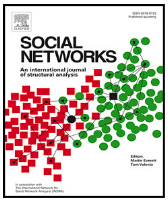
Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet



Understanding networks with exponential-family random network models

Zeyi Wang [a,*], Ian E. Fellows [b], Mark S. Handcock [a]

[a] University of California, Los Angeles, 8125 Mathematical Sciences Building, Los Angeles, 90095-1554, CA, United States
[b] Fellows Statistics, San Diego, 92107, CA, USA

ARTICLE INFO

Keywords:
Endogenous covariates
Social network modeling
ERGM
ERNM
Goodness of fit

ABSTRACT

The structure of many complex social networks is determined by nodal and dyadic covariates that are endogenous to the tie variables. While exponential-family random graph models (ERGMs) have been very successful in modeling social networks with exogenous covariates, they are often misspecified for networks where some covariates are stochastic. Exponential-family random network models (ERNMs) are an extension of ERGM that retain the desirable properties of ERGM, but allow the joint modeling of tie variables and covariates. We compare ERGM to ERNM to show how conclusions of ERGM modeling are improved by consideration of the ERNM framework. In particular, ERNM simultaneously represents the effects of social influence and social selection processes, while commonly used models do not.

1. Introduction

Social network analysis has been highly valued in the social sciences in recent decades. Statistical models are widely used in various fields to represent network structure. The well-known exponential-family random graph model (ERGM) is widely applied, where random graphs consist of a selection of nodes with some fixed nodal or dyadic covariates and random connections (edges) between nodes (Frank and Strauss, 1986; Hunter and Handcock, 2006; Lusher et al., 2013).

While many covariates, such as age, may be exogenous with the random connections, it is common for some of the covariates to be endogenous with the connections or other covariates (e.g., self-esteem, gender). ERGM is a work-horse family for modeling social networks with covariates, partially because of their generality and flexibility (Frank and Strauss, 1986; Hunter and Handcock, 2006; Lusher et al., 2013) and partially because of the quality software environments for their use (Handcock et al., 2021; Morris et al., 2008). However, they assume that the covariates are exogenous or, at least, only model the network structure conditional on the observed covariates. Hence they are misspecified for networks where some covariates are stochastic. The processes of edge and covariate formation commonly occur simultaneously (Leenders, 1997).

As an important motivating case, social and psychological theory often presumes that an individual's social relationships and their self-esteem influence each other (Leary, 2023/06/17). Harris and Orth (2020) conducted an analysis of the empirical evidence for this theory and found "that the link between people's social relationships and their level of self-esteem is truly reciprocal in all developmental stages across the life span, reflecting a positive feedback loop between the constructs". Hence treating a person's self-esteem as a fixed, unchangeable characteristic rather than representing its co-dependence with the person's social relationships will likely misrepresent the structure of the social relationships and also the relevance of self-esteem.

A generalization of ERGM called the exponential-family random network model (ERNM) was developed in Fellows (2012) and Fellows and Handcock (2012). ERNMs are flexible and interpretable models that can represent endogenous edge and node dynamics in cross-sectional data. It represents both "social selection" and "social influence" processes, where the former states that the social connections are determined by the nodal attributes (Robins et al., 2001a; Friemel, 2015) and the latter holds that the nodal attributes are determined by the social connections (Robins et al., 2001b). ERNM represents a joint exponential-family model, where some or all the nodal/dyadic attributes and social connections (edges) are treated as endogenous.

The research on the interdependence of network structure of social connections and nodal attributes has been extensively conducted in social sciences. Most statistical models that can represent this interdependence require longitudinal data (We discuss exceptions in Section 4). As such, ERNM is a valuable addition to this field.

This paper is structured as follows. In the next section (Section 2), we introduce the ERGM and ERNM classes with their model specifications, and we discuss their model interpretations. Section 3 focuses on some interesting network statistics and model estimation. In Section 4, we compare ERGM and ERNM conceptually. We show how ERNMs can correctly model the joint effects of tie variables and covariates,

[*] Corresponding author.
E-mail address: andrea.wang@ucla.edu (Z. Wang).

https://doi.org/10.1016/j.socnet.2023.07.003

while commonly used models fail when the covariates are endogenous. A case-study is conducted in Section 5, which includes a detailed modeling and analyzing study using both ERGM and ERNM for an adolescent health dataset (Harris et al., 2007). Section 6 discusses the results of the comparisons and concludes the paper.

2. ERGM and ERNM classes

We consider the situation where the network is the result of a social process modeled stochastically. For a social network (X, Y), with n nodes, $Y \subset \mathbb{R}^{n\times n} \in \mathcal{Y}$ is the graph with $X \subset \mathbb{R}^{n\times n\times q} \in \mathcal{X}$ as dyadic attributes. The space of tie variables, $\mathcal{Y}$, can be arbitrary although here we focus on binary tie variables:

$$Y_{ij} = \begin{cases} 1 & \text{if actor } i \text{ is connected to actor } j \\ 0 & \text{otherwise,} \end{cases}$$

$i, j = 1, \ldots, n$. For undirected networks, $Y_{ij} = Y_{ji}$. Y is often called an adjacency matrix. The dyadic covariates, X_{ijk}, are measures on the (i, j)th pair. An important special case is covariates that depend only on i or j, that is, nodal covariates denoted by X_{ik} or X_{jk}. Examples of dyadic covariates include a homophily term:

$$X_{ijk} = \begin{cases} 1 & \text{if actor } i \text{ and actor } j \text{ have identical values} \\ & \text{on the } k\text{th nodal characteristic } (X_{ik} = X_{jk}) \\ 0 & \text{otherwise,} \end{cases}$$

$k = 1, \ldots, q$, for each of q separate nodal characteristics. Some of these dyadic covariates can be stochastic, that is, covarying with different realizations of the tie variables, and some of them can be exogenous (that is, not varying with different realizations of the tie variables). In the case study we consider in Section 5, for example, the tie variables are friendships between students within a high school. There are covariates of the nodes (students), such as age, sex, and grade, that are appropriately modeled as non-stochastic or exogenous. However, there is also a covariate that indicates if the student smokes. This is likely covarying with the friendship tie variables due to social selection processes and/or social influence processes. Hence it is appropriate to model it as jointly stochastic with the tie variables (that is, endogenously). In most circumstances, the set of dyadic covariates will contain both exogenous and endogenous covariates. This paper is mainly interested in the situation where at least one endogenous covariate exists.

2.1. ERGM specification

The basic construction of an ERGM includes a graph $Y \in \mathcal{Y}$ that can be explained by some sufficient statistics defined by a d-vector valued function $g(\mathcal{Y})$. A general form of ERGMs that describes a probability distribution of undirected graphs with n nodes:

$$P_\eta(Y = y) = \frac{1}{c(\eta, \mathcal{Y})} \exp\{\eta \cdot g(y)\} \quad y \in \mathcal{Y}, \tag{1}$$

where $\eta \in \mathbb{R}^d$ is a parameter vector associated with a d-vector valued function $g(\mathcal{Y})$ and $c(\eta, \mathcal{Y})$ is the normalizing constant which ensures that this is a proper probability distribution. The family of models has the property of having the maximum entropy among all probability distributions that satisfy the mean constraint on $g(y)$, where $\mathbb{E}_\eta[g(y)] = \mu$. Different choices of $g(\mathcal{Y})$ determine different models within the ERGM family.

Different from traditional statistical models that measure observations with some predefined response variables and explanatory variables separately, exponential-family random graph models (ERGMs) consist of explanatory variables that are functions of response variables themselves. More specifically, in a network, the response variables are typically defined as the state of a tie y – either formation or dissolution. In general ERGMs (1), the graph statistics $g(y)$ are configurations of ties, where $g(\mathcal{Y})$ are jointly sufficient for the model. The observations in network data also consist of nodal attributes x, for example, the age of nodes. The nodal attributes can be included in ERGMs as exogenous predictors (Fienberg and Wasserman, 1981; Wasserman and Pattison, 1996). Writing this explicitly:

$$\begin{aligned} &P_\eta(Y = y|X = x) \\ &= \frac{1}{c(\eta, x, \mathcal{Y})} \exp\{\eta \cdot g(y|x)\} \quad y \in \mathcal{Y}, x \in \mathcal{X}, \end{aligned} \tag{2}$$

where $\eta \in \mathbb{R}^d$ is a d-vector of parameters. $g(y|x)$ is a d-vector of graph statistics, where $g(\mathcal{Y}|x)$ are jointly sufficient statistics. The normalization constant is $c(\eta, x, \mathcal{Y}) = \sum_{y\in\mathcal{Y}} \exp\{\eta \cdot g(y|x)\}$.

