

---

# Hierarchical Models for Independence Structures of Networks

Kayvan Sadeghi<sup>1</sup> | Alessandro Rinaldo<sup>2</sup>

<sup>1</sup>University College London

<sup>2</sup>Carnegie Mellon University

## Correspondence

Department of Statistical Science,  
University College, Gower Street London,  
WC1E 6BT, United kingdom  
Email: k.sadeghi@ucl.ac.uk

## Funding information

AFOSR, Grant Number: FA9550-14-1-014

We introduce a new family of network models, called hierarchical network models, that allow us to represent in an explicit manner the stochastic dependence among the dyads (random ties) of the network. In particular, each member of this family can be associated with a graphical model defining conditional independence clauses among the dyads of the network, called the dependency graph. Every network model with dyadic independence assumption can be generalized to construct members of this new family. Using this new framework, we generalize the Erdős-Rényi and  $\beta$ -models to create hierarchical Erdős-Rényi and  $\beta$ -models. We describe various methods for parameter estimation as well as simulation studies for models with sparse dependency graphs.

**Keywords** — dependency graph, exponential random graph models, graphical models, log-linear models, social network analysis

## 1 | INTRODUCTION

The statistical analysis of network data is concerned with modeling relational data taking the form of random graphs, or networks, where the nodes represent units in a population of interest and the random ties (i.e. dyads) encode the complex of interactions among them.

Many well-known network models and, in fact, most models that can be fit to large networks either rely on the often unrealistic but theoretically convenient assumption of *dyadic independence*, that is probabilistic independence among the dyads, or induce complete dependence among the dyads. Developing network models that explicitly accounts for more complex forms of marginal or conditional independencies among the dyads, also called the *independence structure of the network*, has proved to be quite difficult, both for computational and theoretical reasons. As a result, most of the

network models proposed in the literature and used in practice allow only for a minimal degree of control over the type and strength of stochastic dependence among dyads.

On the other hand, in the field of graphical models, there is a vast body of work on the independence structure of sets of random variables. In graphical models, graphs, generally called *Markov dependency graphs* or simply *dependency graphs* consist of vertices that correspond to random variables and edges that correspond to some types of conditional dependencies between their endpoints; see for example Lauritzen (1996). Throughout, in order to avoid confusion, we use the terms "node" and "tie" for networks, and "vertex" and "edge" for dependency graphs, whereas a potential random tie is called a "dyad". In order to distinguish between different possible conditional independencies among random variables, different typologies of dependency graphs have been proposed in the graphical model literature, each of which corresponding to a particular class of models. One of the most popular class of graphical models is that of *undirected graph models* (Darroch et al., 1980), which assume conditional independence between random variables corresponding to two non-adjacent vertices in the dependency graph given all other variables (vertices of the graph).

There is a natural duality between networks and dependency graphs: dyads of the network are (binary) random variables, and hence can be considered to be vertices of a dependency graph. Edges of the dependency graph would then determine the conditional independence among these variables, i.e. the independence structure among the dyads of the network. This duality was first noted and used for modeling purposes in the seminal work of Frank and Strauss (1986), where network dyads are modeled using a very specific type of dependency graph that assumes dyads to be conditionally independent if they do not share a node; see also Chapter 7 of Lusher et al. (2012). Recently, this duality was used for sampling and model fitting in certain network models in Thiemichen and Kauermann (2017), or for modeling *exchangeable* random networks in Lauritzen et al. (2018, 2019). Also, there have been other approaches to deal with certain "local" types of dependency in networks by using *nodal attributes*, i.e. extra information on the individuals presented by nodes of the network (Schweinberger and Handcock, 2012) and Fellows and Handcock (2012).

The main goal of this paper is to leverage the duality between networks and dependency graphs in order to propose new, tractable and interpretable network models that allow for specific independence structures. We accomplish this by generalizing existing and well-known network models that rely on the dyadic independence assumption to hierarchical log-linear models that additionally conform to a given set of conditional independencies among the dyads. Towards that end, we work with undirected graphical models, which imply that our proposed network models are of linear exponential family form and, therefore, are instances of the class of *exponential random graph models* (ERGMs); see Holland and Leinhardt (1981) (but see also its discussion, Fienberg and Wasserman, 1981) and Frank and Strauss (1986), and, for recent developments, Hunter and Handcock (2006); Robins et al. (2007). As baseline models, we consider both the Erdős-Rényi models, defined by Erdős and Rényi (1959) and studied vastly in the literature of networks and random graph theory and the  $\beta$ -models, defined by Blitzstein and Diaconis (2010); Chatterjee et al. (2011) and studied by Rinaldo et al. (2013). We call the resulting models the hierarchical Erdős-Rényi and hierarchical  $\beta$  model respectively. Our approach could also accommodate, in the same manner, for the class of  $p_1$  models, introduced by Holland and Leinhardt (1981) for directed networks.

We also provide a method based on the gradient descent algorithm to estimate the maximum of the likelihood function for hierarchical Erdős-Rényi models. In principle, this method can be generalized to other families of hierarchical models, but the computational difficulties should be examined in more detail. We also provide simulation studies to apply the proposed method for the maximum likelihood estimation, and to compare hierarchical Erdős-Rényi with Erdős-Rényi.

In the next section, we formally define networks and dependency graphs, introduce the independence structures for undirected dependency graphs, and illustrate the duality between networks and dependency graphs.

In Section 3.1, we introduce hierarchical log-linear models for undirected graphical models, and demonstrate how to

use them to define hierarchical network models. In Section 3.2, we apply the Erdős-Rényi parametrization to hierarchical network models, defined in Section 3.1, write the models in exponential family form, and provide the corresponding normalizing constant in closed form in the sense that they do not depend on summation over all networks. In Section 3.3, we conduct a similar study as in section 3.2 for the  $\beta$ -model parametrization instead of Erdős-Rényi's. In Section 4, we study the maximum likelihood estimation and provide algorithms for this purpose for (sparse) dependency graphs for hierarchical Erdős-Rényi. We also provide relevant simulation studies, which show that the models significantly take the simulated dependencies in the networks into account.

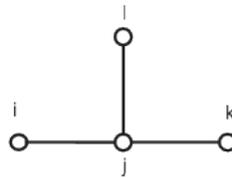
In Section 5, we present several problems related to the proposed models for further work. In particular we briefly discuss model selection for the dependency graph, and the existence of the maximum likelihood estimator for these models, as well as the selection of dependency graphs. We finally discuss analogous models based on marginal independence.

## 2 | NETWORKS AND DEPENDENCY GRAPHS

Graphical models (see, e.g. Lauritzen, 1996) are statistical models expressing conditional independence statements among a collection of random variables  $X = (X_1, \dots, X_M)$  indexed by a finite set  $N$ . A graphical model is determined by a *dependency graph*  $D = (N, F)$  over the indexing set  $N$ , whose edge set  $F$  (which may include edges of undirected, directed or bidirected type) encodes conditional independence relations among the variables, or *Markov properties*. For a non-empty set  $A \subset N$ , let  $X_A = (X_i, i \in A)$  be the corresponding sub-vector of  $X$ . Given disjoint subsets  $A, B$  and  $C$  of  $N$ , with  $A$  and  $B$  non-empty, we express the clause that  $X_A$  is conditional independent of  $X_B$  given  $X_C$  using the notation  $A \perp\!\!\!\perp B \mid C$ . In particular, if  $C$  is the empty set, this will reduce to simply marginal independence of  $X_A$  and  $X_B$ , written as  $A \perp\!\!\!\perp B$ .

For an *undirected graph*  $D$ , where all edges are depicted as full lines, if, for any two non-adjacent vertices  $i$  and  $j$ , it holds that  $i \perp\!\!\!\perp j \mid V \setminus \{i, j\}$ , i.e.  $i$  and  $j$  are conditionally independent given the rest of the vertices then we say that the *pairwise Markov property* is satisfied. If, for any three disjoint subsets  $A, B$  and  $C$  of the vertex set,  $A \perp\!\!\!\perp B \mid C$  when every path between  $A$  and  $B$  has a vertex in  $C$  then we say that the *global Markov property* is satisfied. It is known that these two conditions are equivalent for positive densities; see Pearl (1988); Lauritzen (1996).

