Modeling Transitivity in Local Structure Graph Models

Emily Casleton, Mark S. Kaiser, Daniel J. Nordman

Abstract: Local Structure Graph Models (LSGMs) describe network data by modeling, and thereby controlling, the local structure of networks in a direct and interpretable manner. Specification of such models requires identifying three factors: a saturated, or maximally possible, graph; a neighborhood structure of dependent potential edges; and, lastly, a model form prescribed by full conditional binary distributions with appropriate "centering" steps and dependence parameters. This last aspect particularly distinguishes LSGMs from other model formulations for network data. In this article, we explore the expanded LSGM structure to incorporate dependencies among edges that form potential triangles, thus explicitly representing transitivity in the conditional probabilities that govern edge realization. Two networks previously examined in the literature, the Faux Mesa High friendship network and the 2000 college football network, are analyzed with such models, with a focus on assessing the manner in which terms reflecting two-way and three-way dependencies among potential edges influence the data structures generated by models that incorporate them. One conclusion reached is that explicit modeling of three-way dependencies is not always needed to reflect the observed level of transitivity in an actual graph. Another conclusion is that understanding the manner in which a model represents a given problem is enhanced by examining several aspects of model structure, not just the number of some particular topological structure generated by a fitted model.

Key words and phrases: Conditionally specified models, Network analysis, Network model assessment, Random graphs, Transitivity

1 Introduction

Transitivity, which is related to the number of realized triangles in a network, has drawn a good deal of attention in network analysis. In computer science and statistical physics, transitivity has been identified as one of three features that graph-generating algorithms should attempt to recreate [Lancichinetti et al., 2008]. In the context of social networks, transitivity can be heuristically be described as "friends of a common friend tend to also be friends." Although there is no single agreed upon quantification of this concept [Kolaczyk, 2009], indices of transitivity typically include ratios of the number of triangles to structures that could be triangles with the addition of one edge, called two-stars.

In some discussions of social networks, the presence of moderate to high levels of transitivity seems to be almost pre-supposed as a ubiquitous feature of network data involving social interactions [Snijders et al., 2006, Vasques Filho and O'Neale, 2020]. However, it also has been demonstrated that certain explicit model terms for transitivity (e.g., triangles) in exponential random graph models (ERGMs) can potentially promote a phenomenon called model degeneracy [Robins et al., 2007], which has led to a number of modifications for modeling transitivity [Hunter et al., 2008a, Hunter and Handcock, 2006, Snijders et al., 2006].

In this article, we examine the representation of transitivity by a class of random graph models, called local structure graph models (LSGMs) [Casleton et al., 2017, 2020], for describing the incidence of graph edges. These models are formulated on binary Markov random fields by using full conditional distributions with "centering" steps intended to separate the effects of mean and dependence parameters. LSGMs, as well as ERGMs, are able to model dependencies among potential edges, but LSGMs aim to directly do so in a certain type of interpretable manner without parameter confounding among model terms. In connection to this, LSGMs allow an investigator to exercise a degree of control over the modeling of edge dependencies through the specification of local neighborhoods for potential edges. While ERGMs can, of course, also consider local dependencies [Morris et al., 2008], formulations of ERGM do not traditionally attempt to avoid parameter confounding in an analogous manner to LSGMs with centering [Wang et al., 2013, Section 4]. The effects of positive dependence in a LSGM on realizing clusters of edges (or clusters of edge absence) have been demonstrated in Casleton et al. [2017] under an assumption of pairwise-only dependence [Besag, 1974]. Here, we examine the issue of triangle formation with models having pairwise-only dependence and a more recent extension of LSGM to include three-way dependence terms [Casleton et al., 2020].

The remainder of the article is organized as follows. Section 2 introduces two example networks that will be considered in the sequel and indicates how these examples have been used in other contexts. A description of the LSGM approach will be presented in Section 3, and three associated devices for modeling potential edges and their dependencies will be described in Section 4 for the two example networks. The results of fitting a number of LSGMs to the examples are presented in Section 5, and Section 6 discusses some of the implications of these results for the representation of transitivity in networks. Additional theoretical results for LSGMs are provided supplementary materials, establishing that the LSGMs considered here belong to a type of curved exponential family and possess certain model stability properties (i.e., so-called S-stability, Schweinberger [2011], Kaplan et al. [2020]) that some ERGMs lack.

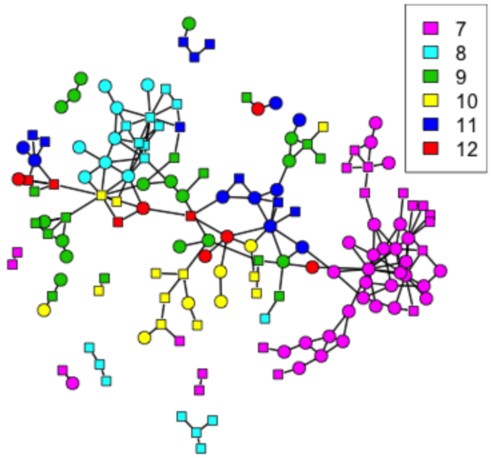
2 Example Networks

The first dataset to be considered here is known as Goodreau's Faux Mesa High friendship network [Handcock et al., 2014]. The 205 nodes represent students in grades 7–12 from one school district in rural, western United States. Undirected edges form between two nodes if both students identified each other as a friend on an in-school survey, where students were given a roster of all students and asked to list up to five of their closest male and five closest female friends. This is known as a mutualized friendship network and is a common conceptualization of friendship in social network analysis [Hunter et al., 2008a].

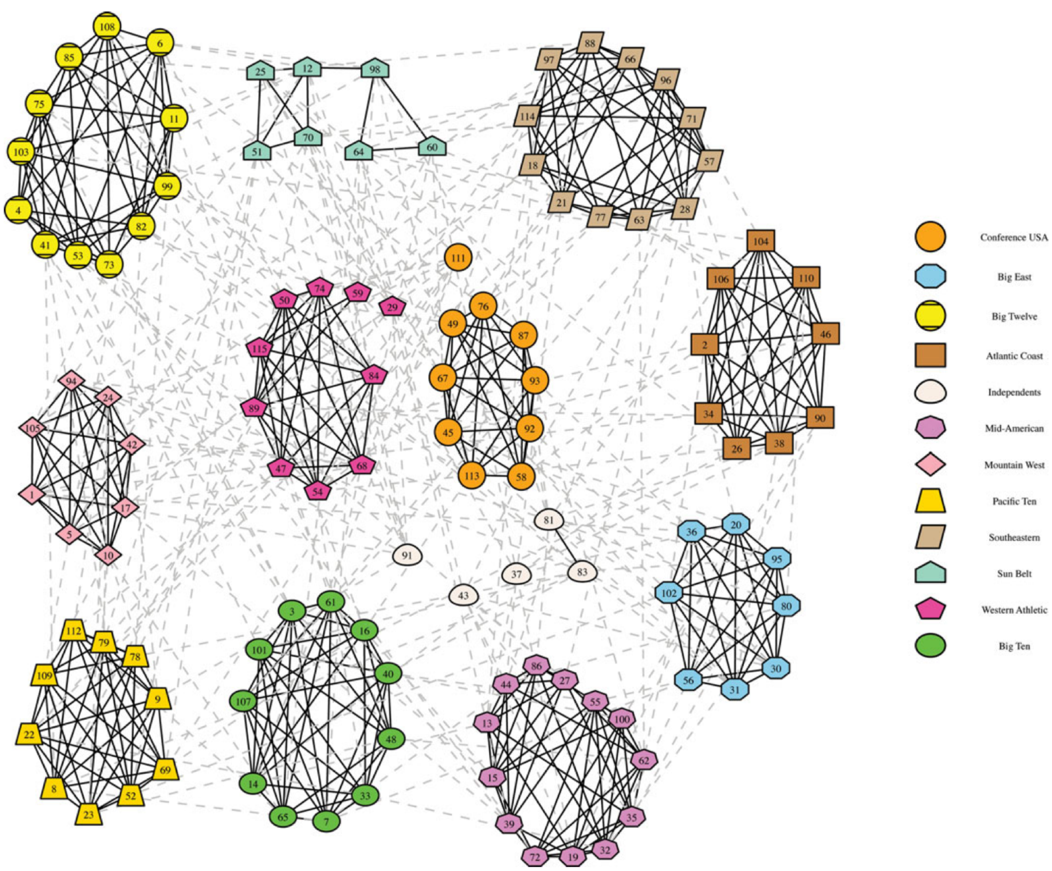
The network is based on School 10 from the first wave of the Add Health longitudinal study, though is simulated to preserve the confidentiality of the students. School 10 was chosen by Hunter et al. [2008a] because its analysis was found to be similar to a simultaneous analysis of 59 schools. For more information on the Add Health survey of adolescent behavior [Resnick et al., 1997], see `http://www.cpc.unc.edu/addhealth/`. Nodal attributes of grade, sex and race were also collected [Hunter et al., 2008a, Handcock et al., 2014]. Students who did not take the survey or were not on the roster were removed, and any missing nodal attributes were imputed with random samples from a weighted distribution of known attributes. An ERGM, having terms to account for density, attribute information and transitivity, was then fit to the complete data, and the Faux Mesa High network represents a single simulation from the fitted ERGM [Goodreau et al., 2008]. These network data are distributed in the `ergm` package for `R` [R Core Team, 2013, Handcock et al., 2014], and a visualization of this network, as it appears from plotting with the `ergm` package [Hunter et al., 2008b], is shown in Figure 1.