Choices of network statistics of interest depend on common knowledge and the social context. Morris et al. (2008) provides examples of the common features.

2.2. ERNM specification

Exponential-family random network models (ERNMs) generalize ERGMs by treating the nodal attributes as endogenous variables (Fellows and Handcock, 2012). This was inspired by Leenders (1997), which argued that the process of social selection and nodal attributes influence are simultaneous. ERNMs model the joint relationship between edges and nodal variates. The ERNM distribution for Y is

$$\begin{aligned} &P_\eta(Y = y, X = x) \\ &= \frac{1}{c(\eta, \mathcal{N})} \exp\{\eta \cdot g(y, x)\} \quad (y, x) \in \mathcal{N}, \end{aligned} \tag{3}$$

where $\mathcal{N}$ is the sample space of Y and X, $\eta \in \Lambda$ is a q-vector of parameters, $g(y, x)$ is a q-vector of network statistics, with $g(Y, X)$ jointly sufficient for the model, and $c(\eta, \mathcal{N})$ is the normalization constant. The formal definition of $c(\eta, \mathcal{N})$ is given in Fellows and Handcock (2012): Let $(N, \mathcal{N}, P_0)$ be a σ-finite measure space with reference measure P_0. Then, a probability measure to this space is an ERNM if it is dominated by P_0. The normalization constant is defined as

$$c(\eta, \mathcal{N}) = \int_{y,x\in\mathcal{N}} \exp\{\eta \cdot g(y, x)\} dP_0(y, x), \tag{4}$$

where $\Lambda \subset \{\eta \in \mathbb{R}^q : c(\eta, \mathcal{N}) < \infty\}$.

2.3. Model interpretation

To interpret the coefficient of ERGM, consider the logit form of exponential family models (2):

$$\text{logit}\left(P_\eta(Y_{ij} = 1|Y^c_{ij} = y^c_{ij})\right) = \eta \cdot \left(g(y^+_{ij}) - g(y^-_{ij})\right), \tag{5}$$

where y^c_{ij} is the set of tie values $y\backslash y_{ij}$, y^+_{ij} and y^-_{ij} correspond to the graphs $(y^c_{ij}, y_{ij} = 1)$ and $(y^c_{ij}, y_{ij} = 0)$, respectively, Y^c_{ij} is the random variable $Y\backslash Y_{ij}$. This is often referred to as the conditional log-odds of a tie Y_{ij}. We see that η has the interpretation of the change in conditional log-odds of a tie Y_{ij} per unit change in the graph statistics were y_{ij} toggled from zero to one.

The interpretation of graph statistics Y_{ij} of ERNM is very similar to ERGM. The only difference is that ERNM models need to consider the covariates X. From (3) we have:

$$\begin{aligned} &\text{logit}\left(P_\eta(Y_{ij} = 1|Y^c_{ij} = y^c_{ij}, X = x)\right) \\ &= \eta \cdot \left(g(y^+_{ij}, x) - g(y^-_{ij}, x)\right), \end{aligned}$$

so that the ERNM parameter has a closely allied interpretation to that of the ERGM parameter as the distribution of Y explicitly conditional on X is the same as the ERGM with implicit conditioning on X. This implies that we are able to interpret the parameter of ERNMs in a similar fashion as ERGMs. See Appendix A.2 for a derivation.

There are other interpretations as well. Fellows and Handcock (2012) discusses the way to interpret coefficients of dyadic variables of ERNM with logistic regression. The dyadic attributes X in Eq. (3)

can be partitioned into two parts, Z and T by defining $Z \in (0,1)$ as a binary dyadic variate of interest (that is an outcome variable) and T as a matrix of regressors, where $Z, T \subset X$. We can rewrite Eq. (3) with the new definition:

$$\begin{aligned} & P_\eta(Y = y, Z = z, T = t) \\ = & \frac{1}{c(\eta, \mathcal{N})} \exp\{\alpha \cdot g(y,t) + \lambda \cdot h(y,z) + z \cdot t\beta\} \\ & (y,x) \in \mathcal{N}, \end{aligned} \tag{6}$$

where $\eta = (\alpha, \beta, \lambda)$ are parameters, $g(y,t)$ and $h(y,z)$ are network statistics, $z \cdot t\beta$ is the relationship of T to Z. We can then derive the logit form of the distribution of z_{ij} from Eq. (6) condition upon the rest of the network (proof details in Appendix A.2):

$$\begin{aligned} & \text{logit}\left(P_\eta(z_{ij} = 1 | z^c_{ij}, t_{ij}, Y = y)\right) \\ & = (t_{ij}\beta) - \left(\lambda \cdot (h(y, z^-_{ij}) - h(y, z^+_{ij}))\right), \end{aligned} \tag{7}$$

where z_{ij} and t_{ij} are the measures on the (i,j)th pair of Z and T, z^c_{ij} is the set of variants $z \setminus z_{ij}$, z^+_{ij} and z^-_{ij} correspond to the variant of z_{ij} where $z_{ij} = 1$ and $z_{ij} = 0$, respectively. Suppose the matrix of regressors t_{ij} changes to t'_{ij}, with all other variables and networks remaining fixed, then the logarithm of odds ratio (R) is

$$\ln R = \beta \cdot (t_{ij} - t'_{ij}). \tag{8}$$

Therefore, the coefficients of the outcome variable z may be interpreted as a conditional logistic regression model. For one unit change in t_{ij}, the log-odds changes by β, keeping all other variables constant.

3. Specification and estimation for ERGM and ERNM

Choices of statistics for modeling are very flexible and case-based. Statistics like edges, mutuality, homophily, and transitivity are primary choices to be included in the model to grasp the major characteristics of the network. For example, the R package `ergm` contains over one-hundred “terms”, each being a coherent set of graph statistics (R Development Core Team, 2022; Handcock et al., 2021; Morris et al., 2008).

The set of network statistics for ERNM includes those for the ERGM, with the difference that they have different roles in the model due to the endogeneity of nodal attributes. Moreover, some statistics, for example, those that involve the nodal characteristics but not the tie values, are specific to ERNM but not to ERGM. An example of such a statistic, one with an important role below in this paper, is the number of students who are smokers in Section 5.2.

3.1. Primary network statistics for ERGM and ERNM

Some fundamental statistics, for example, edges, density, and mutuality measure the overall propensity for a tie in the network, and more sophisticated terms can be found in Morris et al. (2008) and Handcock et al. (2021).

Here we focus on two interesting features: homophily and transitivity. Homophily measures the tendency of individuals with similar attributes to connect compared to individuals with dissimilar attributes. There is uniform homophily and differential homophily. Uniform homophily counts the number of ties where the attributes of the two incident nodes are the same. Differential homophily accounts for each value of the identical attributes, so it will give k statistics if there are k unique values of the attribute. This is called `nodematch` in `ergm` nomenclature.

Informally, transitivity is the tendency of people to cluster together. A famous analogy illustrates this as “the friend of my friend is my friend”. Hence, this term exhibits a triad-closure feature. This tendency can be quantified: consider a three-number summary of the triads in an undirected graph or network being the number of ties, $\frac{1}{3}\sum_{ijk} y_{ij} + y_{jk} + y_{ik}$, the number of two-stars, $\sum_{ijk} y_{ij}y_{ik}$, and the number of triangles, $\sum_{ijk} y_{ij}y_{ik}y_{jk}$. A sophistication of the number of triangles is the edgewise shared partner statistics, ESP(k), representing that the number of unordered pairs $\{i,j\}$ such that an edge exists between i and j, i and j have exactly k common neighbors. A highly transitive graph would have a lot of triangles relative to the number of two-stars. It would seem natural to include these three statistics in a model and use them to measure transitivity. However, models with these terms in them have been shown to have bad statistical properties, referred to as model degeneracy. So we will instead use a measure of transitivity that does have better properties: the geometrically weighted edgewise shared partner statistics (GWESP) (Hunter and Handcock, 2006). A tie that closes triangles is more likely to form than a tie that does not close triangles. As more shared partners of an edge exist, the effect on GWESP decays in a geometric sequence, down-weighting the influence on GWESP as a measure of transitivity of these clustered triangles.

Homophily measures the tendency of individuals to connect with similar individuals, and it is more fundamental than transitivity. To some degree, homophily produces transitivity. Hence, we expect the inclusion of transitivity terms in the model to measure above and beyond what homophily does.

We will include both homophily and GWESP terms under ERGMs and ERNMs in the case-study. The performance in capturing transitivity and homophily of the two models will be analyzed.