For example, in the undirected graph of Fig. 1(a), the pairwise Markov property implies that  $i \perp\!\!\!\perp k \mid \{j, l\}$  and the global Markov property implies that  $\{i, l\} \perp\!\!\!\perp k \mid j$ .



**FIGURE 1** An undirected dependency graph.

We define a *random network* to be the random graph  $G = (V, E)$ , where the node set  $V$  consists of labeled individuals and the random tie set  $E$ , also called the set of *dyads*, consists of binary random variables taking values in  $\{0, 1\}$ . In a realization of a random network, nodes  $i$  and  $j$  are connected if the random variable corresponding to the tie  $ij$  takes the value 1, and they are disconnected otherwise.

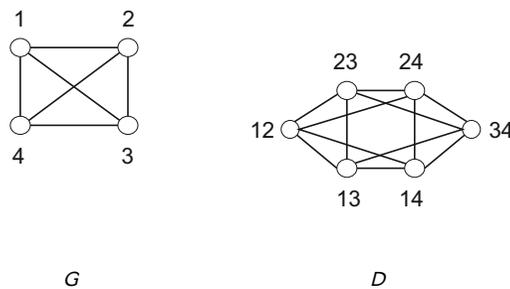
Now suppose that we are interested in networks with  $n = |V|$  nodes. Then, for a random network  $G$  (which is

a *complete* random network in the sense that all dyads are existent) on  $n$  nodes, it holds that its number of dyads is  $|E| = \binom{n}{2} := m$ . Conditional independencies among the dyads of  $G$  can then be expressed through a dependency graph  $D$  on  $m$  vertices, each vertex corresponding to a dyad in  $G$ . We will only be concerned with certain types of conditional independencies, which we introduce next. Let  $i, j, k$ , and  $l$  be labels for the nodes of the random network  $G$ . We say that a dependency graph satisfies the *Markov dependence property* if when ties  $ij$  and  $kl$  do not share a common node in  $G$  then  $ij \sim kl$  in  $D$ , i.e. vertices  $ij$  and  $kl$  are not adjacent. Notice also that the definition requires only non-neighboring ties of  $G$  to be non-adjacent vertices in  $D$  and not vice versa; therefore, the dependency graphs we propose could be any subgraph of the *line graph* of  $G$ : the line graph  $L(G)$  of a graph  $G$  is the intersection graph of the tie set  $E$ , i.e. its vertex set is  $E$  and  $e_1 \sim e_2$  if and only if  $e_1$  and  $e_2$  have a common endpoint (West, 2001, p. 168). For example, for networks with 4 nodes, all dependency graphs that satisfy the Markov dependence property are the subgraphs of the dependency graph  $D$  depicted in Fig. 2. In addition to begin amenable to theoretical analysis, there are practical justifications for adopting such restrictions; see the discussion in Section 5.

The type of restrictions on the dependency graphs described above is directly inspired by the Markov properties for networks put forward by Frank and Strauss (1986) in their seminal paper. Our modeling choice is, however, different in the two following ways: 1) In Frank and Strauss (1986), a unique dependency graph (namely *the line graph* of the complete graph) is used to model networks. Here, on the other hand, we model any possible subgraph of the graph used in Frank and Strauss. Therefore, we deal with different possible independence structures that might occur for networks. 2) In Frank and Strauss (1986), they assume exchangeability among the dyads of the network in order to reduce the number of parameters, whereas here we combine the graphical model with the known network models in the literature to obtain fewer parameters. This also ensures that our models inherit the desired properties of the baseline network model.

As we shall see in the next section, we are particularly interested in the *cliques*, i.e. complete subgraphs, of the dependency graph. A *triangle* in network  $G$  is a subgraph consisting of nodes  $i, j, k$  and ties  $ij, jk, ik$ . A  *$r$ -star* in  $G$  is a subgraph consisting of nodes  $i, i_1, \dots, i_r$  and ties  $ii_1, \dots, ii_r$ ; we call the node  $i$  the *hub* of the star  $C$ , and write  $i = \text{hub}(C)$ .

The following observation plays an important role in the paper: under Markov dependence property, cliques in  $D$  correspond to stars and triangles in the random network  $G$ : A clique of size 1, i.e. a vertex in  $D$ , is of form  $ij$ , and hence corresponds to the tie  $ij$  in  $G$ , i.e. a 1-star; a clique of size 2, i.e. an edge in  $D$ , is of form  $ij, ik$ , and hence, because of the Markov dependence property, corresponds to the 2-star with hub  $i$  and other nodes  $j, k$  in  $G$ ; a clique of size 3 that is of form  $ij, ik, il$  in  $D$  corresponds to a 3-star in  $G$ ; and a clique of size  $r, r \geq 4$ , corresponds to an  $r$ -star in  $G$ . A clique of size 3 that is of form  $ij, ik, jk$  in  $D$  corresponds to a triangle in  $G$ ; we call such cliques of  $D$  *hubless*.



**FIGURE 2** A random network  $G$  with 4 nodes and its corresponding line graph  $D$ .

Henceforth in this paper, we assume that a dependency graph is given, and the independence structure for the network is determined by this corresponding dependency graph.

### 3 | NETWORK MODELS BASED ON UNDIRECTED HIERARCHICAL MODELS

#### 3.1 | Undirected graphical models for networks

Henceforth, let  $\mathcal{G}_n$  be the set of all possible realizations of a network on  $n$  nodes. We do not distinguish between the observed network  $x \in \mathcal{G}_n$  itself and the binary vector of its dyads  $x = (x_{12}, \dots, x_{1n}, x_{23}, \dots, x_{2n}, \dots, x_{n-1,n}) \in \{0, 1\}^m$ , where every dyad is between two labeled nodes with labels  $i$  and  $j$  and  $m = \binom{n}{2}$ .

Our goal is to model  $P(x)$ , the probability of observing  $x$ , given an undirected dependency graph  $D$  on the set of  $m$  dyads conforming to the Markov dependence property defined in the previous section. For this purpose, we will use *hierarchical log-linear models*, which have been comprehensively studied for modeling undirected graphs in the graphical models literature; see Lauritzen (1996); Bishop et al. (2007). In detail, let  $C$  be the set of all cliques in  $D$ , and let  $C_0 = C \cup \{\emptyset\}$ . A hierarchical log-linear model corresponding to  $C_0$  can be written as

$$\log(P(x)) = \sum_{C \in C_0} u_C(x), \quad x \in \{0, 1\}^m, \quad (1)$$

where each  $u_C(x) = u_C(x_C)$  is an appropriate function of  $x \in \{0, 1\}^m$  that depends only on the coordinates in  $C$ . In particular,  $u_\emptyset$  is a constant function ensuring that  $\sum_{x \in \{0, 1\}^m} P(x) = 1$ . The hierarchical assumption requires that, if a term  $u_A(x)$  is set to zero for all  $x$ , so are all the terms  $u_B(x)$  such that  $A \subseteq B$ . Hence, the maximal cliques correspond to the maximal interaction terms not set to zero. These are also called the *generators* of the model.

We should first warn the reader not to confuse these models with hierarchical exponential random graph models, proposed by Schweinberger and Handcock (2012). Here, as will be seen in this section, the goal is to use hierarchical log-linear models to model networks with dependencies among the dyads.

Since dyads are binary variables, the model (1) can be parametrized as follows. Set, for each  $C \in C_0$  and  $x \in \{0, 1\}^m$

$$u_C(x) = \gamma_C \prod_{c \in C} x_c, \quad (2)$$

where  $\gamma_C \in \mathbb{R}$ . Thus, for every clique, there exists only one parameter,  $\gamma_C = u_C(1_C)$ . There are  $2^m - 1$  equations (for every  $x \in \mathcal{G}_n$  subject to the probabilities adding up to 1) and  $|C|$  parameters in the model. It is easy to show that among all dependency graphs with  $m$  vertices, the complete graph yields the largest number of parameters with  $2^m - 1$  parameters. Combining (1) and (2) yields the log-linear representation

$$P(x) = \bar{u} \exp\left\{ \sum_{C \in C} \gamma_C \prod_{c \in C} x_c \right\}, \quad x \in \mathcal{G}_n, \quad (3)$$

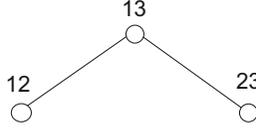
where  $\bar{u} = u_\emptyset$  is the normalizing constant, ensuring that the probabilities add up to 1.

