



Title	A Mixture of Experts Latent Position Cluster Model for Social Network Data
Authors(s)	Gormley, Isobel Claire, Murphy, Thomas Brendan
Publication date	2010-05
Publication information	Gormley, Isobel Claire, and Thomas Brendan Murphy. "A Mixture of Experts Latent Position Cluster Model for Social Network Data" 7, no. 3 (May, 2010).
Publisher	Elsevier
Item record/more information	http://hdl.handle.net/10197/7116
Publisher's statement	This is the author's version of a work that was accepted for publication in Statistical Methodology. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Statistical Methodology (VOL 7, ISSUE 3, (2010)) DOI: 10.1016/j.stamet.2010.01.002.
Publisher's version (DOI)	10.1016/j.stamet.2010.01.002

Downloaded 2024-04-19 06:12:03

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)



© Some rights reserved. For more information

A Mixture of Experts Latent Position Cluster Model for Social Network Data

Isobel Claire Gormley and Thomas Brendan Murphy

University College Dublin

{`claire.gormley`, `brendan.murphy`}@ucd.ie

Abstract

Social network data represent the interactions between a group of social actors. Interactions between colleagues and friendship networks are typical examples of such data.

The latent space model for social network data locates each actor in a network in a latent (social) space and models the probability of an interaction between two actors as a function of their locations. The latent position cluster model extends the latent space model to deal with network data in which clusters of actors exist — actor locations are drawn from a finite mixture model, each component of which represents a cluster of actors.

A mixture of experts model builds on the structure of a mixture model by taking account of both observations and associated covariates when modeling a heterogeneous population. Herein, a mixture of experts extension of the latent position cluster model is developed. The mixture of experts framework allows covariates to enter the latent position cluster model in a number of ways, yielding different model interpretations.

Estimates of the model parameters are derived in a Bayesian framework using a Markov Chain Monte Carlo algorithm. The algorithm is generally computationally expensive — surrogate proposal distributions which shadow the target distributions are derived, reducing the computational burden.

The methodology is demonstrated through an illustrative example detailing relations between a group of lawyers in the USA.

Key words: Clustering; covariates; latent space; mixture models; mixture of experts models; social network data; surrogate proposal distributions.

1. Introduction

Social network data has recently attracted much attention from a broad spectrum of research communities including sociology, statistics, mathematics, physics and computer science. Social network data record the interactions and relationships between a group of social entities or actors. For example, a social network data set may detail the friendship links among a group of colleagues, or it may detail the level of international trade between countries. Network data may be binary, indicating the presence/absence of a link between two actors, or it may be non-binary indicating the level of interaction between two actors. The aim of social network analysis (SNA) is to explore the structure within the network, to aid understanding of underlying phenomena and the relations that may or may not exist within the network.

Recent advances in social network analysis focus on the idea of using a latent social space in which actors are located to model the interactions between actors. Hoff et al. (2002) develop the idea of a latent social space and define the probability of a link between two actors as a function of their separation in the latent space. Moreover, conditional on the latent locations of two actors, the presence or absence of a link between them is considered to be independent of all other links in the network.

The latent position cluster model (Handcock et al., 2007) further develops the idea of a latent social space by extending the model to accommodate clusters of actors in the latent space. Under the latent position cluster model the latent location of each actor is assumed to be drawn from a finite mixture model, each component of which represents a cluster of actors. The latent position cluster model offers a more flexible version of the latent space model for modeling heterogeneous social networks. Section 2.1 provides an introduction to the latent position cluster model.

Both the latent space model and the latent position cluster model provide a framework in which covariates or attributes of the actors may be explicitly included in the model — in these models the probability of a link between two actors may be modeled as a function of both their separation in the latent space and of their relative covariates; this modeling extension facilitates homophily by attributes. However, the covariates may contribute more to the structure of the network model than solely through the link probabilities. For example, perhaps the covariates influence both the cluster membership of an actor and their link probabilities.

Herein, the latent position cluster model is extended to provide a model

38 with the flexibility to allow covariates to contribute to the network structure
39 in a number different ways. The inclusion of covariates is achieved using a
40 mixture of experts (Jacobs et al., 1991) modeling framework. In its original
41 form, the mixture of experts model is a mixture of generalized linear models
42 where additionally the probability of belonging to a cluster is modeled as a
43 logistic function of covariates. Under the latent position cluster model the
44 latent location of an actor is assumed to be drawn from a finite mixture of
45 multivariate normal distributions. The mixture of experts latent position
46 cluster model extends the latent position cluster model such that the mixing
47 proportions are modeled as functions of the actor covariates. Thus, the
48 covariates of an actor may influence the probability of a link with another
49 actor, but additionally covariates may influence the cluster membership of
50 the actor; this network modeling idea was originally proposed in Gormley
51 and Murphy (2007a).

52 The mixture of experts latent position cluster model has an intuitive mo-
53 tivation — the covariates of an actor may influence their cluster membership,
54 their cluster membership influences their latent location, and in turn their
55 latent location determines their link probabilities; covariates may also in-
56 fluence the link probabilities directly. An outline of the background to the
57 mixture of experts modeling framework is given in Section 2.2. The mixture
58 of experts latent position cluster model is developed in Section 2.3.

