*Network Science* **4** (1): 48–80, 2016. 
doi:10.1017/nws.2015.29

# *Mixed-membership of experts stochastic blockmodel*

ARTHUR WHITE
*School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland*
*(e-mail:* arwhite@tcd.ie*)*

THOMAS BRENDAN MURPHY
*School of Mathematical Sciences, University College Dublin, Dublin 4, Ireland*
*(e-mail:* brendan.murphy@ucd.ie*)*

**Abstract**

Social network analysis is the study of how links between a set of actors are formed. Typically, it is believed that links are formed in a structured manner, which may be due to, for example, political or material incentives, and which often may not be directly observable. The stochastic blockmodel represents this structure using latent groups which exhibit different connective properties, so that conditional on the group membership of two actors, the probability of a link being formed between them is represented by a connectivity matrix. The mixed membership stochastic blockmodel extends this model to allow actors membership to different groups, depending on the interaction in question, providing further flexibility.

Attribute information can also play an important role in explaining network formation. Network models which do not explicitly incorporate covariate information require the analyst to compare fitted network models to additional attributes in a post-hoc manner. We introduce the mixed membership of experts stochastic blockmodel, an extension to the mixed membership stochastic blockmodel which incorporates covariate actor information into the existing model. The method is illustrated with application to the Lazega Lawyers dataset. Model and variable selection methods are also discussed.

**Keywords:** *stochastic blockmodel, mixed membership model, node attributes, community finding, model based clustering, covariate information, social selection model*

## 1 Introduction

Social network analysis (SNA) is the study of how links between a set of actors are formed. Typically, it is believed that links are formed in a structured manner, so that the Erdős–Rényi model (Erdős & Rényi, 1959), whereby links occur independently with a constant probability throughout the network, fails to capture many aspects of real-world datasets. Reasons for this structure may be due to, for example, political or material incentives, and often may not be directly observed.

Several classes of statistical methods have been proposed to examine this structure. Exponential family random graph models (ERGM) (Holland & Leinhardt, 1981; Snijders, 2002; Robins *et al.*, 2006) examine whether subgraph summary statistics occur significantly more frequently than by random chance in an unstructured network. If this is the case, then this is treated as evidence of a particular underlying mechanism in the network structure. For example, a larger number of triangles

than could reasonably be expected by chance occurring in a network is evidence of transitivity, whereby a mutually shared link to an actor increases the probability of a link between two actors.

Two other approaches represent network structure using latent variables. The stochastic blockmodel (SBM) (Holland *et al.*, 1983; Snijders & Nowicki, 1997; Daudin *et al.*, 2008) introduces $G$ latent groups underlying the network, so that conditional on the group membership of two actors, the probability of a link being formed between them is represented by a $G \times G$ connectivity matrix. The latent space network model (Hoff *et al.*, 2002) maps actors onto a $d$-dimensional space so that the probability of a link being formed between two actors becomes a function of their distance from each other. The latent position cluster model (Handcock *et al.*, 2007) then extends this model so that the positions of actors in this space are determined by a mixture of spherical multivariate normal distributions. Both the SBM and latent position cluster models can be thought of as types of mixture model applied to network data.

A key difference between the models is that the latent space model is constrained to cluster together actors with strong connections with each other but weak connections to other actors in the network, a behavior known as *affiliation*. Conversely, the SBM has no such constraints and can represent this behavior, as well as *disassociative mixing*, whereby disparate actors connect strongly to a distinct set of actors but only weakly with each other (Latouche *et al.*, 2011). Airoldi *et al.* (2008) and Latouche *et al.* (2011) develop extensions to the SBM, introducing mixed-membership of stochastic blockmodels (MMSBM) and overlapping SBMs respectively. These models allow actors membership of different groups, depending on the actors with which they are interacting, further extending the flexibility of the SBM.

Attribute information can also play an important role in helping to explain how a particular network structure has occurred. For example, high-school students might be more likely to form friendships with others in the same class as them, while gender plays an important role in the formation of sexual networks. This belief, referred to as "homophilly by attributes" is reflected by Breiger (1974), who notes the "metaphor which has often appeared in sociological literature," that "groups ... are collectivities based on the shared interests, personal affinities, or ascribed status of members who participate regularly in collective activities."

Network models which do not incorporate covariate information require the analyst to compare fitted network model clusterings to additional attributes in a post-hoc manner (Handcock *et al.*, 2007; Airoldi *et al.*, 2008). Mariadassou *et al.* (2010) and Gormley & Murphy (2010) respectively extend the SBM and latent space models to incorporate covariates, at link and actor-specific levels. Examples of actor-specific covariates include gender and age, while link-specific covariates relate additional information about the relationship between actors, such as the physical distance between actor locations. Mariadassou *et al.* (2010) also introduce specifications for SBMs fitting for types of interaction beyond binary link types. These models can explicitly investigate the impact which concomitant covariate information has on network structure.

In this paper, we present a method which incorporates actor attribute information into the MMSBM, the mixed-membership of experts stochastic blockmodel (MMESBM). This method makes use of the mixture of experts terminology framework introduced by Jacobs *et al.* (1991) to allow model parameters to depend

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

on covariate information; we adapt the terminology since the covariates are incorporating into a mixed-membership rather than mixture model framework. This model may be thought of as a type of social selection model (e.g., Fellows & Handcock, 2012), in that it is assumed that actor characteristics influence network formation, while the attribute information itself is assumed fixed and known. Models where the converse applies, so that social ties are seen as influencing actor characteristics, are referred to as social influence models. Other approaches include that of Zanghi *et al.* (2010a), who introduce a model whereby the latent structure of the model explains both network and actor attribute information, and Zhang *et al.* (2013), who develop a community detection method which reweights weighing edges according to feature similarities on their terminal nodes to improve performance.

The rest of the paper is structured as follows: both the SBM and MMSBM are briefly reviewed, before the MMESBM is introduced in Section 2. A variational Bayes method for inference similar to that proposed by Airoldi *et al.* (2008) is then described in Section 3. Model selection and validation methods are also discussed in this section. The model is applied to the Lazega Lawyers dataset in Section 5. The results are interpreted and some goodness fit diagnostics are also performed. Possible further extensions to the model are then discussed in Section 6. Some additional details on model inference are provided in the Appendix.

## 2 Model specification

Relational data consists of a set of actors $a_1, \ldots, a_N$, and the links which they share with each other. In this paper, we assume that the links are binary valued, i.e., that they are present or absent. Let the adjacency matrix $\mathbf{Y}$ represent the interaction between pairs of actors in a network. An interaction between any pair of actors $a_i$ and $a_j$ can then be represented as

$$Y_{ij} = \begin{cases} 1 & \text{if a link exists between actors } a_i \text{ and } a_j; \\ 0 & \text{otherwise.} \end{cases}$$

If the link type is thought of as being shared, or symmetric, then the network is said to be undirected, with $Y_{ij} = Y_{ji}$. Otherwise, it is said to be directed. In some settings, such as protein–protein interactions, self-interaction is possible, i.e., $Y_{ii}$ can take values. This property is referred to as reflexivity. In other cases, such as when friendship between high-school students is being considered, such an interaction is not considered meaningful, making the network irreflexive, and as such the diagonal entries of $\mathbf{Y}$ are considered undefined. For the purposes of this paper, we consider only the case when a network is directed and irreflexive.

### *2.1 Stochastic blockmodel*

The SBM assumes that $G$ latent groups underlie the data. Conditional on their memberships to groups $g$ and $h$ respectively, the interaction between two actors $a_i$ and $a_j$ is then modeled by a $G \times G$ interaction matrix $\boldsymbol{\Theta}$, such that $\mathbb{P}(Y_{ij} = 1) = \Theta_{gh}$. Let $\boldsymbol{\tau}$ denote the mixing proportions of the groups, so that $\mathbb{P}(\text{Group } g) = \tau_g$. Each actor $a_i$ is assigned a group membership indicator $\mathbf{Z}_i$, such that

$$Z_{ig} = \begin{cases} 1 & \text{if actor } a_i \text{ belongs to Group } g; \\ 0 & \text{otherwise.} \end{cases}$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Each $\mathbf{Z}_i$ then follows a multinomial distribution, with one trial and probability vector $\boldsymbol{\tau}$. The choice of conjugate priors ensures that $\boldsymbol{\Theta}$ and $\boldsymbol{\tau}$ follow beta and Dirichlet distributions respectively (Snijders & Nowicki, 1997), or inference can be performed in a frequentist framework (Daudin *et al.*, 2008). Inference for the SBM is possible using a variational approximation (Daudin *et al.*, 2008) or a collapsed Gibbs sampler (McDaid *et al.*, 2012). Gibbs sampling on the fully parameterized SBM is also possible (Nowicki & Snijders, 2001), although at substantial additional computational cost.

### 2.2 Mixed-membership stochastic blockmodel

The MMSBM (Airoldi *et al.*, 2008) extends the SBM to allow actors membership to multiple groups depending on the actor with which they interact. Within this framework, each actor $a_i$ is assigned an individual mixing parameter $\boldsymbol{\tau}_i$, denoting their propensity for group membership. Indicator vectors $\mathbf{Z}_{ij}^1$ and $\mathbf{Z}_{ij}^2$ (note the superscript indices) denote the group membership of actors $a_i$ (sender) and $a_j$ (receiver) during an interaction $Y_{ij}$. Conditional on this additional model complexity, actor interaction is again modeled by a matrix $\boldsymbol{\Theta}$ in a similar manner to the SBM.[1] Choosing a Dirichlet prior distribution with hyperparameter $\boldsymbol{\delta}$ ensures that each mixing parameter $\boldsymbol{\tau}_i$ also follows the same distribution. A beta distribution can also be specified for $\boldsymbol{\Theta}$ with the choice of a conjugate prior, otherwise it may be treated as a nuisance parameter (Airoldi *et al.*, 2008).

### 2.3 Mixed-membership of experts stochastic blockmodel

The MMSBM can be further extended by allowing the parameters of the model to be functions of concomitant covariate data. The terminology used in the mixture of experts literature refer to functions of covariates and mixing parameters as "gating networks"[2] and functions of covariates and conditionally distributive parameters as "experts" (Gormley & Murphy, 2010). In this paper, we restrict our analysis to actor-specific attributes $\mathbf{W}_i = W_{i1}, \ldots, W_{iP}$, which are incorporated into the prior distribution of the individual-level mixing parameters $\boldsymbol{\tau}$. The hyperparameter $\boldsymbol{\delta}_i$ is treated as a function of $\mathbf{W}$ and parameter $\boldsymbol{\beta}_g = \beta_{g1}, \ldots, \beta_{gP}$, $g = 1, \ldots, G$ such that $\delta_{ig}(\mathbf{W}_i) = \exp(\sum_{p=1}^{P} W_{ip}\beta_{gp})$.

As well as the actor covariates, we also include an intercept term, which quantifies aspects of the network structure not explained by the available covariate information. For convenience, we include this term as the first variable of the actor-specific attributes, so that $W_{i1} = 1, i = 1, \ldots, n$. Hence, we could alternatively write $\delta_{ig}(\mathbf{W}_i) = \exp(\beta_{g1} + \sum_{p=2}^{P} W_{ip}\beta_{gp})$. Note that inference when including link-specific attributes in a mixed-membership setting may be treated in a similar fashion to the mixture framework described in Mariadassou *et al.* (2010). The data generative process for the MMESBM is outlined in Figure 1.

[1] Airoldi *et al.* (2008) also introduce an additional sparsity parameter in order to distinguish between the case where interactions in the network are in general quite rare, and when non-interaction is due to particularly low-level connection between groups. In our own experiments, we did not find that the parameter's inclusion lead to improved performance. For this reason, and for simplicity of interpretation, we excluded this parameter from our analysis.

[2] Note that in this terminology, the network in question refers to the graphical model specification, and is not to be confused with the network data under investigation.

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

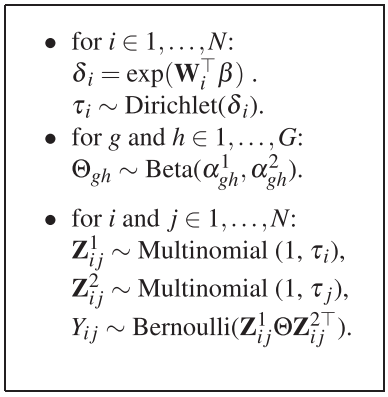



Fig. 1. Data generative process for the MMESBM.

The joint distribution of the model can be decomposed thus[3]

$$p(\mathbf{Y},\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\Theta}|\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,\boldsymbol{\beta},\mathbf{W}) = \prod_{i=1}^{N}\prod_{j=1,j\neq i}^{N} p(Y_{ij}|\mathbf{Z}^1_{ij},\mathbf{Z}^2_{ij},\boldsymbol{\Theta})p(\mathbf{Z}^1_{ij}|\boldsymbol{\tau_i})p(\mathbf{Z}^2_{ij}|\boldsymbol{\tau_j}) \times \prod_{n=1}^{N} p(\boldsymbol{\tau}_n|\boldsymbol{\beta},\mathbf{W})p(\boldsymbol{\Theta}|\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2), \quad (1)$$

where

$$p(Y_{ij}|\mathbf{Z}^1_{ij},\mathbf{Z}^2_{ij},\boldsymbol{\Theta}) = \prod_{g=1}^{G}\prod_{h=1}^{G}\left\{\Theta_{gh}{}^{Y_{ij}}(1-\Theta_{gh})^{1-Y_{ij}}\right\}^{Z^1_{ijg}Z^2_{ijh}}$$

$$p(\mathbf{Z}^1_{ij}|\boldsymbol{\tau}_i) = \prod_{g=1}^{G}\tau_{ig}^{Z^1_{ijg}}$$

$$p(\mathbf{Z}^2_{ij}|\boldsymbol{\tau}_j) = \prod_{g=1}^{G}\tau_{jg}^{Z^2_{ijg}}$$

$$p(\boldsymbol{\tau}_n|\boldsymbol{\beta},\mathbf{W}) = \frac{\Gamma\left(\sum_{h=1}^{G}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ph}\right)\right)}{\prod_{h=1}^{G}\Gamma\left(\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ph}\right)\right)}\prod_{g=1}^{G}\tau_{ng}^{\exp\left(\sum_{p=1}^{P}W_{np}\beta_{pg}\right)-1}$$

$$p(\boldsymbol{\Theta}|\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2) = \prod_{g=1}^{G}\prod_{h=1}^{G}\frac{\Gamma(\alpha^1_{gh}+\alpha^2_{gh})}{\Gamma(\alpha^1_{gh})\Gamma(\alpha^2_{gh})}\Theta_{gh}^{\alpha^1_{gh}-1}(1-\Theta_{gh})^{\alpha^2_{gh}-1}.$$

Note that we again use superscript indices for the hyperparameters $\boldsymbol{\alpha}^1$ and $\boldsymbol{\alpha}^2$. In what follows in Sections 4 and 5, we set $\alpha^1_{gh} = \alpha^2_{gh} = 1$ for $g,h = 1,\ldots,G$, in other words we use a uniform prior for $\boldsymbol{\Theta}$. Experimentally, it was found that different reasonable choices of vague prior had little impact on results.

