LA-UR-16-20622 (Accepted Manuscript)

A local structure model for network analysis

Casleton, Emily Michele
Nordman, Daniel
Kaiser, Mark

To be published in: Statistics and Its Interface

DOI to publisher's version: 10.4310/SII.2017.v10.n2.a15

Permalink to record: http://permalink.lanl.gov/object/view?what=info:lanl-repo/lareport/LA-UR-16-20622

Statistics and Its Interface Volume 0 (0000) 1–13

A Local Structure Model for Network Analysis

Emily Casleton*, Daniel Nordman and Mark Kaiser

The statistical analysis of networks is a popular research topic with ever widening applications. Exponential random graph models (ERGMs), which specify a model through interpretable, global network features, are common for this purpose. In this paper we introduce a new class of models for network analysis, called local structure graph models (LSGMs). In contrast to an ERGM, a LSGM specifies a network model through local features and allows for an interpretable and controllable local dependence structure. In particular, LSGMs are formulated by a set of full conditional distributions for each network edge, e.g., the probability of edge presence/absence, depending on neighborhoods of other edges. Additional model features are introduced to aid in specification and to help alleviate a common issue (occurring also with ERGMs) of model degeneracy. The proposed models are demonstrated on a network of tornadoes in Arkansas where a LSGM is shown to perform significantly better than a model without local dependence.

AMS 2000 subject classifications: Primary 90B15, 05C80; secondary 62H11, 62P12.
Keywords and phrases: Conditional Distributions, Exponential Random Graphs, Markov random fields, neighborhoods, network analysis.

1. INTRODUCTION

Applications of networks appear in a wide variety of disciplines. For example, sociologists use graph models to represent social networks, economists have used networks for studying relations between countries [17], biologists to represent brain connectivity [44], [39], zoologists for examining animal social behavior [28], and computer scientists for representing connections on the internet.

Much literature is devoted to algorithmic construction methods with a goal to quickly and accurately simulate networks mimicking certain properties of interest [27]. Such methods are often not statistical in nature in the sense that the algorithms involve no probability models producing tractable likelihood inference. In contrast, some network analysis approaches allow for explicit probabilistic modeling and related likelihood inference. In a review, Hunter et. al. [20] categorize probabilistic modeling of networks into the exponential random graph models (ERGMs) and latent variable models (LVMs). Of these two, ERGMs have been more widely used and extensively studied. Their popularity can be attributed to the ability to incorporate graph topology as terms of a joint (log-linear) distribution that allows for complex dependencies [24]. Although ERGMs allow for complex dependencies, such dependencies are typically induced rather than directly specified. That is, dependencies in ERGMs are a consequence of graph topologies chosen to be included in the joint distribution. Latent variable models encompass a broad class of models that are hierarchical in nature. Here variables representing edges are commonly specified as having conditional distributions that are conditionally independent given some latent variable defined on the nodes, such as group membership [40], [31] or position within a social space [16], [15].

In this paper, we introduce an approach to specifying a model for network analysis which we call local structure graph models (LSGMs). As a key characteristic, the LSGMs begin model formulation based on a set of full conditional distributions for each potential edge in the network, the distribution for the presence/absence of an edge given the outcomes of all other potential edges. As a further critical characteristic, each conditional distribution is specified in terms of a flexible neighborhood structure, explicitly identifying a set of other network edges on which an edge of interest is "locally" dependent. Under certain conditions, conditional specifications and neighborhood definitions allow construction of a global or joint probability model for the network, having a dependence structure which is interpretable and introduced in a controlled, local manner.

As a consequence of its formulation, LSGMs have characteristics of *both* ERGMs and LVMs, a feature of what has been called "the next generation" of network analysis [42]. Similarly to LVMs, LSGMs are specified through conditional distributions. However, the conditional distributions in the LSGM are not defined in terms of latent variables for nodes, such as group memberships, but rather in terms of neighborhoods involving other network *edges*. Hence, potential network edges are conditionally dependent on other edges belonging to a neighborhood. Like ERGMs, LSGMs result in joint distributions that have a Gibbsian form for random graphs. But in LSGMs the joint distribution results from a set of specified conditioned distributions for edges, while in ERGMs the joint is formulated directly. Consequently, the dependence structure of an ERGM is often induced, while

arXiv: math.PR/0000000
*Corresponding author

that of the LSGM can be more directly and explicitly defined.

Because the joint distributions of LSGMs are similar to those for ERGMs, some details of ERGMs are discussed in Section 2. The main features of LSGMs, involving conditional specifications and neighborhood definitions, are detailed in Section 3, along with a numerical demonstration of a model. Two additional features of LSGMs, the ability to simply incorporate potential spatial information about nodes, and the definition of a "saturated graph," are also introduced in Section 3. These features can help keep the potential sizes of LSGM neighborhoods manageable which is useful for minimizing model degeneracy issues. In Section 4, a LSGM is applied to an example network consisting of tornado outbreaks in the state of Arkansas. Two simulation-based model comparison techniques are also presented in this section and used to compare the fit of the LSGM to that of a model lacking local dependence. The LSGM is shown to provide a graph model of tornado occurrences that better supports observed tornado patterns in important local ways with regard to space and time. Section 5 provides some concluding remarks.

2. EXPONENTIAL RANDOM GRAPH MODEL (ERGM)

A network, or graph, is defined by a set of n nodes and m edges, where the networks of interest here are undirected and simple, with unweighted edges and no self-loops. To construct a random graph model, assign to each of the $\binom{n}{2}$ possible edges a binary random variable $Y(\mathbf{s}_i)$, where the marker $\mathbf{s}_i = \{c_i, r_i\}$ indicates the two nodes, denoted as c_i and r_i, that a potential edge would join. Edge values are collected into $\mathbf{Y}$, an $n \times n$ adjacency matrix, and each entry designates the presence, $y(\mathbf{s}_i) = 1$, or absence, $y(\mathbf{s}_i) = 0$, of an edge between each node pair in the graph. For undirected, simple networks, $\mathbf{Y}$ will be symmetric with $\text{diag}(\mathbf{Y}) = 0$. A realization of the network will be represented as $\mathbf{y}$.

Specification of an ERGM involves identifying the number of elements of $\mathbf{Y}$ that correspond to edges of certain types, which are often called topological features of the graph. For example, the well-studied triad model of [10] includes the topological features of density, or the expected proportion of realized edges, 2-stars, and triangles. Let the classes of edge types to be included in an ERGM be indexed by $j = 1, \ldots, q$. For any possible realization $\mathbf{y}$, let $g_j(\mathbf{y})$ denote the number of occurrences of edge class j present in $\mathbf{y}$. The joint distribution of $\mathbf{y}$ is then specified as

$$\Pr(\mathbf{Y} = \mathbf{y}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{\sum_{j=1}^{q} \theta_j g_j(\mathbf{y})\right\} \tag{1}$$

where θ_j is a model parameter corresponding to topological graph feature of type $j = 1, \ldots, q$.

The summation

$$Q(\mathbf{y}) = \sum_{j=1}^{q} \theta_j g_j(\mathbf{y}) \tag{2}$$

is often referred to as the negpotential function, or Hamiltonian, and $Z(\boldsymbol{\theta}) = \sum_{\mathbf{y}} \exp\{Q(\mathbf{y})\}$ is a normalizing constant for the discrete distribution in (1).

Specification of an ERGM involves identifying topological graph features of interest for defining the negpotential (2) with statistics, $g_j(\mathbf{y})$, as counts of such features. Let N be the set of all edges which share a common node and C be the set of all edges which could potentially form a triangle. The negpotential function for the triad model of [10] can be written as

$$\begin{aligned} Q(\mathbf{y}) =& \rho \sum_i y(\mathbf{s}_i) + \sigma \sum_{\mathbf{s}_i, \mathbf{s}_j \in N} y(\mathbf{s}_i) y(\mathbf{s}_j) \\ &+ \tau \sum_{\mathbf{s}_i, \mathbf{s}_j \mathbf{s}_k \in C} y(\mathbf{s}_i) y(\mathbf{s}_j) y(\mathbf{s}_k) \end{aligned} \tag{3}$$

where the sufficient statistics correspond to topological features as the number $g_1(\mathbf{y}) = \sum_i y(\mathbf{s}_i)$ of edges, the number $g_2(\mathbf{y}) = \sum_{i,j} y(\mathbf{s}_i) y(\mathbf{s}_j)$ of 2-stars, and number $g_3(\mathbf{y}) = \sum_{i,j,k} y(\mathbf{s}_i) y(\mathbf{s}_j) y(\mathbf{s}_k)$ of triangles. In (3), ρ represents a density parameter for the graph, σ represents a clustering parameter [10], and τ is a parameter for transitivity. Hence, the dependence structure of an ERGM is defined by the choice of graph features included in the specification (1) or (2); see [11] for details on choosing topological configurations.

