On the goodness-of-fit of an exponential random graph model to a social network sample

J.W.G. Simons

Utrecht University

On the goodness-of-fit of an exponential random graph model to a social network sample

Introduction

A principal objective of social network research is to establish knowledge on the antecedents of social network structures. Such knowledge is typically induced by deriving hypotheses from theory and evaluating whether these hold with respect to a social network sample. Hypotheses can be represented as parameters in a statistical social network model. Network models can be used to examine the consistency of a set of hypotheses by emulating an observed network structure within a theoretically delimited hypothetical space (Cox, 2006). The degree to which a network model is able to imitate such a structure is defined as its goodness-of-fit (GOF). External validity describes the degree to which network model results can be applied to the population from which it was drawn (Crano, Brewer, & Lac, 2014). External validity stipulates that the results of a network model will not generalize to the population if the network sample to which it was applied does not sufficiently reflect the characteristics of that population. Study results are defined as being externally valid if the network model is able to emulate the structure of the network sample and if that sample captures the characteristics of the network population. Externally valid results consequently express the efficacy of a theory for explaining a social network structure in the population.

Multiple methods have been formulated for estimating network model parameters and assessing the goodness-of-fit (GOF) of a network model with respect to a single observed social network structure. No rigorous method however exists for establishing valid results with respect to a sample of observed social network structures, which is the first condition for external validity. In the extant literature model parameters over a network sample are typically obtained by way of a meta-analysis. This procedure averages over the parameters of a set of network models with an identical specification, where this identical model has been applied to each network in the sample. In the more recent literature, GOF is then used to quantify the degree to which this network model fits on the level of each individual network and by extension that of the network sample.

Model modifications are made when the fit of the network model is found to be unsatisfactory. After any such modifications a network model over the sample is analogously obtained by averaging over the parameters of the resulting set of network models. A drawback of using a meta-analysis to obtain a general network model is that the procedure assumes parameters to be comparable across sample networks. If this assumption is violated, the parameter estimates of the general network model are likely inaccurate, which jeopardizes external validity. Even in instances where GOF is used to offset potential model misfit, it remains unclear whether consequent model averaging results in a valid network model.

Examples of studies where a meta-analysis is used to obtain a social network model over a sample of social networks are Lubbers and Snijders (2007) and Rambaran, Dijkstra, and Veenstra (2020). Lubbers and Snijders (2007) utilize a meta-analysis to obtain an exponential random graph model (ERGM) over a sample of 102 student networks in 57 junior high school classes. ERGMs are statistical social network models for cross-sectional social network structures (Lusher, Koskinen, & Robins, 2013). Since goodness-of-fit (GOF) measures were not yet developed at the time, Lubbers and Snijders (2007) do not diagnose potential misfit of the ERGM to the sample. Furthermore, due to a lack of knowledge on the conditions under which a meta-analysis induces a valid ERGM with respect to a network sample, it is unclear whether the ERGM obtained by Lubbers and Snijders (2007) provides a valid description of the student network sample. Rambaran et al. (2020) utilize a meta-analysis to obtain a stochastic actor-oriented model (SAOM) with respect to a sample of 19 classrooms of 481 students over two observation periods. SAOMs are longitudinal statistical social network models for social network structure (Snijders, 2001). Rambaran et al. (2020) start by fitting a SAOM with the same specification to each classroom combined over the two observation periods. They consequently inspect the fit of each separate SAOM. A good fit is obtained for 14 of the 19 classrooms (Rambaran et al., 2020). For two of these 14 classrooms additional effects needed to be added to obtain a good fit (Rambaran et al., 2020).

The remaining five classrooms did not have a good fit (Rambaran et al., 2020). These were nonetheless included in the final model because sensitivity analysis indicated that leaving them out did not substantially alter the results (Rambaran et al., 2020). Rambaran et al. (2020) then use a meta-analysis with respect to each classroom specific SAOM to obtain a SAOM over the sample of classrooms. As stated earlier, a meta-analysis assumes parameter estimates to be comparable across networks in the sample. This assumption is potentially violated in Rambaran et al. (2020) because the parameter estimates of the meta-analysed SAOM are a function of parameters from SAOMs with different effect specifications and varying degrees of fit. Put differently, the separate networks in the sample might differ to such a degree that the meta-analysed SAOM is unable to provide a valid description of the sample. Thus, even though Rambaran et al. (2020) use GOF to quantify and partially ameliorate misfit, it remains unclear whether doing so resulted in a SAOM which provides a valid description of the classroom sample. Other examples of articles where a a meta-analysis is used to obtain a general network model over a network sample and where similar issues abound are Van Rossem and Vlegels (2009), Rambaran, van Duijn, Dijkstra, and Veenstra (2019), Wittek, Kroneberg, and Lämmermann (2020), McDonald and Benton (2017), McKay, Grygiel, and Karwowski (2017), and Daniel, Santos, Peceguina, and Vaughn (2013).

Based on the presented line of reasoning, the objective of this study becomes to identify the conditions under which a meta-analysis can be used to average over a set of exponential random graph models (ERGM) applied to a sample of completely observed networks. It seeks to do so by investigating the conditions under which a subset of the currently available goodness-of-fit (GOF) measures are able to identify sub-par GOF on the level of the network sample. The ERGM is chosen as the social network model of interest because it is arguably the most dominant model class for modelling cross-sectional network data. Note however that the relevance of the topic is not limited to the cross-sectional modelling framework. On the contrary, the outlined problem is also relevant to longitudinal network models such as stochastic actor-oriented models (SAOMs) (Snijders, 2001), as apparent by its use in Rambaran et al. (2020).

