A Survey of Statistical Network Models
A Survey of Statistical Network Models

Anna Goldenberg
University of Toronto
Canada
anna.goldenberg@utoronto.ca

Alice X. Zheng
Microsoft Research
USA
alicez@microsoft.com

Stephen E. Fienberg
Carnegie Mellon University
USA
fienberg@stat.cmu.edu

Edoardo M. Airoldi
Harvard University
USA
airoldi@fas.harvard.edu

Full text available at: http://dx.doi.org/10.1561/2200000005
Foundations and Trends® in
Machine Learning
Volume 2 Issue 2, 2009
Editorial Board

Editor-in-Chief:
Michael Jordan
Department of Electrical Engineering and Computer Science
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720-1776

Editors

Peter Bartlett (UC Berkeley)
Yoshua Bengio (Université de Montréal)
Avrim Blum (Carnegie Mellon University)
Craig Boutilier (University of Toronto)
Stephen Boyd (Stanford University)
Carla Brodley (Tufts University)
Inderjit Dhillon (University of Texas at Austin)
Jerome Friedman (Stanford University)
Kenji Fukumizu (Institute of Statistical Mathematics)
Zoubin Ghahramani (Cambridge University)
David Heckerman (Microsoft Research)
Tom Heskes (Radboud University Nijmegen)
Geoffrey Hinton (University of Toronto)
Aapo Hyvärinen (Helsinki Institute for Information Technology)
Leslie Pack Kaelbling (MIT)
Michael Kearns (University of Pennsylvania)
Daphne Koller (Stanford University)
John Lafferty (Carnegie Mellon University)
Michael Littman (Rutgers University)
Gabor Lugosi (Pompeu Fabra University)
David Madigan (Columbia University)
Pascal Massart (Université de Paris-Sud)
Andrew McCallum (University of Massachusetts Amherst)
Marina Meila (University of Washington)
Andrew Moore (Carnegie Mellon University)
John Platt (Microsoft Research)
Luc de Raedt (Albert-Ludwigs Universitaet Freiburg)
Christian Robert (Université Paris-Dauphine)
Sunita Sarawagi (IIT Bombay)
Robert Schapire (Princeton University)
Bernhard Schoelkopf (Max Planck Institute)
Richard Sutton (University of Alberta)
Larry Wasserman (Carnegie Mellon University)
Bin Yu (UC Berkeley)

Full text available at: http://dx.doi.org/10.1561/2200000005
Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2009, Volume 2, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.
A Survey of Statistical Network Models

Anna Goldenberg\(^1\), Alice X. Zheng\(^2\), Stephen E. Fienberg\(^3\) and Edoardo M. Airoldi\(^4\)

\(^1\) Center for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, M5S 3G4, Canada, anna.goldenberg@utoronto.ca
\(^2\) Microsoft Research, One Microsoft Way, Redmond, WA, 98052, USA, alicez@microsoft.com
\(^3\) Department of Statistics, Machine Learning Department, Cylab, and iLab Carnegie Mellon University, Pittsburgh, PA, 15213-3890, USA, fienberg@stat.cmu.edu
\(^4\) Department of Statistics & FAS Center for Systems Biology, Harvard University, 1 Oxford Street, Cambridge, MA, 02138, USA, airoldi@fas.harvard.edu

Abstract

Networks are ubiquitous in science and have become a focal point for discussion in everyday life. Formal statistical models for the analysis of network data have emerged as a major topic of interest in diverse areas of study, and most of these involve a form of graphical representation. Probability models on graphs date back to 1959. Along with empirical studies in social psychology and sociology from the 1960s, these early works generated an active “network community” and a substantial literature in the 1970s. This effort moved into the statistical literature in the late 1970s and 1980s, and the past decade has seen a burgeoning network literature in statistical physics and computer science. The growth
of the World Wide Web and the emergence of online “networking communities” such as Facebook, MySpace, and LinkedIn, and a host of more specialized professional network communities has intensified interest in the study of networks and network data.

Our goal in this review is to provide the reader with an entry point to this burgeoning literature. We begin with an overview of the historical development of statistical network modeling and then we introduce a number of examples that have been studied in the network literature. Our subsequent discussion focuses on a number of prominent static and dynamic network models and their interconnections. We emphasize formal model descriptions, and pay special attention to the interpretation of parameters and their estimation. We end with a description of some open problems and challenges for machine learning and statistics.
# Contents

