Network Science **9** (4): 478–498, 2022
doi:10.1017/nws.2021.22

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Iterative estimation of mixed exponential random graph models with nodal random effects

S. Kevork* and G. Kauermann

Institut für Statistik, Ludwig-Maximilians-Universität München, München, Germany.
*Corresponding author. Email: sevag.kevork@stat.uni-muenchen.de

Action Editor: Stanley Wasserman

Abstract

The presence of unobserved node-specific heterogeneity in exponential random graph models (ERGM) is a general concern, both with respect to model validity as well as estimation instability. We, therefore, include node-specific random effects in the ERGM that account for unobserved heterogeneity in the network. This leads to a mixed model with parametric as well as random coefficients, labelled as mixed ERGM. Estimation is carried out by iterating between approximate pseudolikelihood estimation for the random effects and maximum likelihood estimation for the remaining parameters in the model. This approach provides a stable algorithm, which allows to fit nodal heterogeneity effects even for large scale networks. We also propose model selection based on the Akaike Information Criterion to check for node-specific heterogeneity.

Keywords: exponential random graph models; random effects; generalized linear mixed models; network data analysis

1. Introduction

The analysis of network data has become a challenging and emerging field in statistics in the last years. Goldenberg et al. (2010), Hunter et al. (2012) and Fienberg (2012) provide comprehensive articles on statistical approaches, challenges and developments in network data analysis. We also refer to Kolaczyk (2009) and Kolaczyk and Csárdi (2014) for a general introduction and the related routines for network data analysis in R. In this study, we concentrate on exponential random graph models (ERGM) originally introduced in Frank and Strauss (1986) and more deeply discussed e.g. in Lusher et al. (2013). Unless node-specific covariates are included in the ERGM, the probability for all possible edges across the graph is assumed to be homogeneous, which also means that any permutation of the node labels will yield the same probability. This is a questionable assumption, in particular in large networks, which also contributes towards stability problems for estimation as discussed e.g. in Schweinberger (2011) or Schweinberger et al. (2017). We follow Thiemichen et al. (2016) and extend the ERGM by incorporating node-specific heterogeneity effects to overcome the homogeneity assumption of ERGM and capture the unobserved heterogeneity in the network.

Consider a network of n actors for which some dyadic relationships have been recorded. These relations can be represented in the form of an $n \times n$ adjacency matrix Y, with elements $Y_{ij} = 1$ if an edge from i to j exists and $Y_{ij} = 0$ otherwise. In undirected networks, we have $Y_{ij} = Y_{ji} \; \forall \, i \neq j$, which we assume throughout this paper. We consider the probability of

© The Author(s), 2022. Published by Cambridge University Press.

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

observing a given network conditional on a set of (sufficient) network statistics is given by the ERGM

$$\mathbb{P}\big(Y = y|\boldsymbol{\theta}\big) = \frac{\exp\{\boldsymbol{\theta}^T\, \boldsymbol{s}(y)\}}{\kappa(\boldsymbol{\theta})} \tag{1}$$

Here, $\boldsymbol{s}(y)$ is the vector of network statistics and $\boldsymbol{\theta}$ is the vector of model coefficients. Vector $\boldsymbol{s}(.)$ includes any structural characteristics of the network and we refer to Snijders et al. (2006) for a general discussion on network statistics, see also Hunter and Handcock (2006). The denominator $\kappa(\boldsymbol{\theta})$ in (1) represents the normalizing factor to ensure that (1) is a legitimate probability mass function. In general, $\kappa(\boldsymbol{\theta})$ is numerically intractable unless for miniature networks. Estimation of $\boldsymbol{\theta}$ in model (1) needs therefore to be carried out simulation based. An early reference for estimation of ERGMs is Snijders (2002). For a general discussion, we refer to Hunter et al. (2012). A numerical stable routine has been proposed in Hummel et al. (2012) using a so-called stepping algorithm. Bayesian estimation is proposed in Caimo and Friel (2011). An important property resulting from equation (1) is that $\boldsymbol{s}(y)$ is a vector of sufficient statistics for the network. This means, two networks, which coincide in $\boldsymbol{s}(y)$, have the same probability. In particular, this means that all possible node-specific heterogeneity in the network is explained by exogenous effects, which may also be included in model (1). This can be seen as questionable assumption. For instance, in a friendship network, we may believe that the formation of friendships (edges) between individuals (nodes) is driven by many factors, observable as well as unobservable. We may suspect that there are quantities, intangible factors specific to each individual (node) that are difficult if not impossible to measure. It seems therefore, advisable to include node-specific heterogeneity to capture possible unobserved heterogeneity of the nodes.

An early model that incorporates node-specific heterogeneity is the so-called p_1 model proposed in Holland and Leinhardt (1981). The model includes parametric sender and receiver effects but no network statistics (except of reciprocity). Random nodal heterogeneity was proposed by Duijn et al. (2004) or Zijlstra et al. (2006) which led to the so-called p_2 model. Thiemichen et al. (2016) combined the approach with ERGMs and proposed Bayesian estimation, which however is infeasible for large networks, i.e. networks with more than about 100 actors. We follow the approach of Thiemichen et al. (2016) and extended the ERGM towards

$$\mathbb{P}\big(Y = y|\boldsymbol{\theta}, \boldsymbol{u}\big) = \frac{\exp\big\{\boldsymbol{\theta}^T\boldsymbol{s}(y) + \boldsymbol{u}^T\boldsymbol{t}(y)\big\}}{\kappa\big(\boldsymbol{\theta}, \boldsymbol{u}\big)} \tag{2}$$

where $\boldsymbol{s}(y) = (s_0(y), s_1(y), \dots)$ is, as above, the p dimensional vector of network statistics with $s_0(y) = \sum_i \sum_{j>i} y_{ij}$ as intercept and $\boldsymbol{t}(y) = \big(\sum_{j\neq 1} y_{1j},\ \sum_{j\neq 2} y_{2j}, \dots\big)$ is the n dimensional vector of node degrees. The normalization now equals $\kappa(\boldsymbol{\theta}, \boldsymbol{u}) = \log \sum_{y\in\mathcal{Y}} \exp\big(\boldsymbol{\theta}^T\boldsymbol{s}(y) + \boldsymbol{u}^T\boldsymbol{t}(y)\big)$, where $\mathcal{Y}$ is the set of n by n networks. Conditional on $\boldsymbol{u}$, we obtain node-specific heterogeneity, which can be seen as follows. We assume now that $\boldsymbol{\theta} \in \mathbb{R}^p$ is a p dimensional parameter vector while $\boldsymbol{u} = (u_1, u_2, \dots)$ is a n dimensional vector of random node-specific coefficients with

$$\boldsymbol{u} \sim \mathcal{N}(0,\ \sigma_u^2\, \boldsymbol{\mathcal{I}}_n) \tag{3}$$

with σ_u^2 as variance and $\boldsymbol{\mathcal{I}}_n$ as n dimensional identity matrix. This leads to a mixed model with fixed and random coefficients, termed in the following as mixed ERGM, or in short mERGM. The reasoning behind the model structure can be seen through the conditional model for a single edge Y_{ij} conditional on the rest of the network denoted as Y_{-ij}. From (2), we obtain

$$\log\left\{\frac{\mathbb{P}(Y_{ij} = 1|y_{-ij}, \boldsymbol{\theta}, \boldsymbol{u})}{\mathbb{P}(Y_{ij} = 0|y_{-ij}, \boldsymbol{\theta}, \boldsymbol{u})}\right\} = \boldsymbol{\theta}^T\ \Delta_{ij}\, \boldsymbol{s}(y)\ + u_i + u_j \tag{4}$$

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Here, $\Delta_{ij}\boldsymbol{s}(y) = \boldsymbol{s}(y_{ij} = 1, y_{-ij}) - \boldsymbol{s}(y_{ij} = 0, y_{-ij})$ is the so-called change statistics where y_{-ij} is the network except of edge y_{ij}. The terms u_i and u_j are the random node-specific coefficients accounting for heterogeneity not captured with $\Delta_{ij}\boldsymbol{s}(y)$. If we assume normality for coefficients u_i, formula (4) resembles a mixed logistic regression model as extensively discussed e.g. in Breslow and Clayton (1993). A similar model to (4) has been proposed by Box-Steffensmeier et al. (2018) taking the coefficients $\boldsymbol{u}$ in (4) as random normal variable with mean zero and unknown variance. For estimation, they apply a pseudolikelihood approach. Though this circumvents the numerical burden of estimation in ERGM, it comes for the price of biased estimation of the paramteric coefficients $\boldsymbol{\theta}$. In other words, their estimation approach is biased even if the random node effects are close or equal to zero. We refer to Schmid and Desmarais (2017) for a general discussion on pseudolikelihood estimation in ERGMs, see also Strauss and Ikeda (1990) or Desmarais and Cranmer (2012).

Model (2) combined with the probability model (3) can also be seen as special case of the exponential-family random network model as proposed in Fellows and Handcock (2012). They propose simulation-based estimation, which is restricted to small network sizes. The restriction to small networks is also pointed out in Thiemichen et al. (2016). Hence, even though model (2) can be considered as ERGM with some prior structure on coefficients $\boldsymbol{u}$ given in (2), estimation based on simulation based methods (see Snijders 2002) becomes infeasible for larger networks. We therefore, propose a different estimation strategy, which is motivated in the next paragraph. At this point, we also note that several approaches to capture unobserved heterogeneity were introduced in network data analysis, for instance, Koskinen (2009) introduced binary latent class indicators, Schweinberger and Handcock (2015) examined local dependence using a Bayesian framework in random graph models and Henry et al. (2020) developed a modeling framework to capture unobserved heterogeneity in the effects of nodal covariates.

The first goal of this paper is to provide an iterative estimation strategy, combining both maximum likelihood and pseudolikelihood estimation techniques. To be more specific, we take model (4) as starting point and make use of pseudolikelihood estimation for the random coefficients $\boldsymbol{u}$ while for estimation of $\boldsymbol{\theta}$ we use the steplength MCMC-MLE approach proposed in Hummel et al. (2012) and the corresponding stepping algorithm Hummel et al. (2012), implemented in the `ergm` package in R (see Hunter et al. 2008). These two steps are used iteratively, leading to feasible estimation. Our estimation strategy allows us to fit large scale networks, and we observe that the inclusion of the nodal effects works towards numerical estimation stability, as demonstrated through examples and simulations. Moreover, as the second goal of this paper, we propose a simple model selection strategy to evaluate nodal heterogeneity. To be specific, we use Akaike's Information Criterion (see Akaike 1974) to select a model with or without nodal effects. The latter is calculated numerically by employing a Laplace approximation.