The following research question is formulated: How can a valid exponential random graph model (ERGM) representation of a network sample be obtained? More specifically: How can goodness-of-fit (GOF) be used to ascertain whether ERGM parameter estimates can validly be combined to represent a general ERGM? The remainder of the paper is made up of five sections. First, an overview is provided of the ERGM and two methods for evaluating GOF on the level of the network and the network sample. Two factors which are hypothesized to influence ERGM parameter precision and associated GOF are then presented and discussed. Second, the analytical strategy and the data that will be used to evaluate the effects of these two factors is discussed. In short, an empirically informed simulation study will be used. The results of this simulation study are consequently reassessed in an empirical setting. The results, conclusion, and discussion sections respectively present, interpret, and discuss the outcomes of this analysis.

Theory

The exponential random graph model

The exponential random graph model on the network level.

Exponential random graph models (ERGMs) are statistical models which can be used to explicate the antecedents of cross-sectional social network structures (Lusher et al., 2013). In short, the objective of the ERGM is to provide a model for the formation of network structure (Lusher et al., 2013). To further elaborate on this aim, the concept of network edge-variables is introduced. Unless indicated otherwise, the information presented in the remainder of this section has been sourced from the book by Lusher et al. (2013). Assume a fixed and predetermined node set n, represented as $N = \{1, ..., n\}$ with $i \in N$. Let J represent the set of possible vertices excluding self-loops for the node set n: $N, J = \{(i, j) : i, j \in N, i \neq j\}$. For any observed network x_{obs}, the set of vertices E realized in x_{obs} is a random subset of J. Thus, for any element $(i, j) \in J$, a random variable X_{ij} can be defined such that $X_{ij} = 1$ if $(i, j) \in E$ and $X_{ij} = 0$ if $(i, j) \notin E$.

Denote this random variable X_{ij} as an edge-variable which can be represented by a stochastic adjacency matrix $X = [X_{ij}]$. The elements of the stochastic matrix X indicate whether pairs of vertices are adjacent or not. Finally denote the space of all possible stochastic adjacency matrices by $\mathbf{X}$, with a realization of X within $\mathbf{X}$ being denoted by $x = [x_{ij}]$. Note that strictly speaking, this definition of a network edge-variable applies to undirected networks, although it will similarly apply to directed networks given some minor notational adjustments.

Assume that an edge between any two nodes X_{ij} can be explained either by attributes of nodes or edge patterns between nodes. Refer to the first type of predictor as an exogenous covariate and the second type as an endogenous covariate. The effect of an exogenous covariate on X_{ij} can then be understood in analogy to a logistic regression model. Start by defining a set of p exogenous covariates and respectively denote these as $w_{ij,1}, w_{ij,2}, ..., w_{ij,p}$. Interpret the elements of p as a function of an attribute for two respective actors i and j. An examples of such an attribute function is whether i and j share the same ethnic background. Another example is the difference between i and j on the attribute age. Since X_{ij} is a dichotomous response variable, a logistic regression model can be used to obtain estimates for the unknown parameter set $\theta_1, \theta_2, ..., \theta_p$ such that these predict X_{ij} as a function of p. Writing the logistic regression function in terms of the logit:

$$\text{logit}\Pr(X_{ij} = 1 \mid \theta) = \log\frac{\Pr(X_{ij} = 1 \mid \theta)}{\Pr(X_{ij} = 0 \mid \theta)} = \theta_1 w_{ij,1} + \theta_2 w_{ij,2} + ... + \theta_p w_{ij,p} \qquad (1)$$

so that the probability of an edge X_{ij} is given by the set of predictors p weighted by a respective set of parameters θ. Positive parameter values in (1) indicate an increased probability of an edge, where negative parameter values indicate a decrease in that probability. The difference in the log-odds for two pairs of nodes (i, j) and (h, m) can consequently be defined as:

$$\frac{\text{logit}\Pr(X_{ij} = 1 \mid \theta)}{\text{logit}\Pr(X_{hm} = 1 \mid \theta)} = \theta_1(w_{ij,1} - w_{hm,1}) + \theta_2(w_{ij,2} - w_{hm,2}) + ... + \theta_p(w_{ij,p} - w_{hm,p}). \quad (2)$$

Assume for illustration purposes that the pairs (i, j) and (h, m) differ only on the node attribute of the exogenous covariate $w_{ij,1}$. In that case equation (2) reduces to θ_1 which is equivalent to the odds-ratio. The larger the value of θ_1, the greater the probability of the presence of an edge between the node pairs which are similar as opposed to non-similar on $w_{ij,1}$, ceteris paribus.

Define endogenous or structural covariates as the second type of predictor of the edge between any two nodes X_{ij}. Endogenous covariates represent counts of network configurations. A network configuration can be understood as a local sub-graph within the larger graph in which it is embedded. Examples of network configurations are counts of the triad census or the number of reciprocated vertices within a larger social network structure. The analogy with logistic regression breaks down when endogenous covariates are introduced because these variables induce dependencies among edge-variables. The issue of dependence is not extensively discussed here, but the observation is made that ignoring dependence among observations negatively affects the precision of conventional statistical models. As such, in order to make inferences about network structures a model needs to be formulated which is able to incorporate dependencies among edge-variables. Exponential random graph models (ERGMs) incorporate possible dependencies amongst edge-variables by predicting the probability of an edge conditional on what is observed in the rest of the network. More formally, the ERGM models each edge-variable X_{ij} conditional on the other vertices in the network X_{-ij}. Write this conditional probability as $\Pr(X_{ij} = 1 \mid X_{-ij} = x_{-ij}, \theta)$. Concentrating on the exogenous predictors, the conditional logit becomes :

$$\log \frac{\Pr(X_{ij} = 1 \mid X_{-ij} = x_{-ij}, \theta)}{\Pr(X_{ij} = 0 \mid X_{-ij} = x_{-ij}, \theta)} = \theta_1 \delta^{+}_{ij,1}(x) + \theta_2 \delta^{+}_{ij,2}(x) + ... + \theta_p \delta^{+}_{ij,p}(x) \tag{3}$$

The functions $\delta^{+}_{ij,k}(x)$ in (3) are defined as change statistics for the kth configuration. These change statistics represent the change in transitioning from a graph for which $X_{-ij} = x_{-ij}$ and $X_{ij} = 0$ to a graph for which $X_{-ij} = x_{-ij}$ and $X_{ij} = 1$. In layman's terms, adding any edge X_{ij} to a network graph results in a possible change in any of the network configurations p.

