Maximum likelihood estimation based on the Laplace approximation for p_2 network regression models

Ruggero Bellio[*], Nicola Soriani[†]

Abstract

The class of p_2 models is suitable for modelling binary relation data in social network analysis. A p_2 model is essentially a regression model for bivariate binary responses, featuring within-dyad dependence and correlated crossed random effects to represent heterogeneity of actors. Despite some desirable properties, these models are used less frequently in empirical applications than other models for network data. A possible reason for this fact may lie in the computational difficulties existing to estimate such models by means of the methods proposed in the literature, such as joint maximization methods and Bayesian methods. The aim of this paper is to investigate maximum likelihood estimation based on the Laplace approximation approach, that can be refined by importance sampling. Practical implementation of such methods can be performed in an efficient manner, and the paper provides details on a software implementation using R. Numerical examples and simulation studies illustrate the methodology.

Key words: Automatic Differentiation; Importance Sampling; Numerical Integration; Random Effects; Social Network Analysis.

Running title: Maximum likelihood estimation for p_2 models.

[*] Dipartimento di Scienze Economiche e Statistiche, Università di Udine, via Tomadini 30/A, 33100 Udine, Italy, ruggero.bellio@uniud.it

[†] Independent Researcher, nsoriani82@gmail.com

1 Introduction

The p_1 model (Holland and Leinhardt, 1981) is a classical model for directed random graphs. It has some interesting features, such as direct modelling of the within-dyad dependence, the inclusion of sender and receiver effects and the possibility of inserting covariates of various kinds, and it still attracts the attention of theoretical statisticians due to the incidental-parameter problem that affect the mathematical properties of estimators (e.g. Yan and Xu, 2013). At the same time, the p_1 model has some remarkable limitations for empirical applications, as it assumes independence between different dyads, including those involving the same actors.

The basic p_2 model was first introduced by Van Duijn (1995), used in Lazega and Van Duijn (1997), and then studied more thoroughly in Van Duijn et al. (2004). It retains all the desirable properties of the p_1 model, but it includes correlated random effects for ties sharing the same actors, resulting in more realistic assumptions. Recent surveys (Hunter et al., 2012; Snijders, 2011) include such model among those models for network data based on latent variables. Unlike other latent variable network models, such as latent cluster random effects models (Krivitsky et al., 2009), stochastic dependence is explicitly assumed for both the ties belonging to the same dyad given the random effects, as well as the sender and receiver effect for the same actor.

The p_2 model has been used much less in empirical studies of network data compared to other models, notably those belonging to the class of exponential random graph models (ERGMs) (Frank and Strauss, 1986), widely described in the aforementioned review papers; see also Robins et al. (2007) and the two recent monographs by Kolaczyk (2009) and Kolaczyk and Csárdi (2014). Nonetheless, p_2 models represent a useful and flexible class, that can be extended in various directions. Two features are probably worth mentioning. First, it is very simple to simulate networks from a given p_2 model, and this can be useful for evaluating the model fit and comparing different fitted models. Indeed, goodness of fit methods for network data models are typically based on simulation (Hunter et al., 2008). Second, and perhaps more importantly, p_2 models can be adapted to multilevel structures, such as networks formed by students in the same class, with classes nested in schools. This is the essence of the multilevel p_2 model (Zijlstra et al., 2006), used for example in Vermeij et al. (2009).

In the paper that first thoroughly analysed the p_2 model, Van Duijn et al. (2004) estimated the model parameters by a Marginal Quasi Likelihood (MQL) approach (Breslow and Clayton, 1993; Goldstein, 1991), also considered by Zijlstra et al. (2009). The literature on random effects clearly points out that joint maximization methods, such as MQL and PQL, may perform poorly for nonlinear models and discrete data (Molenberghs and Verbeke, 2005, Ch. 14), and this was clearly shown in the simulation studies reported in Zijlstra et al. (2009). Indeed, the latter authors proposed a Bayesian approach, sampling from the posterior distribution by Markov Chain Monte Carlo (MCMC) methods. In particular, they used a slightly informative prior for model parameters, and compared several sampling algorithms. The results obtained with such approach were good, even when assessed from a frequentist perspective. Nonetheless, a maximum likelihood approach may appeal to many users. The maximum likelihood approach is straightforward to apply and fast, and it may also have good scalability properties, with the possibility to readily estimate models for large networks.

The aim of this paper is to demonstrate that the implementation of approximate maximum likelihood estimation of p_2 models using open-source statistical software is a feasible task. The resulting software is then used to study the properties of the proposed estimation method, both by means of some numerical examples and with simulations studies. The latter include a comparison with the numerical results of Zijlstra et al. (2009), and some simulations with larger networks.

The plan of the paper is as follows. Section 2 gives some background details, and Section 3 illustrate approximated maximum likelihood estimation of the model parameters. Section 4 gives some details on the computational implementation, and in particular on the `R` (R Core Team, 2019) package `p2model` that can be used to apply the methods proposed in this paper. Section 5 illustrates some examples based on some well-known data sets, and Section 6 reports the results of some simulation studies. Some concluding remarks are provided in the final section.

2 Background

A directed graph represents the ties between a certain set of nodes, that in social networks represent different actors. A pair of actors, along with their realized ties, is a dyad. Here the focus is on binary ties, represented by the variables y_{ij}, with 1 denoting the presence of a tie from actor i to actor j, with $i, j = 1, \dots, g$.

The basic assumption of the p_2 model is that for a given dyad

$$\begin{aligned} & P(Y_{ij} = y_1, Y_{ji} = y_2 | a_i, b_i, a_j, b_j) \\ = & \frac{\exp\{y_1 (\mu_{ij} + \alpha_i + \beta_j) + y_2 (\mu_{ij} + \alpha_j + \beta_i) + \rho_{ij} \, y_1 y_2\}}{1 + \exp(\mu_{ij} + \alpha_i + \beta_j) + \exp(\mu_{ji} + \alpha_j + \beta_i) + \exp(\mu_{ij} + \mu_{ji} + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho_{ij})} \, , \end{aligned} \tag{1}$$

where y_1 and y_2 are binary values, α_i is the sender parameter of actor i, β_i is receiver parameter of actor i, whereas μ_{ij} and ρ_{ij} are the density and the reciprocity parameters for dyad (i, j). These parameters depend on some covariates and random effects in the following way

$$\alpha_i = x_{1i}^\top \gamma_1 + a_i \, , \qquad \beta_i = x_{2i}^\top \gamma_2 + b_i \, , \qquad \mu_{ij} = \mu + z_{1ij}^\top \delta_1 \, , \qquad \rho_{ij} = \rho + z_{2ij}^\top \delta_2 \, .$$

Here X_1 and X_2 are design matrices with g rows for covariates measured on actors, whereas Z_1 and Z_2 contain dyad-specific covariates. More precisely, Z_1 is a three-dimensional array of size $k_d \times g \times g$, obtained by stacking together k_d matrices of size $g \times g$. Similarly, Z_2 is a three-dimensional array of size $k_c \times g \times g$, but notice that the k_c matrices forming it are always symmetric. The vectors $u_i = (a_i, b_i)^T$ are random effects, which are assumed to be normally distributed independent random variables. Namely,

$$U_i \sim N_2(0, \Sigma) \, , \qquad \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \, . \tag{2}$$