59 Estimation of the mixture of experts latent position cluster model is
60 achieved within a Bayesian framework using a Markov Chain Monte Carlo
61 algorithm. Both the Gibbs sampler (Geman and Geman, 1984) and the
62 Metropolis-Hastings algorithm (Metropolis et al., 1953; Chib and Greenberg,
63 1995) are employed to obtain parameter estimates. Typically, within the con-
64 text of network models, such methods are computationally expensive and re-
65 quire lengthy run times to achieve sufficient mixing of the chain. An inherent
66 problem is the selection of a suitable proposal distribution when employing
67 the Metropolis-Hastings algorithm. Additionally, tuning the parameters of
68 the selected proposal distribution in order to achieve sufficient mixing can be
69 a difficult and time consuming process, leading to inefficient model fitting.
70 Often the target distribution from which we wish to sample has a complex
71 form and standard proposal distributions perform poorly, necessitating long
72 run times to achieve sufficient mixing. The inclusion of covariates in the la-
73 tent position cluster model makes such methods even more computationally
74 expensive. Here, multivariate Taylor expansions of terms within the target
75 distributions are employed to derive suitable proposal distributions which are

76 good surrogates for the target distributions. Details of the derivation of such
77 surrogate proposal distributions are given in Section 3.2.

78 Modeling issues such as identifiability, label switching and model selection
79 are dealt with in Section 4.

80 An illustrative example of the methodology is detailed in Section 5 —
81 the mixture of experts latent position cluster model is fitted to two social
82 network data sets detailing friendship and co-worker relations between a
83 set of 71 attorneys in a northeastern USA corporate law firm. A number
84 of covariates are available for each attorney including seniority, age, office
85 location and law school. The mixture of experts latent position cluster model
86 is fitted over a range of dimensions of the latent social space and over a range
87 of numbers of clusters. Influential covariates are selected using a variable
88 selection procedure. Overall model performance is guided by the AICM (the
89 Akaike Information Criterion – Monte (Carlo)) (Raftery et al., 2007). A
90 number of covariates are deemed to be influential, offering further insight
91 to the structure of the network than was available under the default latent
92 position cluster model.

93 Section 6 comments on the mixture of experts latent position cluster
94 model in general, and suggests directions for future research.

95 **2. The mixture of experts latent position cluster model**

96 Social network data take the form of a set of relations $\{y_{i,j}\}$ from a group
97 of $i, j = 1, \dots, n$ actors, represented by an $n \times n$ sociomatrix \mathbf{Y} . The relation
98 $y_{i,j}$ between actor i and actor j may be a binary relation, indicating the
99 presence or absence of a link between the two actors. In such a case, the
100 matrix \mathbf{Y} can be thought of as a graph in which nodes are actors and edges
101 indicate a link between two nodes. The mixture of experts latent position
102 cluster model is developed within the context of binary valued relations, but
103 the methodology is easily extended to other forms of relation (such as count
104 data). Covariate data $\underline{w}_i^T = (w_{i1}, \dots, w_{ip})$ associated with actor i may also
105 be available, where p denotes the number of observed covariates.

106 We focus on the recent advances in social network analysis based on the
107 idea of a latent social space (Hoff et al., 2002) in which actors each have a
108 location. The probability of a link or relation between two actors is modeled
109 as a function of their separation in the latent space. The idea of a latent social
110 space provides the underlying framework for the latent position cluster model
111 Handcock et al. (2007).

112 *2.1. The latent position cluster model*

113 For each actor, $i = 1, 2, \dots, n$, the latent position cluster model asso-
 114 ciates a location $\underline{z}_i^T = (z_{i1}, \dots, z_{id})$ in a d -dimensional latent social space.
 115 The model assumes that the probability of a link between two actors is inde-
 116 pendent of all other links in the network, conditional on the latent locations
 117 of the actors. Hence, it follows that the likelihood function of the data is

$$\mathbf{P}(\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}) = \prod_{i=1}^n \prod_{j \neq i} \mathbf{P}(y_{i,j} | \underline{z}_i, \underline{z}_j, \underline{x}_{i,j}, \underline{\beta}) \quad (1)$$

118 where \mathbf{Z} is the $n \times d$ matrix of latent locations, \mathbf{X} is the matrix of dyadic
 119 specific covariates associated with the n actors and $\underline{\beta}$ is a vector of param-
 120 eters. In this specification $\underline{x}_{i,j}^T = (x_{i,j,1}, \dots, x_{i,j,p})$ denotes a p vector of dyadic
 121 specific covariates where typically $x_{i,j,k} = d(w_{ik}, w_{jk})$, where $d(w_{ik}, w_{jk})$ is a
 122 measure of the difference or similarity in the value of the k th covariate for
 123 actors i and j , for $k = 1, \dots, p$. The probability of a link between actors i
 124 and j is in turn modeled using a logistic regression model where

$$\log \left\{ \frac{\mathbf{P}(y_{i,j} = 1)}{\mathbf{P}(y_{i,j} = 0)} \right\} = \beta_0 + \beta_1 x_{i,j,1} + \dots + \beta_p x_{i,j,p} - \|\underline{z}_i - \underline{z}_j\| \quad (2)$$

125 The logistic regression model models the probability of a link between two actors
 126 as a function of both their dyadic specific covariates, and the Euclidean
 127 distance between them in the latent social space. The logistic regression
 128 model is appropriate for modeling binary valued relations; for alternative
 129 edge types the logistic regression term can be changed to an appropriate al-
 130 ternative generalized linear model (McCullagh and Nelder, 1983). To identify
 131 the model, as specified in Krivitsky and Handcock (2008), the regression co-
 132 efficient associated with the Euclidean distance is constrained to be -1 while
 133 the latent locations remain unconstrained.