This change is captured by the relevant change statistic, which when weighted by a respective parameter results in a change in the probability for $X_{ij} = 1$. As was the case for the exogenous covariates, whether an increase or decrease in the configuration count results in a respective increase or decrease in the probability for $X_{ij} = 1$ depends on the sign of the parameter estimate.

Having outlined the key fundamentals of the exponential random graph model (ERGM), write an equivalent form of the exponential random graph model (ERGM) such that it gives a probability expression for all edge-variables simultaneously. Define this as the joint form of the ERGM :

$$\Pr(X = x \mid \theta) \equiv P_\theta(x) = \frac{1}{\kappa(\theta)} \exp\{\theta_1 z_1(x) + \theta_2 z_2(x) + ... + \theta_p z_p(x)\}. \tag{4}$$

The functions $z_k(x)$ are counts of configurations in the graph x, such that the corresponding change statistic for $z_k(x)$ is $\delta^+_{ij,k}(x) = z_k(\Delta^+_{ij}x) - z_k(\Delta^-_{ij}x)$, where $\Delta^+_{ij}x(\Delta^-_{ij}x)$ denotes a matrix x for which x_{ij} is constrained to be equal to one (zero). In analogy to what was outlined earlier, the parameters weight the relative importance of their respective configurations. The normalizing constant $\kappa(\theta) = \sum_{y \in X} \exp\{\theta_1 z_1(y) + \theta_2 z_2(y) + ... + \theta_p z_p(y)\}$ finally ensures that the sum of $P_\theta(x)$ over all graphs is equal to one. Put differently, its ensures hat it is a proper probability mass function.

Note that equation (4) describes a probability distribution for all graphs with n nodes. Formally, the exponential random graph model (ERGM) obtains the space of all possible stochastic adjacency matrices $\mathbf{X}$ for a set of nodes n, one of which is the empirically observed graph x_{obs}. This distribution of graphs over $\mathbf{X}$ implies a parallel distribution of network statistics. In general, the inferential goal of the ERGM is to center this distribution of statistics over those of the observed network. Put differently, its goal is to maximize the probability of the observed graph. Define the distribution as centered on the observed values when the expected value of the networks statistics $E_\theta(z(X))$ is equal to the network statistics of the observed network x_{obs}: $E_\theta(z(X)) = z(x_{obs})$). Equivalently, write this as the moment equation $E_\theta(z(X)) - z(x_{obs}) = 0$. Solving the moment equation for θ consequently provides the parameter values that provide maximal support to the observed network x_{obs}.

This resulting set of parameter values are also known as the maximum likelihood estimates (MLEs).

Note that the moment equation does not have an analytical solution for most exponential random graph models. If pairs of dyads in the network graph can be assumed to be independent, the maximum pseudo-likelihood estimator (MPLE) of an ERGM has been shown to equal its true likelihood (Hunter, Handcock, Butts, Goodreau, & Morris, 2008). This implies that in such cases the MLE can be found by using logistic regression as a computational vehicle (Hunter, Handcock, et al., 2008). MPLE has however been criticized because its statistical properties are not well understood when the independence assumption is violated (Van Duijn, Gile, & Handcock, 2009). In most instances, Markov chain Monte Carlo (MCMC) methods are therefore required to obtain a solution to the moment equation. The general principle behind MCMC is to take a provisional parameter vector θ and use a Metropolis sampler to obtain a sample of graphs $x^{(1)}, x^{(2)}, ..., x^{(M)}$. In short, a Metropolis sampler generates a sequence of m graphs by randomly nominating a pair of nodes i and j at each iteration m. It then proposes to update the corresponding vertex-variable such that the proposal graph x^* at iteration m is equal to $x^{(m-1)}$, except that $x_{ij}^{(m-1)}$ is set to $1 - x_{ij}^{(m-1)}$. Given the target distribution $P_\theta(x)$, the Metropolis sampler accepts the proposal with probability

$$\min\left\{1, \frac{P_\theta(x^*)}{P_\theta(x^{(m-1)})}\right\}. \tag{5}$$

If the proposal is accepted, $x^{(m)} = x^*$; otherwise, $x^{(m)} = x^{(m-1)}$. This procedure is repeated until a sample of size m which approximates the graph space $\mathbf{X}$ of $P_\theta(x)$ is obtained. Given this sample, calculate the sample equivalent $\bar{z_\theta} = \frac{1}{M}(z(x^1) + z(x^2) + ... + z(x^M))$ of $E_\theta(z(X))$, and evaluate the moment-equation $\bar{z}_\theta - z(x_{obs}) = 0$. If $\bar{z}_\theta - z(x_{obs}) \neq 0$, choose another value θ and repeat the process until a θ is found for which $\bar{z}_\theta - z(x_{obs}) = 0$, the MLE.

A number of MCMC algorithms exist for obtaining the MLE. The method used here was originally proposed by Geyer and Thompson (1992).

This approach solves the moment equation by representing the graph space $\mathbf{X}$ through a fixed sample of graphs for a provisional value of the parameter vector θ. As outlined previously, such a fixed sample is obtained from the model $P_{\tilde{\theta}}(x)$ with a Metropolis sampler. The method then applies Fisher scoring to find θ such that $\bar{z}_\theta - z(x_{obs}) = 0$. Since the sample of graphs is an approximation of $\mathbf{X}$, a weighted average of the networks statistics is used to determine $\bar{z_\theta}$. The sample average $\bar{f_\theta} = w^{(1)}f(x^{(1)}) + w^{(2)}f(x^{(2)}) + ... + w^{(M)}f(x^{(M)})$, of a function $f(x)$, with weights

