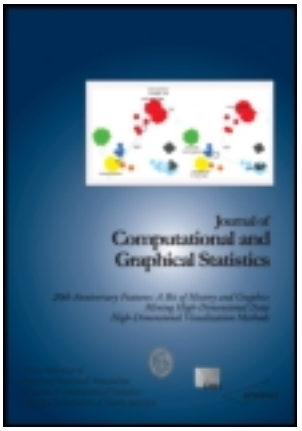

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

A Brief History of Statistical Models for Network Analysis and Open Challenges

Stephen E. Fienberg [a]

[a] Department of Statistics, Machine Learning Department, Heinz College, and Cylab, Carnegie Mellon University, Pittsburgh, PA, 15213-3890

Accepted author version posted online: 19 Oct 2012.Version of record first published: 14 Dec 2012.

To cite this article: Stephen E. Fienberg (2012): A Brief History of Statistical Models for Network Analysis and Open Challenges, Journal of Computational and Graphical Statistics, 21:4, 825-839

To link to this article: http://dx.doi.org/10.1080/10618600.2012.738106

A Brief History of Statistical Models for Network Analysis and Open Challenges

Stephen E. FIENBERG

Networks are ubiquitous in science. They have also 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 "social science 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 coming out of statistical physics and computer science. In particular, the growth of the World Wide Web and the emergence of online "networking communities" such as Facebook, Google +, MySpace, LinkedIn, and Twitter, and a host of more specialized professional network communities have intensified interest in the study of networks and network data. This article reviews some of these developments, introduces some relevant statistical models for static network settings, and briefly points to open challenges.

Key Words: Community discovery; Exponential random graph models; Latent network structure; Maximum likelihood estimation; MCMC; Network experimentation.

1. INTRODUCTION

With the advent of widespread and enveloping online social network communities such as Facebook, Google +, MySpace, LinkedIn, and Twitter, it has become fashionable to analyze and report on network data and structure (see, e.g., Watts 1999, 2003; Barabási 2002; Buchanan 2002; Christakis and Fowler 2009). Entire conferences are devoted to the study of social media and the interactions of individuals who use them, and network visualizations appear in newspapers and popular magazines.

Many scientific fields, especially in the social sciences, have used network representations for their problems. The "statistically" oriented literature on the analysis of networks derives from a handful of seminal contributions, such as those by Simmel and Wolff (1950) at the turn of the last century, and by Moreno (1934), who in the 1930s invented the sociogram—a diagram of points and lines used to represent relations among

Stephen E. Fienberg is Professor (E-mail: *fienberg@stat.cmu.edu*), Department of Statistics, Machine Learning Department, Heinz College, and Cylab, Carnegie Mellon University, Pittsburgh, PA 15213-3890.

Journal of Computational and Graphical Statistics, Volume 21, Number 4, Pages 825–839
DOI: 10.1080/10618600.2012.738106

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. Goldenberg et al. (2010) provided an extensive review of this literature and its more modern incarnations, from which we draw the following overview. Their article also includes extensive discussions of both static and dynamic network modeling; here we focus simply on the static models.

2. HISTORICAL OVERVIEW OF NETWORK MODELS AND ANALYSIS

There was early recognition of network analysis as a subdiscipline in the social science community. The journal *Social Networks* began to publish in 1978 and several articles on network analysis appeared in the *Journal of the American Statistical Association* in the 1980s. The literature has burgeoned with the emergence of online social networks and contributions to the field by social scientists and statisticians are now outnumbered by those of computer scientists and statistical physicists.

2.1 SIZE MATTERS

What distinguishes much of the current work on networks and their analysis from that of the twentieth century is scale. Facebook claims to have a billion users. Protein interaction networks typically involve many hundreds or even thousands of nodes. The analysis of cocitation networks, two decades ago, would typically involve less than 100 articles or authors but today could include a thousand or more. And much of the current effort with different styles of models is focused on computations with networks of such size.

The early network studies in sociology almost all dealt with relatively small sets of subjects whose interconnections could be studied in depth, or at least whose connections allowed communication across a small set of links in a larger network setting. Stanley Milgram (Milgram 1967; Travers and Milgram 1969) 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! For example, in the Boston study (Travers and Milgram 1969), only 64 of 296 chains were completed. His studies provided the title for the play and movie *Six Degrees of Separation*, which ignored the complexity of his results due to the censoring. White (1970) and Fienberg and Lee (1975) 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, Muhamad, and Watts (2003) 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 estimated the median length for completions to be 7, by assuming that attrition occurs at random.