134 To complete the specification of the latent position cluster model, it is
 135 assumed that the latent locations \underline{z}_i are drawn from a finite mixture model.
 136 This mixture modeling framework accounts for clustering of actor locations
 137 in the latent space. Specifically, the latent locations are assumed to be drawn
 138 from a mixture of G multivariate normal distributions where G denotes the
 139 unknown number of clusters of actors:

$$\underline{z}_i \sim \sum_{g=1}^G \lambda_g \text{MVN}(\underline{\mu}_g, \sigma_g^2 \mathbf{I}).$$

140 The probability of belonging to cluster g , or the g th mixing proportion, is
141 denoted λ_g ($\sum_g \lambda_g = 1$). Each component of the mixture model, or cluster,
142 is characterized by a multivariate normal distribution with a d dimensional
143 mean vector $\underline{\mu}_g$ and a diagonal covariance matrix $\sigma_g^2 \mathbf{I}$ where \mathbf{I} is the $d \times d$
144 identity matrix. Modeling the latent locations in this way relates to model-
145 based clustering of observed variables (Banfield and Raftery, 1993; Fraley
146 and Raftery, 2002).

147 A Bayesian framework can be employed for parameter estimation. As
148 is standard in Bayesian estimation of mixture models (Diebolt and Robert,
149 1994; Hurn et al., 2003) the problem is greatly simplified by augmenting the
150 observed data with an indicator variable K_i for each actor i where $K_i = g$
151 if actor i belongs to cluster g . The indicator variable K_i therefore has a
152 multinomial distribution with a single trial and probabilities equal to λ_g for
153 $g = 1, \dots, G$.

154 Parameter estimation could also be achieved via maximum likelihood,
155 but in practice this is challenging. Handcock et al. (2007) derive a two-stage
156 maximum likelihood approach to parameter estimation which first computes
157 the maximum likelihood estimates of the standard latent space model of Hoff
158 et al. (2002), and then computes the maximum likelihood estimates for the
159 mixture model applied to the resulting estimated latent locations. While
160 this approach is fast and algebraically simpler than the Bayesian approach
161 taken herein, it does not take advantage of the clustering information when
162 estimating the latent locations and was observed in practice by Handcock et
163 al. (2007) to be out-performed by the Bayesian approach.

164 In the original specification of the latent position cluster model, actor
165 covariates only play a part in determining the probability of a link between
166 two actors. It is intuitive to also allow covariates to influence other parts of
167 the model — it is conceivable that the covariates of an actor influence which
168 cluster they belong to and, conditional on their cluster membership, they
169 then have a propensity for forming relations. To facilitate such modeling
170 flexibility, a mixture of experts modeling framework is employed.

171 *2.2. Mixture of experts models*

172 Mixture of experts models originally appear in the computer science lit-
173 erature (Jacobs et al., 1991) as a mixture of generalized linear models, where
174 the mixing proportions are also modeled using generalized linear model the-
175 ory. The components of the mixture model are termed ‘expert networks’,

176 with the mixing proportions known as ‘gating networks’ leading to the name
 177 mixture of experts models.

178 In the mixture of experts model, the mixing proportions are viewed as
 179 success probabilities from a multinomial logistic regression where the prob-
 180 ability of belonging to each of $G - 1$ clusters compared to a baseline cluster
 181 is a function of the covariates of an observation. In this framework each ob-
 182 servation has a cluster membership probability, which is a direct parameter
 183 of the model. Observation i ’s mixing proportions $\underline{\lambda}_i^T = (\lambda_1(\underline{w}_i), \dots, \lambda_G(\underline{w}_i))$
 184 are modeled as a logistic function of their p covariates \underline{w}_i i.e. for $g = 2, \dots, G$

$$\log \left\{ \frac{\lambda_g(\underline{w}_i)}{\lambda_1(\underline{w}_i)} \right\} = \tau_{g0} + \tau_{g1}w_{i1} + \dots + \tau_{gp}w_{ip} \quad (3)$$

185 where cluster 1 is treated as a baseline cluster, and τ_{g0} is an intercept term.

186 In its original form the mixture of experts model also used a generalized
 187 linear model as the model within each mixture component. By changing the
 188 form of the model in the components to suit the problem at hand a flexible
 189 suite of models is available. In a univariate setting a simple linear regression
 190 model could be used; employing a logistic regression model offers a classi-
 191 fication model. Mixture of experts models have therefore been applied and
 192 developed in a wide range of areas ranging from speech recognition prob-
 193 lems (Peng et al., 1996), to modeling rank data (Gormley and Murphy, 2008,
 194 2009a). Jordan and Jacobs (1994) estimate the mixture of experts model
 195 within a classical framework via the EM algorithm; Peng et al. (1996) em-
 196 ploy a Bayesian approach. Additionally Jordan and Jacobs (1994) develop
 197 a hierarchical mixture of experts model which is a mixture of experts model
 198 which has more than one layer of mixing (ie. more than one layer of gating
 199 networks).