Initially, ERGMs included parameters to represent the density, transitivity, and k-stars of the network [10], of which the triad model (3) is a special case. This set of parameters leads to a Markovian dependence structure where two potential edges are conditionally dependent if they share a common node. ERGMs can also be expanded to incorporate more complicated graph topologies [47], exogenous covariate information [12], or summaries of distributions of graph statistics [43].

An ERGM specified through a joint distribution (1) involving a choice of statistics or parameters corresponding to graph topological features in the negpotential function (2) will be referred to here as a traditional ERGM. This specification requires an explicit identification of global network features thought to reflect important aspects of graph topology that potentially have scientific interpretations. For example, the social network interpretation of the transitivity parameter is that friends of friends are more likely to also be friends. Thus, the strength of traditional ERGMs is the ability to describe the graph in terms of understandable global features.

To end this section, we mention that fitting ERGMs to realized networks has been demonstrated to be a difficult task, particularly to a network with a large number of nodes. The model can become degenerate, or place most of its probability on a few, disparate graphs, none of which resemble the

observed network. A large amount of research has been devoted to identifying the cause of this behavior [6], [14], [33], [34], recognizing when it has occurred [37], [19], and proposing modifications to the ERGM to avoid the issue [43], [36], [18]. One hypothesized cause of the behavior are large and growing neighborhoods [38], [35], which lead to the local dependence dominating the global structure. LSGMs are not immune to this behavior, although increased interpretability of the dependence through controlled neighborhoods and saturated graphs can permit an easier identification of when such degeneracy will occur, which we explain in the following section.

3. LOCAL STRUCTURE GRAPH MODEL (LSGM)

Local structure graph models (LSGMs) are a new class of graph models, having a global or joint distribution defined in terms of interpretable and controllable local dependence structures. Two defining characteristics of LSGMs are the specification, for each potential edge marker $\mathbf{s}_i$, of a full conditional distribution, $\Pr(y(\mathbf{s}_i)|y(\mathbf{s}_j); j \neq i)$, for the probability of edge presence or absence ($y(\mathbf{s}_i) = 1$ or 0) and a neighborhood, $N_i = \{\mathbf{s}_j : \mathbf{s}_j \text{ is a neighbor of } \mathbf{s}_i\}$ consisting of graph edges which are "local" to $\mathbf{s}_i$. These two features, together with an assumption of Markov dependence induce a direct functional dependence between graph edges defined to be neighbors,

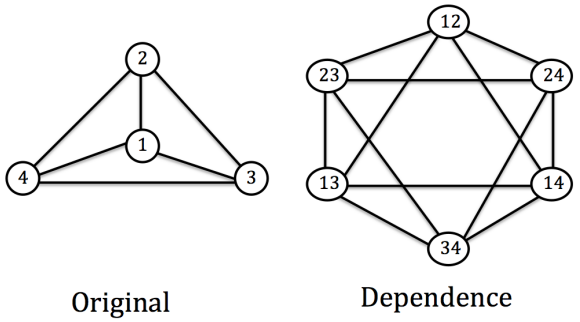
$$\Pr(y(\mathbf{s}_i)|y(\mathbf{s}_j); j \neq i) = \Pr(y(\mathbf{s}_i)|y(\mathbf{s}_j); \mathbf{s}_j \in N_i).$$

This allows the probability of the presence of an edge to be dependent upon the outcomes $y(\mathbf{s}_j)$ of its neighboring edges, $\mathbf{s}_j \in N_i$.

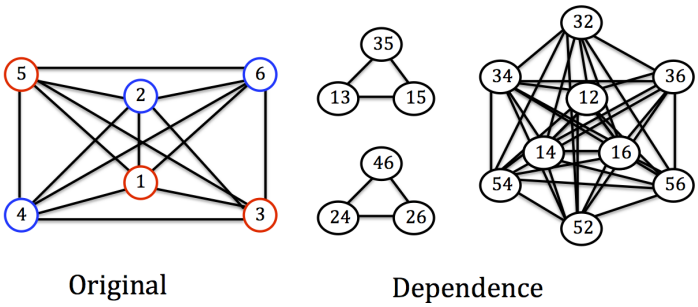
The LSGM can be motivated by a Markov Random Field (MRF) model defined on graph edges. MRF models are commonly encountered in the analysis of spatial data, where effects of spatial dependence are specified conditionally on spatial location information. Intuitively, a response at a particular spatial site might be most heavily influenced by those sites which are spatially neighboring. Through the definition of neighborhoods and conditional specification, a MRF for spatial data allows dependence to be defined through specification of a local structure.

Most common applications of binary MRF models to spatial data can be associated with undirected graphs [23]. In these spatial problems, graph nodes correspond to binary random variables and edges connect nodes which are neighbors. To apply a MRF model to LSGMs, potential edges in the original graph become random variables (locations) in the MRF and neighborhoods are composed of sets of edges that are "near" each other according to some metric. Nodes of the original graph do not appear explicitly in the binary MRF model other than through their role in the edge markers.

To make this connection clearer, consider the neighborhood structure of a LSGM as represented through a dependence graph; see also [10]. Each node in the dependence graph corresponds to a potential edge in the original graph where a connection in the dependence graph indicates the corresponding random variables are conditionally dependent. Two example networks and dependence structures with resulting MRF dependence graphs are shown in Figure 1. The first example appeared in [10] and demonstrates the Markovian dependence as two edges are conditionally dependent if they are incident, or share a node. The second example demonstrates the potential flexibility in the definition of a neighborhood for edges. For this dependence structure, two edges are conditionally dependent if they connect the same number of red, or odd-numbered nodes. The dependence graph here is composed of three disconnected, yet internally fully connected, subgraphs of edges that join the same number of red nodes. In other words, because the nodes of the dependence graph represent potential edge occurrences as random variables in the original graph, the LSGM is placing a MRF on the nodes of the dependence graph.



(a) Incidence definition of dependence. Image recreated from [10]



(b) Two edges are conditionally dependent if they connect the same number of red, or odd numbered, nodes.

Figure 1. Two example networks and dependence structures with resulting dependence graphs. The nodes of the dependence graph corresponds to the edges of the original graph. An edge in the dependence graph indicates conditional dependence between two random variables (i.e., two edges) in the original graph.

The idea of a neighborhood in network analysis has also been used elsewhere, although the definition of a neighborhood has not been consistent. Within LSGMs, a neighborhood defines edges which are conditionally dependent. Neighborhoods are often overlapping and a neighborhood is defined for each potential edge in the network. Our use of the term "neighborhood" (in connection to MRFs and LSGMs) differs from other common uses of this term in network analysis involving block models or community detection [38], where a "neighborhood" often implies a partitioning set of the nodes of the network.

