Milena Mihail

Research Statement

Complex networks is an exciting research area whose study includes networks that find crucial applications in science and technology, including the Internet, the WWW, Peer-to-Peer networks, ad-hoc networks, as well as networks arising in molecular biology and the social sciences. My main contribution in this area is in isolating and quantifying network metrics which have direct impact on network *function* and network *performance*.

My work in this area brings the celebrated theory of *algebraic methods* and *expander graphs* into the study of complex networks. Over the years, regular expander graphs have been used in theory and networking to construct networks with excellent sampling, congestion and non-blocking properties. Algebraic methods have provided certificates of expansion (such as eigenvalues) and primitives of clustering (found in principal eigenvectors). The power of algebraic methods derives from the computational efficiency with which principal eigenvalues and eigenvectors can be computed, even for very large data sets. My work on complex networks extends the theory of regular expanders to graphs with arbitrary degree sequences, such as power-law graphs observed in many current communication, information and social networks. I have shown that several power-law models and other heavy-tailed models posses properties analogous to strong expansion, thus performance of fundamental processes running on these networks, such as *routing* or *searching*, become well characterized. I have also shown how the use of eigenvectors for spectral clustering can be adjusted in graphs with arbitrary degree sequences. The networking significance of this work is that it helps define extremal traffic patterns (intracluster and intercluster) which are important in network simulation.

On the complementary side, my work also aims to develop efficient distributed algorithms that enhance a network topology with good expansion, hence improve the performance of protocols running on these networks. In very recent work, I am also developing distributed algorithms that enhance a network with topology awareness. For example, in a completely local and asynchronous way, we wish for links to "know" if they lie across critical network cuts.

Expansion, conductance and spectral gap are tools that I had previously worked on in the context of convergence rates of *Markov chains*, around which my early research was centered. I was one of a small handful of researchers who, in the late 80's pioneered these studies, and all my early work is highly quoted to the current date. Since then, Markov chains have found a wide range of important applications in sampling and counting of combinatorial structures, in statistical physics, in software testing by simulation and, most recently, in fundamental Internet issues related to information retrieval, content distribution and performance characterization.

My research since the early 90's can be classified as applications driven theory. Beyond complex networks, I have worked on many problems arising in classical telecommunication networks, such as cost effective and reliable *network design* and multiwavelength *optical networks*. One important contribution of my work was in bridging the gap between sophisticated algorithmic ideas developed by theoreticians, and their utilization in practical situations. My work on network design spanned all the way from clean mathematical formulations of problems arising in practice; design of algorithms for these problems; modification of these algorithms into pragmatic heuristics that address the exact practical problem; coding up the solution; and all the way to deploying commercial products.

1. Complex Networks and Large Scale Data

1.1 Scaling, Performance and Algorithms in Complex Networks: Expanders are sparse, regular and highly homogeneous graphs supporting routing with near-optimal throughput and congestion. Expanders have been studied extensively in theory and used in communication networks by providing regular topologies with very good congestion, throughput and non-blocking properties. How does congestion and throughput scale in sparse inhomogeneous topologies, such as power law graphs, whose degrees vary widely? In "Conductance and Congestion in Power law Graphs" and "Certain Connectivity Properties of the Internet Topology" we showed that power law random graphs can also support routing with near-optimal throughput and congestion (upto $\log n$ multiplicative factor), by establishing strong conductance properties in the core of such graphs. The significance of these results is that they provide a rigorous method to argue about the rate of growth of "business" or traffic in the core of the Internet. Once conductance is established a range of other graph properties follow (e.g. separation of the second eigenvalue of the stochastic normalization) and virtually all issues that have been known for regular graphs (e.g. reliability, cover times, hitting times, crawling, searching, information retrieval) are now amenable to analysis.

In papers "On the Random Walk Method for P2P networks", "Hybrid Search Schemes for unstructured P2P networks" and "Random Walks in Power Law Random Graphs" we explore the power of the random walk method as a sampling primitive, and hence a primitive for designing very efficient algorithms in Peer-to-Peer networks.

In my very recent papers "Towards Topology Aware Networks" and "A Local Exchange Markov Chain for Graphs with Given Degrees and Application to Connectivity of Peer-to-Peer Networks", the aim is to design fully distributed algorithms that enhance a network with topology awareness and connectivity properties (such as good expansion, or conductance), which improve the performance of typical protocols running on the underlying network.

1.2 Internet Models: Most of the algorithms and protocols used in the Internet are quite sensitive to its topology. For the purpose of simulation and testing, it is important to characterize the main features of this topology and be able to generate graphs and demand patterns that "look like" the Internet. In an influential paper, Faloutsos et. al. gave a characterization in terms of power-law statistics on the *degree sequence*, *eigenvalues* and path lengths of the Internet at the Autonomous System level; this is the routing fabric of the network.

In paper "On the Eigenvalue Power Law" Papadimitriou and I gave the first mathematical explanation of the eigenvalue power law. For graphs in which high degree nodes are "star-like", which is true of the Internet graph, the latter follows from the former. This result has particular relevance to clustering and information retrieval. By a detailed analysis of the eigenvalues and eigenvectors of matrices related to the adjacency matrix of the Internet topology, in "Spectral Methods for Internet Topologies" we have, for the first time, isolated clusters of Autonomous Systems with semantic correlations (geography, business, etc.). We showed that such clusters can be used to define "intracluster" and "intercluster" demand patters, corresponding to best case and worst case network loads respectively.

The current method of choice for generating graphs that "look like" the Internet is the following: Generate a random graph whose degree sequence matches the given power law target sequence. In "Generating Random Graphs with Given Degree Sequences" and ""A Local Exchange Markov Chain for Graphs with Given Degrees and Application to Connectivity of Peer-to-Peer Networks", we give an efficient method of accomplishing this by simulating a suitably defined Markov chain.