Similarly, in the 1970s, several sociologists chose to study "blockmodels" and "structurally equivalent" groups of individuals again with small network datasets such as those arising in Sampson's (1968) study of the relationships among 18 novices in a monastery, at the time of considerable upheaval in the Roman Catholic Church more broadly following Vatican II. Sampson's data have become a canonical example to illustrate new methods ranging from blockmodel algorithms (Breiger, Boorman, and Arabie 1975; White, Boorman, and Breiger 1976), to the p_1 model of Holland and Leinhardt (1981) and its generalizations (Fienberg, Meyer, and Wasserman 1985), to mixed-membership stochastic blockmodels by Airoldi et al. (2008), and it is a centerpiece in the Hunter, Krivitsky, and Schweinberger (2012) article in this issue. Typically, authors focus on a single extract from the Sampson study instead of its full multivariate and longitudinal structure. Other such examples include Zachary's karate club network of friendships between 34 members of a karate club at a U.S. university in the 1970s (Zachary 1977) and Lazega's study of relationships among 72 partners and associates in a law firm, for example, see Lazega and van Duijn (1997).

2.2 SOME BASIC MODELS

Erdös and Rényi (1960) and the contemporaneous article by Gilbert (1959) introduced the basic probability model for network analysis where all edges essentially have the same probability of occurrence, independently from one another. Erdös and Rényi worked with a 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 as the fixed probability for choosing edges. Although their descriptions might at first appear to be static in nature, we can think in terms of adding edges sequentially and thus turn the model into a dynamic one. In this alternative binomial version of the Erdös–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 (1990) imported ideas and results from stochastic processes into the random graph literature. Aldous (1997) made the link between stochastic processes and random graphs, giving rise to a fundamental distribution in population genetics. See also the book by Durrett (2006).

Holland and Leinhardt's (1981) p_1 model extended the Erdös–Rényi–Gilbert model to allow for differential attraction (popularity) and expansiveness, as well as an additional effect due to reciprocation. Fienberg and Wasserman demonstrated that the p_1 model was log-linear in form, and this alternate representation allowed for easy computation of maximum likelihood estimates (MLEs) using a contingency table formulation of the model (Fienberg and Wasserman 1981a, b).

In the late 1990s, physicists began to work on network models and study their properties in a form similar to the macrolevel descriptions of statistical physics. Barabási, Newman, and Watts, among others, produced what we can think of as variations on the Erdös–Rényi–Gilbert model that 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 (1925) and Simon (1955) (see also Mitzenmacher 2004)—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 claimed that these properties are ubiquitous in realistic networks. To model networks with the small-world phenomenon, it is natural to use 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 articles are included in an edited collection by Newman, Barabási, and Watts (2006). More recently, this style of statistical physics models has been used to detect community structure in networks, for example, see Girvan and Newman (2002) and Backstrom et al. (2006), a phenomenon that 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, Watts and Strogatz (1998) and others used epidemic models to capture general characteristics of the evolution of these new variations on random networks. Durrett's (2006) book-length treatment on the topic includes 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 (2006) provided a complementary treatment of these models and their probabilistic properties.

One of the principal problems with this diverse network literature is that, with some notable exceptions, the statistical tools for estimating 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, for example, 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. Barabási (2005) claimed that the dynamics of a number of human activities are scale free, that is, 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 follows a power law with exponent -1, and he proposed a priority-queuing process as an explanation of the bursty nature of human activity. Stouffer, Malmgren, and Amaral (2008) 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, Shalizi, and Newman (2009).

Machine learning approaches emerged in several forms over the past decade with the empirical studies of Faloutsos, Faloutsos, and Faloutsos (1999) and Kleinberg (2000a, b, 2001), who introduced a model for which the underlying graph is a grid—the graphs generated do not have a power-law degree distribution, and each vertex has the same expected degree. The strict requirement that the underlying graph be a cycle or grid renders the model inapplicable to webgraphs or biological networks. Durrett (2006) treated variations on this

model as well. More recently, a number of authors have looked to combine the stochastic blockmodel ideas from the 1980s with latent space models, model-based clustering (Handcock, Raftery, and Tantrum 2007), or mixed-membership models (Airoldi et al. 2008), to provide generative models that scale in reasonable ways to substantial-sized networks. The class of mixed-membership models resembles a form of soft clustering (Erosheva, Fienberg, and Lafferty 2004) and includes the latent Dirichlet allocation model (Blei, Ng, and Jordan 2003) from machine learning as a special case.

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, such as Sampson's examination of the novices in the monastery, have involved the collection, reporting, and analysis of network data at multiple points in time, thus allowing for the study of the evolution of the network, that is, network dynamics. The focus on relatively small networks reflected the state of the art of computation, but it was sufficient to trigger the discussion of how we might assess the fit of a network model. Should we focus on "small-sample" properties and exact distributions given some form of minimal sufficient statistic, as we often do in other areas of statistics, or should we look at asymptotic properties, where there is a sequence of networks of increasing size? If so, how should we characterize the limiting process and thus the asymptotics? 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.