The representation (3) holds, of course, for any arbitrary binary graphical model. In the present setting however, where each point  $x$  correspond to a network realization and the dependency graph satisfies the Markov dependence property defined above, the model (3) can be interpreted using network statistics: Recall that under Markov dependence property, cliques in  $D$  correspond to stars and triangles in the random network  $G$ . Indeed, in the representation (3), for each clique  $C$ , the term  $\prod_{c \in C} x_c$  is non-zero if and only if the subgraph of  $G$  corresponding to the dyads  $\{c \in C\}$  is

either a triangle or a star.

For example for the dependency graph in Fig. 3, we have that

$$P(x) = \bar{u} \bar{u}_{12}(x) \bar{u}_{13}(x) \bar{u}_{23}(x) \bar{u}_{12,13}(x) \bar{u}_{13,23}(x) = \bar{u} \exp\{\gamma_{12}x_{12} + \gamma_{13}x_{13} + \gamma_{23}x_{23} + \gamma_{12,13}x_{12}x_{13} + \gamma_{13,23}x_{13}x_{23}\}.$$



**FIGURE 3** A dependency graph corresponding to a network with node set  $\{1, 2, 3\}$ .

In the next sections, we use the above model to generalize the Erdős-Rényi and  $\beta$ -models in order to deal with the independence structure implied by the dependency graphs. This is done by putting constraints on the parameters  $\gamma_C$  in (3) that come from the mentioned network models. The method is independent of the choice of the network models and can be applied to other network models that assume dyadic independence.

### 3.2 | Hierarchical Erdős-Rényi models

In Erdős-Rényi models, it is assumed that ties occur independently, and the probabilities  $p_{ij}$  of observing a tie between nodes  $i$  and  $j$  are all equal to  $p$ . Hence the probability of observing a network is

$$P(x) = \prod_{i < j} p^{x_{ij}} (1-p)^{1-x_{ij}} = (1-p)^{\binom{n}{2}} \prod_{i < j} \exp\{x_{ij} \log(\frac{p}{1-p})\}, \quad x \in \mathcal{G}_n.$$

In order to come up with an Erdős-Rényi type model that captures the independence structure implied by a given  $D$ , we generalize this model in the sense that the model for the baseline, where the dependency graph is the null graph, i.e. a graph with no edges, is the same as the model above. This implies that, in this case, one can consider  $\bar{u}_{ij} = e^{q x_{ij}}$ , where  $q = \log(p/(1-p))$ , and leave  $(1-p)^{\binom{n}{2}}$  in the normalizing constant  $\bar{u}$ .

In order to define the hierarchical Erdős-Rényi model, we set the following constraints on parameters in the model in (3) that conform with the constraints in Erdős-Rényi model. The remaining parameters after setting the constraints are denoted by the vector of parameters  $q$  and a single parameter  $t$ .

$$\gamma_C = \begin{cases} q^{(r)}, & \text{if } C = \{i_1, i_2, \dots, i_r\}; \\ t, & \text{if } C = \{ij, ik, jk\}. \end{cases} \quad (4)$$

An interpretation of the parameters is provided below. The hierarchical feature of the model is such that  $q^{(r)} = 0$  implies that  $q^{(r+1)} = 0$ . In the saturated model, which corresponds to the line graph of the random network  $K_n$ , the number of parameters is  $n$ , and for dependency graphs with maximal clique of size  $d < n$ , the number of parameters is  $d + 1$ .

In the example of Fig. 3, under this model, we have

$$P(x) = \tilde{u} \exp\{q^{(1)}x_{12} + q^{(1)}x_{13} + q^{(1)}x_{23} + q^{(2)}x_{12}x_{13} + q^{(2)}x_{13}x_{23}\}.$$

This, for example, implies that  $12 \perp\!\!\!\perp 23 \mid 13$ , as is also implied by the dependency graph, since

$$P(x) = \tilde{u} \exp\{q^{(1)}(x_{12} + x_{13}) + q^{(2)}x_{12}x_{13}\} \exp\{q^{(1)}x_{23} + q^{(2)}x_{13}x_{23}\}.$$

Let  $d$  be the size of the largest clique in  $D$ . By using (4), (3) can be written in exponential family form:

$$P(x) = \exp\left\{\sum_{r=1}^d q^{(r)} s_{C^{(r)}}^{(r)}(x) + t s'_{\tau}(x) - \psi(q, t)\right\}, \quad x \in \mathcal{G}_n, \quad (5)$$

where  $C^{(r)}$  is the set of all cliques with  $r$  vertices in  $D$ ,  $q^{(r)} \in \mathbb{R}$ , and  $s_{C^{(r)}}^{(r)}(x)$  is the number of  $r$ -stars in  $x$  whose edges form a member of  $C^{(r)}$ ; similarly,  $\tau$  is the set of all cliques with 3 vertices in  $D$  of form  $\{ij, jk, ik\}$ , and  $s'_{\tau}(x)$  is the number of triangles in  $x$  whose edges form a member of  $\tau$ .

Notice that  $s_{C^{(1)}}^{(1)}(x)$  is simply the number of ties of  $x$  since all vertices of  $D$  are considered cliques.

Therefore, since parameter  $q^{(r)}$  corresponds to higher order interactions (of dimension  $r$ ) in the dependency graph, it can be interpreted as propensity for the network to possess specific  $r$ -stars related to the cliques of the dependency graph. Hence,  $q^{(1)}$  can be interpreted in the same way as the parameter  $q$  in Erdős-Rényi. Similarly,  $t$  can be interpreted as propensity for the network to possess triangles related to the dependency graph. Indeed, the sufficient statistics are correlated with each other, and the value 0 for a parameter, say  $q^{(r)}$ , means that given the value of other parameters, the number of cliques of size  $r$  is close to the average number of possible cliques of size  $r$ .

Notice also that for the saturated model, we have that

$$P(x) = \exp\left\{\sum_{r=1}^{n-1} q^{(r)} s^{(r)}(x) + t s'(x) - \psi(q, t)\right\}, \quad x \in \mathcal{G}_n,$$

where  $s^{(r)}(x)$  is the number of  $r$ -stars; and  $s'(x)$  is the number of triangles.

For example, the model corresponding to the graph in Fig. 3, can be written in exponential family form as

$$P(x) = \exp\{q^{(1)}e(x) + q^{(2)}s_{\{\{12,13\},\{13,23\}\}}^{(2)}(x) - \psi(q)\}.$$

Obtaining the normalizing constant in a closed form, for models in exponential family, in principle allows us to apply optimization methods for obtaining the maximum likelihood estimator. Notice that except in very few cases (such as Erdős-Rényi and  $\beta$ -models), the normalizing constant in ERGMs is typically not in closed form.

Here we sum over all possible values of the binary vector  $x$  in (3) after inserting the parameters in (4), and set it equal to 1 in order to calculate the normalizing constant  $\psi(q, t)$ :

$$\psi(q, t) = \log\left(1 + \sum_{r=1}^{n(n-1)/2} \sum_{H \in \mathcal{D}^{(r)}} \exp\left\{\sum_{r'=1}^{\min(d,r)} c^{(r')}(H)q^{(r')} + c'(H)t\right\}\right), \quad (6)$$

where  $\mathcal{D}^{(r)}$  is the set of all subgraphs of  $D$  with  $r$  vertices, and  $c^{(r')}(H)$  is the number of cliques of size  $r'$  in  $H$ ; and similarly

$c'(H)$  is the number of cliques of size 3 in  $H$  of form  $(ij, jk, ik)$ . This could be written as  $\psi(q) = \log((1 + e^{q^{(1)}})^{\binom{2}{2}} + f(q))$ , where the term  $\log(1 + e^{q^{(1)}})^{\binom{2}{2}}$  corresponds to cliques of size 1, which are the vertices of the dependency graph. By neglecting the other term in the logarithm, we obtain the normalizing constant for the Erdős-Rényi model.

For example, for the dependency graph in Fig. 3, we obtain

$$\psi(q) = \log(1 + 3e^q + e^{2q} + 2e^{2q+q^{(2)}} + e^{3q+2q^{(2)}}).$$

### 3.3 | Hierarchical $\beta$ -models

Next, we apply an approach analogous to the one described in the previous section to the  $\beta$ -model. The directed version of  $\beta$ -model is the  $p_1$ -model (Holland and Leinhardt, 1981), and the following approach can further generalize for  $p_1$ -model with few minor additional technicalities. For brevity, we have not included this in this paper.