Although the Faux Mesa High network is simulated, it has been argued to be a realistic representation of an adolescent friendship network [Hunter et al., 2008a] and has been widely used as an example network. Morris et al. [2008] and Bender-deMoll et al. [2008] have applied this mutualized friendship network to demonstrate various aspects of ERGMs. By also using the Faux Mesa High network, a method for mitigating a virus attack was tested by Kashirin and Dijkstra [2013], a graph generation algorithm using hyperplane features was demonstrated by Lunga and Kirshner [2011], and a visualization method for disease transmission was considered by Lofgren [2012].

The second network of interest here is constructed from American football games played between NCAA Division I universities during the 2000 season. Nodes represent the 115 schools with a Division I college football team during that season, and an edge exists between two nodes if the teams competed against each other. Most college football programs are



(a) Faux Mesa High network. Colors represent the grade and symbol shapes represent sex, where males are squares and females are circles. Figure adapted from Hunter et al. [2008b].



(b) NCAA Football network. Color and shape of nodes represent the conference. In- and out-of-conference edges are represented differently. Figure taken from Guo et al. [2013].

Figure 1: Visualizations of the (a) Faux Mesa High and (b) College Football networks.

members of an athletic conference, or a group of teams who predominately compete against each other. An exception are schools classified as Independent, who do not belong to any conference. A plot of the nodes, with realized edges and conference designations, appears in Figure 1 as obtained from Guo et al. [2013].

This Football network was compiled by Girvan and Newman [2002] and has been used as a test bench to evaluate community detection techniques for uncovering the conference structures from edge configurations in the network [Guo et al., 2013]. Our goal is not to determine these features, but rather the conference structures will be considered as known attributes for providing information toward modeling the presence of edges among nodes. Edges will be categorized as either in-conference (for games between two schools from the same conference) or out-of-conference (for games between schools in different conferences or games involving an Independent school).

3 Conditionally Specified Models and Centering

Mathematically, we define a network (or random graph) by a fixed set of n nodes involving m possible edges. Each of the m possible edges is assigned a binary random variable $Y(\mathbf{s}_i)$, where an edge marker $\mathbf{s}_i = \{c_i, r_i\}$ indicates the two nodes, c_i and r_i, that an edge could potentially join. Edge values are binary and designate the presence $Y(\mathbf{s}_i) = 1$, or absence $Y(\mathbf{s}_i) = 0$, of an edge at a marker $\mathbf{s}_i$. Covariate information on the nodes can also be associated with the marker $\mathbf{s}_i$ and may be designated as a possible vector-valued $\mathbf{x}(\mathbf{s}_i)$.

For examining the two example networks (Section 2), the random graph model used is a local structure graph model (LSGM) approach. For each edge variable $Y(\mathbf{s}_i)$, this network model involves specifying a full conditional distribution, given by $P(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\{y(\mathbf{s}_j) : j \neq i\}), y(\mathbf{s}_i) \in \{0, 1\}$, along with an associated set of dependent edges, known as a neighborhood $N_i \subset \{\mathbf{s}_j\}_{j=1}^m \backslash \{\mathbf{s}_i\}$. A Markov dependence assumption simplifies the conditional distributions

$$P(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\{y(\mathbf{s}_j) : j \neq i\}) = P(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i)), \quad i = 1, \ldots, m,$$

so that these depend functionally only on the observed values $\mathbf{y}(N_i) \equiv \{y(\mathbf{s}_j) : \mathbf{s}_j \in N_i\}$ of neighbors to the edge variable at $\mathbf{s}_i$. The advantage of a conditional specification with an explicit neighborhood definition is control over the interpretation of the local structures in the network, particularly in combination with parameter centering to follow along with neighborhood-size adjustments to dependence effects (e.g., (3)–(5)). The application of these features to network analysis was introduced in Casleton et al. [2017] and extended to include higher-order dependence in Casleton et al. [2020].

LSGMs can be interpreted as an alternate method of specifying another more common class of random graphs known as exponential random graph models (ERGMs). In contrast to LSGMs, traditional formulations of ERGMs specify a model for a network through a joint distribution, often by identifying particular global topological graph features to be included as statistics in the log-linear term of the joint distribution [Kolaczyk, 2009]. Effects frequently included are edge density, transitivity, block effects or covariate effects. Sets of dependent edge random variables and conditional distributions are *induced* by the terms included in a ERGM, rather than *explicitly specified* as in a LSGM. Both the traditional formulations of ERGMs and LSGMs have joint distributions in Gibbsian form [Casleton et al., 2020], but these joint distributions are not central to the prescription of a LSGM here. However, the joint distribution for a LSGM may be constructed, for example, from the set of specified full conditional distributions under certain non-restrictive conditions [Kaiser and Cressie, 2000]; see the supplementary materials for more details on this joint distribution.

A LSGM involves an application of a binary Markov Random Field (MRF) model to the graph edges. This model, originally referred to as the auto-logistic model, was introduced in Besag [1974] and is commonly used to analyze spatially geo-referenced binary data due to its ability to model dependence. Consider the conditional binary distribution written in exponential family form as

$$\Pr(Y(\mathbf{s}_i) = y(\mathbf{s}_i)|\mathbf{y}(N_i)) = \exp[y(\mathbf{s}_i)A_i\{\mathbf{y}(N_i)\} - B_i\{\mathbf{y}(N_i)\}], \quad y(\mathbf{s}_i) = 0, 1,$$

where $A_i\{\mathbf{y}(N_i)\}$ is referred to as the natural parameter function and where $B_i\{\mathbf{y}(N_i)\}$ is a function of A_i, given as $B_i\{\mathbf{y}(N_i)\} = \log[1 + \exp(A_i\{\mathbf{y}(N_i)\})]$ for an auto-logistic model. The natural parameter function of Besag [1974] is given by

$$A_i\{\mathbf{y}(N_i)\} = \alpha_i + \sum_{\mathbf{s}_j \in N_i} \eta_{ij} y(\mathbf{s}_j), \quad i = 1, \ldots, m, \tag{1}$$

where α_i are leading constants and the $\eta_{ij} = \eta_{ji}$ are dependence parameters between pairs of random variables. This formula assumes pairwise-only dependence so that dependence is only modeled between pairs of dependent edges.

Let $S \equiv \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m\}$ be the collection of all edge markers and let V denote any non-empty subset of S. In order to incorporate higher-order dependence terms by explicitly modeling dependence between sets of random variables of size greater than two, Lee et al. [2001] presented a necessary form for the natural parameter function of binary conditional distributions as

$$A_i\{\mathbf{y}(N_i)\} = \alpha_i + \sum_{V:\mathbf{s}_i \in V} \left[\theta_V \prod_{\mathbf{s}_j \in V \setminus \{\mathbf{s}_i\}} y(\mathbf{s}_j)\right], \quad i = 1, \ldots, m, \tag{2}$$

where α_i are similar leading constants and where the θ_V represent dependence parameters between sets of random variables that must be invariant to any permutation of the indices in V. Although this allows for dependence to be modeled between any sized set of dependent edges, all possible subsets of S are hardly ever considered. One reason is that the Hammersly-Clifford Theorem [Cressie, 1993, p. 417] implies $\theta_V = 0$ unless the edges in V represent a clique. A clique is a single random variable or any set of random variables such that all random variables in the set are neighbors of every other random variable in the set. Thus, the neighborhood definitions specified for the edges of the network play a large role in which terms of the natural parameter function are included.