This paper is organized as follows. In Section 2, we discuss the estimation for the underlying model in detail and introduce our algorithm. Furthermore, in Section 3, we present a simulation-based method for evaluating our model, which allows us to calculate the AIC value for the mERGM and compare it with the AIC value of a corresponding ERGM. In Section 4, we present a simulation-based study with the corresponding results, in Section 5, we then apply our approach to three data examples. Finally, Section 6 closes with a discussion.

2. Model and estimation

We consider model (2) where the nodal heterogeneity effects $u_1, \ldots, u_n$ are assumed to be random with σ_u^2 as variance and $\boldsymbol{\mathcal{I}}_n$ as n dimensional identity matrix. The aim is to fit parameter $\boldsymbol{\theta}$ taking nodal heterogeneity into account. Moreover, we need to estimate σ_u^2, which in fact quantifies the amount of nodal heterogeneity.

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

In principle, we need to maximize the marginal log-likelihood

$$l(\boldsymbol{\theta},\ \sigma_u^2) = \log \int \frac{\exp\{\boldsymbol{\theta}^T \boldsymbol{s}(y) + \boldsymbol{u}^T \boldsymbol{t}(y)\}}{\kappa(\boldsymbol{\theta}, \boldsymbol{u})} \cdot \frac{1}{(2\pi\sigma_u^2)^{\frac{n}{2}}} \cdot \exp\Big(-\frac{1}{2}\,\frac{\boldsymbol{u}^T\boldsymbol{u}}{\sigma_u^2}\Big) \prod_{i=1}^{n} du_i$$

$$= \boldsymbol{\theta}^T \boldsymbol{s}(y) - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_u^2) + \log \int \exp\Big(g(\boldsymbol{u}, \boldsymbol{\theta}, \sigma_u^2)\Big) \prod_{i=1}^{n} du_i \tag{5}$$

where

$$g(\boldsymbol{u}, \boldsymbol{\theta}, \sigma_u^2) = \boldsymbol{u}^T \boldsymbol{t}(y) - \log(\kappa(\boldsymbol{\theta}, \boldsymbol{u})) - \frac{1}{2} \cdot \frac{\boldsymbol{u}^T\boldsymbol{u}}{\sigma_u^2} \tag{6}$$

We may approximate the integral in (5) using a Laplace approximation. Let therefore $\hat{u}$ be the maximizer of $g(\boldsymbol{u}, \boldsymbol{\theta}, \sigma_u^2)$, which apparently depends on both, $\boldsymbol{\theta}$ and σ_u^2. This leads to the approximate log-likelihood

$$l(\boldsymbol{\theta}, \sigma_u^2) \approx \boldsymbol{\theta}^T \boldsymbol{s}(y) + \hat{u}^T \boldsymbol{t}(y) - \log(\kappa(\boldsymbol{\theta}, \hat{u})) - \frac{1}{2} \cdot \frac{\hat{u}^T\hat{u}}{\sigma_u^2} - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_u^2)$$

$$- \frac{1}{2}\ \log\left|\frac{\partial^2 g(\hat{u}, \boldsymbol{\theta}, \sigma_u^2)}{\partial u\, \partial u}\right| \tag{7}$$

Note that the likelihood can be considered as profile likelihood, where $\boldsymbol{u}$ is "estimated" through maximizing (6). If we now treat $\hat{u}$ as given, then maximization of $l(\boldsymbol{\theta}, \sigma_u^2)$ with respect to $\boldsymbol{\theta}$ corresponds to maximizing the likelihood of the probability model

$$\mathbb{P}\big(Y = y|\boldsymbol{\theta}, \hat{u}\big) = \frac{\exp\{\boldsymbol{\theta}^T \boldsymbol{s}(y) + \hat{u}^T \boldsymbol{t}(y)\}}{\kappa(\boldsymbol{\theta}, \hat{u})}$$

where $\hat{u}^T \boldsymbol{t}(y)$ is fixed as given offset. The terminology offset means here that the quantity $\hat{u}^T \boldsymbol{t}(y)$ is treated as fixed. In other words, setting the random coefficients to $\hat{u}$ simplifies the estimation of $\boldsymbol{\theta}$ to Maximum Likelihood estimation in an ERGM with offset $\hat{u}^T \boldsymbol{t}(y)$. This is numerically available with standard software packages (e.g. Hunter et al. 2008) and the stepping algorithm proposed in Hummel et al. (2012). Let us therefore first look in more detail how to obtain $\hat{u}$ if we keep $\boldsymbol{\theta}$ as fixed. Note that $\hat{u}$ results by solving

$$\frac{\partial\, g(\boldsymbol{u}, \boldsymbol{\theta}, \sigma_u^2)}{\partial \boldsymbol{u}} = 0$$

Apparently, this is numerically problematic, since $\kappa(\boldsymbol{\theta}, \boldsymbol{u})$ is numerically intractable. Differentiation yields

$$\frac{\partial \log \kappa(\boldsymbol{\theta}, \boldsymbol{u})}{\partial \boldsymbol{u}} = \mathbb{E}(\boldsymbol{t}(y)|\boldsymbol{u})$$

which in principle can be approximated using simulation based approaches (see Snijders 2002). However, this is a numerically challenging task, since $\boldsymbol{u}$ is high dimensional, namely n dimensional. We therefore propose to approximate the estimation step of $\boldsymbol{u}$ by pseudolikelihood estimation. To do so, we look at the model for a single edge given the rest of the network. That is to say we take model (4), but now we fix $o_{ij} := \boldsymbol{\theta}^T\, \Delta_{ij}\, \boldsymbol{s}(y)$ as offset and ignore the dependence on

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

y. This leads to a pseudolikelihood approach as discussed for instance in van Duijn et al. (2009). The pseudo-log-likelihood thereby equals

$$l_{\text{pseudo}}(\boldsymbol{u}) = \sum_i \sum_{j>i} y_{ij}\{o_{ij} + u_i + u_j\} - \log\{1 + \exp(o_{ij} + u_i + u_j)\} \tag{8}$$

where a priori coefficient $\boldsymbol{u}$ follows a normal distribution as given in (3). The pseudo-log-likelihood (8) together with (3) leads to (pseudo) generalized linear mixed effects model, so that we can borrow estimation strategies from this field. In particular, we make use of Breslow and Clayton (1993) who propose to approximate the resulting marginal likelihood using a Laplace approximation, similar to (7) above. This allows to estimate the a priori variance σ_u^2 and predict the random coefficients $\boldsymbol{u}$. Note, if o_{ij} is independent of y_{-ij}, then the Laplace approach (7) is equal to the estimation proposed in Breslow and Clayton (1993). Apparently, this is the case for the so-called p_2 model (see Duijn et al. 2004). To be specific, the Laplace approximated pseudolikelihood with $o_{ij} = (o_{1\,2}, o_{1\,3}, \ldots, o_{n-1\,n})$ as offset and constant terms omitted results through

$$l_{\text{pseudo}}(\sigma_u^2) \approx -\frac{1}{2} \cdot \frac{\check{u}^T \check{u}}{\sigma_u^2} - \frac{n}{2} \log(\sigma_u^2) \tag{9}$$

where $\check{u}$ is the maximizer of

$$\check{g}(\boldsymbol{u}, \sigma_u^2; o) = \boldsymbol{u}^T \boldsymbol{t}(y) - \sum_i \sum_j \log\{1 + \exp(o_{ij} + u_i + u_j)\} - \frac{1}{2} \cdot \frac{\boldsymbol{u}^T \boldsymbol{u}}{\sigma_u^2}$$

We call $\check{u}$ an estimate subsequently, even though of course it is a predictor given that $\boldsymbol{u}$ is considered as random. This is implemented in multiple R packages, see e.g. Faraway (2016). To be specific, for estimation of $\boldsymbol{u}$ we use the `mgcv` package (see Wood 2011). Denoting with $\check{\boldsymbol{u}} = (\check{u}_1, \ldots, \check{u}_n)$ the resulting estimates, we set $\check{\boldsymbol{u}}^T \boldsymbol{t}(y)$ in (2) as offset and estimate parameter $\boldsymbol{\theta}$ using simulation based techniques. For this step we use the `ergm` package (see Hunter et al. 2008). Both estimation steps are used iteratively until convergence. That is we take the current estimate $\hat{\boldsymbol{\theta}}_{(t)}$ and update $\check{\boldsymbol{u}}$ with pseudolikelihood leading to $\check{\boldsymbol{u}}_{(t+1)}$. This in turn allows to update $\hat{\boldsymbol{\theta}}$ after replacing the offset by $\check{\boldsymbol{u}}_{(t+1)}^T \boldsymbol{t}(y)$. Our algorithmic steps work in detail as follows:

Algorithm 1: Fit ERGM with nodal random effect components

Step 0: Obtain a prediction for $\boldsymbol{u}$ and estimate σ_u^2:
- a. Fit the model `logit` $\mathbb{P}(y_{ij} = 1 | y_{-ij}, \boldsymbol{\theta}, \boldsymbol{u}) = \theta_{(0)} + u_i + u_j$ to the data, where $1 \le i < j \le n$
- b. extract the vector of the predicted random effects $\check{u}_{(0)}$ as offset and set $t = 0$

Step 1: Estimate $\boldsymbol{\theta}$ with ERGM and take $\check{u}_{(t)}^T \boldsymbol{t}(y)$ as an offset parameter:
- a. Fit the model $\mathbb{P}(Y = y | \boldsymbol{\theta}) \propto \exp\{\boldsymbol{\theta}_{(t+1)}^T \boldsymbol{s}(y) + \underbrace{\check{u}_{(t)}^T \boldsymbol{t}(y)}_{\text{offset}}\}$ using maximum likelihood and simulation based methods
- b. extract $o_{ij} = \boldsymbol{\theta}_{(t+1)}^T \Delta_{ij} \boldsymbol{s}(y)$ as new offset for $1 \le i < j \le n$

Step 2: Update $\check{u}_{(t+1)}$ and $\hat{\sigma}_{u(t+1)}^2$ now taking o_{ij} as offset parameter:
- a. Fit the model `logit` $\mathbb{P}(y_{ij} = 1 | y_{-ij}, \boldsymbol{\theta}, \boldsymbol{u}) = o_{ij} + u_i + u_j$ with priori structure (3) to the data
- b. extract the vector of the predicted random effects $\check{u}_{(t+1)}$ as new offset

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Set $t = t + 1$ and iterate between step 1 and 2 until $\boldsymbol{\theta}_{(t)}$ converges. Convergence is achieved if $|\boldsymbol{\theta}_{(t)} - \boldsymbol{\theta}_{(t+1)}| \leq \epsilon = 0.05$.