3.1 Specification

To formulate a LSGM, one must specify the form of the conditional distributions and a neighborhood or dependence structure. For simple networks, the goal is to model the presence or absence of edges and thus the conditional distributions are binary, as with the initial LSGM described next. A binary conditional distribution expressed in exponential family form is given by

$$\begin{aligned}\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i)) &= \exp\left[y(\mathbf{s}_i)A_i(\mathbf{y}(N_i))\right. \\ &\quad \left. - B_i(\mathbf{y}(N_i))\right], \quad y(\mathbf{s}_i) = 0, 1 \end{aligned} \tag{4}$$

where A_i is a natural parameter function and $B_i = \log[1 + \exp(A_i(\mathbf{y}(N_i)))]$. In (4), $\mathbf{y}(N_i)$ represents values of the binary random variables (here edges) in the neighborhood of $y(\mathbf{s}_i)$; note $y(\mathbf{s}_i) = 1$ indicates edge occurrence at the marker $\mathbf{s}_i$ of a potential edge. Dependence among random variables is modeled through the natural parameter function, A_i, and a function B_i of A_i. For binary conditionals, a form of the natural parameter function is

$$A_i(\mathbf{y}(N_i)) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i} \eta_{ij}[y(\mathbf{s}_j) - \kappa_j] \tag{5}$$

with m denoting the number of edge markers or total number of possible edges in the network, the sets of parameters, $\{\kappa_i : i = 1, \ldots, m\}$ and $\{\eta_{ij} : i = 1, \ldots, m; \mathbf{s}_j \in N_i\}$, represent global and local structure features of the network model, respectively, and will be discussed in detail in Section 3.2. The parameterization of the natural parameter function in (5) involves centering by global parameters κ_j, as introduced by [7] and [22]. This centered parameterization has been shown to separate global from local structure in (4)–(5) leading to increased interpretation of all model parameters for reasonable amounts of statistical dependence.

Specification of a collection of full conditional distributions must be done in such a way to guarantee the existence of a compatible joint distribution. That is, specifying a graph model based on local structures must be done in a manner such that a valid global model exists and is consistent with the specified local model structure. Our approach based on (4) and (5) will meet this requirement if two conditions are satisfied. First, we assume that the presence or absence of any potential edge can occur with any combination of other edge realizations. This is the so-called positivity condition of [4]. Second, we require that $\eta_{i,j} = \eta_{j,i}$ for all possible pairs of edges indexed by i and j. Then the conditions of Theorem 3 in [23] are satisfied and a joint distribution having the specified conditionals exists. An explicit proof of this for exponential family conditional distributions having natural parameter functions for which (5) is a special case is given in Proposition 1 of [21]. Note that these conditions are slightly stronger than conditions that are necessary for the existence of a joint, which are given in [2]. But, as demonstrated as a constructive procedure in [23], they are necessary for the joint to both exist and be correctly identified up to a constant of proportionality through use of the negpotential function, (2). Using this constructive procedure allows us to identify the global model that corresponds to a specified LSGM, which we do presently.

Section 2 discussed how an ERGM is defined by specifying global topological graph features, or parameters and statistics, to include in the negpotential function (2). Specification of the negpotential function is equivalent to the specification of a joint distribution (1). Graph features chosen to be included in an ERGM implies a set of induced full conditional distributions, but, because an ERGM focuses on global network features, these conditional distributions are not directly modeled or often even identified (and, in fact, such conditional distributions might not even be "local" by depending functionally on every edge in the graph). In contrast, LSGMs are defined by specifying a set of full conditional distributions (typically again with local neighborhood structures) that leads to a constructed negpotential function and thus joint distribution. This relationship between the two different methods of model specification is demonstrated in Figure 2. Using the binary conditionals from (5), a constructed negpotential function for a LSGM can be shown to be [22]

$$\begin{aligned} Q^C(\mathbf{y}) &= \sum_{i=1}^{n} \left[\log\left(\frac{\kappa_i}{1-\kappa_i}\right) - \sum_{\mathbf{s}_j \in N_i} \eta_{ij}\kappa_j\right] y(\mathbf{s}_i) \\ &\quad + \sum_{i=1}^{n} \sum_{\mathbf{s}_j \in N_i} \eta_{ij} y(\mathbf{s}_i) y(\mathbf{s}_j) \end{aligned} \tag{6}$$

determining the joint distribution (1) for $\mathbf{Y}$ under these conditionals; above the superscript C denotes a negpotential Q^C constructed from full conditionals, e.g. (4), in contrast to a direct negpotential formulation (2). The functional form (6) implies a LSGM here can be represented as an ERGM (1) with Markovian dependence, and thus our proposed approach provides an alternate specification of a type of ERGM.

Network model features can generally be divided into those that affect the global structure and those that affect the local structure of random graphs [38], [11]. The global structure can be defined through patterns prevalent in the

Model specified as negpotential (ERGM)

$Q(\mathbf{y}) \Leftrightarrow \Pr(\mathbf{Y} = \mathbf{y}) \rightarrow \{\Pr^I(y(\mathbf{s}_i)|y(\mathbf{s}_j); j \neq i) : i = 1, \ldots, n\}$

Negpotential	$\Leftrightarrow$	Joint distribution	$\rightarrow$	Induced full conditional distributions

Model specified as conditional distributions (LSGM)

$\{\Pr(y(\mathbf{s}_i)|y(\mathbf{s}_j); j \neq i) : i = 1, \ldots, n\} \rightarrow Q^c(\mathbf{y}) \Leftrightarrow \Pr^c(\mathbf{Y} = \mathbf{y})$

Full conditional distributions	$\rightarrow$	Constructed Negpotential	$\Leftrightarrow$	Constructed joint distribution

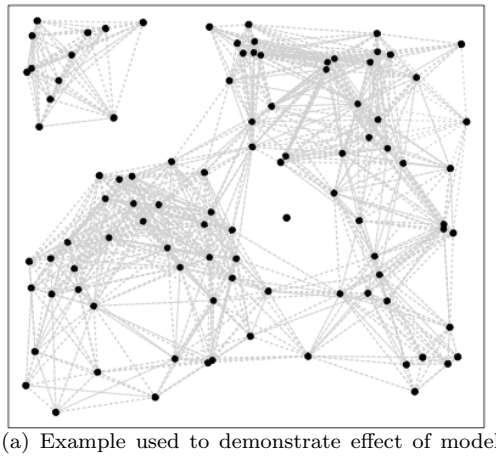
Figure 2. Relationship between the negpotential, joint distribution, and full conditional distributions when either the model is specified as the negpotnetial (ERGM) or full conditionals (LSGM).

overall network, such as density. Features that allow for departures from the global structure at a local level would be classified as local structure. An example of the local structure is transitivity, or tendency towards the closure of individual triangles. By specifying LSGMs through conditional distributions, we are able to directly model such local structure.

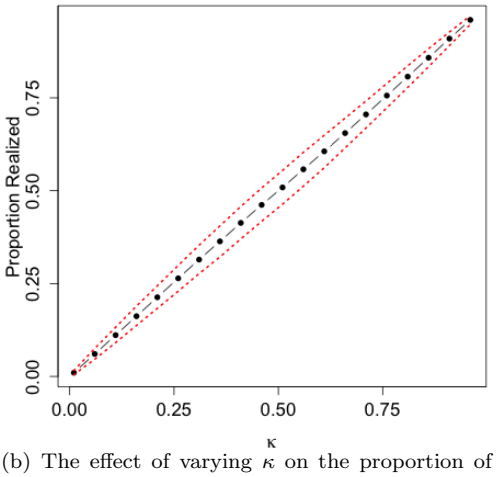
3.2 Model Parameters

Recall that, in LSGMs (4), two sets of parameters control the natural parameter function (5), $\{\kappa_i : i = 1, \ldots, m\}$ and $\{\eta_{ij} : i = 1, \ldots, m; \mathbf{s}_j \in N_i\}$. In its most general form, this model could allow for a different κ_i for every potential edge $y(\mathbf{s}_i)$ and a different η_{ij}, with $\eta_{ij} = \eta_{ji}$, for every $\mathbf{s}_j \in N_i$. However, restrictions are typically placed on these sets of parameters for model identifiability. The effect of the model parameters will be demonstrated for the simplest case where $\kappa_i = \kappa$ and $\eta_{ij} = \eta$ for every i, j in the example network displayed in the first panel of Figure 3. The network consists of 97 nodes and 824 possible edges, where only those pairs of nodes connected in Figure 3 are assigned a random variable for the potential occurrence of an edge and thus indicate those network edges with a positive probability of being realized.