The parameterization of the natural parameter functions in (1) and (2) is often referred to as the original, or uncentered, parameterization. This form has been shown to lead to confounding and interpretation issues with the parameters, particularly in separating the large and small scale model structures (e.g., overall mean, α_i vs. dependence effects, η_{ij} in (1)) and when attribute information is included in the model (e.g., covariates, $x(\mathbf{s}_i)$ in α_i). To allow a more uniform and separable interpretation of parameters across reasonable levels of statistical dependence, Caragea and Kaiser [2009] proposed a centered parameterization of the natural parameter function for binary conditional distributions. For simplicity, we will assume common dependence parameters for each type of clique size and then adjust (i.e., scale) to accommodate for potentially varying sizes among neighborhoods and other dependence sets [Casleton et al., 2017]. The parameterization presented in Caragea and Kaiser [2009], also assuming pairwise-only dependence, can be written as

$$A_i\{\mathbf{y}(N_i)\} = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\mathbf{s}_j \in N_i} \left[\frac{\eta_2}{|N_i| + |N_j|}\right] (y(\mathbf{s}_j) - \kappa_j), \tag{3}$$

where the parameter $\kappa_i \in (0,1)$ represents the large scale structure, η_2 is a real-valued dependence parameter between pairs of edges, and $|N_\ell|$ denotes the size of the neighborhood for the edge variable at $\mathbf{s}_\ell$. Hence, (3) is an important type of modification of (1) using dependence parameters $\eta_{ij} = \eta_{ji} = \eta_2/(|N_i| + |N_j|)$ that allow the value of η_2 to have the same interpretation across neighborhoods of varying sizes and maintain invariance of dependence parameters to permutation of their indices (analogously to the parameters θ_V in (2)), which is one condition needed for the joint distribution to be identified through use of the negpotential function [Kaiser and Cressie, 2000]. The centered parameterization was extended by Casleton et al. [2020] to include cliques of size three with resulting natural parameter function

$$\begin{aligned} A_i\{\mathbf{y}(N_i)\} = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) &+ \sum_{\mathbf{s}_j \in N_i} \left[\frac{\eta_2}{|N_i| + |N_j|}\right] (y(\mathbf{s}_j) - \kappa_j) \\ &+ \sum_{\{\mathbf{s}_j, \mathbf{s}_k\} \in \mathcal{C}_i^3} \left[\frac{\eta_3}{|\mathcal{C}_i^3| + |\mathcal{C}_j^3| + |\mathcal{C}_k^3|}\right] (y(\mathbf{s}_j)y(\mathbf{s}_k) - \kappa_j\kappa_k), \end{aligned} \tag{4}$$

where $\mathcal{C}_i^3 \equiv \{\{\mathbf{s}_j, \mathbf{s}_k\} : \mathbf{s}_i, \mathbf{s}_j \text{ and } \mathbf{s}_k \text{ are neighbors of each other}\}$ represents the collection of all cliques of size three involving the random variable $Y(\mathbf{s}_i)$ and where $|\mathcal{C}_i^3|$ denotes the cardinality. Note that (4), like (3), corresponds to a form of $A_i\{\mathbf{y}(N_i)\}$ as in (2). While the LSGM framework does not directly use a joint data density, the joint implied by the centered binary conditions (3)-(4) can be shown to belong to a class of curved ERGMs [Snijders et al., 2006, Hunter, 2007] and possess features of model stability; the supplementary materials provide a formal treatment of these properties.

By definition, a clique of size three is a set of three possible edges, at markers $\{\mathbf{s}_i, \mathbf{s}_j, \mathbf{s}_k\}$ say, which are all mutual neighbors, i.e., $\mathbf{s}_u \in N_v \cap N_w$ for $u, v, w \in \{i, j, k\}$ with $u \neq v, w$. Hence, potential cliques of size three are determined by the specification of neighborhoods. A common neighborhood definition in network analysis is incidence, where two potential edges are considered neighbors (i.e., $\mathbf{s}_i \in N_j$ and $\mathbf{s}_j \in N_i$) if they share a common node: markers $\mathbf{s}_i \equiv \{c_i, r_i\}$ and $\mathbf{s}_j \equiv \{c_j, r_j\}$ have a non-empty intersection with respect to some nodes c_i, r_i, c_j, r_j. Configurations of edges, or subgraphs [Frank and Strauss, 1986], that lead to cliques of size three under an incidence definition of dependence are 3-stars and triangles (see Figure 2). The dependence term in (4) may be partitioned based on the type of subgraph, or only a particular subgraph can be modeled. For example, if only cliques of size three corresponding to triangles are considered, the natural parameter function can be written as

$$\begin{aligned} A_i\{\mathbf{y}(N_i)\} = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) &+ \sum_{\mathbf{s}_j \in N_i} \left[\frac{\eta_2}{|N_i| + |N_j|}\right] (y(\mathbf{s}_j) - \kappa_j) \\ &+ \sum_{\{\mathbf{s}_j, \mathbf{s}_k\} \in \mathcal{T}_i} \left[\frac{\eta_3}{|\mathcal{T}_i| + |\mathcal{T}_j| + |\mathcal{T}_k|}\right] (y(\mathbf{s}_j) y(\mathbf{s}_k) - \kappa_j \kappa_k), \end{aligned} \tag{5}$$

where $\mathcal{T}_i \equiv \{\{\mathbf{s}_j, \mathbf{s}_k\} : \mathbf{s}_u \cap \mathbf{s}_v \neq \emptyset \text{ for } u, v \in \{i, j, k\}\}$ gives the set of all triangle-type cliques involving the random variable $Y(\mathbf{s}_i)$, with corresponding size denoted as $|\mathcal{T}_i|$.

Note that it is not necessary to include all lower order dependence terms in the natural parameter functions (4) or (5). For example, a valid model will result when excluding the pairwise dependence term (i.e., taking $\eta_2 = 0$ there). Finally, we note that the large-scale

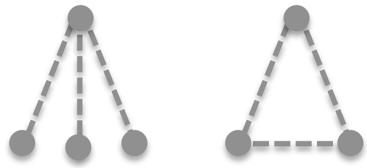


Figure 2: Subgraphs corresponding to cliques of size 3 given an incidence definition of dependence: a 3-star (left) and triangle (right). Note here that the 3-star and triangle subgraphs visualize and represent relations of model dependence among potential edges, which follow from neighborhood definitions. This aspect differs from the realization, or actual incidence, of edges among nodes that may be observed in a random graph.

parameter $\kappa_i \in (0,1)$ appearing in a natural parameter function (3)–(5) for an edge $Y(\mathbf{s}_i)$ could be further modeled based covariate information $x(\boldsymbol{s}_i)$ when available, e.g., using logit link $\log(\kappa_i/(1-\kappa_i)) = \beta' x(\boldsymbol{s}_i)$ with a regression parameter vector β (cf. Casleton et al. [2020]). Alternatively, we might specify κ_i categorically, with each κ_i associated with a type or group of other edges, e.g., $\kappa_i = \kappa_G$ for all $i \in G$; the latter is considered in Section 5. For reasonable levels of dependence (η_2, η_3) in the centered parameterizations (3)–(5), one would anticipate a large scale parameter κ_i to approximately match the unconditional expectation $E[Y(s_i)]$ or marginal probability of an edge [Caragea and Kaiser, 2009, Kaiser et al., 2012a]. The extent to which $E[Y(s_i)] \approx \kappa_i$ fails to hold, particularly across a common edge type $\kappa_i = \kappa_G$, may be examined in simulation as one possible check on degeneracy or other inadequacies in fitted models based on (3)–(5) (as illustrated in Section 5).

4 Graph Structure in LSGM

There are three remaining factors involved in the formulation of a LSGM that influence the overall random network allowed by the model in a problem. In this section, we describe those factors briefly in turn (Sections 4.1-4.3) as well as their specification for the two network examples (Section 4.4).

4.1 The Saturated Graph

Although not required, specification of a saturated graph, or a maximally allowable graph in terms of edges, can influence the overall potential density of a graph, reduce the computational burden, and keep the magnitude of estimated parameters in more easily interpretable ranges. In a graph with n nodes, a saturated graph arises from restricting the edges with positive probability to a meaningful subset of the $n(n-1)/2$ possible pairs of nodes. Such restrictions are intuitive for situations that involve physically impossible edges, but can also be used to focus an analysis on types of edges that are of primary interest. For example, in a graph for which nodes are morphological variants or subspecies of a given type of salamander and edges represent inter-breeding, we might allow potential edges only among those subspecies that overlap substantially in a geographical range. This does not mean that individuals of two subspecies with disjunct ranges may never inter-breed, just that such an occurrence would be a rarity, and inclusion of all possible edges of this type would not benefit the analysis.

4.2 Neighborhood Specification

While the use of a saturated graph that contains less than $n(n-1)/2$ potential edges influences global, or large-scale, aspects of graph topology, the specification of neighborhoods directly influences local graph structure. The use of neighborhoods to model the conditional probability that potential edges are realized is, in fact, the origin of *local structure* in a LSGM.

There are few, if any, general principles that guide the specification of neighborhood structures in a LSGM, and the process is highly specific to whatever substantive problem is under consideration. This is similar to the specification of neighborhoods in spatial applications of MRF models. There, default neighborhood specifications on regular lattices are often taken to be either four-nearest or eight-nearest structures. In a similar vein, if one is concerned with triangle realizations and transitivity exhibited in a graph, neighborhoods defined by incidence (cf. Section 3) are a reasonable default specification, and one we will use in the following analyses. Regarding neighborhood specification, other related notions of local dependence have traditionally existed in the ERGM literature [Frank and Strauss, 1986, Snijders et al., 2006, Morris et al., 2008, Wang et al., 2013] which can also be suggestive of neighborhood structures, as discussed further in Section 4.4.