In  $\beta$ -models, it is also assumed that ties occur independently, and the probability  $p_{ij}$  of observing a tie between nodes  $i$  and  $j$  is parameterized as follows:

$$p_{ij} = \frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}}, \quad \forall i \neq j, \quad \beta_1, \dots, \beta_n \in \mathbb{R}^n,$$

where  $\beta_i$  can be interpreted as the propensity of node  $i$  to have ties. The probability of observing a network is

$$P_\beta(x) = \prod_{i < j} p_{ij}^{x_{ij}} (1 - p_{ij})^{1 - x_{ij}} = \prod_{i < j} e^{(\beta_i + \beta_j)x_{ij}} \frac{1}{1 + e^{(\beta_i + \beta_j)}}, \quad x \in \mathcal{G}_n.$$

In this case, the model above, which is the model for the baseline, can be considered to be  $\tilde{u}_{ij} = e^{(\beta_i + \beta_j)x_{ij}}$  and  $1/(1 + e^{(\beta_i + \beta_j)})$  can be left in the normalizing constant  $\tilde{u}$ .

Now again suppose that there is a dependency graph  $D$  that satisfies the Markov dependence property, and is modeled by the hierarchical model (3).

We have observed in Section 2 that cliques in  $D$  correspond to stars and triangles in  $G$ . In order to define the hierarchical  $\beta$ -model, we set the following constraints on the parameters in the model in (3) that conform with the constraints in the  $\beta$ -model. The remaining parameters, after setting the constraints, are denoted by vectors of parameters  $(\beta)_i$  and a single vector of parameter  $\eta_i$ .

$$\gamma_C = \begin{cases} \beta_i^{(1)} + \beta_j^{(1)}, & \text{if } C = \{ij\} \\ \beta_i^{(r)}, & \text{if } C = \{ii_1, ii_2, \dots, ii_r\}, \quad r \geq 2 \\ \eta_i + \eta_j + \eta_k, & \text{if } C = \{ij, ik, jk\}. \end{cases} \quad (7)$$

An interpretation of the parameters is provided below. As before,  $\beta_i^{(r)}$  are hierarchical in the sense that if  $\beta_i^{(r)} = 0$  then  $\beta_i^{(r+1)} = 0$ . In the saturated model, the number of parameters is  $n^2$ , but when the maximal clique size in the dependency graph is of size  $d$ , the number of parameters is  $n \cdot d$ .

In the example of Fig. 3, under this model, we have

$$P(x) = \tilde{u} \exp\{(\beta_1^{(1)} + \beta_2^{(1)})x_{12} + (\beta_1^{(1)} + \beta_3^{(1)})x_{13} + (\beta_2^{(1)} + \beta_3^{(1)})x_{23} + \beta_1^{(2)}x_{12}x_{13} + \beta_3^{(2)}x_{13}x_{23}\}. \quad (8)$$

This, for example, implies that  $12 \perp 23 \mid 13$ , as is also implied by the dependency graph, since

$$P(x) = \bar{u} \exp\{(\beta_1^{(1)} + \beta_2^{(1)})x_{12} + (\beta_1^{(1)} + \beta_3^{(1)})x_{13} + \beta_1^{(2)}x_{12}x_{13}\} \exp\{(\beta_2^{(1)} + \beta_3^{(1)})x_{23} + \beta_3^{(2)}x_{13}x_{23}\}.$$

By using (7), (3) can be written in exponential family form:

$$P(x) = \exp\left\{\sum_{i=1}^n \sum_{r=1}^n \beta_i^{(r)} d_{i,C_i^{(r)}}^{(r)}(x) + \eta_i d_{i,\tau_i}(x) - \psi(\beta, \eta)\right\}, \quad x \in \mathcal{G}_n,$$

where  $C_i^{(r)}$  is the set of all cliques with  $r$  vertices in  $D$  such that all their vertices share  $i$ , and  $d_{i,C_i^{(r)}}^{(r)}(x)$  is the number of  $r$ -stars in  $x$  with hub  $i$  such that its endpoints pairing with  $i$  form a member of  $C_i^{(r)}$ ; similarly,  $\tau_i$  is the set of all cliques with 3 vertices in  $D$  of form  $\{ij, jk, ik\}$ , and  $d_{i,\tau_i}(x)$  is the number of triangles in  $x$  that contain  $i$  and two other vertices such that they form a member of  $\tau_i$ .

Notice that  $d_{i,C_i^{(1)}}^{(1)}(x)$  is simply the degree of node  $i$  since all vertices of  $D$  are considered cliques.

Therefore, the parameter  $\beta_i^{(r)}$  can be interpreted as the propensity of node  $i$  in the network to be the hub of specific  $r$ -stars related to the cliques of the dependency graph. Hence,  $\beta_i^{(1)}$  can be interpreted in the same way as the parameter  $\beta_i$  in the  $\beta$ -model. Similarly,  $\eta_i$  can be interpreted as propensity for node  $i$  in the network to possess triangles related to dependency graph.

For the saturated model, the model is

$$P(x) = \exp\left\{\sum_{i=1}^n \sum_{r=1}^n \beta_i^{(r)} d_i^{(r)}(x) + \eta_i d_i'(x) - \psi(\beta, \eta)\right\}, \quad x \in \mathcal{G}_n,$$

where  $d_i^{(r)}(x)$  is the number of  $r$ -stars with  $i$  as the hub; and  $d_i'(x)$  is the number of triangles that contain  $i$ . In this case the sufficient statistics  $d_i^{(r)}(x)$  are determined for  $r \geq 2$  by  $d_i^{(1)}(x)$ . However, sufficient statistics in the submodels of the saturated model,  $d_{i,C_i^{(r)}}^{(r)}(x)$ , can be arbitrary. For dense dependency graphs, the correlation between sufficient statistics can be high, and in some cases there might even be linear dependencies. Verifying this requires a case by case verification, generally, but for sparser dependency graphs this is not an issue.

For example, the model corresponding to the graph in Fig. 3, can be written in exponential family form as

$$P(x) = \exp\{\beta_1^{(1)} d_1^{(1)}(x) + \beta_2^{(1)} d_2^{(1)}(x) + \beta_3^{(1)} d_3^{(1)}(x) + \beta_1^{(2)} d_{1,\{\{12,13\}\}}^{(2)}(x) + \beta_3^{(2)} d_{3,\{\{13,23\}\}}^{(2)}(x) - \psi(\beta)\}.$$

As in the hierarchical Erdős-Rényi case, we are interested in writing the normalizing constant in a closed form in order to be able to apply optimization methods for obtaining the maximum likelihood estimator. We sum over all possible values of the binary vector  $x$  in (3) after inserting the parameters in (7), and set it equal to 1 in order to calculate the normalizing constant  $\psi(\beta, \eta)$ .

$$\psi(\beta, \eta) = \log\left(\prod_{i < j} (1 + e^{\beta_i + \beta_j}) + \sum_{r=2}^{n(n-1)/2} \sum_{H \in \mathcal{D}^{(r)}} (e^{\sum_{d=2}^r \sum_{C \in \mathcal{C}^{(d)}(H)} \beta_{\text{hub}(C)}^{(d)} + \sum_{C \in \tau(H)} (\eta_{c_1} + \eta_{c_2} + \eta_{c_3})} - 1) e^{\sum_{v \in H} \beta_{v_1} + \beta_{v_2}}\right), \quad (9)$$

where  $\mathcal{D}^{(r)}$  is the set of all subgraphs of  $D$  with  $r$  vertices,  $V(H)$  is the vertex set of  $H$ ,  $\mathcal{C}^{(d)}(H)$  is the set of all cliques in  $H$  with  $d$  vertices except the hubless cliques,  $\tau(H)$  is the set of all hubless cliques, which are of form  $\{(i, j), (i, k), (j, k)\}$ , and

we denote such  $i, j$ , and  $k$  by  $c_1, c_2$ , and  $c_3$ . In addition, we write  $v$  as  $(v_1, v_2)$ . In (9), the term  $\prod_{i < j} (1 + e^{\beta_i + \beta_j})$  corresponds to cliques of size 1, which are the vertices of the dependency graph, and neglecting the other term in the logarithm, we obtain the normalizing constant for  $\beta$ -model,  $\sum_{i < j} \log(1 + e^{\beta_i + \beta_j})$ .