In ongoing work, together with my students, we are trying to develop network models with varying parameters (of expansion, conductance and spectral gap), which will provide network simulation with extreme, as well as average, cases of network topology configurations.

I believe that it is very significant for the network simulation community, to have models representing extremal network instances. In particular, since eigenvalues are indicative of the performance of fundamental network processes, we would like to develop models where the spectrum of the graph is a parameter of the model. In on-going work I am developing such models.

Beyond networking, I am exploring algorithmic problems related to the World Wide Web and other very large data corpuses (e.g. NSA sanitized data). For example, in the context of crawling the WWW, I am working on efficient strategies when servers present arbitrary and unknown delays. Together with my students, we are developing efficient queuing strategies, and an extension of the theory of random walks for the case where nodes have arbitrary and unknown delays.

I also see the primitives of how this experience may bridge with biology. For example, what we see as communities of interest or small clusters in communications networks correspond to "motifs" in metabolic networks. Biologists need to characterize what constitutes a "motif". We can tackle this problem by studying the statistical properties of small patterns on random complex networks (mean and variance) and characterize a motif by the deviation of the corresponding pattern on the biological network. As another example, the question has been raised on how epidemics spread in networks with very skewed degrees. I am exploring how the spectral gap, or "clustering" of these networks affects the spread of epidemics. As a final example, I am looking at gene and protein interaction networks of the drosophila and the human genome, and try to explain the significance of nodes with very low degree but very high centrality. A lot of these open questions covered in the graduate course on Algorithms for Complex Networks that I taught in Spring 2005 and Spring 2007 and can be found on my homepage.

2. Rapidly mixing Markov chains, expanders and their applications

The simulation of Markov chains is an exciting algorithmic paradigm that has many computational applications. In the late 80's, ingenious techniques were introduced for bounding the mixing times of Markov chains (and therefore the running times of the corresponding Monte Carlo algorithms) in terms of a structural property of the Markov chain, known as its expansion or conductance. The first bounds, due to Alon, and Jerrum and Sinclair, used algebraic arguments: they established a connection between expansion and the second largest eigenvalue of the Laplacian of the Markov chain, which has been known to dominate the mixing rate of the chain.

In my PhD thesis and my FOCS paper "Conductance and Convergence of Markov Chains: A Combinatorial Treatment of Expanders", I gave the first truly combinatorial (non-algebraic) argument for bounding mixing time in terms of expansion, thus extending results known only for undirected graphs and reversible Markov chains to directed graphs and non-reversible Markov chains.

In "Number of Eulerian Orientations", Winkler and I gave the first efficient Monte Carlo algorithm for approximating the so-called Z_{ICE} function, which is second in importance only to the Ising model in statistical physics. The main ingredient of this algorithm was a Markov chain for randomly sampling Eulerian orientations of a given undirected graph.

In "Polytopes, Permanents and Graphs with Large Factors", Dagum, Luby, Vazirani and I derived an approximation algorithm for computing permanents of graphs with large factors. This

work extended Jerrum and Sinclair's novel canonical path technique for lower bounding expansion to random paths selected from suitably defined probability spaces.

In "Balanced Matroids" Feder and I deal with the problem of sampling the bases of a matroid. This very general problem includes certain forms of the network reliability problem as a special case. Our paper initiated work on negative correlations, which was followed up by probabilists as well as computer scientists (Robin Pemantle, Yuval Peres, Mark Jerrum).

In "On the Random Walk Method for Protocol testing", Papadimitriou and I studied Markov chains in the context of testing software protocols on very large state spaces. We provided the first rigorous analysis of the random simulation method for protocol testing.

3. Network Design

During the days of design and early deployment of the information super highway, fundamental technological choices were being made by a consortium of telecommunications companies, under DARPA contracts, of which Bellcore was a part. I headed the optimization group. Several of the algorithms and software developed were incorporated in Bellcore commercial toolkits for automated design of next generation backbone networks. Two examples of theoretical and implementation work done in this context are:

THE STEINER NETWORK PROBLEM: An important consideration in network design is survivability under link failures. The Steiner network problem, which is a generalization of the classical Steiner tree problem to higher connectivity requirements, is an abstraction of this issue. In "A Primal-Dual Algorithm for the Steiner Network Problem" we gave the first approximation algorithm for this problem. This involved defining a mechanism of relaxing complementary slackness conditions for the use of the primal-dual schema, an idea which was later used in numerous works. A heuristic adaptation of this algorithm was implemented and the resulting code forms the core of the Bellcore product for automated design of Common Channel Signaling networks.

PATH COLORING: In the WDM (wavelength division multiplexing) high speed networking technology, several optical waves can travel through a single strand of fiber; however, they must be of different frequencies. This gives rise to new path coloring problems, which can be viewed as generalizations of the edge coloring problem. Approximation algorithms for minimizing the number of frequencies used are presented in "Efficient Access to Optical Bandwidth".

4. Information Retrieval

Caching is a widely used technique for improving efficiency of data delivery on the Web and in advanced telecommunications networks. In "Cashing with Expiration Times for Internet and WWW Applications" we study natural adaptations of LRU to the problem of caching with expiration times, where data is tagged with the time beyond which it is not valid.

A commonly used preprocessing technique to improve query response time in very large databases and data warehouses is to materialize, in memory, suitably chosen "views" (summaries) of the database. In an award winning SIGMOD '96 paper, Harinarayan, Rajaraman and Ullman gave a combinatorial formalization of this operation and a greedy heuristic for it. In the "Complexity of the View Selection Problem", we pointed out that the actual critical optimization parameter cannot be approximated, assuming P$\neq$NP.