4.3 Clique Sizes (Parameters) Included in Model

The final factor that exerts an influence over the manner in which a LSGM represents graph topologies is specification of the clique sizes included in modeling the natural parameter functions of conditional binary distributions for potential edges. In the models considered here, the choice is between (3) and (4). The former model addresses dependence only among pairs of neighboring potential edges, while the latter model includes dependence among cliques of three potential edges, that is, among triples of edges that are all neighbors of each other. Relative to the phenomenon of transitivity, cliques of size three contain edges that form potential triangles.

Note that inclusion of the last term in (4) differs from the common practice with MRF models of partitioning an overall neighborhood to deal with, for example, directional spatial dependencies. Here, the last right hand term in (4) is not mutually exclusive of the pairwise-dependence term in the first right hand sum. In fact, each clique of size three will result in two terms being included in the first right hand sum and one term in the second right hand sum of (4). Ultimately, there are four specific models that arise from (4). An independence model results from restricting $\eta_2 = \eta_3 = 0$; a model with pairwise-only dependence results from restricting $\eta_3 = 0$; a model with no pairwise dependence but dependence among triples results from $\eta_2 = 0$; and a model with both pairwise and three-way dependencies results from not restricting either η_2 or η_3. All of these models will, in fact, generate some triangles in a random network realization. The question becomes which model might be more in concert with a given network.

4.4 Specifications for Example Networks

Possible and realized topological features and dependence structures are next contrasted between the Faux Mesa High and Football networks, particularly in light of the specification of saturated graphs and neighborhood assignments discussed in previous sections.

For the Faux Mesa High friendship network, a saturated graph will be specified that allows edges to form only between two students (nodes) in the same grade. More than 80% of all realized edges in the network are captured by this saturated graph, but the number of potential edges to model is reduced from $20,910$ for an unrestricted situation to only $4,174$ under this definition of a saturated graph. The neighborhood definition for the Faux Mesa High network models involves both incidence and homophily, so that two potential edges are neighbors if they share a node and connect two nodes of the same sex and race. For example, the neighborhood of an edge which potentially connects student A to student B, where both students are female and Native American, consists of all edges which either connect student A to other Native American females or student B to other Native American females. See Figure 3 for an illustration of edge neighbors and non-neighbors. Node characteristics, including particularly homophily, have been shown to be important in predicting edge formation in social networks [Hunter et al., 2008a], and the general notion of limiting dependence based on attribute information is also common with ERGMs [Morris et al., 2008]. Note that edges may still form between nodes that do not exhibit the same sex and race, this neighborhood definition only affects the amount of dependence in the models. The resulting dependence structure is summarized in Table 1. Edges have between 0 and 38 neighbors, with an average of 18.95. The number of potential cliques of size three is $162,229$; on average an edge is a member of 116.60 cliques of size three. Regarding triangle-type subgraphs of dependence, the number of cliques of size three that are triangles is 3,093, so only 2% of cliques of size three resulting from this neighborhood definition are triangles. In a network realization from a model however, note that triangles may form between sets of nodes that do not constitute a clique of size three, even in models that incorporate dependence among such cliques explicitly.

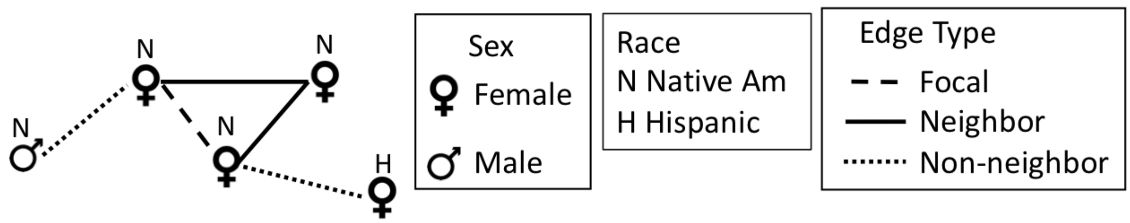


Figure 3: Some example edges for the Faux Mesa High School network to illustrate neighbor relations. Two edges are neighbors if the nodes which they connect are incident and share the same race and sex. Thus, for the focal edge connecting two Native American females, neighbors include the two potential edges connecting the nodes to another Native American female. The dotted lines are not neighbors because they do not share both the race and sex attributes of the focal edge.

In the Football network, no restrictions will be placed on the saturated graph so that, with 115 nodes (teams), there are $6,555$ potential edges. Neighborhoods in this example will be defined by any potential edges that are both incident and join two nodes within the same conference; this reflects a notion of local dependence within a block structure as in Guo et al. [2013], Wang et al. [2013], Schweinberger and Handcock [2012]. A result of this specification is that $6,066$ of the possible edges will have no neighbors. Table 1 displays resulting neighborhood statistics for those edges that have at least one neighbor (i.e., positive neighborhood sizes), where the average neighborhood size is then 17.12. For the Football network, the definition of neighborhoods leads to $12,507$ possible cliques of size three, $1,395$ of which are triangle-subgraphs but, again, this does not necessarily constrain the number of possibly observable triangles.

Notice the distinction between these two examples in their uses of saturated graphs and neighborhood specifications. In the Faux Mesa High network, the number of potential edges have been reduced through the specification of a saturated graph, and edge neighborhoods have been defined to include pairs of edges that may lead to triangle-cliques and also satisfy other similarity constraints. In models that incorporate these specifications, positive dependence promotes transitivity and homophily. A lack of dependence, on the other hand,

Table 1: Comparison of the modeled dependence of the Faux Mesa High and football networks

	Football	Faux Mesa High
# nodes	115	205
Result of Saturated Graph:		
# potential edges	6555	4174
Result of Neighborhood Definition:		
Number of neighbor-less edges	6066	21
Average neighborhood size	17.12*	18.95
Average number of triangles	8.56*	11.13*
Number of cliques of size 3	12,507	162,229
Number of triangles	1395	3093
Number of unique 2-stars	4185	39,546

*Average of only the positive values.

suggests these phenomena are not major factors in determining graph topology. In the Football network, the full set of potential edges are allowed. The neighborhood definition again involves edges that are incident, but implicitly captures only edges that could reasonably be expected to result in triangle-type cliques, given the uneven number of games each team plays against in-conference versus out-of-conference opponents.

5 Analysis

The remaining factor in a LSGM that determines the structure of networks under the model is parameter specification, i.e., the manner in which the κ_i are modeled in (4) and how dependence parameters (η_2, η_3 or both) are included. We fit three models to each of the Faux Mesa High and Football networks. Within each example, all models treated the large-scale parameters (the κ_i) in the same manner, where each $\kappa_i \in (0, 1)$ in the centered parameterizations of (3)-(4) intends to reflect the marginal/overall probability of an edge $Y(\mathbf{s}_i) = 1$, often varying by edge type. Consequently, for the Faux Mesa High network, we used

$$\kappa_i = \begin{cases} \kappa_{FF} & \text{for Female-Female edges} \\ \kappa_{MM} & \text{for Male-Male edges} \\ \kappa_{FM} & \text{for Female-Male or Male-Female edges.} \end{cases}$$

For the Football network, these large-scale parameters were specified as

$$\kappa_i = \begin{cases} \kappa_{In} & \text{for intra-conference games} \\ \kappa_{Out} & \text{for inter-conference games.} \end{cases}$$

The difference between the three models within each example was in how the small-scale, or dependence, structure was represented. Model 1 assumed pairwise-only dependence, with natural parameter function $A_i\{\mathbf{y}(N_i)\}$ from (3) or, equivalently, (4) with $\eta_3 = 0$. Dependence between both pairs and triples of dependent edges were included in Model 3 (i.e., η_2, η_3). For the Football network, the Model 3 natural parameter function $A_i\{\mathbf{y}(N_i)\}$ corresponded to (4); for the Faux Mesa High network, triangle-subgraphs were instead the only cliques of size three used so that the Model 3 natural parameter function was that of (5). Finally, what we will call Model 2 for each network was the same as Model 3, but with $\eta_2 = 0$ in each case.

5.1 Fitted Models and Large-Scale Parameters

5.1.1 Parameter Estimation

Parameter estimates were obtained by maximizing the log-pseudolikelihood, which for a LSGM is

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

where $p_i(N_i) \equiv P(Y(\mathbf{s}_i) = 1|\mathbf{y}(N_i)) = \exp(A_i\{\mathbf{y}(N_i)\})/[1 + \exp(A_i\{\mathbf{y}(N_i)\})]$ is the conditional probability of an edge at $\mathbf{s}_i$ and $A_i\{\mathbf{y}(N_i)\}$ is the corresponding natural parame-

Table 2: Parameter estimates and 90% percentile parametric bootstrap confidence intervals for three models fit to the Faux Mesa High network.