3. CONDITIONALLY INDEPENDENT EDGES AND DYADS

For a binary graph with conditionally independent edges, each edge outcome X_{ij} can be expressed as a Bernoulli binary random variable with probability of existence p_{ij}. The simplest of this class of network models is the Erdös–Rényi–Gilbert random graph model (Erdös and Rényi 1959, 1960; Gilbert 1959) (sometimes referred to as the Erdös–Rényi model, or "the" random graph model), in which any given edge exists with probability p. This model extends immediately to directed graphs, where the existence of any edge has the same probability p, irrespective of the direction.

The *beta* model is the natural extension to the Erdös–Rényi model, which allows for heterogeneous edge probabilities:

$$\log \frac{P(X_{ij})}{1 - P(X_{ij})} = \beta_i + \beta_j, \tag{1}$$

and there has been a flurry of articles recently on its properties, for example, see Diaconis and Sly (2011). The minimal sufficient statistic is the degree sequence, and Rinaldo, Petrovic, and Fienberg (2011) established conditions for the existence of the MLEs of the $\{\beta_i\}$.

The p_1 model of Holland and Leinhardt (1981) begins with a directed version of the beta model and proposes that three factors affect the outcome of a dyad with directed edges: (1) the "gregariousness" α of an individual, that is, how likely one is to have outgoing ties; (2) the "popularity" β of an individual, that is, how likely one is to have incoming ties; and (3) "reciprocity" ρ, that is, the tendency to which the two arcs in a dyad are identical, taking into account their existing characteristics. Given a parameter for the overall density

of edges θ, the form of the joint likelihood is

$$\log P(X = x) \propto \theta x_{++} + \sum_i \alpha_i x_{i+} + \sum_j \beta_j x_{+j} + \rho \sum_{ij} x_{ij} x_{ji}, \tag{2}$$

where $K(\rho, \theta, \alpha, \beta)$ is a normalizing constant to ensure that the total probabilities add to 1. The minimum sufficient statistics are the in-degree and out-degree for each node and the number of dyads with reciprocated edges. Holland and Leinhardt (1981) presented an iterative proportional fitting method for maximum likelihood estimation for this model and discussed the complexities involved in assessing goodness of fit. Fienberg and Wasserman (1981a) provided a contingency table and log-linear representation of this simple version of p_1 and extended the model to allow for node-specific reciprocation where we replace ρ by $\rho + \rho_i + \rho_j$. The log-linear formulation also led to various generalizations for characterizing multidimensional network structures, for example, by Fienberg, Meyer, and Wasserman (1985).

A natural extension of the p_1 model is the case of tightly linked "blocks" of nodes, within which the α parameters are equated or at least taken to be exchangeable and similarly for the β parameters, suggesting an equivalence between members of the same block. The inference for and discovery of "communities" in networks has become especially popular in recent network literature in a variety of different applications; see Newman (2004) for an example. See the further discussion of blockmodels in the next section.

3.1 Ensemble Models and Topological Motivations—ERGMs

Rather than focusing on the dyads as independent units, some classes of models consider topological features of interest in the network as the main measure of the model. The best known of these is the exponential random graph model, or ERGM, introduced by Frank and Strauss (1986), also referred to as the p^* class of models (Anderson, Wasserman, and Faust 1992), which extends the p_1 class of model by adding statistical summaries of topological relevance. For example, the number of three cycles or triangles in a graph is equal to

$$T(x) = \sum_{i,j,k} x_{ij} x_{jk} x_{ki}. \tag{3}$$

We can then add an additional parameter into the likelihood of Equation (2), for which the triad count (3.1) is the corresponding multiplier, as in

$$P(X = x) \propto \exp\left(\tau T(x) + \rho m + \theta x_{++} + \sum_i \alpha_i x_{i+} + \sum_j \beta_j x_{+j}\right). \tag{4}$$

Due to the computational intractability of the normalizing constant for Equation (4), which is of the form $K(\tau, \rho, \theta, \alpha, \beta)$, Strauss and Ikeda (1990) introduced a pseudolikelihood approximation that built on independent logistic regression components. A trio of articles building on their pseudolikelihood procedures for ERGMs by Wasserman and Pattison (1996), Pattison and Wasserman (1999), and Robins, Pattison, and Wasserman (1999) led to the widespread use of ERGMs in a descriptive form for cross-sectional network structures or cumulative links for networks. Much of the recent literature on ERGMs uses Markov chain Monte Carlo (MCMC) methods for implementing maximum likelihood estimation of

the model parameters (Snijders 2002). Additionally, these models often have degenerate or near-degenerate solutions, as explained by Handcock et al. (2003) and Rinaldo, Fienberg, and Zhou (2009). Hunter, Krivitsky, and Schweinberger (2012) gave further details for these models and Handcock et al. (2008) provided software for fitting them.