The nodecount term counts the number of nodes with the certain attribute value of a covariate variable. This term is specifically for ERNM because this count is invariant in ERGM. The typical ERNM would include nodecount terms since they have the same role for the prevalence of covariate values as the Edges term does for tie density. As the ERNM allows endogenous nodal attributes, the nodecount term on the stochastic covariate is used to capture its random prevalence.

3.2. Degeneracy and MCMC diagnostics

The inference on ERGM parameters typically employs Markov Chain Monte Carlo (MCMC) procedures to compute the maximum likelihood estimation (MLE) (Hunter and Handcock, 2006). Likelihood-based inference for ERGMs with only dyadic independence terms can be computed easily and deterministically. For models with dyad dependence, as for the models considered here, certain combinations of terms result in the model not placing sufficient probability mass on realistic graphs and networks. This is called the model degeneracy problem (Handcock, 2003a,b; Schweinberger, 2011; Schweinberger et al., 2020). It is important to look at the diagnostics of the algorithms used to approximate the likelihood function to ensure they are computationally accurate and the model is realistic.

4. Comparing ERGM to ERNM conceptually

A main goal of social network analysis is to model the relationship between social ties in the context of nodal attributes. Two types of processes are commonly considered: social selection and social influence. In social selection processes, individuals form social ties on the basis of attributes, theirs, and others (Robins et al., 2001a; Friemel, 2015). In social influence processes, the direction is reversed, where the network structure influences the attributes of the individuals in the network; that is, an individual’s attributes may be changed by other individuals whom they share social ties with (Robins et al., 2001b). We follow the definition of the social selection and social influence processes in Leenders (1997):

1. Social selection process: Conventional network statistical models represent the network structure stochastically, measure the dynamic change of the network, and treat the nodal attributes as independent variables (or explanatory variables), usually stable

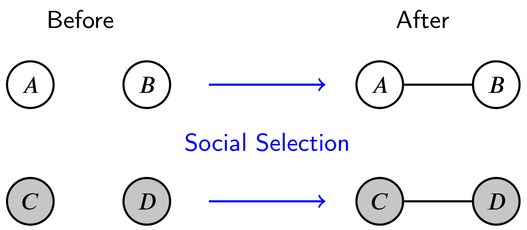


Fig. 1. Illustration of social selection: Color of nodes: nodal attributes.

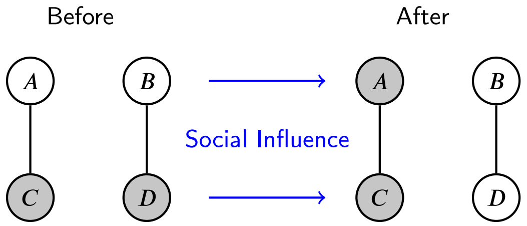


Fig. 2. Illustration of social influence: Color of nodes: nodal attributes.

and fixed. The nodal attributes in this context refer to some fixed actor characteristics, such as age, gender, and race. The involved process is called the social selection process, where the network structure is determined by some fixed actor attributes.

2. Social influence process: Conversely, actors may alter their attributes because of the influence of the network structure they are embedded in. In other words, the network structure is conditioned on and regarded as exogenous and is invariant over time, whereas the actor attributes are modeled stochastically. This process is called the social influence process (some literature uses the term "contagion").

The two processes can be illustrated using network examples shown in Figs. 1 and 2. Both figures contain networks with four nodes (actors), A, B, C, D, and social ties. In Fig. 1, two new ties are formed between nodes with the same color (A and B, C and D). It is the tendency of which an individual makes connections with other individuals that share the same attributes, such as age or habits, which is referred to as social selection. In Fig. 2, node A and node D adjust their color to the nodes which they share ties with (B and D, respectively). Hence, social influence processes describe the tendency that individual may change their attributes influenced by other related individuals.

It is essential to note that these are rarely disjoint processes. Specifically, we would expect that a mixture of social selection and social influence processes occur simultaneously. It is extensively argued that the two processes are not mutually exclusive: the social ties affect the nodal attributes and vice versa (Erickson, 1988; Leenders, 1997). As we shall see, this is precisely the situation that ERNM represents, while ERGM and autologistic actor attribute models (ALAAM) (Robins et al., 2001b; Lusher et al., 2013, Chapter 9) do not. ALAAMs are developed as alternatives to ERGMs to capture social influence processes. ALAAMs model how network relationships influence the nodal attributes, for instance, how friendship of adolescents may influence their smoking behaviors (nodal attributes). Hence, ERNM can be thought of as jointly ERGM and ALAAM, plus more. Explicitly, ERGM and ALAAM are each given by conditional views of ERNM. ERGM and ALAAM are each special cases of ERNM, a point we make here and one developed more fully in Fellows and Handcock (2012). ERNM represents the joint connection between the social selection and social influence processes. Consequently, both nodal attributes and social connections are treated as endogenous and stochastic variables, reflecting reality. Moreover, despite what many network models assume, it is implausible to have invariant nodal attributes in the network, as ERGM assumes. Although some reference nodal attributes in social selection processes, such as sex, age, and race, are invariant in social influence processes, many other attributes, such as smoking and drinking behaviors, may be altered by the network structure. Despite this, the two social network processes are widely studied in the literature separately, the mutual interdependence between the two processes being rarely jointly modeled.

ERNMs are exponential-family graph models that model the joint behavior of edges and nodal attributes. The model (3) can be rewritten as

$$P_\eta(Y=y, X=x) = P_\eta(Y=y|X=x)P_\eta(X=x), \tag{9}$$

where

$$P_\eta(X=x) = \frac{c(\eta;x)}{c(\eta;\mathcal{N})} \quad x \in \mathcal{X}. \tag{10}$$

The first component of (9) can be viewed as an ERGM that is conditional on nodal attributes X (Frank and Strauss, 1986; Hunter and Handcock, 2006):

$$\begin{aligned} &P_\eta(Y=y|X=x)\\ =&\frac{1}{c(\eta;\mathcal{N}(x),x)}\exp\{\eta\cdot g(y,x)\} \quad y\in\mathcal{N}(x), \end{aligned} \tag{11}$$

where $\mathcal{N}(x)=\{y:(y,x)\in\mathcal{N}\}$. The second component $P_\eta(X=x)$ is the marginal distribution of the nodal variate X, which does not necessarily belong to a non-trivial exponential family. The rewritten form (9) illustrates the difference between ERNM and ERGM: the former models the joint behavior of Y and X, whereas the latter models the conditional distribution of Y given X.

In the ERGM (2), the graph statistics $g(y|x)$, equivalent to $g(y,x)$ in (11), model the network conditioning on the nodal attributes. In other words, the formation or dissolution of an individual's social ties is influenced by other individuals' fixed nodal attributes. Hence, the nodal attributes are treated as exogenous to the model, and in many real situations, this assumption is inappropriate. On the contrary, ERNMs use the network statistics $g(y,x)$, which brings more flexibility to the modeling. Moreover, ERNMs take care of both nodal attributes and dyadic variables, different from ERGMs, which stochastically model the tie variables only. As a consequence, ERNMs are conceptually able to model both the social selection and the social influence processes with endogenous nodal attributes. To further illustrate the two models under the context of the two social processes, consider homophilous selection, which is measured by homophily terms in ERGM and ERNM:

ERGM homophily selection:

$$\begin{aligned} &P_\eta(Y=y|X=x)\\ =&\frac{1}{c(\eta;\mathcal{N}(x),x)}\exp\{\eta\cdot \text{homophily}(y|x)\} \quad y\in\mathcal{N}(x). \end{aligned} \tag{12}$$

ERNM Homophily selection:

$$\begin{aligned} &P_\eta(Y=y, X=x)\\ =&P_\eta(Y=y|X=x)P_\eta(X=x)\\ =&\frac{1}{c(\eta;\mathcal{N}(x),x)}\exp\{\eta\cdot \text{homophily}(y|x)\}\\ &E_Y[P_\eta(X=x|Y)] \qquad (y,x)\in\mathcal{N}. \end{aligned} \tag{13}$$

Both ERGM and ERNM capture social selection. By controlling for other alternative mechanisms, we can achieve a more accurate homophilous selection result (Steglich et al., 2010). Because ERGM treats the nodal attributes as exogenous, it does not represent any social influences. ERNM is able to reflect the social influence at the same time since the nodal attributes are free to vary on the basis of the fixed network structure:

$$P_\eta(Y=y, X=x) = P_\eta(X=x|Y=y)P_\eta(Y=y)$$

$$P_\eta(X=x|Y=y) = \frac{1}{c(\eta;\mathcal{N}(y),y)}\exp\{\eta\cdot g(y,x)\}, \tag{14}$$

$$x \in \mathcal{N}(y) \qquad\qquad (y,x) \in \mathcal{N}$$

where $\mathcal{N}(y) = \{x : (y,x) \in \mathcal{N}\}$. The term (14) represents the ALAAM class (Robins et al., 2001b). This decomposition makes the relationship between ALAAM and ERNM transparent. ALAAM is the ERNM conditional over the network tie structure. So it is a special case of ERNM.