Parameter	Model 1	Model 2	Model 3
κ_{FF}	0.063 (0.049, 0.078)	0.063 (0.048, 0.079)	0.060 (0.043, 0.078)
κ_{FM}	0.025 (0.020, 0.031)	0.027 (0.021, 0.033)	0.026 (0.020, 0.032)
κ_{MM}	0.033 (0.024, 0.043)	0.033 (0.023, 0.044)	0.032 (0.020, 0.044)
η_2	8.40 (4.155, 10.732)	–	7.15 (0.093, 10.273)
η_3	–	36.51 (29.10, 51.16)	24.19 (18.076, 42.028)

ter function. This fitting approach is particularly numerically tractable given the centered conditional specifications and the locally adjusted dependence parameters in (3)–(5), and pseudo-likelihood estimation can be shown to be consistent here under mild assumptions, similarly to spatial data contexts (cf. Besag [1974, 1975], Guyon [1995])

Confidence intervals were obtained through a percentile parametric bootstrap procedure. For each model, 1000 simulations were obtained from a Gibbs Sampling algorithm using a burn-in and thinning of 10,000 networks. Point estimates and confidence intervals for the Faux Mesa High are displayed in Table 2. All 1,000 simulations are represented in the confidence intervals of Model 1; however, the estimation algorithm failed to produce estimates for 14 simulations from Model 2 and 33 from Model 3 so that the intervals of Table 2 for these models are based on 986 and 967 simulations, respectively. The estimates for the κ parameters are relatively close to the overall proportions in the data (by edge type), with confidence intervals that are fairly narrow. That is, the κ parameter estimates do not change dramatically over Models 1-3 incorporating different dependence types, which is intended by centered parameterizations (3)–(5). None of the confidence intervals for dependence parameters contain zero, indicating a substantial amount of dependence between both pairs of neighboring edges as well as triples of neighboring edges that include triangles.

Estimates of the parameters for the three models fit to the Football network are displayed in Table 3, where again (due to centering) κ parameter estimates are stable over the Models 1-3 though dependence parameter estimates vary greatly. Confidence intervals are again computed using percentile parametric bootstrap, using $1,000$ simulated data sets. For Model

Table 3: Parameter estimates and 90% percentile parametric bootstrap confidence intervals for three models fit to the football network.

Parameter	Model 1	Model 2	Model 3
κ_{out}	0.034 (0.031, 0.038)	0.034 (0.031, 0.038)	0.034 (0.030, 0.038)
κ_{in}	0.830 (0.796, 0.863)	0.831 (0.794, 0.873)	0.832 (0.801, 0.861)
η_2	3.54 (-3.299, 7.354)	–	-142.23 (-209.16, 2.23)
η_3	–	3.97 (-1.75, 7.79)	127.79 (-1.75, 186.45)

3, 161 of those data sets failed to produce estimation convergence and so the confidence intervals for Model 3 are based on 839 simulated networks. Again, the estimates for the large-scale parameters are in concert with the realized data proportions in the network, and confidence intervals appear to be symmetric and narrow around the estimates. In contrast with the Faux Mesa High example, confidence intervals for the dependence parameters do not indicate strongly significant dependence and are not as symmetric, particularly those for Model 3. In fact, it appears that this network is dominated by the proportion of within-conference games played by each team (related to κ_{in}) and this alone is largely responsible for the structure of the realized graph.

5.1.2 Checks of Degeneracy and Large-Scale Parameter Effects

An important consideration when fitting network models is the issue of model degeneracy. This phenomenon occurs when a model places all or most of its probability on only a few possible network realizations that often do not resemble the network of interest. This model failure has been widely studied in the network analysis literature for ERGMs [Handcock, 2003, Schweinberger, 2011], has been recognized in a more general class of models for interactive systems [Strauss, 1986], and is associated with long-range dependence in Ising models [Snijders, 2002]. To identify model degeneracy, Kaiser et al. [2012a] suggest simulating from the fitted model and verifying that proportions of realized edges, among different types of edges in a simulated network, are reasonably in concert with those from the original network data [1]. If not in concert, this may also suggest other model inadequacies regarding the fitted model being unable to account for large-scale (mean) effects in combination with the fitted dependence parameters [Caragea and Kaiser, 2009].

It should be noted that there are three related aspects to what is often called model degeneracy. First is what might be considered absolute degeneracy, in which a given model contains only a few points of non-negligible probability in its joint support. A second related concept of model stability [Schweinberger, 2011] concerns dramatic changes in probability among joint configurations differing in only one or a small number of data points. Finally, and largely unrecognized, is a connection with the lack of a unique decomposition between large-scale and small-scale structures in models. All models considered here can be shown to be stable (see supplementary materials) so that artifacts related to degeneracy are caused by excessive magnitude of dependence parameters [Kaiser et al., 2012a]. But poor representations of large-scale (mean) data structure can also result from the combined effects of large-scale (κ_i) and small-scale (η_1, η_2, η_3) model parameters and this is, again, exacerbated when an (estimated) dependence parameter becomes overly large. Unfortunately, what overly large means is not easily determined and depends on many factors such as the sizes of neighborhoods and values of other dependence parameters that may be present in a model, among others.

For both example networks, Figure 4 displays overall and category-wise proportions of edges found in each of 1,000 (bootstrap) network simulations from the three fitted models per network example; each set of 1,000 proportions from simulation appears in a Normal quantile-quantile plot. A dashed horizontal line represents an edge proportion as realized in the corresponding original network example.

[1] The binary conditionals with centered parameterizations in (3)–(5) induce a curved exponential joint distribution for the LSGMs here and simulations from models fit by pseudo-likelihood (or maximum likelihood) estimation may not produce proportions of realized edges that match, on average, those from the original data. Due to the curved exponential form, the natural parameter space of these LSGMs is also not of the same dimension as the true parameter space; for the Faux Mesa High network, Models 1,2,3 have 4,4,5 parameters, respectively, while the dimensions of a full rank minimally sufficient statistic are 7,10,14 in these models. See the supplementary materials for more details on these joint distributions and minimal sufficiency.

The first row of Figure 4 summarizes simulations of the Faux Mesa High network obtained from the three fitted models. Note that the saturated graph determines the total number of possible edges, so the denominator for each proportion. Simulations from Models 2 and 3 overestimate both the overall and same gender proportions of edges. This feature is not necessarily indicative of model degeneracy, as model degeneracy occurs when the simulations result in only a few possible networks. Rather, the overestimation indicates that these models are not adequately describing a feature of the network. That is, Models 2-3 fail to re-create these gender proportions observed in the data (relating to large-scale/mean behavior), even though estimates of large-scale parameters (κ's in Table 2) are close to these observed sample proportions, which suggests issues in how these models are attempting to incorporate dependence structure. However, all three models are able to recreate the distribution of Female–Male edge proportions with the value realized from the original network near the center of all distributions. Note, though, that the realized proportion of Female-Male edges in the data was quite small. The interesting implication of Figure 4 for the Faux Mesa High network is that the model with pair-wise only dependence (Model 1) was able to capture the overall proportion of each type of edge, while models that included a dependence term for cliques of size three, either in addition (Model 3) or alone (Model 2), vitiated this performance. This illustrates the aforementioned interaction between large-scale and small-scale model parameters in determining model behavior, as the large-scale parameters were nearly the same for all three models.

The second row of Figure 4 displays the simulated edge proportions from the three models fit to the Football network. Out-of-conference edges form independently under all three models, and thus the simulated proportions are nearly identical regardless of the model fit. Proportions are also similar between the three models for the in-conference edge proportions (and thus for the overall proportions as well), with the middle of the distributions aligning with the realized proportions. Thus, none of the models raise concerns of model degeneracy or inadequacy, even though the estimates from Model 3 do not seem intuitive.