If we move back to the undirected network setting and treat the nodes symmetrically, then Equation (4) reduces to Frank and Strauss' (1986) homogeneous Markov random graph model:

$$P(X = x) \propto \exp\left(\sum_{k=1}^{n-1} \theta_k S_k(x) + \tau T(x) + \psi(\theta, \tau)\right) \quad x \in \mathcal{X}, \tag{5}$$

where $\{S_k\}$ and T are count-specific, minimal sufficient statistics structures, such as edges, triangles, and k-stars:

number of edges: $S_1(x) = \sum_{1 \le i \le j \le n} x_{ij}$,

number of k-stars ($n - 1 \le k \ge 2$): $S_k(x) = \sum_{1 \le i \le n} \binom{x_{i+}}{k}$,

number of triangles: $T(x) = \sum_{1 \le i \le j \le h \le n} x_{ij}\, x_{ih}\, x_{jh}$,

$\theta = \{\theta_k\}$ and τ are the parameters, and $\psi(\theta, \tau)$ is the normalizing constant. Note that there is a hierarchical dependence structure to the parameters of the Markov random graph model of Equation (5), with edges being contained in 2-stars, and 2-stars being contained in both triangles and 3-stars, etc. The statistical attractiveness of the Markov random graph model comes in part from its "usual graphical model" interpretation in terms of conditional independence, when one moves from the standard graph representation to an edge graph representation, where the edges become nodes, and the new edges correspond to k-stars in the original graph. How to exploit this interpretation and Markov structure for theoretical and computational purposes remains an open problem.

One of the more subtle ideas surrounding network models is well captured by the structure of ERGMs. For ERGMs that go beyond dyadic independence, there is no simple generative process that allows us to write the likelihood function as a product of independent components. Thus, we are, in essence, in an $n = 1$ situation. In the real world, however, we often only observe a piece of a full network, and possibly only a sample. The question then arises as to whether inferences from a subnetwork using ERGMs generalize to the full network. Conversely, if an ERGM describes the full network, will it also be appropriate for a subnetwork? Stumpf, Wiuf, and May (2005) answered these questions in the negative for scale-free networks and Wiuf and Stumpf (2006) provided a more general characterization. Shalizi and Rinaldo (in press) addressed these questions for ERGMs. They said that a model is *projective* when the same parameters can be used for the full network and for any of its subnetworks. They demonstrated that ERGMs are projective essentially only for models involving dyadic independence, and that Markov random graph models and other ERGMs involving parameters that lead to the counting of triangles are not projectable. The Shalizi–Rinaldo results also explain the sense in which one can or cannot get the consistency of maximum likelihood estimation for ERGMs.

4. SOME ALTERNATIVE APPROACHES TO ERGMs

The network modeling literature has a multiplicity of techniques and approaches that go beyond ERGMs. We describe a few of these briefly.

From the 1970s onward, for example, see Lorrain and White (1971), Breiger, Boorman, and Arabie (1975), and White, Boorman, and Breiger (1976), there has been a near obsession with blocks or community structure within networks. A blockmodel relies on the intuitive notion of the *structural equivalence* of sets of nodes, that is, based on their connectivity. Blockmodels are typically characterized as having dense ties within a collection of nodes (a block) and sparse ties between blocks, that is, people within a community interact more than between communities. More general versions are essentially framed as a partition of the nodes in the graph such that the resulting blocks, either within or between, are either dense or sparse, for example, see Bickel and Chen (2009). When the edges or dyads in directed networks are taken to be independent, these stochastic blockmodels are simply variants of those we described in the preceding section and, while implicit in earlier work, they are explicit in Holland, Laskey, and Leinhardt (1983). For a description of the extensive related literature on the search for community structure, see Goldenberg et al. (2010).

In the historical overview in Section 2, we mentioned two Bayesian variations on this theme, involving latent space models (Hoff, Raftery, and Handcock 2002) and mixed-membership stochastic blockmodels (Airoldi et al. 2008). In the latent space approach, the nodes in the network are embedded in a low-dimensional Euclidean space with placement set as a function of covariates, and then with clustering added on to determine the blocks (Handcock, Raftery, and Tantrum 2007). MCMC is the mechanism for estimating the posterior distribution. The mixed-membership stochastic blockmodel approach is similar to the latent structure model approach in a way, but the nodes are now set in a low-dimensional simplex, where the vertices are the counterpart of blocks or clusters and every node has a membership vector that describes what proportion of it belongs to each vertex. The data-generating process for the mixed-membership stochastic blockmodel is the following:

1. For each node $p \in N$,
   (a) sample mixed membership $\vec{\pi}_p \sim \text{Dirichlet}_K(\vec{\alpha})$.
2. For each pair of nodes $(p, q) \in N \times N$,
   (a) sample membership indicator, $\vec{z}_{p\to q} \sim \text{mult}_K(\vec{\pi}_p)$,
   (b) sample membership indicator, $\vec{z}_{p\leftarrow q} \sim \text{mult}_K(\vec{\pi}_q)$,
   (c) sample interaction, $Y(p, q) \sim \text{Bern}(\vec{z}^{\top}_{p\to q} B \vec{z}_{p\leftarrow q})$.

