H. Wee and J. Kubiatowic, Selfish caching in distributed systems: A game-theoretic analysis, *PODC'04*, (2004), 21-30.

[16] R. Cole, Y. Dodis and T. Roughgarden, Pricing network edges for heterogeneous selfish users, *STOC*, (2003), 521-530.

[17] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the Internet topology, *Proceedings of the ACM SIGCOM Conference*, Cambridge, MA, (1999), 251-262.

[18] A. Fabrikant, E. Koutsoupias, C. H. Papadimitriou, Heuristically optimized trade-offs: A new paradigm for power laws in the Internet, *Preceedings of the 29th International Colloquium on Automata, Languages and Programming, Lecture Notes in Computer Science 2380*, pp. 110-122, Berlin-Heidelberg, Springer-Verlab, 2002.

[19] R. Feldmann, M. Gairing, T. Lücking, B. Monien and M. Rode, Selfish routing in non-cooperative networks: A survey, *Lecture Notes in Computer Science*, Volume 2747, (2003), 21-45.

[20] B. Green and T. Tao, The primes contain arbitrarily long arithmetic progressions, *Annals of Math*, to appear.

[21] C. Gkantsidis, M. Mihail, A. Saberi, Conductance and congestion in power law graphs, *SIGMETRICS*, (2003), 148-159.

[22] S. Har-Peled and S. Mazumdar, On coresets for k-means and k-median clustering, *STOC*, (2004), 291-300.

[23] M. R. Henzinger, Algorithmic challenges in web search engines, *Internet Mathematics*, **1** (2003), 115-123.

[24] Internet Software Consortium, http://www.isc.org/

[25] S. Irani, Page replacement with multi-size pages and applications to web caching, *Algorithmic*, 33, no. 3, July 2002, 384-409.

[26] M. Karlsson, and M. Mahalingam, Do we need replica placement algorithms in content delivery networks? *Proceedings of Web Content Caching and Distribution (WCW)*, 2002.

[27] J. Komlos and M. Simonovits, Szemerédi's regularity lemma and its applications in graph theory, *Combinatorics, Paul Erdős is Eighty*, (K. Mikilos et. al. eds.), Bolyai Society Mathematical Studies, **2**, (1996), 295-352.

[28] B. Krishnamurthy and J. Rexford, *Web Protocols and Practice: HTTP/1.1, Networking Protocols, Caching and Traffic measurement*, Addison-Wesley, Reading, MA., 2001.

[29] A. N. Langville and C. D. Meyer, Deeper inside pagerank, *Internet Mathematics*, **1** (2003-4), 335-380.

[30] F. T. Leighton, *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*, Morgan Kaufmann, San Mateo, CA, 1992.

[31] R. J. Lipton, E. Markakis and A. Mehta, Playing large games using simple strategies, *EC*, (2003), 36-41.

[32] L. Lovász and M. Simonovits, The mixing rate of Markov chains, an isoperimetric inequality and computing the volume, *FOCS*, (1990), 346-354.

[33] A. Meyerson, K. Munagala and S. Plotkin, Web caching using access statistics, *SODA*, (2001), 354-363.

[34] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, *Internet Mathematics*, **1**, (2004), 226-251.

[35] M. Molloy and B. Reed, A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, Vol. **6**, no. 2 and 3 (1995), 161-179.

[36] M. Molloy and B. Reed, The size of the giant component of a random graph with a given degree sequence, *Combin. Probab. Comput.* **7**, no. 3, (1998), 295-305.

[37] C. Papadimitriou, Computing correlated equilibria in multi-player games, STOC, (2005), 49-56.

[38] S. Podlipnig and L. Böszörmenyi, A survey of web cache replacement strategies, *ACM Computing Surveys*, (2003), 35, no. 4, (2003), 374-398.

[39] T. Roughgarden and E. Tardos, How bad is selfish routing? *Journal of the ACM*, 49, no. 2, (2002), 236-259.

[40] R. Savani and B. von Stengel, Exponentially many steps for finding a Nash equilibrium in a bimatrix game, *FOCS*, (2004), 258-267.

[41] D. Spielman and S.-H. Teng, Nearly linear time algorithms for graph partitioning, graph sparsification and solving linear systems, *STOC*, (2003), 81-90.

[42] M. Thorup, Quick k-median, k-center, and facility location for sparse graphs, *SIAM J. Comput.*, **34**, (2005), 405-432.

[43] Y. Xiao, X. Shan and Y. Ren, Game theory models for IEEE 802.11 DCF in wireless ad hoc networks, *IEEE Radio Communications*, (2005), S22-S26.

[44] N. Young. On-line file caching, *Algorithmica*, 33, (2002), 371-383.