[3] Note that since the hyperparameters $\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \boldsymbol{\beta}$, and covariate data $\mathbf{W}$ are all fixed non-random quantities, it is technically more correct to use the notation, e.g., $p(\boldsymbol{\tau}_n;\boldsymbol{\beta},\mathbf{W})$ to distinguish from the case where random quantities are being conditioned upon, e.g., $p(\mathbf{Z}^1_{ij}|\boldsymbol{\tau_i})$. For simplicity of exposition, we will use the conditional notation e.g., $p(\boldsymbol{\tau}_n|\boldsymbol{\beta},\mathbf{W})$, in both cases, although we hope it is clear to the reader which case applies.

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

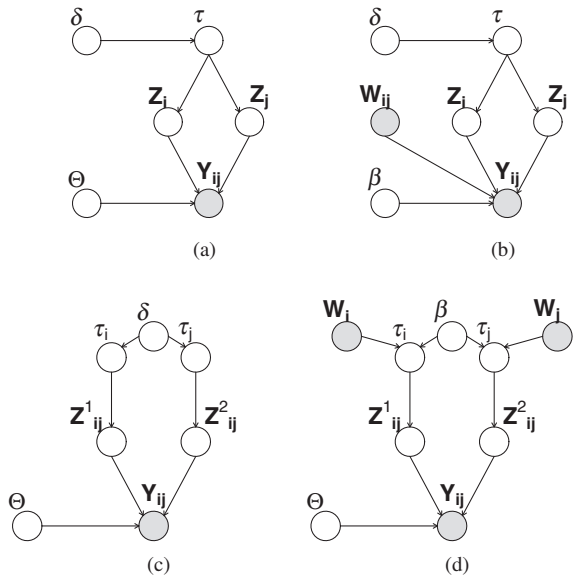



Fig. 2. Graphical model representations of the (a) SBM, (b) SBM with edge covariates introduced by Mariadassou *et al.* (2010), (c) MMSBM, and (d) MMESBM. Note that for simplicity, a prior distribution for $\boldsymbol{\theta}$ is not shown.

Graphical model representations of the SBM, the SBM with edge covariates introduced by Mariadassou *et al.* (2010), MMSBM, and MMESBM are provided in Figures 2(a)–(d).

## 3 Model inference

In a similar fashion to Airoldi *et al.* (2008), we estimate the model parameters by employing a variational Bayes approximation. For estimation of the hyperparameter $\boldsymbol{\beta}$, a fixed non-random quantity, we employ a Newton–Raphson algorithm, which, as (Blei, 2014) notes, in this setting can be thought of as an empirical Bayes method (Robbins, 1956; Efron & Morris, 1973; Efron, 2013).

Variational methods have previously proved useful in both network (Daudin *et al.*, 2008; Salter-Townshend & Murphy, 2013) and mixed-membership settings (Blei *et al.*, 2003; Rogers *et al.*, 2005; Erosheva *et al.*, 2007). See Beal (2003), Bishop (Chapter 10, 2006) and Ormerod & Wand (2010) for overviews of the method at differing levels of intensity. The main idea is to approximate the posterior $p(\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}|\mathbf{Y})$ with a set of distributions $q(\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta})$ which have a nice form. Then, the log of the marginal distribution $\log p(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$ can be re-written as

$$\begin{aligned}
\log p(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}) &= \log \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\tau}} \sum_{\mathbf{Z}^1} \sum_{\mathbf{Z}^2} p(\mathbf{Y}, \mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}) \frac{q(\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta})}{q(\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta})} d\boldsymbol{\tau}\, d\boldsymbol{\theta} \\
&\geqslant \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\tau}} \sum_{\mathbf{Z}^1} \sum_{\mathbf{Z}^2} q(\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}) \log \frac{p(\mathbf{Y}, \mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})}{q(\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta})} d\boldsymbol{\tau}\, d\boldsymbol{\theta}, \\
&= \mathbb{E}^q_{\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}} \left[ \log p(\mathbf{Y}, \mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}) \right] \\
&\quad - \mathbb{E}^q_{\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}} \left[ \log q(\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}) \right], \\
&= \mathscr{L}.
\end{aligned}$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Here, the concavity of the logarithmic function has been exploited to ensure that $\mathscr{L}$ is a lower bound to $\log p(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W})$, with the discrepancy in the inequality being equal to the Kullback–Liebler divergence (Kullback & Leibler, 1951) $\mathscr{KL}(q||p)$ between the true and approximate distributions $p$ and $q$. Note that the superscript $q$ is here used to denote that the expectation $\mathbb{E}^q$ is taken with respect to the approximate distribution $q$ rather than the true joint distribution $p$.

If we then restrict the set of distributions $q$ such that they can be factorized independently, then the optimal (i.e., the Kullback–Liebler divergence minimizing) form of each distribution will be the same as the conditional distribution of its respective parameter

$$q(\mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta}) = q(\boldsymbol{\Theta}|\boldsymbol{\zeta}^{\mathbf{1}}, \zeta^2) \prod_{i=1}^{N} q(\boldsymbol{\tau}_i|\boldsymbol{\gamma}_i) \prod_{j=1}^{N} q(\mathbf{Z}^1_{ij}|\boldsymbol{\phi}^1_{ij}) q(\mathbf{Z}^2_{ij}|\boldsymbol{\phi}^2_{ij}),$$

where $q(\mathbf{Z}^1_{ij}|\boldsymbol{\phi}^1_{ij})$ and $q(\mathbf{Z}^2_{ij}|\boldsymbol{\phi}^2_{ij})$ are multinomial distributions, $q(\boldsymbol{\tau}_i|\boldsymbol{\gamma}_i)$ is a Dirichlet distribution, $q(\boldsymbol{\Theta}|\boldsymbol{\zeta}^{\mathbf{1}}, \zeta^2)$ is a beta distribution, and we have introduced the variational parameters $\boldsymbol{\phi}^1, \boldsymbol{\phi}^2, \boldsymbol{\zeta}^{\mathbf{1}}, \zeta^2$, and $\boldsymbol{\gamma}$.

Much like for an expectation–maximization algorithm (Dempster *et al.*, 1977), these parameters can be updated in a stepwise manner which iteratively optimizes $\mathscr{L}$, and by extension $\log p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{W}, \boldsymbol{\alpha})$. Updates are as follows:

$$\begin{aligned}
\zeta^1_{gh} &= \alpha_{gh} + \sum_{i=1}^{N}\sum_{j=1}^{N} \phi^1_{ijg}\phi^2_{ijh} Y_{ij}, \\
\zeta^2_{gh} &= \beta_{gh} + \sum_{i=1}^{N}\sum_{j=1}^{N} \phi^1_{ijg}\phi^2_{ijh}(1 - Y_{ij}), \\
\gamma_{ig} &= \exp\left(\sum_{p=1}^{P} \beta_{gp} W_{ip}\right) + \sum_{j=1}^{N} (\phi^1_{ijg} + \phi^2_{jig}), \\
\phi^1_{ijg} &\propto \exp\left(\Psi(\gamma_{ig}) - \Psi\left(\sum_{k=1}^{G} \gamma_{ik}\right)\right) \\
&\quad \times \exp\left\{\sum_{h=1}^{G} \phi^2_{ijh}\left[Y_{ij}\left(\Psi(\zeta^1_{gh}) - \Psi(\zeta^1_{gh} + \zeta^2_{gh})\right)\right.\right. \\
&\quad \left.\left. + \ (1 - Y_{ij})\left(\Psi(\zeta^2_{gh}) - \Psi(\zeta^1_{gh} + \zeta^2_{gh})\right)\right]\right\}, \\
\phi^2_{ijg} &\propto \exp\left(\Psi(\gamma_{jg}) - \Psi\left(\sum_{k=1}^{G} \gamma_{jk}\right)\right) \\
&\quad \times \exp\left\{\sum_{h=1}^{G} \phi^1_{ijh}\left[Y_{ij}\left(\Psi(\zeta^1_{hg}) - \Psi(\zeta^1_{hg} + \zeta^2_{hg})\right)\right.\right. \\
&\quad \left.\left. + \ (1 - Y_{ij})\left(\Psi(\zeta^2_{hg}) - \Psi(\zeta^1_{hg} + \zeta^2_{hg})\right)\right]\right\},
\end{aligned}$$

for $i, j = 1, \ldots, N$ and $g, h = 1, \ldots, G$, and where $\Psi$ denotes the digamma function (Abramowitz & Stegun, 1965).

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

### *3.1 Estimating $\hat{\boldsymbol{\beta}}$*

It remains to estimate $\hat{\boldsymbol{\beta}}$. Inference via a closed form solution is not possible (Blei *et al.*, 2003). Instead, we make use of a Newton–Raphson algorithm to maximize $\mathscr{L}$, by updating $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - H^{-1}\nabla$ until the algorithm has deemed to converge. The gradient and Hessian take the following values:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \beta_{iq}} &= \sum_{n=1}^{N} W_{nq} \exp\left(\sum_{p=1}^{P} W_{np}\beta_{ip}\right) \\
&\quad \times \left\{ \Psi\left[\sum_{h=1}^{G} \exp\left(\sum_{p=1}^{P} W_{np}\beta_{hp}\right)\right] - \Psi\left[\exp\left(\sum_{p=1}^{P} W_{np}\beta_{ip}\right)\right] \right. \\
&\quad \left. + \ \Psi(\gamma_{ng}) - \Psi(\sum_{h=1}^{G} \gamma_{nh}) \right\}, \\
\frac{\partial^2 \mathscr{L}}{\partial \beta_{iq}\partial \beta_{jr}} &= \sum_{n=1}^{N} W_{nq} W_{nr} \exp\left(\sum_{p=1}^{P} W_{np}(\beta_{ip} + \beta_{jp})\right) \\
&\quad \times \left\{ \Psi'\left[\sum_{h=1}^{G} \exp\left(\sum_{p=1}^{P} W_{np}\beta_{hp}\right)\right] - \mathbb{I}_{i=j}\Psi'\left[\exp\left(\sum_{p=1}^{P} W_{np}\beta_{ip}\right)\right] \right\} \\
&\quad + \mathbb{I}_{i=j}\left( W_{nq} W_{nr} \exp\left(\sum_{p=1}^{P} W_{np}\beta_{ip}\right) \right. \\
&\quad \times \left\{ \Psi\left[\sum_{h=1}^{G} \exp\left(\sum_{p=1}^{P} W_{np}\beta_{hp}\right)\right] - \Psi\left[\exp\left(\sum_{p=1}^{P} W_{np}\beta_{ip}\right)\right] \right. \\
&\quad \left. \left. + \ \Psi(\gamma_{ni}) - \Psi\left(\sum_{h=1}^{G} \gamma_{nh}\right) \right\} \right).
\end{aligned}$$

Experimental results found that the estimates obtained by the Newton–Raphson algorithm can vary wildly depending on the initial parameter settings. One strategy is to initialize the parameters using a method of moments approach proposed by Minka (2012) when estimating the parameters of a Dirichlet distribution. Our goal is slightly different, in that we wish to estimate the parameters $\hat{\boldsymbol{\beta}}$ with respect to the expected log of the probabilities $\mathbb{E}[\log \boldsymbol{\tau}]$ rather than the usual observed log of the probabilities. Nevertheless, the initialization method still proves to be effective. In short, we initially assume that the covariates provide no additional information about the prior probability of group membership, before then setting $\beta_{1g}^{(1)} = \log(\mathbb{E}[\delta_g]\Sigma_{h=1}^{G}\delta_h)$, where $\Sigma_{h=1}^{G}\delta_h = (\mathbb{E}[\delta_1] - \mathbb{E}[\delta_1]^2)/(\mathbb{E}[\delta_1^2] - \mathbb{E}[\delta_1]^2)$. Intuitively, we can think of this initialization as starting from a position of skepticism; i.e., we assign weights to the covariate parameters only if it increases the lower bound. The method of moments approach serve as a reasonable initial estimate which the Newton–Raphson algorithm can then improve on.

Another difficulty which was encountered when using the estimator experimentally was that the estimated values of coefficients for covariates with only a small number

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

of observations tended to infinity. This may be related to an issue known as separability in logistic regression models (Albert & Anderson, 1984), which typically occurs for smaller datasets, whereby for certain patterns of data points maximum likelihood estimates do not exist. While methods have been suggested to remedy this problem for logistic regression models (Heinze & Schemper, 2002), as we have noted, this model is not as straightforward as other regression models, and it is not clear whether a similar approach will prove fruitful.