The differences between ERNM and ERGM go beyond homophily terms involving endogenous covariates. It applies to all terms in the ERNM/ERGM (e.g., k-stars, degrees, GWESP, GWDSP). This is direct for terms involving endogenous covariates but also indirectly via interactions between model statistics. For example, the presence of direct terms changes the interpretation and coefficients of the other terms in the model.

It is essential to note that ERNM models the association between ties and nodal attributes and is not a causal model. With cross-sectional data and no causality specified, we cannot preclude social selection from other mechanisms (Steglich et al., 2010). For example, if we are trying to model the adolescent connections on smoking behavior, the homophilous selection modeled on the smoking attribute may preclude the transitivity or reciprocity processes. Other mechanisms, like similarity in drinking behaviors, may also be masked. Although we present a way to jointly model the social selection and social influence processes in a cross-sectional context, in order to disentangle the two processes, longitudinal data is needed (Leenders, 1997).

Any specific joint probability model for Y and X can be represented by a member of the ERNM family by the appropriate choice of network statistics. However, this is not anywhere as strong as it seems, as the statistics are unknown and could be of arbitrary complexity and number. Because of this, there is a great need and opportunity for highly structured models that represent much of the complex structure of the network in a relatively simple fashion. For example, Almquist and Butts (2014) introduces a vertex process temporal ERGM for modeling joint edge and behavior dynamics but makes limiting assumptions so that the model is tractable. Fosdick and Hoff (2015) presents a latent variable model that contains both additive and multiplicative latent effects and is able to represent complex network structures, including within-dyad correlation. This model can be motivated by the concept of an underlying social space with the model a reduced dimension representation of the network structure (Hoff, 2005). It jointly models nodal covariates and ties variables via this latent space. Like the ERNM, it requires careful specification of the latent and manifest variable structure. Weng (2020) develop a separable model for the tie variables and endogenous covariates and introduce individual-specific random effects to represent individual unobserved heterogeneity influencing both network formation and the covariates. These models make quite different assumptions than ERNM, and a comparison would need to be in-depth and application specific.

5. Case-study indicating the need for ERNM over ERGM

5.1. Introduction to the adolescent health data

Much network data on school friendships were collected by the National Longitudinal Study of Adolescent Health (Harris et al., 2007) (Add Health). This nationally representative study includes a longitudinal sample of more than 20,000 adolescents in grades 7 to 12 who were surveyed with in-school questionnaires in the US in 1994 and 1995. Their smoking behaviors are recorded by asking whether they have ever smoked at least once. In our study, we used four networks of grades 9 to 12 (Clark and Handcock, 2022). The smoking behavior is coded as a binary variable, where 1 was used for students who reported that they have ever smoked at least once and 0 otherwise.

Students were asked to nominate up to 5 other students who were their best female friends and also up to 5 other students who were their best male friends. We build the network of weak friendship ties, that is, an undirected tie if either student nominated the other. A visualization of the networks is shown in Fig. 3. The edges are undirected, as a connection between nodes A and B may represent A nominated B, B nominated A, or both. Table 1 shows the summary of each network. It is clear that the higher the grade, the greater the proportion of smoking adolescents, which is to be expected.

Table 1
Network summary.

	Network summary				
	Nodes	Edges	Non-smoker	Smoker	Smoking ratio
Grade 9	256	617	168	88	0.3438
Grade 10	228	498	138	90	0.3947
Grade 11	192	416	118	74	0.3854
Grade 12	193	413	101	92	0.4767

We are specifically interested in smoking behavior, especially its interconnection to the network structure. Unlike other measurements like sex, age, and race that are exogenous to the network, smoking behavior may be influenced by social connections and is expected to be endogenous.

5.2. Models

We fitted the same model terms for both ERGM and ERNM. The computational aspects are discussed in Section 5.3. In the nomenclature of that section, these are:

- ERGM: `Network` ~ `edges` + `ESP(0, 1, 2)`
  +`GWESP(0.5)` +`GWDegree(0.5)`
  +`nodefactor(smoke)`
  +`nodematch(smoke)`
- ERNM: `Network` ~ `edges` + `ESP(0, 1, 2)`
  +`GWESP(0.5)` +`GWDegree(0.5)`
  +`nodefactor(smoke)`
  +`nodematch(smoke)` | `smoke`

For grade 12, only ESP(0) is used. In the first analysis, all the statistics included in ERGM and ERNM are the same. What causes the difference between the two models is that ERNM treats the smoke indicator variable as stochastic, whereas ERGM treats it as fixed. Because of this feature of ERNM, we can, and do, fit another model to ERNM by adding the node count of smokers statistic to the first model:

- ERNM-Count: `Network` ~ `edges` + `ESP(0, 1, 2)`
  +`GWESP(0.5)` +`GWDegree(0.5)`
  +`nodefactor(smoke)` + `nodematch(smoke)`
  +`nodecount(smoke)` | `smoke`

The `edges` term represents the overall density of the network. The ESP term models the lower end of the shared partner distribution. The GWESP term with decay parameter 0.5 is a much more robust measure of transitivity than triangles (as discussed in Section 3.1). The `nodefactor` term counts the number of times a node appears in an edge for each value of the attribute. `GWDegree` stands for geometrically weighted degree distribution, specifically representing some aspects of the degree distribution (Morris et al., 2008). The `nodematch` term on smoking behavior measures the number of edges with the same smoking behaviors on two ends. In other words, it counts the edge $(i-j)$ when i and j are both smokers or non-smokers. Note that this model is saturated on smoking based mixing behavior (Handcock et al., 2021). The node count of smokers counts the number of smokers and this is specifically for ERNM because the number of smokers is invariant in ERGM.

