Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

23-10

Marginalized Curved Exponential Random Graph Models

Thomas Suesse

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Marginalized Curved Exponential Random Graph Models

Thomas Suesse[1]

Centre for Statistical and Survey Methodology

School of Mathematics and Applied Statistics

University of Wollongong

NSW 2522, Australia

email: tsuesse@uow.edu.au

tel: (61) 2 4221 4173; fax: (61) 2 4221 4845

[1]Thomas Suesse is a postdoctoral research fellow at the Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia (Email: *tsuesse@uow.edu.au*)

Abstract

Curved Exponential random graph models (CERGMs) are a popular tool for modeling social networks representing relational data, such as working relations or friendships. Additionally to the network itself some exogenous variables are often collected, such as gender, age, etc. CERGMs allow modeling of the effects of such exogenous variables on the joint distribution, but not on the marginal probabilities of observing a relation. In this paper, we consider a modification of CERGMs that uses a CERGM to model the joint distribution of a network, which is then subject to a marginal logistic regression model for the marginal probabilities of observing a relation. Explanatory variables depend on the exogenous variables, such as the difference in age between two nodes. This model approach, termed a marginalized CERGM, is a natural extension of a CERGM that allows a convenient interpretation of parameters in terms of log odds ratios. Several algorithms to obtain ML estimates and solutions to reduce the computational burden are presented. The methods are illustrated using the working relations between 36 partners in a New England law firm.

Key Words: Social network; Exponential random graph model; Marginalized models; Odds ratio; Markov chain Monte Carlo; Maximum Likelihood

1 Introduction

Networks, or mathematical graphs, are a useful tool to represent relational data between interacting actors or nodes. In particular, social relationships occurring in everyday life, such as best friends, can be represented in its simplest form in a network, often called *social network*. Modeling of social networks and their inherent social structure has become increasingly important in a number of fields, for example anthropology, biology, communication studies, economics, geography, information science, organizational studies, social psychology, and sociolinguistics.

In a network each actor or node represents a social group or individual and each link represents the presence of a relationship between two actors. The network consists of the set of nodes and the set of edges, where an edge is said to exist between two nodes, if there is a link between the two nodes.

The network can be represented by a $n \times n$ matrix $\mathbf{Y} = (Y_{ij})_{i,j=1}^{n}$, where n refers to the number of nodes and Y_{ij} is a binary indicator, which is one if an edge or link exists between nodes i and j and zero otherwise. A node has no link to itself, i.e. $Y_{ii} = 0$ for all nodes i. The pairs of nodes are often called *dyads*. Networks can be directed or undirected, the latter is characterized by a symmetric matrix $\mathbf{Y}$, i.e. $Y_{ij} = Y_{ji}$. For example the network formed by the social relationship "best friends" would be undirected, because both parties would need to confirm being best friends to each other, whereas the network formed by the transmission of a disease is certainly directed, because if i is infected by j then j was already infected and cannot be infected by i ($Y_{ij} = 1$ but $Y_{ji} = 0$). The number of dyads in an undirected network is $N = \binom{n}{2}$, and in an directed network $N = 2\binom{n}{2}$. Theoretically the matrix $\mathbf{Y}$ can contain several relationships, but for simplicity, we only consider binary variables Y_{ij} and undirected graphs $Y_{ij} = Y_{ji}$ in this article.

Exponential random graph models (ERGMs) (Holland and Leinhardt 1981; Strauss and Ikeda 1990; Snijders 2002; Hunter and Handcock 2006) (the last reference abbreviated as HH06) are currently the most popular statistical models for social networks. The first model of this class was proposed by Holland and Leinhardt (1981), called the p_1 or $p*$ model, which treats all dyads as independent. In general, maximum likelihood (ML) estimation is complicated because the sample space is of size $\exp(\binom{n}{2} \log 2) = 2^N$, an incredible large number even for small n (say $n = 10$ gives $N = 2^{45} = 3.518437 \cdot 10^{13}$). For convenience we will also use $\mathbf{Y}$ to denote the vector $(Y_{12}, Y_{13}, \ldots, Y_{n-1,n})^T$ and the associated realized vector is $\mathbf{y} = (y_{12}, y_{13}, \ldots, y_{n-1,n})^T$.

The probability distribution of an ERGM is of the following form

$$\Pr(\mathbf{Y} = \mathbf{y}; \boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^T \mathbf{Z}(\mathbf{y}) - \kappa(\boldsymbol{\eta})\right), \tag{1}$$

which is a canonical exponential family model (Barndorff-Nielsen 1978). Here $\mathbf{y}$ stands for the observed network and $\mathbf{Y}$ for the random vector. The vector $\mathbf{Z}(\mathbf{y}) \in \mathbb{R}^q$ consists of network statistics, $\boldsymbol{\eta} \in \mathbb{R}^q$ is the corresponding canonical parameter vector and κ is the familiar normalizing constant associated with this distribution. Curved ERGMs (CERGMs) are an extension of ERGMs where $\boldsymbol{\eta}$ is replaced by $\boldsymbol{\eta}(\boldsymbol{\theta})$, a mapping from $\mathbb{R}^p$ to $\mathbb{R}^q$ with $p < q$. The distribution of a CERGM is a member of the curved exponential family.

The vector of statistics $\mathbf{Z}(\mathbf{y})$ induces a dependence structure and its choice depends on the network data and the associated research question. The number of potential network statistics is huge, see Morris et al. (2008) for a detailed description of network statistics available in the `R` (R-Development-Core-Team 2006) package `ergm` (Handcock et al. 2010).

Until recently CERGMs have been fitted by a pseudo-ML (PML) method (Strauss and Ikeda 1990), a naive approach that has been criticized (Wasserman and Robins 2004). The properties of such estimates are not well understood. An alternative to pseudo-ML is a stochastic approximation of the true log-likelihood based on Markov chain Monte Carlo (MCMC) algorithms. Snijders (2002) considered a Robbins-Monroe-type algorithm (Robbins and Monroe 1951), which aims to find a solution of the likelihood equations. HH06 discussed, in detail, ML estimation for CERGMs, which is based on a stochastic approximation of the likelihood using MCMC methods. Such estimates are then called MCMCMLE (MCMC ML estimates).

The statistic $\mathbf{Z}$ may also contain exogenous variables, for example node attributes denoted by $\mathbf{X}_i$. Consider the statistic

$$\sum_{i<j}^{n} y_{ij} f(\mathbf{X}_i, \mathbf{X}_j), \tag{2}$$

where $f(\cdot)$ is a symmetric function of the nodal covariate vectors $\mathbf{X}_i$ and $\mathbf{X}_j$. As an example

consider a network of collaborative working relations between 36 partners in a New England law firm described in detail by Lazega and Pattison (1999) and Lazega (2001). An edge between two partners exists if both partners indicate collaboration with each other. The data also contain a number of attributes of each partner: seniority (rank number of entry into the firm), practice (litigation=0, corporate law=1), gender and office (3 offices in different cities). The data has been analyzed by HH06 using main effects: seniority and practice defined by $f_1(\mathbf{X}_i, \mathbf{X}_j) = \text{seniority}_i + \text{seniority}_j$ and $f_2(\mathbf{X}_i, \mathbf{X}_j) = \text{practice}_i + \text{practice}_j$, and similarity effects of practice, gender and office, where similarity of e.g. gender is defined by $f_4(\mathbf{X}_i, \mathbf{X}_j) = I(\text{gender}_i = \text{gender}_j)$ ($I(\cdot)$ is an indicator function), similarly f_3 and f_5 for similarity effects of practice and office.

A reasonable approach to model the probability $\Pr(Y_{ij} = 1)$, ignoring the induced dependence by the social relations, would be to apply a logistic regression model

$$\text{logit}\left(\Pr(Y_{ij} = 1)\right) = \beta_0 + \sum_{k=1}^{5} \beta_k f_k(\mathbf{X}_i, \mathbf{X}_j), \tag{3}$$

and naively assume independence between all nodes. Such a model can be fitted by standard statistical software that can fit logistic regression models. The resulting parameter estimates for the logistic regression model can be found in Table 1 along with estimates of the log-odds ratio. The advantage of this approach is that it allows marginal interpretation of parameters, for example the odds of collaboration of partners of the same gender are $\exp(\beta_4) = \exp(1.128) = 3.09$ times higher than the odds of collaboration of partners of different gender.

Such a logistic model is very useful for social scientists, because of its frequent use and the familiarity of the odds ratio in making interpretations about marginal probabilities. Model (3) is equivalent to an ERGM when the vector $\mathbf{Z}(\mathbf{y})$ consists of the edges statistic and the statistics of exogenous variables of form (2) with previously defined f_k, $k = 1, 2, 3, 4, 5$, such that $\boldsymbol{\eta} = \boldsymbol{\beta}$. Unfortunately, adding typical network statistics to $\mathbf{Z}(\mathbf{y})$ to account for

variables	estimates – β_k (s.e)	p-value	odds-ratio – $\exp(\beta_k)$
intercept	-8.306 (0.953)	$< 2e-16$	0.002
main seniority	0.044 (0.009)	$8.9e-07$	1.045
main practice	0.902 (0.163)	$3.1e-08$	2.464
sim practice	0.879 (0.231)	0.00014	2.408
sim gender	1.128 (0.348)	0.00121	3.089
sim office	1.653 (0.254)	$7.6e-11$	5.222

Table 1: Logistic regression for Law Firm Data Set

the network dependence does not allow such a marginal interpretation of the associated parameters ($\boldsymbol{\eta} \neq \boldsymbol{\beta}$).