$$w^m = \frac{e^{(\theta_1-\tilde{\theta_1})z_1(x^{(m)})+(\theta_2-\tilde{\theta_2})z_2(x^{(m)})+...+(\theta_p-\tilde{\theta_p})z_p(x^{(m)})}}{\sum_{k=1}^{M} e^{(\theta_1-\tilde{\theta_1})z_1(x^{(m)})+(\theta_2-\tilde{\theta_2})z_2(x^{(m)})+...+(\theta_p-\tilde{\theta_p})z_p(x^{(m)})}} \tag{6}$$

is an asymptotic approximation to the true expected value $E_\theta(f(X))$ when $\theta \approx \tilde{\theta}$. To solve the moment equation, a sequence of parameters $\theta^{(0)} = \tilde{\theta}, \theta^{(1)}, \theta^{(2)}, ..., \theta^{(g)}$ is generated using Fisher scoring. The updating step of $\theta^{(g)}$ in the Fisher scoring algorithm is of the form:

$$\theta^{(g)} = \theta^{(g-1)} - \frac{1}{D(\theta^{(g-1)})} \left\{ \sum_{m=1}^{M} w^m z(x^m) - z(x_{obs}) \right\} \tag{7}$$

and iterates until that $\theta^{(g-1)}$ for which $\sum_{m=1}^{M} w^m z(x^m) = E_\theta(z(X))$ such that $E_\theta(z(X)) - z(x_{obs}) = 0$, which is the MLE. The term $D(\theta)$ in the Fisher scoring algorithm is a scaling matrix for the difference between observed values and the expected values of the network statistics. In short, the network statistics need to be scaled because they differ in their sensitivity to changes in the parameter vector θ. The matrix $D(\theta)$ is set to the weighted sample covariance $\sum_m w^{(m)} z(x^{(m)}) z(x^{(m)})^T - \left[\sum_m w^{(m)} z(x^{(m)})\right] \left[\sum_m w^{(m)} z(x^{(m)})\right]^T$.

The exponential random graph model on the sample level. Having introduced the exponential random graph model (ERGM) for a single observed social network, let us now define the ERGM at the level of the network sample. One method for obtaining an ERGM over a network sample is a meta-analysis (Lubbers, 2003). The idea behind a meta-analysis is to use a weighted average to summarize over the results of multiple statistical models, here ERGMs (Lubbers, 2003). To illustrate the method, let us imagine a scenario where we want to apply an ERGM to a network sample of size $\{2, ..., m\}$.

Initially, an ERGM is fitted to each network in the sample, resulting in a set of ERGMs with size equal to m, each with its own set of parameters and standard errors. In order to summarize these separate ERGMs, a meta-analytic approach as outlined in Lubbers (2003) can be applied. This method splits the coefficients of the separate ERGMs into an average coefficient and a network-dependent deviation (Lubbers, 2003). Let θ denote a parameter included in each ERGM. The meta-level regression equation for this parameter can then be written as:

$$\hat{\theta} = \mu_\theta + U_m + E_m \tag{8}$$

where $\hat{\theta}$ denotes the estimated parameter value for network m, μ_θ is the average parameter coefficient in the population, U_m is the true deviation of network m, which has a mean of 0 and a variance σ^2_θ, and E_m denotes the estimation error associated with the true parameter value θ_m. The variance of E_m is assumed to be known and can be obtained as the squared standard error of θ_m (Lubbers, 2003). The meta-analysis consequently seeks to differentiate between the true and error variance, in order to obtain more precise estimators for μ_θ and σ^2_θ. Put differently, by differentiating between true and error variances the method more accurately weights μ_θ based on differences in standard errors between networks. Lubbers (2003) suggests using iteratively re-weighted least squares (IRLS) for estimating (8), such that networks with large standard errors have less influence on μ_θ than networks with small standard errors. This results in a weighted vector of parameter means, each with an associated standard error. These results can then be interpreted as representing a general ERGM description of the network sample. The reader is referred to Lubbers (2003) for a complete overview and practical example of the meta-analytic procedure.

Goodness-of-fit for the exponential random graph model

Having thus defined the form and inferential procedure of the exponential random graph model (ERGM) on the level of the single network and the network sample, let us now formulate an approach for evaluating the fit of the ERGM on both these levels.

Goodness-of-fit (GOF) indices are typically used for diagnosing the fit of an ERGM (Lusher et al., 2013). On the single network level, GOF indices can be used to identify sub-optimal fit of the ERGM to the observed network. This information can consequently be used to improve the fit of the ERGM through parameter modification. On the level of the network sample, GOF indices can be used to evaluate the degree to which the ERGM fits to the sample as a whole. Here, GOF can thus be used to ascertain whether conclusions about ERGM parameters can reasonably be extended to the population. Note that there exists an interplay between the GOF indices on the two levels. Sub-optimal GOF on the network level is generally an indicator of sub-optimal GOF on the sample level, and vice versa the same holds. Let us start by defining the GOF of an ERGM with respect to a single, completely observed social network.

Goodness-of-fit of the exponential random graph model on the level of the network. In general, the goodness-of-fit (GOF) of a statistical model can be defined as the fit of that model to a set of observations. It is quantified by way of an index which summarizes the discrepancy between the observed values and the values expected under the statistical model. In the context of the exponential random graph model (ERGM), GOF can thus be understood as the fit of an ERGM to an observed social network structure. This fit is quantified by an index which summarizes the difference between what is observed in the network and expected based on the ERGM. An index expressing GOF can consequently be formulated in a number of different ways. Examples of such indices are auxiliary statistics (Hunter, Goodreau, & Handcock, 2008), information criteria such as Aikake's information criterion (AIC) (Akaike, 1974) and Schwarz's information criterion (SIC) (Schwarz, 1978), approximate Bayesian GOF (Lusher et al., 2013) and outlier analysis (Koskinen, Wang, Robins, & Pattison, 2018). Here, the auxiliary statistics and information criteria indices are used to quantify the GOF of an ERGM to an observed social network. The auxiliary statistics index is chosen because it is the standard method in most of the social network literature for diagnosing the GOF of a social network model.