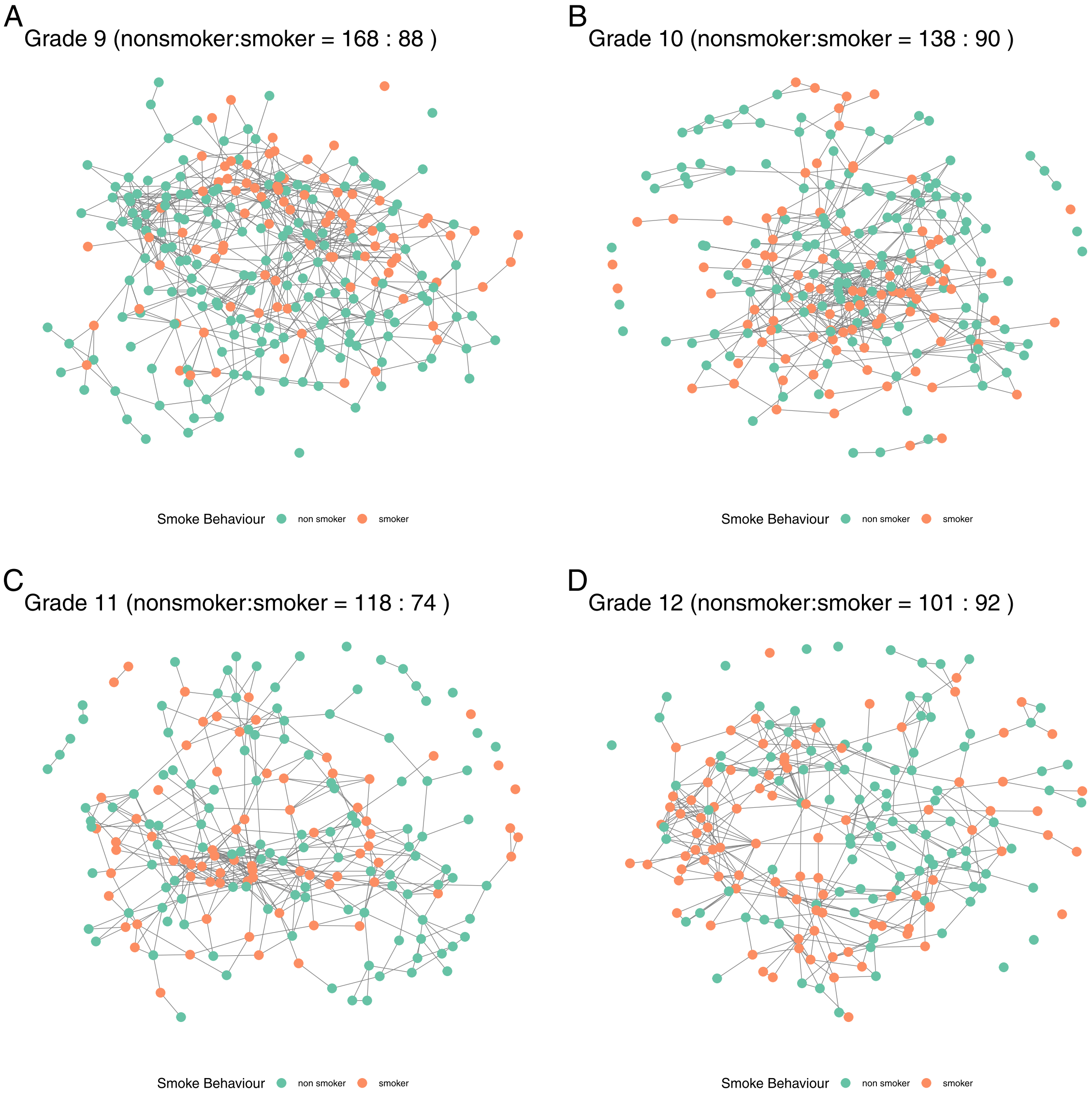

A
Grade 9 (nonsmoker:smoker = 168 : 88)
B
Grade 10 (nonsmoker:smoker = 138 : 90)
Smoke Behaviour ● non smoker ● smoker
C
Grade 11 (nonsmoker:smoker = 118 : 74)
D
Grade 12 (nonsmoker:smoker = 101 : 92)
Smoke Behaviour ● non smoker ● smoker

Fig. 3. Addhealth network visualization.

5.3. Computational aspects

All models in this paper are fit with the open-source user-friendly R packages `ergm` (Handcock et al., 2021) or `ernm` (Fellows, 2014; R Development Core Team, 2022). The easy availability of powerful, sophisticated community supported software allows broad accessibility of both these modeling classes for researchers. In particular, the `ergm` package is a part of the `statnet` community of packages (Krivitsky et al., 2003–2020). Together these allow robust MCMC based maximum likelihood estimation of ERGM and ERNM model parameters. In addition, they offer powerful models and computational diagnostic tools that we applied here and are available to all. We do not focus on these computational aspects here but refer the reader to the extensive material in the references.

As we discussed earlier in Section 3.2, due to the intricate dependence feature of ERGM and ERNM, computation of approximate maximum likelihood estimates of the parameters may be complicated by model degeneracy. MCMC diagnostics are needed to check the appropriation of the model, in other words, whether the model converges. From the results of MCMC diagnostics (Appendix A.2), the trace plots of simulated statistics from the fitted model indicate low dependency and Markov chains convergent to the stationary distribution for both ERGM and ERNM. The MCMC samplers mix well.

5.4. ERGM and ERNM fits

We show the results of ERGM fit under the suggested model of four networks (corresponding to grades 9, 10, 11, and 12) in Table 2. We can interpret the coefficients using the log-odds definition in Section 2.3. The combination of Edges, Diff-homophily-smoke and Homophily-non-smoker represents the propensity for forming a tie between all the possible combinations of smoking attributes between paring of nodes. The baseline (Edges) corresponds to a heterogeneous pairing. The Homophily-non-smoke term represents the homophily for non-smokers. All networks exhibit a positive estimated coefficient on the homophily of non-smokers (although Grade 11 is not significant). To interpret this result, taking Grade 9 as an example, the positive coefficient estimate of homophily on non-smoking (0.40) suggests that students who have not smoked are more likely to nominate as friends others who have not smoked (all else held constant). The Diff-homophily-smoke coefficients give us the differential homophily for smokers. In other words, it represents the excess (or differential) homophily for smokers over that for non-smokers. Summing the Homophily-non-smoker and Diff-homophily-smoke coefficients give us the homophily for smokers. All networks exhibit stronger homophily for smokers (although Grade 12 is only marginally significant). To interpret this result, taking Grade

Table 2
Summary of the fit of the ERGM in Section 5.2 on four grades.

Coefficient	ERGM			
	Grade 9	Grade 10	Grade 11	Grade 12
Edges	−8.97(2.99)∗∗	−19.55(3.06)∗∗∗	9.04(3.16)∗∗	−6.73(0.28)∗∗∗
ESP-0	3.35(2.99)	14.01(3.05)∗∗∗	3.87(3.15)	1.21(0.23)∗∗∗
ESP-1	0.67(1.11)	4.77(1.17)∗∗∗	0.81(1.18)	NA
ESP-2	−0.23(0.40)	1.38(0.45)∗∗	−0.05(0.43)	NA
GWESP	3.85(1.88)∗	10.44(1.89)∗∗∗	4.06(1.97) .	2.40(0.18)∗∗∗
GWDegree	2.98(0.48)∗∗∗	1.58(0.32)∗∗∗	1.66(0.35)∗∗∗	1.73(0.33)∗∗∗
Diff-homophily-smoke	0.10(0.02)∗∗∗	0.06(0.01)∗∗∗	0.10(0.03)∗∗∗	0.06(0.03) .
Homophily-non-smoker	0.39(0.06)∗∗∗	0.40(0.06)∗∗∗	0.12(0.08)	0.33(0.08)∗∗∗

Homophily-non-smoker is the number of ties between nodes with the same smoker activity;
Diff-homophily-smoke is the number of ties incident on a non-smoking node;
∗∗∗p < 0.001, ∗∗p < 0.01, ∗p < 0.05, . p < 0.1.

Table 3
Summary of the fit of the ERNM in Section 5.2 on four grades.

	ERNM			
	Grade 9	Grade 10	Grade 11	Grade 12
Edges	−7.95(2.95)∗∗∗	−19.36(3.12)∗∗∗	−9.54(3.19)∗∗∗	−6.82(0.27)∗∗∗
ESP-0	2.23(2.94)	13.79(3.12)∗∗∗	4.24(3.18)	1.24(0.24)∗∗∗
ESP-1	0.28(1.10)	4.69(1.20)∗∗∗	0.94(1.20)	NA
ESP-2	−0.33(0.40)	1.36(0.47)∗∗∗	−0.01(0.45)	NA
GWESP	3.12(1.85)∗	10.30(1.93)∗∗∗	4.30(1.99)∗∗	2.42(0.18)∗∗∗
GWDegree	3.05(0.48)∗∗∗	1.59(0.33)∗∗∗	1.63(0.35)∗∗∗	1.68(0.33)∗∗∗
Diff-homophily-smoke	0.01(0.03)	0.02(0.03)	−0.01(0.04)	0.00(0.04)
Homophily-non-smoker	0.37(0.08)∗∗∗	0.40(0.08)∗∗∗	0.14(0.11)	0.35(0.09)∗∗∗

Homophily-non-smoker is the number of ties between nodes with the same smoker activity;
Diff-homophily-smoke is the number of ties incident on a non-smoking node;
∗∗∗p < 0.001, ∗∗p < 0.01, ∗p < 0.05, . p < 0.1.

Table 4
Summary of the fit of the ERNM including the count of the number of smokers on four grades (Section 5.2).

	ERNM-Count			
	Grade 9	Grade 10	Grade 11	Grade 12
Edges	−8.80(2.95)∗∗∗	−19.56(3.12)∗∗∗	−8.97(3.22)∗∗∗	−6.73(0.28)∗∗∗
ESP-0	3.18(2.94)	14.02(3.11)∗∗∗	3.80(3.20)	1.21(0.24)∗∗∗
ESP-1	0.62(1.10)	4.77(1.19)∗∗∗	0.78(1.21)	NA
ESP-2	−0.24(0.40)	1.38(0.47)∗∗∗	−0.05(0.45)	NA
GWESP	3.73(1.85)∗∗	10.45(1.93)∗∗∗	4.02(2.01)∗∗	2.40(0.18)∗∗∗
GWDegree	2.97(0.47)∗∗∗	1.58(0.32)∗∗∗	1.66(0.35)∗∗∗	1.72(0.34)∗∗∗
Diff-homophily-smoke	0.10(0.03)∗∗∗	0.06(0.03)∗∗	0.10(0.05)∗∗	0.06(0.05)
Homophily-non-smoker	0.39(0.08)∗∗∗	0.40(0.08)∗∗∗	0.11(0.11)	0.34(0.09)∗∗∗
Nodecount-smoker	−0.62(0.20)∗∗∗	-0.45(0.18)∗∗	−0.85(0.25)∗∗∗	−0.42(0.24)∗

Homophily-non-smoker is the number of ties between nodes with the same smoker activity;
Diff-homophily-smoke is the number of ties incident on a non-smoking node;
Nodecount-smoker is the number of students that report smoker activity;
∗∗∗p < 0.001, ∗∗p < 0.01, ∗p < 0.05, . p < 0.1.