In this article, we propose a unifying approach that models the joint distribution by a CERGM to account for the induced dependence structure subject to the marginal logistic model (3). This approach, here termed marginalized CERGM, is similar to the approach considered by Fitzmaurice and Laird (1993), labeled in the remainder as FL93, who modeled multivariate binary data using a log-linear model subject to a marginal model. Their approach is simpler because the log-likelihood can be computed exactly and involves only relatively simple statistics - higher order statistics of a log-linear model. Here we need to approximate the log-likelihood by MCMC algorithms and the computation of network statistics also induces more complexity. Marginalized models are common in statistics, other examples are marginalized generalized linear mixed models (Heagerty 1999; Wang and Louis 2004) and generalized estimating equations (Liang and Zeger 1986).

In the next section, we outline the theory behind CERGMs and show how the parameters can be used to make conditional, but not marginal, interpretations. In Section 3, we introduce the marginalized CERGM and derive likelihood equations using a Fisher-scoring scheme. However, ML estimation is even more complicated for a marginalized CERGM than for a CERGM. In fact, to apply an iteration of the Fisher-scoring scheme, we need to solve another set of equations. Section 4 describes the details of ML estimation and gives two alternative methods to solve the set of equations in each step. The next section revisits

the Lazega (2001) data set and gives parameter estimates for our proposed model, comparing estimates with those of logistic regression and CERGM. This article finishes with a discussion.

2 Curved Exponential Random Graph Models

The distribution of an exponential random graph model (ERGM) is specified by (1). Some examples of network statistics are: the number of *edges* denoted by $E(\mathbf{y}) = \sum_{1\leq i<j\leq n} y_{ij}$, the number of nodes of *degree* i denoted by $D_i = \sum_{j=1}^{n} y_{ij}, i = 1, \ldots, n-1$, and the *edgewise shared partner statistic* with k common neighbors denoted by $EP_k, k = 0, \ldots, n-2$, which is the number of edges $(i, j)(i < j)$ (implies i and j must be neighbors of each other, i.e. $y_{ij} = 1$), that share exactly k neighbors in common. Two actors i and j are said to share a neighbor l if $y_{il} = 1$ and $y_{jl} = 1$.

An ERGM can be characterized by the vector of network statistics $\mathbf{Z}(\mathbf{y}) = (E, D_1, \ldots, D_{n-1}, EP_0, \ldots, EP_{n-2})^T$ and the associated parameter vector $\boldsymbol{\eta}$ then contains $1+(n-1)+(n-1) = 2n-1$ parameters. Exogenous variables can also be added, using the network statistics in (2), for example main effects for practice and seniority, and similarity effects for gender, practice and office, as described previously. This adds five more parameters to the model.

Consider the model with $\mathbf{Z}(\mathbf{y}) = (E, \sum_{ij} f_1(\mathbf{X}_i, \mathbf{X}_j)y_{ij}, \cdots, \sum_{ij} f_5(\mathbf{X}_i, \mathbf{X}_j)y_{ij})$

$$\Pr(\mathbf{Y} = \mathbf{y}; \boldsymbol{\eta}) = \exp\left(\eta_0 E(\mathbf{y}) + \sum_{k=1}^{5}\sum_{ij} \eta_k f_k(\mathbf{X}_i, \mathbf{X}_j)y_{ij} - \kappa(\boldsymbol{\eta})\right), \tag{4}$$

with f_k defined above. This model can be re-expressed as a standard logistic model (3) with $\eta_k = \beta_k, k = 0, 1, 2, \ldots, 5$ and with dyads representing independent binary observations. Therefore, for this ERGM, the parameters can be used to make inference about marginal probabilities. However, including typical complex network statistics, such as D_i and EP_i, to account for network dependence, parameters η_k (ERGM) and β_k (logistic model) associated

with $f_k(\mathbf{X}_i, \mathbf{X}_j)$ are different.

The parameters η_{ij} in $\boldsymbol{\eta}$ only allow a conditional interpretation. Consider a dyad Y_{ij} and let the "rest of the graph" (without Y_{ij}) be denoted by Y_{ij}^c. Then

$$\frac{\Pr(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{\Pr(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \exp\left(\boldsymbol{\eta}^T \Delta(\mathbf{Z}(\mathbf{y}))_{ij}\right), \tag{5}$$

where $\Delta(\mathbf{Z}(\mathbf{y}))_{ij}$ denotes the difference in $\mathbf{Z}(\mathbf{y})$ between $y_{ij} = 1$ and $y_{ij} = 0$ while Y_{ij}^c remains fixed. Given the $\Delta(\mathbf{Z}(\mathbf{y}))_{ij}$, the parameters η_{ij} have a conditional interpretation, which might be useful in certain instances, but generally is of very limited use only when we want to make predictions without conditioning on all except one link.

Recently, ERGMs have been extended to CERGMs (HH06), which have the following form

$$\Pr(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}) = \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{Z}(\mathbf{y}) - \kappa\left\{\boldsymbol{\eta}(\boldsymbol{\theta})\right\}\right), \tag{6}$$

which is technically a member of the curved exponential family. Here $\boldsymbol{\eta}(\boldsymbol{\theta})$ is a mapping from $\mathbb{R}^p$ to $\mathbb{R}^q$ with $p < q$. A number of useful statistics have been introduced recently (Snijders et al. 2006, HH06), for example the geometrically weighted edgewise shared partner statistic (GWESP)

$$\exp(\theta_2) \sum_{i=1}^{n-2} \left\{1 - \left(1 - \exp(-\theta_2)^i\right)\right\} EP_i(\mathbf{y}), \tag{7}$$

the geometrically weighted degree statistic and the geometrically weighted dyadwise shared partner statistic. An ERGM with the statistics EP_i, $(i = 0, \ldots, n-2)$ has $n-2$ parameters $\eta_0, \ldots, \eta_{n-2}$. Using the GWESP statistic, the $n-1$ parameters $\eta_0, \ldots, \eta_{n-2}$ depend on two parameters θ_1 and θ_2 only, defined by the relationship

$$\eta_i = \theta_1 \exp(\theta_2) \left\{1 - \left(1 - \exp(-\theta_2)^i\right)\right\}.$$

When using these statistics, the model is not of form (1) but of form (6). Fitting a CERGM is difficult and similar to fitting an ERGM.

The pseudo-ML (PML) method assumes $\Pr(Y_{ij} = 1|Y_{ij}^c = y_{ij}^c) \approx \Pr(Y_{ij} = 1)$ and that dyads Y_{ij} are independent, in which case (5) becomes

$$\text{logit}\left(\Pr(Y_{ij} = 1)\right) = \boldsymbol{\eta}^T \Delta(\mathbf{Z}(\mathbf{y}))_{ij}$$

This is a standard logistic regression model and standard software can be used to obtain ML estimates for logistic regression. PML has been proposed by many authors (Frank and Strauss 1986; Strauss and Ikeda 1990; Frank 1991), but the properties of such estimates are not well understood. ML and PML have been compared by Corander et al. (1998) and van Duijn et al. (2009) among others. Wasserman and Robins (2004) reported that PML estimates are biased and intrinsically highly dependent on the observed network. A simulation study by van Duijn et al. (2009) showed that PML estimates are biased and that a bias correction method reduced this bias. The authors reported a strong under-coverage of the 95% confidence interval of the PML method, whereas the coverage of the ML method was nearly 95%.

Table 2 shows parameter estimates for the Lazega (2001) data set for a CERGM with the same variables as the logistic model, but with extra parameters θ_1 and θ_2 referring to the two parameters associated with the GWESP statistic. Comparing Tables 1 and 2 shows that main effects and similarity effects are different. Parameters of a CERGM do not have the convenient interpretation of odds ratios, as illustrated by the additional column $\exp(\beta_k)$ in Table 1, which is omitted in Table 2. The results for the CERGM are based on conditioning on the sufficient statistics for the edges parameter and are obtained from HH06, because `R` package `ergm` (Handcock et al. 2010) fails to fit the full model.

3 Marginalized Curved ERGM

As described above, when using only exogenous variables and the edges statistic E, the CERGM is equivalent to a logistic regression model with convenient interpretation of pa-

Variables	Covariates with GWESP estimates (s.e)	p-value
edges	– (–)	–
main seniority	0.023 (0.006)	0.00012
main practice	0.390 (0.117)	0.00085
sim practice	0.757 (0.194)	$9.5e-05$
sim gender	0.688 (0.248)	0.00553
sim office	1.123 (0.194)	$7.1e-09$
GWESP (θ_1)	0.878 (0.279)	0.00164
GWESP (θ_2)	0.814 (0.196)	$3.2e-05$

Table 2: ERGM estimates for Law Firm Data Set

rameters that allows prediction of marginal probabilities $\Pr(Y_{ij} = 1)$. When adding more complex network statistics, then the marginal relationship

$$\pi_{ij}(\boldsymbol{\beta}) := \text{logit}(\Pr(Y_{ij} = 1)) = \beta_0 + \sum_{k=1}^{K} \beta_k f_k(\mathbf{X}_i, \mathbf{X}_j), \tag{8}$$

does not hold anymore. Instead, the marginal probabilities are determined by the joint distribution and can only be approximated by simulating from the CERGM. The relationship between marginal probabilities and parameters of an CERGM cannot be easily expressed by a simple formula, such as formula (8).

We pursue another approach by assuming that the marginal edge probabilities $\Pr(Y_{ij} = 1)$ are governed by equation (8) for K exogenous variables re-expressed in the more compact form by

$$\text{logit}(\boldsymbol{\pi}) = \tilde{\mathbf{X}}\boldsymbol{\beta} =: \boldsymbol{\nu} \tag{9}$$

with $N \times (K+1)$ design matrix $\tilde{\mathbf{X}}$ ($N = \binom{n}{2}$ dyads and K exogenous variables plus intercept) and $\boldsymbol{\pi} := (\pi_{12}, \ldots, \pi_{n,n-1})^T$. Any other standard link function $g(\cdot)$ for a binary variable, such as the probit link or complementary log-log are also possible.