Furthermore, since the Monte Carlo scheme outlined earlier produces an approximate maximum likelihood estimate, the AIC and SIC can be used to select that ERGM which minimizes the value for these two respective indices. Let us proceed by respectively defining these two respective GOF indices.

Auxiliary statistics as a goodness-of-fit index. The auxiliary statistics goodness-of-fit index was introduced by Hunter, Goodreau, and Handcock (2008). It quantifies the GOF of an exponential random graph model (ERGM) to an observed social network by simulating from the ERGM to investigate graph configurations which were not modeled explicitly. Put differently, the index summarizes the discrepancy between what is observed in the social network graph and what is expected under the ERGM for a graph configuration not included in the model. For example, is including an edge parameter in the ERGM sufficient for also explaining the degree distribution in the observed graph? Note that Hunter, Goodreau, and Handcock (2008) distinguish between the fit of graph configurations explicitly modeled by the ERGM and graph configurations not explicitly modeled by the ERGM. The first are indicators of the convergence of the MCMC algorithm where the second are indicators of the GOF of the ERGM to the observed network.

The auxiliary statistics goodness-of-fit (GOF) index is calculated by simulating a distribution of graphs from an exponential random graph (ERGM) for the auxiliary graph statistic(s) of interest and determining the position of the observed auxiliary statistic in that distribution. Note that the choice of auxiliary graph statistic(s) determines which structural aspects of the networks are important for assessing the GOF of the ERGM. Hunter, Goodreau, and Handcock (2008) propose using the degree, the triad census, and the geodesic distance in most instances. Let us observe that if the fitted ERGM is sufficient for explaining a particular graph configuration in the observed social network, then the position of the associated observed auxiliary statistic $S_k(x_{obs})$ should not be extreme in the simulated distribution of graphs.

One can thus quantify GOF with regards to an observed auxiliary statistic by contextualizing its position in the distribution of simulated graphs by way of a t-statistic: $t = [S_k(x_{obs}) - \bar{S_k}]$ / $SD(S_k(x))$. Here, $\bar{S_k}$ and $SD(S_k(x))$ are the mean and standard deviation of the distribution of the statistic over the simulated sample of graphs. A p-value can consequently be obtained as the observed value of the statistic with respect to the distribution of simulated networks as: $p_{S_k(x)} = \Pr(S_k(x) \leq S_k(x_{obs})$. If the resulting p-value is smaller than a particular α then the observed statistic is far from what is expected under the ERGM. A value for α of 0.05 is typically taken as being extreme. This would then be indicative of a poor fit of the ERGM to the observed social network structure for that particular graph configuration.

Information criteria as a goodness-of-fit index. In general, the objective of an information criterion is to select that statistical model which best approximates the underlying, unobserved process which generated the observed data (Wit, Heuvel, & Romeijn, 2012). Formally, this is achieved by minimizing the Kullback–Leibler divergence with respect to the observed data (Wit et al., 2012). A more intuitive definition is that information criteria seek to maximize the fit of the model to the data, while simultaneously minimizing the complexity of that model that is specified (Wit et al., 2012). This so-called bias-variance trade-off prescribes that one wants to avoid including too many parameters in an exponential random graph model (ERGM) because this will cause the out of sample prediction error to become unacceptably large. On the other hand it is also desirable to avoid including too few parameters in the ERGM because this will heavily bias statistical inference by the ERGM. As such, information criteria define goodness-of-fit (GOF) as a quantity which seeks to balance complexity and fit, where lower values indicate a model which is better able to strike this balance. Note that information criteria are a fundamentally relative quantity. Put differently, they should be interpreted as expressing the relative efficacy of two or more models for striking a balance between complexity and fit.

As stated earlier, the information criteria formulated by Akaike (1974) - Aikake's information criterion (AIC) - and (Schwarz, 1978) - Schwarz's information criterion (SIC) - will here be used to quantify GOF. The AIC and SIC can be defined respectively as

$$\text{AIC(M)} = -2\log P_{\hat{\theta}} + 2p, \tag{9}$$

$$\text{SIC(M)} = -2\log P_{\hat{\theta}} + \log S. \tag{10}$$

Here, M refers to a particular ERGM, $P_{\hat{\theta}}$ refers to the maximized likelihood under M, p refers to the number of parameters in the ERGM, and S refers to the size of the sample (Wit et al., 2012). For a derivation of these two formulae with respect to the Kullback–Leibler divergence the reader is referred to Wit et al. (2012).

Note that there are a number of disadvantages associated with using information criteria over auxiliary statistics as goodness-of-fit (GOF) indices for exponential random graph models (ERGMs). The first is that information criteria assume an independent and identically distributed (IDD) sample (Wit et al., 2012). As outlined earlier, independence assumptions generally do not hold for social network data. Moreover, as was discussed earlier, it is not possible to evaluate the likelihood function directly for most ERGMs. Since a likelihood value is required for calculating information criteria, quantifying GOF with the AIC or SIC is approximate at best (Wit et al., 2012). The auxiliary statistics approach to GOF is finally more more informative than the AIC and SIC results in the sense that it shows which graph configurations are fit well and which are not (Wit et al., 2012). For these reasons, auxiliary statistics are generally favoured over information criteria as an index for quantifying GOF.

Note however that as a goodness-of-fit (GOF) index, information criteria have one distinct advantage over auxiliary statistics: exponential random graph model (ERGM) comparison. More specifically, information criteria enable comparison of the fit of differently parameterized ERGMs to a particular observed network. The auxiliary statistics GOF index does not have this feature, which makes the inclusion of information criteria as a GOF index relevant.

Goodness-of-fit of the exponential random graph model on the level of the network sample. Having thus defined a procedure for obtaining the goodness-of-fit (GOF) of an exponential random graph model (ERGM) on the level of a single network, the next step is to quantify the GOF of a meta-analyzed ERGM with respect to a network sample. A methodology for doing so is here presented for the auxiliary statistic and information criterion GOF indices.