The inclusion of random effects induces a correlation among all the ties sent or received for a given actor. Moreover, the two random effects for the same actor, entering the linear predictor for sender and receiver parameters respectively, are allowed to be correlated. All the different parameters of this model are collected together in the vector θ

$$\theta = (\gamma_1^T, \gamma_2^T, \mu, \delta_1^T, \rho, \delta_2^T, \sigma_A^2, \sigma_{AB}, \sigma_B^2)^T \, .$$

Van Duijn et al. (2004) and Zijlstra et al. (2009) noticed that the p_2 model is conveniently formulated as a multinomial regression model with random effects. Although this fact is important and it may be useful for some computer implementations, it will not be exploited in what follows. Instead, note that from (1) we can readily obtain both the distribution of a given component, say Y_{ij}, as well as the conditional distribution of the other component (Y_{ji}) given the first one, keeping in either case the random effects as fixed. Both these two results are simple since the two response variables are binary. The resulting formulas are

$$P(Y_{ij}=1|a_i,b_i,a_j,b_j) \quad = \quad \frac{\exp{(\eta_{ij})}+\exp{(\eta_{ij}+\eta_{ji}+\rho_{ij})}}{1+\exp{(\eta_{ij})}+\exp{(\eta_{ji})}+\exp{(\eta_{ij}+\eta_{ji}+\rho_{ij})}} \tag{3}$$

$$P(Y_{ji}=1|Y_{ij}=y_1,a_i,b_i,a_j,b_j) \quad = \quad \frac{\exp\{y_1\,\eta_{ij}+\eta_{ji}+y_1\,\rho_{ij}\}}{\exp\{y_1\,\eta_{ij}\}+\exp\{y_1\,\eta_{ij}+\eta_{ji}+y_1\,\rho_{ij}\ \}}\,, \tag{4}$$

where

$$\eta_{ij} \quad = \quad \mu_{ij}+\alpha_i+\beta_j\,.$$

A remarkable consequence of (3)-(4) is that simulation of a network from a p_2 model is quite simple, i.e. after simulating the random effects, it just requires the simulation of two binary variables for each dyad.

3 Maximum likelihood estimation

The likelihood function for the model defined by (1) and (2) is obtained by integrating out the random effects. After setting

$$p(y|u;\theta)=\prod_{i=1}^{g-1}\prod_{j=i+1}^{g}p(y_{ij},y_{ji}|u_i,u_j)\,,$$

where $p(y_{ij},y_{ji}|u_i,u_j)$ corresponds to (1), it follows that

$$L(\theta)=\int_{{\rm I\!R}^{2\,g}}p(y|u;\theta)\,\left\{\prod_{i=1}^{g}\phi_2(u_i;0,\Sigma)\right\}du\,, \tag{5}$$

being $\phi_2(\cdot)$ the bivariate normal density (2).

3.1 Estimation based on the Laplace approximation

The high-dimensional integral in (5) can only be evaluated numerically, as the correlated random effects u_i have a crossed structure, therefore there is no possible reduction of the dimension of integration. A doable approach to the numerical approximation to (5) is via the Laplace's method of integration, as proposed in Skaug (2002). Let $h(u; \theta, y)$ be defined as

$$h(u; \theta, y) = \log p(y|u; \theta) + \sum_{i=1}^{g} \log \phi_2(u_i; 0, \Sigma) ,$$

the first-order Laplace approximation to $L(\theta)$ is given by

$$L^*(\theta) = \exp\{h(\widehat{u}_\theta; \theta, y)\} \, |H(\theta)|^{-1/2} , \tag{6}$$

with $\widehat{u}_\theta = \underset{u}{\text{argmax}} \; h(u; \theta, y)$ and $H(\theta)$ given by minus the Hessian matrix of $h(u; \theta, y)$ at the maximum

$$H(\theta) = - \left. \frac{\partial^2 h(u; \theta, y)}{\partial u \, \partial u^T} \right|_{u=\widehat{u}_\theta} .$$

3.2 Estimation properties

There exists some encouraging published evidence about the good performances of the estimator $\widehat{\theta}^*$ of θ obtained from the maximization of $L^*(\theta)$ for mixed models with crossed random effects structures (Noh and Lee, 2007; Skaug, 2002). Besides, the recent paper by Ogden (2017) provides some theoretical support about the properties of $\widehat{\theta}^*$. Using the results of the latter paper, together with those in Shun and McCullagh (1995), it is possible to verify that

i) $\widehat{\theta}^*$ is consistent when $g \to \infty$, being the number of nodes the right asymptotic index for network data (Krivitsky and Kolaczyk, 2015);

ii) $\widehat{\theta}^*$ has the same limiting distribution as the exact MLE $\widehat{\theta}$ which maximizes (5).

The implications of the above properties is that one can safely adopt $\widehat{\theta}^*$ as the estimator of choice for p_2 models, and estimating its standard error using the observed information matrix obtained from $\log L(\theta)$. Further usages of the latter quantity include computation of likelihood ratio tests and likelihood-based model selection criteria, that are also covered by the results in Ogden (2017).

Some residual concerns about $\widehat{\theta}^*$ may arise for small-sized networks, where it might be sensible to compare it with another estimator which is theoretically superior. A possible route to this is to approximate $L(\theta)$ by importance sampling, following Skaug (2002) and Brinch (2012). The latter author provided the apt denomination *explicitly parameter dependent Laplace importance sampling.* The idea is to take as the importance sampling distribution a normal distribution with mean vector $\widehat{u}_\theta$ and covariance matrix $H(\theta)^{-1}$. If $u^{(1)}, \ldots, u^{(M)}$ is random sample of size M from such a distribution, the resulting approximation is given by

$$L^\dagger(\theta) = \frac{1}{M} \sum_{j=1}^{M} \frac{\exp\{h(u^{(j)}; \theta, y)\}}{\phi_{2g}\{u^{(j)}; \widehat{u}_\theta, H(\theta)^{-1}\}} \, . \tag{7}$$

The random draws $u^{(j)}$ can be generated as

$$u^{(j)} = \widehat{u}_\theta + C(\theta)\, v^{(j)} \, ,$$

where $C(\theta)$ is the Cholesky factor of $H(\theta)^{-1}$ and $v^{(j)}$ is a vector of independent standard normal draws, so that $L^\dagger(\theta)$ can be conveniently expressed in a form suitable for direct computation

$$L^\dagger(\theta) = |C(\theta)| \, \exp\{h(\widehat{u}_\theta; \theta, y)\} \, \frac{1}{M} \sum_{j=1}^{M} \exp\left[h\{\widehat{u}_\theta + C(\theta)\, v^{(j)}; \theta, y\} - h(\widehat{u}_\theta; \theta, y) + 0.5\, \|v^{(j)}\|^2\right] \, . \tag{8}$$

Following Skaug (2002), in order to facilitate the maximization of $L^\dagger(\theta)$ it is advisable to use the same set of random draws $v^{(1)}, \ldots, v^{(M)}$ to generate $u^{(1)}, \ldots, u^{(M)}$, for all values of θ. The maximization of $\log L^\dagger(\theta)$ can then be repeated for various choices of M, to check whether the resulting estimates become stable for large values of M.