Equation (9) can be used to compute the normalizing constant explicitly, although depending on the size of the network and the density of  $D$ , the computation of the sum could become intractable. For example, for the dependency graph in Fig. 3, and from (8) we obtain

$$\begin{aligned} \psi(\beta) &= \log\left(\sum_x \exp\{(\beta_1^{(1)} + \beta_2^{(1)})x_{12} + (\beta_1^{(1)} + \beta_3^{(1)})x_{13} + (\beta_2^{(1)} + \beta_3^{(1)})x_{23} + \beta_1^{(2)}x_{12}x_{13} + \beta_3^{(2)}x_{13}x_{23}\}\right) \\ &= \log(1 + e^{\beta_1^{(1)} + \beta_2^{(1)}} + e^{\beta_1^{(1)} + \beta_3^{(1)}} + e^{\beta_2^{(1)} + \beta_3^{(1)}} + e^{\beta_1^{(1)} + \beta_2^{(1)} + \beta_1^{(1)} + \beta_3^{(1)} + \beta_1^{(2)}} + e^{\beta_1^{(1)} + \beta_2^{(1)} + \beta_2^{(1)} + \beta_3^{(1)}} \\ &\quad + e^{\beta_1^{(1)} + \beta_3^{(1)} + \beta_2^{(1)} + \beta_3^{(1)} + \beta_3^{(2)}} + e^{\beta_1^{(1)} + \beta_2^{(1)} + \beta_1^{(1)} + \beta_3^{(1)} + \beta_2^{(1)} + \beta_3^{(1)} + \beta_1^{(2)} + \beta_3^{(2)}}) \\ &= \log((1 + e^{\beta_1^{(1)} + \beta_2^{(1)}})(1 + e^{\beta_1^{(1)} + \beta_3^{(1)}})(1 + e^{\beta_2^{(1)} + \beta_3^{(1)}}) + (e^{\beta_1^{(2)}} - 1)e^{2\beta_1^{(1)} + \beta_2^{(1)} + \beta_3^{(1)}} \\ &\quad + (e^{\beta_3^{(2)}} - 1)e^{2\beta_3^{(1)} + \beta_1^{(1)} + \beta_2^{(1)}} + (e^{\beta_1^{(2)} + \beta_3^{(2)}} - 1)e^{2\beta_1^{(1)} + 2\beta_2^{(1)} + 2\beta_3^{(1)}}). \end{aligned}$$

The method proposed above for Erdős-Rényi and beta is not restricted to these models. In general, if there is a model that assumes dyadic independence then this method can be applied in the following manner:

Every linear exponential random graph model can be written in the form expressed in (3). The idea is that, for every clique  $C$ , the parameter  $\gamma_C$  is further parametrized based on the baseline network model such that for every clique size  $r$  there is a family of parameters  $\{\theta_{i,k}\}_k \in \mathcal{K}$ , where  $\mathcal{K}$  is the set of parameters in the baseline network model. The baseline model corresponds to the empty dependency graph, and  $\gamma_{ij}$ , for every  $i, j$ , is reparametrized by the first order parameters  $\{\theta_{i,k}\}_k \in \mathcal{K}$  in order to obtain the baseline network model.

In general, by using the reparametrization of  $\gamma_C$ , (3) can be written in exponential family form and the sufficient statistics will show up in the model. In addition, for these models, if the normalizing constant of the baseline network model is in closed form then, by summing over all possible values of the binary vector  $x$  in the reparametrized version of (3), the normalizing constant can be written in closed form, although it still contains a sum over subgraphs of the dependency graph as opposed to a sum over all networks with  $n$  nodes.

## 4 | PARAMETER ESTIMATION AND SIMULATION STUDIES

### 4.1 | Maximum likelihood estimation

One important difference between the proposed models in this paper and other exponential random graph models that do not assume dyadic independence is that the normalizing constants (see (6) and (9)) in our proposed models are in *closed form* in the sense that they do not depend on summation over all networks. The models inherit this property from the models on which they are based (i.e. the Erdős-Rényi and  $\beta$ -models). However, the normalizing constants depend on summation over subgraphs of the corresponding dependency graph, which can still be computationally demanding. We will, however, show below that for sparse dependency graphs some computations are manageable. This is essential for implementing the ML estimation for these models. This is in contrast to other exponential random graph models without dyadic independence assumption, which usually require some type of Markov chain Monte Carlo methods in order to compute the normalizing constant; see for example Hunter et al. (2008).

Henceforth, for brevity, we assume that there is no triangle of form  $\langle ij, ik, jk \rangle$  in the dependency graph  $D$ . This implies that the parameter  $t$  can be removed from the model. However, one can trivially generalize all the calculations

and algorithms below to the case that  $t$  is existent in the model.

By (5), the log-likelihood of the Erdős-Rényi model can be written as

$$l(q) = \sum_{r=1}^{n-1} q^{(r)} s^{(r)}(x) - \psi(q),$$

where  $\psi(q)$  is defined in (6).

This function is obviously concave. The goal is to apply the *gradient descent* method (for example, see Snyman (2005)) to find its maximum. An element of the gradient of the log-likelihood, for  $m = n(n-1)/2$ , is

$$\partial l(q) / \partial q^{(i)} = s^{(i)}(x) - \frac{\sum_{r=1}^m \sum_{H \subseteq D^{(r)}} c^{(i)}(H) \exp\{\sum_{r'=1}^{\min(d,r)} c^{(r')}(H) q^{(r')}\}}{1 + \sum_{r=1}^m \sum_{H \subseteq D^{(r)}} \exp\{\sum_{r'=1}^{\min(d,r)} c^{(r')}(H) q^{(r')}\}}, \quad (10)$$

where  $d$  is the maximal clique size in  $D$ . The main problem in computing the gradient vector is the large sum in both numerator and denominator, whose number of terms in the worst case is of order  $O(n^3 d 2^m)$ . Here we will address this issue.

By calculating the term  $r = 1$  separately in (10), we obtain for  $2 \leq i \leq d$ ,

$$\partial l(q) / \partial q^{(i)} = s^{(i)}(x) - \frac{\sum_{r=2}^m e^{r q^{(1)}} \sum_{H \subseteq D^{(r)}} c^{(i)}(H) \exp\{\sum_{r'=2}^{\min(d,r)} c^{(r')}(H) q^{(r')}\}}{1 + m e^{q^{(1)}} + \sum_{r=2}^m e^{r q^{(1)}} \sum_{H \subseteq D^{(r)}} \exp\{\sum_{r'=2}^{\min(d,r)} c^{(r')}(H) q^{(r')}\}}; \quad (11)$$

and

$$\partial l(q) / \partial q^{(1)} = s^{(1)}(x) - \frac{m e^{q^{(1)}} + \sum_{r=2}^m r e^{r q^{(1)}} \sum_{H \subseteq D^{(r)}} \exp\{\sum_{r'=2}^{\min(d,r)} c^{(r')}(H) q^{(r')}\}}{1 + m e^{q^{(1)}} + \sum_{r=2}^m e^{r q^{(1)}} \sum_{H \subseteq D^{(r)}} \exp\{\sum_{r'=2}^{\min(d,r)} c^{(r')}(H) q^{(r')}\}}. \quad (12)$$

Suppose that  $D$  is *sparse* in the sense that the number of non-isolated vertices in  $D$ , denoted by  $m'$ , is of order of a constant, i.e. not growing with  $n$ , or it is increasing by a rate slower than  $n$ . This implies that, based on our assumption, there is only a finite number of individuals in the model to which the connected ties are dependent. By increasing  $n$ , the dependency graph converges to the empty graph, and the model converges to Erdős-Rényi. Denote the induced subgraph by non-isolated vertices by  $D'$ . An expression in the last term in the denominator of (11) or (12) is expanded as follows:

$$\sum_{H \subseteq D^{(r)}} \exp\left\{ \sum_{r'=2}^{\min(d,r)} c^{(r')}(H) q^{(r')} \right\} = \sum_{r''=2}^{\min(m',r)} \binom{m-m'}{r-r''} \sum_{H \subseteq D^{(r'')}} \exp\left\{ \sum_{r'=2}^{\min(d,r'')} c^{(r')}(H) q^{(r')} \right\} + \binom{m-m'}{r} + m' \binom{m-m'}{r-1},$$

where  $\binom{a}{b} := 0$ , for  $b > a$ .