200 2.3. The mixture of experts latent position cluster model

201 The latent position cluster model can be extended within a mixture of
 202 experts framework. Under the latent position cluster model, the latent lo-
 203 cations \underline{z}_i ($i = 1, \dots, n$) are assumed to be drawn from a finite mixture of
 204 multivariate normal distributions with mixing proportions $\underline{\lambda}^T = (\lambda_1, \dots, \lambda_G)$.
 205 In the mixture of experts latent position cluster model framework, the mix-
 206 ing proportions are actor specific and can be modeled as a logistic function
 207 of their covariates, that is,

$$\underline{z}_i \sim \sum_{g=1}^G \lambda_g(\underline{w}_i) \text{MVN}(\underline{\mu}_g, \sigma_g^2 \mathbf{I}) \quad (4)$$

208 where $\sum_g \lambda_g(\underline{w}_i) = 1$ and by (3)

$$\mathbf{P}\{K_i = g | \underline{w}_i\} = \lambda_g(\underline{w}_i) = \frac{\exp(\tau_{g0} + \tau_{g1}w_{i1} + \dots + \tau_{gp}w_{ip})}{\sum_{g'=1}^G \exp(\tau_{g'0} + \tau_{g'1}w_{i1} + \dots + \tau_{g'p}w_{ip})} \quad (5)$$

209 where $\underline{\tau}_1^T = (0, 0, \dots, 0)$.

210 Thus the mixture of experts latent position cluster model is the latent
 211 position cluster model of Handcock et al. (2007) where the mixing propor-
 212 tions are modeled as a function of the actor covariates. Figure 1 provides
 213 an illustration of the dependencies within the original latent position cluster
 214 model and the mixture of experts latent position cluster model using
 215 graphical models.

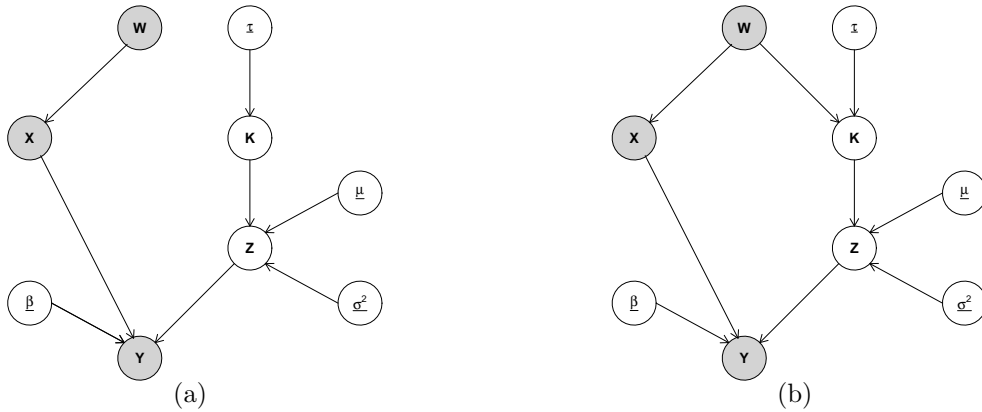


Figure 1: Graphical models illustrating the dependencies (a) within the latent position cluster model and (b) within the mixture of experts latent position cluster model.

216 The actor covariates can influence the cluster membership model (5), the
 217 link probability model (2), both models or neither model. This yields a suite
 218 of four possible network models within the mixture of experts framework.
 219 The graphical model representations for these four options can be developed

220 by including or deleting the edges from the \mathbf{W} node to the \mathbf{X} node (and
 221 from the \mathbf{X} node to the \mathbf{Y} node) and from the \mathbf{W} node to the \mathbf{K} node in
 222 Figure 1(b).

223 3. Model fitting

224 A Bayesian approach using Markov Chain Monte Carlo is utilized for
 225 fitting the mixture of experts latent position cluster model. Interest lies in
 226 inferring the latent locations \underline{z}_i of each actor i , the link probability regres-
 227 sion parameters $\underline{\beta}$, the (mixing proportion) cluster membership regression
 228 parameters $\underline{\tau}_g$, the cluster mean $\underline{\mu}_g$ and variance σ_g^2 of each cluster g and the
 229 actor cluster memberships $\underline{K} = (K_1, \dots, K_n)$.

230 3.1. Model specification

231 Prior distributions on the model parameters are required and are speci-
 232 fied as follows:

$$\begin{aligned}
 234 \quad \underline{\beta} &\sim \text{MVN}(\underline{\xi}, \Psi) & \underline{\tau}_g &\sim \text{MVN}(\underline{\gamma}, \Phi) \text{ i.i.d } g = 2, \dots, G. \\
 235 \quad \underline{\mu}_g &\sim \text{MVN}(\underline{0}, \Omega) \text{ i.i.d } g = 1, \dots, G. & \sigma_g^2 &\sim \sigma_0^2 \text{Inv}\chi_\alpha^2 \quad \text{i.i.d } g = 1, \dots, G.
 \end{aligned}$$

236 where $\underline{\xi}, \Psi, \Omega, \sigma_0^2, \alpha, \underline{\gamma}$ and Φ are hyperparameters which need to be specified
 237 by the practitioner. Krivitsky and Handcock (2008) employ a hierarchical
 238 Bayesian model for the latent position cluster model and specify hyperpriors
 239 on the hyperparameters. In this work, as in Handcock et al. (2007), the
 240 hyperparameters are specified to be $\underline{\xi} = \underline{\gamma} = \underline{0}$ and $\Psi = \Phi = \Omega = 2\mathbf{I}$ which
 241 postulates a wide range of values for both the link probability regression pa-
 242 rameters $\underline{\beta}$ and the mixing proportion regression parameters $\underline{\tau}_g$ and provides
 243 a relatively flat prior for the cluster means. For the cluster covariance pa-
 244 rameter σ_g^2 , the same prior as that in Handcock et al. (2007) is used with
 245 $\alpha = 2$ and $\sigma_0^2 = 0.103$.