Auxiliary statistics as a goodness-of-fit index. The auxiliary statistic goodness-of-fit (GOF) index for an exponential random graph model (ERGM) on the level of the network sample is a multivariate realization of its definition on the level of a single network. Let us recall that the auxiliary statistics GOF index for an ERGM was determined by generating sampling distributions from that ERGM for an a-priori chosen set of auxiliary statistics. In line with Hunter, Goodreau, and Handcock (2008), the degree, the triad census, and the geodesic distance will here be used as the set of auxiliary statistics. Then, for each network in the sample the t-value associated with the position of that sample network's auxiliary statistic in the general ERGM sampling distribution is calculated. These t-values are translated into p-values which can then be used to inspect the GOF of each individual ERGM on these structural features. Given these components, one can consequently calculate a Mahalanobis distance for each auxiliary statistic over the network sample. These Mahalanobis distances can then be translated into p-values which can be used to evaluate the fit of each statistic on the level of the network sample.

Information criteria as a goodness-of-fit index. A naive methodology is secondly adopted to quantify the goodness-of-fit (GOF) of an exponential random graph model (ERGM) on the level of the network sample by way of the AIC and SIC information criteria. Given a network sample of size m and a set of fitted ERGMs of the same size, start by using the approximated likelihood to determine an AIC and SIC for each ERGM.

Average over the resulting AIC and BIC vectors to obtain mean AIC and SIC values for the meta-analyzed ERGM:

$$\mu_{\text{AIC}} = \frac{1}{m}(\sum_{i=1}^{m} \text{AIC}_i) = \frac{\text{AIC}_1 + \text{AIC}_2 + ... + \text{AIC}_m}{m},$$

$$\mu_{\text{SIC}} = \frac{1}{m}(\sum_{i=1}^{m} \text{SIC}_i) = \frac{\text{SIC}_1 + \text{SIC}_2 + ... + \text{SIC}_m}{m}.$$

Calculate the standard error as the standard deviation over the respective AIC and SIC vectors:

$$\sigma_{\text{AIC}} = \sum_{i=1}^{m} \frac{(\text{AIC}_i - \mu_{\text{AIC}})^2}{m},$$

$$\sigma_{\text{SIC}} = \sum_{i=1}^{m} \frac{(\text{SIC}_i - \mu_{\text{SIC}})^2}{m}.$$

Finally express the GOF of the meta-analyzed ERGM to the network sample as a mean AIC or SIC with an associated standard error, i.e., $\mu_{\text{AIC}} \pm \sigma_{\text{AIC}}$ and $\mu_{\text{SIC}} \pm \sigma_{\text{SIC}}$, respectively.

Note that the auxiliary statistics goodness-of-fit (GOF) index should be used for adding or removing parameters from the exponential random graph model (ERGM) on the single network level to increase the fit of the meta-analyzed ERGM to the sample. The information criteria GOF indices should conversely be used for identifying the meta-analyzed ERGM parameter specification which provides the best fit to the sample. For instance, given that sub-optimal GOF has been identified on the sample level, a researcher should (re-)inspect GOF on the individual network level and modify those ERGMs for which sub-optimal GOF is identified. After any such modifications have been made, GOF should then be reinspected on the sample level. This procedure should subsequently be repeated until satisfactory GOF is achieved on the sample level. Given an acceptable fit, information criteria can then be used on the sample level to identify the overall best fitting meta-analyzed ERGM. Put differently, the objective is optimize the fit of the meta-analyzed ERGM to the sample by way of the information criteria GOF indices. This can be achieved by adding or removing parameters from the ERGMs on the individual network level by way of the auxiliary statistics GOF index.

On the precision of the exponential random graph model

Having thus defined the form of the exponential random graph model (ERGM) and the goodness-of-fit (GOF) indices on the level of the individual network and the network sample, let us now return to the research question: How can a valid ERGM representation of a network sample be obtained? More specifically: How can GOF be used to ascertain whether ERGM parameter estimates can validly be combined to represent a general ERGM? Note that it has been demonstrated that a meta-analysis can be used to obtain an ERGM representation of a network sample. It has furthermore been demonstrated that, given that the sample sufficiently captures the salient characteristics in the population, GOF indices can be used to diagnose and optimize the fit of the ERGM to the sample. Let us now sketch an ideal scenario where an ERGM over a network sample is specified with GOF as a diagnostic and optimization instrument such that its parameter estimates are identical to those in the population.

Start by assuming that a researcher has obtained a network sample which sufficiently captures the characteristics of the population from which it was drawn. Based on the outlined method, the researcher would proceed by fitting an exponential random graph model (ERGM) to each network in the sample. The researcher would then obtain a meta-analyzed ERGM over the resulting set of ERGMs which provides parameter estimates and standard errors on the level of the network sample. The researcher would consequently need to inspect the degree to which the ERGM fits on the level of the individual networks and the network sample. Ideally, the goodness-of-fit (GOF) indices function as reliable indicators for the detection of optimal or sub-optimal fit of an ERGM on both the individual network and sample level. Under the assumption that the GOF indices are in fact reliable, the researcher can utilize GOF to make modifications to the ERGM if sub-par GOF is identified. In an ideal scenario, after any such modifications are made, the final meta-analyzed ERGM would then show an optimal fit to the network sample. By extension, this would make the meta-analyzed ERGM parameter estimates identical to the population parameters.

Since the sample was an adequate representation of the population, the researcher could then establish external validity. Put differently, the researcher would be able to evaluate theoretically derived hypotheses and determine whether the effect of a certain actor attribute or network feature is salient in the real world.

Note that in practice, it is generally not known when the parameter estimates of an exponential random graph model (ERGM) are decently accurate at the sample level. This is because the population parameters are unknown for empirical data. As such, even when goodness-of-fit (GOF) is used to evaluate and improve the precision of the meta-analyzed ERGM, it is not clear whether an acceptable level of precision has been achieved. One method for making inferences about the precision of an ERGM over a network sample is a simulation study. Such an approach can provide insight into the factors which are important for determining the precision of the ERGM on the level of the network sample. As such, the objective of this paper is to identify and evaluate factors which are hypothesized to affect the precision of the parameter estimates of the meta-analyzed ERGM. In short, this study identifies two types of factors which are hypothesized to affect the precision of the ERGM model parameters on the level of the network sample: the characteristics of the sample and the complexity of the model.