This issue meant that we were unable to generate data using the naive implementation outlined in Figure 1 for the simulation studies carried out in Section 4. Although the sample size in this case was relatively large, it was found that if the covariates corresponded too closely to the underlying group structure then a similar effect would occur. With regards to our application in Section 5, including interaction terms proved difficult, and we were forced to omit one covariate, office location, entirely, since only five actors in the dataset practiced in one of the three locations.

While the Newton–Raphson step described obtains the optimal parameter values $\hat{\boldsymbol{\beta}}$, it is necessary to obtain some estimate of the uncertainty of these parameter values before the impact of the covariates may be assessed. One approach is to consider the diagonal entries of the inverse Hessian $\mathbf{H}^{-1}$ specified in Section 3.1 in order to approximate the observed information matrix. The diagonal entries of this matrix should somewhat approximate the standard errors of $\hat{\boldsymbol{\beta}}$. However, this approach is limited by two facts: first, that we differentiate $\mathscr{L}$ and not the true log-posterior, and secondly, that we do not obtain the full Hessian matrix whereby $\mathscr{L}$ is twice differentiated with respect to all parameters, the dimension of which creates computational difficulties. Nevertheless, some information about the curvature of the parameters is obtained using this method.

A second approach is to exploit the generative properties of the MMESBM to estimate the behavior of $\boldsymbol{\beta}$ using a parametric bootstrapping method. Each bootstrap replication is obtained by first generating a network from the fitted model parameters using the process previously specified in Figure 1, with the estimated variational parameters $\boldsymbol{\gamma}$, $\boldsymbol{\zeta}^1$, and $\boldsymbol{\zeta}^2$ in place of the hyperparameters $\boldsymbol{\delta}$, $\boldsymbol{\alpha}^1$, and $\boldsymbol{\alpha}^2$. The model is then refitted to the simulated data and the results recorded. While this approach may be more reliable than the first outlined, it is worth noting that the typical under-estimation of parameter uncertainty in variational Bayes methods will be reflected in the bootstrapped values for $\boldsymbol{\beta}$. It therefore follows that while this method allows us to dismiss unimportant covariates with a high degree of certainty, care must be taken when interpreting and selecting which attributes appear to have a meaningful impact on the network.

### *3.2 Algorithm initialization*

Experimentally, it was found that the clustering solutions obtained by the MMESBM was highly dependent on the starting values used to initialize the algorithm. We employed the followed initalization approach, which we found to give sensible results in most cases. First, an SBM was fit to the network using the mixer package (Daudin *et al.*, 2008; Zanghi *et al.*, 2010b; Latouche *et al.*, 2012) in R (R Core Team, 2015).

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Then $\hat{\mathbf{Z}}^*$, the expected group membership for each actor in the SBM framework, was used as the initial estimate for $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$, so that $\boldsymbol{\phi}^1_{ij} = \boldsymbol{\phi}^2_{ij} = \hat{\mathbf{Z}}^*_i$, for $i = 1, \ldots, N$.

The MMESBM algorithm was then run, but with the Newton–Raphson step omitted for the first 500 iterations, or until the algorithm increases fails to increase $\mathscr{L}$ by some small margin. In other words, the parameters in the MMSBM model are estimated first. Finally, the full MMESBM algorithm was run to convergence, or for a maximum of 1,000 iterations.

### *3.3 Model selection*

While model assumptions require the number of profiles $G$ to be fixed and known, in reality this is not the case. We therefore run the model over a range of values of $G' = 1, \ldots, G^{\max}$, and compare the models post-hoc. The variational approximation to Equation (1) provides only a lower bound to the integrated likelihood, making the use of criteria such as the Bayesian information criterion (Kass & Raftery, 1995) difficult to obtain. Other difficulties, such as determining the effective sample size of the data, also occur in this setting (Hunter *et al.*, 2008).

Approaches for model selection for the SBM include the integrated complete-data likelihood (ICL) (Daudin *et al.*, 2008) and the integrated likelihood variational Bayes (ILvb) (Latouche *et al.*, 2012), with the ILvb in particular being a suitable method for small networks since it does not depend on an asymptotic approximation (Latouche *et al.*, 2012). However, it is not clear how these criteria perform in a mixed-membership setting.

Alternatively, cross-validation methods can prove useful when performing model selection in a model-based setting (Smyth, 2000; Hoff, 2008; Airoldi *et al.*, 2008). Note that in this setting, individual links, rather than all links associated with individual actors are removed. In this instance, the method takes the following steps:

1. Divide the network edges $\mathbf{Y}$ into $k$ folds of roughly equal size. Let $Y_{(-k)}$ be the data with the $k$th fold removed.
2. Drop a single fold and fit the MMESBM to the remaining data—it is straightforward to estimate the posterior parameters $\boldsymbol{\Theta}, \boldsymbol{\tau}$, and $\boldsymbol{\delta}$; the values of $\mathbf{Z}^1$ and $\mathbf{Z}^2$ for missing edges can simply be ignored during the estimation procedure.
3. Compare the fitted model parameters against the out-of-sample data. Conditional on the fitted posterior estimates, the link probabilities for an out of sample data point $Y_{ij}$ take the form

$$p(Y_{ij}|\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\tau}}_i, \hat{\boldsymbol{\tau}}_j) = \sum_{h=1}^{G} \sum_{g=1}^{G} \hat{\tau}_{ig} \hat{\tau}_{jh} \hat{\Theta}_{gh}^{Y_{ij}} (1 - \hat{\Theta}_{gh})^{1-Y_{ij}}.$$

   Here, $\hat{\Theta}_{gh}, \hat{\boldsymbol{\tau}}_i,$ and $\hat{\boldsymbol{\tau}}_j$ denote the posterior mean of $\Theta_{gh}, \boldsymbol{\tau}_i,$ and $\boldsymbol{\tau}_j$, conditional on the training data $Y_{(-k)}$, and $p(Y_{ij}|\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\tau}}_i, \hat{\boldsymbol{\tau}}_j)$ has been integrated over $\mathbf{Z}^1_{ij}$ and $\mathbf{Z}^2_{ij}$.
4. Repeat for each fold in turn. Once this has been completed, the model with highest average hold out log-likelihood, taking into account the uncertainty in the estimation, is deemed to be most suitable. In this way, we can also assess

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Table 1. *Actor covariates for Simulation Study 1.*

| $i = 1,\ldots,40$ | | $i = 41,\ldots,100$ | |
|---|---|---|---|
| $W_{i1} = 1$ | | $W_{i1} = 1$ | |
| $W_{i2} \sim$ Binomial$(1, 0.8)$ | | $W_{i2} \sim$ Binomial$(1, 0.2)$ | |
| $W_{i3} \sim$ Binomial$(1, 0.8)$ | | $W_{i3} \sim$ Binomial$(1, 0.2)$ | |
| $W_{i4} \sim \begin{cases} \text{Normal}(-2,1) \\ \text{Normal}(2,1) \end{cases}$ | $\begin{matrix} \text{with probability } 0.9 \\ \text{with probability } 0.1 \end{matrix}$ | $W_{i4} \sim \begin{cases} \text{Normal}(-2,1) \\ \text{Normal}(2,1) \end{cases}$ | $\begin{matrix} \text{with probability } 0.1 \\ \text{with probability } 0.9 \end{matrix}$ |

goodness of fit for the model, by e.g., checking the total predicted data against the total observed data; this is described in further detail in Section 5.

## 4 Simulation study

The approach is first demonstrated on simulated data. Two simulation studies are performed. In both cases, the total number of actors is set to be $N = 100$, and the number of underlying groups is set to be $G = 3$. First, three covariates are used to inform the model. These covariates distinguish Group 1 from Groups 2 and 3, but do not distinguish between Groups 2 and 3. In the second study, an additional three non-informative variables are also included. The second network (9.6% density) is sparser than the first (26%).

### *4.1 Simulation study 1*

In the first simulation study, community-like behavior was specified for all three groups. We divide the actors into subsets, such that the individual mixing parameters are likely to strongly favor one group over the others. This is done by setting the hyperparameter $\boldsymbol{\delta}$ in the following manner:

$$\boldsymbol{\delta}_i = \begin{cases} (0.8, 0.2, 0.2) & 1 \leqslant i \leqslant 40 \\ (0.2, 0.8, 0.2) & 41 \leqslant i \leqslant 80 \\ (0.2, 0.2, 0.8) & 81 \leqslant i \leqslant 100. \end{cases}$$

Together with an intercept term, covariates $\mathbf{W}$ were then generated to distinguish the first subset of actors ($1 \leqslant i \leqslant 40$) from the rest of the dataset. These consisted of two binary and one continuous variables. Actors in the first subset were more likely to be assigned positive values for $p = 2, 3$ (i.e., $W_{ip} = 1$) and negative values (i.e., $W_{ip} < 0$) for $p = 4$. See Table 1 for full details of how the covariates were generated.

Conditional on group structure, the probabilities of within group links are high, and between group links are low:

$$\boldsymbol{\Theta} = \begin{pmatrix} 0.60 & 0.05 & 0.01 \\ 0.05 & 0.70 & 0.10 \\ 0.01 & 0.01 & 0.80 \end{pmatrix}.$$

The data was then generated in a similar manner to the process outlined in Figure 1, except that probabilities for $\boldsymbol{\Theta}$ were kept fixed, rather than generated from a prior distribution, and the hyperparameter $\boldsymbol{\delta}_i$ was used directly in place of $\exp(\mathbf{W}_i^\top \boldsymbol{\beta})$. This was done for the reasons discussed in Section 3.1.

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

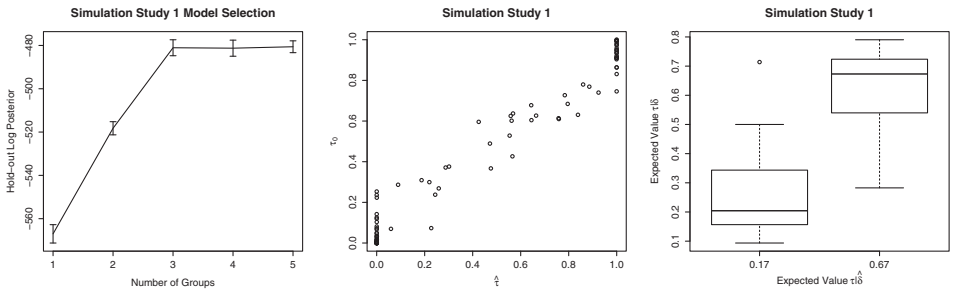



Fig. 3. Figure 3(a) shows the hold-out log-likelihood for the first simulation study. Figure 3(b) is a scatterplot comparing the estimated value of $\hat{\tau}_{11},\ldots,\hat{\tau}_{1N}$ against the true value $\tau_{11},\ldots,\tau_{1N}$. Figure 3(c) is a boxplot comparing the *a priori* expected value of individual mixing parameters conditional on the estimated hyperparameter $\hat{\boldsymbol{\delta}}, \mathbb{E}[\tau_{11}|\boldsymbol{\delta_1}],\ldots,\mathbb{E}[\tau_{1N}|\boldsymbol{\delta_N}]$, against the expected value using the true value $\boldsymbol{\delta}$.

MMESBM models were fitted to the data over a range of values, from $G = 1,\ldots,5$. The 10-fold cross-validated log-likelihood is shown in Figure 3(a). While the 1 and 2 group model are clearly inferior, the 4 and 5 group models do not appear to substantively improve the fit achieved by the 3 group model.

The 3 group model captures the underlying mixed membership structure of the network well. Figure 3(b) compares the estimated value of the individual membership of each actor to Group 1, $\hat{\tau}_{11},\ldots,\hat{\tau}_{1N}$, to the true value $\tau_{11},\ldots,\tau_{1N}$. The close correspondence between the values is reflected in the extremely high correlation ($r = 98\%$). There is similar agreement between the estimated and true values of individual membership to Groups 2 and 3 (again, $r = 98\%$ in both cases).

The estimated values $\hat{\boldsymbol{\delta}}$, where $\hat{\boldsymbol{\delta}}_i = \exp(\mathbf{W}_i^\top \hat{\boldsymbol{\beta}})$, did not correspond directly to the true values of $\delta$. However, it was found that the *a priori* expected value of individual mixing parameters conditional on the estimated hyperparameter $\hat{\boldsymbol{\delta}}, \mathbb{E}[\tau_{ig}|\boldsymbol{\delta_i}]$, corresponded somewhat to the expected value using the true value $\boldsymbol{\delta}$. This was particularly true for Group 1 ($r = 81\%$), although the association was poorer for Groups 2 and 3 ($r = 55\%$ and 25% respectively). This makes sense, since the covariates provide us with more information about Group 1 than Groups 2 and 3. Figure 3(c) compares the *a priori* expected value of individual mixing parameters for Group 1 using $\hat{\boldsymbol{\delta}}$ against the expected value using the true value, $\boldsymbol{\delta}$.