246 The posterior distribution for the model parameters given the observed
 247 data is formed by combining the likelihood in equations (1) – (5) with the
 248 specified prior distributions. The full conditional posterior distributions of
 249 the model parameters are specified as follows:

250

$$\underline{z}_i | K_i = g, \dots \propto \mathbf{P}\{\mathbf{Y} | \mathbf{Z}, \mathbf{X}, \underline{\beta}\} \text{MVN}(\underline{z}_i | \underline{\mu}_g, \sigma_g^2 \mathbf{I}) \quad \text{for } i = 1, \dots, n \\ \text{and } g = 1, \dots, G$$

$$\underline{\beta} | \dots \propto \mathbf{P}\{\mathbf{Y} | \mathbf{Z}, \mathbf{X}, \underline{\beta}\} \text{MVN}(\underline{\beta} | \underline{\xi}, \Psi)$$

$$\underline{\tau}_g | \dots \propto \mathbf{P}\{\underline{K} = (g, \dots, g) | \underline{\tau}_g, \mathbf{W}\} \text{MVN}(\underline{\tau}_g | \underline{\gamma}, \Phi) \quad \text{for } g = 2, \dots, G$$

251

$$\underline{\mu}_g | \dots \propto \text{MVN}\left(\frac{m_g \bar{z}_g}{m_g + \sigma_g^2 / \omega^2}, \frac{\sigma_g^2}{m_g + \sigma_g^2 / \omega^2} \mathbf{I}\right) \quad \text{for } g = 1, \dots, G$$

$$\sigma_g^2 | \dots \propto (\sigma_0^2 + s_g^2) \text{Inv } \chi_{\alpha + m_g d}^2 \quad \text{for } g = 1, \dots, G$$

$$K_i | \dots \propto \text{Multinomial}\left[1, \left\{ \frac{1}{C} \lambda_1(\underline{w}_i) \text{MVN}(\underline{z}_i | \underline{\mu}_1, \sigma_1^2 \mathbf{I}), \right. \right. \\ \left. \left. \dots, \frac{1}{C} \lambda_G(\underline{w}_i) \text{MVN}(\underline{z}_i | \underline{\mu}_G, \sigma_G^2 \mathbf{I}) \right\} \right] \quad \text{for } i = 1, \dots, n$$

252 where \dots denotes all other parameters, latent locations and cluster member-253 ships in the model and \mathbf{W} is the $n \times p$ matrix of actor specific covariates.

254 Also

$$m_g = \sum_{i=1}^n \mathbf{1}\{K_i = g\} \quad s_g^2 = \sum_{i=1}^n (\underline{z}_i - \underline{\mu}_g)^T (\underline{z}_i - \underline{\mu}_g) \mathbf{1}\{K_i = g\} \\ \bar{z}_g = \frac{1}{m_g} \sum_{i=1}^n \underline{z}_i \mathbf{1}\{K_i = g\} \quad C = \sum_{g'=1}^G \lambda_{g'}(\underline{w}_i) \text{MVN}(\underline{z}_i | \underline{\mu}_{g'}, \sigma_{g'}^2 \mathbf{I}).$$

255

256 The full conditional posterior distributions of some model parameters are
 257 readily available and sampling parameter values in these cases is straightfor-
 258 ward using the Gibbs sampler. However the full conditional posterior distri-
 259 butions of some model parameters are not in a recognizable form and require
 260 a Metropolis-Hastings sampling procedure. A Metropolis-within-Gibbs sam-
 261 pler (eg. O'Hagan and Forster, 2004) is therefore required — this algorithm
 262 embeds a Metropolis step (or steps) within a Gibbs algorithm. Carlin and
 263 Louis (2000) discuss convergence issues associated with such an algorithm.

264 As with any Metropolis-Hastings algorithm a suitable proposal distribu-
 265 tion must be specified from which to draw proposed parameter values. Ide-
 266 ally, such proposal distributions shadow the target distribution from which
 267 sampling is required — the rate at which the algorithm converges to sampling

268 from the target distribution depends crucially on the relationship between
 269 the proposal distribution and the target distribution (Gilks et al., 1996). De-
 270 termining a ‘good’ proposal distribution often relies on craftsmanship and
 271 initial experimentation to determine both the form of a good proposal distri-
 272 bution and its associated parameters. In high-dimensional models, such as
 273 the mixture of experts latent position cluster model, such an experimental
 274 approach involving choosing a suitable distribution and tuning parameters
 275 in the proposal is unpractical and time consuming. Herein, multivariate
 276 Taylor expansions of terms within the target distributions are employed to
 277 aid the construction of *surrogate proposal distributions*. A surrogate pro-
 278 posal distribution is an approximation of a target distribution, derived by
 279 approximating terms within the target distribution via multivariate Taylor
 280 expansions. Similar ideas are employed in Gormley and Murphy (2009b).