We need to mention that pseudolikelihood estimation in network data analysis is biased, in particular, if dyadic statistics are involved. Our setting here, however, is slightly different since we use pseudolikelihood estimation only for the degree statistics $\boldsymbol{t}(y)$. Moreover, we consider the coefficients $\boldsymbol{u}$ not to be fixed but random so that in general, we are more interested in the variance of coefficients u_i and less interested in their true value. We refer to Schmid and Desmarais (2017) or Cranmer and Desmarais (2011) for further discussion on pseudolikelihood estimation in network data analysis.

3. Inference through model selection and variance estimation

The central question in network data analysis is to explain the dominating factors in the network, i.e. the sufficient statistics describing the network structure. If we allow for node-specific heterogeneity, we are additionally faced with the problem of model selection. In other words, we need to describe whether the network data at hand shows evidence with respect to heterogeneous nodes or whether the homogeneity assumption of ERGM seems valid. We tackle this question by approximate calculation of the Akaike Information Criterion (AIC). To do so, we assume for simplicity that the determinant component in (7) depends only weakly on $\boldsymbol{\theta}$ so that we can ignore it subsequently. This is in line with the arguments proposed in Breslow and Clayton (1993) who suggest the use of Laplace approximation in generalized linear mixed models. Note that

$$\frac{\partial^2 g(\boldsymbol{u}, \boldsymbol{\theta}, \sigma_u^2)}{\partial \boldsymbol{u}\, \partial \boldsymbol{u}^{\boldsymbol{T}}} = -Var(\boldsymbol{t}(Y)\,|\boldsymbol{\theta}, \boldsymbol{u}) - \frac{1}{\hat{\sigma}_u^2}\,\mathcal{I}_n \tag{10}$$

Hence, ignoring the determinant component in (7) is justified if we assume that the variance matrix of the degree vector $\boldsymbol{t}(Y)$ depends only weakly on $\boldsymbol{\theta}$. We refer to Breslow and Clayton (1993) for a deeper discussion and motivation which justifies to pursue this simplification. Generally, the variance is difficult to calculate or even infeasible for large networks, so that we make use of simulations to estimate (10). To do so, we simulate networks in order to obtain a simulation based approximation for $\kappa(\hat{\boldsymbol{\theta}}, \breve{u})$. We make further use of the simulated networks to obtain a simulation based approximation of $Var(\boldsymbol{t}(y)|\hat{\boldsymbol{\theta}}, \breve{u})$. To be specific, let $y^{\star(1)}, \ldots, y^{\star(N)}$ be a set of (independent) network simulations derived from model (2) with $\boldsymbol{\theta}$ set to $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{u}$ set to $\breve{u}$. We estimate $Var(\boldsymbol{t}(y)\,|\,\hat{\boldsymbol{\theta}},\ \breve{\boldsymbol{u}})$ through

$$\frac{1}{N}\sum_{j=1}^{N}\left[\boldsymbol{t}(y^{\star(j)}) - \bar{t}(y^{\star})\right]\left[\boldsymbol{t}(y^{\star(j)}) - \bar{t}(y^{\star})\right]^T$$

where $\bar{t}(y^{\star})$ is the arithmetic mean of the simulated values.

With these prerequisites, we can now approximate all quantities in (7). This also holds for the normalization constant, which is estimated through

$$\hat{\kappa}(\hat{\boldsymbol{\theta}}, \breve{u}) = \frac{1}{N}\sum_{j=1}^{N}\exp\big(\hat{\boldsymbol{\theta}}^T\boldsymbol{s}(y^{\star(j)}) + \breve{u}^T\boldsymbol{t}(y^{\star(j)})\big)$$

Model comparison can now be carried out with the AIC. Setting p as the number of parameters in $\boldsymbol{\theta}$ the AIC results to

$$AIC_{mERGM} = -2\,l(\hat{\boldsymbol{\theta}}, \hat{\sigma}_u^2) + 2\cdot(p+1) \tag{11}$$

Note that (11) resembles the marginal AIC, that is after integrating out $\boldsymbol{u}$. We refer to Greven and Kneib (2010) or Vaida and Blanchard (2005) for a deeper discussion of applying AIC

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

in random effects models. In our case, formula (11) serves as approximation, relying on the pseudolikelihood estimation for $\boldsymbol{u}$.

We compare the AIC in the mERGM to the resulting AIC in the case of node homogeneity, that is by setting $\sigma_u^2 = 0$. This is carried out in a similar way, but we set $\boldsymbol{u} = 0$.

In other words, we use the likelihood in (1) by calculating $\kappa(\boldsymbol{\theta})$ simulation based from N draws $y^{\star(1)}, \ldots, y^{\star(N)}$ from model (1) with $\boldsymbol{\theta}$ set to the ML estimate in model (1). We call this

$$AIC_{ERGM} = -2\, l_{ERGM}(\hat{\boldsymbol{\theta}}) \; + \; 2p \tag{12}$$

where l_{ERGM} is the log-likelihood in the ERGM resulting when $\boldsymbol{u} \equiv 0$.

Though the focus of the paper lies on model comparison, we shortly discuss how to calculate the variance of the estimate if the algorithm above is used. In the ERGM, we obtain the Fisher information matrix

$$I(\boldsymbol{\theta}) = Var(\boldsymbol{s}(Y)|\boldsymbol{\theta}) \tag{13}$$

This can be estimated simulation based, that is we simulate from model (1) and calculate $Var(\boldsymbol{s}(Y))$ based on the simulated values. For the mERGM, we need to take into account that coefficients $\boldsymbol{u}$ are considered to be random so that in principle, we need to calculate the (inverse) Fisher information of the log-likelihood. Assuming that the determinant in (7) does depend only weekly on $\boldsymbol{\theta}$ and ignoring for simplicity the dependence of $\breve{u}$ on $\boldsymbol{\theta}$ we obtain again (13). This is, of course, an approximation since we ignored the dependence of $\breve{u}$ and $\boldsymbol{\theta}$ as well as estimation variability of σ_u^2. In other words, exact variance calculation in the mERGM is complicated and we here provide a rough approximation only. However, inference can be carried out by model selection via the AIC, which is what we pursue in simulations and data examples below. Still, we can make use of the simulations from above to obtain an estimate of the Fisher information and hence a variance estimate for the estimates.

4. Simulation study

In our simulation study, we want to explore the estimation results of the model parameters and the model selection step based on two network sizes. Small networks with 50 nodes, and large networks with 500 nodes. For each network size, we use network settings with different levels of nodal heterogeneity σ_u^2. For each network setting and network size, we simulate 50 networks using the simulation routines from the `ergm` package (Hunter et al., 2008). Each network setting has the same structural effects $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_{edges},\ \theta_{gwesp},\ \theta_{2-stars}) = (-1,\ 0.2,\ -0.3)$, but the nodal heterogeneity takes six different levels $\sigma_u^2 = (0,\ 0.1,\ 0.2,\ 0.5,\ 0.8,\ 1)$, where $\boldsymbol{u}$ is randomly drawn for each simulated network and setting from a normal distribution following (3). For each network size of these six heterogeneity levels, we fit an ERGM and a mERGM to the 50 simulated networks. Note that the tuning parameters for the θ_{gwesp} term are the same for both ERGMs and mERGMs with (`decay = 0.8, fixed = TRUE`). Additionally, we provide the results of two more simulation studies as supplementary material, where we perturb the parameter vector $\boldsymbol{\theta}$ with different configurations for both small and large networks. Furthermore, we also provide supplementary material illustrations of how well σ_u^2 is recovered in the simulation study.

Our first focus is on the performance of the estimates. In Tables 1 and 2, we summarize the results of our simulation study for small (50 nodes) and large (500 nodes) networks respectively, distinguishing all six different levels of nodal heterogeneity. Let us first look at the case $\sigma_u^2 = 0$, the fitted parametric coefficients of the ERGMs in both cases, small and large networks, show a stable estimation performance; however, this stability is more evident in large networks. The mERGMs, on the other hand, show some estimation variability and are outperformed by ERGMs concerning parameter estimation. The trend changes increasingly in favour of mERGMs, with

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Table 1. Resulting means, standard deviations, the medians, 0.1 and 0.9 quantiles of the estimated coefficients of network size 50 nodes and for all six σ_u^2 levels