### *4.2 Simulation study 2*

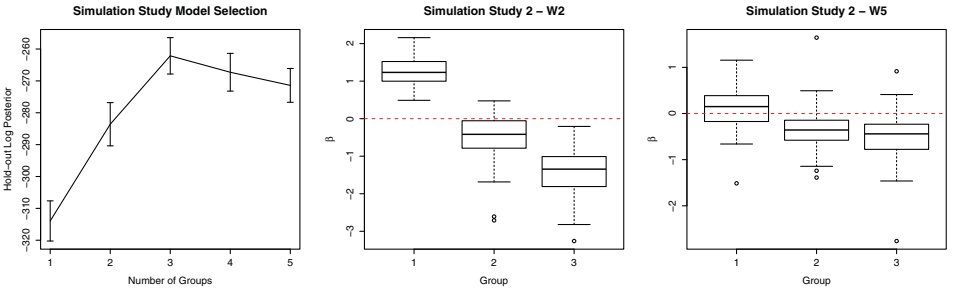
In the second simulation study, community-like behavior was specified for Groups 1 and 3, with Group 2 exhibiting disassociative mixing. Again, we divide the actors into subsets, such that the individual mixing parameters are likely to strongly favor one group over the others; however, in this case the subsets all have unequal size and the third subset is much smaller than the others. We set the hyperparameter $\boldsymbol{\delta}$ in the following manner:

$$\boldsymbol{\delta}_i = \begin{cases} (0.8, 0.2, 0.2) & \text{if } 1 \leqslant i \leqslant 50 \\ (0.2, 0.8, 0.2) & \text{if } 51 \leqslant i \leqslant 90 \\ (0.2, 0.2, 0.8) & \text{if } 91 \leqslant i \leqslant 100. \end{cases}$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Table 2. *Actor covariates for simulation study 2.*

| $i = 1,\ldots,50$ | $i = 51,\ldots,100$ |
|---|---|
| $W_{i1} = 1$ | $W_{i1} = 1$ |
| $W_{i2} \sim$ Binomial$(1, 0.8)$ | $W_{i2} \sim$ Binomial$(1, 0.2)$ |
| $W_{i3} \sim$ Binomial$(1, 0.8)$ | $W_{i3} \sim$ Binomial$(1, 0.2)$ |
| $W_{i4} \sim \begin{cases} \text{Normal}(-2,1) & \text{with probability } 0.9 \\ \text{Normal}(2,1) & \text{with probability } 0.1 \end{cases}$ | $W_{i4} \sim \begin{cases} \text{Normal}(-2,1) & \text{with probability } 0.1 \\ \text{Normal}(2,1) & \text{with probability } 0.9 \end{cases}$ |
| $i = 1,\ldots,100$ | |
| $W_{i5} \sim$ Binomial$(1, 0.5)$ | |
| $W_{i6} \sim$ Binomial$(1, 0.5)$ | |
| $W_{i7} \sim$ Normal$(0, 1)$ | |




(a) (b) (c)

Fig. 4. Figure 4(a) shows the hold-out log-likelihood for the second simulation study. Figures 4(b) and 4(c) show boxplots of parametric bootstrap samples of covariate parameter estimates $\hat{\boldsymbol{\beta}}$ for the 3 group MMESBM, for covariates $\mathbf{W}_2$ (informative) and $\mathbf{W}_5$ (noise). The dashed line occurs at zero. (color online)

Conditional on group structure, the probabilities of a link occurring is set to be

$$\boldsymbol{\Theta} = \begin{pmatrix} 0.30 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.10 \\ 0.01 & 0.20 & 0.60 \end{pmatrix}.$$

Data was then generated in the same way as for Simulation Study 1.

As for Simulation Study 1, three covariate variables $\mathbf{W}_2,\ldots,\mathbf{W}_4$ were generated to distinguish the first subset of actors ($1 \leqslant i \leqslant 50$) from the rest. These were generated in the same manner as for the previous study. Additionally, three non-informative covariates, $\mathbf{W}_5,\ldots,\mathbf{W}_7$, again consisting of two binary and one continuous variable, were included. See Table 2 for full details of how the covariates were generated.

Several aspects of this simulation study are similar to the first. The 10-fold cross-validated log-likelihood for MMESBM models fitted to the data for $G = 1,\ldots,5$ are shown in Figure 4(a). Again, there appears to be little evidence for the 1 and 2 group models, while the performance appears to be broadly comparable across the 3, 4, and 5 group models, although in this case the 3 group model's performance appears best.

The 3 group model again captures the underlying mixed membership structure of the network well. In this case, the correspondence between the estimated value

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Table 3. *Estimates of the covariate parameters* $\boldsymbol{\beta}$. *95% bootstrap quantile ranges (2.5% and 97.5%) are included in parentheses. Estimates whose quantile range does not include zero are highlighted in bold.*

| | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| $\mathbf{W}_1$ | **−3.53** (−4.25, −2.81) | **−3.06** (−3.82, −2.28) | **−1.31** (−2.14, −0.35) |
| $\mathbf{W}_2$ | **1.25** (0.53, 2.09) | −0.45 (−1.64, 0.42) | **−1.42** (−2.56, −0.47) |
| $\mathbf{W}_3$ | **1.32** (0.40, 3.81) | 0.02 (−1.10, 1.59) | −0.22 (−1.57, 2.29) |
| $\mathbf{W}_4$ | **−0.34** (−0.58, −0.14) | 0.12 (−0.07, 0.33) | −0.08 (−0.35, 0.17) |
| $\mathbf{W}_5$ | 0.12 (−0.53, 0.83) | −0.38 (−1.12, 0.41) | −0.50 (−1.39, 0.33) |
| $\mathbf{W}_6$ | 0.20 (−0.46, 0.98) | **−0.75** (−1.60, −0.02) | −0.60 (−1.47, 0.34) |
| $\mathbf{W}_7$ | −0.04 ( −0.37, 0.33) | −0.07 (−0.47, 0.23) | 0.41 (−0.06, 0.92) |

of the individual membership of each actor compared to the true value resulted in correlation scores of $r = 96\%$ for each group. Similarly, the *a priori* expected value of individual mixing parameters conditional on the estimated hyperparameter $\hat{\boldsymbol{\delta}}$, corresponded closely to the expected value using the true value $\boldsymbol{\delta}$ for Group 1 (r = 90%), but less well for Groups 2 and 3 (r = 76% and 30% respectively).

Estimates of $\hat{\boldsymbol{\beta}}$ were obtained from 100 parametric bootstrap replications of the fitted model. Parameter estimates with bootstrap quantiles at the 95% level are provided in Table 3. For each of the informative covariates $\mathbf{W}_2, \ldots, \mathbf{W}_4$ at least one of the parameter quantile ranges does not contain 0. This result is similar to that found for Simulation Study 1. Of the non-informative covariates, $\mathbf{W}_5, \ldots, \mathbf{W}_7$, only one of the terms appears to be significant. On closer inspection, the upper bound for this quartile range, −0.02, is particularly close to 0. Boxplots of the parametric bootstrap samples of $\hat{\boldsymbol{\beta}}$ for $\mathbf{W}_2$ and $\mathbf{W}_5$ are provided in Figures 4(b) and (c).

## 5 Lazega lawyers application

We apply our method to the Lazega Lawyers dataset[4], obtained from a network study of corporate law partnership carried out in a Northeastern U.S. law firm. Several features make the data of interest, the most notable being that the lack of a strong formal working structure, coupled with large incentives to behave opportunistically create an interesting environment for the formation of network structure. Three types of network link are available from the study: strong co-worker, basic advice, and friendship networks. In this paper, we focus on the friendship network, which consists of 71 actors and 575 links. Actor attribute information is also available, and described in Table 4.

Care must be taken when incorporating qualitative variables into the MMESBM framework so that each modality is associated to a specific parameter. In particular, it was necessary to re-code the nominal variable Law School, which in its original format can take one of three values, into a set of two binary variables. Each of these variables then serves as a comparison between one of nominal categories (Law School = University of Connecticut, or Law School = Other) to the baseline category (Law School = Harvard or Yale).

[4] The dataset is available to download at https://www.stats.ox.ac.uk/~snijders/siena/Lazega_lawyers_data.htm.

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

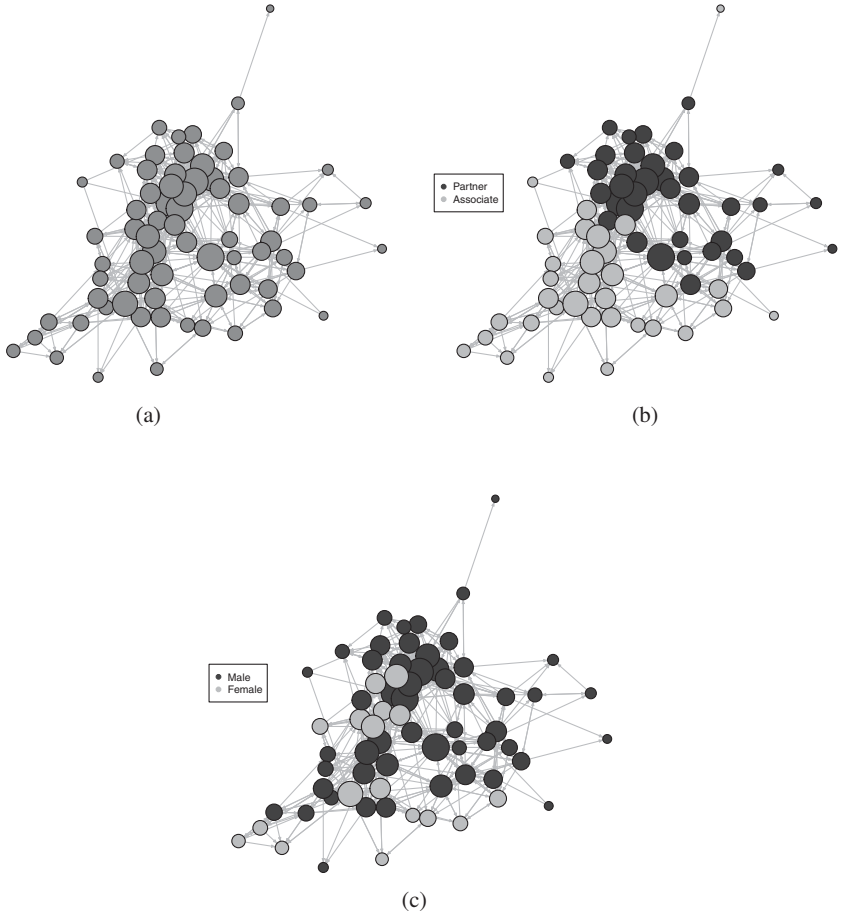



Fig. 5. Lazega Lawyers friendship network represented using a Fruchtermann–Reingold algorithm. Node size is used to give an indication of the number of links sent and received by each actor. Figures (b) and (c) color nodes with respect to gender and status respectively. (a) Lazega lawyers. (b) Status. (c) Gender.

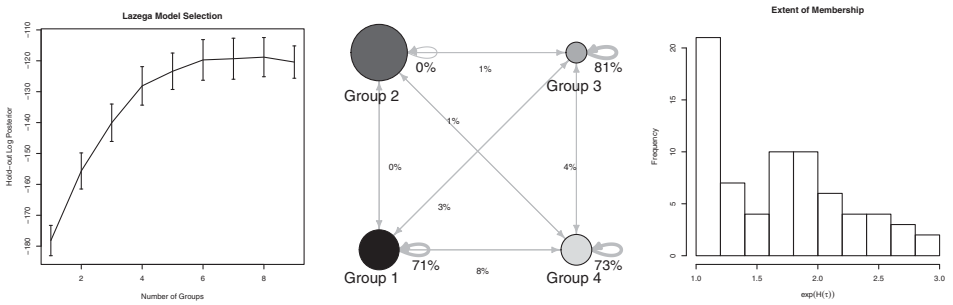
The actor attributes have previously been incorporated into studies conducted by Gormley & Murphy (2010) and Snijders *et al.* (2006); the former found that office location and years with the firm had a significant impact on a latent position cluster model when included as covariates for group membership, while Snijders *et al.* (2006) found evidence that seniority, practice, and location affected the 36-actor network of partners beyond other structural effects in the data using an ERGM-based approach.

The network is visualized in Figure 5(a) using a Fruchterman–Reingold algorithm. Two actors, who are not connected to any others in the network, are not plotted. (They are still included in the analysis.) The size of each node in the graph is representative of the overall number of links each actor has formed in the network. Note that several of the covariates are correlated. This is partly visualized in Figure 5(b) and (c). These figures compare gender and status; from these figures it is clear that there are more men (53) than women (18) in the firm, and that women are more likely to be associates than partners (there are only three female partners in the dataset). Seniority is also highly (negatively) correlated with both age and years with

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Table 4. *Actor attribute information for Lazega Lawyers.*

| Attribute | Description (where necessary) |
|---|---|
| Seniority | Rank in firm, where 1 is the highest, 71 the lowest. |
| Status | Indicates partner or associate in firm, with 0 = partner and 1 =associate. |
| Gender | 0 = man; 1 = woman. |
| Years with firm | |
| Age | In years. |
| Practice | 0 = litigation; 1 = corporate. |
| Law school | 0 = Yale or Harvard; 1= University of Connecticut; or 2 = Other. |
| Office | Excluded from analysis. |
| Recoded variables | |
| Law school—Connecticut | 0 = (Yale or Harvard) or Other; 1 = University of Connecticut; |
| Law school—Other | 0 = (Yale or Harvard) or University of Connecticut; 1 = Other. |




(a) (b) (c)

Fig. 6. Figure 6(a) shows the hold-out log-likelihood for the Lazega Lawyers Friendship data. Figure 6(b) is a visualization of the blocks behavior. Figure 6(c) is a histogram of EoM for actors in the network. Over half the actors display mixed membership between two groups.

the firm. In what follows, the continuous attributes have been standardized to have mean= 0 and standard deviation= 1, to facilitate interpretation of the covariate parameters $\hat{\boldsymbol{\beta}}$.

### *5.1 Fitting the model*

MMESBM models were fitted to the Lazega Lawyers data over a range of values, from $G = 1,\ldots,9$. While the 10-fold cross-validated log-likelihood shown in Figure 6(a) is maximized for an 8 group model, the error bars for models with four or more groups models all overlap, suggesting a somewhat limited improvement in performance from the inclusion of additional groups. After considering the competing models, the 4 group model was deemed a satisfactory fit to the data.