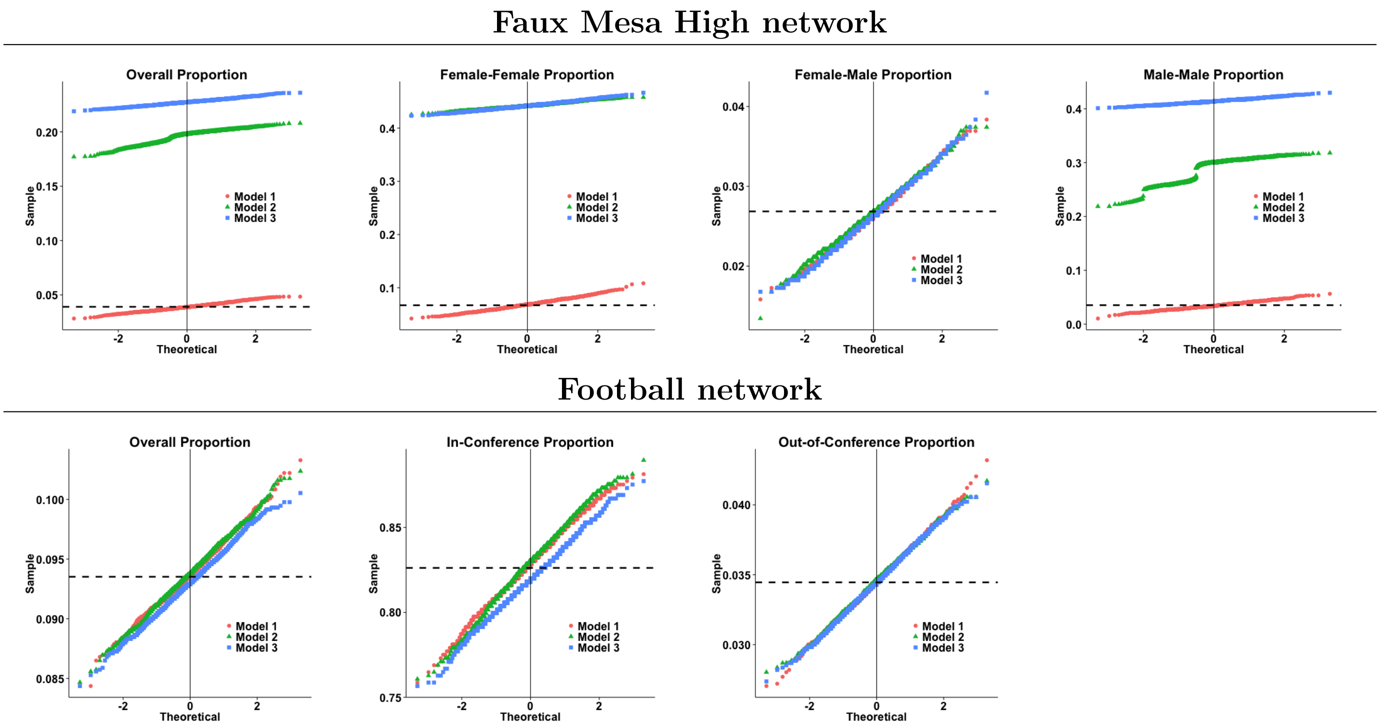


Figure 4: Normal quantile-quantile plots of the edge proportions (by different edge types) found in each of 1,000 simulated networks from three fitted models, per network example. The first row represents simulations from the models fit to the Faux Mesa High network and the second row to the Football network. Dashed horizontal lines represent edge proportions from the original network data. A vertical line at the theoretical quantile of zero has been drawn for reference.

5.2 Fitted Models and Dependence Effects

5.2.1 Examining Conditional Probability Structures

To further investigate the behavior of these models, estimated conditional probabilities $P(Y(\mathbf{s}_i) = 1|\mathbf{y}(N_i)) = E(Y(\mathbf{s}_i)|\mathbf{y}(N_i))$ of edges (or, equivalently, conditional expectations) were explored. Estimated conditional and marginal probabilities are displayed for Models 1 and 3 fit to the Faux Mesa High network in Table 4. Results of Models 2 and 3 are similar, thus values for Model 2 are not included. Because conditional probabilities depend on the values and number of neighboring edges, these probabilities will be computed for a focal edge, represented as a dashed line, for the neighborhood configurations displayed in Figure 5. Although these particular configurations are not necessarily common in the Faux Mesa High network, these are used for purposes of illustration due to their simplicity to visualize. Marginal probabilities are approximately the estimate of the corresponding κ (because of the centered parameterization) and only depend on the sex of the two nodes prescribing the focal edge in Figure 5. The neighbor configuration on the left of Figure 5 is used to compute the conditional probabilities of an edge connecting nodes of the same sex. Due to the incidence-homophily neighborhood structure (cf. Section 4.4), a potential Female–Male edge cannot similarly belong to a triangle-type clique so that a comparable neighborhood configuration is displayed on the right of Figure 5, where the focal edge has four neighbors.

For both Models 1 and 3 and for each type of edge, as the number of neighbors realized increases so does the conditional probability of edge occurrence. When zero, one or two neighbors that do not form a triangle-clique (in Model 3) are realized, the conditional probabilities for same-sex edges are similar between the two models. However, for same-sex edges, if there are two realized neighboring edges where the subsequent occurrence of the focal edge would realize a triangle in the network, then the probability of a same-sex focal edge forming is nearly unity under Model 3. The modeling implication for same-sex edges is that triangles in the network appear under Model 1 simply through the increase in clusters caused by dependence among pairs of potential edges (i.e., three and four positive neighbors), while

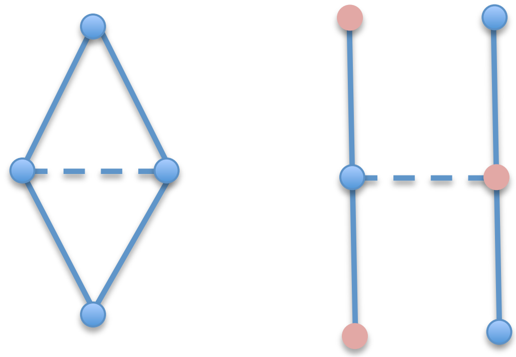


Figure 5: Neighborhood configurations used to compute the conditional probabilities for Female–Female and Male–Male (left) and Female–Male (right) in Table 4. The focal edge, for which the conditional probability is computed, appears as the dashed line in both.

under Model 3 triangles are realized almost whenever they are possible, with two, three or four positive neighbors. For Female–Male edges, the two models result in similar conditional probabilities among all possible outcomes of neighboring values.

To compare the fit of the three models to the Football network, the conditional probabilities of a typical in-conference edge are computed. These conditional probabilities depend on the number of neighbors and also, for Models 2 and 3, on the resulting number of cliques of size three. Twenty is the most common number of neighbors in these network models, so that the denominator of the pairwise dependence η_2 term will be 2×20 in (3) or (4) for purposes of computing conditional probabilities. Similarly, edges commonly belong to about 100 cliques of size three in Models 2-3, so that the denominator of the three-way dependence η_3 term will be treated as 3×100 in (4) for determining conditional probability. The resulting conditional probabilities for all three models are plotted in Figure 6 against the possible number of positive/realized neighboring edges, $\{0, 1, \ldots, 20\}$ here, along with the approximate marginal probability from each model as a dashed, horizontal line.

The conditional probabilities for Model 1 increase monotonically with the number of positive neighboring edges. Because Models 2 and 3 also include the dependence from triples of edges, there are multiple conditional probabilities possible for edge occurrence based on a given number of positive neighbors, as depicted in Figure 6 (e.g., Model 1 shows one proba-

	Female–Female	Male–Male		Female–Male
Model 1				
Marginal Prob, $P(Y(\mathbf{s}_i) = 1) \approx \kappa_i$	0.063	0.033		0.025
Conditional Prob, $P(Y(\mathbf{s}_i) = 1 \vert \mathbf{x}(\mathbf{s}_i), \mathbf{y}(N_i))$				
0 neighbors realized	0.047	0.029		0.023
1 neighbor realized	0.142	0.089		0.071
2 neighbors realized	0.354	0.245		0.203
3 neighbors realized	0.645	0.519		0.458
4 neighbors realized	0.858	0.781		0.737
Model 3				
Marginal Prob, $P(Y(\mathbf{s}_i) = 1) \approx \kappa_i$	0.060	0.032		0.026
Conditional Prob, $P(Y(\mathbf{s}_i) = 1 \vert \mathbf{x}(\mathbf{s}_i), \mathbf{y}(N_i))$				
0 neighbors realized	0.046	0.028		0.023
1 neighbor realized	0.117	0.074		0.072
2 neighbors realized				0.203
From same potential triangle	0.994	0.989		
From different potential triangle	0.269	0.180		
3 neighbors realized	0.998	0.996		0.456
4 neighbors realized	1	1		0.734

Table 4: Marginal and conditional probabilities for edges with the neighborhood configuration shown in Figure 5.