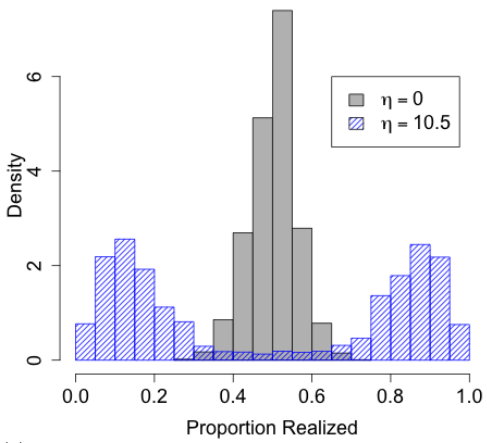
Large-scale structure in (4)-(5) is represented by the first parameter to be discussed, $\kappa \in (0, 1)$. The parameter κ controls the density or proportion of realized edge variables in the overall network and is interpreted as the marginal probability a randomly chosen potential edge will be realized, $y(\mathbf{s}_i) = 1$. As a demonstration of the effect of κ, 10,000 networks were simulated for 20 values of κ and a fixed $\eta = 5$. Due to the conditional specification, a network from a LSGM is naturally simulated with a Gibbs sampler where each potential edge is sampled from its conditional, $\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i))$, in turn. (Given randomly initialized values for all edges, the Gibbs sampler was run with



(a) Example used to demonstrate effect of model parameters. Gray lines indicate all possible edges.



(b) The effect of varying κ on the proportion of possible edges realized. Points represent the median of 10,000 simulations and red, dashed lines 95% envelopes.



(c) Proportion of the 64 possible edges in the disconnected clump of the northwest corner realized for $\kappa = 0.5$ in 10,000 simulations obtained from two values of η.

Figure 3. Example network and a demonstration of the effect of model parameters.

a burn-in of 10,000 complete graph iterations after which sample graphs were retained from subsequent rounds of 500 iterations for thinning.) For the retained simulations, the proportion of realized edges out of those possible was computed for each graph. The second panel of Figure 3 plots the median proportion of realized edges for each κ against κ and red, dashed lines enclose 95% of the simulated proportions. A strong, monotonic relationship exists between κ and the proportion of realized edges with little variability, especially towards the boundaries of the parameter space.

More subtle is the effect on local structure determined by the dependence parameter, η. This parameter quantifies the strength of dependence between neighboring edges and thus controls the extent to which sets of edges either exhibit a neighboring effect or behave independently. When $\eta = 0$, the summation term in the natural parameter function (5) that incorporates the value of neighboring edges is absent, i.e., $A_i(\mathbf{y}(N_i)) = \log\left(\frac{\kappa}{1-\kappa}\right)$, so that each edge formation consequently occurs according to an independent Bernoulli trial with success probability κ (the conditional probability of edge realization is equivalent to the marginal one, as expected under independence). In contrast, larger values of η induce neighbor effects on edge probabilities which can lead to groups of edges behaving in the same manner, e.g., all realized or all not realized. To illustrate, we again simulated 10,000 networks for each of two LSGMs: $\eta = 0$ and $\eta = 10.5$, both with a fixed $\kappa = 0.5$. Now considering the 64 possible edges in the disconnected northwest clump of the example network (Figure 3), we computed the proportion of realized edges among this local subset from each simulation run and the last panel of Figure 3 displays a histogram of these proportions across the 10,000 simulations. When $\eta = 0$, edge probabilities are unaffected by the rest of the network resulting in a distribution of proportions which are symmetric and centered at κ, as displayed in the gray, solid histogram of Figure 3. Few simulations resulted in less than 40% or more than 60% of the edges in this northwest subset being realized. However, for the larger dependence parameter value, $\eta = 10.5$, an induced dependence between neighboring edges is clear. Neighboring edges tended to behave in a group fashion, with edges among this northwest subgroup either mostly all present or mostly all absent, resulting in a bimodal distribution of proportions, as displayed by the blue, dashed histogram of Figure 3. Note that the histogram for this strong dependence scenario is still centered at the expected global proportion of realized edges, $\kappa = 0.5$. That is, even when the local dependence is strong, the marginal mean κ in LSGMs is preserved over multiple simulations, due to the centered parameterization of the natural parameter function (5).

Additional modeling of the local dependence parameter is often necessary. In application of a LSGM, we recommend that this term be adjusted to account for unequal neighborhood sizes. It is common in spatial statistics for neighborhoods of random variables in a MRF to be similar in size, such as occurs, for example, with a four-nearest neighbor structure for a regular spatial lattice. However, neighborhoods for potential edges in LSGMs will often not result in equally-sized neighborhoods (see Figure 9 in Section 4 for an example). To allow the summation term in the natural parameter function (5) to have a uniform effect on edges of varying neighborhood size, we modify dependence parameters as

$$\eta_{ij} = \frac{\eta}{|N_i| + |N_j|} \tag{7}$$

where $|N_k|$ represents the size of the neighborhood of edge $y(\mathbf{s}_k)$. The summation of neighborhood sizes in the denominator of (7) assures that $\eta_{ij} = \eta_{ji}$, guaranteeing the identification of a joint distribution through construction of a negpotential function [23].

A practical parameter space for $\eta \in \mathbb{R}$ is not as well defined as the large-scale parameter, $\kappa \in (0, 1)$. When the local structure of the model overwhelms the global structure, e.g., $|\eta|$ is "too large" compared to κ, the model will become degenerate and place most of its probability on unrealistic network realizations. As a demonstration, the proportion of realized edges in 10,000 simulations of the example network for a LSGM with parameter values $\kappa = 0.5$ and $\eta = 35$ is shown in Figure 4. Almost all edges are realized in all simulations, as the model places most of its probability on the nearly complete graph. This behavior has been recognized for the ERGM and, more generally, in a class of models for interactive systems [45], and is similar to long-range dependence observed in the Ising model [41]. In the ERGM context, large and growing neighborhoods have been identified as a potential cause of degenerate model behavior [38]. A large dependence parameter in LSGMs produces the same essential effect of model degeneracy as having overly large neighborhoods in the ERGM. Both result in summation terms in a negpotential function (e.g. (3), (5)) that dominate terms for marginal probability, which undermines any concept of dependence in the model (as a departure from independence), and thereby ruins the overall model. As a further complication related to similar degeneracy issues in ERGMs, the values for dependence parameters which are inappropriately large in a LSGM, leading to degeneracy, can change between data applications. A recommendation from [22] is to simulate from the fitted LSGM to assure that the simulations appear reasonable given an observed network. Further work in this area is a topic of ongoing research, but the structures of clearly defined neighborhoods in a LSGM can help in diagnosing and treating degeneracy issues related to edge dependence.

3.3 Additional Features

An important issue in formulating LSGMs is how to define meaningful neighborhoods which capture an appropriate dependence structure. Two additional modeling features can be used to assist with this choice, a potentially latent

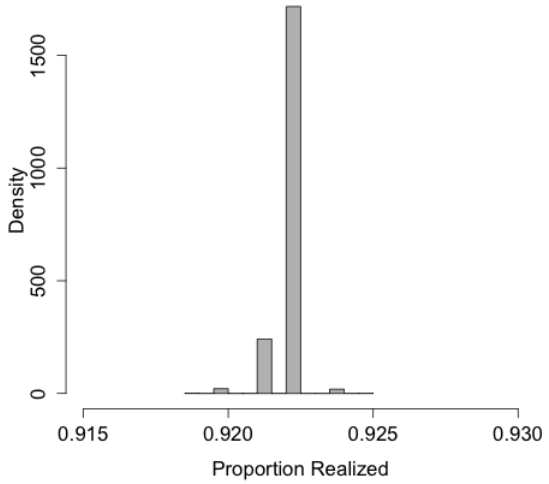
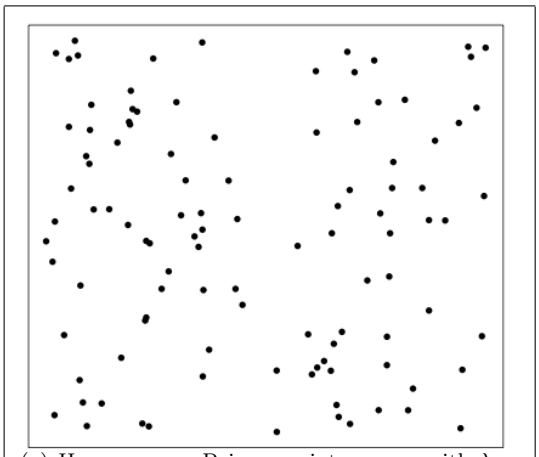


Figure 4. Proportion of realized edges in 10,000 simulations when $\kappa = 0.5$ and $\eta = 35$. The proportion realized does not correspond to the large-scale parameter, $\kappa = 0.5$. This is an example of an area of the parameter space where the model is degenerate.

