The Annals of Applied Statistics
2010, Vol. 4, No. 1, 1–4
DOI: 10.1214/10-AOAS346

INTRODUCTION TO PAPERS ON THE MODELING AND ANALYSIS OF NETWORK DATA

BY STEPHEN E. FIENBERG

Carnegie Mellon University

In today's world, networks seem to appear everywhere. There are social networks, communication networks, financial transaction networks, gene regulatory networks, disease transmission networks, ecological food networks, mobile telephone and sensor networks and more. We, our professional colleagues, our friends and family, and especially our students, are often part of online networks such as *Facebook*, *LinkedIn* and now *Google Buzz*. Some network structures are static and others are dynamically evolving. Networks are usually represented in terms of graphs with the nodes representing entities, for example, people, and the edges representing ties or relationships. Edges may be directed or undirected depending on the application and substantive question of interest. In terms of statistical science, a network model is one that accounts for the structure of the network ties in terms of the probability that each network tie exists, whether conditional on all other ties, or as considered part of the distribution of the ensemble of ties.

Ideas and language from graph theory abound in the technical literature on networks. A typical representation involves a network with N nodes, having $\binom{N}{2}$ unordered pairs of nodes, and hence $2\binom{N}{2}$ possible directed edges. If the labels on edges reflect the nodes they link, as (i, j), Y_{ij} represents the existence of an edge from individual i to j, and $\{\mathbf{Y}\} = \{Y_{12}, Y_{13}, \ldots, Y_{(N-1)N}\}$ represents the ties in the graph. The simplest network models assume the edges to be independent, while a statistically more interesting class of models treats the dyadic structures for pairs of nodes to be independent.

In an extensive review of the statistical literature on network modeling, Goldenberg et al. (2010) note:

> Almost all of the "statistically" oriented literature on the analysis of networks derives from a handful of seminal papers. In social psychology and sociology there is the early work of Simmel (1950) at the turn of the last century and Moreno (1934) in the 1930s, as well as the empirical studies of Milgram (1967) and Travers and Milgram (1969) in the 1960s; in mathematics/probability there is the Erdös–Rényi work on random graph models [Erdös and Rényi (1959, 1960), and a closely related *Annals of Mathematical Statistics* paper by Gilbert (1959)]. There are of course other papers that dealt with these topics contemporaneously or even earlier. But these are the ones that appear to have had lasting impact.

Received March 2010.

Statistical work in the late 1970s and early 1980s emphasized models that exploited dyadic independence, for example, in the work of Holland and Leinhardt (1981). More complex exponential random graph models (ERGMs) then drew considerable attention; for example, see Frank and Strauss (1986). But the estimation of parameters for such models turns out to have been more problematic than expected; for example, see the discussion in Rinaldo, Fienberg and Zhou (2009).

The network modeling literature has "taken off" in the past decade, in part because of the interest in structures associated with the internet, and there are contributors from many different disciplines, including biology, computer science, statistical physics, sociology and, of course, statistics. Kolacyzk (2009) provides a book length treatment of a selection of approaches and Airoldi et al. (2007) provides a compilation of relevant papers. In addition there is the probabilistic literature that has derived from the Erdös–Rényi–Gilbert formulations much of which is described in Chung and Lu (2006) and Durrett (2006).

Methods for the analysis of network data now take at least as many forms as the applications in which they arise. While the original examples of networks analyzed in the literature were typically small (e.g., $n = 18$ nodes corresponding to monks in a monastery), the size of networks analyzed with more modern methodology has grown exponentially. Networks with 1000 nodes are common, for example, in the study of protein–protein interaction, and online networks such as *Facebook* include hundreds of million nodes. An interesting statistical question we can ask is whether there is a relevant asymptotics associated with network models as we move into such high dimensions. A recent paper by Bickel and Chen (2009) opens the door to such important statistical issues by linking back to ideas in the probabilistic network literature.

The response to our initial call for papers on the topic of network modeling was so overwhelming that we are dividing the special section into two parts, with the first appearing in this issue of *The Annals of Applied Statistics* (Volume 4, No. 1), and the remainder in the next issue (Volume 4, No. 2).

In Part I of this special section, we include a diverse collection of papers with applications spanning sampling of rare populations, internet flows, gene networks, online e-loyalty networks, document-as-nodes links induced from text, and more. The methodologies begin with ERGMs but include sparse regression models and state space models.