The distribution of $\mathbf{Y}$ is characterized by a CERGM of the form

$$\Pr(\mathbf{Y} = \mathbf{y}) = \exp\left(\boldsymbol{\Psi}^T\mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbf{Z}(\mathbf{y}) - \kappa\left\{\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{\Psi})\right\}\right), \tag{10}$$

where $\boldsymbol{\Psi} = (\Psi_{12}, \ldots, \Psi_{n-1,n})^T$ is a vector of additional parameters and $\mathbf{Z}(\mathbf{y})$ contains network statistics, but no first order statistics, such as edges or exogenous variables, due to identifiability constraints. Equations (9) and (10) describe the full network model. Such an approach is quite common in statistics, for example generalized estimating equations (Liang and Zeger 1986) and marginalized models for categorical data (FL93, Wang and Louis 2004; Lee and Mercante 2010).

The vector $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\theta}, \boldsymbol{\beta})$ depends on model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ implicitly, such that marginal probabilities follow equation (9). The parameters of $\boldsymbol{\Psi}$ can be considered as nuisance parameters.

This approach has the advantage of easy and direct interpretation of edge probabilities, but the disadvantage of being highly complex. In this article we focus on the estimation of such a marginalized CERGM (denoted by MCERGM) following HH06, who describe ML estimation for CERGMs.

4 Likelihood Equations

The log-likelihood for (10) is

$$l(\boldsymbol{\Psi}, \boldsymbol{\theta}; \mathbf{y}) = \boldsymbol{\Psi}^T \mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{Z}(\mathbf{y}) - \kappa \left\{ \boldsymbol{\Psi}, \boldsymbol{\eta}(\boldsymbol{\theta}) \right\}$$

with the normalizing constant defined by

$$\kappa \left\{ \boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{\Psi}) \right\} = \log \left\{ \sum_{\mathbf{y} \in Y} \exp(\boldsymbol{\Psi}^T \mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{Z}(\mathbf{y})) \right\},$$

where the summation goes over all possible networks $\mathbf{y} \in Y$.

Let $\bigtriangledown \boldsymbol{\eta}(\boldsymbol{\theta})$ denote the $q \times p$ matrix of partial derivatives of $\boldsymbol{\eta}$ with respect to $\boldsymbol{\theta}$. Also let $\mathbf{D} = \text{Diag}(\text{Var}(\mathbf{Y}))^{-1}$ for logistic regression, for general link function $g(\cdot)$ it is defined as $\mathbf{D} = \partial \boldsymbol{\pi} / \partial \boldsymbol{\nu}$.

The likelihood equations have the following form

$$\tilde{\mathbf{X}}^T \mathbf{D} \mathbf{C}_{\mathbf{Y}}^{-1} (\mathbf{Y}^{obs} - \boldsymbol{\pi}) = \mathbf{0} \tag{11}$$

and

$$\bigtriangledown \boldsymbol{\eta}(\boldsymbol{\theta}) \left\{ -\mathbf{C}_{\mathbf{Z},\mathbf{Y}} \mathbf{C}_{\mathbf{Y}}^{-1} (\mathbf{Y}^{obs} - \boldsymbol{\pi}) + \mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z} \right\} = \mathbf{0} \tag{12}$$

where $\mathbf{C}_{\mathbf{A},\mathbf{B}} := \mathrm{Cov}(\mathbf{A}, \mathbf{B})$ and $\mathbf{C}_{\mathbf{A}} := \mathbf{C}_{\mathbf{A},\mathbf{A}}$. The appendix shows the details of how these equations are derived.

Likelihood equations (11) and (12) are similar to those derived by FL93 for log-linear models subject to a marginal model. The first set of equations (11) has the standard form of generalized estimating equations (GEE) (Liang and Zeger 1986).

The likelihood equations for CERGM are

$$\bigtriangledown \boldsymbol{\eta}(\boldsymbol{\theta}) \left(\mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z} \right) \tag{13}$$

and reduce to $\mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z} = \mathbf{0}$ for an ERGM. Equation (12) is different from (13) containing another term associated with the first set of equations given by (11).

The covariance of ML estimates $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)^T$ can be approximated by the inverse of the Fisher information matrix:

$$\begin{aligned} \widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})} := \hat{I}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})^{-1} &= \left\{ \mathbb{E} \begin{pmatrix} \partial l / \partial \boldsymbol{\beta} \\ \partial l / \partial \boldsymbol{\theta} \end{pmatrix} \begin{pmatrix} \partial l / \partial \boldsymbol{\beta} \\ \partial l / \partial \boldsymbol{\theta} \end{pmatrix}^T \right\}^{-1} \\ &= \begin{pmatrix} (\tilde{\mathbf{X}}^T \mathbf{D} \mathbf{C}_{\mathbf{Y}}^{-1} \mathbf{D} \tilde{\mathbf{X}})^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\bigtriangledown \boldsymbol{\eta}(\boldsymbol{\theta}) \{ \mathbf{C}_{\mathbf{Z}} - \mathbf{C}_{\mathbf{Z},\mathbf{Y}} \mathbf{C}_{\mathbf{Y}}^{-1} \mathbf{C}_{\mathbf{Z},\mathbf{Y}}^T \} \bigtriangledown \boldsymbol{\eta}(\boldsymbol{\theta})^T \right)^{-1} \end{pmatrix}. \end{aligned}$$

The off-diagonal parts of the matrix are zero implying that $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are orthogonal to each other (Cox and Reid 1987). Therefore the consistency of $\hat{\boldsymbol{\beta}}$ does not depend on the correct specification of the joint model, which is in contrast to a CERGM for which a

mis-specification of the full model implies inconsistency for the effects $\boldsymbol{\beta}$ of the exogenous variables. This is a big advantage when we are mainly interested in the estimation of $\boldsymbol{\beta}$.

The ML estimates can be obtained by a Fisher-scoring scheme where the difference between old and new iterates ($\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} + \Delta(\boldsymbol{\beta})$ and $\boldsymbol{\theta}^{new} = \boldsymbol{\theta}^{old} + \Delta(\boldsymbol{\theta})$) is given by

$$\Delta(\boldsymbol{\beta}) = (\tilde{\mathbf{X}}^T \mathbf{D} \mathbf{C}_{\mathbf{Y}}^{-1} \mathbf{D} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{D} \mathbf{C}_{\mathbf{Y}}^{-1} (\mathbf{Y}^{obs} - \boldsymbol{\pi}) \tag{14}$$

$$\begin{aligned} \Delta(\boldsymbol{\theta}) = &\left(\bigtriangledown \boldsymbol{\eta}(\boldsymbol{\theta}) \{ \mathbf{C}_{\mathbf{Z}} - \mathbf{C}_{\mathbf{Z},\mathbf{Y}} \mathbf{C}_{\mathbf{Y}}^{-1} \mathbf{C}_{\mathbf{Z},\mathbf{Y}}^T \} \bigtriangledown \boldsymbol{\eta}(\boldsymbol{\theta})^T \right)^{-1} \times \\ &\bigtriangledown \boldsymbol{\eta}(\boldsymbol{\theta}) \left\{ -\mathbf{C}_{\mathbf{Z},\mathbf{Y}} \mathbf{C}_{\mathbf{Y}}^{-1} (\mathbf{Y}^{obs} - \boldsymbol{\pi}) + \mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z} \right\}. \end{aligned} \tag{15}$$

There is one main problem with applying this scoring algorithm. The scoring equations do not only depend on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ but also on $\boldsymbol{\Psi}$. There are no closed-form expressions for $\boldsymbol{\Psi}$ depending on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. FL93 circumvented this problem by applying the iterative proportional fitting (IPF) algorithm in each step to first obtain a solution for the complete joint distribution from which all other quantities needed in (14) and (15) can be computed. Even though they were dealing with only a few binary observations (in this case the joint distribution consists of only a few joint probabilities), the IPF algorithm is time consuming. For our network data, for which the joint distribution consists of 2^N probabilities, this method is not applicable, even if n is relatively small, say $n = 5$, then $2^N = 2^{\binom{5}{2}} = 1024$ would already be close to infeasibility for the IPF algorithm. Therefore, other methods need to be considered. The next section addresses how ML estimation can still be achieved.

5 ML Estimation

5.1 Preliminaries

The following section discusses ML estimation for marginal curved ERGM (MCERGM).

First define $\boldsymbol{\alpha} =: (\boldsymbol{\Psi}^T, \boldsymbol{\theta}^T)^T$ for notational convenience. Consider two distinct parameter

vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^0$ and write:

$$r(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0) := l(\boldsymbol{\alpha}) - l(\boldsymbol{\alpha}^0) = (\boldsymbol{\Psi} - \boldsymbol{\Psi}^0)^T \mathbf{y}^{obs} + (\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^0))^T \mathbf{Z}^{obs} - (\kappa(\boldsymbol{\alpha}) - \kappa(\boldsymbol{\alpha}^0)) \quad (16)$$

Now we express the last term on the right hand side as a function of $\boldsymbol{\alpha}$ for known and fixed $\boldsymbol{\alpha}^0$. Then

$$\begin{aligned} \exp\left(\kappa(\boldsymbol{\alpha}) - \kappa(\boldsymbol{\alpha}^0)\right) &= \sum_{\mathbf{y} \in Y} \exp\left[(\boldsymbol{\Psi} - \boldsymbol{\Psi}^0)^T \mathbf{y} + (\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^0))^T \mathbf{Z}\right] \left(\frac{(\boldsymbol{\Psi}^0)^T \mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^0)^T \mathbf{Z}}{\kappa(\boldsymbol{\alpha}^0)} \right) \\ &= \mathbb{E}_{\boldsymbol{\alpha}^0} \left\{ \exp\left[(\boldsymbol{\Psi} - \boldsymbol{\Psi}^0)^T \mathbf{y} + (\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^0))^T \mathbf{Z}\right] \right\} \quad (17) \end{aligned}$$

This expectation might be approximated by a sample $\mathbf{y}_1, \ldots, \mathbf{y}_m$ from the random graph distribution for given $\boldsymbol{\alpha}^0$ and may be obtained by a Markov chain Monte Carlo (MCMC) algorithm.