3.3 Penalized estimation

In our own experience, estimated p_2 models will often have estimated covariance of sender and receiver random effects σ_{AB} with negative sign. This is also found in several published results, though there are also instances of positive estimated correlation, as will be seen in the second example of the following section. Occasionally, it is possible to encounter data sets where the estimate of the correlation $\rho_{AB} = \sigma_{AB}/(\sigma_A\,\sigma_B)$ is extreme, with estimated random effects matrix close to singularity. The implications of this fact are twofold. First, the fact that the estimated Σ is close to singular complicates the parameter estimation, since it is typically associated to a profile likelihood function that is nearly flat for the parameters that specify Σ. This can be overcome with some care in the implementation, following for example Bates et al. (2015, Section 4), but it may lead to occasional failure of fitting routines. Another consequence of (almost) singular estimated Σ is that the empirical Bayes prediction of random effects would lie on a line, which is somewhat unattractive.

A possible resolution to these issues is suggested by Chung et al. (2015), that propose to penalize the log-likelihood by a weakly informative prior. Namely, they suggest to maximize the log-likelihood plus a penalty function for Σ,

$$\log L(\theta) + \log p(\Sigma)\,, \tag{9}$$

where for the case of two random effects a possible default choice for $p(\Sigma)$ is

$$p(\Sigma) \propto |\Sigma|^{1/2} = \sigma_A\,\sigma_B\,\sqrt{1-\rho_{AB}^2}\,.$$

Such a modification is of order $O(1)$, thus it does not alter the asymptotic properties of the MLE, and the estimator defined as the maximizer of (9) will be typically close to the ordinary MLE; indeed, in all the experiments we made, the inclusion of the penalty had a modest effect on the model estimates. At the same time, the penalty prevents the maximizer of (9) to achieve a maximum at $\rho_{AB} \pm 1$. Although Chung et al. (2015) consider the case of linear mixed models, the same properties are valid more generally. We suggest the recourse to the penalized estimator whenever the estimated Σ is nearly singular, and the `R` software accompanying this article has options for this. Finally, note that Chung et al. (2015) also illustrate how to employ more informative penalties for Σ to incorporate additional prior information, moving closer towards Bayesian inference while remaining within the frequentist realm.

4 Software implementation

The methods of the previous sections have been implemented in the `R` package `p2model`. The package, which is available at `https://github.com/rugbel/p2model`, makes use of software for Automatic Differentiation, as endorsed by Cudeck (2005), and first used for random effects models by Skaug (2002) and Skaug and Fournier (2006). The usage of such software in `p2model` is totally transparent to the user, and this should make the usage of the package appealing also to social scientists.

In particular, the `R` package `Template Model Builder (TMB)` (Kristensen et al., 2016) has been employed for approximate maximum likelihood estimation based on the Laplace approximation and Laplace importance sampling. This package, available at the CRAN (see also `https://github.com/kaskr/adcomp`) operates by means of `C++` templates implementing the log-likelihood function, taking advantage of many available options which greatly simplify the task with respect of full implementation in a low-level programming language. `TMB` is well integrated with `R`, and it is particularly effective for large models. At the time of writing, the Laplace importance sampler automatically provided by `TMB` is not exactly the same one described in Section 3, and so our own implementation is made available in the package and it has been employed for the examples that follow. The bulk of the computation of (8), consisting in obtaining the estimated effects $\widehat{u}_\theta$, the Cholesky factor $C(\theta)$ and the function $h(\cdot)$ is based on functions supplied by `TMB`.

Automatic differentiation software has two distinct useful features for the implementation of the methods of interest in this article. First, estimation of θ based on $L^*(\theta)$ is very fast and computationally efficient, taking advantage of numerical methods for sparse linear systems implemented in the software. This ensures a good degree of scalability, and actually the software provided can readily fit models to networks with about 1,000 nodes or even larger, a task that would have been unfeasible for several other implementations. The second important feature is the possibility to easily extend the basic model formulation, to include for example multilevel p_2 models (e.g. Vermeij et al., 2009; Zijlstra et al., 2006), a point we will return on in the concluding section.

5 Data examples

5.1 High-tech managers data

As a first example, we consider the high-tech managers data (Krackhardt, 1987), already used by several authors, including Wasserman and Faust (1994). The data are available within the `statnet` project (Handcock et al., 2003), and they are about the friendship relations among 21 managers of a firm. The data set includes four actor attributes, namely age in years (`Age`), years spent in the organization (`Tenure`), level in the corporate hierarchy (`Level`) and department of the employer (`Department`). The latter two are categorical variables.

A plausible model for this data set has been fitted using the methods of Section 3, and the results are reported in Table 1. The fitted model includes both sender and receiver covariate effects, along with some density effects.

[Table 1 about here.]

In this example, approximate maximum likelihood estimates obtained from $L^{\dagger}(\theta)$ stabilizes very quickly with the value of M, and actually very little variation is found in both the estimates and the maximized likelihoods obtained with M in the range 1,000-50,000. The results are very close to those based on $L^{*}(\theta)$, confirming the theoretical properties of the latter method.

Figure 1 reports a plot of the network along with the estimated sender and receiver effects, given by the empirical Bayes estimates of the random effects. The latter are computed by replacing θ by $\widehat{\theta}^{*}$ in the expression of $\widehat{u}_{\theta}$. The plot was inspired by the proposal in Thiemichen et al. (2016, Figure 2), who introduced an interesting model for undirected networks combining together the exponential random graph specification with nodal random effects.

[Figure 1 about here.]

Here the estimated correlation is $\widehat{\rho}^{*}_{AB} = -0.79$ (0.20), which is substantial, though not so much as to suggest the recourse to penalized estimation. The negative correlation between sender and receiver effects is apparent from the shading of the nodes in Figure 1. It should be noted, however, that the estimated sender and receiver effects do not simply mirror the out-degree and in-degree distribution of the network, respectively. This is demonstrated by Figure 2, displaying two caterpillar plots of estimated random effects. The effects are sorted by increasing size of the out- and in-degree respectively, showing that whereas the estimated sender effects follow the out-degree distribution to a good extent, this is less the case for the estimated receiver effects. One of the reasons for this is the inclusion of actor covariates in the model.

[Figure 2 about here.]