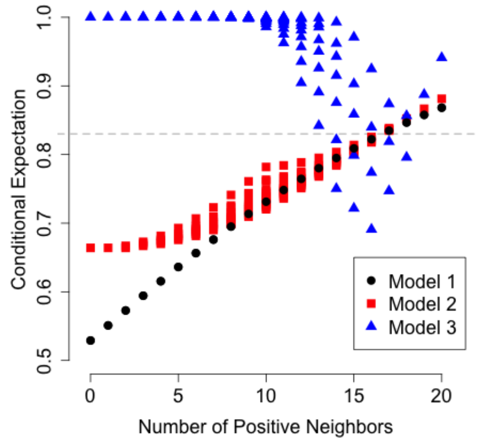


Figure 6: Conditional expectations/probabilities under the three fitted models as a function of the number of positive neighbors $(1, \ldots, 20)$. The approximate marginal expectation/probability for each model is plotted as a gray, dashed horizontal line.

bility for a given number of positive neighbors, while Models 2-3 show several probabilities). For example, if three neighbors are positive, then the occurrence of a focal edge may result in all positive edges among either 1, 2 or 3 cliques of size three. Conditional probabilities for Models 1 and 2 are similar in Figure 6, particularly once the number of positive neighbors is at least 7, possibly due to the similarity of the estimates of η_2 and η_3 in each model.

The pattern of conditional probabilities from Model 3 in Figure 6 is not intuitive, as this value is almost 1 even when few neighbors are realized and then this probability decreases once half of the neighbors are realized. This aspect is due to the fact that the estimate of η_2 is negative for this model and, as the number of positive neighbors increases, the number of two-stars increases more rapidly than does the number of cliques of size three. Thus, as the number of positive neighbors grows, the influence of η_2 on conditional probabilities becomes greater than the influence of η_3.

5.2.2 Further Examinations of Estimated Dependence

Lastly, the three fitted models were examined as to how well they are able to recreate two-stars and triangles realized in the actual networks. We consider these features only with regard to those model edges included in at least one neighborhood, as these are the edges for which the models explicitly prescribe dependence. Based on repeated network simulations from the three fitted models, Figure 7 demonstrates the proportions of realized two-stars and dependent triples of edges (found in each simulated network) among those possible under the models. The top two plots show the results from the models fit to the Faux Mesa High network. Simulations from the fitted Models 2 and 3 result in too many two-stars and triangles, which is intuitive as edges from these models were over simulated in general (see Figure 4). Model 1 does not include a term that explicitly describes transitivity, but was most able to recreate the number of realized triangles among edges included in the dependence terms of Models 2-3. The use of a saturated graph and the particular neighborhood definition in the Faux Mesa High network results in 3,093 possible triangles, where 6 (or 0.2%) are actually realized. For comparison, an unrestricted saturated graph for the Faux Mesa High network would entail $\binom{205}{2} = 20,910$ potential edges with $4,244,730$ two-star neighbors and $1,414,910$ triangle-type neighbors. Among the latter, there are 62 realized triangles in the actual data, or 0.004% of such triangles. It is not the use of a restricted saturated model that drives the results of the Faux Mesa High network but, rather, the low level of transitivity exhibited by the data.

In the Football network, the capacity of the three models to reproduce realized edges among subgroups of two-stars and triples of dependent edges is indicated in the second row of Figure 7. Again, all three models behave similarly. Simulated networks from the fitted models tend to over recreate edges among two-stars and cliques of size three, on average, compared to those found in the actual network. Thus, it may be that the strength of the dependence between pairs and triples of dependent edges is slightly overestimated in all three models.

These results for the Football network suggest that inclusion of both η_2 and η_3 in Model 3 may be redundant. This is further supported by the plot of estimated values of these parameters in Figure 8 which have a correlation of -0.997. Estimated values in this plot are found

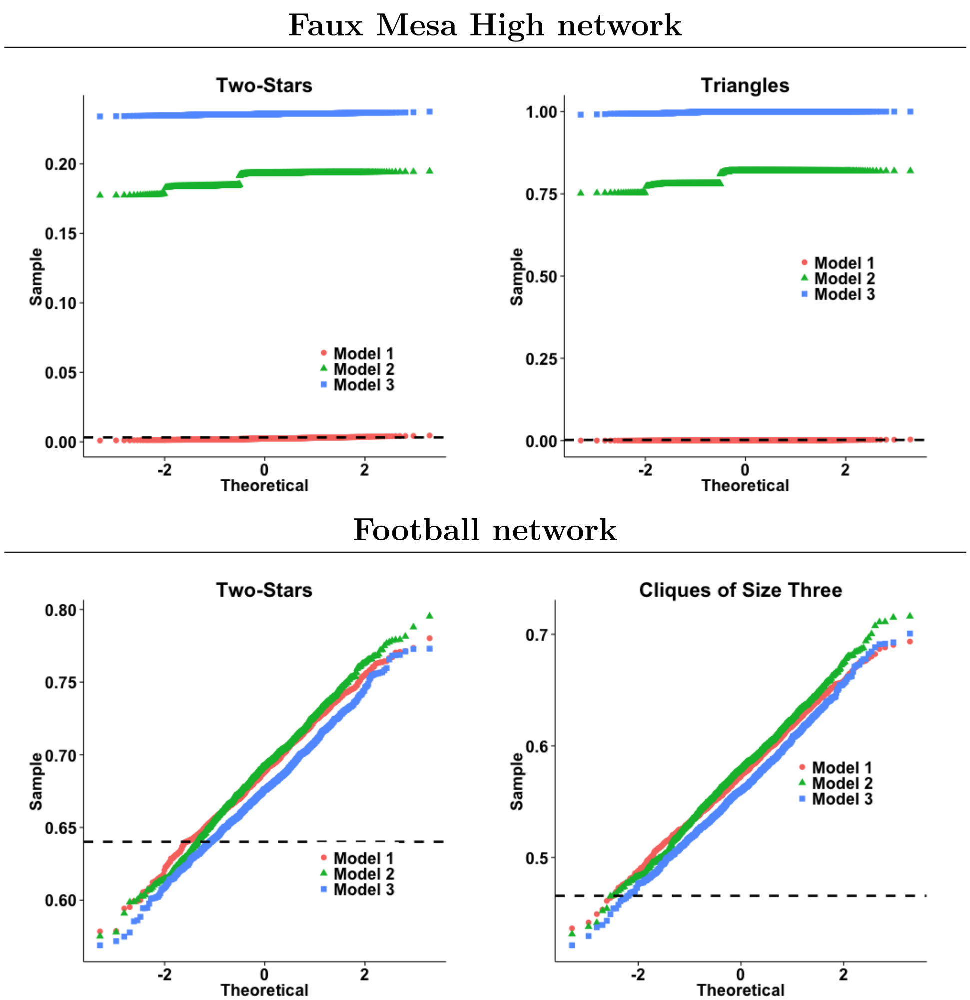


Figure 7: Normal quantile-quantile plots which demonstrate the ability of the three fitted models to recreate realized proportions of two-stars and triples of dependent edges among those modeled in each of the Faux Mesa High and Football networks; one network simulation from a fitted model results in a proportion indicated above. Horizontal, dashed lines correspond to the actual proportions in these networks.

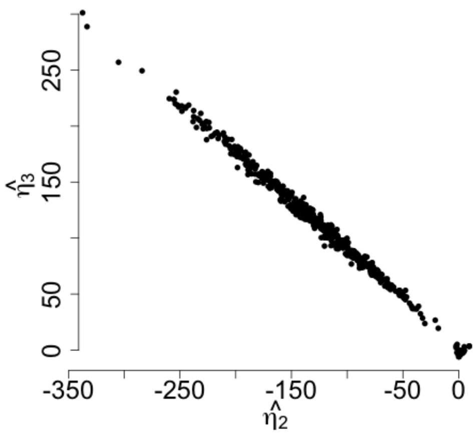


Figure 8: Scatterplot of the estimates of η_2 against η_3 from 839 (bootstrap) network simulations from the fitted Model 3 for the Football network.

from the bootstrap samples used to compute interval estimates in Table 3. A choice between Model 1 and Model 2 to describe this example (both as reductions of Model 3) may depend on factors not considered in this article.

6 Discussion

We described features associated with the use of a LSGM to account for structural components of realized networks. The use of *specified* neighborhoods for potential edges is the distinguishing characteristic of the LSGM approach. While ERGM specifications *imply* certain conditioning sets of edges in the full conditional distributions, a joint or global model specification may offer little guidance about how to control the structure of those neighborhoods or how to tailor neighborhood structures to particular problems. Given that full conditional distributions are explicitly formulated in a LSGM, an investigator has the additional flexibility of including pairwise-only or higher order interactions among edges in the model.

All of the controlling factors in a LSGM can impact the manner in which a model represents the phenomenon of transitivity in networks, which has been the focus of this article. We have relied on two well-known example networks which exhibit marked differences in overall topology. The Faux Mesa High friendship network is sparse, with only minimal topological features of interest. The representation of this network using pairwise-only dependence on a restricted saturated graph, and with neighborhoods chosen to exploit homophily, is adequate for describing this network. Model terms explicitly targeted at increasing the level of triangles (e.g., relative to open two-stars) not only fail to offer improvement in this example, but actually prove detrimental to adequate representation of the observed data. Our conclusions for this example differ from those of Hunter et al. [2008a], who used a traditional ERGM structure and found that there was a need to include a host of parameters for effects of grade, sex, and race, as well as a highly contrived term constructed to reflect transitivity (geometrically weighted edge-wise shared partner) in order to adequately represent this same measure in the actual data.