Define $U(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0, \mathbf{y}) := \exp\left[(\boldsymbol{\Psi} - \boldsymbol{\Psi}^0)^T \mathbf{y} + (\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^0))^T \mathbf{Z}(\mathbf{y})\right]$. Approximate $r(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0)$ by

$$\hat{r}_m(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0) := \log U(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0, \mathbf{y}^{obs}) - \log \left[\frac{1}{m} \sum_{i=1}^{m} U(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0, \mathbf{y}_i) \right]. \quad (18)$$

The term $\hat{r}_m(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0)$ converges almost surely to $r(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0)$ as $m \to \infty$. For fixed sample size m, maximization of $\hat{r}_m(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0)$ for fixed $\boldsymbol{\alpha}^0$ as a function of $\boldsymbol{\alpha}$ provides an approximation of the maximum likelihood estimator $\tilde{\boldsymbol{\alpha}}$. This procedure called MCMC maximum likelihood estimation (MCMCMLE) was developed by Geyer (1992) and suggested by HH06 to fit CERGMs.

The ratio $\hat{r}_m(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0)$ can be used to compute the log likelihood. Note $l(\mathbf{0}) = -\log M$ with $M = N \log 2$, now

$$\hat{l}(\boldsymbol{\alpha}) := \hat{r}_m(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0) - \hat{r}_m(\mathbf{0}, \boldsymbol{\alpha}^0) - \log M. \quad (19)$$

Despite the formula's simplicity, reliably estimating $r(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0)$ and $r(\mathbf{0}, \boldsymbol{\alpha}^0)$ is difficult.

5.2 Obtaining MCMC Sample

Let $\mathbf{y}_{current}$ be given a network and assume we want to sample another network. Then we use a stochastic or deterministic process to determine a pair (i, j) and then decide whether $Y_{ij} = 1$ or $Y_{ij} = 0$. When fixing the rest of the graph $Y_{ij}^c = y_{ij}^c$

$$\frac{\Pr(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{\Pr(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \exp\left(\Psi_{ij} + \boldsymbol{\eta}(\boldsymbol{\theta})^T \Delta(\mathbf{Z}(\mathbf{y}))_{ij}\right), \tag{20}$$

where $\Delta(\mathbf{Z}(\mathbf{y}))_{ij}$ is the change statistic for $\mathbf{Z}$ defined previously. This formula is slightly different from the formula for CERGM, because it contains additionally parameter Ψ_{ij}. This method is called Gibbs sampling. An alternative is the Metropolis algorithm for which we need to propose transitions from $\mathbf{y}_{current}$ to $\mathbf{y}_{proposed}$. The algorithm accepts $\mathbf{y}_{proposed}$ with probability

$$\min\left(1, \frac{\Pr(\mathbf{Y} = \mathbf{y}_{proposed})}{\Pr(\mathbf{Y} = \mathbf{y}_{current})}\right).$$

The ratio is:

$$\frac{\Pr(\mathbf{Y} = \mathbf{y}_{proposed})}{\Pr(\mathbf{Y} = \mathbf{y}_{current})} = \exp\left[\boldsymbol{\Psi}^T(\mathbf{y}_{proposed} - \mathbf{y}_{current}) + \boldsymbol{\eta}(\boldsymbol{\theta})^T(\mathbf{Z}(\mathbf{y}_{proposed}) - \mathbf{Z}(\mathbf{y}_{current}))\right]. \tag{21}$$

When $\mathbf{y}_{proposed}$ and $\mathbf{y}_{current}$ only differ by a single edge, then the right hand side of (21) reduces to $\exp(\Psi_{ij} + \boldsymbol{\eta}(\boldsymbol{\theta})^T \Delta(\mathbf{Z}(\mathbf{y}))_{ij})$, see equation (20). When $\mathbf{y}_{current}$ and $\mathbf{y}_{proposed}$ differ substantially, then we can consider a sequence of networks, two consecutive networks only differing by one pair (i, j), and the sequence starting with the current network and finishing with the proposed network. For each step, the ratio is a simple function of change statistics making the ratio (21) relatively easily computable.

For our purposes we use the `R` (R-Development-Core-Team 2006) package `ergm` (Handcock et al. 2010) to simulate a MCMC sample, because all common network statistics are implemented and also the efficient calculation of change statistics makes fast simulation of the MCMC sample possible. However, `ergm` does not allow specification of $\boldsymbol{\Psi}$ directly, so

we apply a manipulation and exchange edge covariates with parameters Ψ_{ij} and set the corresponding single parameter $\eta(\theta)_{ij}$ to one. Another problem occurs, because the package only returns m network statistics $\mathbf{Z}(\mathbf{y}_1), \ldots, \mathbf{Z}(\mathbf{y}_m)$ and the last network $\mathbf{y}_m$, but it does not return the sequence $\mathbf{y}_1, \ldots, \mathbf{y}_m$ needed for the estimation of the model specified by equations (9) and (10). This means we needed to call the `ergm` function m times instead of calling it just once. As a consequence the sampling process for one MCMC sample took roughly $1-5$ minutes for the Lazega data set instead of roughly 10 seconds using a PC with processor Intel C2Q Q9550 2.83GHz. Only a slight modification of the internal C function (part of `ergm` package) would be needed to return all m networks and the whole sampling process would take no more than just a few seconds.

5.3 Fitting Algorithm

Firstly, be aware of the two different parameterizations of the full model: two equivalent sets of parameters are $\boldsymbol{\alpha} = (\boldsymbol{\Psi}^T, \boldsymbol{\theta}^T)^T$ and $\boldsymbol{\zeta} := (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$, where $\boldsymbol{\alpha}$ contains the vector of nuisance parameters $\boldsymbol{\Psi}$, that is of little interest, but is needed to obtain a new MCMC sample.

Second, initial parameter values for which the MCMC sample was generated are denoted by $\boldsymbol{\alpha}^0$ (needs to satisfy (9) with $\boldsymbol{\beta}^0$) and parameters of the kth iteration still based on this MCMC sample generated at $\boldsymbol{\alpha}^0$ are denoted by $\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\Psi}^{(k)}$ and for $k = 0$: $\boldsymbol{\beta}^{(0)} := \boldsymbol{\beta}^0$, $\boldsymbol{\theta}^{(0)} := \boldsymbol{\theta}^0$ and $\boldsymbol{\Psi}^{(0)} := \boldsymbol{\Psi}^0$.

The main algorithm, basically a Fisher-scoring algorithm, is presented next, followed by a detailed explanation of the steps.

Main Algorithm

1 select initial values $\boldsymbol{\beta}^0$, $\boldsymbol{\theta}^0$ and $\boldsymbol{\Psi}^0$ satisfying (9) and (10); obtain MCMC sample $\mathbf{y}_1, \ldots, \mathbf{y}_m$ along with $\mathbf{Z}(\mathbf{y}_1), \ldots, \mathbf{Z}(\mathbf{y}_m)$ for given $\boldsymbol{\alpha}^0$, set k:=0, go to step 3;

2 obtain $\boldsymbol{\Psi}^{(k+1)}$ that solves

$$\mathbb{E}\mathbf{Y}(\boldsymbol{\Psi}^{(k+1)}, \boldsymbol{\theta}^{(k)}) = \text{expit}(\tilde{X}\boldsymbol{\beta}^{(k)})$$

(expit := logit^{-1}) for given MCMC sample generated at $\boldsymbol{\alpha}^0$; if $\widehat{\text{Var}}_{MC}(\hat{r}_m)$ in equation (22) is too large, say $\widehat{\text{Var}}_{MC}(\hat{r}_m) > c$, then set $k := 0$ and obtain new MCMC sample;

3 estimate $\mathbf{C}_{\mathbf{Y}}$, $\mathbf{C}_{\mathbf{Z},\mathbf{Y}}$, $\mathbf{C}_{\mathbf{Z}}$, $\mathbb{E}\mathbf{Z}$ from MCMC sample for given $\boldsymbol{\alpha}^0$ and $\boldsymbol{\alpha}^{(k)}$ needed for formulas (14) and (15), see below for details of estimation of moments;

4 apply iteration scheme (14) and (15), new estimates are obtained via

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \gamma_{\boldsymbol{\beta}^{(k)}}\Delta(\boldsymbol{\beta}^{(k)})$$

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \gamma_{\boldsymbol{\theta}^{(k)}}\Delta(\boldsymbol{\theta}^{(k)});$$

5 if converged consider $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(k+1)}$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(k+1)}$ as MCMCMLE, stop;

6 $k := k + 1$, return to step 2.

Here $\gamma_{\boldsymbol{\beta}^{(k)}}$ and $\gamma_{\boldsymbol{\theta}^{(k)}}$ are step sizes, ideally equal to one, but our experience has shown that step-sizes should rather be smaller, e.g. $0.2 - 0.5$, making the iteration scheme more stable.

Computation of Expectations in Step 3 and of $\widehat{\text{Var}}_{MC}(\hat{r}_m)$ in step 2

Previously we defined $U(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0, \mathbf{Y})$. We write $U_1, \ldots, U_m$ for $U(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0, \mathbf{y}_1), \ldots, U(\boldsymbol{\alpha}, \boldsymbol{\alpha}^0, \mathbf{y}_m)$ for the given MCMC sample $\mathbf{y}_1, \ldots, \mathbf{y}_m$ at $\boldsymbol{\alpha}^0$, similarly $\mathbf{Z}_i$ for $\mathbf{Z}(\mathbf{y}_i)$.