9 as an example, this indicates that the homophily for smokers is more than that for non-smokers by about 0.10 on the log-odds scale, and all else held constant. In other words, friendship formed between two smokers is more likely than a friendship formed between two non-smokers. The result of the two terms is evidence that the homophilies between smokers and between non-smokers are both stronger than the heterogamy of smokers. In other words, non-smokers are less likely to make friends with smokers (or vice versa), all other aspects held constant. Moreover, the model indicates that the homophily of smokers is generally higher than the homophily of non-smokers.

The GWESP and ESP terms together model the edgewise shared partner distribution and represent the transitivity in the network (not represented by the other terms, especially the homophily terms). The three ESP terms account for the number of pairs of nodes having 0, 1, and 2 alter in common. The adjustment for any deviation in the lower end of the edgewise shared partner distribution for the additional transitivity implied by the GWESP term. We see that the total effects of these terms are positive and significant in all but the Grade 11 network. This indicates that there is generally transitivity above and beyond that implied by the level of homophily in the network.

From the ERGM fit, we conclude that the social connections of 9 to 12-grade adolescents generally show a tendency for transitivity. The homophily of non-smokers and smokers are both positive, with that of smokers higher than non-smokers. Adolescents with different smoking behaviors are less likely to make friends with each other.

5.4.1. *ERNM fit for the same terms as the ERGM*

We fitted the same networks with ERNM with the same terms, and the results are shown in Table 3. As we discussed earlier, the qualitative interpretation of graph statistics of the ERNM is comparable to that of the comparable ERGM. The dyadic variables of ERNM can be interpreted under the conditional logistic regression as we show in Section 2.3. The parameter estimations of the basic terms (Edges, ESP, GWESP, and GWDegree) of ERNM are similar to those of the ERGM fit for each of the four networks. The Homophily-non-smoker term is also very similar to the ERGM coefficient. The estimated coefficients of homophily on non-smokers are positive and significant in all four networks (except Grade 11). The standard errors are on the same scale as the ERGM (around 0.1). To interpret the coefficients, take

Grade 9 as an example: the log-odds of a homogeneous tie of a non-smoker (that is a tie between non-smoker and non-smoker) versus a heterogeneous tie (that is a tie between a non-smoker and a smoker) is 0.4 higher, holding all else fixed. This suggests that there is a higher statistically significant probability of a tie between two smokers than a tie between one smoker and one non-smoker (holding all else constant). This coincides with the conclusion of ERGM. What stands out is the difference between the homophilies for smokers and non-smokers (that is, the Diff-homophily-smoke term). Unlike ERGM, which has statistically significantly higher homophily for smokers compared to non-smokers, ERNM has coefficients of homophily of smokers very close to homophily of non-smokers, which suggests *uniform homophily*. The difference in homophily on smoking is close to zero for each grade. This suggests that there is no big difference in the tendency to make friends between non-smokers and non-smokers compared to smokers and smokers. We will discuss this result in detail in the next section as it sheds light on the difference between the two models.

5.4.2. *ERNM fit when a count of the number of smokers is included*

Another ERNM (ERNM-Count) is fitted by adding the node count of the smoke term to the first model, and the results are shown in Table 4. Note that this is equivalent to the model counting the number of non-smokers as the total number of nodes/students is fixed for each grade/network.

Note that the coefficients of all terms except the added term are very close to those in the ERGM of Table 2 (which excludes the smoker count term). The reason for this is a geometric feature of exponential family models. Consider conditioning on the number of smokers in the ERNM-Count model:

$$P_\eta(Y = y, X = x) = \frac{1}{c(\eta)} \exp\{\eta \cdot g(y, x) + \eta_c n_{smokers}(x)\}$$

$$P_\eta(n_{smokers}(X) = n_{smokers}) = \frac{c(\eta; n_{smokers})}{c(\eta)}$$

$$P_\eta(Y = y, X = x | n_{smokers}(X) = n_{smokers}) = \frac{1}{c(\eta; n_{smokers})} \exp\{\eta \cdot g(y, x; n_{smokers})\} \quad (15)$$

The Eq. (15) represents the exponential family form of the ERNM-Count model. By specifying the number of smokers, it acts as if conditioning on the nodal attributes X representing the smoking behavior. Recall the functional form of ERGM (2), and we would expect the parameter estimates η of ERGM and ERNM-Count relating to homophilous smoking to be close.

As Table 4 shows, the majority of the results from four networks are consistent with the previous two models, except for the homophily terms. In particular, we observe differential homophily on smoking in ERNM-Count (that is the statistically significant positive estimated coefficients of Diff-homophily-smoker term). A positive coefficient suggests that the homophily of smokers is more than the homophily of non-smokers. The Nodecount-smoker term is significant (or marginally significant in Grade 12) with negative coefficient estimates. Taking Grade 9 as an example, we see that the log-odds of a student being a smoker is −0.62, holding the rest constant. This suggests that there are fewer smokers than expected based on the social structure of the friendship ties.

Comparing the two models of ERNM, the discrepancy in homophily measures gives an illuminating finding. Without the node count on smokers, we find uniform homophily between smokers and non-smokers. However, after taking into account the node count of smokers, the homophily of smokers is driven up, for example, 0.38 to 0.49 in Grade 9, which indicates that the homophily of smokers is more than the homophily of non-smokers. And the conditional log-odds of a tie for a smoker who chooses a smoker is 0.09 higher after adding the node count of the smoker.

5.5. *Assessing goodness of fit*

It is necessary to check the goodness of fit (GOF) of fitted models to verify the rationality of the model. Hunter et al. (2008) introduced a procedure for goodness of fit, which generates simulations on target network statistics and compares them to the observed graph statistics.

The construction of the tests is as follows: for both ERGM and ERNM, we generate 1000 simulations from the fitted models and compare their simulated distributions to the observed statistics. The GOF plot (Appendix A.2) consists of statistics that are included in the model and statistics (Degree(0:20), ESP(3:10)) that are not included in the model. We provide side-by-side box plots, including the statistics mentioned above, in order to compare the models without the node count term fitted by ERNM and ERGM. We also compare the model with the node count of the smoker fitted by ERNM and without the node count model fitted by ERGM. Both ERGM and ERNM simulate distribution aligning closely to the observed statistics, which is under expectation. Although there are some deviations in the degree distribution, it does an adequate job, as no degree terms are included in the model. The comparison of the count of smokers and non-smokers between ERNM and ERNM-Count is shown in Fig. 4. The model which sets the number of smokers relative to the observed social structure (ERNM) thinks there should typically be more smokers than observed, whereas ERNM-Count suggests that we actually have fewer smokers than suggested by the ERNM model (and consistent with the observed number).

In addition we introduce an alternative Goodness of Fit measure: the Kullback–Leibler (KL) divergence (relative entropy), which is a statistical distance that accounts for how one probability distribution P is different from the second probability distribution Q (MacKay, 2002). It is denoted as

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

In our analysis, we are interested in the relative entropies from ERNM-Count to ERNM and from ERNM-Count to ERGM to demonstrate the fitting of the three models. Although the sample spaces of the ERGM and the ERNM are different, we can analyze the relative entropy of the conditional probability of network statistics in the model. Using the previously generated 1000 simulations of fitted models, we compute the estimate of the KL divergence using the univariate Gaussian approximation. We also used the bias-corrected (basic) bootstrap to produce estimates of the divergences and their standard errors (using 5000 resamplings). The bias-corrected means and standard errors are shown in Tables 5 and 6. The Kullback–Leibler divergence can be interpreted as the expected log-likelihood ratio for rejecting the hypothesis that the variable was drawn from ERNM-Count based on a single network. The interpretation of Table 5 is the expected excess surprise from using ERGM as a model when the actual distribution is ERNM-Count. The results for ERGM show that the impact on some statistics is small (most GWDegree, Homophily-non-smoker terms). However, there are large impacts on most Edges, Diff-homophily-smoke terms, and the GWESP terms, indicating that the ERGM provides a poor fit for key structural terms in the model. Note that the ERGM misspecifies the number of smokers as a constant, while the standard deviation of the number of smokers is 6.9 (Fig. 4). We see in Table 6 that the lack-of-fit for the ERNM model is small, especially relative to the ERGM.