- In *Modeling Social Networks from Sampled Data*, Handcock and Gile develop the conceptual and computational statistical framework for likelihood inference for ERGMs based on sampled network information, especially for data from adaptive network designs. They motivate and illustrate these ideas by analyzing the effect of link-tracing sampling designs on the collaborative working relations between 36 partners in a New England law firm.
- In *Analysis of Dependence Among Size, Rate and Duration of Internet Flows*, Park, Hernãndez-Campos, Marron, Jeffay and Smith use Pearson's correlation

coefficient and extremal dependence analysis to study the flows of packet traces from three internet networks. The correlations between size and duration turn out to be much smaller than one might expect and can be strongly affected by applying thresholds to size or duration. Using extremal dependence analysis, they draw a similar conclusion, that is, near independence for extremal values of size and rate.

- Peng, Zhu, Han, Noh, Pollack and Wang work with sparse regression approaches in *Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer*. They apply their methods to genome wide RNA transcript levels and DNA copy numbers were measured for 172 tumor samples.
- In *Optimal Experiment Design in a Filtering Context with Application to Sampled Network Data*, Singhal and Michailidis examine the problem of optimal design in the context of filtering multiple random walks on networks. They apply their methodology to tracking network flow volumes using sampled data where the design variable corresponds to controlling the sampling rate, and they relate their approach to the steady state optimal design for state space models.
- Political networks and gene regulatory networks are the primary focus of application in *Estimating Time-Varying Networks* by Kolar, Song, Ahmed and Xing. They describe an approach that builds on a temporally smoothed l^1-regularized logistic regression formalism that can be cast as standard convex-optimization problem and solved efficiently using generic solvers scalable to large networks.
- Working with scientific citation networks, hyperlinked web pages and geographically tagged news articles, Chang and Blei develop a *Hierarchical Relational Model of Document Networks*. They develop a hierarchical model of both network structure where the attributes of each document are its words, and for each pair of documents, the model is their link as a binary random variable that is conditioned on their contents. They derive efficient inference and estimation algorithms based on variational methods that take advantage of sparsity and scale with the number of links.
- Jank and Yahav focus on a dataset involving 30,000 auctions from one of the main consumer-to-consumer online auction houses. They propose a novel measure of e-loyalty via the associated network of transactions between bidders and sellers. In *E-Loyalty Networks in Online Auctions*, they employ ideas from functional principal component analysis to derive, from this network, the distribution of perceived loyalty of every individual seller and associated loyalty scores. In the process, they confront the clustering feature of loyalty networks, with a few high-volume sellers accounting for most of the individual transactions.

Part II of this special section will explore another diverse collection of network models and applications.

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E., GOLDENBERG, A., XING, E. P. and ZHENG, A. X., eds. (2007). *Statistical Network Analysis: Models, Issues and New Directions. Lecture Notes in Computer Science* **4503**. Springer, Berlin.

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

CHUNG, F. and LU, L. (2006). *Complex Graphs and Networks*. Amer. Math. Soc., Providence, RI. MR2248695

DURRETT, R. (2006). *Random Graph Dynamics*. Cambridge Univ. Press. MR2271734

ERDÖS, P. and RÉNYI, A. (1959). On random graphs, I. *Publ. Math. Debrecen* **6** 290–297. MR0120167

ERDÖS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** 17–61. MR0125031

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

GILBERT, E. N. (1959). Random graphs. *Ann. Math. Statist.* **30** 1141–1144. MR0108839

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning* **2** 129–233.

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

KOLACYZK, E. D. (2009). *Statistical Analysis of Network Models*. Springer, New York.

MILGRAM, S. (1967). The small world problem. *Psychology Today* **1** 60–67.

MORENO, J. (1934). *Who Shall Survive?* Nervous and Mental Disease Publishing Company, Washington, DC.

RINALDO, A., FIENBERG, S. E. and ZHOU, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **3** 446–484. MR2507456

SIMMEL, G. and WOLFF, K. H. (1950). *The Sociology of Georg Simmel*. The Free Press, New York.

TRAVERS, J. and MILGRAM, S. (1969). An experimental study of the small world problem. *Sociometry* **32** 425–443.

DEPARTMENT OF STATISTICS AND
MACHINE LEARNING DEPARTMENT
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: fienberg@stat.cmu.edu