We make use of this example to illustrate the important point of goodness-of-fit evaluation of a fitted p_2 model. Goodness-of-fit procedures for network data are based on the seminal work by Hunter et al. (2008), that developed a procedure for evaluating the model goodness of fit by simulation. The output of the procedure consists in some plots, which have been produced here using the estimates based on $L^*(\theta)$, as reported in Figure 3. The figure was obtained by adapting a portion of the `statnet` suite of `R` packages for network analysis (Handcock et al., 2003), and in particular the `gof.ergmm` function of the `latentnet` package (Krivitsky and Handcock, 2008; Krivitsky and Handcock, 2018). The plots are obtained by simulating several networks from the fitted model and then estimating the sample distribution of some network statistics. The observed value of these statistics are displayed as solid lines, and they are then compared with the simulated distributions, with the latter summarized by means of boxplots. The first two plots show that the in- and out-degree distributions are adequately captured by the sender and receiver effects. The other plots refer to other summary statistics, including some that are not directly parameterized by the p_2 model. Yet the goodness of fit seems acceptable, pointing to the capability of the model to capture higher-order dependences, at least for the data at hand.

[Figure 3 about here.]

5.2 Lazega friendship network

We consider here the Lazega's associates friendship network (Lazega, 2001), fitting the same three models reported in Van Duijn et al. (2004). In particular, Model 0 is an empty model including the intercepts μ, ρ and the variance terms, while Model 1 and Model 2 contain also some terms for density and reciprocity effects. Moreover, the networks on advice and collaboration are set as covariates for the density parameter in Model 2. More details on the model specification can be found in Van Duijn et al. (2004).

The results obtained from the approximate maximum likelihood methodology proposed in this paper are compared with those obtained with the `StOCNET` software (Boer et al., 2006) for Windows. This software is capable of estimating a p_2 model using both the MQL approach and the Bayesian approach of Zijlstra et al. (2009). Among the methods implemented by `StOCNET`, we report in Table 2 the results based on the IGLS-3 algorithm for the MQL approach, and on the MCMC Random Walk sampler for the Bayesian approach. The table includes also the estimates obtained with the Laplace methods, where M=10,000 was used for the importance sampling in $L^{\dagger}(\theta)$. Note that for Model 1 the density covariate defined as the difference of sending and receiving actor seniority values has been dropped due to collinearity problems (which are an issue for all the methods), retaining only the absolute differences of the same quantities. Therefore, the comparison with the results of Van Duijn et al. (2004) has to be taken with some care.

[Table 2 about here.]

We observe that regression coefficients estimated by MQL are generally attenuated with respect to the other methods. The results for the remaining methods are instead in good agreement. Plots on the model goodness of fit for all the three models are included in the Supplementary Materials.

5.3 Dutch social behavior study

As a final example, we consider the data from the Dutch Social Behavior Study (Baerveldt and Snijders, 1994), already analysed in Baerveldt et al. (2004) and Zijlstra et al. (2005). The interest was on a social network of reported received emotional support among a group of high-school students. Zijlstra et al. (2005) employed a Bayesian approach based on MCMC, taking the first network of 62 students as calibration sample used to obtain prior distributions for the analysis sample, the second network of 39 students. In particular, these authors first fitted a Bayesian model with diffuse priors for the calibration sample, and then used the results to define moderately informative priors for the analysis sample. Model selection was performed for the analysis sample using the Bayes factor, selecting a model ('Model 4') among a set of five possible alternative models. Here we replicate their analysis for what concerns model selection following the likelihood approach, using only the analysis sample.

[Table 3 about here.]

It is found that importance sampling performs a modest adjustment to the standard Laplace approximation, especially for the estimation of variance parameters. Maximum likelihood estimates obtained from $L^{\dagger}(\theta)$ are similar for a broad set of values for M, and actually little variation is found in both the estimates and the maximum log likelihood values for M in the range 5,000-50,000. Table 3 reports the maximized log likelihood values along with AIC and BIC values for the five models defined in Zijlstra et al. (2005), with the BIC computed using the number of nodes as the sample size. A comforting finding is that likelihood-based model selection points to the same model selected by the Bayesian method used by Zijlstra et al. (2005), as Model 4 has the lowest AIC (and BIC) values with either $L^{*}(\theta)$ or $L^{\dagger}(\theta)$. Plots on the model goodness of fit of Model 4 are included in the Supplementary Materials.

6 Simulation studies

The examples of the previous sections suggest that the two approximate maximum likelihood methods give very similar results, and tend to be close to the results obtained with the Bayesian approach, while much larger differences exist with respect to the MQL method. Further information on the properties of the proposed methodology can be gleaned from some simulation studies.

6.1 Simulation study for small-size networks

At first, we replicated the simulation study of Zijlstra et al. (2009). These authors considered three model settings for two network sizes (20 and 40 nodes). Following their description, Model 1 is an empty model, with density and reciprocity parameters equal to $\mu = -2$ and $\rho = 2$ respectively, and independent standardized random effects. Model 2 is similar to Model 1, but it also has a dyadic covariate for the density and a sender covariate. The density covariate has regression coefficient 0.5, and it is a network (*net1*) generated from Model 1. The sender covariate has a regression coefficient 0.05, and it equals the actor's rank number $(1, \ldots, g)$. Model 3 has a receiver covariate, two density covariates and one reciprocity covariate. The receiver covariate has regression coefficient -0.1, and it is a binary variable drawn from a coin flip. The first density covariate is the same used in Model 2, (*net1*), and it has regression coefficient 0.5. The second density covariate (*fc*) has regression coefficient 0.2, and it contains the absolute differences of an actor covariate drawn from a multinomial distribution having as sample space the set $\{1, 2, 3, 4, 5\}$ and five equal probabilities. The latter variable is also used as reciprocity covariate, with regression coefficient 0.05. The random effects in Model 3 are negatively correlated ($\sigma_{AB} = -0.5$), with sender variance $\sigma_A^2 = 1.5$, and receiver variance $\sigma_B^2 = 0.75$.

The results for the simulation studies focus in particular on the approximate maximum likelihood methods, taking for the sake of comparison the results from a MQL-type method (RIGLS-3) and a Bayesian one (RW) from Zijlstra et al. (2009). The results for Model 3 are reported in Figure 4, that displays the relative bias of each estimator (that is, the bias divided by the absolute value of the true parameter value), along with the root mean squared error. The results for the other two models are not included here, since they are very similar to those of Model 3.

From these results, it is apparent that the MQL approach produces estimates with very large bias, that renders such method totally unappealing for practical use. This is not surprising, and relevant with a vast body of literature on random effects modeling. Both the Bayesian approach and our proposals seem instead to perform well, with no appreciable differences between the Laplace and the Laplace Importance Sampling methods. At times the two approximate maximum likelihood estimation methods appear to be slightly more efficient than the Bayesian method, but in general one can safely say that they are largely comparable. As a general trend, the results improve considerably for larger networks, for all the methods but MQL. The coefficient of the reciprocity covariate (*fc*) is the most difficult to estimate, and here the Laplace and the Laplace Importance Sampling methods have an edge over the Bayesian method.

[Figure 4 about here.]