The 4 group MMESBM is represented in Figure 6(b). Each node in the diagram represents a group and is labeled accordingly. Node size reflects the overall (weighted) membership of the group, while arrow sizes loosely correspond to interaction

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Table 5. *Estimates for blockmodel interaction* $\hat{\boldsymbol{\Theta}}$.

| | Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|---|
| Group 1 | 0.71 | 0.00 | 0.01 | 0.01 |
| Group 2 | 0.00 | 0.00 | 0.01 | 0.01 |
| Group 3 | 0.03 | 0.01 | 0.81 | 0.04 |
| Group 4 | 0.08 | 0.01 | 0.01 | 0.73 |

levels, with larger arrows indicating higher levels of interaction. Selected interaction probabilities are also included in the figure. The within group interaction terms are printed in large font size beside the relevant group, while the larger of the two interaction probabilities between each pair of groups is included in smaller font size. Each between group probability is printed roughly half way between the relevant groups.

Inspecting the figure, Groups 1, 3, and 4 can be characterized as exhibiting community-like behavior. In each case, the probability of within group interaction occurring is far higher than would be expected in the network under the null Erdős–Rényi model, whereby links between actors occur independently with probability 11.5% in this case. Group 2 is a highly antisocial group, with no interaction probabilities exceeding 1%. The fitted values for $\hat{\boldsymbol{\Theta}}$ are provided in Table 5. Between group interaction occurs with low probability (less than 10%) in all cases.

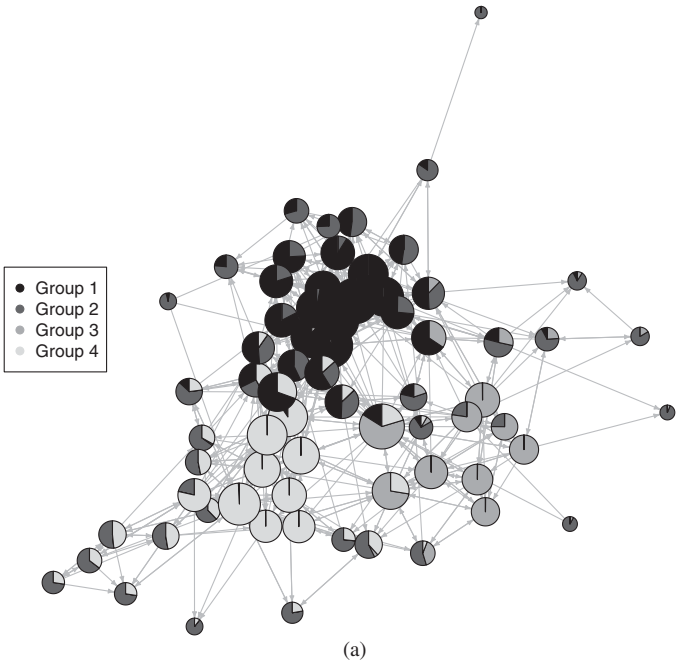
One way to check an actor's propensity for mixed membership is to inspect their extent of profile membership (EoM) score (Hill, 1973; White *et al*., 2012):

$$\mathrm{EoM}_i = \exp(H(\hat{\boldsymbol{\tau}}_i)),$$

where $H$ denotes the entropy function, $H(\hat{\boldsymbol{\tau}}_i) = -\sum_{g=1}^{G} \hat{\tau}_{ig} \log \hat{\tau}_{ig}$. A histogram of each actor's EoM score is shown in Figure 6(c). Over half (42) of the actors EoM scores are over 1.5, suggestive of at least some amount of mixed membership. Of the 20 actors with the lowest EoM score, six belong to Group 1, five to Group 3, and seven to Group 4. These actors can be viewed as being most highly involved in their respective groups and exhibit almost no mixed membership. The three actors who belong most strongly to Group 2 possess a single (received) link in the network between them. Thirteen of the fourteen actors with the highest EoM scores exhibit activity across three groups; one actor has a small amount of membership to all four groups. With one exception these actors all have some membership of Group 2, indicating that they are not full participants in the other groups that they have membership of; only one actor appears to be highly social with Groups 1, 3, and 4.

Figure 7(a) visualizes the model using the same network layout as Figure 5(a)–(c), with each of the plotted nodes assigned a pie chart representing their mixed membership to different groups. The colors in each pie chart are consistent with those in Figure 6(b). Inspecting this plot, it is clear that a large amount of mixed membership is exhibited by actors in this model, corroborating the EoM statistics reported in Figure 6(c). Recall that the size of each node in the graph is related to the popularity of the actor in question. The smaller nodes in the graph contain prominent dark gray sections, representing Group 2, while the largest nodes display

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press




(a)

Fig. 7. Visualization of the 4 group MMESBM fitted to the Lazega Lawyers friendship dataset.

membership to the community-like Groups 1, 3, and 4, colored darkest gray and light and lightest gray respectively. In particular, two types of mixed membership are occurring: actors moving between the Groups 1, 3, and 4, and actors whose membership is split between Group 2, and another group, indicating diminished involvement. Note that the pie charts in the figure do not represent the uncertainty classification of each actor, which has been the purpose of similar plots produced by Handcock *et al.* (2007).

### *5.2 Covariate parameters*

We now investigate the impact of covariates in the model. Recall that the continuous attributes have been standardized to facilitate interpretation of the covariate parameters.

Estimates of $\hat{\boldsymbol{\beta}}$ were obtained from 100 parametric bootstrap replications of the fitted model. These were in broad agreement with the estimates obtained by taking the approximate Hessian matrix. The two methods mainly disagreed on the significance of terms related to Group 3, the group with the least participation, with the Hessian term finding almost all covariate terms for this group important, whereas only two terms, the intercept, and the status of actors, appear to be meaningful based on the bootstrap estimates. Parameter estimates with bootstrap quantiles at the 95% level are provided in Table 6, while boxplots of the parametric bootstrap samples of $\hat{\boldsymbol{\beta}}$ are provided in Figure 8(a)–(i).

The effect of the covariates not only has an influence on the propensity for observations to belong to groups, through the mean of the mixed membership

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Table 6. *Estimates of the covariate parameters* $\boldsymbol{\beta}$. *95% bootstrap quantile ranges (2.5% and 97.5%) are included in parentheses. Estimates whose quantile range does not include zero are highlighted in bold.*

| | Group 1 | | Group 2 | | Group 3 | | Group 4 | |
|---|---|---|---|---|---|---|---|---|
| Intercept | −1.62 | (−3.01, 1.00) | −0.42 | (−2.19, 1.97) | **−1.58** | (−6.30, −0.42) | **−2.42** | (−7.14, −0.39) |
| Seniority | **−2.28** | (−3.61, −0.40) | 0.16 | (−0.50, 2.61) | −1.21 | (−2.82,0.49) | 0.88 | (−0.13, 3.71) |
| Status | **−0.90** | (−6.73, −0.72) | 0.02 | (−2.78, 0.88) | **−1.63** | (−5.55, −1.13) | 0.91 | (−2.29, 3.97) |
| Gender | **1.85** | (1.48, 7.22) | 0.15 | (−0.55, 2.07) | −1.00 | (−1.34, 1.43) | −0.23 | (−1.14, 1.28) |
| Years | **−0.90** | (−3.48, −0.27) | **−1.08** | (−3.57, −0.30) | −1.76 | (−4.00, 0.50) | **−0.77** | (−5.28, −0.48) |
| Age | −0.35 | (−0.55, 1.83) | **1.22** | (0.73 , 3.92) | −0.29 | (−2.47, 0.90) | **0.69** | (0.13, 3.68) |
| Practice | 0.21 | (−0.80, 0.82) | 0.38 | (−0.34, 1.20) | 0.81 | (−0.36, 1.94) | −0.66 | (−1.27, 0.71) |
| UConn | **−1.17** | (−3.68, −0.22) | **−2.26** | (−4.54, −0.92) | −0.78 | (−2.09, 3.68) | −0.21 | (−3.23, 1.06) |
| Other | **−1.15** | (−3.49, −0.11) | **−1.09** | (−3.49, −0.02) | −1.07 | (−2.86, 3.75) | −0.52 | (−3.56, 0.62) |

distribution. The covariates also have an influence on how much the mixed membership values vary around the mean. Thus, a large positive element in the coefficient vector implies that the mixed membership values become less dispersed as the covariate increases, and a large negative element implies that the mixed membership values become more dispersed as the covariate increases. In the case where a coefficient has similar magnitude for each group, then the value of the covariate does not have a large impact on the mean of the mixed membership distribution, but it does have an impact on the variance of the mixed membership distribution around the mean. While several covariates appear to influence the network structure, only the parameters associated with Gender and Seniority appear to differ significantly between groups.

While our interest in the parameters directly related to covariate terms is perhaps more obvious, the behavior of the intercepts are also worth considering; namely, intercept terms far from zero would indicate that the group membership in the network is poorly explained by the available covariate information. Of the four groups, the intercepts of Groups 3 and 4 are consistently below zero, although with quite high variance. These groups have the fewest significant covariate terms, which also suggests that their structure is only partly explained by the covariate information.

At least one covariate appears to play some part in explaining each group's structure. Gender appears to have the most sizable effect, in particular on membership to Group 1, where several other covariates are also influential, including seniority, status, years with the firm and type of law school. It is interesting to note that despite the fact that the covariates Seniority and Years with the Firm are negatively correlated, their respective parameters for Group 1 are in agreement. This reflects the difference in distribution between the covariates. Whereas Seniority is inherently evenly distributed across the data due to its ranked nature, the Years with the Firm covariate is strongly positively skewed, reflecting the firm's tendency to recruit many junior staff and retain only the most successful. In terms of Group 1, the group consists mainly of the more senior and long-established actors in the network, yet the very oldest and most experienced actors in the network are less involved in the network. Thus, the parameter penalizes the very oldest actors in the network from strong membership to Group 1, where the standardized value of Years with the Firm is further from the standardized mean (2.23 standard deviations) than Seniority (1.69 standard deviations).

A similar effect occurs in Group 2, where Years with the firm and Age disagree despite their positive correlation in the data. The difference in distribution between these covariates is less pronounced; however, Age is not so strongly positively skewed as Years with the Firm. The values of the parameters mean that younger actors with relatively little experience are assigned high prior probability to Group 4, while the older actors with highest experience are assigned with high prior probability to Group 2. In several cases, actors are assigned relatively high prior probability to both groups.

Note that this interpretation is contingent on whether or not all groups share the same parameter value for Age. If no differences between the parameters existed, an alternative scenario is appropriate. In this case, large values of Age would make the membership distribution less dispersed around the existing prior probabilities,

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Table 7. *Estimates of difference between group parameters* $\boldsymbol{\beta}$ *for Age covariate. 95% bootstrap quantile ranges (2.5% and 97.5%) are included in parentheses. Estimates whose quantile range does not include zero are highlighted in bold.*

| | Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|---|
| Group 1 | – | – | – | – |
| Group 2 | **−1.57** (−2.46, −1.15) | – | – | – |
| Group 3 | −0.06 (−0.18, 2.97) | **1.51** (1.43, 5.37) | – | – |
| Group 4 | **−1.04** (−2.35, −0.31) | 0.53 (−0.27, 1.18) | **−0.98** (−5.14, −0.68) | – |

rather than biasing membership towards group 2. To investigate further, we inspected estimates with bootstrap quantiles at the 95% level of the between group differences for these parameters, shown in Table 7. These values suggest that the parameter values for Groups 1 and 3 are distinct from those for Groups 2 and 4, indicating that the former interpretation in this case is appropriate.

Group 3 are the group perhaps least well explained by the covariates. Almost all of the continuous covariates for actors assigned to Group 3 with high prior probability were within one standard deviation of the mean. Noticeably, however, almost all of these actors have partner status, the one covariate parameter which appears significant based on the bootstrap estimates.

While the law-school parameters appear significant in this analysis, it must be noted that the upper quantiles for these parameters are close to zero, and that the variance for these parameters is large, particularly the comparison between actors attending other law schools and those attending Harvard or Yale. If the uncertainty surrounding these terms were even slightly underestimated, then it seems likely that at least two of these terms would no longer seem significant. An exception is the negative impact which attending the law school at the University of Connecticut has on membership to Group 2 in comparison with the baseline law school of Harvard or Yale,where the impact seems to be quite large. Finally, we note that the type of practice engaged in by the actors appears to have little impact on whom they form friendships with in this model setting.

### *5.3 Goodness of fit*

Properties of the fitted model are now examined so as to determine how well the model fits the data. Hoff *et al.* (2002) note that one desirable property of a model is that its predictive probabilities for links and non-links be well separated. Figure 9(a) shows boxplots of the predicted probabilities for links and non-links of the data based on the fitted parameters outlined in Section 5. The two boxplots show a high degree of separation, with the lower quartile of the observed link probabilities at roughly the same level as the top whisker of the observed non-links. Another approach is to evaluate how well the data predicts links in a hold-out modeling approach using a receiver operating characteristic (ROC) curve (Hoff, 2008). This is shown in Figure 9(b). Again, the model appears to perform quite well, with a total area under the curve (AUC) score of almost 0.86.