Network size: 50 Nodes								
σ_u^2	Model type	Parameter	Real value	Mean	SD	Q 0.1	Median	Q 0.9
0	ERGM	θ_{edges}	−1	−0.71	0.79	−1.65	−0.74	0.37
		θ_{gwesp}	0.2	0.06	0.78	−0.06	0.17	0.38
		$\theta_{2-stars}$	−0.3	−0.35	0.13	−0.50	−0.34	−0.20
0	mERGM	θ_{edges}	−1	0.52	1.18	−0.89	0.39	0.31
		θ_{gwesp}	0.2	0.06	0.81	−0.06	0.17	0.38
		$\theta_{2-stars}$	−0.3	−0.58	0.19	−0.86	−0.57	−0.35
0.1	ERGM	θ_{edges}	−1	−1.11	0.77	−2.13	−1.12	−0.15
		θ_{gwesp}	0.2	0.16	0.16	0.03	0.19	0.34
		$\theta_{2-stars}$	−0.3	−0.25	0.12	−0.39	−0.24	−0.10
0.1	mERGM	θ_{edges}	−1	0.06	0.89	−0.96	−0.03	0.98
		θ_{gwesp}	0.2	0.16	0.16	−0.02	0.19	0.34
		$\theta_{2-stars}$	−0.3	−0.45	0.14	−0.60	−0.44	−0.28
0.2	ERGM	θ_{edges}	−1	−1.63	0.52	−2.21	−1.72	−0.88
		θ_{gwesp}	0.2	0.17	0.21	−0.06	0.17	0.39
		$\theta_{2-stars}$	−0.3	−0.17	0.08	−0.31	−0.16	−0.06
0.2	mERGM	θ_{edges}	−1	−0.73	0.61	−1.44	−0.85	0.21
		θ_{gwesp}	0.2	0.18	0.21	−0.07	0.19	0.40
		$\theta_{2-stars}$	−0.3	−0.33	0.11	−0.49	−0.31	−0.20
0.5	ERGM	θ_{edges}	−1	−2.05	0.43	−2.51	−2.10	−1.47
		θ_{gwesp}	0.2	0.21	0.12	0.03	0.21	0.37
		$\theta_{2-stars}$	−0.3	−0.07	0.06	−0.13	−0.06	−0.01
0.5	mERGM	θ_{edges}	−1	−0.71	0.62	−1.35	−0.77	0.03
		θ_{gwesp}	0.2	0.21	0.12	0.03	0.22	0.36
		$\theta_{2-stars}$	−0.3	−0.26	0.08	−0.38	−0.26	−0.17
0.8	ERGM	θ_{edges}	−1	−2.90	0.27	−3.17	−2.93	−2.58
		θ_{gwesp}	0.2	0.21	0.14	−0.01	0.21	0.34
		$\theta_{2-stars}$	−0.3	0.02	0.04	−0.04	0.03	0.07
0.8	mERGM	θ_{edges}	−1	−1.02	0.22	−1.33	−1.04	−0.74
		θ_{gwesp}	0.2	0.20	0.16	−0.03	0.23	0.37
		$\theta_{2-stars}$	−0.3	−0.39	0.14	−0.49	−0.33	−0.29
1	ERGM	θ_{edges}	−1	1123.83	9739.40	−186.45	−12.76	4.27
		θ_{gwesp}	0.2	−1112.36	4651.95	−737.19	−15.23	0.36
		$\theta_{2-stars}$	−0.3	785.38	10336.06	−2118.73	−3.77	104.20
1	mERGM	θ_{edges}	−1	−1.23	0.40	−1.68	−1.23	−0.74
		θ_{gwesp}	0.2	0.26	0.18	0.01	0.30	0.49
		$\theta_{2-stars}$	−0.3	-00.37	0.10	−0.49	−0.35	−0.26

increasing heterogeneity. At a heterogeneity of $\sigma_u^2 = 0.8$, the mERGMs excel with better results in small and large networks than the ERGMs, which show substantial variability in the results. At a heterogeneity of $\sigma_u^2 = 1$, the results indicate severe stability problems of the ERGM estimates, especially in the small networks, but the more the size of the network increases, the less unstable

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Table 2. Resulting means, standard deviations, the medians, 0.1 and 0.9 quantiles of the estimated coefficients of network size 500 nodes and for all six σ_u^2 levels

Network size: 500 nodes								
σ_u^2	Model type	Parameter	Real value	Mean	SD	Q 0.1	Median	Q 0.9
0	ERGM	θ_{edges}	−1	−0.96	0.29	−1.31	−0.98	−0.55
		θ_{gwesp}	0.2	0.19	0.05	0.12	0.20	0.27
		$\theta_{2-stars}$	−0.3	−0.30	0.03	−0.34	−0.30	−0.27
0	mERGM	θ_{edges}	−1	0.18	0.51	−0.53	0.27	0.81
		θ_{gwesp}	0.2	0.19	0.05	0.13	0.20	0.27
		$\theta_{2-stars}$	−0.3	−0.16	0.04	−0.21	−0.16	−0.11
0.1	ERGM	θ_{edges}	−1	−1.99	0.24	−2.34	−1.99	−1.73
		θ_{gwesp}	0.2	0.19	0.05	0.13	0.20	0.25
		$\theta_{2-stars}$	−0.3	−0.22	0.02	−0.23	−0.21	−0.18
0.1	mERGM	θ_{edges}	−1	−0.13	0.07	−0.22	−0.14	−0.04
		θ_{gwesp}	0.2	0.19	0.05	0.14	0.20	0.25
		$\theta_{2-stars}$	−0.3	−0.16	0.09	−0.28	−0.17	−0.03
0.2	ERGM	θ_{edges}	−1	−2.59	0.29	−2.93	−2.63	−2.16
		θ_{gwesp}	0.2	0.19	0.06	0.12	0.20	0.26
		$\theta_{2-stars}$	−0.3	−0.16	0.02	−0.19	−0.15	−0.13
0.2	mERGM	θ_{edges}	−1	−0.57	0.26	−0.92	−0.54	−0.22
		θ_{gwesp}	0.2	0.19	0.07	0.12	0.19	0.27
		$\theta_{2-stars}$	−0.3	−0.45	0.04	−0.49	−0.45	−0.42
0.5	ERGM	θ_{edges}	−1	−3.65	0.16	−3.82	−3.68	−3.45
		θ_{gwesp}	0.2	0.19	0.04	0.14	0.20	0.25
		$\theta_{2-stars}$	−0.3	−0.06	0.01	−0.08	−0.06	−0.05
0.5	mERGM	θ_{edges}	−1	−0.48	0.28	−0.79	−0.48	−0.19
		θ_{gwesp}	0.2	0.19	0.05	0.14	0.18	0.25
		$\theta_{2-stars}$	−0.3	−0.34	0.02	−0.37	−0.35	−0.32
0.8	ERGM	θ_{edges}	−1	−4.29	0.11	−4.41	−4.29	−4.15
		θ_{gwesp}	0.2	0.19	0.04	0.14	0.20	0.24
		$\theta_{2-stars}$	−0.3	−0.01	0.01	−0.02	−0.01	−0.001
0.8	mERGM	θ_{edges}	−1	−1.45	0.24	−1.77	−1.45	−1.16
		θ_{gwesp}	0.2	0.20	0.04	0.15	0.20	0.24
		$\theta_{2-stars}$	−0.3	−0.26	0.02	−0.28	−0.26	−0.23
1	ERGM	θ_{edges}	−1	−4.58	0.08	−4.67	−4.57	−4.48
		θ_{gwesp}	0.2	0.19	0.04	0.15	0.19	0.23
		$\theta_{2-stars}$	−0.3	0.01	0.01	0.01	0.01	0.02
1	mERGM	θ_{edges}	−1	−1.22	0.21	−1.84	−1.41	−1.29
		θ_{gwesp}	0.2	0.19	0.05	0.14	0.19	0.24
		$\theta_{2-stars}$	−0.3	−0.26	0.02	−0.26	−0.25	−0.23

the estimates of the ERGMs become. This fact is not surprising since a heterogeneity of $\sigma_u^2 = 1$ in a network with 50 nodes is a different claim than in a network with 500 nodes. Nevertheless, the mERGMs in both cases, small and large networks, show an appropriate and stable performance.

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

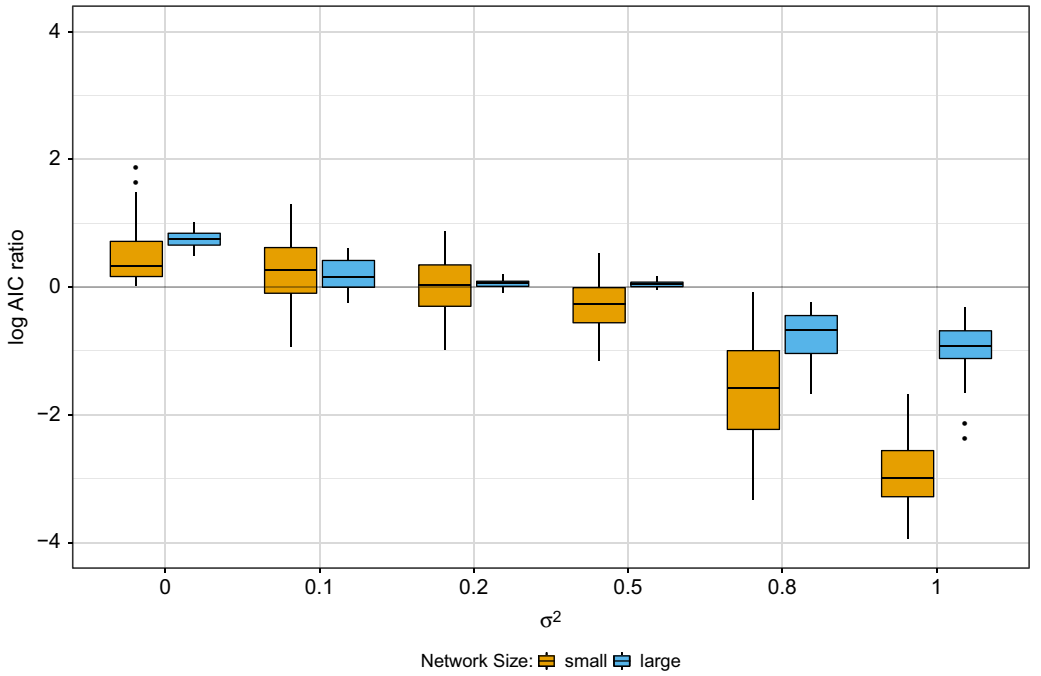


Figure 1. Resulting log AIC ratios of mERGM and ERGM for each network setting and size. Orange boxplots indicate small networks (50 nodes), blue boxplots large networks (500 nodes).

Hence, including nodal heterogeneity in the model increases the stability of the mERGM fit. This is a welcome effect of the model extension from ERGM to mERGM.

As a second point, we consider the performance of the model selection based on the AIC. To do so, we calculate for each of the 50 simulations in each setting the log ratio

$$\log\left(\frac{AIC_{mERGM_k}}{AIC_{ERGM_k}}\right) \quad k = 1, \ldots, 50$$

If the log ratio is positive, it speaks in favour of a model without nodal heterogeneity. In contrast, if the log ratio is negative, there is an indication of model heterogeneity. Figure 1 visualizes the log AIC ratio for different strengths of nodal heterogeneity and both network sizes. In small (50 nodes) and large (500 nodes) networks, the ERGM was correctly preferred in the case of missing nodal heterogeneity, that is $\sigma_u^2 = 0$. With increasing nodal heterogeneity level in the network, the mERGM becomes more appropriate, and from a heterogeneity level of $\sigma_u^2 = 0.5$ and $\sigma_u^2 = 0.8$ for small networks and large networks, respectively, the mERGM gets clearly selected based on the AIC. We, therefore, can conclude that the AIC allows for model selection in case of node-specific heterogeneity.