Furthermore, using the simulation results it is also possible to estimate the accuracy of standard errors computed using the observed information matrix for the two approximate maximum likelihood estimation methods. To this end, the estimated standard errors seem to be a reliable approximation to the standard deviation of parameter estimates, since the average ratio between mean of estimated standard errors and standard deviation of parameter estimates is equal to about 0.97 for either method with $g = 20$, and about 0.98 with $g = 40$.

6.2 Simulations for large networks

As a further study, we focus on moderately large networks, considering only the simple method based on the first-order Laplace approximation. In particular, 1,000 networks of size $g = 200$ and $g = 400$ were generated by the same setting of Model 3 of the previous section, and the $\hat{\theta}^*$ estimator was obtained. The estimator based on $L^\dagger(\theta)$ has not been considered, since for network large network sizes it is virtually indistinguishable from $\widehat{\theta}^*$. Figure 5 reports the empirical distribution for μ, ρ and for the three elements of Σ, along with the parameter with the largest bias in the previous study, namely the coefficient of the reciprocity covariate *fc*. It is apparent that the bias for the reciprocity coefficient observed for $g = 20, 40$ disappears for $g = 200$ and $g = 400$; similar results were obtained for the other coefficients. Likewise, the marked asymmetry in the finite-sample distribution of the estimators of variance components visible for $g = 20, 40$ nearly disappears with larger network size. The ratio between mean of estimated standard errors and standard deviation of parameter estimates is very close to 1 for larger sizes, confirming that the standard errors suggested by maximum likelihood estimation theory are rather accurate for $\hat{\theta}^*$.

[Figure 5 about here.]

7 Concluding remarks

The results obtained with the approximate maximum likelihood estimation methods based on the Laplace approximation for the class of p_2 models are rather encouraging. Indeed, the simple simulation-free approach given by the first-order Laplace approximation seems to perform rather well, and it surely constitutes a fast and simple solution for fitting the p_2 model.

The real advantage of our proposal with respect to the Bayesian approach is mainly simplicity of usage and computational efficiency. On the other hand, the Bayesian approach facilitates the usage of prior information when this is available, and the use of weakly informative priors may help in those cases when the estimated matrix of random effects is close to singularity. We acknowledge that the usage of the Bayesian approach may be appealing to some researchers, but others may find conceptually simpler a well-performing frequentist method like maximum likelihood estimation. For practical implementation, Bayesian methods can be implemented using some publicly-available software, such as the BUGS engine (see Lunn et al., 2000); another possibility is the recent `R` package `dyads` (Zijlstra, 2017). As further point, notice that the Bayesian approach may also be the most natural resolution for extending the model in those cases where the assumption of normality for random effects is deemed to be a limitation.

As illustrated in the `p2model` package that accompanies this paper, the methods proposed here can be simply implemented using freely available and multi-platform software. Both the Bayesian approach and the Laplace-based ones can be extended to more complex data structures, such as the multilevel data set studied in Vermeij et al. (2009). Such task would be feasible for the approach proposed here, starting from the `TMB` template which is part of the `p2model` package. An interesting open research topic would be the comparison of the multilevel p_2 model with alternative approaches, such as the Hierarchical Network Model introduced in Sweet et al. (2013), or other approaches for multilevel structures; see Snijders (2016) for further details.

This work was mainly focused on the small-size networks already proposed in the literature for p_2 models. As stated in Section 4, fitting p_2 models to networks with a few hundreds of nodes, or even a few thousands ones, is not an issue using the implementation adopted here, and indeed this was demonstrated in the second simulation study. The practical relevance of the p_2 model for such large scale networks is open to discussion, yet the scalability of the procedure surely instils some confidence for the extension to complex settings, such as those where several covariates are available.

The p_2 models have some important features, including simplicity of interpretation and proximity with random effects models. Another useful feature is the possibility of generating data in a very straightforward manner, which can be used for goodness-of-fit procedures. Regarding the latter point, there will be instances where the p_2 model may provide a fit less good than competing models, such as ERGM models, as the latter may include some terms that were explicitly developed to capture high-order dependences and transitivity. Despite this fact, the higher simplicity of the p_2 models makes them appealing nonetheless, and the availability of reliable estimation methods may lead to a more widespread usage of these models by practitioners.

Acknowledgements

We would like to thank dr. Bonne Zijlstra for providing us the data of the Dutch Social Behavior Study, and dr. Chris Baerveldt for the permission to their usage. A very preliminary version of this work was presented at the 7th Conference on Statistical Computation and Complex Systems, held in Padua in September 2011. The authors are grateful to the Editors and the Referees for their detailed comments and suggestions that led to an improved version of the paper. This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003.

References

Baerveldt, C. and Snijders, T. A. B. (1994). Influences on and from the segmentation of networks: Hypotheses and tests. *Social Networks*, 16(3):213–232.

Baerveldt, C., Van Duijn, M. A. J., Vermeij, L., and Van Hemert, D. A. (2004). Ethnic boundaries and personal choice. Assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils' networks. *Social Networks*, 26(1):55–74.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using `lme4`. *Journal of Statistical Software*, 67(1):1–48.

Boer, P., Huisman, M., Snijders, T. A. B., Steglich, C., Wichers, L. H. Y., and Zeggelink, E. P. H. (2006). *StOCNET: An Open Software System for the Advanced Statistical Analysis of Social Networks. Version 1.7.*

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

Brinch, C. (2012). Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling. *Computational Statistics*, 27(1):13–28.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics*, 40(2):136–157.

Cudeck, R. (2005). Fitting psychometric models with methods based on automatic differentiation. *Psychometrika*, 70(4):599–617.

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

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78(1):45–51.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2003). statnet: Software tools for the statistical modeling of network data. Seattle, WA.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76(373):33–65.

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

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

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

Kolaczyk, E. D. and Csárdi, G. (2014). *Statistical Analysis of Network Data with R*. Springer.

Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, 9(2):109–134.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. J., and Bell, B. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21.

Krivitsky, P. N. and Handcock, M. S. (2008). Fitting position latent cluster models for social networks with `latentnet`. *Journal of Statistical Software*, 24(5).

Krivitsky, P. N. and Handcock, M. S. (2018). *latentnet: Latent Position and Cluster Models for Statistical Networks*. The Statnet Project (`http://www.statnet.org`). R package version 2.9.0.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213.

Krivitsky, P. N. and Kolaczyk, E. D. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statistical Science*, 30(2):184–198.

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

Lazega, E. and Van Duijn, M. A. J. (1997). Position in formal structure, personal characteristics and choices of advisors in a law firm: a logistic regression model for dyadic network data. *Social Networks*, 19(4):375–397.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Comput-*

ing, 10(4):325–337.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, 98(5):896–915.

Ogden, H. E. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika*, 104(1):153–164.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks*, 29(2):173 – 191.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B: Methodological*, 57(4):749–760.

Skaug, H. J. (2002). Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. *Journal of Computational and Graphical Statistics*, 11(2):458–470.

Skaug, H. J. and Fournier, D. A. (2006). Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics and Data Analysis*, 51(2):699–709.

Snijders, T. A. B. (2011). Statistical models for social networks. *Annual Review of Sociology*, 37:131–153.