The matrix B in (c) above provides the block structure involving K blocks. A key feature of this model is the fact that the group membership of each node is *context dependent*, that is, each node may assume different membership when interacting with or being interacted by different peers in the network. Statistically, each node is an admixture of group-specific interactions. The two sets of latent group indicators are denoted by $\{\vec{z}_{p\to q} : p, q \in \mathcal{N}\} =: Z_{\to}$ and $\{\vec{z}_{p\leftarrow q} : p, q \in \mathcal{N}\} =: Z_{\leftarrow}$. The pairs of group memberships that underlie

interactions need not be equal; this fact is useful for characterizing asymmetric interaction networks. We can enforce quality when modeling symmetric interactions.

MCMC works for small numbers of nodes but approximations are necessary for the mixed-membership model approach to scale. In particular, the variational approximation approach by Airoldi et al. (2008) scales easily to thousands of nodes. Deciding on the appropriate number of blocks, K, can be done in a natural fashion in the Bayesian framework, but unless one uses an automated approach such as the maximum a posteriori (MAP) rule to choose K—methods for choosing K are computationally intensive. There are also time-varying versions of the mixed-membership stochastic blockmodels, although they tend to treat time as discrete and look at network structure in a sequence of epochs, for example, see Kolar et al. (2010).

For networks generated from a stochastic blockmodel, Rohe, Chatterjee, and Yu (2011) bound the number of "misclustered" nodes using spectral clustering methods. Their asymptotic results allow the number of clusters in the model to grow with the number of nodes.

The primary purpose of this section has been to single out statistically valid, model-based approaches to incorporate block structure into network models as opposed to the computationally oriented ones, especially those in the literature on "community structure" in network settings.

5. SOME CHALLENGING NETWORK RESEARCH PROBLEMS

Goldenberg et al. (2010) listed a number of open and challenging research problems in the domain of statistical models. What follows is a somewhat related description of five areas I think are especially worthy of statistical attention and which have either direct or indirect links to computation and/or visualization.

5.1 VISUALIZATION TIED TO STATISTICAL MODELS

One of the more popular ways to visualize network adjacency data is force-directed simulation algorithms that assign forces among the set of edges and the set of nodes, for example, as if the edges were springs and the nodes were electrically charged particles (di Battista et al. 1999; Kaufmann and Wagner 2001). The simulation applies forces to the nodes, pulling them closer together or pushing them further apart, until the computation achieves a steady state.

A related spring-like approach exploits some variant of multidimensional scaling to find a "good" visualization layout, for example, see Eades (1984). Yet other approaches use general purpose clustering algorithms. While each approach has a different form of visual appeal, none of them are tied to the kinds of models we review in this article. Developing effective visual displays linked to classes of statistical models such as ERGMs remains a major challenge.

Graphical displays are more often artifacts of the algorithms used as opposed to illumination about models and their interpretation. Methods such as mixed-membership stochastic blockmodels and latent space models do point a natural pathway toward visualization, but we need a systematic way to think about model-based or model-associated visual network displays.

5.2 STATISTICAL PROPERTIES OF MLES AND OTHER NETWORK MODEL ESTIMATORS AND ASSESSING GOODNESS OF FIT

Given that the basic network models described in Section 3 have been in use for several decades, it is surprising that their statistical properties did not develop somehow in parallel. As we noted, it has only been very recently that we have seen a number of new theoretical results for network models of practical relevance: (1) Chatterjee, Diaconis, and Sly (2011) on asymptotics for the simple beta model; (2) Rinaldo, Petrovic, and Fienberg (2011) on the existence of MLEs for the beta and p_1 models; and (3) Shalizi and Rinaldo (in press) on the consistency of maximum likelihood estimation for ERGMs. The algebraic statistics representation by Petrović, Rinaldo, and Fienberg (2010) and Fienberg, Petrović, and Rinaldo (2011) that draws on the polynomial structure of p_1 models is a potential first step toward new approaches to assessing model fit. But there remain many important open questions regarding the properties of statistical approaches to network modeling that require attention.

One might not think that such work bears highlighting in a journal devoted to computational issues, but the fact is that many of the methods proposed for scaling statistical estimation to large networks may compute answers that make little or no sense, because the likelihood function is poorly behaved and maximized on or near the boundary.

5.3 NETWORK EXPERIMENTATION

A key challenge is to develop the next generation of statistical methods and software tools for the design and analysis of experiments when observations are not independent due to network interactions and influences. Unfortunately, almost all of formal statistical theory for randomized experiments assumes that experimental units are independent, and that one unit's outcomes are unaffected by another unit's treatment assignment, referred to as the stable unit-treatment value assumption (SUTVA), for example, see Cox (1958) and Rubin (1980).

Networks by their very nature involve dependence among the nodes as we have illustrated in the presentation above, that is, the units in a network are dependent even if they are linked by independent or conditionally independent edges. We expect treatments randomly allocated to nodes to propagate via network links and network dynamics over time. Thus, we expect substantial violations of SUTVA and treatment interference. As a consequence, we ask: what does randomization buy us in experiments on networks? Or alternatively, how should one design a network-based experiment with randomization so that we can statistically analyze the results?