Additionally, we compare our approach with the approach proposed by Box-Steffensmeier et al. (2018) (FERGM) to illustrate the point of biased estimation of the parametric coefficients $\boldsymbol{\theta}$ we mentioned in Section (1). We apply the same simulation approach as above, distinguishing the size of the networks and compare the results of the mERGM with them of FERGM fitted to the 50 simulated networks with setting $\sigma_u^2 = 1$. The results of our comparison are given in Table 3. As we can conclude from Table 3, regardless of the network size, the performance of FERGM is worse. Nevertheless, we have to note that Box-Steffensmeier et al. (2018) also points out that their approach might only be suitable for certain research questions or certain network types.

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Table 3. Resulting estimated means, standard deviations, the medians, 0.1 and 0.9 quantiles of the parameters for network setting with nodal heterogeneity $\sigma_u^2 = 1$

Network size	Model type	Parameter	Real value	Mean	SD	Q 0.1	Median	Q 0.9
50 Nodes	mERGM	θ_{edges}	−1	−1.23	0.40	−1.68	−1.23	−0.74
		θ_{gwesp}	0.2	0.26	0.18	0.01	0.30	0.49
		$\theta_{2-stars}$	−0.3	−0.37	0.10	−0.49	−0.35	−0.26
50 Nodes	FERGM	θ_{edges}	−1	37.99	6.51	35.29	38.89	42.06
		θ_{gwesp}	0.2	0.12	0.19	−0.11	0.10	0.38
		$\theta_{2-stars}$	−0.3	-7.58	1.23	−8.43	−7.65	−7.08
500 Nodes	mERGM	θ_{edges}	−1	−1.22	0.21	−1.84	−1.41	−1.29
		θ_{gwesp}	0.2	0.19	0.05	0.14	0.19	0.24
		$\theta_{2-stars}$	−0.3	−0.26	0.02	−0.26	−0.25	−0.23
500 Nodes	FERGM	θ_{edges}	-[1	-[5.18	0.04	-[5.24	-[5.18	-[5.12
		θ_{gwesp}	0.2	0.18	0.04	0.14	0.19	0.23
		$\theta_{2-stars}$	-[0.3	0.06	0.01	0.05	0.06	0.06

Table 4. Models with sufficient network statistics for the Facebook network data

Model type	Model	θ_{edges}	θ_{gwesp}	$\theta_{2-stars}$	θ_{gwnsp}	Nodal effects
ERGM	1	✓	✓	✓		
	2	✓	✓		✓	
mERGM	3	✓	✓	✓		✓
	4	✓	✓		✓	✓

5. Examples

5.1 Facebook network

As a first data example, we look at Facebook (undirected) network data, which is publicly accessible (McAuley and Leskovec, 2012). The entire network comprises 4039 nodes. We analyze a subset of 250 nodes to demonstrate the performance of our routine, e.g. we take the first 251 nodes of the network and remove the "center" node (the "ego"). Figure 2 gives a visual impression of the network. Just by looking at the network we can easily conclude that an ERGM which assumes nodal homogeneity is more than questionable. The mERGM, therefore, appears as a possible alternative. The aim of our analysis is to evaluate and compare the two models: the mERGM and the standard ERGM, with the intention to quantify the evidence for the presumed favour of the mERGM.

We fitted four models to the data, two mERGMs and two ERGMs. Table 4 describes the models by listing the sufficient network statistics. As network statistics, we included the number of edges, the number of two-stars and two weighted statistics, i.e. geometrically weighted edgewise shared partners (gwesp) and geometrically weighted nonedgewise shared partners (gwnsp). For the exact definitions of the weighted statistics, we refer to Snijders et al. (2006). The number of iterations for the mERGMs was set at 50 to ensure convergence.

Table 5 shows the resulting estimates for the models. Note that the standard deviations given for the mERGM estimates result from the ERGM fit taking the random effects as fixed. As noted above, these estimates are not reliable as they ignore the uncertainty of the estimated random effects. We, therefore, give these values for completeness only but do not interpret them. We see that the gwesp coefficient is always positive, indicating that the probability for an edge between two partners increases with the number of shared partners for the considered edge. The effect is however generally smaller in the mERGM, that is, if node-specific heterogeneity is taken into account. To make the models comparable, we calculated the AIC values for both the ERGMs and

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Table 5. Model fitting results for the Facebook network data. Standard errors listed in the mERGM are not accurate since they ignore the variability resulting through node heterogeneity

Model type	Model	Parameter	Estimate	SE	AIC
ERGM	1	θ_{edges}	−7.178	0.091	6049.086
		θ_{gwesp}	1.875	0.046	
		$\theta_{2-stars}$	0.052	0.094	
ERGM	2	θ_{edges}	−5.973	1.698	70977.08
		θ_{gwesp}	2.076	0.881	
		θ_{gwnsp}	−0.034	1.211	
mERGM	3	θ_{edges}	−6.021	*0.468*	4260.723
		θ_{gwesp}	1.186	*0.192*	
		$\theta_{2-stars}$	−0.008	*0.001*	
mERGM	4	θ_{edges}	−3.943	*0.698*	2867.874
		θ_{gwesp}	0.486	*0.305*	
		θ_{gwnsp}	−0.053	*0.008*	

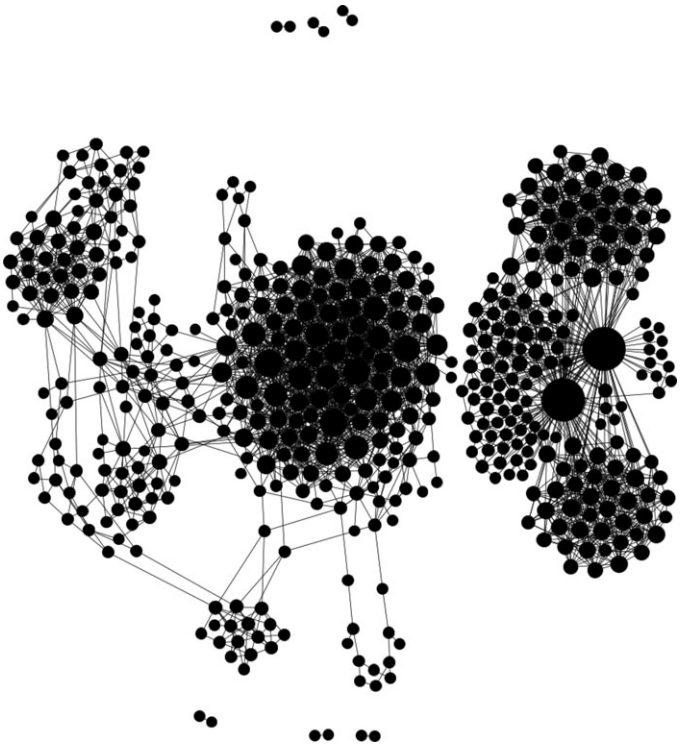


Figure 2. Facebook Network Data. Large nodes indicate nodes with high degrees.

the mERGMs according to the proposed approach as described in Section 3. For the calculation of the AIC values, we used 1000 simulations for both ERGMs and mERGMs, respectively.

Looking at the AIC values of the four models in Table 5, we see that both mERGMs outperform the ERGMs. This gives clear evidence of existing node-specific heterogeneity, and hence the proposed models with nodal random effects are preferable. This is also apparent in the goodness-of-fit plots. For instance, in Figure 3, we can see that model 3 fitted with mERGM gives a better

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

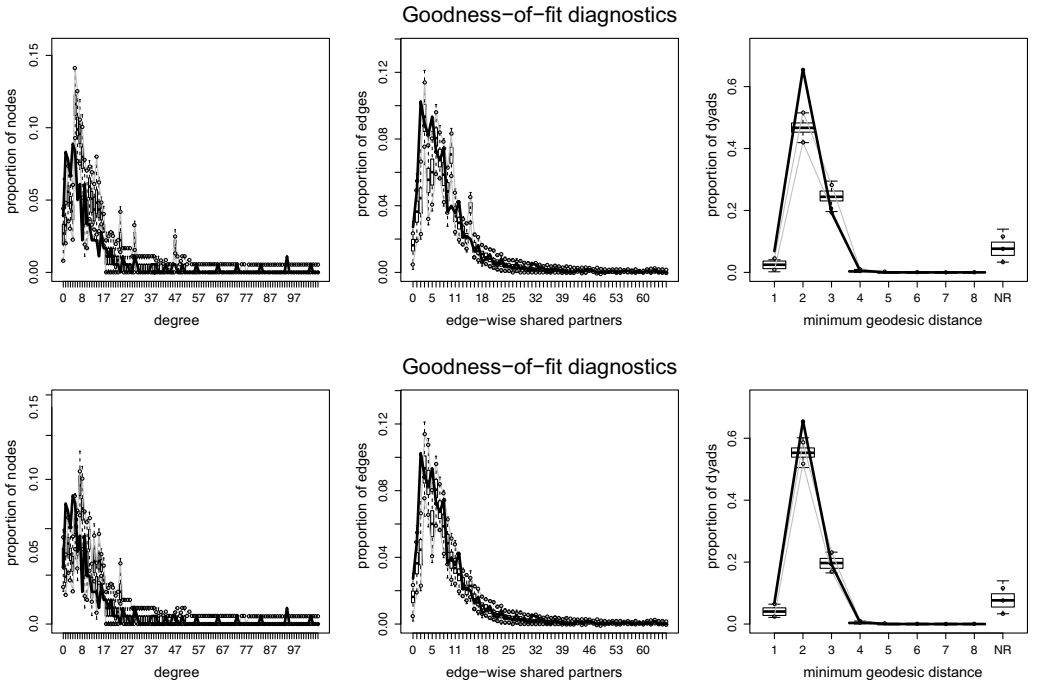


Figure 3. Goodness-of-fit diagnostics for model 1 fitted with ERGM (top row) and for model 3 fitted with mERGM (bottom row) for the Facebook network data.

fit compared to model 1 fitted with ERGM, which is in line with the corresponding AIC values. Furthermore, the AIC value of model 2 in Table 5 fitted with ERGM is exceptionally huge compared to model 4 fitted with mERGM, and we observe this issue well reflected in the goodness-of-fit plots in Figure 4, where model 2 struggles with huge convergence issues. Overall, model 4 appears to be the most suitable among the four fitted models to describe the data.