Therefore, by the same argument for numerators, (11) and (12) can be written as

$$\partial l(q) / \partial q^{(i)} = s^{(i)}(x) - \frac{\sum_{r'=2}^{m'} \sum_{H \subseteq D^{(r'')}} c^{(i)}(H) \exp\{\sum_{r'=2}^{\min(d,r'')} c^{(r')}(H) q^{(r')}\} (\sum_{r=2}^{m'} \binom{m-m'}{r-r''} e^{r q^{(1)}})}{1 + \sum_{r=1}^m e^{r q^{(1)}} \left( \binom{m-m'}{r} + m' \binom{m-m'}{r-1} \right) + \sum_{r''=2}^{m'} \sum_{H \subseteq D^{(r'')}} \exp\{\sum_{r'=2}^{\min(d,r'')} c^{(r')}(H) q^{(r')}\} (\sum_{r=2}^m \binom{m-m'}{r-r''} e^{r q^{(1)}})}; \quad (13)$$

and

$$\partial l(q) / \partial q^{(1)} = s^{(1)}(x) - \frac{\sum_{r=1}^m r e^{r q^{(1)}} \left( \binom{m-m'}{r} + m' \binom{m-m'}{r-1} \right) + \sum_{r''=2}^{m'} \sum_{H \subseteq D^{(r'')}} \exp\{\sum_{r'=2}^{\min(d,r'')} c^{(r')}(H) q^{(r')}\} (\sum_{r=2}^m r \binom{m-m'}{r-r''} e^{r q^{(1)}})}{1 + \sum_{r=1}^m e^{r q^{(1)}} \left( \binom{m-m'}{r} + m' \binom{m-m'}{r-1} \right) + \sum_{r''=2}^{m'} \sum_{H \subseteq D^{(r'')}} \exp\{\sum_{r'=2}^{\min(d,r'')} c^{(r')}(H) q^{(r')}\} (\sum_{r=2}^m \binom{m-m'}{r-r''} e^{r q^{(1)}})}, \quad (14)$$

where  $\binom{a}{b} := 0$ , for  $b < 0$ .

By fixing  $q^{(1)}$ , and assuming  $m'$  is of order of a constant, since  $d$  can no longer grow with  $n$ , the number of terms in the sum in both numerator and denominator is of order  $O(C2^n)$ , where  $C$  is a constant. Now for every value of  $q_*^{(1)}$ , by the gradient descent method, the largest value for  $g(q_*^{(1)}) = \max_{q^{(2)}, \dots, q^{(d)}} l(q_*^{(1)}, q^{(2)}, \dots, q^{(d)})$  can be computed. (If  $f(x, y)$  is concave then so is  $f(x, y_*)$  for a fixed value of  $y_*$ .) Hence, one can use iterative methods in optimization to find the value of  $q^{(1)}$  within a certain range with a certain precision that maximizes the likelihood function.

## 4.2 | Computational and simulation study

We have written code that calculates the maximum likelihood estimator by the optimization method discussed above, although, at the moment it cannot computationally carry large networks ( $> 50$  nodes) (the computational difficulty depends also on the density of the dependency graph).

We have also incorporated the corresponding statistics in the package `ERGM` (Hunter et al., 2007), by coding up the change statistics in the package `ERGM.usertertterms` (Hunter et al., 2013), and can now utilize the functionality provided in the `ERGM` package such as those for approximating the maximum likelihood estimator by Markov chain Monte Carlo methods and simulating networks based on exponential random graph models.

For a given network  $x$  and dependency graph  $D$ , and in order to have some base for comparisons regardless of the difference in the density of the simulated networks, we find the *likelihood ratio* under the maximum likelihood estimator of hierarchical Erdős-Rényi and Erdős-Rényi (the latter can be exactly computed easily):

$$S = 2l_{HER}(\hat{q}_1, \dots, \hat{q}_r; x, D) - 2l_{ER}(\hat{q}; x). \quad (15)$$

Since the Erdős-Rényi is a submodel of hierarchical Erdős-Rényi, it holds that  $l_{HER} > l_{ER}$ , and hence  $S > 0$ . In addition, for two fixed hierarchical models, denoted by 1 and 2, and the same network, the difference between the log-likelihoods is proportional to  $S_1 - S_2$ .

We first provide a method to simulate a vector of size  $m$  of binary variables with conditional correlations (linear dependencies) induced by a given undirected dependency graph  $D$ . By this method the dependencies induced by the dependency graph are preserved, but there might be independencies that turn into non-linear dependencies among the simulated binary data:

We simulate an  $m \times m$  symmetric positive-definite matrix  $M$  (e.g., using of The QR Decomposition of a Matrix in LAPACK (Anderson et al., 1999)). We then change the values of the matrix in order for zeros to correspond to missing edges of  $D$  as follows. For  $A \perp\!\!\!\perp B \mid C$ , it holds that  $\text{cov}(A, B) = \text{cov}(A, C)\text{cov}(C, C)^{-1}\text{cov}(C, B)$ . Thus, start with  $M$  as a covariance matrix and cycle through all pairs corresponding to the missing edges of  $D$  by forcing the formula to hold for  $(i, j, V \setminus \{i, j\})$ . We stop when the sum of the deviations from the concentrations fitted to the missing edges (i.e. the sum of the corresponding elements on the generated matrix) is smaller than some default. For more details, see Speed and Kiiveri (1986). Call the resulting matrix  $L$ .

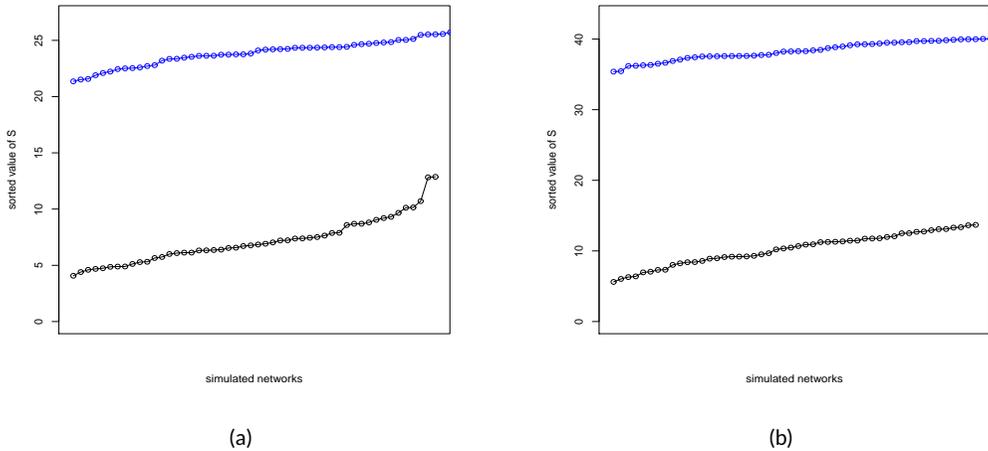
We can now consider  $K = L^{-1}$  to be the *concentration matrix*, i.e. the inverse of the covariance matrix, of a Gaussian distribution with 0 mean that is Markov to  $D$ . We generate binary random variables by thresholding a normal distribution, i.e. by setting the simulated value of a variable to 0 if the corresponding value in the Gaussian distribution is negative, and to 1 if the corresponding value in the Gaussian distribution is positive. The linear conditional dependencies are preserved under thresholding; see Leisch et al. (1998) and the R package `bindata`, introduced in its appendix.

We now assume that a simulated  $x$  is an observed network with every element corresponding to a tie. We find an approximation of the maximum likelihood estimator for the hierarchical Erdős-Rényi model by the `ERGM` package,

and also the maximum likelihood estimator for Erdős-Rényi. We then calculate the likelihood ratio  $S$ . We discard the simulated networks for which the maximum likelihood estimator does not seem to exist (as implied by ERGM).