In as yet unpublished work by several different authors, there are suggestions for how to approach this using ideas from cluster experimentation and hierarchical modeling of spillover effects, but we need much more focused attention on the issue. There are many empirical issues associated with network experimentation that require focused statistical attention as well.

5.4 DYNAMIC NETWORK-BASED MODELS FOR TRANSACTIONAL DATABASES THAT INCORPORATE GEOGRAPHY AND OTHER MEASURES OF LOCALITY

More often than not, commercial and large-scale research data arise in a networked environment, especially in today's digital online world, but where the network structure is latent. What we observe is either manifestations of links, for example, e-mail messages, or transactions attributable to individuals locatable in time and/or space. In such circumstances, we are often interested in measuring the latent network structure, for example, see Eagle, Pentland, and Lazer (2009) and Crandall et al. (2010), as well as its influence on some other phenomenon, for example, purchase behavior (Ugander et al. 2012). This type of modeling needs to be better rooted in the context of latent versions of some of the models described or referred to in this article.

5.5 NETWORK PRIVACY

A review of network challenges would be remiss without at least a brief reference to the issue of privacy and network data. Network privacy is a topic that is regularly in the news, in part because of GPS (global positioning system) location tracking of cell phones, and in part because of the concern that current large-scale social networks offer little of it and that users of them have little understanding of how their data are used by the owners of the network for commercial purposes as well as by other users who are able to access what should be private information directly or indirectly (Backstrom, Dwork, and Kleinberg 2007). Rather than providing yet another long list of references on the topic, we simply note that this is an active area of research that is made especially difficult precisely because of the dependence of the nodes in the network, and nodes with high numbers of in-degrees or out-degrees, that is, high levels of linkage, are especially vulnerable to attack. But if network data are to be shared in some form as part of the study of social networks, we wish to avoid having the protection of privacy of individuals so badly distort the inferences one can reach from the released data, so that little is to be learned from their analyses.

ACKNOWLEDGMENTS

This research was supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office and by grant FA9550-12-1-0392 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA).

[Received September 2012. Revised September 2012.]

REFERENCES

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [827,829,832,833]

Aldous, D. (1997), "Brownian Excursions, Critical Random Graphs and the Multiplicative Coalescent," *The Annals of Probability*, 25, 812–854. [827]

Anderson, C. J., Wasserman, S. S., and Faust, K. (1992), "Building Stochastic Blockmodels," *Social Networks*, 14, 137–161. [830]

Backstrom, L., Dwork, C., and Kleinberg, J. (2007), "Wherefore Art Thou r3579x? Anonymized Social Networks, Hidden Patterns, and Structural Steganography," in *Proceedings of the 16th International Work Wide Web Conference (WWW 2007)*, pp. 181–191, New York: ACM Press. [835]

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006), "Group Formation in Large Social Networks: Membership, Growth, and Evolution," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press, pp. 44–54. [828]

Barabási, A.-L. (2002), *Linked: The New Science of Networks*, Cambridge, MA: Perseus. [825]

——— (2005), "The Origin of Bursts and Heavy Tails in Human Dynamics," *Nature*, 435, 207–211. [828]

Bickel, P. J., and Chen, A. (2009), "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities," *Proceedings of the National Academy of Sciences*, 106, 21068–21073. [832]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [829]

Breiger, R., Boorman, S., and Arabie, P. (1975), "An Algorithm for Clustering Relational Data With Applications to Social Network Analysis and Comparison With Multidimensional Scaling," *Journal of Mathematical Psychology*, 12, 328–383. [827,832]

Buchanan, M. (2002), *Nexus: Small Worlds and the Groundbreaking Science of Networks*, New York: W. W. Norton & Company. [825]

Chatterjee, S., Diaconis, P., and Sly, A. (2011), "Random Graphs With a Given Degree Sequence," *The Annals of Applied Probability*, 21, 1400–1435. [829,834]

Christakis, N. A., and Fowler, J. H. (2009), *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*, New York: Little, Brown and Co. [825]

Chung, F., and Lu, L. (2006), *Complex Graphs and Networks*, Providence, RI: American Mathematical Society. [828]

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009), "Power-Law Distributions in Empirical Data," *SIAM Review*, 51, 661–703. [828]

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley. [834]

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. (2010), "Inferring Social Ties From Geographic Coincidences," *Proceedings of the National Academy of Sciences*, 107, 22436–22441. [835]

di Battista, G., Eades, P., Tamassia, R., and Tollis, I. G. (1999), *Graph Drawing: Algorithms for the Visualization of Graphs*, Upper Saddle River, NJ: Prentice Hall. [833]

Dodds, P. S., Muhamad, R., and Watts, D. J. (2003), "An Experimental Study of Search in Global Social Networks," *Science*, 301, 827–829. [826]