281 The idea of surrogate proposal distributions is loosely related to the con-
 282 cepts behind MM algorithms which facilitate optimization of problematic
 283 functions by iteratively optimizing a surrogate function (Lange et al., 2000;
 284 Hunter and Lange, 2004; Hunter, 2004). Additionally, surrogate proposal
 285 distributions and variational methods have a similar ethos in that they both
 286 form an approximation to a problematic posterior distribution, but in the
 287 case of variational methods no sampling is required. Variational approaches
 288 to fitting graphical models are detailed in Wainwright and Jordan (2008);
 289 Salter-Townshend and Murphy (2009) specifically develop a variational ap-
 290 proach to inference for the original latent position cluster model of Handcock
 291 et al. (2007).

292 3.2. *Surrogate proposal distributions*

293 Surrogate proposal distributions are constructed by approximating a term
 294 (or terms) within the target distribution by a quadratic multivariate Taylor
 295 expansion. The quadratic multivariate Taylor expansion of a function $g(\underline{\theta})$
 296 with parameter vector $\underline{\theta}$ about $\bar{\underline{\theta}}$,

$$g(\underline{\theta}) \approx g(\bar{\underline{\theta}}) + (\underline{\theta} - \bar{\underline{\theta}})^T \{\nabla g(\bar{\underline{\theta}})\} + 0.5(\underline{\theta} - \bar{\underline{\theta}})^T \{\nabla^2 g(\bar{\underline{\theta}})\}(\underline{\theta} - \bar{\underline{\theta}}), \quad (6)$$

297 provides a quadratic approximation or *surrogate function* for $g(\underline{\theta})$.

298 The idea of approximating a term within a function to produce a surro-
 299 gate function is employed to construct surrogate proposal distributions which
 300 approximate problematic full conditional distributions from which sampling
 301 is required. When sampling a parameter $\underline{\theta}$ at each iteration of the MCMC

302 algorithm, a surrogate function is formed about a parameter value $\bar{\theta}$ which
 303 in practice is the previously sampled value of the parameter. Thus the sur-
 304 rogate proposal distribution is updated each time a new parameter value is
 305 sampled, maintaining the improved approximation throughout the run time
 306 of the algorithm.

307 Since the choice of both the form and the parameters of the surrogate pro-
 308 posal distribution is automated, model fitting is efficient as the user need not
 309 spend time experimenting with different proposal distributions and tuning
 310 the associated parameters until sufficient mixing is achieved. The Metropolis-
 311 Hastings algorithm no longer involves user input as it is driven by the surro-
 312 gate proposal distribution and its parameters, automatically determined by
 313 the model itself and the observed data.

314 Within the mixture of experts latent position cluster model, Metropolis-
 315 Hastings steps are required to sample the latent locations \underline{z}_i , the link proba-
 316 bility regression parameters $\underline{\beta}$ and the cluster membership regression param-
 317 eters $\underline{\tau}_g$. When sampling the latent locations, a multivariate normal proposal
 318 distribution centered on the current estimate and with a fixed diagonal co-
 319 variance matrix performs sufficiently well; employing such a proposal for the
 320 regression parameters does not perform well, and convergence is crucially de-
 321 pendent on both starting values and the choice of the parameters employed
 322 within the proposal distributions. Thus surrogate proposal distributions are
 323 constructed for the regression parameters only.

324 The mathematical derivation of the surrogate proposal distributions is
 325 given in Appendix A; here it suffices to say that the resulting surrogate
 326 proposal distributions are multivariate normal distributions. The parameters
 327 of these surrogate proposal distributions depend on the current location of
 328 the Markov Chain and are therefore updated at each step of the algorithm.
 329 Thus the construction of proposal distributions is efficient, as the choice of a
 330 distributional form and associated parameter values is automated, reducing
 331 the time required to fit the mixture of experts latent position cluster model.

332 3.3. The Metropolis-within-Gibbs sampler

333 The Metropolis-within-Gibbs sampler proceeds as follows:

- 334 1. Update the latent locations \underline{z}_i for $i = 1, \dots, n$ using a Metropolis-
 335 Hastings step.

336

- 337 (a) Propose $\underline{z}_i^* \sim \text{MVN}(\underline{z}_i, \sigma_g^2 \mathbf{I})$.

(b) If $u \sim U[0, 1]$ is such that

$$u \leq \min \left\{ \frac{\mathbf{P}\{\mathbf{Y}|\mathbf{Z}^*, \mathbf{X}, \underline{\beta}\} \text{ MVN}(\underline{z}_i^* | \underline{\mu}_{K_i}, \sigma_{K_i}^2 \mathbf{I})}{\mathbf{P}\{\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}\} \text{ MVN}(\underline{z}_i | \underline{\mu}_{K_i}, \sigma_{K_i}^2 \mathbf{I})}, 1 \right\}.$$

338 set latent location \underline{z}_i to be \underline{z}_i^* .

339 2. *Update the link probability regression parameters $\underline{\beta}$ using a Metropolis-*
 340 *Hastings step.*

341

342 (a) Propose $\underline{\beta}^* \sim \text{MVN}(\underline{\delta}_{\underline{\beta}}, \Delta_{\underline{\beta}})$ where $\underline{\delta}_{\underline{\beta}}$ and $\Delta_{\underline{\beta}}$ are as defined in
 343 Appendix A, equations (9) and (8) respectively.