6. Discussion

Despite the wide success of exponential-family random graph models in representing complex network data, they treat the nodal and dyadic covariates as exogenous. This is not true for many realistic social processes. In this paper, we show that this treatment misspecifies the social structure of network processes. We also provide evidence that Exponential-family random network models represent a much

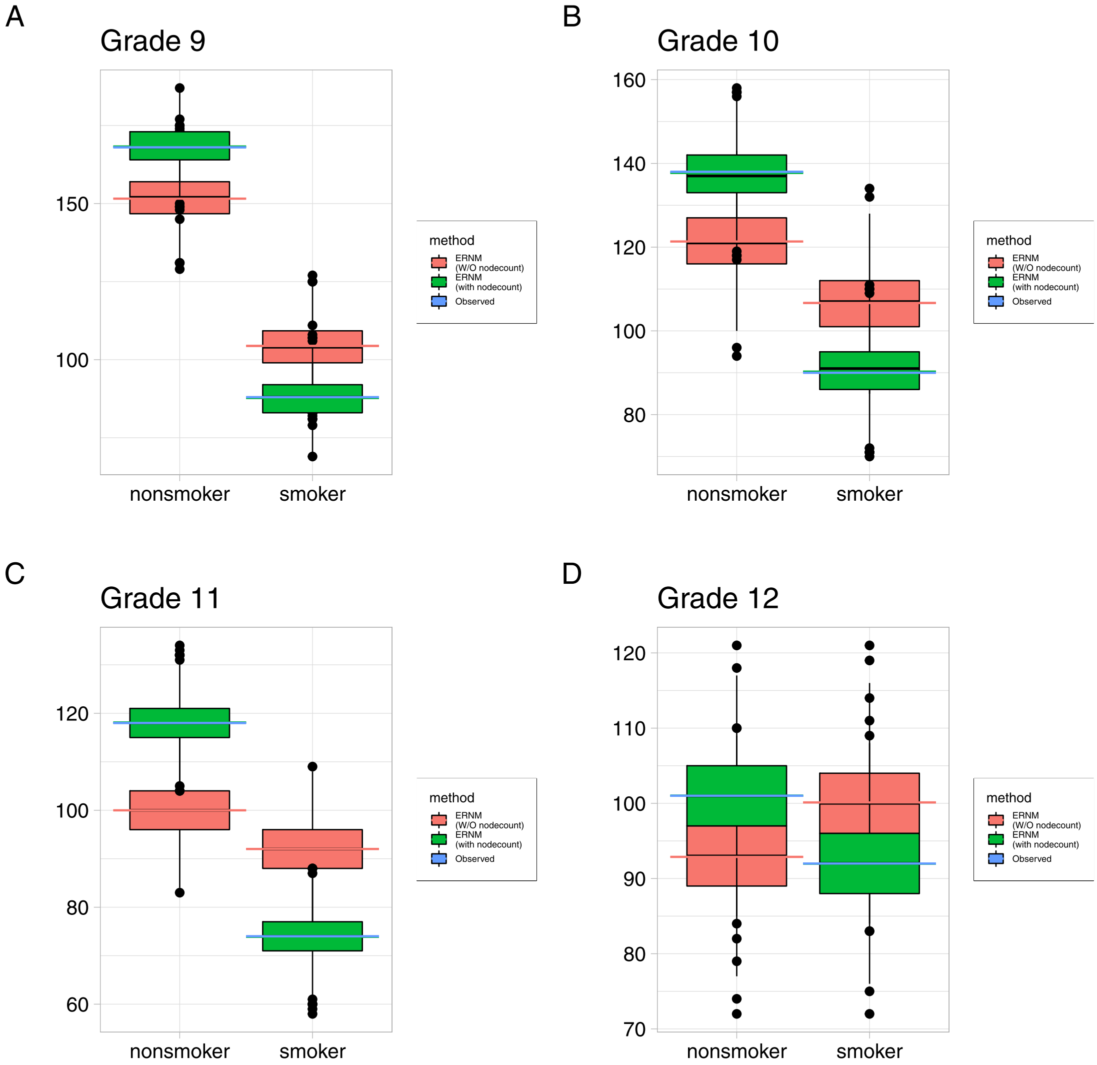


Fig. 4. Comparison between two ERNMs of count of smokers and non-smokers.

Table 5
KL divergence of ERNM-Count to ERGM.

	D_{KL}(ERNMCount ‖ ERGM)			
	Grade 9	Grade 10	Grade 11	Grade 12
Edges	0.152(0.030)	0.337(0.045)	0.008(0.005)	0.003(0.002)
ESP-0	0.005(0.003)	0.006(0.004)	0.006(0.004)	0.020(0.008)
ESP-1	0.012(0.006)	0.022(0.009)	0.004(0.003)	0.005(0.003)
ESP-2	0.003(0.002)	0.003(0.002)	0.009(0.005)	0.011(0.006)
GWESP	0.168(0.033)	0.352(0.044)	0.016(0.007)	0.013(0.006)
GWDegree	0.003(0.002)	0.013(0.007)	0.009(0.006)	0.005(0.003)
Diff-homophily-smoke	0.391(0.061)	0.837(0.086)	0.062(0.018)	2.064(0.223)
Homophily-non-smoker	0.005(0.004)	0.013(0.007)	0.008(0.005)	0.009(0.005)

The entries are the Kullback–Leibler divergences between the ERGM distribution of each variable from the corresponding ERNM-Count distribution. The values in parentheses are the standard errors of those values (computed via the bootstrap).

Table 6
KL Divergence of ERNM-Count to ERNM.

	D_{KL}(ERNMCount ‖ ERNM)			
	Grade 9	Grade 10	Grade 11	Grade 12
Edges	0.026(0.010)	0.019(0.009)	0.009(0.005)	0.003(0.002)
ESP-0	0.003(0.002)	0.004(0.003)	0.007(0.004)	0.009(0.005)
ESP-1	0.004(0.003)	0.017(0.008)	0.017(0.008)	0.012(0.006)
ESP-2	0.006(0.004)	0.004(0.003)	0.004(0.003)	0.011(0.006)
GWESP	0.022(0.010)	0.023(0.010)	0.010(0.006)	0.005(0.004)
GWDegree	0.003(0.003)	0.010(0.006)	0.003(0.002)	0.003(0.002)
Diff-homophily-smoke	0.039(0.014)	0.079(0.017)	0.055(0.014)	0.310(0.035)
Homophily-non-smoker	0.004(0.003)	0.006(0.004)	0.009(0.006)	0.006(0.004)

The entries are the Kullback–Leibler divergences between the ERNM distribution of each variable from the corresponding ERNM-Count distribution. The values in parentheses are the standard errors of those values (computed via the bootstrap).

better class of models for representing processes with nodal and dyadic covariates that are endogenous to the tie variables. ERNMs have many advantages. First, they are also in the exponential-family class of models, which have been shown to be able to represent complex social structures. Exponential-family classes of models have been extensively studied, and their properties have been explored. Because of this, the extensive knowledge and software platforms that have been developed for ERGM can, and have been, extended to ERNM.

In this paper, we compare ERGM and ERNM, with a special interest in situations where at least some of the covariates are stochastic. We use as a case-study: a friendship network among students within a school from the National Longitudinal Study of Adolescent Health. Within this network, the student's smoking behavior is likely endogenous to their friendship ties. Both ERNM and ERGM models represent the four friendship networks well, as evidenced by the goodness-of-fit and MCMC diagnostics. The coefficient estimates, and interpretations of the ERGM and ERNM are very similar after adding the node count term into ERNM (Tables 2 and 4). Although both models show significant differential homophilies on smoke, the node count term on smoke is a notable addition to the ERNM fitting. Combining the result of ERNM and ERNM-Count model, we find there are fewer smokers in the network than expected due to chance: the simulation results of the ERNM model gives a higher number of smokers compared to the observed statistics and the ERNM-Count (Fig. 4) simulated statistics; The negative coefficient of the node count variable in the ERNM-Count model also suggests this. It illustrates the importance of treating smoking status as endogenous, rather than exogenous, as in ERGM.

The impact of the endogeneity is throughout the model. As the case study shows, primary properties, such as the presence of differential homophily, can be misspecified by ERGM. Coefficients can be both under and overestimated by ERGM and standard errors can be affected both ways. In our case-study, this can be seen by comparing the coefficients and standard errors in Table 2 (ERGM) to those in Table 3 (ERNM).