In contrast, the Football network is characterized by a large overall degree among the predominant edge category (in-conference contests), leading to many two-stars and triangles. But, contrary to what one might anticipate, inclusion of modeling terms to explicitly account for these features, namely the inclusion of cliques of size three in natural parameter functions, offers no improvement in description of the network over a pairwise-only dependence model. Here, however, the cause is quite different than in the Faux Mesa High example. In the Football network, models with parameters for both pairwise and group interactions are essentially modeling the same structure twice. One could use a model with only cliques of size three or one could use a model with only cliques of size two (pairwise-only dependence), but using both leads to uninterpretable parameter estimates. In fact, if one is guided solely by non-simultaneous interval estimates, the possibility that an independence model might be reasonable suggests itself. In their analysis of the same network, Guo et al. [2013] found the that a hierarchical block model with a parameter for in-conference contests for larger (more than 9 team) conferences and a separate parameter for smaller conferences was the most appropriate fit to the data. The dependence structure of the LSGM is able to capture this feature of the Football network more intuitively, without delineating in-conference games based on conference size.

Consideration of the two example networks used in this article also provides a cautionary tale to fitting network models without adequate examination. After fitting the model with the most dependence terms (Model 3) to the Faux Mesa High network, if only the parameter estimates and resulting bootstrap confidence intervals had been examined, then the adequacy of the model would most likely not have been questioned. It is when simulations from this model are examined and compared to the actual network that it becomes clear that the model is not adequately describing features of interest (e.g., occurrence proportions among types of edges). Similarly, an adequate model fit might have been declared for the Football network if Model 3 alone had been fit and if only simulation of network features had been examined, without regard to the behavior of parameter estimates and conditional probabilities. Interval estimates of parameters and estimated conditional expectations indicate a non-intuitive model fit for this network example, further verified by the strong linear relationship in estimated dependence parameters η_2 and η_3 among bootstrap samples from the fitted model.

The use of a LSGM to examine the structure of networks involves choices in model formulation that will impact conclusions made on the basis of statistical analysis. These choices are the most defensible when made on the basis of scientific understanding of the network under study. Absent that, however, there are procedures that the investigator can use to determine how parts of the overall model structure interact with parts of the overall data structure. We suggest that such procedures include assessment of both marginal graph features (number of topological features generated by a fitted model) and conditional features (conditional probabilities of edge realization under specific neighborhood configurations).

The conditional formulation of LSGMs may also be useful for potentially new types of goodness-of-fit assessments for network models. For example, Kaiser et al. [2012b] proposed a general testing procedure for spatial Markov random field models based on the notion of concliques (groups of non-neighboring observations) and certain generalized residuals defined on such concliques. Extension of this procedure to binary models, and LSGMs in particular, has the potential to produce new perspectives to formal model assessments for network data. Additionally, formal assessments of neighborhood structure in networks may also be possible based on the development of conditional moment tests, such as those in Kaiser and Nordman [2012] for spatial data.

7 Acknowledgements

The work was supported in part by the Sandia National Laboratories Laboratory-Directed Research and Development Program. Research was partially supported by NSF DMS-2015390. This work has been approved for public release, LA-UR-20-23875.

References

Skye Bender-deMoll, Martina Morris, and James Moody. Prototype packages for managing and animating longitudinal network data: `dynamicnetwork` and `rSoNIA`. *Journal of Statistical Software*, 24(7), 2008.

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

Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, 24:179–195, 1975.

Petruţa C. Caragea and Mark S. Kaiser. Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(3):281–300, September 2009. ISSN 1085-7117. URL http://www.springerlink.com/index/10.1198/jabes.2009.07032.

Emily Casleton, Daniel Nordman, and Mark Kaiser. A local structure model for network analysis. *Statistics and Its Interface*, 10(2):355–367, 2017.

Emily Casleton, Daniel Nordman, and Mark Kaiser. Local structure graph models with higher-order dependence. *The Canadian Journal of Statistics*, to appear, 2020.

N Cressie. *Statistics for Spatial Data*. Wiley-Interscience, New York, 1993.

Ove Frank and David Strauss. Markov Graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

Steven M. Goodreau, Mark S. Handcock, David R. Hunter, Carter T. Butts, and Martina Morris. A statnet tutorial. *Journal of statistical software*, 24(9):1, 2008.

Jiqiang Guo, Alyson G Wilson, and Daniel J Nordman. Bayesian nonparametric models for community detection. *Technometrics*, 55(4):390–402, 2013.

Xavier Guyon. *Random fields on a network: modeling, statistics, and applications*. Springer, 1995.

Mark S Handcock. Assessing degeneracy in statistical models of social networks. Working Paper 39, Center for Statistics and the Social Sciences, University of Washington, Seattle, 2003.

Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, Skye Bender-deMoll, and Martina Morris. *statnet: Software tools for the Statistical Analysis of Network Data*. The Statnet Project (`http://www.statnet.org`), 2014. URL `CRAN.R-project.org/package=statnet`. R package version 2014.2.0.

David R Hunter. Curved exponential family models for social networks. *Social networks*, 29 (2):216–230, 2007.

David R. Hunter and Mark S. Handcock. Inference in Curved Exponential Family Models for Networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, September

2006. ISSN 1061-8600. doi: 10.1198/106186006X133069. URL http://pubs.amstat.org/doi/abs/10.1198/106186006X133069.

David R. Hunter, Steven M. Goodreau, and Mark S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008a.

David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29, 2008b.

Mark S. Kaiser and Noel Cressie. The construction of multivariate distributions from Markov random fields. *Journal of Multivariate Analysis*, 73(2):199–220, 2000.

Mark S Kaiser and Daniel J Nordman. Blockwise empirical likelihood for spatial markov model assessment. *Statistics and Its Interface*, 5(3):303–318, 2012.

Mark S Kaiser, Petruţa C Caragea, and Kyoji Furukawa. Centered parameterizations and dependence limitations in Markov random field models. *Journal of Statistical Planning and Inference*, 142(7):1855–1863, 2012a.

Mark S Kaiser, Soumendra N Lahiri, Daniel J Nordman, et al. Goodness of fit tests for a class of markov random field models. *The Annals of Statistics*, 40(1):104–130, 2012b.

Andee Kaplan, Daniel J Nordman, and Stephen B Vardeman. On the s-instability and degeneracy of discrete deep learning models. *Information and Inference: A Journal of the IMA*, 9(3):627–655, 2020.

VV Kashirin and LJ Dijkstra. A heuristic optimization method for mitigating the impact of a virus attack. *Procedia Computer Science*, 18:2619–2628, 2013.

E.D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):1–5, October 2008. ISSN 1539-3755. doi: 10.1103/PhysRevE.78.046110. URL http://link.aps.org/doi/10.1103/PhysRevE.78.046110.

Jaehyung Lee, Mark S. Kaiser, and Noel Cressie. Multiway dependence in exponential family conditional distributions. *Journal of Multivariate Analysis*, 79(2):171–190, 2001.

Eric Lofgren. Visualizing results from infection transmission models. *Epidemiology*, 23(5): 738–741, 2012.

Dalton Lunga and Sergey Kirshner. Generating similar graphs from spherical features. In *Ninth Workshop on Mining and Learning with Graphs (MLG '11)*, San Diego, CA, Aug 2011.

Martina Morris, Mark S Handcock, and David R Hunter. Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. *Journal of Statistical Software*, 24(4):1548–7660, January 2008. ISSN 1548-7660. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2481518&tool=pmcentrez&rendertype=abstract.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

Michael D Resnick, Peter S Bearman, Robert Wm Blum, Karl E Bauman, Kathleen M Harris, Jo Jones, Joyce Tabor, Trish Beuhring, Renee E Sieving, Marcia Shew, et al. Protecting adolescents from harm: findings from the national longitudinal study on adolescent health. *Jama*, 278(10):823–832, 1997.

G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29(2):192–215, 2007.

Michael Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370, 2011.

Michael Schweinberger and Mark S Handcock. Hierarchical exponential-family random graph models with local dependence. 2012.

Tom A. B. Snijders. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.

Tom A. B. Snijders, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36:99–153, 2006.

David Strauss. On a General Class of Models for Interaction. *SIAM Review*, 28(4):513–527, 1986.

Demival Vasques Filho and Dion RJ O'Neale. Transitivity and degree assortativity explained: The bipartite structure of social networks. *Physical Review E*, 101(5):052305, 2020.

Peng Wang, Garry Robins, Philippa Pattison, and Emmanuel Lazega. Exponential random graph models for multilevel networks. *Social Networks*, 35(1):96–115, 2013.