5.2 Zachary's Karate club network

As a second data example, we look at a well known dataset, the Zachary's karate club (Zachary, 1977). This undirected network data represents the friendship among 34 members of a university karate club. Figure 5 shows the network graph of Zachary's Karate Club. One can easily see that in this network, there are few nodes with high degree, while the remaining nodes have only few edges, so again, we assume that the mERGM should be a suitable approach to capture the unobserved nodal heterogeneity in the network. We fitted three different ERGMs and three mERGMs to this data. To make the models comparable, we included the same network statistics. Table 6 gives an overview of the different models. In Table 7, we summarize the results of our models including the AIC values. The iteration steps for the mERGMs was set to 50. For the calculation of the AIC values, we used 1000 simulations for both ERGMs and mERGMs.

We can see that model 1 fitted with ERGM struggles with convergence issues. This is mirrored in invalid variance estimates, resulting from a badly conditioned Fisher matrix. We, therefore, indicate this as "$\star''$" in Table 7, which also means, of course, that the estimate itself is not reliable at all. We refer to Hunter et al. (2008) for further explanations. On the other hand, model 4 fitted with mERGM with the same model parameters as model 1 does not show any convergence issues, which also means that the mERGM can deal with estimation degeneracy issues. The inclusion of node-specific heterogeneity works towards numerical stabilization. To explore this in more depth,

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Table 6. Models with sufficient network statistics for the Zachary network data

Model type	Model	θ_{edges}	θ_{gwesp}	$\theta_{2-stars}$	θ_{gwnsp}	$\theta_{gwdegree}$	Nodal effects
ERGM	1	✓	✓	✓			
	2	✓	✓		✓		
	3	✓	✓			✓	
mERGM	4	✓	✓	✓			✓
	5	✓	✓		✓		✓
	6	✓	✓			✓	✓

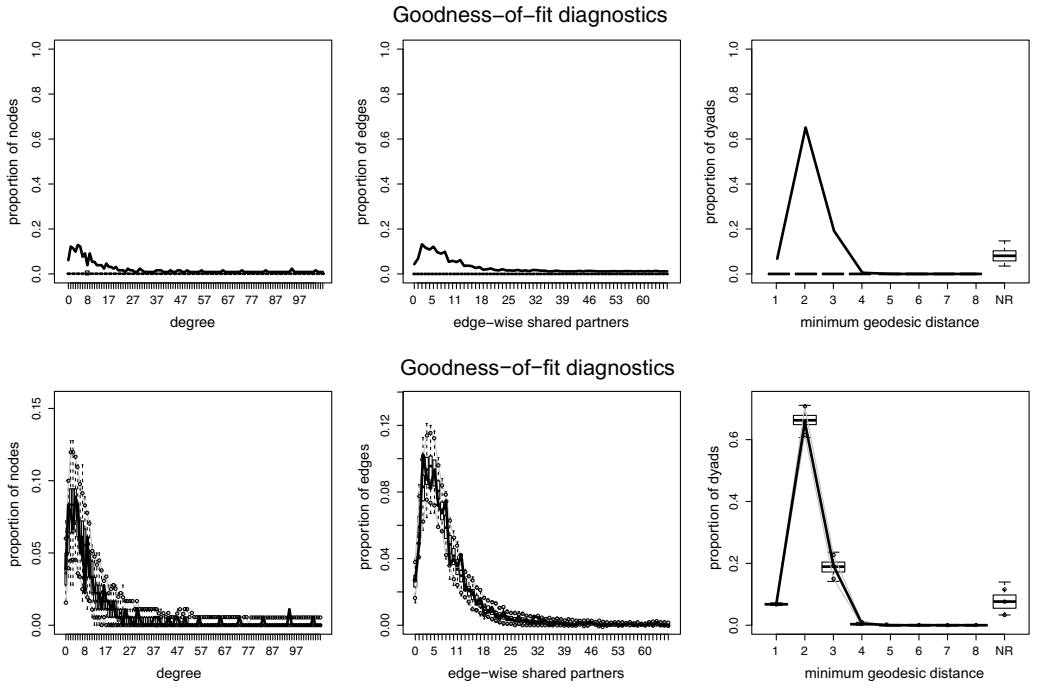


Figure 4. Goodness-of-fit diagnostics for model 2 fitted with ERGM (top row) and for model 4 fitted with mERGM (bottom row) for the Facebook network data.

we look in Figure 6 at the goodness-of-fit plots for model 1 fitted with ERGM. Figure 6 shows the same diagnostics results for model 4 fitted with mERGM. Remember that these two models include the same sufficient network statistics. Boxplots of the distributions of degree, edge-wise shared partners and minimum geodesic distance for the resulting simulated networks are shown in the plots where the bold line indicates the values of the original karate club dataset. In the diagnostics plots of model 1 in Figure 6, we can clearly see that ERGM fails to fit the model, whereas the diagnostics plots of model 4 in Figure 6 gives a good evidence of an appropriate fit.

We extend the model exploration to the other models. Figure 7 shows the goodness-of-fit diagnostics plots of both model 2 fitted with ERGM and model 5 fitted with mERGM, respectively. For model 2, we see some problems in Figure 7 concerning all the three diagnostics, the degree distribution, the edgewise-shared partners distribution and the minimum geodesic distance, which indicate the poorness of the model. For model 4, in contrast, we can see in Figure 7 again a much better performance of the fit.

Finally, Figure 8 shows the diagnostics plots of model 3 fitted with the ERGM, this model is the best ERGM fitted to this data according to the AIC value and also the diagnostics plots

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Table 7. Model fitting results for the Zachary network data. Standard errors listed in the mERGM are not accurate since they ignore the variability resulting through node heterogeneity

Model type	Model	Parameter	Estimate	SE	AIC
ERGM	1	θ_{edges}	−4.893	⋆⋆⋆	⋆⋆⋆
		θ_{gwesp}	0.642	⋆⋆⋆	
		$\theta_{2-stars}$	0.689	⋆⋆⋆	
ERGM	2	θ_{edges}	−3.635	0.241	496.351
		θ_{gwesp}	0.596	0.117	
		θ_{gwnsp}	0.153	0.026	
ERGM	3	θ_{edges}	−3.183	0.513	442.883
		θ_{gwesp}	0.716	0.181	
		$\theta_{gwdegree}$	−0.519	0.8990	
mERGM	4	θ_{edges}	−1.214	*0.386*	311.304
		θ_{gwesp}	0.236	*0.174*	
		$\theta_{2-stars}$	−0.159	*0.042*	
mERGM	5	θ_{edges}	−1.776	*0.366*	337.894
		θ_{gwesp}	−0.144	*0.109*	
		θ_{gwnsp}	−0.089	*0.043*	
mERGM	6	θ_{edges}	−4.464	*0.724*	303.217
		θ_{gwesp}	0.213	*0.176*	
		$\theta_{gwdegree}$	4.427	*1.341*	

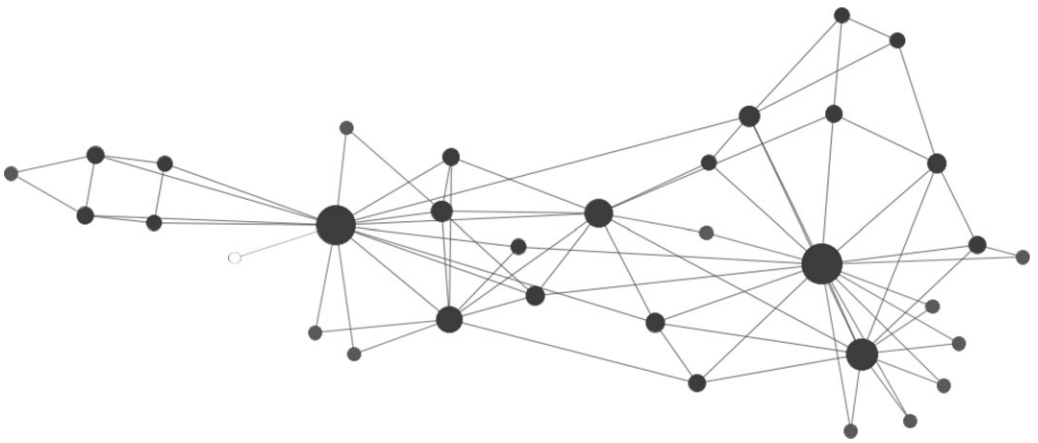


Figure 5. Zachary's Karate Club Network Data. Large nodes indicate nodes with high degrees.

are reasonable. On the other hand, model 6 fitted with mERGM, including the same sufficient network statistics as model 3, is the best mERGM fitted to this data according to the AIC value. However, the goodness-of-fit of model 6 shown in Figure 8 visually looks better than of model 3, which also justifies with a smaller AIC value.

5.3 High school friendship network

As a third data example, we investigate a real-world dataset, a high school friendship network in Marseilles, France, provided by Mastrandrea et al. (2015). This network data represents the friendship among 134 high school students of specific classes. These specific classes, unique to

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

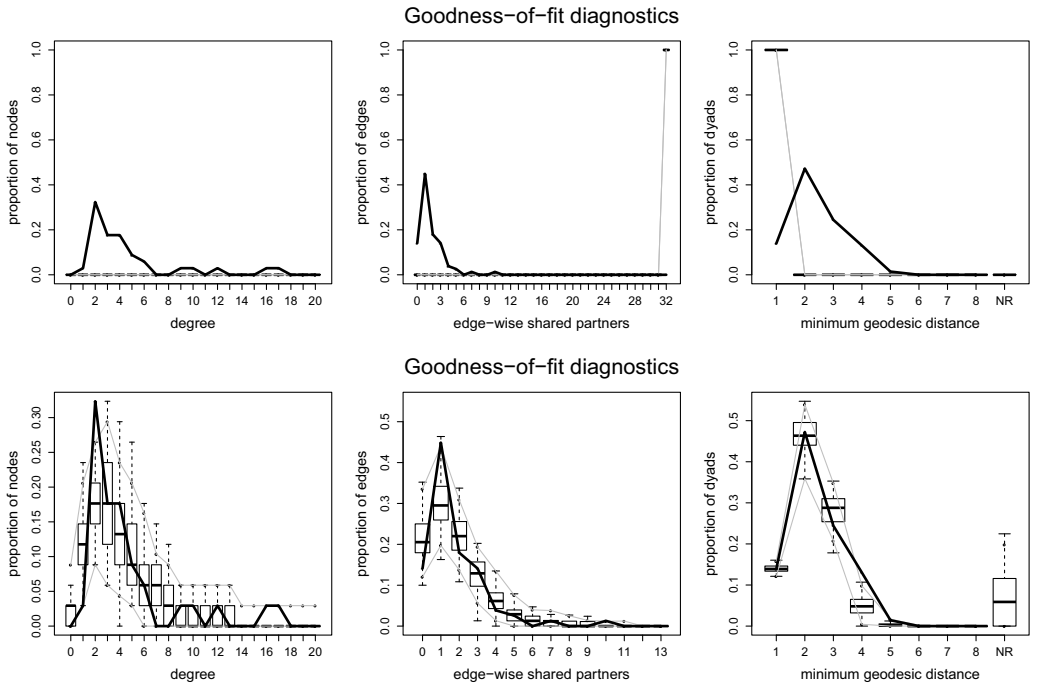


Figure 6. Goodness-of-fit diagnostics for model 1 fitted with ERGM (top row) and for model 4 fitted with mERGM (bottom row) for the karate club data.