Durrett, R. (2006), *Random Graph Dynamics*, Cambridge: Cambridge University Press. [827,828]

Eades, P. (1984), "A Heuristic for Graph Drawing," *Congressus Numerantium*, 42, 149–160. [833]

Eagle, N., Pentland, A., and Lazer, D. (2009), "Inferring Friendship Network Structure by Using Mobile Phone Data," *Proceedings of the National Academy of Sciences*, 106, 15274–15278. [835]

Erdös, P., and Rényi, A. (1959), "On Random Graphs, I," *Publicationes Mathematicae*, 6, 290–297. [829]

——— (1960), "The Evolution of Random Graphs," *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5, 17–61. [827,829]

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. (2004), "Mixed-Membership Models of Scientific Publications," *Proceedings of the National Academy of Sciences*, 101(Suppl. 1), 5220–5227. [829]

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999), "On Power-Law Relationships of the Internet Topology," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '99)*, New York: ACM Press, pp. 251–261. [828]

Fienberg, S. E., and Lee, S. K. (1975), "On Small World Statistics," *Psychometrika*, 40, 219–228. [826]

Fienberg, S. E., Meyer, M. M., and Wasserman, S. S. (1985), "Statistical Analysis of Multiple Sociometric Relations," *Journal of the American Statistical Association*, 80, 51–67. [827,830]

Fienberg, S. E., Petrović, S., and Rinaldo, A. (2011), "Algebraic Statistics for Random Graph Models: Markov Bases and Their Uses," in *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland* (Vol. 202 of Lecture Notes in Statistics), eds. N. J. Dorans and S. Sinharay, New York: Springer, pp. 21–38. [834]

Fienberg, S. E., and Wasserman, S. S. (1981a), "Categorical Data Analysis of Single Sociometric Relations," in *Sociological Methodology 1981*, ed. S. Leinhardt, San Francisco: Jossey-Bass, pp. 156–192. [827,830]

——— (1981b), "An Exponential Family of Probability Distributions for Directed Graphs: Comment," *Journal of the American Statistical Association*, 76, 54–57. [827]

Frank, O., and Strauss, D. (1986), "Markov Graphs," *Journal of the American Statistical Association*, 81, 832–842. [830,831]

Gilbert, E. N. (1959), "Random Graphs," *The Annals of Mathematical Statistics*, 30, 1141–1144. [827,829]

Girvan, M., and Newman, M. E. J. (2002), "Community Structure in Social and Biological Networks," *Proceedings of the National Academy of Sciences*, 99, 7821–7826. [828]

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. [826,832,833]

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2008), "statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data," *Journal of Statistical Software*, 24, 12–25. [831]

Handcock, M. S., Raftery, A. E., and Tantrum, J. (2007), "Model-Based Clustering for Social Networks" (with discussion), *Journal of the Royal Statistical Society,* Series A, 170, 301–354. [829,832]

Handcock, M. S., Robins, G. L., Snijders, T. A. B., Moody, J., and Besag, J. (2003), "Assessing Degeneracy in Statistical Models of Social Networks," *Journal of the American Statistical Association*, 76, 33–50. [831]

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098. [832]

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137. [832]

Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs" (with discussion), *Journal of the American Statistical Association*, 76, 33–65. [827,829,830]

Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012), "Computational Statistical Methods for Social Network Models," *Journal of Computational and Graphical Statistics*, 21, 856–882, DOI: 10.1080/10618600.2012.732921. [827,831]

Kaufmann, M., and Wagner, D. (2001), *Drawing Graphs: Methods and Models* (Lecture Notes in Computer Science, vol. 2025), New York: Springer. [833]

Kleinberg, J. M. (2000a), "Navigation in a Small World—It Is Easier to Find Short Chains Between Points in Some Networks Than Others," *Nature*, 406, 845. [828]

——— (2000b), "The Small-World Phenomenon: An Algorithmic Perspective," in *Proceedings of the 32nd ACM Symposium on Theory of Computing*, New York: ACM Press, pp. 163–170. [828]

——— (2001), "Small-World Phenomena and the Dynamics of Information," in *Advances in Neural Information Processing Systems (NIPS)* (Vol. 14), pp. 431–438, Cambridge, MA: MIT Press. [828]

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010), "Estimating Time-Varying Networks," *The Annals of Applied Statistics*, 4, 94–123. [833]

Lazega, E., and van Duijn, M. (1997), "Position in Formal Structure, Personal Characteristics and Choices of Advisors in a Law Firm: A Logistic Regression Model for Dyadic Network Data," *Social Networks*, 19, 375–397. [827]

Lorrain, F., and White, H. C. (1971), "Structural Equivalence of Individuals in Social Networks," *Journal of Mathematical Sociology*, 1, 49–80. [832]

Milgram, S. (1967), "The Small World Problem," *Psychology Today*, 1, 60–67. [826]