Define the weights

$$\omega_i^{(k)} := \frac{U_i}{\sum_{j=1}^m U_j}.$$

Clearly these weights are equal, i.e. $\omega_i^{(k)} = 1/m$, if $\boldsymbol{\alpha}^{(k)} = \boldsymbol{\alpha}^0$.

$$\widehat{\mathrm{Var}}_{MC}(\hat{r}_m) := \frac{1}{m^2 \bar{U}^2} \sum_{k=-K}^{K} (m - |k|)\phi_k, \tag{22}$$

where $\phi_k = \phi_{-k}$ is the auto-covariance of the sequence $U_1, \ldots, U_m$ and $\bar{U} := \frac{1}{m}\sum_{i=1}^{m} U_i$. This is the same equation as in HH06. The expectations can be estimated by weighted sums given by

$$\widehat{\mathbb{E}\mathbf{Z}} = \sum_{i=1}^{m} \omega_i^{(k)} \mathbf{Z}_i \qquad\qquad \hat{\mathbf{C}}_{\mathbf{Z},\mathbf{Y}} = \sum_{i=1}^{m} \omega_i^{(k)} \mathbf{Z}_i \mathbf{Y}_i^T - \mathbb{E}\mathbf{Z}(\mathbb{E}\mathbf{Y})^T, \tag{23}$$

and similarly for all other expectations. These estimated expectations are identical to sample means and sample covariances for a fresh MCMC sample, when $\boldsymbol{\alpha}^{(k)} = \boldsymbol{\alpha}^0$. Equation (23) enables estimation of moments without generating new MCMC samples. Step 2 ($\widehat{\mathrm{Var}}_{MC}(\hat{r}_m) > c$) says, when the variation in $U_1, \ldots, U_m$ is too large, then we need to generate a new MCMC sample, because the old sample is too unreliable in obtaining estimates of the expectations.

Least Squares Algorithm for Solving for $\boldsymbol{\Psi}$

Step 2 of the main algorithm is complex. We need to obtain $\boldsymbol{\Psi}^{(k+1)}$ that solves

$$\boldsymbol{\pi}(\boldsymbol{\beta}^{(k)}) = \boldsymbol{\pi}(\boldsymbol{\Psi}^{(k+1)}, \boldsymbol{\theta}^{(k)}).$$

with

$$\boldsymbol{\pi}(\boldsymbol{\beta}) := \mathrm{expit}\left(\tilde{\mathbf{X}}\boldsymbol{\beta}\right)$$

and

$$\boldsymbol{\pi}(\boldsymbol{\Psi}, \boldsymbol{\theta}) := \sum_{\mathbf{y} \in Y} \mathbf{y} \exp\left(\boldsymbol{\Psi}^T \mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{Z}(\mathbf{y}) - \kappa\left\{\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{\Psi})\right\}\right).$$

The vector of probabilities $\boldsymbol{\pi}(\boldsymbol{\beta})$ is given by the marginal logistic regression model (9) and $\boldsymbol{\pi}(\boldsymbol{\Psi}, \boldsymbol{\theta})[= \boldsymbol{\pi}(\boldsymbol{\alpha})]$ is given by the joint model (10). Define $\mathbf{g}(\boldsymbol{\Psi}) := \boldsymbol{\pi}(\boldsymbol{\beta}) - \boldsymbol{\pi}(\boldsymbol{\Psi}, \boldsymbol{\theta})$. The equation $\mathbf{g} = \mathbf{0}$ can be solved iteratively (using index j for sub-iterations for kth step of main algorithm) by non-linear least squares

$$\Delta(\boldsymbol{\Psi})^{(k,j)} = \Big(\bigtriangledown \mathbf{g}(\boldsymbol{\Psi}^{(k,j)})^T \mathbf{W} \bigtriangledown \mathbf{g}(\boldsymbol{\Psi}^{(k,j)}) \Big)^{-1} \bigtriangledown \mathbf{g}(\boldsymbol{\Psi}^{(k,j)})^T \mathbf{W} \mathbf{g}(\boldsymbol{\Psi}^{(k,j)}), \tag{24}$$

where $\boldsymbol{\Psi}^{(k,j)}$ is the current iterate of $\boldsymbol{\Psi}$ and $\bigtriangledown \mathbf{g}(\boldsymbol{\Psi}^{(k,j)})$ is the derivative of $\mathbf{g}$ with respect to $\boldsymbol{\Psi}$ at $\boldsymbol{\Psi}^{(k,j)}$. This derivative reduces to $\bigtriangledown \mathbf{g}(\boldsymbol{\Psi}) = \partial \boldsymbol{\pi}(\boldsymbol{\alpha}) / \partial \boldsymbol{\Psi} = \mathbf{C}_{\mathbf{Y}}(\boldsymbol{\Psi})$. Matrix $\mathbf{W}$ is a weight matrix, for simplicity let $\mathbf{W} = \mathbf{I}$. Equation (24) then reduces to

$$\Delta(\boldsymbol{\Psi}^{(k,j)}) = \mathbf{C}_{\mathbf{Y}}^{-1}(\boldsymbol{\Psi}^{(k,j)}) \mathbf{g}(\boldsymbol{\Psi}^{(k,j)}). \tag{25}$$

We suggest the following algorithm for step 2 of the main algorithm to obtain $\boldsymbol{\Psi}^{(k+1)}$ for given $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$:

Step 2 of Main Algorithm

2.0 given is a MCMC sample $\mathbf{y}_1, \ldots, \mathbf{y}_m$ at $\boldsymbol{\alpha}^0 = ((\boldsymbol{\Psi}^0)^T, (\boldsymbol{\theta}^0)^T)^T$, set $j := 0$ and set $\boldsymbol{\Psi}^{(k,j)} := \boldsymbol{\Psi}^{(k)}$;

2.1 estimate $\mathbf{C}_{\mathbf{Y}}^{-1}(\boldsymbol{\Psi}^{(k,j)})$ and $\mathbf{g}(\boldsymbol{\Psi}^{(k,j)})$ from MCMC sample;

2.2 adjust step-size $\gamma_{\boldsymbol{\Psi}^{(j)}}$ and $\mathbf{C}_{\mathbf{Y}}^{-1}$ according to $\|\mathbf{g}(\Psi^{(j)})\|$, $\|\cdot\|$ is some norm;

2.3 apply equation (25) for current $\Psi^{(k,j)}$ and obtain $\Psi^{(k,j+1)}$ by

$$\Psi^{(k,j+1)} = \Psi^{(k,j)} + \gamma_{\boldsymbol{\Psi}^{(k,j)}} \Delta(\boldsymbol{\Psi}^{(k,j)});$$

2.4 update weights $\omega_i^{(j+1)}$;

2.5 if $\widehat{\mathrm{Var}}_{MC}(\hat{r}_m)$ in equation (22) is too large, say $\widehat{\mathrm{Var}}_{MC}(\hat{r}_m) > c$, then $\boldsymbol{\Psi}^0 := \boldsymbol{\Psi}^{(k,j+1)}$ and obtain a new MCMC sample for $\boldsymbol{\Psi}^0$ and $\boldsymbol{\theta}^0 := \boldsymbol{\theta}^{(k)}$; set $\boldsymbol{\beta}^0 := \boldsymbol{\beta}^{(k)}$, $j := 0$ and $k := 0$;

2.6 convergence is achieved when

$$T := m(\widehat{\mathbf{g}}(\mathbf{\Psi}^{(k,j+1)})^T(\widehat{\mathbf{C}}_{\mathbf{Y}}(\mathbf{\Psi}^{(k,j+1)}))^{-1}\widehat{\mathbf{g}}(\mathbf{\Psi}^{(k,j+1)}) \leq T_2^{\epsilon}\left[\binom{n}{2}, m-1\right] =: T_2^{\epsilon},$$

if converged stop and consider $\mathbf{\Psi}^{(k+1)} := \mathbf{\Psi}^{(k,j+1)}$ as the solution of step 2, otherwise continue with 2.7;

2.7 go to step 2.2, j:=j+1.

Due to the estimation of $\mathbf{g}$ based on the MCMC sample we cannot determine exactly whether indeed $\mathbf{g} = \mathbf{0}$, i.e. whether $\boldsymbol{\pi}(\mathbf{\Psi}^{(k+1)}, \boldsymbol{\theta}^{(k)}) = \boldsymbol{\pi}(\boldsymbol{\beta}^{(k)})$ due to the stochastic approximations. Let $\boldsymbol{\pi}(\mathbf{\Psi}^{(k+1)}, \boldsymbol{\theta}^{(k)})$ denote the true vector of probabilities and $\hat{\boldsymbol{\pi}}(\mathbf{\Psi}^{(k+1)}, \boldsymbol{\theta}^{(k)}) := \sqrt{m}\sum_{i=1}^{m}\mathbf{y}_i/m$ the estimate based on the MCMC sample $\mathbf{y}_1, \ldots, \mathbf{y}_m$. We know $\sqrt{m}\hat{\boldsymbol{\pi}}(\mathbf{\Psi}, \boldsymbol{\beta}) \sim \text{Bin}(\boldsymbol{\pi}(\mathbf{\Psi}, \boldsymbol{\beta}), m)$, hence $\sqrt{m}\hat{\boldsymbol{\pi}}(\mathbf{\Psi}, \boldsymbol{\beta}) \sim_d N(\boldsymbol{\pi}(\mathbf{\Psi}, \boldsymbol{\beta}), \mathbf{C}_{\mathbf{Y}})$. In step 2.6, we apply Hotelling's T^2 square test, since $\mathbf{C}_{\mathbf{Y}}$ also needs to be estimated. $T_2^{\epsilon} := T_2^{\epsilon}[\binom{n}{2}, m-1]$ is the ϵ-quantile of the T^2 distribution with parameters $\binom{n}{2}$ and $m-1$. Note $\frac{a-b+1}{ab}T_{a,b}^2 = F_{a,b-a+1}$, where $F_{a,b}$ is the F distribution with parameters a and b. We chose $\epsilon = 0.1$. Be aware $\mathbf{y}_1, \ldots, \mathbf{y}_m$ are not independent due to the MCMC technique, therefore the T^2 distribution applies only approximately. This test is only a criterion to determine when to stop.