1 Introduction 1

1.1 Overview of Modeling Approaches 2

1.2 What This Survey Does Not Cover 7

2 Motivation and Data-Set Examples 11

2.1 Motivations for Network Analysis 11

2.2 Sample Data Sets 13

3 Static Network Models 25

3.1 Basic Notation and Terminology 26

3.2 The Erdős–Rényi–Gilbert Random Graph Model 26

3.3 The Exchangeable Graph Model 28

3.4 The $p_1$ Model for Social Networks 33

3.5 $p_2$ Models for Social Networks and Their Bayesian Relatives 36

3.6 Exponential Random Graph Models 37

3.7 Random Graph Models with Fixed Degree Distribution 40

3.8 Blockmodels, Stochastic Blockmodels and Community Discovery 42

3.9 Latent Space Models 46
Many scientific fields involve the study of networks in some form. Networks have been used to analyze interpersonal social relationships, communication networks, academic paper co-authorships and citations, protein interaction patterns, and much more. Popular books on networks and their analysis began to appear a decade ago (see, e.g., \cite{24, 50, 68, 318, 319}), and online “networking communities” such as Facebook, MySpace, and LinkedIn are an even more recent phenomenon.

In this work, we survey selective aspects of the literature on statistical modeling and analysis of networks in social sciences, computer science, physics, and biology. Given the volume of books, papers, and conference proceedings published on the subject in these different fields, a single comprehensive survey would be impossible. Our goal is far more modest. We attempt to chart the progress of statistical modeling of network data over the past 70 years and to outline succinctly the major schools of thought and approaches to network modeling and to describe some of their interconnections. We also attempt to identify major statistical gaps in these modeling efforts. From this overview one might then synthesize and deduce promising future research directions. Kolaczyk \cite{177} provides a complementary statistical overview.
Introduction

The existing set of statistical network models may be organized along several major axes. For this monograph, we choose the axis of static vs. dynamic models. Static network models concentrate on explaining the observed set of links based on a single snapshot of the network, whereas dynamic network models are often concerned with the mechanisms that govern changes in the network over time. Most early examples of networks were single static snapshots. Hence static network models have been the main focus of research for many years. However, with the emergence of online networks, more data are available for dynamic analysis, and in recent years there has been growing interest in dynamic modeling.

In the remainder of this chapter we provide a brief historical overview of network modeling approaches. In subsequent chapters we introduce some examples studied in the network literature and give a more detailed comparative description of select modeling approaches.

1.1 Overview of Modeling Approaches

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 and Woff [268] at the turn of the last century and Moreno [222] in the 1930s as well as the empirical studies of Stanley Milgram [216, 298] in the 1960s; in mathematics/probability there is the Erdös–Rényi paper on random graph models [94]. There are other papers that dealt with these topics contemporaneously or even earlier. But these are the ones that appear to have had lasting impact.

Moreno [222] invented the sociogram — a diagram of points and lines used to represent relations among persons, a precursor to the graph representation for networks. Luce and others developed a mathematical structure to go with Moreno’s sociograms using incidence matrices and graphs (see, e.g., [11, 200, 201, 202, 203, 244, 282]), but the structure they explored was essentially deterministic. Milgram gave the name to what is now referred to as the “Small-World” phenomenon — short paths of connections linking most people in social spheres — and his experiments had provocative results: the shortest path between any
two people for completed chains has a median length of around 6; however, the majority of chains initiated in his experiments were never completed! (His studies provided the title for the play and movie Six Degrees of Separation, ignoring the complexity of his results due to the censoring.) White [321] and Fienberg and Lee [100] gave a formal Markov chain like model and analysis of the Milgram experimental data, including information on the uncompleted chains. Milgram’s data were gathered in batches of transmission, and thus these models can be thought of as representing early examples of generative descriptions of dynamic network evolution. Recently, Dodds et al. [86] studied a global “replication” variation on the Milgram study in which more than 60,000 e-mail users attempted to reach one of 18 target persons in 13 countries by forwarding messages to acquaintances. Only 384 of 24,163 chains reached their targets but they estimate the median length for completions to be 7, by assuming that attrition occurs at random.

The social science network research community that arose in the 1970s was built upon these earlier efforts, in particular the Erdős–Rényi–Gilbert model. Research on the Erdős–Rényi–Gilbert model (along with works by Katz et al. [166, 167, 168]) engendered the field of random graph theory. In their papers, Erdős and Rényi worked with fixed number of vertices, \( N \), and number of edges, \( E \), and studied the properties of this model as \( E \) increases. Gilbert studied a related two-parameter version of the model, with \( N \) as the number of vertices and \( p \) the fixed probability for choosing edges. Although their descriptions might at first appear to be static in nature, we could think in terms of adding edges sequentially and thus turn the model into a dynamic one. In this alternative binomial version of the Erdős–Rényi–Gilbert model, the key to asymptotic behavior is the value \( \lambda = pN \). There is a “phase change” associated with the value of \( \lambda = 1 \), at which point we shift from seeing many small connected components in the form of trees to the emergence of a single “giant connected component.” Probabilists such as Pittel [243] imported ideas and results from stochastic processes into the random graph literature.