Here we conduct our study for several dependency graphs with complete connected components of different sizes (including isolated vertices); see also Section 5 for a short discussion on model selection (based on nodal attribute or exchangeability). We performed four simulations experiments.

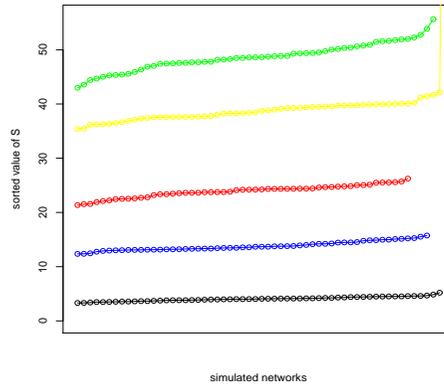
1) We simulate networks on 50 nodes from the same dependency graph, but both using low values of the partial correlations  $\rho_{ij,ik}$ , for all possible vertices  $ij$  and  $ik$ , among the dependent variables ( $|\rho_{ij,ik}| < 0.1$ ), and also high values of the partial correlations ( $|\rho_{ij,ik}| < 0.9$ ). We consider two dependency graphs: one consisting of a clique of size 25 and isolated vertices and the other of a clique of size 40 and isolated vertices. For each simulated network in each of the scenarios we compute the likelihood ratio statistic as in (15), and plot the ordered values of the corresponding statistics, in order to give a sense of the spread and tail of the their distribution. The results are presented in Figure 4. We clearly observe that the model performs better for higher correlations, which is expected.



**FIGURE 4** Ordered likelihood-ratios of networks simulated from a conditional independence structure implied by a dependency graph with 1225 vertices and (a) with one clique of size 25 together with isolated vertices, low correlations (black, at the bottom) and high correlations (blue, on top); (b) with one clique of size 40 together with isolated vertices, low correlations (black, at the bottom) and high correlations (blue, on top).

2) We simulate networks on 50 nodes with different dependency graphs with medium level correlations. The dependency graph again consists of one cliques and isolated nodes, for increasing values of the clique size (and therefore of the number of parameters in the model). It is obvious that the likelihood ratio should increase as well, which is confirmed by our study. Figure 5 shows the results of the experiment for cliques of size 5, 15, 25, 40, and 50.

3) The following study further suggests that the model works. We again simulate networks of the same size, but with different dependency graphs of the same maximal clique size. Hence, the number of parameters is the same for different dependency graphs. Here we compare models for two dependency graphs, one of which contains only one clique and isolated vertices, and the other has several disconnected cliques of the same size. In particular, we depict the values of the likelihood ratio for networks with 50 nodes simulated from two dependency graphs that, along with isolated vertices, have respectively 1 maximal clique of size 5, and 10 maximal cliques of size 5 (part (a) of Fig. 6), and 1



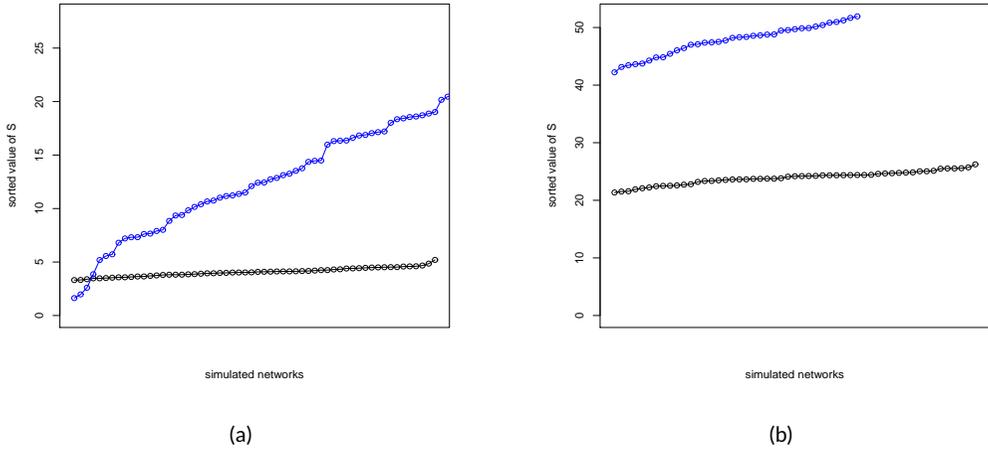
**FIGURE 5** Ordered likelihood-ratios of networks simulated from a conditional independence structure implied by a dependency graph with 1225 vertices and one clique of size 5 (black, at the bottom), 15 (blue, second from the bottom), 25 (red, in the middle), 40 (yellow, second from the top), and 50 (green, on top), and isolated vertices.

component of size 25, and 2 components of size 25 (part (b) of Fig. 6). In this experiment we discard cases where the maximum likelihood estimator does not exist (or at least cannot be computed by this method). It is seen in Fig. 6 (b) that the number of simulated networks, for which the maximum likelihood estimator seems to exist is visibly lower for the denser dependency graph. We observe the same trend in our simulation studies, where the simulated networks based on the models with denser dependency graphs are more affected by the issue of non-existence of the maximum likelihood estimator. This can be due to the fact that the value for certain statistics in some simulated networks is 0, or due to degeneracy issues in other cases.

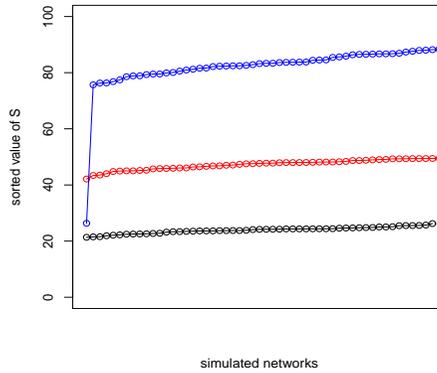
As it is often the case with exponential random graph models, the issues of existence of the maximum likelihood estimator and degeneracy (Schweinberger, 2011; Rinaldo et al., 2009; Handcock, 2003) affect also the models we propose. In fact, in our simulations we have encountered several cases in which the likelihood function does not appear to be strongly convex, as the estimated Fisher information matrix becomes nearly singular along optimizing sequences of parameters with norms diverging to infinity. This is a clear indication of a nonexistent maximum likelihood estimator and manifests itself in failed convergence of the Markov chain Monte Carlo maximum likelihood estimation procedure and slow mixing. While the investigation of these important issues is clearly outside the scope of this paper, nonetheless we recommend using small dependency graphs and a careful monitoring of the convergence of the optimizing procedure of choice.

In addition, the model with more cliques seems to have a higher variance than the model with only one clique.

4) We increase the number of nodes and try to keep the density of dependency graph (whose size increases of order of  $n^2$ ) unchanged. As expected, we see that the value of  $S$  increases when increasing the number of nodes. Here we depict examples of this value for simulated networks of size 50, 100, and 200 in the graph of Fig. 7 with dependency graphs that contain a clique of size  $n/2$  along with isolated vertices.



**FIGURE 6** Ordered likelihood-ratios of networks simulated from a conditional independence structure implied by a dependency graph with 1225 vertices and (a) one clique of size 10 (black, at the bottom) and 5 cliques of size 10 (blue, on top) together with isolated vertices; (b) one clique of size 25 (black, at the bottom) and 2 cliques of size 25 (blue, on top) together with isolated vertices.



**FIGURE 7** Ordered likelihood-ratios of networks simulated from a conditional independence structure implied by a dependency graph with  $m = \binom{n}{2}$  vertices and one clique of size  $n/2 = 25$  (black, at the bottom),  $n/2 = 50$  (red, in the middle), and  $n/2 = 100$  (blue, on top) together with isolated vertices.

## 5 | DISCUSSION AND FUTURE WORK

Our aim in this paper was to introduce a new approach for proposing a new family of models, where each model is the combination of a known network model and the hierarchical model for a given dependency graph. There are, however,

still many open questions regarding the proposed models as well as natural ways to generalize such approach.

As discussed before, for these hierarchical models, if the normalizing constant of the baseline network model is in closed form then the normalizing constant is in closed form, i.e. it does not contain a sum over all networks of size  $n$ . This is an essential aspect of these models, which makes exact parameter estimation possible. However, the normalizing constant still contains a sum over subgraphs of the dependency graph, which depending on the size and density of dependency graph could be computationally intractable. As discussed in Section 4, because of this computational demands, it is not possible to only apply standard optimization methods to find the maximum likelihood estimator. In this paper, we provided a fairly more sophisticated computational techniques that work for "sparse" dependency graphs.