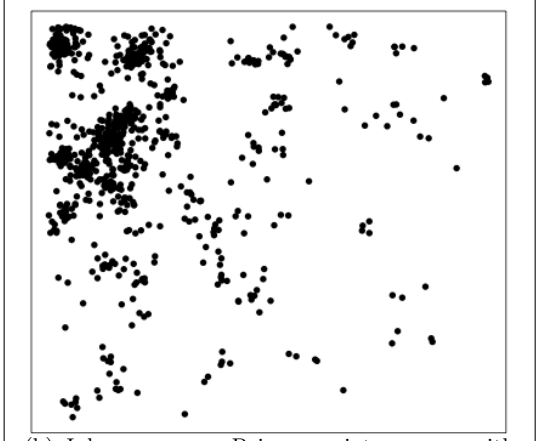
spatial location of nodes (for defining neighborhoods given relevant spatial information) and a saturated graph (for restricting the total number of graph edges).

Recall that a LSGM, with its conditional specification and explicit neighborhood definition, incorporates ideas from the MRF model, a common tool used to analyze georeferenced data. If the nodes of the network have an observed spatial location, such as the location of the buses in the electric power grid [48], the routers of the Internet [30], or the formation site of tornadoes, LSGMs provide a natural way to incorporate this spatial information. Networks for which nodes do not have spatial locations can also be modeled with a LSGM. One option is to impose a latent spatial structure, and such types of spatial locations for nodes could be applied in defining neighborhoods. As an illustration, three example point processes and the resulting node placements are displayed in Figure 5. In this LSGM formulation, latent node locations could potentially be estimated iteratively as a step in a Gibbs sampler. As another example of imposing a spatial location for the nodes, a latent variable on the nodes might be imposed based on auxiliary information. That is, nodal covariate information could potentially be incorporated to define spatial locations for nodes in some unobserved "social space" (cf. [16] and [15]). However, to avoid introducing additional complicating factors, the application of latent point processes will not be considered in the current work; rather, the tornado example presented in Section 4 illustrates how spatial information may be incorporated to formulate a LSGM.

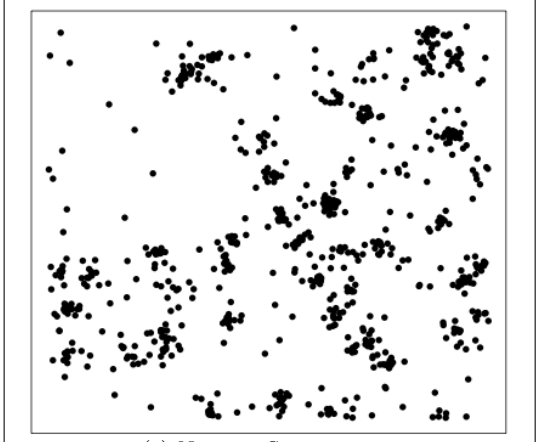
A saturated graph is a second additional LSGM feature that can assist in the specification of meaningful and useful



(a) Homogeneous Poisson point process with $\lambda = 100$.



(b) Inhomogeneous Poisson point process with $\lambda(x, y) = \exp(2 - 4x + 3y)$



(c) Neyman-Scott process

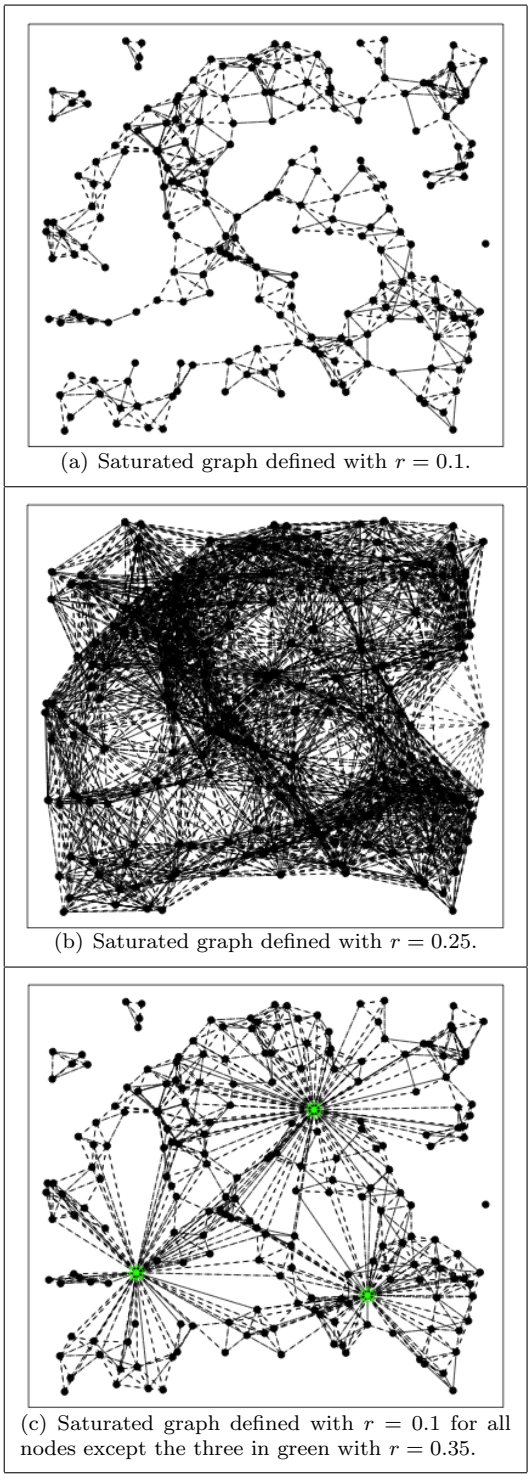
Figure 5. Examples of random node placements through different point processes.

neighborhoods for a network. A saturated graph is defined as those network edges having a positive probability of being realized, so that a saturated graph represents the maximal network realization under consideration. The network displayed in the first panel of Figure 3 is an example of a saturated graph. In some applications, it is reasonable to impose some types of cutoffs in potential edge formations, to produce a meaningful saturated graph for n nodes that is significantly smaller than a graph allowing $\binom{n}{2}$ edges. For instance, Sensor-Actuator Networks (SAN) have a common transmission range which is the maximum distance possible between two connected nodes [32] and, in biological networks, growth factors and diffusible signaling concentration decrease as a function of distance making "long distance" edges improbable [44].

For networks composed of nodes with an observed or latent spatial setting, an intuitive approach for defining saturated graphs is to use a method similar to the formation of a unit disk graph [26]. Given a radius, r, an edge between two nodes within distance r will be defined to have positive probability of being realized. To illustrate, three example saturated graphs with consistent node locations on the unit square are displayed in Figure 6. Radius size is held constant at $r = 0.1$ and $r = 0.25$, respectively, for all nodes in the first two panels. Smaller radius size leads to a graph that is not completely connected with two clusters of nodes disconnected from the majority and one isolated node. In contrast, the resulting saturated graph from the larger radius size is completely connected with no isolated nodes. Additionally in this manner, hubs of nodes could be permitted and modeled by varying the radius size between nodes. The saturated graph displayed in the bottom panel of Figure 6 was created with a radius of $r = 0.1$ for all except the three nodes highlighted in green which had radius $r = 0.35$.

One advantage to imposing a saturated graph is a decrease in the number of random variables to be modeled. As an illustration consider the three saturated graphs of Figure 6, which contain 213 nodes. The small radius of $r = 0.10$ results in 668 possible edges. When the radius is increased to $r = 0.25$, the number of possible edges jumps to 3,467; the combination of radius sizes results in 873 possible edges. Without a saturated graph, edges could form between all pairs of nodes which would result in $\binom{213}{2} = 22,578$ random variables to model. In a small example this may be plausible. However, the direction of current research is to analyze networks with a large number of nodes [9] so that modeling an edge between all pairs of nodes can be computationally prohibitive and perhaps physically unreasonable. Defining a saturated graph based on contextual information for the problem under consideration can be a beneficial modeling strategy.