The findings from the ERGM and ERNM are fundamentally different: Either the smokers have similar homophily to non-smokers (Table 4), or the number of smokers is lower than we would expect (Table 4). This is missed by the ERGM and clear (and statistically significant) in the ERNM. Note that the ERGM model is conceptually wrong in this case as it is a pure social selection model where ERNM allows both social selection and social influence.

On the one hand, based on the result of ERNM and ERGM, this is consistent with a process of social selection. Smokers tend to connect with smokers, and non-smokers tend to connect with non-smokers. If this argument holds empirically, then this would be a selection causal mechanism that leads to it. However, there might be other tie formation processes, such as different social contexts (Feld, 1981). Take a simple example: since smoker A and smoker B go to the same tobacco shop, they meet there a lot of times and become friends. This cannot be seen as a causal relationship of the homophily, rather, it is the social context that leads to the formation of friendship. Hence, we can conclude an association between nodal attributes and social networks instead of causality. On the other hand, AddHealth networks can be explained by the process of social influence. Given the result of ERNM-Count, smokers who are connected with non-smokers may choose to quit smoking and become non-smokers. If this holds, then this would be an influence causal mechanism that leads to it, which is a mechanism that only involves influence. Steglich et al. (2010) has tested the existence of such a mechanism. However, it may also be possible that smokers quit smoking because they want to make connections with non-smokers. Hence, the tie formation is a homophily or social selection process. Although we cannot disentangle the social selection and influence mechanisms, based on our conceptual knowledge, we tend to believe that the social influence process is less credible under this context, and it is much more likely that it is the selection mechanism that dominates.

The availability of powerful user-friendly open-source software allows broad accessibility and use of both ERGM and ERNM (Krivitsky et al., 2003–2020; Fellows, 2014). The analysis in this paper supports the notion that ERNM is preferred when networks have stochastic covariates.

Finally, we note the ERNM provides a way to specify the complex dependency structures that would empower autologistic actor attribute models (ALAAM) (Robins et al., 2001b). This connection to ALAAM will be the topic of future investigation.

Acknowledgments

The authors gratefully acknowledge support from the National Science Foundation grants IIS-1546259, CCF-2200197, and from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, population research infrastructure grants P2C-HD041041 and P2C-HD041022 and training grant T32-HD007545.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.socnet.2023.07.003. In this supplementary, we provide some derivation results for Section 2.3. We also show the goodness-of-fit plots and MCMC diagnostics results for both the ERGM and ERNM fits to the adolescent health network data.

References

Almquist, Z.W., Butts, C.T., 2014. Logistic network regression for scalable analysis of networks with joint edge/vertex dynamics. Sociol. Methodol. 44 (1), 273–321, URL: https://doi.org/10.1177/0081175013520159. PMID: 26120218.

Clark, D.A., Handcock, M.S., 2022. An approach to causal inference over stochastic networks. http://dx.doi.org/10.48550/arXiv.2106.14145.

Erickson, B.H., 1988. The relational basis of attitudes. In: Wellman, B., Berkowitz, S.D. (Eds.), Social Structures: A Network Approach. Cambridge University Press, Cambridge, pp. 99–121.

Feld, S.L., 1981. The focused organization of social ties. Am. J. Sociol. 86 (5).

Fellows, I.E., 2012. Exponential Family Random Network Models (Ph.D. thesis). University of California, Los Angeles, Advisor: Mark S. Handcock.

Fellows, I.E., 2014. Ernm: Exponential-family random network models. URL: https://github.com/fellstat/ernm. R package version 1.1.

Fellows, I., Handcock, M.S., 2012. Exponential-family random network models. http://dx.doi.org/10.48550/ARXIV.1208.0121, URL: https://arxiv.org/abs/1208.0121.

Fienberg, S.E., Wasserman, S.S., 1981. Categorical data analysis of single sociometric relations. Sociol. Methodol. 12, 156–192.

Fosdick, B.K., Hoff, P.D., 2015. Testing and modeling dependencies between a network and nodal attributes. J. Amer. Statist. Assoc. 110 (511), 1047–1056, URL: https://doi.org/10.1080/01621459.2015.1008697.

Frank, O., Strauss, D., 1986. Markov graphs. J. Amer. Statist. Assoc. 81 (395), 832–842, URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478342.

Friemel, T.N., 2015. Influence versus selection: A network perspective on opinion leadership. Int. J. Commun. 9, 1002–1022, URL: https://ijoc.org/index.php/ijoc/article/view/2806.

Handcock, M.S., 2003a. Assessing Degeneracy in Statistical Models of Social Networks. Working paper #39, Center for Statistics and the Social Sciences, University of Washington, URL: https://csss.uw.edu/Papers/wp39.pdf.

Handcock, M.S., 2003b. Statistical models for social networks: Inference and degeneracy. In: Breiger, R., Carley, K., Pattison, P. (Eds.), Dynamic Social Network Modeling and Analysis. National Academies, Washington, DC, pp. 229–252.

Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., Krivitsky, P.N., Morris, M., 2021. ERGM: Fit, Simulate and Diagnose Exponential-Family Models for Networks. The Statnet Project, https://statnet.org. URL: https://CRAN.R-project.org/package=ergm. R package version 4.0-6406.

Harris, K., Halpern, C., Smolen, A., Haberstick, B., 2007. The national longitudinal study of adolescent health (add health) twin data. Twin Res. Hum. Genet. Off. J. Int. Soc. Twin Stud. 9, 988–997.

Harris, M.A., Orth, U., 2020. The link between self-esteem and social relationships: A meta-analysis of longitudinal studies. J. Personal. Soc. Psychol. 119 (6), 1459–1477, URL: https://doi.org/10.1037/pspp0000265.

Hoff, P.D., 2005. Bilinear mixed-effects models for dyadic data. J. Amer. Statist. Assoc. 100 (469), 286–295, URL: https://doi.org/10.1198/016214504000001015.

Hunter, D.R., Handcock, M.S., 2006. Inference in curved exponential family models for networks. J. Comput. Graph. Statist. 15 (3), 565–583, URL: https://doi.org/10.1198/106186006X133069.

Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M., 2008. Ergm: A package to fit, simulate and diagnose exponential-family models for networks. J. Stat. Softw. 24 (3), 1–29.

Krivitsky, P.N., Handcock, M.S., Hunter, D.R., Butts, C.T., Klumb, C., Goodreau, S.M., Morris, M., 2003–2020. Statnet: Software Tools for the Statistical Modeling of Network Data. Statnet Development Team, URL: http://statnet.org.

Leary, M.R., 2023/06/17. Handbook of Theories of Social Psychology: Volume 2. SAGE Publications Ltd, London; London, pp. 141–159, URL: https://sk.sagepub.com/reference/hdbk_socialpsychtheories2.

Leenders, R., 1997. Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. Evol. Soc. Netw. 1, 165–184.

Lusher, D., Koskinen, J., Robins, G., 2013. Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications. Cambridge University Press.

MacKay, D.J.C., 2002. Information Theory, Inference and Learning Algorithms. Cambridge University Press, USA.

Morris, M., Handcock, M.S., Hunter, D.R., 2008. Specification of exponential-family random graph models: Terms and computational aspects. J. Stat. Softw. 24 (4), URL: http://www.jstatsoft.org/v24/i04/.

R Development Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0, URL: http://www.R-project.org/.

Robins, G., Elliott, P., Pattison, P., 2001a. Network models for social selection processes. Social Networks 23, 1–30.

Robins, G., Pattison, P., Elliott, P., 2001b. Network models for social influence processes. Psychometrika 66 (2), 161–189.

Schweinberger, M., 2011. Instability, sensitivity, and degeneracy of discrete exponential families. J. Amer. Statist. Assoc. 106 (496), 1361–1370, URL: https://doi.org/10.1198/jasa.2011.tm10747. PMID: 22844170.

Schweinberger, M., Krivitsky, P.N., Butts, C.T., Stewart, J.R., 2020. Exponential-family models of random graphs: Inference in finite, super and infinite population scenarios. Statist. Sci. 35 (4), 627–662, URL: https://doi.org/10.1214/19-STS743.

Steglich, C., Snijders, T.A., Pearson, M., 2010. Dynamic networks and behavior: Separating selection from influence. Sociol. Methodol. 40 (1), 329–393.

Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. Psychometrika 61 (3), 401–425.

Weng, H., 2020. A Social Interaction Model with Endogenous Network Formation (Ph.D. thesis). University of Cincinnati, University of Cincinnati, URL: http://rave.ohiolink.edu/etdc/view?acc_num=ucin159317152899108.