The question of how to efficiently find solutions for gradients (13) and (14) is indeed an interesting optimization problem. By skipping the 1 in the denominator, a general form for the gradient that one should deal with is  $\sum c_i f_i / \sum f_i$ . The other term in the gradient is computationally manageable.

Throughout this paper we worked under the assumption that there is a given dependency graph. In general, the model depends on the labeling of the network, i.e. the individuals represented by the network. In practical cases, a sparse dependency graph may normally be provided by experts, by marking certain individuals whose relations with other individuals affect the whole or some parts of the network.

If such an expert opinion is not available, there may be several methods to select a dependency graph based on the nature of the network or the model: It is common to observe some nodal attributes along with an observed network. In this case these attributes can be used for selecting a dependency graph.

In addition, features of networks could lead to specific dependence graphs. These include *homophily*, i.e. individuals with similar characteristics being more likely to relate (Krivitsky et al., 2009), *heterophily*, where individuals relate to those with different characteristics (as seen in some social networks (Lozares et al., 2014)), *exchangeability*, in the sense that the model is invariant under relabeling of the nodes, and *transitivity*, which states that given that we know whether  $j$  and  $k$  are friends (i.e. adjacent), knowing whether  $i$  is a friend of  $j$  would impact the probability of  $i$  being a friend of  $k$ .

For exchangeable ERGMs, it was shown in (Lauritzen et al., 2018) that the corresponding dependence graph could be either empty, complete, the line graph of the complete graph, or its complement.

For the case of transitivity, for three nodes, this corresponds to  $ij \perp\!\!\!\perp k \mid jk$  not being satisfied, but the marginal independence  $ij \perp\!\!\!\perp k$  being satisfied. This can be represented by the *bidirected dependency graph*  $ij \longleftrightarrow jk \longleftrightarrow ik$ . This motivates introducing a parallel theory as in this paper to define a set of network models that preserve the independence structures of bidirected graphical models as opposed to undirected graphical models: for a *bidirected graph*  $D$ , where all edges are depicted as bidirected edges,  $\longleftrightarrow$ , the pairwise Markov property states that for two non-adjacent vertices  $i$  and  $j$ ,  $i \perp\!\!\!\perp j$ , i.e.  $i$  and  $j$  are marginally independent; whereas the global Markov property states that  $A \perp\!\!\!\perp B \mid C$  when every path between disjoint vertex subsets  $A$  and  $B$  has no vertex in  $C$ .

Suppose that a given dependency graph captures the independence structure of a network of friendships. It is plausible to assume that the existence of a tie between individuals  $i$  and  $j$  and the existence of a tie between  $k$  and  $l$  are marginally independent, i.e., knowing whether  $i$  and  $j$  are friends would not change the probability of friendship between  $k$  and  $l$  when there is no information available on the existence of other friendships in the network. This conforms with the missing edge between  $ij$  and  $kl$  in a bidirected dependency graph that satisfies the Markov dependence property. Moreover, a possible edge between  $ij$  and  $ik$  in the dependency graph indicates that knowing whether  $i$  has a friend  $j$  would impact the probability of  $i$  having another friend  $k$ .

For the purpose of modeling ERGMs based on bidirected dependency graphs, *marginal binary models* of Drton and Richardson (2008) can be used. The parameters of this model are marginal probabilities for every connected subgraph of the dependency graph (as opposed to every clique in hierarchical models). These models are *curved* exponential families. In order to make the parameter estimation more practical, one may use other marginal models introduced in

graphical models that are in curved exponential family form; see, for example, the *multivariate logistic transformation*, introduced in Lupporelli et al. (2009), and *log-mean linear models*, introduced in Roverato et al. (2013). Of course, instead of the plain marginal binary models, the idea is to apply these models to networks.

## ACKNOWLEDGEMENTS

The authors are grateful to Sonja Petrović, Thomas Richardson, Despina Stasi, and Ryan Tibshirani for very helpful discussions.

## REFERENCES

- E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, Jack J. Dongarra, J. Du Croz, S. Hammarling, A. Greenbaum, A. McKenney, and D. Sorensen. *LAPACK Users' Guide (Third Ed.)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999. ISBN 0-89871-447-8.
- Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Springer-Verlag, New York, USA, 2007.
- Joseph Blitzstein and Persi Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics*, 6(4):489–522, 2010.
- Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. *Annals of Applied Probability*, 21(4):1400–1435, 2011.
- J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, 8:522–539, 1980.
- Mathias Drton and Thomas S. Richardson. Binary models for marginal independence. *Journal of the Royal Statistical Society Series B*, 70(2):287–309, 2008.
- P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.
- Ian Fellows and Mark S. Handcock. Exponential-family random network models. 2012. <http://arxiv.org/abs/1208.0121>.
- S. E. Fienberg and S. Wasserman. Discussion of an exponential family of probability distributions for directed graphs by holland and leinhardt. *Journal of the American Statistical Association*, 76:54–57, 1981.
- Ove Frank and David Strauss. Markov graphs. *Journal of the American Statistical Association*, 81:832–842, 1986.
- Mark Handcock. Assessing degeneracy in statistical models for social networks. Technical Report 39, Center for Statistics in the Social Sciences, University of Washington, 2003.
- Paul Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76:33–50, 1981.
- 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, 2006.
- 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, 12 2007.
- David R. Hunter, Steven M. Goodreau, and Mark S. Handcock. Goodness of fit for social network models. *Journal of the American Statistical Association*, 103:248–258, 2008.

- David R. Hunter, Steven M. Goodreau, and Mark S. Handcock. *ergm.userterms*: A template package for extending statnet. *Journal of Statistical Software*, 52(2):1–25, 2013.
- Pavel N. Krivitsky, Mark S. Handcock, Adrian E. Raftery, and Peter D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31:204–213, 2009.
- S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, United Kingdom, 1996.
- Steffen L. Lauritzen, Alessandro Rinaldo, and Kayvan Sadeghi. Random networks, graphical models, and exchangeability. *Journal of the Royal Statistical Society Series B*, 80(3):481–508, 2018.
- Steffen L. Lauritzen, Alessandro Rinaldo, and Kayvan Sadeghi. On exchangeability in network models. *Journal of Algebraic Statistics*, 10(1):85–113, 2019.
- Friedrich Leisch, Andreas Weingessel, and Kurt Hornik. On the generation of correlated artificial binary data. 1998.
- Carlos Lozares, Joan Verd, Irene Cruz, and Oriol Barranco. Homophily and heterophily in personal networks. from mutual acquaintance to relationship intensity. *Quality & Quantity*, 48(5):2657–2670, 2014.
- M. Lupparelli, G. M. Marchetti, and W. P. Bergsma. Parameterization and fitting of discrete bi-directed graph models. *Scandinavian Journal of Statistics*, 36:559–576, 2009.
- Dean Lusher, Johan Koskinen, and Garry Robins. *Introduction*. Structural Analysis in the Social Sciences. Cambridge University Press, 2012. doi: 10.1017/CBO9780511894701.001.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems: networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, CA, USA, 1988.
- Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics*, 3:446–484, 2009.
- Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. Maximum likelihood estimation in the beta model. *Annals of Statistics*, 41(3):1085–1110, 2013.
- G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph ( $\rho^*$ ) models for social networks. *Social Networks*, 29:192–215, 2007.
- A. Roverato, M. Lupparelli, and L. La Rocca. Log-mean linear models for binary data. *Biometrika*, 100(2):485–494, 2013.
- 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. <http://sites.stat.psu.edu/mus47/publications/h.ergm.pdf>.
- Jan A. Snyman. *Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms*. Applied optimization. Springer, New York, 2005. ISBN 0-387-24348-8. URL <http://opac.inria.fr/record=b1132592>.
- T. P. Speed and H. T. Kiiveri. Gaussian markov distributions over finite graphs. *The Annals of Statistics*, 14(1):pp. 138–150, 1986.
- S. Thiemichen and G. Kauermann. Stable exponential random graph models with non-parametric components for large dense networks. *Social Networks*, 49:67 – 80, 2017.
- D. B. West. *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ, USA, 2001.