Step 2.2 includes a step size $\gamma_{\mathbf{\Psi}}$. As for $\gamma_{\boldsymbol{\beta}}$ and $\gamma_{\boldsymbol{\theta}}$ in the main algorithm, this step-size is ideally one, but we found that tuning these and other parameters made the algorithm more stable. The norm $\|\mathbf{g}(\mathbf{\Psi})\|$ can be any distance measure. For our implementation we use T, because this Mahalanobis distance is already used as a stopping criteria in step 2.6. When T is large, say $T > 2 \cdot T_2^{\epsilon}$ then we use $\gamma_{\mathbf{\Psi}} = 0.5$, otherwise the step-size is reduced in each step by, say, 5% until it reaches a minimal step size, say 0.005.

This might be exactly the opposite of what one might expect, but a large step-size (e.g. 0.9) near the solution resulted frequently in a big jump away from the solution. We also modified $\mathbf{C}_{\mathbf{Y}}^{-1}$ in equation (25) according to $\|\mathbf{g}(\mathbf{\Psi})\|$ respectively T. In fact we exchanged

$\mathbf{C_Y}$ with

$$\lambda \mathbf{C_Y} + (1-\lambda)\text{Var}(\mathbf{Y}),$$

where $\text{Var}(\mathbf{Y})$ is the diagonal matrix with variance of $\mathbf{Y}$ on its diagonal. $\lambda \in [0,1]$ should be 1 when T is close to T_2^ϵ, say $T < 1.2 \cdot T_2^\epsilon$. If far apart, say $T > 2 \cdot T_2^\epsilon$, then we set $\lambda = 0$. If $1.2 \cdot T_2^\epsilon \leq T \leq 2 \cdot T_2^\epsilon$, then $\lambda = 1 - \max\{(T - T_2^\epsilon)/T_2^\epsilon, 0\}$. This tuning worked well for the Lazega (2001) data set.

Sequential Solving for $\mathbf{\Psi}$

As an alternative to the proposed method in step 2, we propose another algorithm which solves each component of $\mathbf{g} = \mathbf{0}$, i.e. $\pi_{ij}(\mathbf{\Psi}^{(k+1)}, \boldsymbol{\theta}^{(k)}) = \pi_{ij}(\boldsymbol{\beta}^{(k)})$ directly as a function of $\Psi_{ij}^{(k+1)}$. The marginal probability

$$\Pr(Y_{ij} = 1) = \sum_{\mathbf{y} \in Y : Y_{ij} = 1} \Pr(\mathbf{Y} = \mathbf{y}) = \sum_{y_{ij}^c \in Y_{ij}^c} \Pr(Y_{ij} = 1, Y_{ij}^c = y_{ij}^c)$$

is difficult to compute, as it needs an estimate of the normalizing constant κ.

Instead we can express the marginal probability as

$$\Pr(Y_{ij} = 1) = \sum_{y_{ij}^c \in Y_{ij}^c} \Pr(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c) \Pr(Y_{ij}^c = y_{ij}^c),$$

which can be estimated from the MCMC sample by (see equation (20))

$$\Pr(Y_{ij} = 1) \approx \sum_{k=1}^{m} \Pr((\mathbf{Y}_k)_{ij} = 1 | (\mathbf{Y}_k)_{ij}^c = (\mathbf{y}_k)_{ij}^c) = \sum_{k=1}^{m} \text{expit}\left(\Psi_{ij} + \boldsymbol{\eta}(\boldsymbol{\theta})^T \Delta(\mathbf{Z}(\mathbf{y}_k))_{ij}\right), \tag{26}$$

where $\mathbf{y}_k$ is the kth network of the MCMC sample and $(\mathbf{y}_k)_{ij}$ is the ij relation of the kth network. Assume we would know all $\Delta(\mathbf{Z}(\mathbf{y}_k))_{ij}$ for the MCMC sample, then we simply need to solve

$$\pi_{ij}(\boldsymbol{\beta}) - \sum_{k=1}^{m} \omega_i \times \text{expit}\left(\Psi_{ij} + \boldsymbol{\eta}(\boldsymbol{\theta})^T \Delta(\mathbf{Z}(\mathbf{y}_k))_{ij}\right) = 0, \tag{27}$$

which can be solved for each Ψ_{ij}, given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, by standard optimization routines.

After we have solved (27) for the pair ij, we can proceed to solve for the pair uv, but weights ω_i need to be updated due to the change from $\Psi_{ij}^{(k+1)}$ to $\Psi_{ij}^{(k+1)}$, implied by solving equation (27).

When $\widehat{\text{Var}}_{MC}(\hat{r}_m)$ in equation (22) is too large, say $\widehat{\text{Var}}_{MC}(\hat{r}_m) > c$, then we need to obtain a new MCMC sample for $\boldsymbol{\Psi}^0 = \boldsymbol{\Psi}^{(j+1)}$ and $\boldsymbol{\theta}^0 := \boldsymbol{\theta}^{(k)}$.

There are two main issues with this algorithm. The first is that we need to solve for $\binom{n}{2}$ parameters in for each k and each large change in $\boldsymbol{\Psi}$ might require generating a new MCMC sample. The second issue is the computation of all $\Delta(\mathbf{Z}(\mathbf{y}_k))_{ij}$ of the complete MCMC sample. We computed this for each network of the MCMC sample directly, i.e. $\binom{n}{2}$ change statistics for each of the m, say $m = 10,000$, networks. Even though we implemented it efficiently, it is still not efficient enough, taking roughly 40 minutes for the Lazega (2001) data set.

However, there is a relatively easy solution. When the MCMC sample is created, each sampled network $\mathbf{y}_k$ differs by the predecessor $\mathbf{y}_{k-1}$ by a few dyads. Assume the change statistics $\Delta(\mathbf{Z}(\mathbf{y}_{k-1}))_{ij}$ are known, then the change statistics $\Delta(\mathbf{Z}(\mathbf{y}_k))_{ij}$ can be computed from $\Delta(\mathbf{Z}(\mathbf{y}_{k-1}))_{ij}$ and the knowledge of the pairs Y_{ij} that have changed. The only challenge remaining is its implementation in the existing `ergm` package along with an output of the networks $\mathbf{y}_1, \ldots, \mathbf{y}_m$, as previously mentioned.

5.3.1 Monte-Carlo Error

Approximating the ML estimates $\hat{\boldsymbol{\zeta}}$ (previously defined by $\boldsymbol{\zeta} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$) by $\tilde{\boldsymbol{\zeta}}$ incurs another error, the Monte Carlo error. Equivalently let $\hat{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\alpha}}$ denote the estimates for the alternative parameterization ($\boldsymbol{\alpha} = (\boldsymbol{\Psi}^T, \boldsymbol{\theta}^T)^T$).

Applying a Taylor series approximation gives

$$m^{1/2}(\tilde{\boldsymbol{\zeta}} - \hat{\boldsymbol{\zeta}}) \approx \left[\bigtriangledown^2 \hat{r}_m(\tilde{\boldsymbol{\zeta}})\right]^{-1} \bigtriangledown \hat{r}_m(\tilde{\boldsymbol{\zeta}}), \tag{28}$$

where $\bigtriangledown^2 \hat{r}_m(\tilde{\boldsymbol{\zeta}})$ denotes the Hessian and $\bigtriangledown \hat{r}_m$ the gradient. Suppose $\mathbf{y}_1, \ldots, \mathbf{y}_m$ arise from a stationary Markov chain for given $\boldsymbol{\alpha}^0$. Then $m^{1/2}\hat{r}_m(\tilde{\boldsymbol{\zeta}})$ converges in distribution as $m \to \infty$ to a $K+1+p$-variate normal distribution with mean $\mathbf{0}$ and covariance

$$\left[\frac{\kappa(\boldsymbol{\alpha}^0)}{\kappa(\hat{\boldsymbol{\alpha}})}\right]^2 \sum_{-\infty}^{\infty} \mathrm{Cov}\left(\mathbf{W}_1(\hat{\boldsymbol{\alpha}}), \mathbf{W}_{1+|k|}(\hat{\boldsymbol{\alpha}})\right),$$

where

$$\mathbf{W}_i(\boldsymbol{\alpha}) = \begin{pmatrix} \tilde{\mathbf{X}}\mathbf{D}\mathbf{C}_{\mathbf{Y}}^{-1}(\mathbf{y}^{obs} - \mathbf{y}_i) \times U_i(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0) \\ \bigtriangledown\boldsymbol{\eta}(\boldsymbol{\theta}) \left\{-\mathbf{C}_{\mathbf{Z},\mathbf{Y}}\mathbf{C}_{\mathbf{Y}}^{-1}(\mathbf{y}^{obs} - \mathbf{y}_i) + \mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z}_i\right\} \times U_i(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0) \end{pmatrix}. \tag{29}$$

The value of $\hat{\boldsymbol{\alpha}}$ is unknown but replaced by $\tilde{\boldsymbol{\alpha}}$. The ratio $\kappa(\boldsymbol{\alpha}^0)/\kappa(\hat{\boldsymbol{\alpha}})$ is approximated by the sample mean and finally (5.3.1) is approximated by

$$\tilde{V} := \frac{1}{m}\left[\sum_{i=1}^{m} U(\tilde{\boldsymbol{\alpha}}, \boldsymbol{\alpha}^0, \mathbf{y}_i)\right]^2 \sum_{k=-K}^{K} \hat{\xi}_k,$$

where $\hat{\xi}_k$ is the sample lag-k auto-covariance matrix of the sequence $\mathbf{W}_1(\tilde{\boldsymbol{\alpha}}), \mathbf{W}_2(\tilde{\boldsymbol{\alpha}}), \ldots, \mathbf{W}_m(\tilde{\boldsymbol{\alpha}})$.