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

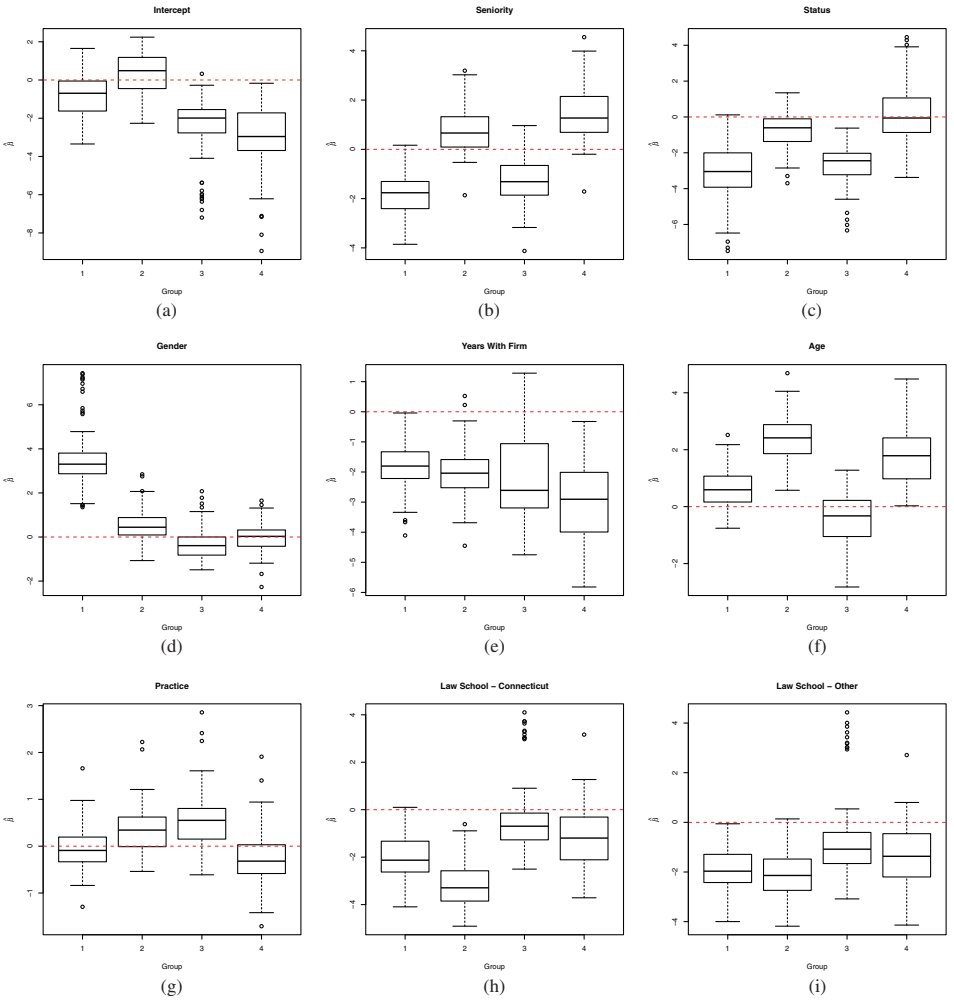
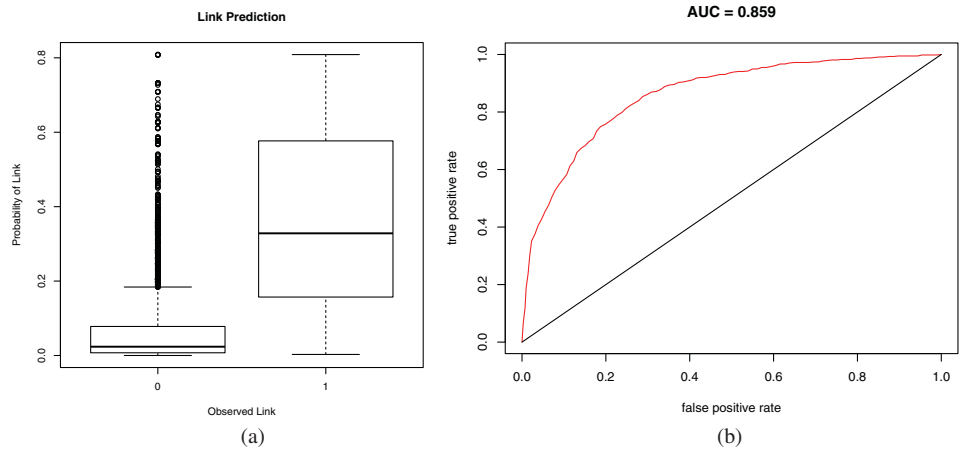



Fig. 8. Boxplots of parametric bootstrap samples of covariate parameter estimates $\hat{\boldsymbol{\beta}}$ for the 4 group MMESBM. The dashed line occurs at zero. (color online)

Another approach to checking model fit is based on network simulation (Hunter *et al*., 2008; Krivitsky & Handcock, 2008; Salter-Townshend & Murphy, 2013). The main idea is to generate networks based on the fitted model parameters and then compare properties of these simulated networks to the observed network. Network properties which are not directly based on model parameters are considered the best indicators of model fit (Hunter *et al*., 2008). Here, the model performs less well than suggested by the link prediction measures.

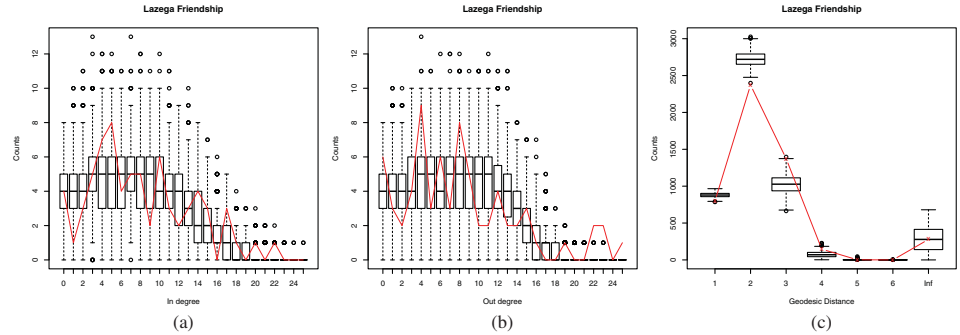
We compare the simulated networks to the observed network with respect to the following summary statistics: in degree, out degree, and geodesic distance. Plots of these statistics are shown in Figures 10(a)–(c). These show the observed network summary statistics as a line superimposed over boxplots of the same statistics obtained from 100 simulations. While the general behavior of the statistics is reasonably well accounted for, the upper and lower quartiles of the in and out degree statistics appear to be too narrow, indicating a lack of variability in

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press




(a) (b)

Fig. 9. Boxplot of link probabilities for the data based on fitted parameters. Note the high degree of separation between the probabilities for present and absent observed links. The plot on the right shows the ROC for link prediction for each of the held-out data samples during the 10-fold cross validation process. (color online)




(a) (b) (c)

Fig. 10. Goodness of fit diagnostics for the 4 group MMESBM. (color online)

the simulated data. The simulated networks also fail to account for the actors in the network with the highest in and out degree. For minimum geodesic distance, while the model correctly predicts that the majority of actors are connected by two degrees of separation, it overestimates this number while underestimating the number of actors connected by three degrees of separation, again suggesting a lack of variability. This may be caused by the underestimation of uncertainty in the data generative process caused by the variational Bayes approximation already discussed in Section 3.

## 6 Conclusion

The large number of network models which have recently been introduced and extended provide the analyst with ever more tools with which to analyze relational data. In the future, it may be of interest to combine the MMESBM with other extensions to MMSBM, such as the dynamic MMSBM of Xing *et al.* (2010). Alternatively, it would be interesting to develop models such as the overlapping SBM of Latouche *et al.* (2011) to include covariates, and to compare such a

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

model's performance to that of the MMESBM. For example, the specification of the overlapping SBM means that fewer parameters must be estimated in comparison to the MMSBM—this could mean that the model is preferable, since its fit to the network is more parsimonious, or perhaps there are aspects of the data which are not captured as successfully as for the MMSBM.

It is interesting to note the flexibility of the mixed membership approach, beyond allowing actors to interact in multiple social circles; for example, a majority of actors in the Lazega Lawyers dataset were assigned partial membership to Group 2, a group characterized by low interaction. This can be interpreted as the model accounting for degree heterogeneity in the dataset, which must be explicitly modeled for (Krivitsky *et al*., 2009) when using a latent position model.

In this paper, an approach to incorporate actor covariates into the MMSBM has been introduced and demonstrated on a dataset. This complements the work of Mariadassou *et al.* (2010), who incorporate link-specific attributes into the SBM. Care must be taken when choosing how to include this information into a model. For example, when incorporating covariates into a latent space model, Gormley & Murphy (2010) found that the manner in which they entered the model had a major impact on the interpretation of the resulting analysis. Where possible, specifying actor-specific covariates is arguably more easily interpretable: Wasserman & Faust (1994, Chapter 16) state that two actors $i$ and $i\prime$ are stochastically equivalent if the probability of an event, in this case a link to an actor $j$, is unchanged by the interchanging of the actors. This is not the case when link-specific attributes must also be considered.

While the variational Bayes method is an effective method for inference, at least from a computational and clustering perspective, in its directly implemented form its computational cost is still of order $O(N^2)$. As currently implemented, the algorithm took several minutes to fit a single model to the Lazega Lawyers data. The outlined cross-validation approach for model selection, and the bootstrapping procedure used to evaluate covariate parameters provide a further computational burden, since the model must effectively be re-fit to the data multiple times. The case-control approximated likelihood approach introduced by Raftery *et al.* (2012) for latent space models, which has been successfully applied in a variational Bayes setting by Salter-Townshend & Murphy (2013) could prove effective when fitting the model to larger networks.

Model choice remains a challenge for network and mixed membership models, within the model based clustering literature and beyond. While the hold out likelihood approach which has been used in this paper gives some idea of which group choices are most suitable for the data, a high level of uncertainty still surrounds the identification of an optimal model. Similarly, care must be taken when determining which covariates appear to impact on the data.

The MMESBM as specified here can be seen to treat the covariate parameters as nuisance parameters, when they are of as much or greater interest as the other parameters in the model. While the introduction of a hyper prior would allow for inference to be performed in a more principled manner, it would also make it much more complicated, as the conjugacy between distributions would be lost. Similarly, certain properties of the Dirichlet distribution may prove too restrictive when modeling the group membership of actors, especially with the introduction

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

of covariates; the use of other distributions, such as a logistic normal distribution may prove useful (Aitchison, 1982; Blei & Lafferty, 2007). Again, this would lead to additional inferential complexity.

While not a particular goal of the paper, it remains unclear how to choose between progressively more complex classes of model such as the SBM and MMSBM, or whether or not to include covariates, when analyzing a given dataset. Potentially, another class of model, such as the latent space or ERGM may be more suitable. Hoff (2008) compares fundamentally different methods by assessing their link predictive properties on hold-out samples of data, and it may be possible to extend the use of the hold-out likelihood method employed by this paper for model selection, not just for the number of groups but also for the class of model. This possibility comes with the caveat that link prediction is expressfully the primary goal of the analyst, when other properties in the network may be viewed as equally or more important.

## Acknowledgment

This material is based upon work completed while both authors were based in University College Dublin, and was supported by the Science Foundation Ireland under Grant No. 08/SRC/I1407: Clique: Graph & Network Analysis Cluster and 12/RC/2289:Insight Research Centre.

## References

Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions* (1st ed.). Mineola, New York: Dover Publications.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., Goldberg, A., Xing, E. P., & Zheng, A. X. (2007). *Statistical network analysis: Models, issues and new directions*, Lecture Notes in Computer Science, vol. 4503. Berlin: Springer.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, **9**, 1981–2014.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**(2), 139–177.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**(1), 1–10.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, University College London.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and its Application*, **1**, 203–232.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *Annals of Applied Statistics*, **1**(1), 17–35.

Blei, D. M., Ng, Andrew Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Breiger, R. L. (1974). The duality of persons and groups. *Social Forces*, **53**(2), 181–190.

Daudin, J.-J., Picard, F., & Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, **18**(2), 173–183. 10.1007/s11222-007-9046-7.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, **39**(1), 1–38.

Efron, B. (2013). *Empirical Bayes modeling, computation, and accuracy*. Tech. rept. Stanford University.

Efron, B., & Morris, C. (1973). Combining possibly related estimation problems. *Journal of the Royal Statistical Society, Series B*, **35**(3), 379–421.

Erdős, P. & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae debrecen*, **6**, 290–297.

Erosheva, E. A., Fienberg, S. E., & Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, **1**(2), 502–537.

Fellows, I., & Handcock, M. S. (2012). Exponential-family random network models. *Arxiv e-prints*.

Gormley, I. C., & Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Statistical Methodology*, **7**(3), 385–405.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A*, **170**(2), 1–22.

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**(16), 2409–2419.

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, **54**(2), 427–432.

Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20*. (pp. 657–664) Cambridge, MA: MIT Press.

Hoff, P., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, **97**(460), 1090–1098.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, **5**(2), 109–137.

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

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

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, **3**(1), 79–87.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Krivitsky, P. N., & Handcock, M. S. (2008). Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software*, **24**(5), 1–23.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., & 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.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**(1), 79–86.

Latouche, P., Birmelé, E., & Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *Annals of Applied Statistics*, **5**(1), 309–336.

Latouche, P., Birmelé, E., & Ambroise, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, **12**(1), 93–115.

Mariadassou, M., Robin, S., & Vacher, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, **4**(2), 715–742.

McDaid, A. F, Murphy, T. B., Friel, N., & Hurley, N. (2012). Model-based clustering in networks with stochastic community finding. A. Colubi, K. Fokianos, E. J. Kontoghiorghes, & G. Gonzáles-Rodríguez (Eds.), *Proceedings of COMPSTAT 2012: 20th International conference on computational statistics*. Limassol, Cyprus: ISI-IASC, pp. 549–560.

Minka, T. P. (2012). *Estimating a Dirichlet distribution*. Online Manuscript.

Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction of stochastic blockstructures. *Journal of the American statistical association*, **96**(455), 1077–1087.

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, **64**(2), 140–153.

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Raftery, A. E., Niu, X., Hoff, P. D., & Yeung, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, **21**(4), 901–919.

Robbins, H. (1956). *An empirical Bayes approach to statistics*. Berkeley, California: University of California Press.

Robins, G., Snijders, T. A. B., Wang, P., & Handcock, M. S. (2006). Recent developments in exponential random graph ($p^*$) models for social networks. *Social Networks*, **29**(2), 192–215.

Rogers, S., Girolami, M., Campbell, C., & Breitling, R. (2005). The latent process decomposition of cDNA microarray datasets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **2**(2), 143–156.

Salter-Townshend, M., & Murphy, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Computational Statistics and Data Analysis*, **57**(1), 661–671.