Holland and Leinhardt’s [150] \( p_1 \) model extended the Erdős–Rényi–Gilbert model to allow for differential attraction (popularity) and expansiveness, as well as an additional effect due to reciprocation.
Introduction

The $p_1$ model was log-linear in form, which allowed for easy computation of maximum likelihood estimates using a contingency table formulation of the model [103, 104]. It also allowed for various generalizations to multidimensional network structures [101] and stochastic blockmodels. This approach to modeling network data quickly evolved into the class of $p^*$ or exponential random graph models (ERGMs) originating in the work of Frank and Strauss [110] and Strauss and Ikeda [287]. A trio of papers demonstrating procedures for using ERGMs [241, 254, 316] led to the widespread use of ERGMs in a descriptive form for cross-sectional network structures or cumulative links for networks — what we refer to here as static models. Full maximum likelihood approaches for ERGMs appeared in the work of Snijders and Handcock and their collaborators, some of which we describe in Section 3.

Most of the early examples of networks in the social science literature were relatively small (in terms of the number of nodes) and involved the study of the network at a fixed point in time or cumulatively over time. Only a few studies (e.g., Sampson’s 1968 data on novice monks in the monastery [259]) collected, reported, and analyzed network data at multiple points in time so that one could truly study the evolution of the network, i.e., network dynamics. The focus on relatively small networks reflected the state-of-art of computation, but it was sufficient to trigger the discussion of how one might assess the fit of a network model. Should one focus on “small sample” properties and exact distributions given some form of minimal sufficient statistic, as one often did in other areas of statistics, or should one look at asymptotic properties, where there is a sequence of networks of increasing size? Even if we have “repeated cross-sections” of the network, if the network is truly evolving in continuous-time we need to ask how to ensure that the continuous-time parameters are estimable. We return to many of these questions in subsequent sections.

In the late 1990s, physicists began to work on network models and study their properties in a form similar to the macro-level descriptions of statistical physics. Barabási, Newman, and Watts, among others, produced what we can think of as variations on the Erdős–Rényi–Gilbert model which either controlled the growth of the network or
allowed for differential probabilities for edge addition and/or deletion. These variations were intended to produce phenomena such as “hubs,” “local clustering,” and “triadic closures.” The resulting models gave us fixed degree distribution limits in the form of power-laws — variations on preferential attachment models (“the rich get richer”) that date back to Yule [320] and Simon [269] (see also [219]) — as well as what became known as “small-world” models. The small-world phenomenon, which harks back to Milgram’s 1960s studies, usually refers to two distinct properties: (1) small average distance and (2) the “clustering” effect, where two nodes with a common neighbor are more likely to be adjacent. Many of these authors claim that these properties are ubiquitous in realistic networks. To model networks with the small-world phenomenon, it is natural to utilize randomly generated graphs with a power-law degree distribution, where the fraction of nodes with degree \( k \) is proportional to \( k^{-a} \) for some positive exponent \( a \). Many of the most relevant papers are included in an edited collection by Newman et al. [204]. More recently this style of statistical physics models has been used to detect community structure in networks, e.g., see Girvan and Newman [122] and Backstrom et al. [20], a phenomenon which has its counterpart description in the social science network modeling literature.

The probabilistic literature on random graph models from the 1990s made the link with epidemics and other evolving stochastic phenomena. Picking up on this idea, Watt and Strogatz [320] and others used epidemic models to capture general characteristics of the evolution of these new variations on random networks. Durrett [91] has provided us with a book-length treatment on the topic with a number of interesting variations on the theme. The appeal of stochastic processes as descriptions of dynamic network models comes from being able to exploit the extensive literature already developed, including the existence and the form of stationary distributions and other model features or properties. Chung and Lu [69] provide a complementary treatment of these models and their probabilistic properties.