The Hessian $\bigtriangledown^2 \hat{r}_m(\tilde{\boldsymbol{\zeta}})$ is difficult to compute and is replaced by the estimated Fisher information matrix $\hat{I}(\tilde{\boldsymbol{\zeta}})$ and finally we obtain

$$\hat{I}(\tilde{\boldsymbol{\zeta}})^{-1}\tilde{V}\hat{I}(\tilde{\boldsymbol{\zeta}})^{-1},$$

as the final approximate covariance matrix for $\tilde{\boldsymbol{\zeta}}$.

6 Example - Lazega Data Set

Table 1 shows the results for the model with covariates only (simple logistic model) and Table 2 the estimates for the curved ERGM with the GWESP statistic added. Here we use the GWESP statistic for two main reasons: firstly, one can extend the framework for

Variables	estimates	Accounting for MCMC error (s.e)	p-value	Not Accounting for MCMC error (s.e)	p-value	odds ratio
edges	-7.383	(1.609)	$4.4e-06$	(1.350)	$4.5e-08$	0.0006
main seniority	0.0387	(0.014)	0.00684	(0.012)	0.00209	1.0395
main practice	0.7643	(0.347)	0.02766	(0.275)	0.00552	2.1475
sim practice	0.8800	(0.262)	0.00078	(0.271)	0.00115	2.4109
sim gender	0.7831	(0.483)	0.10529	(0.429)	0.06841	2.1882
sim office	1.7572	(0.393)	$7.8e-06$	(0.369)	$1.9e-06$	5.7962
GWESP (θ_1)	0.8780	(0.269)	0.00108	(0.288)	0.00232	–
GWESP (θ_2)	0.8627	(0.212)	$4.9e-05$	(0.214)	$5.7e-05$	–

Table 3: Estimates for the Marginalized curved ERGM for Law Firm Data Set

CERGMs and secondly, one can illustrate the difference of MCERGMs and CERGMs effectively, because the data set has been analyzed in previous papers, e.g. HH06.

Table 3 shows the results for the same data set fitting the marginalized curved ERGM. The parameter estimates of the exogenous effects (main effects, similarity effects and intercept) in Table 3 are substantially different from those of Tables 1 and 2. We started with sampling $m = 10,000$ networks and finished with $30,000$ to obtain higher accuracy. We also increased the step-size (also known as thinning factor) from $1,000$ at the beginning of the algorithm to $3,000$ at the end of the algorithm for the MCMC chain to obtain a MCMC sample with less dependence.

Table 3 also shows the standard errors and p-values when not accounting for the MCMC error. Ignoring this error might lead to potentially incorrect messages in regards to significance. We believe that our marginal model approach is to be preferred when the main focus is on the effect of exogenous variables on the edge probabilities.

Our algorithm needed roughly 20 main iterations and each iteration needed roughly $10 - 50$ sub-iterations to solve for $\mathbf{\Psi}$ with the proposed least squares algorithm, in total roughly 10 hours. We often created a new MCMC sample, even though this was not needed because $\widehat{\text{Var}}_{MC}(\hat{r}_m)$ was small. As we already outlined, we did not create our own algorithm to obtain a MCMC sample, but used the `ergm` package that does not create all the output

	resiudal deviance	deviance	residual df	p-value
Null deviance	598.78	-	-	-
Covariates only (GLM)	501.80	96.98	5	0.000
marginal model	482.31	19.49	2	$5.858e-05$
curved ERGM	456.21	26.10	0	$3.241e-07$

Table 4: Deviances for null model, logistic model, MCERGM and CERGM

needed effectively.

Another issue is finding good starting values $\boldsymbol{\Psi}$ in step 2 of main algorithm for given $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$ to solve $\mathbf{g}(\boldsymbol{\Psi}) = \mathbf{0}$. Finding a $\boldsymbol{\Psi}$ such that $\mathbf{g}(\boldsymbol{\Psi}) \approx \mathbf{0}$ would speed up the algorithm dramatically.

One can start with $\boldsymbol{\eta}(\boldsymbol{\theta}^0) = \mathbf{0}$, because this implies that $\boldsymbol{\Psi}^0 = \tilde{\mathbf{X}}\boldsymbol{\beta}^0$ ($\text{expit}(\Psi_{ij}^0) = \pi_{ij} = (\text{expit}(\tilde{\mathbf{X}}\boldsymbol{\beta}^0))_{ij}$) solves $\mathbf{g}(\boldsymbol{\Psi}^0) = \mathbf{0}$. However, the next step of the main algorithm will usually require a large $\Delta\boldsymbol{\theta}$, which implies that $\boldsymbol{\Psi}^0$ is far away from solving $\mathbf{g}(\boldsymbol{\Psi}^0) = \mathbf{0}$.

It seems better to obtain an initial estimate $\boldsymbol{\theta}^0 \neq \mathbf{0}$ by fitting a CERGM (possibly including the edges statistic E as an "intercept"), because $\boldsymbol{\theta} = \mathbf{0}$ implying dyad independence seems unrealistic. But again finding a proper $\boldsymbol{\Psi}$ needed for the MCERGM is difficult.

Efficient implementation will speed up the algorithm provided $\binom{n}{2}$ is not too large, simply because a $\binom{n}{2} \times \binom{n}{2}$ matrix for estimating $\boldsymbol{\beta}$ needs to be inverted, see (11). For the Lazega (2001) data set ($n = 36$) the matrix is of size 630×630, however for large n, e.g. $n = 1,000$, inverting a $\binom{n}{2} \times \binom{n}{2}$ is not feasible with current computer technologies.

Table 4 shows the deviances for the models considered in this paper. In this example, the CERGM with the GWESP statistic provides a better fit than the MCERGM. While the fit of the CERGM is better than the fit of the MCERGM, our primary focus is on the marginal model and not on the joint model.

This situation is similar to other statistical problems, where the main interest is often on the marginal distribution. A simple model for the joint distribution is used to account for dependence between observations. For example consider longitudinal binary observations. If the focus was on the joint distribution, a log-linear model would be appropriate. Alter-

natively, a marginal logistic regression model might be preferred, and to account for the dependence among observations on the same subject one might choose to apply the GEE method to obtain parameter estimates of the logistic regression model. This can be seen as a crude way of modeling the joint distribution. Usually, the implied joint model would not represent the best possible fit. As an alternative to GEE one might apply a GLMM, which implies conditional independence, again a simple model assumption that is unlikely to describe the joint distribution correctly. The investigators usually do not give great attention to the modeling of the underlying joint distribution.

7 Discussion

In this paper, we proposed a marginal model for the marginal probabilities $\Pr(Y_{ij} = 1)$ for a network, with covariates defined by the attributes, subject to the joint distribution which is specified by a curved ERGM. Here we used a logistic link but any other standard link function is also possible. Such modeling approaches are quite common, for example FL93 proposed a marginal logistic model subject to a log-linear model describing the joint distribution and Heagerty (1999) and Wang and Louis (2004) proposed a marginal logistic regression model subject to a generalized linear mixed model. So far ML estimation has not been addressed, because it was assumed that such a problem was infeasible.

The main advantage of the approach considered here is the convenient interpretation of model parameters $\boldsymbol{\beta}$ in terms of log-odds ratio. For example, for the Lazega (2001) data set the similarity effect of gender allows interpretation of the odds of observing an edge for actors of the same sex relative to the odds of observing an edge for actors of different gender. The estimated effect was 0.7831, hence the odds of observing an edge for equal gender are $\exp(0.7831) = 2.1882$ higher than the odds for different gender. The current methodology for curved ERGM does not allow such an interpretation.

When comparing results for logistic regression and marginalized ERGM, increased stan-

dard errors relative to logistic regression can be observed for the proposed approach. This is similar to wrongly applying a linear model to clustered data ignoring the dependence within clusters. Estimated standard errors of the naive approach assuming independence will be smaller than standard errors using a linear mixed model.

Another advantage is that ML estimates of the effects of exogenous variables in a MCERGM are consistent provided the mean model is correctly specified. That is, the joint model mainly characterized by $\boldsymbol{\theta}$ can be specified incorrectly, because asymptotically ML estimates of these exogenous effects are orthogonal to those of $\boldsymbol{\theta}$. This is not the case for a CERGM, for which the joint model characterized by exogenous effects and $\boldsymbol{\theta}$ must be specified correctly.

The drawback of the proposed methodology is that the estimation method is relatively complex. We outlined some algorithms that make estimation possible, but it is very time-consuming, currently a matter of hours. However, we also outlined simple solutions to reduce the computation time. The existing MCMC sampler of the `ergm` package needs only minor modifications to make this happen. We believe modifying the current `ergm` sampler is the best option in order to make use of the seemingly hundreds of implemented network statistics and to avoid re-implementing.

Another problem of the proposed methodology is that it can only be applied for relatively small networks with small n, because a $\binom{n}{2} \times \binom{n}{2}$ matrix has to be inverted. Modern computers allow inversion of matrices with $\binom{n}{2} \approx 1,000 - 10,000$. Therefore a marginalized CERGM cannot be fitted for very large networks. The same problem applies to CERGM but to a lesser extent. CERGMs have been fitted for $n \approx 2,000$ nodes, and clearly this number of nodes is too large to fit a marginalized CERGM.

We do not claim that the proposed algorithm is numerically optimal. Hopefully future research for more efficient algorithms will be stimulated by our article. Another important question is whether multiple independent and multiple dependent (repeated) networks can be fitted. We are currently in preparation of papers to address these important questions.

We believe our marginalized CERGM approach is preferable over the existing CERGM

approach if the main focus is on exogenous effects.

Acknowledgments

We want to thank Prof. David Steel, Prof. Ray Chambers, Tung Pham, PhD and PhD candidate Sarah Neville for their helpful comments. The research was supported by the ARC discovery grant LX0883143.

References

Barndorff-Nielsen, O. (1978), *Information and Exponential Families*, New York: John Wiley.

Corander, J., Dahmstroem, K., and Dahmstroem, P. (1998), "Maximum Likelihood Estimation for Markov Graphs," Tech. rep., Department of Statistics, University of Stcokholm.