Salter-Townshend, M., White, A., Gollini, I., & Murphy, T. B. (2012). Review of statistical network analysis: Models, algorithms, and software. *Statistical Analysis and Data Mining*, **5**(4), 243–264.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, **10**(1), 63–72.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, **3**(2), 1–40.

Snijders, T. A. B., & Nowicki, K. (1997). Estimation and prediction for stochastic bockmodels for graphs with latent block structure. *Journal of Classification*, **14**(1), 75–100.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, **36**(1), 99–153.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

White, A., Chan, J., Hayes, C., & Murphy, T. B. (2012). Mixed membership models for exploring user roles in online fora. In N. Ellison, J. G. Shanahan, & Z. Tufekci (Eds.), *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM 2012)*, Dublin, Ireland, pp. 599–602.

Xing, E. P., Fu, W., & Song, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *Annals of Applied Statistics*, **4**(2), 535–566.

Zanghi, H., Picard, F., Miele, V., & Ambroise, C. (2010b). Strategies for online inference of model-based clustering in large and growing networks. *Annals of Applied Statistics*, **4**(2), 687–714.

Zanghi, H., Volant, S., & Ambroise, C. (2010a). Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, **31**(9), 830–836.

Zhang, Y., Levina, E., & Zhu, J. (2013). *Community detection in networks with node features*. Lake Tahoe, Nevada.

## Appendix: Estimating model parameters

We know provide details for how the estimates given in Section 3 were derived. The general idea, for a given distribution with parameters $\Omega_1, \ldots, \Omega_J$, is to approximate a posterior $p(\Omega) = p(\Omega_1, \ldots, \Omega_J)$ with a set of distributions $q(\Omega_1), \ldots, q(\Omega_J)$ which can factorized independently, such that

$$p(\Omega_1, \ldots, \Omega_J) \approx q(\Omega_1) \times \cdots \times q(\Omega_J).$$

It can be shown (Bishop, 2006) that the optimal (i.e., the Kullback–Liebler divergence minimizing) form for $q(\Omega_j)$ can be found by setting

$$q(\Omega_j) \propto \exp\left(\mathbb{E}^q_{i \neq j}\left[\log p(\Omega)\right]\right).$$

We make the following approximation:

$$p(\mathbf{Y}, \mathbf{Z}^1, \mathbf{Z}^2, \boldsymbol{\tau}, \boldsymbol{\Theta} | \boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \boldsymbol{\beta}, \mathbf{W}) \approx q(\mathbf{Z}^1 | \boldsymbol{\phi}^1) q(\mathbf{Z}^2 | \boldsymbol{\phi}^2) q(\boldsymbol{\tau} | \boldsymbol{\gamma}) q(\boldsymbol{\theta} | \boldsymbol{\zeta}^{\mathbf{1}}, \boldsymbol{\zeta}^2),$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

where we have introduced the variational parameters $\boldsymbol{\phi}^1, \boldsymbol{\phi}^2, \boldsymbol{\zeta}^{\mathbf{1}}, \boldsymbol{\zeta}^2$, and $\boldsymbol{\gamma}$.

Keeping $\boldsymbol{\beta}$ fixed, and setting each $\delta_{ig} = \exp(\sum_{p=1}^{P} W_{ip}\beta_{gp})$, inference for $q(\boldsymbol{\tau}|\boldsymbol{\gamma})$ is as follows:

$$\begin{aligned} q(\boldsymbol{\tau}_i|\boldsymbol{\gamma}_i) &\propto \exp\left\{ \mathbb{E}^q_{\mathbf{Z}^1,\mathbf{Z}^2}\left[ \sum_{j=1}^{N} \log p(\mathbf{Z}^1_{ij}|\boldsymbol{\tau}_i) + \log p(\mathbf{Z}^2_{ji}|\boldsymbol{\tau}_i) + \log p(\boldsymbol{\tau}_i|\boldsymbol{\delta}) \right] \right\} \\ &\propto \prod_{g=1}^{G} \tau_{ig}^{\delta_{ig}-1} \times \exp\left\{ \sum_{j=1}^{N}\sum_{h=1}^{G} \left( \mathbb{E}^q_{\mathbf{Z}^1}\left[Z^1_{ijh}\right] \log \tau_{ig} + \mathbb{E}^q_{\mathbf{Z}^2}\left[Z^2_{jih}\right] \log \tau_{ig} \right) \right\} \\ &= \prod_{g=1}^{G} \tau_{ig}^{\delta_{ig}-1+\sum_{j=1}^{N} \mathbb{E}^q_{\mathbf{Z}^1}\left[Z^1_{ijg}\right] + \mathbb{E}^q_{\mathbf{Z}^2}\left[Z^2_{jig}\right]}, \end{aligned}$$

which we can recognize as a Dirichlet distribution. It is also straightforward to see that $q(\boldsymbol{\theta}|\boldsymbol{\zeta}^{\mathbf{1}}, \boldsymbol{\zeta}^2)$ is a beta distribution:

$$\begin{aligned} q(\theta_{gh}|\zeta^1_{gh}, \zeta^2_{gh}) &\propto \exp\left\{ \mathbb{E}^q_{\mathbf{Z}^1,\mathbf{Z}^2}\left[ \sum_{i=1}^{N}\sum_{j=1}^{N} \log p(Y_{ij}|Z^1_{ij}, Z^2_{ij}, \theta_{gh}) + \log p(\theta|\alpha^1_{gh}, \alpha^2_{gh}) \right] \right\} \\ &= \theta_{gh}^{\zeta^1_{gh}} (1-\theta_{gh})^{\zeta^2_{gh}}, \end{aligned}$$

where

$$\begin{aligned} \zeta^1_{gh} &= \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}^q_{\mathbf{Z}^1_{ij}}\left[Z^1_{ijg}\right] \mathbb{E}^q_{\mathbf{Z}^2{}_{ij}}\left[Z^2_{ijh}\right] Y_{ij} + \alpha^1_{gh} \\ \zeta^2_{gh} &= \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}^q_{\mathbf{Z}^1_{ij}}\left[Z^1_{ijg}\right] \mathbb{E}^q_{\mathbf{Z}^2{}_{ij}}\left[Z^2_{ijh}\right] (1-Y_{ij}) + \alpha^2_{gh}. \end{aligned}$$

Note that the calculation of $q(\boldsymbol{\tau}_i|\boldsymbol{\gamma}_i)$ did not require taking the expectation of the log of the joint distribution with respect to $\boldsymbol{\theta}$, and vice versa. This is because the parameters are conditionally independent of one another due to the presence of the indicator variable $\mathbf{Z}$. This is perhaps most clearly seen in the diagram in Figure 2(b).

Calculating $q(\mathbf{Z}^1_{ij}|\boldsymbol{\phi}^1_{ij})$ is a little trickier, since we must calculate $\mathbb{E}^q_{\boldsymbol{\theta}}\left[\log \theta_{gh}\right]$ and $\mathbb{E}^q_{\boldsymbol{\tau}}\left[\log \tau_{ig}\right]$:

$$\begin{aligned} q(\mathbf{Z}^1_{ij}|\boldsymbol{\phi}^1_{ij}) &\propto \exp\left\{ \mathbb{E}^q_{\boldsymbol{\tau},\boldsymbol{\theta},\mathbf{Z}^2}\left[\log p(Y_{ij}|\mathbf{Z}^1_{ij}, \mathbf{Z}^2_{ij}, \boldsymbol{\theta}) + \log p(\mathbf{Z}_{ij}|\boldsymbol{\tau}_i)\right] \right\} \\ &= \exp\left\{ \sum_{g=1}^{G} Z^1_{ijg} \left( \sum_{h=1}^{G} \mathbb{E}^q_{\mathbf{Z}^2{}_{ij}}\left[Z^2_{ijh}\right] \left( Y_{ij}\mathbb{E}^q_{\theta_{gh}}\left[\log \theta_{gh}\right] + (1-Y_{ij})\mathbb{E}^q_{\theta_{gh}}\left[\log(1-\theta_{gh})\right] \right) \right.\right. \\ &\quad \left.\left. + \ \mathbb{E}^q_{\boldsymbol{\tau}_i}\left[\log \tau_{ig}\right] \right) \right\} \\ &= \prod_{g=1}^{G} \left\{ \prod_{h=1}^{G} \left[ \exp\left( \mathbb{E}^q_{\theta_{gh}}\left[\log \theta_{gh}\right] \right)^{Y_{ij}} \exp\left( \mathbb{E}^q_{\theta_{gh}}\left[\log(1-\theta_{gh})\right] \right)^{1-Y_{ij}} \right]^{\mathbb{E}^q_{\mathbf{Z}^2_{ij}}\left[Z^2_{ijh}\right]} \right. \\ &\quad \left. \times \exp\left( \mathbb{E}^q_{\boldsymbol{\tau}_i}\left[\log \tau_{ig}\right] \right) \right\}^{Z^1_{ijg}}. \end{aligned}$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

Similarly,

$$q(\mathbf{Z}_{ij}^2|\boldsymbol{\phi}_{ij}^2) \propto \prod_{h=1}^{G}\left\{\prod_{g=1}^{G}\left[\exp\left(\mathbb{E}_{\theta_{gh}}^{q}\left[\log\theta_{gh}\right]\right)^{Y_{ij}}\exp\left(\mathbb{E}_{\theta_{gh}}^{q}\left[\log(1-\theta_{gh})\right]\right)^{1-Y_{ij}}\right]^{\mathbb{E}_{\mathbf{Z}_{ij}^1}^{q}[Z_{ijg}^1]} \times \exp\left(\mathbb{E}_{\boldsymbol{\tau}_j}^{q}\left[\log\tau_{jh}\right]\right)\right\}^{Z_{ijh}^2}.$$

We can recognize both $q(\mathbf{Z}_{ij}^1|\boldsymbol{\phi}_{ij}^1)$ and $q(\mathbf{Z}_{ij}^2|\boldsymbol{\phi}_{ij}^2)$ to be multinomial distributions.

Since the approximate distributions all have tractable form, we can calculate the required expectations, and give updates in fully parametric form:

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}_{ij}^1}^{q}\left[Z_{ijg}^1\right] &= \phi_{ijg}^1\\
\mathbb{E}_{\mathbf{Z}_{ij}^2}^{q}\left[Z_{ijg}^2\right] &= \phi_{ijg}^2\\
\mathbb{E}_{\boldsymbol{\tau}_i}^{q}\left[\log\tau_{ig}\right] &= \Psi(\gamma_{ig}) - \Psi\left(\sum_{k=1}^{G}\gamma_{ik}\right)\\
\mathbb{E}_{\theta_{gh}}^{q}\left[\log\theta_{gh}\right] &= \Psi(\zeta_{gh}^1) - \Psi(\zeta_{gh}^1+\zeta_{gh}^2)\\
\mathbb{E}_{\theta_{gh}}^{q}\left[\log(1-\theta_{gh})\right] &= \Psi(\zeta_{gh}^2) - \Psi(\zeta_{gh}^1+\zeta_{gh}^2).
\end{aligned}$$

Parameter updates then become

$$\begin{aligned}
\zeta_{gh}^1 &= \sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{ijg}^1\phi_{ijh}^2 Y_{ij} + \alpha_{gh}^1,\\
\zeta_{gh}^2 &= \sum_{i=1}^{N}\sum_{j=1}^{N}\phi_{ijg}^1\phi_{ijh}^2(1-Y_{ij}) + \alpha_{gh}^2,\\
\gamma_{ig} &= \delta_{ig} + \sum_{j=1}^{N}(\phi_{ijg}^1 + \phi_{jig}^2)\\
\phi_{ijg}^1 &\propto \exp\left(\Psi\left(\gamma_{ig}\right) - \Psi\left(\sum_{k=1}^{G}\gamma_{ik}\right)\right),\\
&\quad\times\exp\left\{\sum_{h=1}^{G}\phi_{ijh}^2\left[Y_{ij}\left(\Psi\left(\zeta_{gh}^1\right) - \Psi\left(\zeta_{gh}^1+\zeta_{gh}^2\right)\right)\right.\right.\\
&\quad\left.\left.+(1-Y_{ij})\left(\Psi\left(\zeta_{gh}^2\right) - \Psi\left(\zeta_{gh}^1+\zeta_{gh}^2\right)\right)\right]\right\},\\
\phi_{ijg}^2 &\propto \exp\left(\Psi(\gamma_{jg}) - \Psi\left(\sum_{k=1}^{G}\gamma_{jk}\right)\right)\\
&\quad\times\exp\left\{\sum_{h=1}^{G}\phi_{ijh}^1\left[Y_{ij}\left(\Psi\left(\zeta_{hg}^1\right) - \Psi\left(\zeta_{hg}^1+\zeta_{hg}^2\right)\right)\right.\right.\\
&\quad\left.\left.+(1-Y_{ij})\left(\Psi\left(\zeta_{hg}^2\right) - \Psi\left(\zeta_{hg}^1+\zeta_{hg}^2\right)\right)\right]\right\}.
\end{aligned}$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