Snijders, T. A. B. (2016). The multiple flavours of multilevel issues for networks. In Lazega, E. and Snijders, T. A. B., editors, *Multilevel Network Analysis for the Social Sciences; Theory, Methods and Applications*, volume 12 of *Methodos Series (Methodological Prospects in the Social Sciences)*, pages 15–46. Springer.

Sweet, T. M., Thomas, A. C., and Junker, B. W. (2013). Hierarchical network models for education research: Hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, 38(3):295–318.

Thiemichen, S., Friel, N., Caimo, A., and Kauermann, G. (2016). Bayesian exponential random graph models with nodal random effects. *Social Networks*, 46:11–28.

Van Duijn, M. A. J. (1995). Estimation of a random effects model for directed graphs. In Snijders, T. A. B., editor, *SSS'95. Symposium Statistische Software, nr. 7. Toeval zit overal: programmatuur voor random-coëfficiënt modellen*, pages 113–131. ProGAMMA, Groningen.

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

Vermeij, L., Van Duijn, M. A. J., and Baerveldt, C. (2009). Ethnic segregation in context: social discrimination among native dutch pupils and their ethnic minority classmates. *Social Networks*, 31(4):230–239.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, New York.

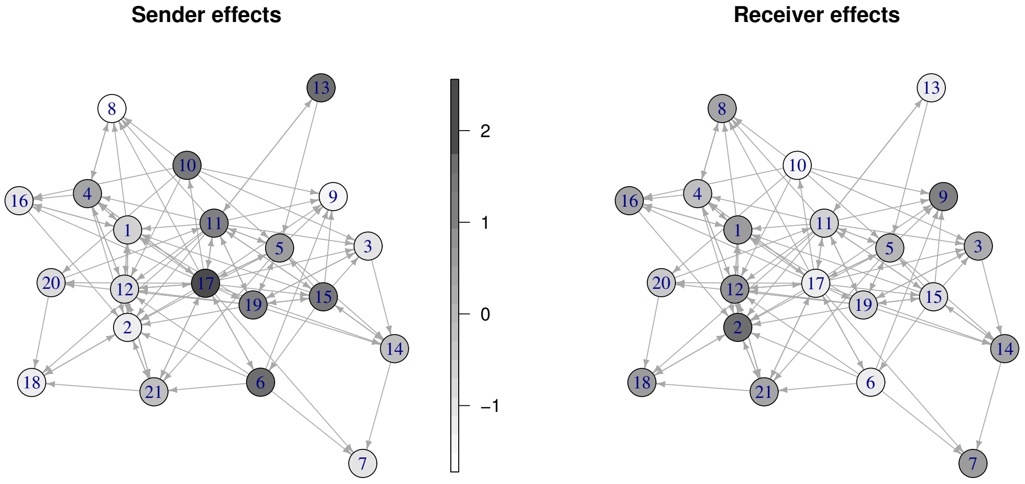
Yan, T. and Xu, J. (2013). A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika*, 100(2):519–524.

Zijlstra, B. J. H. (2017). *dyads: Dyadic Network Analysis*. R package version 1.1.

Zijlstra, B. J. H., Van Duijn, M. A. J., and Snijders, T. A. B. (2005). Model selection in random effects models for directed graphs using approximated Bayes factors. *Statistica Neerlandica*, 59(1):107–118.

Zijlstra, B. J. H., Van Duijn, M. A. J., and 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.

Zijlstra, B. J. H., Van Duijn, M. A. J., and Snijders, T. A. B. (2009). MCMC estimation for the p_2 network regression model with crossed random effects. *British Journal of Mathematical and Statistical Psychology*, 62(1):143–166.



Sender effects

Receiver effects

Figure 1: `high-tech managers` data: Network representation with nodes shaded according to the size of estimated sender and receiver effects.

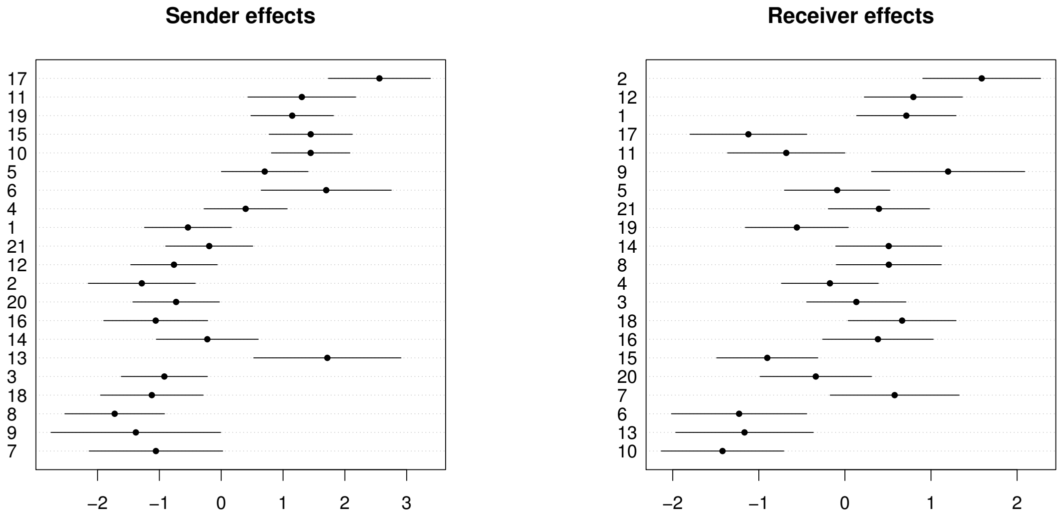


Figure 2: `high-tech managers` data: Caterpillar plots of sender and receiver estimated random effects, with segments extending to ± 1 estimated standard errors. The effects are sorted by observed out- and in-degree sizes, respectively.