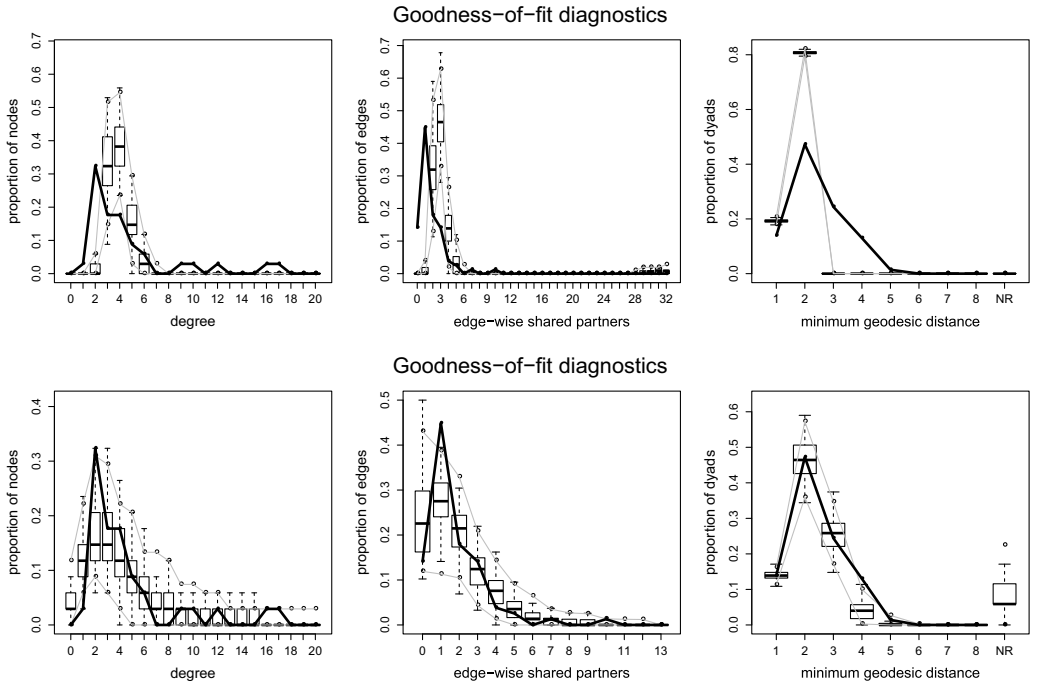


Figure 7. Goodness-of-fit diagnostics for model 2 fitted with ERGM (top row) and for model 5 fitted with mERGM (bottom row) for the karate club data.

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Table 8. Number of students according to the types of the classes in the data

Classes	MP	MP*1	MP*2	PC	PC*	PSI*	2BIO1	2BIO2	2BIO3
No. of Students	21	3	7	21	10	15	10	19	28

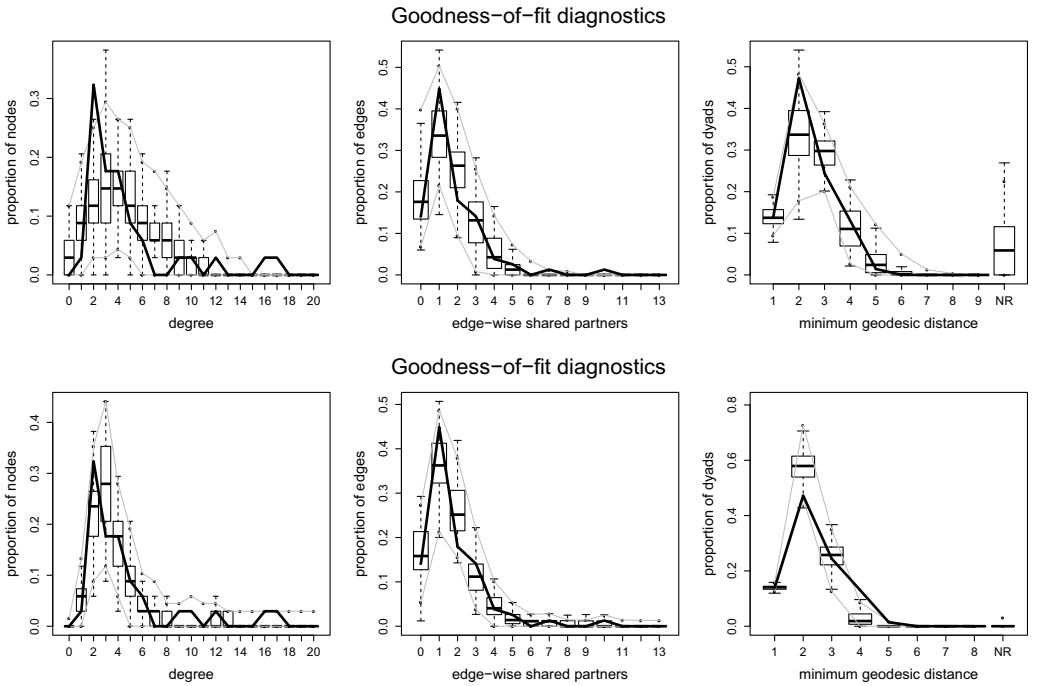


Figure 8. Goodness-of-fit diagnostics for model 3 fitted with ERGM (top row) and for model 6 fitted with mERGM (bottom row) for the karate club data.

the French educational system, gather students for studies that last two years after completing the usual high school studies. These classes are in a different part of the building, so the students are somehow isolated from the "regular" high school students. As a result, they form an almost closed population with little contact with the outside world, at least during schooldays. At the end of these 2 years, these students take competitive exams to gain admission to various higher educational institutions. The classes have different specializations: "MP" classes focus more on mathematics and physics, "PC" classes on physics and chemistry, "PSI" classes on engineering, and "BIO" classes on biology. Furthermore, there are three classes of type "MP", two of type "PC", one of type "PSI" and three of type "BIO". Due to the class sizes in the dataset, as demonstrated in Table 8, we decided to merge the types of classes to get an appropriate fit for our models, e.g., we do not distinguish between the types of the classes.

Additionally, we also have the gender information of the students in the data, with 79 female and 55 male students. Figure 9 shows the friendship network graph, the colour of the nodes represents the different classes, where the shape of the nodes the gender of the students. Assuming that there might be unobserved nodal heterogeneity, which can not be captured exclusively by the nodal covariates, the mERGM could be a reasonable approach.

We fit an ERGM and a mERGM to this data, including dyad-dependent network statistics such as the `GWESP` (geometrically weighted edgewise shared partner distribution) with a fixed decay parameter equal to 0.25 and dyad-independent terms such as the `nodefactor` and `nodematch` parameters. The `nodematch` is the homophily parameter in ERGM, where we also allow for each

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Table 9. Model estimates for the high school friendship network. Standard errors listed in the mERGM are not accurate since they ignore the variability resulting through node heterogeneity

Model type	Parameter	Estimate	SE	Model type	Parameter	Estimate	SE
	Edges	−7.606	0.403		Edges	−9.021	*0.415*
	GWESP (0.25)	2.026	0.165		GWESP (0.25)	0.538	*0.215*
	Male	0.157	0.052		Male	−0.494	*0.068*
	Gender match	0.226	0.089		Gender match	0.813	*0.099*
	MP	0.627	0.272		MP	2.355	*0.238*
ERGM	PC	0.309	0.272	mERGM	PC	1.253	*0.238*
	PSI	1.054	0.276		PSI	3.042	*0.221*
	BIO class match	2.562	0.402		BIO class match	4.403	*0.295*
	MP class match	1.547	0.337		MP class match	2.814	*0.349*
	PC class match	2.279	0.393		PC class match	3.572	*0.379*
	PSI class match	1.504	0.379		PSI class match	2.047	*0.355*
AIC	3880.681			AIC	1097.836		

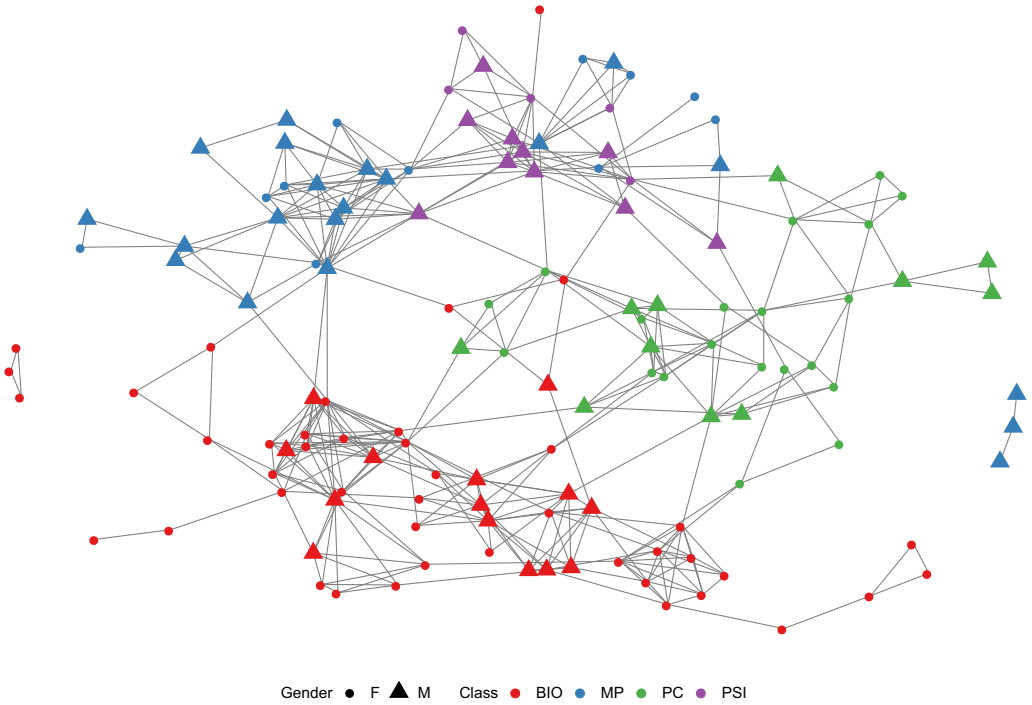


Figure 9. High School Friendship Network Data. The Colour of the nodes represents the different classes. The shape of the nodes describes the gender of the students.