As alluded to in Section 2, a further consequence of the saturated graph is reasonably sized neighborhoods. Consider an incidence definition of dependence, where two edges which share a node are conditionally dependent in the example network of Figure 6. In the absence of a saturated



(a) Saturated graph defined with $r = 0.1$.

(b) Saturated graph defined with $r = 0.25$.

(c) Saturated graph defined with $r = 0.1$ for all nodes except the three in green with $r = 0.35$.

Figure 6. Examples of saturated graph on same set of nodes for various radius sizes.

graph, each of the 22,578 edge random variables would be dependent upon $2(213-2) = 422$ neighbors, and thus each summation in the natural parameter (5) would include 422 terms. For the same incidence dependence structure, the use of a saturated graph allows the neighborhood size to vary and depend on the number of "nearby" edges. The average neighborhood size for edges in the first panel of Figure 6 is 12.5, 68 for the second panel, and for the final panel each edge is dependent upon 30.67 neighbors, on average. Thus, the use of a saturated graph not only decreases computational time, but also can alleviate overly large neighborhood sizes, which was identified by [38] as a source of model degeneracy.

4. APPLICATION

The atmospheric and meteorological processes that lead to the formation of tornadoes remains an active topic of research. Thunderstorm supercells, which are thunderstorms that contain rotating updrafts, are often the focus of research into tornadogenesis (e.g., [29]) although a non-negligible number of tornadoes seem to occur in conjunction with other strong convection events [46]. In addition, the evolution of supercell mesocyclones and possible production of multiple vortices within one cell have led to the concepts of tornado families and long-track tornadoes that "skip" (e.g.,[1]). Some of the evidence used by atmospheric scientists to develop and evaluate theories of tornadogenesis involves proximity of tornadoes in space and time [1] and viable theories must be able to explain patterns in tornado occurrence. Our interest here lies in determining whether network analysis offers one potential way to characterize such patterns. In particular, we are interested in whether a random graph model of tornado occurrence can be used to generate patterns of tornadoes that agree with observed tornado events in important ways. The data pattern of concern in this application was the proximity in both space and time of tornadoes spawned by the same storm system or event.

4.1 The Network

The data consist of locations and times of origin for tornadoes that occurred in Arkansas during April, 2011. These data were obtained from the United States National Oceanic and Atmospheric Administration (NOAA) National Data Center *Storm Data* severe weather report database. Graph nodes were defined as locations of origin for tornadoes and edges connect two tornadoes from the same storm system. In total, 13 storm systems generated 59 tornados during this time period. The observed graph, presented in Figure 7, resulted from defining two tornadoes to be from the same storm system if they originated within 80 kilometers and two hours of each other (80 kilometers is roughly the fastest that thunderstorms are thought to travel in an hour [25]).

To formulate a LSGM here, we first defined a relatively large saturated graph for the model (i.e., the largest possible graph for realization) having possible edges defined as joining tornadoes that originated within 80 kilometers of each other. This saturated graph contained a total of 292 possible edges and is shown in Figure 8. Neighborhoods for the LSGM were specified as groups of potential edges that were incident in the saturated graph (i.e., all shared a common tornado node). These neighborhoods contained from 2 to 39 potential edges and the frequency distribution of neighborhood sizes is given in Figure 9. In addition to the LSGM, we also considered an independence model with potential edges in the saturated graph conceptualized as occurring through independent, identically distributed Bernoulli trials.

Note that the realized graph is defined to have edges between tornados that are within certain distances of each other in both space and time. In contrast, the saturated graph contains potential edges that are defined in terms of only proximity in space. Both the LSGM and the independence model take possible edges to be given by the saturated graph. The independence model represents the occurrence of these potential edges as equally likely. The LSGM represents the occurrence of edges in a neighborhood group as being related to the occurrence of other edges in the same group. These neighborhood groups consist of tornados clustered even more tightly in space than the saturated graph. If the LSGM can reproduce the spatial and temporal patterns of edges in the realized graph, then partial information on spatial proximity plus the use of neighborhoods may be sufficient to represent the pattern of tornados arising from a series of storm cells over time. This use of neighborhoods is motivated by the concept of families of tornados, in which a single storm cell spawns a sequence of tornados over a limited region in space.

4.2 The Fit of the LSGM

Here we consider the LSGM with a single marginal mean, κ, and single dependence parameter, η, for the Arkansas tornado network. The dependence parameter is adjusted to account for unequal neighborhood sizes as in (7). Point estimates of the model parameters are obtained through a maximization of the log pseudo-likelihood (PL) [5], the summation of the log of the conditional distributions,

$$\log \mathrm{PL} = \sum_i \{y(\mathbf{s}_i) \log[p_i(N_i)] + (1 - y(\mathbf{s}_i)) \log[1 - p_i(N_i)]\}$$

where $p_i(N_i) = E_i(Y(\mathbf{s}_i)|\mathbf{y}(N_i))$ represents the conditional expectation for edge $Y(\mathbf{s}_i)$ given neighboring values $\mathbf{y}(N_i)$,

$$p_i(N_i) = \frac{\exp(A_i(N_i))}{1 + \exp(A_i(N_i))}.$$

Point estimates obtained by maximizing the PL function are known to be generally consistent and asymptotically normal for Markov random field models [13], including the LSGM here. Interval estimates were obtained using parametric bootstrap percentile intervals [8, Chapter 5.3] with

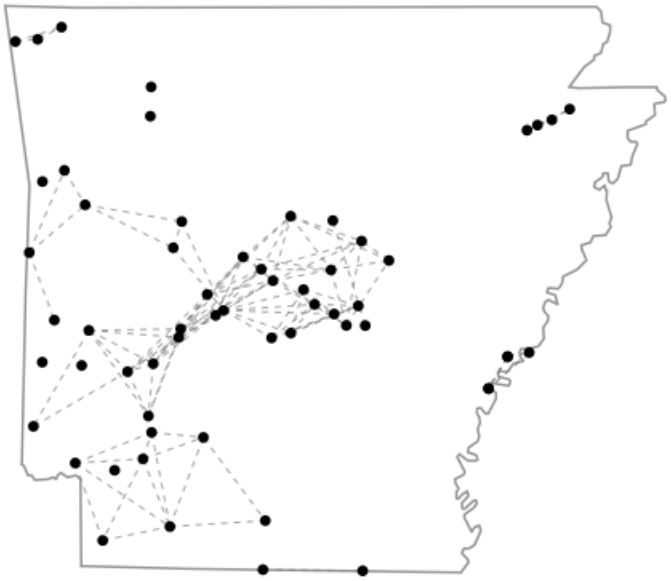


Figure 7. Nodes and realized edges of the Arkansas tornado network.

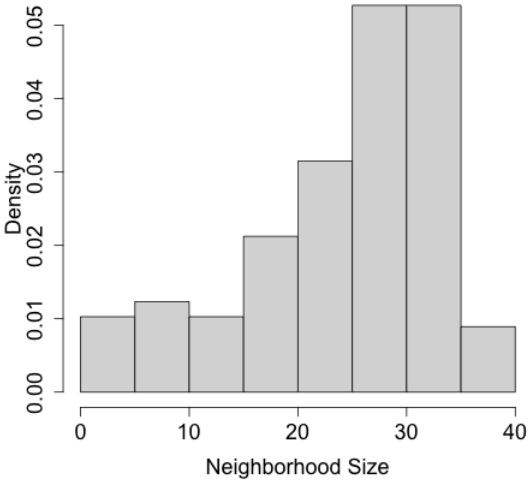


Figure 9. Neighborhood sizes when a saturated graph of $r = 80$ kilometers is used in the analysis of the Arkansas tornado network.

	$\hat{\kappa}$	$\hat{\eta}$
LSGM	0.27 (0.15, 0.75)	8.60 (4.93, 11.07)
Independence	0.43 (0.38, 0.48)	—

Table 1. Point estimates and 90% percentile parametric bootstrap interval estimates for the LSGM and independence model fits to the Arkansas tornado network.