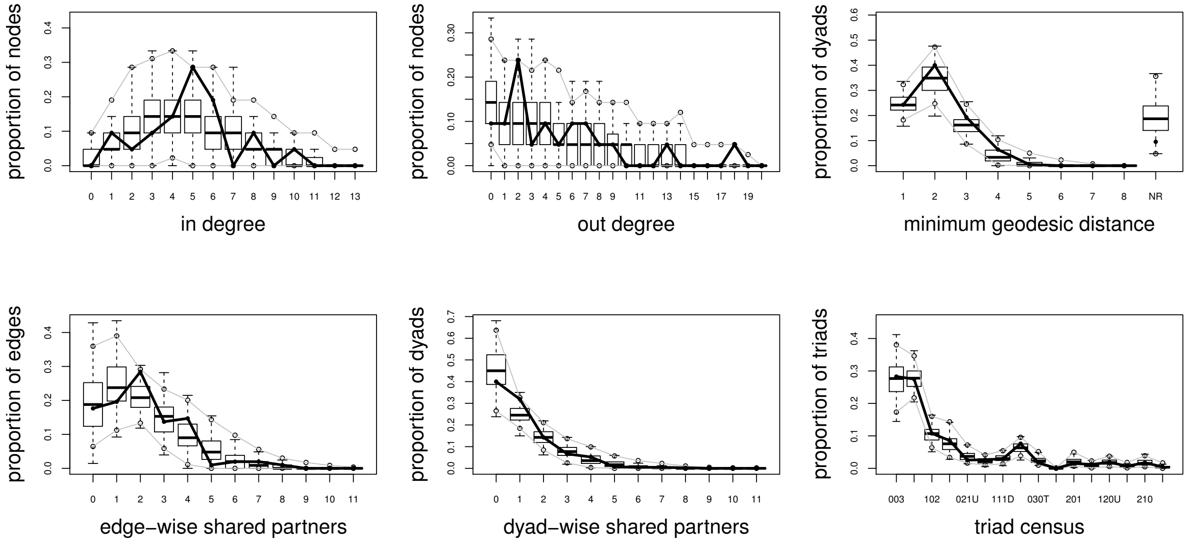


Figure 3: `high-tech managers` data: goodness-of-fit plots based on 100 simulated networks. Solid lines show the observed value of some network statistics, and the boxplots summarize their empirical distribution obtained from the simulated samples. The outer gray lines highlight the range where 95% of the simulated values fall.

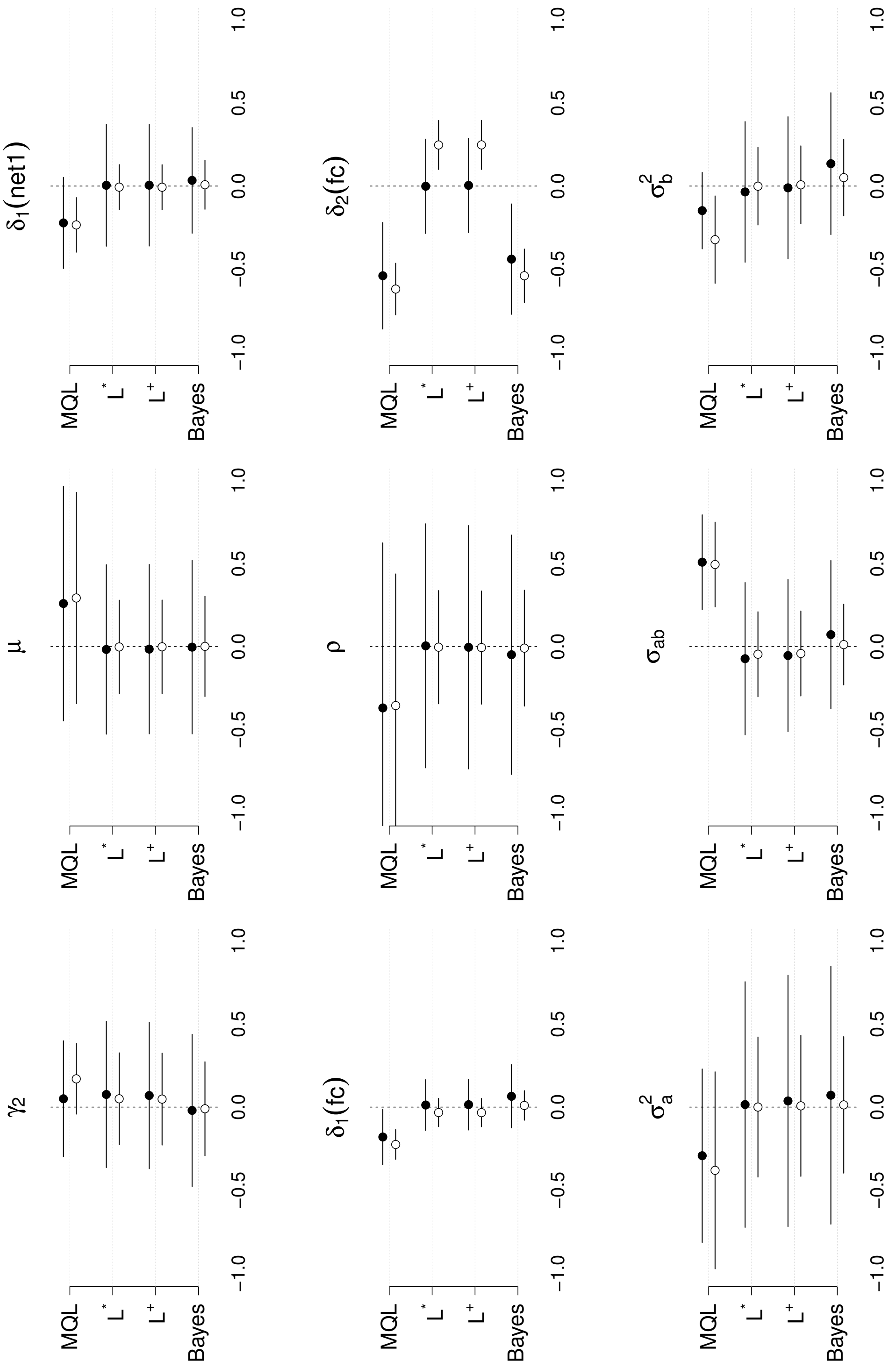


Figure 4: Summary of simulation results for Model 3. Relative bias and root mean squared error for each parameter, for various methods. Full circles are for $g = 20$ and empty circles for $g = 40$. Based on 1,000 simulations, with the results for the MQL and Bayesian methods taken from Zijlstra et al. (2009).

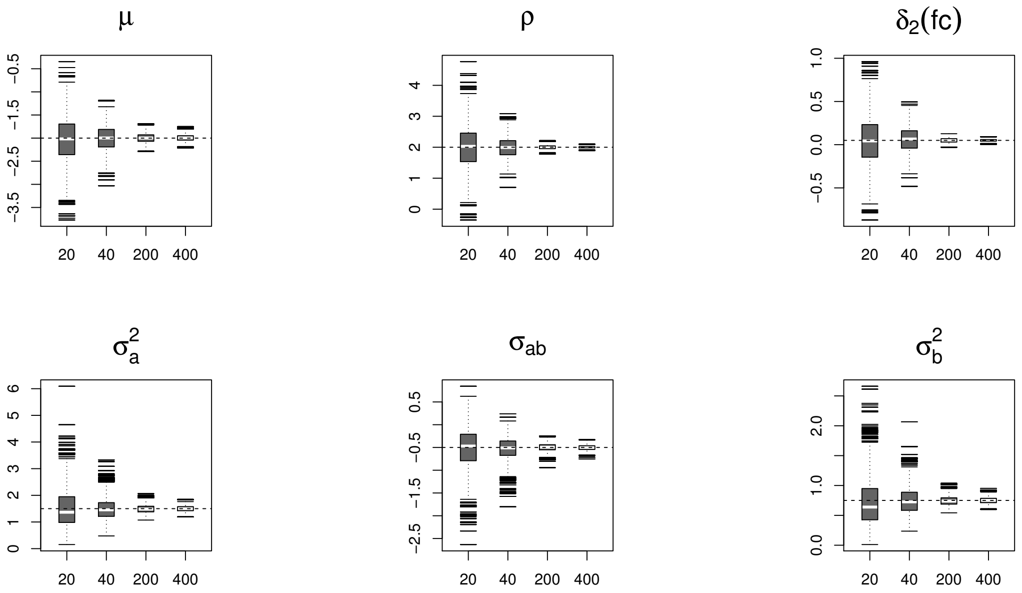


Figure 5: Summary of simulation results for Model 3. Boxplots of the simulated distribution of $\hat{\theta}^*$ for $g = 20, 40, 200, 400$, for selected parameters; true parameter values are given by horizontal lines, so that the y-scales are parameter-specific. Based on 1,000 simulations.