(b) If $u \sim U[0, 1]$ is such that

$$u \leq \min \left\{ \frac{\mathbf{P}\{\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}^*\} \text{ MVN}(\underline{\beta}^* | \underline{\xi}, \Psi) \text{ MVN}(\underline{\beta} | \underline{\delta}_{\underline{\beta}^*}, \Delta_{\underline{\beta}^*})}{\mathbf{P}\{\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}\} \text{ MVN}(\underline{\beta} | \underline{\xi}, \Psi) \text{ MVN}(\underline{\beta}^* | \underline{\delta}_{\underline{\beta}}, \Delta_{\underline{\beta}})}, 1 \right\}.$$

344 set the link probability regression parameters $\underline{\beta}$ to be $\underline{\beta}^*$.

345 3. *Update the mixing proportion regression parameters $\underline{\tau}_g$ for $g = 2, \dots, G$*
 346 *using a Metropolis-Hastings step.*

347

348 (a) Propose $\underline{\tau}_g^* \sim \text{MVN}(\underline{\delta}_{\underline{\tau}_g}, \Delta_{\underline{\tau}_g})$ where $\underline{\delta}_{\underline{\tau}_g}$ and $\Delta_{\underline{\tau}_g}$ are as defined in
 349 Appendix A, equations (12) and (11) respectively.

(b) If $u \sim U[0, 1]$ is such that

$$u \leq \min \left\{ \frac{\mathbf{P}\{\underline{K} = (g, \dots, g) | \underline{\tau}_g^*, \mathbf{W}\} \text{ MVN}(\underline{\tau}_g^* | \underline{\gamma}, \Phi) \text{ MVN}(\underline{\tau}_g | \underline{\delta}_{\underline{\tau}^*}, \Delta_{\underline{\tau}^*})}{\mathbf{P}\{\underline{K} = (g, \dots, g) | \underline{\tau}_g, \mathbf{W}\} \text{ MVN}(\underline{\tau}_g | \underline{\gamma}, \Phi) \text{ MVN}(\underline{\tau}_g^* | \underline{\delta}_{\underline{\tau}}, \Delta_{\underline{\tau}})}, 1 \right\}.$$

350 set the mixing proportion regression parameters $\underline{\tau}_g$ to be $\underline{\tau}_g^*$.

351 4. *Update the remaining parameters using Gibbs sampling.*

352 Update \underline{K} , $\underline{\mu}_g$ and σ_g^2 for $g = 1, \dots, G$ from the expressions detailed in
 353 Section 3.1.

354 4. Model issues

355 Prior to examining the approximate posterior distributions output by the
 356 Metropolis-within-Gibbs sampler some model issues must be taken into ac-
 357 count. Modifications need to be made to the samples drawn prior to summa-
 358 rizing their posterior distributions through the calculation of posterior means
 359 or modes. Additionally the issue of model performance must be addressed.

360 *4.1. Model identifiability*

361 The latent locations of actors enter the likelihood function only as a func-
362 tion of their Euclidean distance. The likelihood is therefore invariant to any
363 distance preserving transformation of the configuration of latent locations.
364 If the non-identifiability of the latent locations is not accounted for, poste-
365 rior estimates of the latent locations will be misleading. Additionally, as in
366 any Bayesian mixture modeling context, the likelihood is invariant to switch-
367 ing the cluster labels. This ‘label switching’ phenomenon leads to posterior
368 distributions which are multimodal or symmetric and again summarizing
369 the posterior using posterior means is inappropriate. The output from the
370 Markov Chain Monte Carlo must be processed to ensure final estimates of
371 all parameters are valid.

372 Procrustean methods are used to aid the eradication of likelihood invari-
373 ance due to transformations of the configuration of latent locations. Pro-
374 crustean methods (Krzanowski, 1988) match one configuration of points to
375 another as well as possible in a least squares sense. Transformations such as
376 translation, rotation, and dilation are used to create the match. In the cur-
377 rent context only translation and orthogonal rotations are applicable due to
378 the definition of the likelihood function — a dilation or oblique rotation of the
379 configuration of latent locations would alter the Euclidean distances between
380 actors therefore changing the likelihood of the model. Orthogonal rotations
381 and translations do not influence the Euclidean distances between actors in
382 the latent space and so are permitted. Thus, a reference configuration of la-
383 tent locations is selected and each sampled configuration is translated and/or
384 rotated to match it.

385 In practice, to obtain a reference configuration of locations, the Metropolis-
386 Hastings step to sample latent locations is initially constrained to accept
387 uphill moves only. This will find at least a local optimum on the poste-
388 rior surface which is employed as a rough approximation of the *maximum a*
389 *posteriori* (MAP) configuration of latent locations. This approximate MAP
390 configuration of latent locations is then used as a reference to which all future
391 sampled configurations are matched. Posterior estimates of the actors’ latent
392 locations derived from the post-uphill samples will therefore be consistently
393 matched to the same approximate MAP configuration.

394 A local optimum which is far from the global optimum of the posterior
395 surface could unintentionally be employed as the approximate MAP config-
396 uration. This is not ideal as a reference configuration, but as the chain con-
397 verges the sampled configurations will be drawn from the true full conditional

398 distribution, regardless of the configuration to which they are mapped to aid
 399 model identifiability. The distances between actors in the sampled configura-
 400 tions influence the likelihood of the model, and as they are not altered by the
 401 transformations, the configuration used as a reference is somewhat irrelevant;
 402 the rough MAP approximation is simply used as a reference to identify the
 403 model – it does not influence the validity of the sampled configurations.