class to have a unique propensity for within-class ties, we refer to Morris et al. (2008) for instance, for more details regarding the ERGM terms and their specifications. The iteration steps for the mERGMs was set to 50. For the calculation of the AIC values, we used 1000 simulations for both ERGMs and mERGMs. In Table 9, we summarize the results of the fitted models. Comparing the

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

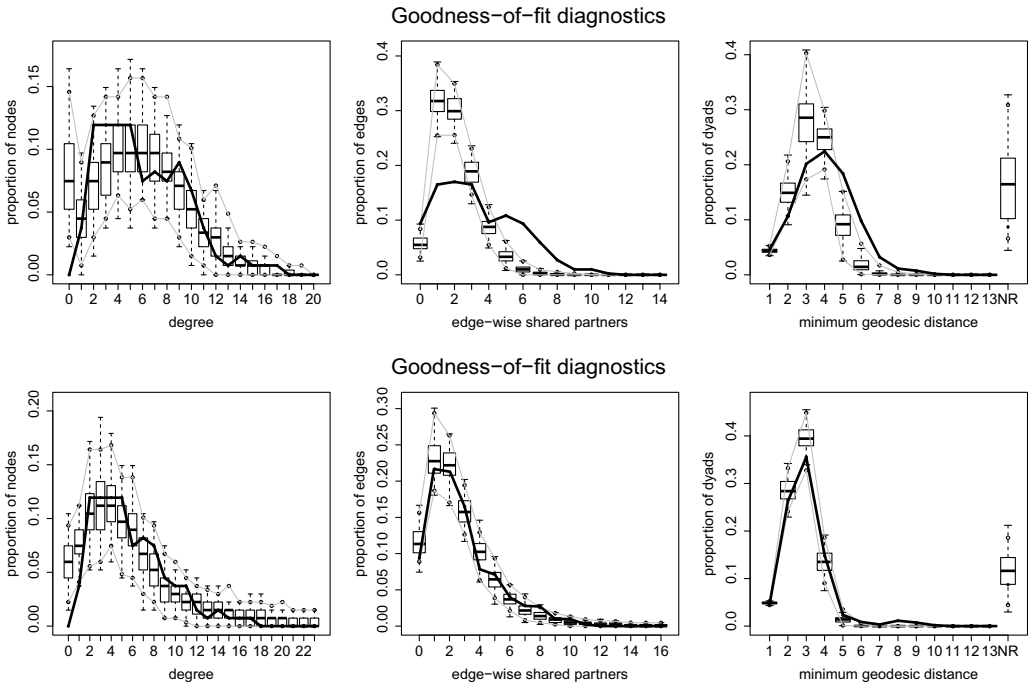


Figure 10. Goodness-of-fit diagnostics for the model fitted with ERGM (top row) and for the model fitted with mERGM (bottom row) for the high school friendship network data.

AIC values of the two models, the model fitted with mERGM is preferred, which is not very surprising since the mERGM additionally takes the "unobserved" nodal-heterogeneity into account, which is not captured by dyad-dependent or independent terms, aka network statistics. For the sake of interpretation, we look, for instance, at the log-odds of a hypothetical tie between two male students attending the MP class that does not close a triangle. In ERGM, the log-odds of such a tie would be -4.264; in mERGM, the interpretation is slightly different since it should be only in the conditional sense. This means, given two male students that attend the MP class, the conditional log-odds of such a tie is -1.671. In Figure 10, we show the diagnostic plots of the two models, we can clearly see that the model fitted with the mERGM performs much better, which is also clear evidence to the corresponding AIC values.

6. Discussion

In most cases, nodal heterogeneity in the network is explained by including known or well-studied nodal covariates, see e.g. Robins et al. (2001). However, the node-specific covariates cannot fully or sufficiently account for unobserved heterogeneity in the network. Our extensions towards mixed ERGMs can therefore be a meaningful approach to model network data by just adding nodal random effects to the model to capture the unobserved nodal heterogeneity.

Our proposed model (4) can in principle also be extended to directed networks, where $u_i^{(s)}$ and $u_j^{(r)}$ would be treated as random sender and random receiver node-specific coefficients, respectively. We consider it beyond the scope of the current paper.

Though the calculation of the AIC value is computationally intensive, our proposed method of estimating and calculating the AIC values allows us to compare the mERGM with the conventional ERGM. Furthermore, as we can see in our simulation study, the mERGM can always be a

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

reasonable approach for modelling networks even if we observe small nodal heterogeneity in the network. Overall, the mERGM works towards stabilizing the fitting routine without adding too much numerical effort.

In this study, neither in the simulation study nor in the examples, we investigate assortative mixing parameters. Therefore, a more thorough investigation in this regard is worth pursuing in future work.

Acknowledgments. The project was supported by the European Cooperation in Science and Technology [COST Action CA15109 (COSTNET)].

Funding. This work has been funded by the German Federal Ministry of Education and Research (BMBF) under grant no. 01IS18036A. The authors of this work take full responsibility for its content.

Competing interests. None.

Ethical statement. The authors also confirm that each author has seen and approved the contents of the submitted manuscript.

Supplementary. For supplementary material for this article, please visit http://dx.doi.org/10.1017/nws.2021.22.

References

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike* (pp. 215–222). Springer.

Box-Steffensmeier, J. M., Christenson, D. P., & Morgan, J. W. (2018). Modeling unobserved heterogeneity in social networks with the frailty exponential random graph model. *Political Analysis*, 26(1), 3–19.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421), 9–25.

Caimo, A., & Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1), 41–55.

Cranmer, S. J., & Desmarais, B. A. (2011). Inferential network analysis with exponential random graph models. *Political Analysis*, 19(1), 66–86.

Desmarais, B., & Cranmer, S. (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications*, 391(4), 1865–1876.

Duijn, M. A. J., Snijders, T. A., & Zijlstra, B. J. H. (2004). p_2: A random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58(2), 234–254.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC Press.

Fellows, I., & Handcock, M. S. (2012). Exponential-family random network models. Retrieved from arXiv preprint arXiv:1208.0121.

Fienberg, S. E. (2012). A Brief History of Statistical Models for Network Analysis and Open Challenges. *Journal of Computational and Graphical Statistics*, 21(4), 825–839.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395), 832–842.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., & Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2), 129–233.

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, 97(4), 773–789.

Henry, T. R., Gates, K. M., Prinstein, M. J., & Steinley, D. (2020). Modeling heterogeneous peer assortment effects using finite mixture exponential random graph models. *Psychometrika*, 85(1), 8–34.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373), 33–50.

Hummel, R. M., Hunter, D. R., & Handcock, M. S. (2012). Improving simulation-based algorithms for fitting ergms. *Journal of Computational and Graphical Statistics*, 21(4), 920–939.

Hunter, D. R. & Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3), 565–583.

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), 1–29.

Hunter, D. R., Krivitsky, P. N., & Schweinberger, M. (2012). Computational statistical methods for social network analysis. *Journal of Computational and Graphical Statistics*, 21(4), 856–882.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Models*. New York: Springer.

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press

Kolaczyk, E. D. & Csárdi, G. (2014). *Statistical analysis of network data with R*. New York: Springer.
Koskinen, J. (2009). Using latent variables to account for heterogeneity in exponential family random graph models.
Lusher, D., Koskinen, J., & Robins, G. (2013). *Exponential random graph models for social networks*. Cambridge: Cambridge University Press.
Mastrandrea, R., Fournet, J., & Barrat, A. (2015). Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one*, 10(9), e0136497.
McAuley, J. J., & Leskovec, J. (2012). In *NIPS*, volume 2012. Citeseer (pp. 548–656).
Morris, M., Handcock, M. S., & Hunter, D. R. (2008). Specification of exponential-family random graph models: terms and computational aspects. *Journal of statistical software*, 24(4), 1548.
Robins, G. L., Elliot, P., & Pattison, P. (2001). Network models for social selection processes. *Social Networks*, 23(1), 1–30.
Schmid, C. S. & Desmarais, B. A. (2017). Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE (pp. 116–121).
Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496), 1361–1370.
Schweinberger, M. and Handcock, M. S. (2015). Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the American Statistical Association*, 77(3), 647.
Schweinberger, M., Krivitsky, P. N., & Butts, C. T. (2017). Foundations of finite-, super-, and infinite-population random graph inference. arXiv preprint arXiv:1707.04800.
Snijders, T. A. B. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2), 1–40.
Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1), 99–153.
Strauss, D. & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409), 204–212.
Thiemichen, S., Friel, N., Caimo, A., & Kauermann, G. (2016). Bayesian exponential random graph models with nodal random effects. *Social Networks*, 46, 11–28.
Vaida, F. & Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, 92(2), 351–370.
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.
Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1), 3–36.
Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4), 452–473.
Zijlstra, B. J. H., Duijn, M. A. J., & Snijders, T. A. B. (2006). The multilevel p_2 model: A random effects model for the analysis of multiple social networks. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(1), 42–47.

Cite this article: Kevork S. and Kauermann G. (2022). Iterative estimation of mixed exponential random graph models with nodal random effects. *Network Science* **9**, 478–498. https://doi.org/10.1017/nws.2021.22

https://doi.org/10.1017/nws.2021.22 Published online by Cambridge University Press