Table 1: Results for `high-tech managers` data.

		$L^*(\theta)$	$L^\dagger(\theta)$, M =5,000	$L^\dagger(\theta)$, M =10,000	$L^\dagger(\theta)$, M =20,000
Effect	Covariate	Est. (s.e)	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)
Sender	`Age`	-0.13 (0.06)	-0.13 (0.06)	-0.13 (0.06)	-0.13 (0.06)
	`Tenure`	0.14 (0.06)	0.14 (0.06)	0.14 (0.06)	0.14 (0.06)
Receiver	`Age`	-0.002 (0.04)	-0.002 (0.04)	-0.002 (0.04)	-0.002 (0.04)
	`Tenure`	0.045 (0.041)	0.046 (0.042)	0.046 (0.042)	0.046 (0.042)
Density	μ	0.039 (1.74)	0.005 (1.76)	0.001 (1.76)	0.001 (1.76)
	`Department`[a]	1.58 (0.35)	1.59 (0.35)	1.59 (0.35)	1.59 (0.35)
	`Level`[a]	1.15 (0.41)	1.17 (0.41)	1.17 (0.41)	1.17 (0.41)
	`Age`[b]	-0.055 (0.025)	-0.055 (0.025)	-0.055 (0.025)	-0.055 (0.025)
Reciprocity	ρ	2.12 (0.63)	2.08 (0.63)	2.08 (0.63)	2.08 (0.63)
Sender Var.	σ_A^2	2.05 (0.95)	2.08 (0.97)	2.11 (0.97)	2.11 (0.98)
Receiver Var.	σ_B^2	1.00 (0.56)	1.01 (0.57)	1.01 (0.57)	1.01 (0.57)
Covariance	σ_{AB}	-1.14 (0.64)	-1.12 (0.64)	-1.12 (0.64)	-1.12 (0.64)

[a] dichotomized difference of sending and receiving actor covariate values

[b] absolute difference of sending and receiving actor covariate values

Table 2: Results for `Lazega friendship` network.

		MQL	$L^*(\theta)$	$L^\dagger(\theta)$, M =10,000	Bayesian
Effect	Covariate	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)	Est. (s.e.)
Model 0					
Density	μ	-2.79 (0.23)	-3.43 (0.30)	-3.43 (0.31)	-3.40 (0.28)
Reciprocity	ρ	3.55 (0.30)	4.04 (0.46)	4.00 (0.45)	3.91 (0.45)
Sender Var.	σ_A^2	1.29 (0.28)	1.11 (0.41)	1.13 (0.42)	1.14 (0.44)
Receiver Var.	σ_B^2	0.85 (0.21)	0.70 (0.30)	0.70 (0.31)	0.70 (0.32)
Covariance	σ_{AB}	-0.52 (0.19)	-0.15 (0.27)	-0.12 (0.27)	-0.056 (0.28)
Model 1					
Density	μ	-1.02 (0.32)	-1.09 (0.41)	-1.07 (0.41)	-1.05 (0.41)
	`Office`	-2.36 (0.46)	-2.76 (0.50)	-2.77 (0.50)	-2.85 (0.51)
	`Seniority`	-0.44 (0.07)	-0.62 (0.09)	-0.62 (0.09)	-0.62 (0.09)
	`Gender`	-0.49 (0.15)	-0.70 (0.19)	-0.70 (0.19)	-0.69 (0.19)
	`Specialty`	-0.40 (0.15)	-0.59 (0.19)	-0.59 (0.19)	-0.59 (0.18)
Reciprocity	ρ	2.68 (0.32)	3.00 (0.48)	2.96 (0.48)	2.85 (0.45)
	`Office`	2.15 (0.96)	1.67 (1.01)	1.69 (1.01)	1.76 (1.00)
Sender Var.	σ_A^2	1.40 (0.31)	1.47 (0.54)	1.47 (0.54)	1.47 (0.55)
Receiver Var.	σ_B^2	0.76 (0.20)	0.68 (0.32)	0.67 (0.32)	0.69 (0.32)
Covariance	σ_{AB}	-0.30 (0.18)	0.003 (0.31)	0.03 (0.31)	0.13 (0.31)
Model 2					
Density	μ	-2.08 (0.36)	-2.36 (0.44)	-2.34 (0.44)	-2.37 (0.45)
	`Location`	-1.42 (0.28)	-1.84 (0.37)	-1.85 (0.37)	-1.94 (0.38)
	`Seniority`	-0.29 (0.08)	-0.56 (0.10)	-0.56 (0.10)	-0.57 (0.11)
	`Gender`	-0.61 (0.17)	-0.72 (0.21)	-0.72 (0.21)	-0.73 (0.22)
	`Advise`	1.54 (0.25)	2.17 (0.32)	2.17 (0.32)	2.23 (0.33)
	`Cowork`	0.37 (0.27)	0.75 (0.33)	0.75 (0.33)	0.76 (0.32)
Reciprocity	ρ	3.19 (0.34)	2.91 (0.47)	2.89 (0.46)	2.78 (0.42)
Sender Var.	σ_A^2	1.75 (0.38)	1.79 (0.68)	1.80 (0.68)	1.88 (0.69)
Receiver Var.	σ_B^2	0.74 (0.21)	0.55 (0.31)	0.55 (0.31)	0.66 (0.33)
Covariance	σ_{AB}	-0.17 (0.21)	0.33 (0.34)	0.35 (0.34)	0.47 (0.35)

Table 3: `Dutch social behavior study` data: Maximized log-likelihood values, AIC and BIC for five models of interest.

Method	Full model	Model 2	Model 3	Model 4	Empty model
$L^*(\theta)$	**-253.2**	-254.3	-264.3	-254.8	-272.3
AIC	540.4	532.6	544.6	**527.5**	554.5
BIC	568.7	552.6	557.9	**542.5**	562.8
$L^\dagger(\theta)$, M =5,000	**-254.0**	-255.1	-265.0	-255.5	-273.1
AIC	542.0	534.1	546.1	**529.0**	556.3
BIC	570.2	554.1	559.4	**544.0**	564.6
$L^\dagger(\theta)$, M =20,000	**-254.0**	-255.0	-265.0	-255.5	-273.1
AIC	541.9	534.1	546.0	**529.0**	556.2
BIC	570.2	554.0	559.3	**543.9**	564.6