Characteristics of the sample. The first factor which is hypothesized to affect the precision of the meta-analyzed exponential random graph model (ERGM) are the characteristics of the sample. Two sample characteristics, its size and its heterogeneity, are considered. The size of the sample simply refers to the number of social networks in the sample. The heterogeneity of the sample refers to the degree to which the networks in the sample show different co-variate and structural properties. Heterogeneity can alternatively be defined as the number of outlying networks in the sample.

Traditional sampling theory prescribes that larger sample sizes result in increased model precision when estimating population parameters (Lohr, 2009). The concept of sample size and its effect on model precision is much more ambiguous in the context of social network analysis (Kolaczyk & Krivitsky, 2015).

Under a more sophisticated definition of the central-limit theorem, the same principle however holds for the application of meta-analyzed exponential random graph models (ERGMs) to social network samples (Kolaczyk & Krivitsky, 2015). On this basis the parameter estimates for a meta-analyzed ERGM are hypothesized to be more precise for larger as opposed to smaller network samples.

As is the case for standard statistical models, the precision of a meta-analyzed exponential random graph model (ERGM) is reduced when it is applied to a sample which consists of many as opposed to a few outlying networks (Koskinen et al., 2018). This is due to the fact that the meta-analyzed ERGM is unable to sufficiently capture all intricacies of all the networks in the sample (Koskinen et al., 2018). In more formal terms, ERGMs are not robust (Rehnberg, 2016). On that basis, it is hypothesized that the precision of the meta-analyzed ERGM is reduced when it is applied to a sample with many as opposed to few outlying networks, ceteris paribus.

A third and final important characteristic of the sample are the sizes of the social networks in the sample. This sample characteristic can be seen as a form of heterogeneity, in that a higher variance in the respective network sizes could be be hypothesized to reduce the precision of the meta-analyzed exponential random graph model (ERGM). Although network size is generally an important feature (Kolaczyk & Krivitsky, 2015), here this characteristic is kept constant in order to limit the scope of the simulation study.

Complexity of the exponential random graph model. The second factor relates to the complexity of the meta-analyzed exponential random graph model (ERGM). This factor can be defined as the number and type of parameters that need to be included in the ERGM to provide an adequate representation of the network sample. A more complex meta-analyzed ERGM tends to provide a better fit at the cost of increased complexity. As such, the objective is to find that ERGM which provides a fit to the network sample which is as simple as possible while still being complex enough to capture the intricacies of each network in the sample.

If the configuration of the meta-analyzed ERGM is too simple or too complex, sample results will not correspond to the characteristics of the network population. Different aspects of the networks in the sample might furthermore also be more or less easy to model. Work by Van Duijn et al. (2009) showed higher parameter precision for covariate as opposed to structural network effects. As such, covariate effects are more likely to be adequately modeled by a relatively simple parameterized ERGM, where a more complex ERGM is required to capture structural effects. Based on the previous line of reasoning, it is hypothesized that the precision of the meta-analyzed ERGM decreases for an increased discrepancy between the complexity of the model and the complexity of the sample. For instance, given a highly complex sample - one with a small sample size and many outliers - a meta-analyzed ERGM with a complex specification will be required to obtain precise estimates. Conversely, for a sample with low complexity - a large sample size and few outliers - a meta-analyzed ERGM with a simple specification will be required to obtain precise estimates. The larger the discrepancy between the relative complexity of the model and the sample, the larger the imprecision of the meta-analyzed ERGM estimates. Note that no hypothesis is formulated about the interaction between the complexity of the meta-analyzed ERGM and the sample. It is furthermore hypothesized that ERGM precision is generally higher for covariate as opposed to structural network effects.

Analytical Strategy

In this section the analytical strategy for performing the simulation study is outlined. To start, the "ergm" R-package will be used for estimating exponential random graph (ERGM) parameters at the single network level (Hunter, Handcock, et al., 2008; Team, 2013). The "metafor" R-package will be used for obtaining the meta-analyzed ERGM at the level of the network sample (Viechtbauer, 2010; Team, 2013). The procedure for calculating auxiliary statistics goodness-of-fit (GOF) indices at the level of the individual network and the network sample will be hand-implemented in R.

The information criteria GOF indices are provided by the "ergm" R-package at the level of the single network (Hunter, Handcock, et al., 2008; Team, 2013). From these values the AIC and the SIC at the level of the network sample are subsequently obtained. The bias and root mean square error of approximation (RMSEA) will furthermore be used to quantify the precision of the ERGM at the level of the network sample.

The second step is to obtain social network samples to which the exponential random graph model (ERGM) and the respective goodness-of-fit (GOF) indices can be applied. A simulation approach will be used to generate social network samples. It should be noted that simulating network variability is a delicate process where networks with nonsensical configurations are quickly induced for certain ERGM parameter combinations. As such, simulated network variability should be informed by empirical network variability. Moreover, a trial and error process should be employed which identifies parameter combinations that generate empirically feasible networks. The Bernard and Killworth fraternity will be used as a basis for simulating empirically informed networks (Bernard, Killworth, & Sailer, 1980). The “ergm” R package will be used to obtain empirically informed ERGM parameter estimates and standard errors for the Bernard and Killworth fraternity network (Hunter, Handcock, et al., 2008; Team, 2013). These estimates are consequently used to generate sampling distributions from which ERGM parameter values can be drawn. In this way, empirical network variability is represented in the simulation process, where each draw from the sampling distributions can be used to simulate a network in the sample. A trial and error process is simultaneously used to evaluate whether the selected parameters generate empirically realistic network structures. An empirical sample will additionally be used to confirm whether expectations based on the simulation study also hold in practice. Currently, one empirical sample by Vermeij (2006) is available for the purposes of this study. This empirical network sample consists of 86 Dutch secondary class networks. Ethical approval has been obtained for both the simulation study and use of the Vermeij data.