Mitzenmacher, M. (2004), "A Brief History of Generative Models for Power Law and Lognormal Distributions," *Internet Mathematics*, 1, 226–251. [828]

Moreno, J. (1934), *Who Shall Survive?* Washington, DC: Nervous and Mental Disease Publishing Company. [825]

Newman, M., Barabási, A.-L., and Watts, D. J. (eds.) (2006), *The Structure and Dynamics of Networks*, Princeton, NJ: Princeton University Press. [828]

Newman, M. E. J. (2004), "Detecting Community Structure in Networks," *European Physics Journal B*, 38, 321–330. [830]

Pattison, P. E., and Wasserman, S. S. (1999), "Logit Models and Logistic Regressions for Social Networks: II. Multivariate Relations," *British Journal of Mathematical and Statistical Psychology*, 52, 169–193. [830]

Petrović, S., Rinaldo, A., and Fienberg, S. E. (2010), "Algebraic Statistics for a Directed Random Graph Model With Reciprocation," in *Algebraic Methods in Statistics and Probability II* (Vol. 516), eds. M. A. G. Vianaand H. P. Wynn, Providence, RI: American Mathematical Society, pp. 261–283. [834]

Pittel, B. (1990), "On Tree Census and the Giant Component in Sparse Random Graphs," *Random Structures and Algorithms*, 1, 311–342. [827]

Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), "On the Geometry of Discrete Exponential Families With Application to Exponential Random Graph Models," *Electronic Journal of Statistics*, 3, 446–484. [831]

Rinaldo, A., Petrovic, S., and Fienberg, S. E. (2011), "Maximum Likelihood Estimation in Network Models," arXiv:1105.6145. [829,834]

Robins, G. L., Pattison, P. E., and Wasserman, S. S. (1999), "Logit Models and Logistic Regressions for Social Networks: III. Valued Relations," *Psychometrika*, 64, 371–394. [830]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [833]

Rubin, D. B. (1980), Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test" by D. Basu, *Journal of the American Statistical Association*, 75, 591–593. [834]

Sampson, F. S. (1968), "A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships," Ph.D. thesis, Cornell University. [827]

Shalizi, C. R., and Rinaldo, A. (in press), "Consistency Under Sampling of Exponential Random Graph Models," *The Annals of Statistics*, 40. [831,834]

Simmel, G., and Wolff, K. H. (1950), *The Sociology of Georg Simmel*, New York: The Free Press. [825]

Simon, H. A. (1955), "On a Class of Skew Distribution Functions," *Biometrika*, 42, 425–440. [828]

Snijders, T. A. B. (2002), "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure*, 2, 1–40. [831]

Stouffer, D. B., Malmgren, R. D., and Amaral, L. A. N. (2008), "Lognormal Statistics in E-mail Communication Patterns" [online]. Available at *http://arXiv.org/abs/physics/0605027*. [828]

Strauss, D., and Ikeda, M. (1990), "Pseudolikelihood Estimation for Social Networks," *Journal of the American Statistical Association*, 85, 204–212. [830]

Stumpf, M. P. H., Wiuf, C., and May, R. M. (2005), "Subnets of Scale-Free Networks Are Not Scale-Free: Sampling Properties of Networks," *Proceedings of the National Academy of Sciences*, 102, 4221–4224. [831]

Travers, J., and Milgram, S. (1969), "An Experimental Study of the Small World Problem," *Sociometry*, 32, 425–443. [826]

Ugander, J., Backstrom, L., Marlow, C., and Kleinberg, J. (2012), "Structural Diversity in Social Contagion," *Proceedings of the National Academy of Sciences*, 109, 5962–5966. [835]

Wasserman, S. S., and Pattison, P. E. (1996), "Logit Models and Logistic Regression for Social Networks: I. An Introduction to Markov Graphs and p^*," *Psychometrika*, 61, 401–425. [830]

Watts, D. J. (1999), *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton, NJ: Princeton University Press. [825]

——— (2003), *Six Degrees: The Science of a Connected Age*, New York: W. W. Norton & Company. [825]

Watts, D. J., and Strogatz, S. H. (1998), "Collective Dynamics of 'Small-World' Networks," *Nature*, 393, 440–442. [828]

White, H. C. (1970), "Search Parameters for the Small World Problem," *Social Forces*, 49, 259–264. [826]

White, H. C., Boorman, S. A., and Breiger, R. L. (1976), "Social Structure From Multiple Networks. I. Blockmodels of Roles and Positions," *The American Journal of Sociology*, 81, 730–780. [827,832]

Wiuf, C., and Stumpf, M. P. H. (2006), "Binomial Subsampling," *Journal of the Royal Society,* Series A, 462, 1181–1195. [831]

Yule, G. U. (1925), "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S," *Philosophical Transactions of the Royal Society of London,* Series B, 213, 21–87. [828]

Zachary, W. W. (1977), "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of Anthropological Research*, 33, 452–473. [827]