One of the principal problems with this diverse network literature that we see is that, with some notable exceptions, the statistical tools for estimation and assessing the fit of “statistical physics” or stochastic
process models are lacking. Consequently, no attention is paid to the fact that real data may often be biased and noisy. What authors in the network literature have often relied upon is the extraction of key features of the related graphical network representation, e.g., the use of power-laws to represent degree distributions or measures of centrality and clustering, without any indication that they are either necessary or sufficient as descriptors for the actual network data. Moreover, these summary quantities can often be highly misleading as the critique by Stouffer et al. [285, 286] of methods used by Barabási [25] and Vázquez et al. [304] suggest. Barabási claimed that the dynamics of a number of human activities are scale-free, i.e., he specifically reported that the probability distribution of time intervals between consecutive e-mails sent by a single user and time delays for e-mail replies follow a power-law with exponent $-1$, and he proposed a priority-queuing process as an explanation of the bursty nature of human activity. Stouffer et al. [286] demonstrated that the reported power-law distribution was solely an artifact of the analysis of the empirical data and used Bayes factors to show that the proposed model is not representative of e-mail communication patterns. See a related discussion of the poor fit of power-laws in Clauset et al. [74]. There are several works, however, that try to address model fitting and model comparison. For example, the work of Williams and Martinez [323] showed how a simple two-parameter model predicted “key structural properties of the most complex and comprehensive food webs in the primary literature”. Another good example is the work of Middendorf et al. [215] where the authors used network motif counts as input to a discriminative systematic classification for deciding which configuration model the actual observed network came from; they looked at power-law, small-world, duplication-mutation and duplication-mutation-complementation and other models (seven in total) and concluded that the duplication-mutation-complementation model described the protein–protein interaction data in Drosophila melanogaster species best.

Machine learning approaches emerged in several forms over the past decade with the empirical studies of Faloutsos et al. [97] and Kleinberg [172, 173, 174], who introduced a model for which the underlying graph is a grid — the graphs generated do not have a power-law degree
1.2 What This Survey Does Not Cover

This survey focuses primarily on statistical network models and their applications. As a consequence there are a number of topics that we touch upon only briefly or essentially not at all, such as:

- **Probability theory associated with random graph models.** The probabilistic literature on random graph models is now truly extensive and the bulk of the theorems and proofs, while interesting in their own right, are largely unconnected with the present exposition. For excellent introductions to this literature, see Chung and Lu [69] and Durrett [91]. For related results on the mathematics of graph theory, see Bollobás [43].

- **Efficient computation on networks.** There is a substantial computer science literature dealing with efficient calculation of quantities associated with network structures, such as shortest paths, network diameter, and other measures of connectivity, centrality, clustering, etc. The edited volume by Brandes and Erlebach [48] contains good overviews of a number of these topics as well as other computational issues associated with the study of graphs.

- **Use of the network as a tool for sampling.** Adaptive sampling strategies modify the sampling probabilities of selection based on observed values in a network structure.
This strategy is beneficial when searching for rare or clustered populations. Thompson and Seber [296] and Thompson [293] discuss adaptive sampling in detail. There is also related work on target sampling [294] and respondent-driven sampling [258] [305].

• Neural networks. Neural networks originated as simple models for connections in the brain but have more recently been used as a computational tool for pattern recognition (e.g., Bishop [38]), machine learning (e.g., Neal [229]), and models of cognition (e.g., Rogers and McClelland [257]).

• Networks and economic theory. A relatively new area of study is the link between network problems, economic theory, and game theory. Some useful entrees to this literature are Even-Dar and Kearns [96], Goyal [131], Kearns et al. [169], and Jackson [160], whose book contains an excellent semi-technical introduction to network concepts and structures.

• Relational networks. This is a very popular area in machine learning. It uses probabilistic graphical models to represent uncertainty in the data. The types of “networks” in this area, such as Bayes nets, dependency diagrams, etc., have a different meaning than the networks we consider in this review. The main difference is that the networks in our work are considered to “be given” or arising directly from properties of the network under study, rather than being representative of the uncertainty of the relationships between nodes and node attributes. There is a multitude of literature on relational networks, e.g., see Friedman et al. [112], Getoor et al. [116], Neville and Jensen [230], Neville et al. [231], and Getoor and Taskar [117].

• Bipartite graphs. These are graphs that represent measurement on two populations of objects, such as individuals and features. The graphs in this context are seldom the best representation of the data, with exception perhaps of binary measurements or when the true populations have comparable sizes. Recent work on exchangeable Rasch matrices is related to this topic and potentially relevant for network analysis.
1.2 What This Survey Does Not Cover

Lauritzen [186][187], Bassetti et al. [29] suggest applications to bipartite graphs.

- **Agent-based modeling.** Building on older ideas such as cellular automata, agent-based modeling attempts to simulate the simultaneous operations of multiple agents, in an effort to re-create and predict the actions of complex phenomena. Because the interest is often on the interaction among the agents, this domain of research has been linked with network ideas. With the recent advances in high-performance computing, simulations of large-scale social systems have become an active area of research, e.g., see [46]. In particular, there is a strong interest in areas that revolve around national security and the military, with studies on the effects of catastrophic events and biological warfare, as well as computational explorations of possible recovery strategies [54][58]. These works are the contemporary counterparts of more classical work at the interface between artificial intelligence and the social sciences [54][55][57].
References


References


Full text available at: http://dx.doi.org/10.1561/2200000005


References


References


References


References


References


References


References


98 References


References


References


References


Full text available at: http://dx.doi.org/10.1561/2200000005
References


References


References


References