Fitzmaurice, G. M. and Laird, N. M. (1993), "A Likelihood-Based Method for Analyzing Longitudinal Binary Responses," *Biometrika*, 80, 141–151.

Frank, O. (1991), "Statistical Analysis of Change in Networks," *Statistica Neerlandica*, 45, 283–293.

Frank, O. and Strauss, D. (1986), "Markov Graphs," *Journal of the American Statistical Association*, 81, 832–842.

Geyer, C. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science*, 7, 473–511.

Handcock, M., Hunter, D. R., Butts, C., Goodreau, S. M., and Morris, M. (2010), "ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. Version 2.2-4." Available at http://statnetproject.org.

Heagerty, P. J. (1999), "Marginally Specified Logistic-Normal Models for Longitudinal Binary Data," *Biometrics*, 55, 688–698.

Holland, P. W. and Leinhardt, S. (1981), "An Exponential Family of Probability-Distributions for Directed Graphs," *Journal of the American Statistical Association*, 76, 33–50.

Hunter, D. R. and Handcock, M. S. (2006), "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

Lazega, E. (2001), *The Collegial Phenomenon: The Social Mechanism of Cooperation Among Peers in a Corporate Law Partnership*, Oxford: Oxford University Press.

Lazega, E. and Pattison, P. (1999), "Multiplexity, Generalized Exchange and Cooperation in Organizations: A Case Study," *Social Networks*, 21, 67–90.

Lee, K. and Mercante, D. (2010), "Longitudinal Nominal Data Analysis Using Marginalized Models," *Computational Statistics & Data Analysis*, 54, 208–218.

Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Morris, M., Handcock, M., and Hunter, D. (2008), "Specication of Exponential-Family Random Graph Models: Terms and Computational Aspects," *Journal of Statistical Software*, 24, 1–24.

R-Development-Core-Team (2006), "R: A Language and Environment for Statistical Computing," Available at http://www.R-project.org.

Robbins, H. and Monroe, S. (1951), "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, 22, 400–407.

Snijders, T. (2002), "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure*, 1–40.

Snijders, T., Pattison, P., Robins, G., and Handcock, M. (2006), "New Specifications for Exponential Random Graph Models," *Sociological Methodology*, 36, 99–153.

Strauss, D. and Ikeda, M. (1990), "Pseudolikelihood Eestimation for Social Networks," *Journal of the American Statistical Association*, 85, 204–212.

van Duijn, M. A. J., Gile, K. J., and 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, 52–62.

Wang, Z. R. and Louis, T. A. (2004), "Marginalized Binary Mixed-Effects Models with Covariate-Dependent Random Effects and Likelihood Inference," *Biometrics*, 60, 884–891.

Wasserman, S. and Robins, G. (2004), "An Introduction to random graphs, dependence graphs, and p*," in *Models and Methods in Social Network Analysis*, eds. Carrington, J. and Wasserman, S., Cambridge: Cambridge University Press, pp. 148–161.

A Derivation of Likelihood Equation

First we express the mean of $\mathbf{Y}$ as

$$\boldsymbol{\pi} = \mathbb{E}\mathbf{Y} = \sum_{\mathbf{y}\in Y} \mathbf{y} \exp\left(\boldsymbol{\Psi}^T\mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbf{Z}(\mathbf{y}) - \kappa\left\{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{\Psi})\right\}\right) = \partial\kappa/\partial\boldsymbol{\Psi}$$

and that of $\mathbf{Z}$ by

$$\mathbb{E}\mathbf{Z}(\mathbf{Y}) = \sum_{\mathbf{y}\in Y} \mathbf{Z}(\mathbf{y}) \exp\left(\boldsymbol{\Psi}^T\mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbf{Z}(\mathbf{y}) - \kappa\left\{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{\Psi})\right\}\right) = \partial\kappa/\partial\boldsymbol{\eta}.$$

Now we obtain

$$\partial\boldsymbol{\pi}/\partial\boldsymbol{\Psi} = \left[\sum_{\mathbf{y}\in Y} \mathbf{y}\mathbf{y}^T \exp\left(\boldsymbol{\Psi}^T\mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbf{Z}(\mathbf{y}) - \kappa\left\{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{\Psi})\right\}\right)\right] - \boldsymbol{\pi}\boldsymbol{\pi}^T = \mathbf{C}_{\mathbf{Y}},$$

$$\partial\boldsymbol{\pi}/\partial\boldsymbol{\eta} = \left[\sum_{\mathbf{y}\in Y} \mathbf{y}\mathbf{Z}(\mathbf{y})^T \exp\left(\boldsymbol{\Psi}^T\mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbf{Z}(\mathbf{y}) - \kappa\left\{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{\Psi})\right\}\right)\right] - \boldsymbol{\pi}\mathbb{E}\mathbf{Z}(\mathbf{y})^T = \mathbf{C}_{\mathbf{Y},\mathbf{Z}}$$

and

$$\partial\boldsymbol{\pi}/\partial\boldsymbol{\eta} = \left[\sum_{\mathbf{y}\in Y}\mathbf{Z}(\mathbf{y})\mathbf{Z}(\mathbf{y})^T \exp\left(\boldsymbol{\Psi}^T\mathbf{y} + \boldsymbol{\eta}(\boldsymbol{\theta})^T\mathbf{Z}(\mathbf{y}) - \kappa\left\{\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{\Psi})\right\}\right)\right] - \mathbb{E}\mathbf{Z}(\mathbf{y})\mathbb{E}\mathbf{Z}(\mathbf{y})^T = \mathbf{C}_{\mathbf{Z}}.$$

We have

$$\begin{pmatrix}\partial l/\partial\boldsymbol{\Psi}\\ \partial l/\partial\boldsymbol{\eta}\end{pmatrix} = \begin{pmatrix}\mathbf{Y}^{obs} - \mathbb{E}\mathbf{Y}\\ \mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z}\end{pmatrix} = \begin{pmatrix}\partial\boldsymbol{\pi}/\partial\boldsymbol{\Psi} & \partial\boldsymbol{\eta}/\partial\boldsymbol{\Psi}\\ \partial\boldsymbol{\pi}/\partial\boldsymbol{\eta} & \partial\boldsymbol{\eta}/\partial\boldsymbol{\eta}\end{pmatrix}\begin{pmatrix}\partial l/\partial\boldsymbol{\pi}\\ \partial l/\partial\boldsymbol{\eta}\end{pmatrix} = \begin{pmatrix}\mathbf{C}_{\mathbf{Y}} & \mathbf{0}\\ \mathbf{C}_{\mathbf{Z},\mathbf{Y}} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\partial l/\partial\boldsymbol{\pi}\\ \partial l/\partial\boldsymbol{\eta}\end{pmatrix}.$$

Therefore

$$\begin{pmatrix}\partial l/\partial\boldsymbol{\pi}\\ \partial l/\partial\boldsymbol{\eta}\end{pmatrix} = \begin{pmatrix}\mathbf{C}_{\mathbf{Y}} & \mathbf{0}\\ \mathbf{C}_{\mathbf{Z},\mathbf{Y}} & \mathbf{I}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{Y}^{obs} - \mathbb{E}\mathbf{Y}\\ \mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z}\end{pmatrix} = \begin{pmatrix}\mathbf{C}_{\mathbf{Y}}^{-1} & \mathbf{0}\\ -\mathbf{C}_{\mathbf{Z},\mathbf{Y}}\mathbf{C}_{\mathbf{Y}}^{-1} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{Y}^{obs} - \mathbb{E}\mathbf{Y}\\ \mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z}\end{pmatrix}.$$

Finally we derive

$$\begin{aligned}\begin{pmatrix}\partial l/\partial\boldsymbol{\beta}\\ \partial l/\partial\boldsymbol{\theta}\end{pmatrix} &= \begin{pmatrix}\partial\boldsymbol{\pi}/\partial\boldsymbol{\beta} & \partial\boldsymbol{\eta}/\partial\boldsymbol{\beta}\\ \partial\boldsymbol{\pi}/\partial\boldsymbol{\theta} & \partial\boldsymbol{\eta}/\partial\boldsymbol{\theta}\end{pmatrix}\begin{pmatrix}\partial l/\partial\boldsymbol{\pi}\\ \partial l/\partial\boldsymbol{\eta}\end{pmatrix}\\ &= \begin{pmatrix}\tilde{\mathbf{X}}^T\partial\boldsymbol{\pi}/\partial\boldsymbol{\eta} & \mathbf{0}\\ \mathbf{0} & \bigtriangledown\boldsymbol{\eta}(\boldsymbol{\theta})\end{pmatrix}\begin{pmatrix}\mathbf{C}_{\mathbf{Y}}^{-1} & \mathbf{0}\\ -\mathbf{C}_{\mathbf{Z},\mathbf{Y}}\mathbf{C}_{\mathbf{Y}}^{-1} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{Y}^{obs} - \mathbb{E}\mathbf{Y}\\ \mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z}\end{pmatrix},\end{aligned}$$

where $\partial\boldsymbol{\pi}/\partial\boldsymbol{\eta} = \mathrm{Diag}(\mathrm{Var}(\mathbf{Y}))^{-1} := \mathbf{D}$ yielding

$$\begin{pmatrix}\partial l/\partial\boldsymbol{\beta}\\ \partial l/\partial\boldsymbol{\theta}\end{pmatrix} = \begin{pmatrix}\tilde{\mathbf{X}}^T\mathbf{D}\mathbf{C}_{\mathbf{Y}}^{-1}(\mathbf{Y}^{obs} - \boldsymbol{\pi})\\ \bigtriangledown\boldsymbol{\eta}(\boldsymbol{\theta})\left\{-\mathbf{C}_{\mathbf{Z},\mathbf{Y}}\mathbf{C}_{\mathbf{Y}}^{-1}(\mathbf{Y}^{obs} - \boldsymbol{\pi}) + \mathbf{Z}^{obs} - \mathbb{E}\mathbf{Z}\right\}\end{pmatrix}.$$