### *A.1 Estimating covariate parameters*

Recall that the log-posterior is intractable, and that we instead maximize a lower bound $\mathscr{L}$:

$$\mathscr{L} = \mathbb{E}^q_{\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\theta}}\left[\log p(\mathbf{Y},\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\theta}|\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,\boldsymbol{\delta})\right] - \mathbb{E}^q_{\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\theta}}\left[\log q(\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\theta})\right],$$

where

$$\begin{aligned}\mathbb{E}^q_{\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\theta}}\left[\log p(\mathbf{Y},\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\theta}|\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,\boldsymbol{\delta})\right] &= \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}^q_{\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\theta}}\left[\log p(Y_{ij}|\mathbf{Z}^1_{ij},\mathbf{Z}^2_{ij},\boldsymbol{\theta})\right]\\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}^q_{\mathbf{Z}^1,\boldsymbol{\tau}}\left[\log p(\mathbf{Z}^1_{ij}|\boldsymbol{\tau}_i)\right]\\ &+ \mathbb{E}^q_{\mathbf{Z}^2,\boldsymbol{\tau}}\left[\log p(\mathbf{Z}^2_{ij}|\boldsymbol{\tau}_j)\right]\\ &+ \sum_{n=1}^{N}\mathbb{E}^q_{\boldsymbol{\tau}}\left[\log p(\boldsymbol{\tau}_n|\boldsymbol{\delta})\right]\\ &+ \mathbb{E}^q_{\boldsymbol{\theta}}\left[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2)\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}^q_{\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\theta}}\left[\log q(\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\tau},\boldsymbol{\theta})\right] &= \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}^q_{\mathbf{Z}^1}\left[\log q(\mathbf{Z}^1_{ij}|\boldsymbol{\phi}^1_{ij})\right] + \mathbb{E}^q_{\mathbf{Z}^2}\left[\log q(\mathbf{Z}^2_{ij}|\boldsymbol{\phi}^2_{ij})\right]\\ &+ \sum_{n=1}^{N}\mathbb{E}^q_{\boldsymbol{\tau}}\left[\log q(\boldsymbol{\tau}_n|\boldsymbol{\gamma}_n)\right] + \mathbb{E}^q_{\boldsymbol{\theta}}\left[\log q(\boldsymbol{\theta}|\boldsymbol{\zeta}^1,\boldsymbol{\zeta}^2)\right].\end{aligned}$$

This is straightforward to calculate:

$$\begin{aligned}\mathbb{E}^q_{\mathbf{Z}^1,\mathbf{Z}^2,\boldsymbol{\theta}}\left[\log p(Y_{ij}|\mathbf{Z}^1_{ij},\mathbf{Z}^2_{ij},\boldsymbol{\theta})\right] &= \sum_{g=1}^{G}\sum_{h=1}^{G}\phi^1_{ijg}\phi^2_{ijh}\left\{Y_{ij}(\Psi(\zeta^1_{gh}) - \Psi(\zeta^1_{gh}+\zeta^2_{gh}))\right.\\ &\left. + (1-Y_{ij})(\Psi(\zeta^2_{gh}) - \Psi(\zeta^1_{gh}+\zeta^2_{gh}))\right\}\\ \mathbb{E}^q_{\mathbf{Z}^1,\boldsymbol{\tau}}\left[\log p(\mathbf{Z}^1_{ij}|\boldsymbol{\tau}_i)\right] &= \mathbb{E}^q_{\mathbf{Z}^1,\boldsymbol{\tau}}\left[\sum_{g=1}^{G}Z^1_{ijg}\log\tau_{ig}\right]\\ &= \sum_{g=1}^{G}\phi^1_{ijg}\times\left\{\Psi(\gamma_{ig}) - \Psi\left(\sum_{k=1}^{G}\gamma_{ik}\right)\right\},\\ \mathbb{E}^q_{\mathbf{Z}^2,\boldsymbol{\tau}}\left[\log p(\mathbf{Z}^2_{ij}|\boldsymbol{\tau}_j)\right] &= \sum_{g=1}^{G}\phi^2_{ijg}\times\left\{\Psi(\gamma_{jg}) - \Psi\left(\sum_{k=1}^{G}\gamma_{jk}\right)\right\},\\ \mathbb{E}^q_{\boldsymbol{\tau}}\left[\log p(\boldsymbol{\tau}_n|\boldsymbol{\delta})\right] &= \left[\log\Gamma\left(\sum_{h=1}^{G}\delta_h\right) - \sum_{k=1}^{G}\log\Gamma(\delta_k)\right.\\ &\left. + \sum_{g=1}^{G}(\delta_g-1)\log\tau_{ng}\right]\end{aligned}$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

$$= \log\Gamma\left(\sum_{h=1}^{G}\delta_h\right) - \sum_{k=1}^{G}\log\Gamma(\delta_k) + \sum_{g=1}^{G}(\delta_g - 1) \times \left\{\Psi(\gamma_{ng}) - \Psi\left(\sum_{k=1}^{G}\gamma_{nk}\right)\right\},$$

$$\mathbb{E}^q_{\boldsymbol{\theta}}\left[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2)\right] = \sum_{g=1}^{G}\sum_{h=1}^{G}\log\Gamma\left(\alpha^1_{gh}+\alpha^2_{gh}\right) - \log\Gamma(\alpha^1_{gh}) - \log\Gamma(\alpha^2_{gh}) + \sum_{g=1}^{G}(\alpha^1_{gh}-1)\left\{\Psi(\zeta^1_{gh}) - \Psi(\zeta^1_{gh}+\zeta^2_{gh})\right\} + \sum_{g=1}^{G}(\alpha 2_{gh}-1)\left\{\Psi(\zeta^2_{gh}) - \Psi(\zeta^1_{gh}+\zeta^2_{gh})\right\}.$$

The terms for the second part of the lower bound are given below:

$$\mathbb{E}^q_{\mathbf{Z}^1}\left[\log q(\mathbf{Z}^1_{ij}|\boldsymbol{\phi}^1_{ij})\right] = \mathbb{E}^q_{\mathbf{Z}^1}\left[\sum_{g=1}^{G} Z^1_{ijg}\log\phi^1_{ijg}\right] = \sum_{g=1}^{G}\phi^1_{ijg}\log\phi^1_{ijg},$$

$$\mathbb{E}^q_{\mathbf{Z}^2}\left[\log q(\mathbf{Z}^2_{ij}|\boldsymbol{\phi}^2_{ij})\right] = \sum_{g=1}^{G}\phi^2_{ijg}\log\phi^2_{ijg},$$

$$\mathbb{E}^q_{\boldsymbol{\tau}}\left[\log q(\boldsymbol{\tau}_{\boldsymbol{n}}|\boldsymbol{\gamma}_{\boldsymbol{n}})\right] = \mathbb{E}^q_{\boldsymbol{\tau}}\left[\log\Gamma\left(\sum_{h=1}^{G}\gamma_{nh}\right) - \sum_{k=1}^{G}\log\Gamma(\gamma_{nk}) + \sum_{g=1}^{G}(\gamma_{ng}-1)\log\tau_{ng}\right]$$

$$= \log\Gamma\left(\sum_{h=1}^{G}\gamma_{nh}\right) - \sum_{k=1}^{G}\log\Gamma(\gamma_{nk}) + \sum_{g=1}^{G}(\gamma_{ng}-1)\times\left\{\Psi(\gamma_{ng}) - \Psi\left(\sum_{k=1}^{G}\gamma_{nk}\right)\right\},$$

$$\mathbb{E}^q_{\boldsymbol{\theta}}\left[\log q(\boldsymbol{\theta}|\boldsymbol{\zeta}^1,\boldsymbol{\zeta}^2)\right] = \sum_{g=1}^{G}\sum_{h=1}^{G}\log\Gamma\left(\zeta^1_{gh}+\zeta^2_{gh}\right) - \log\Gamma(\zeta^1_{gh}) - \log\Gamma(\zeta^2_{gh}) + \sum_{g=1}^{G}(\zeta^1_{gh}-1)\left\{\Psi(\zeta^1_{gh}) - \Psi(\zeta^1_{gh}+\zeta^2_{gh})\right\} + \sum_{g=1}^{G}(\zeta^2_{gh}-1)\left\{\Psi(\zeta^2_{gh}) - \Psi(\zeta^1_{gh}+\zeta^2_{gh})\right\}.$$

To estimate $\hat{\boldsymbol{\beta}}$, we make us of a Newton–Raphson step to iteratively maximize $\mathscr{L}$. It's simpler to first calculate the gradient and Hessian functions in terms of

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

$\boldsymbol{\delta}$:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \delta_i} &= \frac{\partial}{\partial \delta_i} \sum_{n=1}^{N} \mathbb{E}_{\tau}^{q} \left[ \log p(\tau_n | \delta) \right] \\
&= \frac{\partial}{\partial \delta_i} \sum_{n=1}^{N} \left[ \log \Gamma \left( \sum_{h=1}^{G} \delta_h \right) - \sum_{h=1}^{G} \log \Gamma(\delta_h) \right. \\
&\quad \left. + \sum_{h=1}^{G} (\delta_h - 1) \left\{ \Psi(\gamma_{ng}) - \Psi \left( \sum_{h=1}^{G} \gamma_{nh} \right) \right\} \right] \\
&= \frac{\partial}{\partial \delta_i} N \left\{ \log \Gamma \left( \sum_{h=1}^{G} \delta_h \right) - \sum_{k=1}^{G} \log \Gamma(\delta_k) \right\} \\
&\quad + \frac{\partial}{\partial \delta_i} \sum_{n=1}^{N} \sum_{g=1}^{G} (\delta_g - 1) \left\{ \Psi(\gamma_{ng}) - \Psi \left( \sum_{h=1}^{G} \gamma_{nh} \right) \right\} \\
&= N \left\{ \Psi \left( \sum_{h=1}^{G} \delta_h \right) - \Psi(\delta_i) \right\} + \sum_{n=1}^{N} \left\{ \Psi(\gamma_{ni}) - \Psi \left( \sum_{h=1}^{G} \gamma_{nh} \right) \right\}. \\
\Rightarrow \frac{\partial^2 \mathscr{L}}{\partial \delta_i \partial \delta_j} &= N \left\{ \Psi' \left( \sum_{h=1}^{G} \delta_h \right) - \mathbb{I}_{i=j} \Psi' (\delta_i) \right\}.
\end{aligned}$$

Now, noting that $\frac{\partial \delta_{ig}}{\partial \beta_{gp}} = W_{ip} \exp(\sum_{p=1}^{P} W_{ip} \beta_{gp})$, we can then maximize the lower bound $\mathscr{L}$ with respect to $\boldsymbol{\beta}$ by making use of the chain rule:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \beta_{iq}} &= \sum_{n=1}^{N} \frac{\partial \mathscr{L}}{\partial \delta_{ni}} \frac{\partial \delta_{ni}}{\partial \beta_{iq}} \\
&= \sum_{n=1}^{N} W_{nq} \exp \left( \sum_{p=1}^{P} W_{np} \beta_{ip} \right) \\
&\quad \times \left\{ \Psi \left[ \sum_{h=1}^{G} \exp \left( \sum_{p=1}^{P} W_{np} \beta_{hp} \right) \right] - \Psi \left[ \exp \left( \sum_{p=1}^{P} W_{np} \beta_{ip} \right) \right] \right. \\
&\quad \left. + \Psi(\gamma_{ng}) - \Psi \left( \sum_{h=1}^{G} \gamma_{nh} \right) \right\}.
\end{aligned}$$

We can then calculate the Hessian matrix, again making use of the product rule:

$$\begin{aligned}
\frac{\partial^2 \mathscr{L}}{\partial \beta_{iq} \partial \beta_{jr}} &= \sum_{n=1}^{N} W_{nq} \exp \left( \sum_{p=1}^{P} W_{np} \beta_{ip} \right) \\
&\quad \times \frac{\partial}{\partial \beta_{jr}} \left\{ \Psi \left[ \sum_{h=1}^{G} \exp \left( \sum_{p=1}^{P} W_{np} \beta_{hp} \right) \right] - \Psi \left[ \exp \left( \sum_{p=1}^{P} W_{np} \beta_{ip} \right) \right] \right. \\
&\quad \left. + \Psi(\gamma_{ni}) - \Psi \left( \sum_{h=1}^{G} \gamma_{nh} \right) \right\}
\end{aligned}$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press

$$
\begin{aligned}
&+\sum_{n=1}^{N}\frac{\partial}{\partial\beta_{jr}}\left(W_{nq}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\right)\\
&\times\left\{\Psi\left[\sum_{h=1}^{G}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{hp}\right)\right]-\Psi\left[\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\right]\right.\\
&\left.+\ \Psi(\gamma_{ni})-\Psi\left(\sum_{h=1}^{G}\gamma_{nh}\right)\right\}\\
=&\sum_{n=1}^{N}W_{nq}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\\
&\times W_{nr}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{jp}\right)\left\{\Psi'\left[\sum_{h=1}^{G}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{hp}\right)\right]\right.\\
&\left.-\mathbb{I}_{i=j}\Psi'\left[\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\right]\right\}\\
&+\mathbb{I}_{i=j}\left(W_{nq}W_{nr}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\right.\\
&\times\left\{\Psi\left[\sum_{h=1}^{G}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{hp}\right)\right]-\Psi\left[\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\right]\right.\\
&\left.\left.+\Psi(\gamma_{ni})-\Psi\left(\sum_{h=1}^{G}\gamma_{nh}\right)\right\}\right).
\end{aligned}
$$

Cleaning this up a little gives the result:

$$
\begin{aligned}
\frac{\partial^2\mathscr{L}}{\partial\beta_{iq}\partial\beta_{jr}}=&\sum_{n=1}^{N}W_{nq}W_{nr}\exp\left(\sum_{p=1}^{P}W_{np}(\beta_{ip}+\beta_{jp})\right)\\
&\times\left\{\Psi'\left[\sum_{h=1}^{G}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{hp}\right)\right]-\mathbb{I}_{i=j}\Psi'\left[\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\right]\right\}\\
&+\mathbb{I}_{i=j}\left(W_{nq}W_{nr}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\right.\\
&\times\left\{\Psi\left[\sum_{h=1}^{G}\exp\left(\sum_{p=1}^{P}W_{np}\beta_{hp}\right)\right]-\Psi\left[\exp\left(\sum_{p=1}^{P}W_{np}\beta_{ip}\right)\right]\right.\\
&\left.\left.+\Psi(\gamma_{ni})-\Psi\left(\sum_{h=1}^{G}\gamma_{nh}\right)\right\}\right).
\end{aligned}
$$

https://doi.org/10.1017/nws.2015.29 Published online by Cambridge University Press