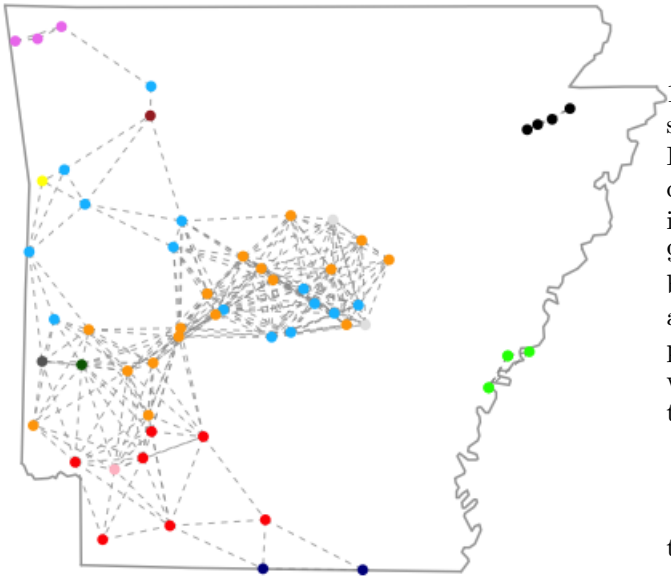


Figure 8. Nodes of the Arkansas tornado network defined by tornadoes that originated in Arkansas during April, 2011. Color corresponds to the event in which the tornado occurred. Edges represent the saturated graph.

10,000 bootstrap renditions of the network data. For each simulated network, parameter estimates were obtained by PL and 90% percentile bootstrap confidence intervals were calibrated from the 5th and 95th percentiles of the resulting empirical distributions of estimates. Point estimates and 90% confidence intervals are shown in the first row of Table 1. For comparison purposes, a maximum PL estimate and parametric bootstrap interval are obtained for the one parameter independence model fit to the tornado network with the dependence parameter η set to zero. The results of this fit are also shown in Table 1.

4.3 Model Assessment

Two methods of model comparison are used to contrast the fits of the LSGM and independence models for the Arkansas tornado network. The first involves a simulation-based analog of the likelihood-ratio test and the second approach attempts to quantify the extent to which the LSGM is able to replicate types of local structure in the observed tornado network.

Comparison of the LSGM and independence models using a likelihood-based test does not fall under the umbrella of regular problems, making identification of an appropriate reference distribution complicated. An alternative is to construct a reference distribution through the use of simu-

lation. We used the difference in maximized log PL as a test statistic, specifically,

$$D = \log \mathrm{PL}(\mathrm{LSGM}) - \log \mathrm{PL}(\mathrm{Indep}). \tag{8}$$

To construct a reference distribution, both the LSGM and independence models were fit to 10,000 networks simulated from the fitted independence model. The p-value for assessing the plausibility of the LSGM relative to the independence model was

$$\frac{1}{10000} \sum_{h=1}^{10000} I(D_h^* > D) \tag{9}$$

based on the test statistics $D_h^*,\ h = 1, \ldots 10,000$, computed for each simulation as in (8); above $I(A)$ denotes the indicator function which takes the value 1 if an event A holds and 0 otherwise. The observed tornado data resulted in a test statistic $D = 14.06$ and associated p-value of 0.0016. Thus, we conclude that the LSGM is superior to the independence model for representing the Arkansas tornado network.

Model assessment may also proceed using an approach similar to that proposed in [19]. In this approach, one chooses a feature of the data that is of interest in the problem, but that is not involved in the manner by which the data inform the estimation procedure used (e.g., sufficient statistics). This data feature is quantified, and the value resulting from the observed data is compared to a distribution of values resulting from data sets simulated from a fitted model.

Our interest here is whether the LSGM can generate data that are more similar to the actual data than data generated by the independence model. The distinguishing feature of the LSGM used in this application is the use of neighborhoods, which were defined externally to the observed data. The effect of including neighboring (potential) edges in the LSGM is to increase or decrease the probabilities of edge realization from the marginal value, depending on whether neighboring edges are realized or not. This causes a certain extent of group behavior because neighborhood membership is symmetric for pairs. That is, if edge j is in the neighborhood of edge k, then edge k is in the neighborhood of edge j as well. Thus, realized potential edges tend to have a large (relative to the marginal) proportion of neighboring edges that are also realized, and similarly for unrealized potential edges. In the independence model, neighborhood information is not used and each potential edge should have roughly the marginally dictated proportion of realized and unrealized neighbors.

The saturated graph defines a potential state space and neighborhoods provide a way to specify relations among potential edges of the saturated graph. The LSGM and independence models provide two mechanisms by which realized graphs can be generated, the LSGM making use of neighborhood information and the independence model ignoring neighborhood information. Given a realization of one of these models, we quantified the manner in which the model reflects local group behavior by computing the proportion of realized neighbors for each realized potential edge from the saturated graph, and the proportion of unrealized neighbors for each unrealized potential edge from the saturated graph. For each potential edge of the saturated graph we then have the proportion of "like" edges in its neighborhood that occur in the particular realization under inspection. The average of these proportions over all potential edges in the saturated graph then provides a measure of the degree of group behavior in the realization. Conducting this procedure for 10,000 simulated realizations from the LSGM results in a reference distribution for that model. Similarly, conducting the procedure for 10,000 simulated realizations from the independence model results in a reference distribution for that model. Conducting the procedure once using the actual tornado network as the realized graph results in a test statistic that can be compared to the two reference distributions in a manner similar to the p-value of expression (9). In this application the test statistic had a value of 0.561, and associated p-values were 0.0002 for the independence model and 0.7481 for the LSGM. We conclude that the independence model is not able to reflect this aspect of local behavior in the tornado network, while the LSGM is able to do so.

The previous model assessment techniques indicate that the patterns in the observed Arkansas tornado network can be appropriately captured by the LSGM, which can detect and accommodate local spatial dependence (i.e., in this case attributable to temporal behavior in tornadoes). To better understand the nature of this local dependence in the LSGM, it is helpful to consider how the conditional probability of an edge (tornado siting) changes under the model as a function of neighboring outcomes. This relationship is plotted in Figure 10. This plots considers an edge, $y(\mathbf{s}_i)$, with 20 neighbors, $|N_i| = 20$, where each of its neighbors also has 20 neighbors, $|N_j| = 20\ \forall \mathbf{s}_j \in N_i$. Under the fitted LSGM, the marginal expectation for this edge being realized is $\hat{\kappa} = 0.27$, regardless of the value of the neighbors (red, dashed vertical line). However, the conditional probability, $p_i(N_i)$, that this edge occurs (black points and line), depends heavily on the number of realized neighboring edges. When all neighboring edges of $y(\mathbf{s}_i)$ are absent, the conditional probability that $y(\mathbf{s}_i) = 1$ is only 0.10. The probability increases monotonically with the number of realized neighboring edges to $p_i(N_i) = 0.89$ when all neighboring edges are realized, i.e., $y(\mathbf{s}_j) = 1\ \forall \mathbf{s}_j \in N_i$.

5. CONCLUSIONS

The goal of this work is to introduce local structure graph models (LSGMs), a new class of models for network analysis, and to demonstrate its use with a simple application. Specification of a LSGM is achieved through conditional distributions which are functions of specified edge neighborhoods, or sets of conditionally dependent edges. An advantage of

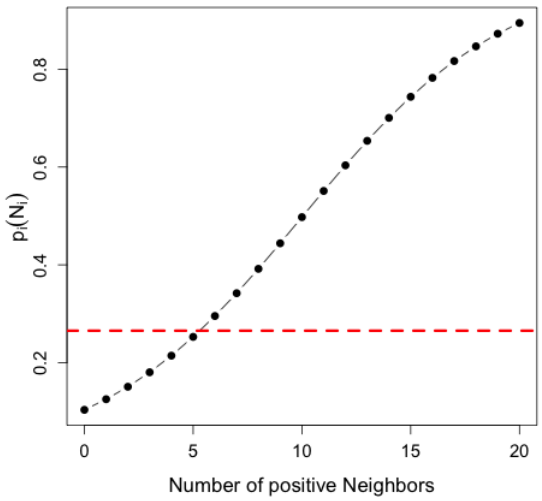


Figure 10. Number of positive neighbors against conditional expectation for a random variable with 20 neighbors. The red, dashed, vertical line represents the marginal expectation of $\hat{\kappa} = 0.27$.

the LSGM approach is an explicit formulation of local dependence in the network, resulting in dependence which is interpretable and controlled by the modeler.