Given the simulation procedure, the study will proceed by varying and assessing the effect of the hypothesized factors on the precision of the meta-analyzed exponential random graph model (ERGM). With regards to the sample size, samples consisting of 10, 20, and 50 respective networks will be generated. It is consequently examined whether the precision of the meta-analyzed ERGM increases with sample size, and whether this is consistent with the respective goodness-of-fit (GOF) indices.

The heterogeneity factor can consequently be operationalised in two ways. It can either be represented as the difference in the distribution of the covariate and the structural features of the networks in the sample, or as the difference in the effect of the parameter for these features. Here, the second approach is adopted, meaning that parameter effects are varied for the networks in the sample. More specifically, three scenario's are identified: one with little variability in the effect of the respective parameters between sample networks, one with moderate such variability, and one with high variability. It is consequently examined whether the precision of the meta-analyzed exponential random graph model (ERGM) increases with decreased heterogeneity, and whether this is consistent with the goodness-of-fit (GOF) indices.

The complexity factor is finally examined by varying the complexity of the exponential random graph model (ERGM) that is applied to the different network samples. Note that a complexity continuum has already defined by way of the network samples which vary in their size and heterogeneity. ERGMs with low, moderate, and high respective complexity are fitted to each of these different samples. As such, with regards to the heterogeneity factor, three different scenario's are identified. It is consequently examined whether the precision of the meta-analyzed ERGM increases for a decrease in the discrepancy between sample and model complexity. It is also assessed whether this is consistent with the goodness-of-fit (GOF) indices. Furthermore, both covariate and structural effects are included in each ERGM, which remain constant regardless of the complexity of the sample to which the ERGM is applied. This enables examining whether covariate parameters estimates are more precise than their structural counterparts.

Ultimately, for the scenario's as they have been defined here, a total of 27 scenario's can be identified. Under these specifications, the effects for all the hypothesized factors can be evaluated.

Complexity	Heterogeneity	Sample size		
		10	20	50
Low	Low	Scenario 1	Scenario 2	Scenario 3
	Moderate	Scenario 4	Scenario 5	Scenario 6
	High	Scenario 7	Scenario 8	Scenario 9
Moderate	Low	Scenario 10	Scenario 11	Scenario 12
	Moderate	Scenario 13	Scenario 14	Scenario 15
	High	Scenario 16	Scenario 17	Scenario 18
High	Low	Scenario 19	Scenario 20	Scenario 21
	Moderate	Scenario 22	Scenario 23	Scenario 24
	High	Scenario 25	Scenario 26	Scenario 27

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Bernard, H. R., Killworth, P. D., & Sailer, L. (1980). Informant accuracy in social network data iv: A comparison of clique-level in behavioral and cognitive network data. *Social Science Research*, *11*, 30–66.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge university press.

Crano, W. D., Brewer, M. B., & Lac, A. (2014). *Principles and methods of social research*. Routledge.

Daniel, J. R., Santos, A. J., Peceguina, I., & Vaughn, B. E. (2013). Exponential random graph models of preschool affiliative networks. *Social Networks*, *35*(1), 25–30.

Geyer, C. J., & Thompson, E. A. (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(3), 657–683.

Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, *103*(481), 248–258.

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. *Journal of statistical software*, *24*(3), nihpa54860.

Kolaczyk, E. D., & Krivitsky, P. N. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *30*(2), 184.

Koskinen, J., Wang, P., Robins, G., & Pattison, P. (2018). Outliers and influential observations in exponential random graph models. *Psychometrika*, *83*(4), 809–830.

Lohr, S. L. (2009). *Sampling: design and analysis*. Nelson Education.

Lubbers, M. J. (2003). Group composition and network structure in school classes: a multilevel application of the p$*$ model. *Social Networks*, *25*(4), 309–332.

Lubbers, M. J., & Snijders, T. A. (2007). A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. *Social networks*, *29*(4), 489–507.

Lusher, D., Koskinen, J., & Robins, G. (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

McDonald, S., & Benton, R. A. (2017). The structure of internal job mobility and organizational wage inequality. *Research in Social Stratification and Mobility*, *47*, 21–31.

McKay, A. S., Grygiel, P., & Karwowski, M. (2017). Connected to create: A social network analysis of friendship ties and creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(3), 284.

Rambaran, J. A., Dijkstra, J. K., & Veenstra, R. (2020). Bullying as a group process in childhood: A longitudinal social network analysis. *Child development*, *91*(4), 1336–1352.

Rambaran, J. A., van Duijn, M. A., Dijkstra, J. K., & Veenstra, R. (2019). Stability and change in student classroom composition and its impact on peer victimization. *Journal of Educational Psychology*.

Rehnberg, Z. (2016). *Exponential random graph models under measurement error.* Washington University in St. Louis.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological methodology*, *31*(1), 361–395.

Team, R. C. (2013). *R: A language and environment for statistical computing.* Vienna, Austria.

Van Duijn, M. A., Gile, K. J., & Handcock, M. S. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, *31*(1), 52–62.

Van Rossem, R., & Vlegels, J. (2009). Ethnic segregation in flemish high schools:

Structure, homophily, ethnic subcultures or interethnic conflict. In *Esa 2009-9th conference of the european sociological association.*

Vermeij, L. (2006). *What's cooking: cultural boundaries among dutch teenagers of different ethnic origins in the context of school*. Utrecht University.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of statistical software*, *36*(3), 1–48.

Wit, E., Heuvel, E. v. d., & Romeijn, J.-W. (2012). 'all models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, *66*(3), 217–236.

Wittek, M., Kroneberg, C., & Lämmermann, K. (2020). Who is fighting with whom? how ethnic origin shapes friendship, dislike, and physical violence relations in german secondary schools. *Social Networks*, *60*, 34–47.