Behavior of LSGMs is controlled by two sets of parameters in a binary model for graph edges: parameters $\{\kappa_i; i = 1, \ldots, m\}$ which represent the global structure for the network model and control the marginal probabilities of edge realization in the network, and parameters $\{\eta_{ij}; i \neq j\}$, which capture the local model structure and can be interpreted as dependence parameters. If dependence parameters become too large, LSGMs can become degenerate, a common modeling consideration for models of interactive systems which encompasses ERGMs. However, because LSGMs are connected to edge neighborhoods through their specification, this aspect may help in formulating and diagnosing models which avoid model degeneracy through controlled, local dependence parameters; this is a topic of on-going investigation. Spatial location of nodes and a saturated graph are introduced to aid in avoiding model degeneracy. These features are not required to the specification of a LSGM, as the form of the conditional distributions and neighborhood structure is all that is necessary.

An extension to LSGMs is the inclusion of auxiliary information into either the global or local (dependence) structure. This can be accomplished through additional modeling of κ, η, or the neighborhoods. Explicit modeling of transitivity, or dependence between triples of random variables, will also require an additional extension. This is due to the fact that the constructed negpotential of a LSGM in (6) includes an assumption of pairwise-only dependence, where dependent sets of random variables of size greater than two are not directly modeled. Although this assumption is often appropriate for the common spatial application of a MRF model, it may be less suitable for the analysis of some networks.

ACKNOWLEDGEMENTS

The authors wish to thank an associate editor and two referees for comments which helped to improve of an earlier version of the paper. The work was supported in part by the Sandia National Laboratories Laboratory-Directed Research and Development Program [3]. Dr. Nordman's research was partially supported by NSF DMS-1406747.

Received 3 March 2015

REFERENCES

[1] Agee, E., Snow, J., and Clare, P. (1976). Multiple vortex features in the tornado cyclone and the occurrence of tornado families. *Monthly Weather Review* **104**, 5, 552–563.

[2] Arnold, B. C. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association* **84**, 405, 152–156.

[3] Berger-Wolf, T., Berry, J. W., Bhowmick, S., Casleton, E., Kaiser, M., Leung, V. J., Nordman, D. J., Phillips, C. A., Pinar, A., Robinson, D. G., and Wilson, A. G. (2014). Statistically significant relational data mining: LDRD report. Tech. Rep. SAND2014-1105. February.

[4] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society* **36**, 2, 192–236.

[5] Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24*, 179–195.

[6] Bhamidi, S., Bresler, G., and Sly, A. (2008). Mixing time of exponential random graphs. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE Computer Society, 803–812.

[7] Caragea, P. C. and Kaiser, M. S. (2009). Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics* **14**, 3 (Sept.), 281–300. http://www.springerlink.com/index/10.1198/jabes.2009.07032.

[8] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

[9] Fienberg, S. E. (2012). Brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics* **21**, 4, 825–839. http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.738106.

[10] Frank, O. and Strauss, D. (1986). Markov Graphs. *Journal of the American Statistical Association* **81**, 395, 832–842.

[11] Goodreau, S. M. (2007). Advances in Exponential Random Graph (p*) Models Applied to a Large Social Network. *Social networks* **29**, 2 (May), 231–248.

[12] Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., and Morris, M. (2008). A statnet tutorial. *Journal of statistical software* **24**, 9, 1.

[13] Guyon, X. (1995). *Random fields on a network: modeling, statistics, and applications*. Springer.

[14] Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. Working Paper 39, Center for Statistics and the Social Sciences, University of Washington, Seattle.

[15] Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society* **170**, 2, 301–354.

[16] 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**, 460 (Dec.), 1090–1098. http://pubs.amstat.org/doi/abs/10.1198/016214502388618906.

[17] Hoff, P. D. and Ward, M. D. (2005). Analyzing dependencies in international relations: commerce, capitalism, conflict, cooperation, and democracy. In *46th Annual Convention of the International Studies Association*. Honolulu, HI, 1–20.
[18] Hunter, D. R. (2007). Curved exponential family models for social networks. *Social networks* **29**, 2, 216–230.
[19] Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association* **103**, 481, 248–258.
[20] Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics* **21**, 4, 856–882.
[21] Kaiser, M., Cressie, N. A., and Lee, J. (2002). Spatial mixture models based on exponential family conditional distributions. *Statistica Sinica* **12**, 2, 449–474.
[22] Kaiser, M. S., Caragea, P. C., and Furukawa, K. (2012). Centered parameterizations and dependence limitations in Markov random field models. *Journal of Statistical Planning and Inference* **142**, 7, 1855–1863.
[23] Kaiser, M. S. and Cressie, N. (2000). The construction of multivariate distributions from Markov random fields. *Journal of Multivariate Analysis* **73**, 2, 199–220.
[24] Kolaczyk, E. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer.
[25] Krider, E. (1999). Thunderstorms and lightning. In *Encyclopaedia Britannica*.
[26] Kuhn, F., Moscibroda, T., and Wattenhofer, R. (2004). Unit disk graph approximation. In *DIALM-POMC*. Philadelphia, Pennsylvania, 17–23.
[27] Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. (2010). Kronecker Graphs : An Approach to Modeling Networks. *Journal of Machine Learning Research 11*, 985–1042.
[28] Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, Suppl 2, S186–S188.
[29] Markowski, P. M. and Richardson, Y. P. (2009). Tornadogenesis: Our current understanding, forecasting considerations, and questions to guide future research. *Atmospheric Research* **93**, 1, 3–10.
[30] Neumayer, S. and Modiano, E. (2010). Network reliability with geographically correlated failures. In *Proceedings IEEE INFOCOM*. IEEE, 1–9.
[31] Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96**, 455, 1077–1087.
[32] Onat, F. and Stojmenovic, I. (2007). Generating random graphs for wireless actuator networks. In *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*. 1–12.
[33] Park, J. and Newman, M. E. (2004). Solution of the two-star model of a network. *Physical Review E* **70**, 6, 066146.
[34] Park, J. and Newman, M. E. (2005). Solution for the properties of a clustered network. *Physical Review E* **72**, 2, 026136.
[35] Pattison, P. and Robins, G. (2002). Neighborhood–based models for social networks. *Sociological Methodology* **32**, 1, 301–337.
[36] Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks* **29**, 2, 192–215.
[37] Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association* **106**, 496, 1361–1370.
[38] Schweinberger, M. and Handcock, M. S. (2012). Hierarchical exponential-family random graph models with local dependence.
[39] Simpson, S. L., Hayasaka, S., and Laurienti, P. J. (2011). Exponential random graph modeling for complex brain networks. *PloS ONE* **6**, 5, e20039.
[40] Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification* **14**, 1, 75–100.
[41] Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* **3**, 2, 1–40.
[42] Snijders, T. A. B. (2007). Contribution to the discussion of Handcock, M. S., A. E. Raftery, and J. M. Tantrum, Model-based clustering for social networks. *Journal of the Royal Statistical Society* **170**, 2, 301–354.
[43] Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology 36*, 99–153.
[44] Sporns, O., Chialvo, D. R., Kaiser, M., Hilgetag, C. C., and others. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences* **8**, 9, 418–425.
[45] Strauss, D. (1986). On a General Class of Models for Interaction. *SIAM Review* **28**, 4, 513–527.
[46] Trapp, R. J., Tessendorf, S. A., Godfrey, E. S., and Brooks, H. E. (2005). Tornadoes from squall lines and bow echoes. part i: Climatological distribution. *Weather and forecasting* **20**, 1, 23–34.
[47] Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks. *Psychometrika* **61**, 3, 401–425.
[48] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**, 6684 (June), 440–442. http://www.ncbi.nlm.nih.gov/pubmed/9623998.

Emily Casleton
Statistical Sciences Group
Los Alamos National Laboratory
Los Alamos, NM 87545
E-mail address: ecasleton@lanl.gov

Daniel Nordman
Department of Statistics
Iowa State University
Ames, IA 50011
E-mail address: dnordman@iastate.edu

Mark Kaiser
Department of Statistics
Iowa State University
Ames, IA 50011
E-mail address: mskaiser@iastate.edu