404 The issue of label switching in Bayesian mixture models has received
 405 much attention in the literature — Richardson and Green (1997) suggest
 406 minimizing the label switching problem by imposing artificial identifiability
 407 constraints such as ordering the mixing proportions or other model param-
 408 eters. The selection of the parameters on which to base the ordering and
 409 indeed the ordering itself is somewhat ad hoc however. Relabeling strategies
 410 (Celeux et al., 2000; Stephens, 2000) based on a decision theoretic approach
 411 are an alternative way to deal with label switching. The decision theoretic
 412 approach involves defining a loss function $\mathcal{L}(\hat{\theta}|\theta)$ which quantifies the loss in-
 413 curred by choosing $\hat{\theta}$ when the true parameter value is θ . The aim is therefore
 414 to choose $\hat{\theta}$ to minimize the posterior expected loss $\mathbf{E}\{\mathcal{L}(\hat{\theta}|\theta)\}$.

415 The mixture of experts latent position cluster model has several param-
 416 eters, any of which could be used as the basis for dealing with the label
 417 switching problem. It seems intuitive to base the loss function on a param-
 418 eter whose full conditional distribution is specifiable, rather than choosing a
 419 parameter which requires a Metropolis-Hastings update step. Values sam-
 420 pled within a Gibbs sampler are guaranteed to be drawn from the required
 421 target distribution. Values sampled by a Metropolis-Hastings sampler how-
 422 ever are only guaranteed to be drawn from the distribution of interest at
 423 convergence, which itself is difficult to confirm and also may take some time
 424 to achieve. It seems appropriate to base the label switching correction on
 425 parameter values sampled from the required target distribution, rather than
 426 values which are not guaranteed to be sampled from the target distribution.
 427 The full conditional distributions are available for the cluster memberships
 428 \underline{K} , the cluster means $\underline{\mu}_g$ and the cluster variances σ_g^2 . The cluster means
 429 and variances appear to be more informative parameters in this context as
 430 they are estimated from several observations, whereas each cluster member-
 431 ship is only estimated from a single observation. As the cluster variances,
 432 by definition, are constrained to be equal across dimensions, it appears to be
 433 desirable to base the loss function on the cluster means $\underline{\mu}_g$.

434 An approximation to the true value of the cluster means $\underline{\mu}^R = (\underline{\mu}_1^R, \dots, \underline{\mu}_G^R)$

435 is used as a reference. This reference value of the cluster means is chosen to
 436 be an approximation of the *maximum a posteriori* (MAP) estimate obtained
 437 after the uphill only Metropolis-Hastings steps which were used to determine
 438 the reference configuration of latent locations. This approximate MAP value
 439 is used as the reference to which each new estimate $\hat{\boldsymbol{\mu}}$ of the means is matched
 440 to correct for any label switching which may have occurred during sampling.
 441 A sum of squares function is employed as the loss function to be minimized
 442 i.e.

$$\mathcal{L}(\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}^R) = \sum_{g=1}^G \sum_{j=1}^d (\hat{\mu}_{gj} - \mu_{gj}^R)^2.$$

443 Once the MAP estimate has been obtained and subsequent to a typical
 444 burn-in period of the Markov chain, the estimated cluster means $\hat{\underline{\mu}}_g^t$ are per-
 445 muted until the loss function is minimized. An online algorithm (Stephens,
 446 2000) which calculates valid posterior means then proceeds as follows:

- 447 1. Generate all $G!$ permutations ν_l for $l = 1, \dots, G!$. Set $t = 0$.
- 448 2. Subsequent to burn-in of the Markov chain, choose ν_l which minimizes

$$\mathcal{L}(\hat{\boldsymbol{\mu}}_{\nu_l}^t|\boldsymbol{\mu}^R) = \sum_{g=1}^G \sum_{j=1}^d (\hat{\mu}_{\nu_l(g)j}^t - \mu_{gj}^R)^2.$$

- 449 3. Calculate the posterior mean parameter estimates using the computa-
 tionally efficient online formula:

$$\theta = \frac{t}{t+1}\theta + \frac{1}{t+1}\hat{\theta}_{\nu_l}^t$$

450 for $\theta = \underline{\mu}_g, \sigma_g^2, \underline{\tau}_g$ for $g = 1, \dots, G$. Set $K_i = \nu_l(K_i)$ for each actor.

- 451 4. Set $t = t + 1$ and repeat steps 2 and 3 subsequent to each run of the
 452 Metropolis-within-Gibbs sampler.

453 Note that as cluster 1 is used as the baseline cluster the mixing proportion
 454 regression parameters $\underline{\tau}_g$ must be re-scaled subsequent to label switching and
 455 prior to evaluating the posterior mean. This is achieved by simply subtracting
 456 the current $\underline{\tau}_1$ values from $\underline{\tau}_g$ $g = 2, \dots, G$ and then evaluating the posterior
 457 mean.

458 Employing both the Procrustes and label-switching corrections aims to
 459 minimize the influence of identifiability issues on the posterior parameter
 460 estimates.

461 *4.2. Model assessment*

462 Model assessment in the Bayesian framework involves evaluating the pos-
463 terior model probability for all models under consideration and then generally
464 choosing the model with largest posterior probability (Kass and Raftery,
465 1995). Evaluating the posterior model probability poses difficulties in the
466 context of the mixture of experts latent position cluster model. Handcock
467 et al. (2007) employ a version of the Bayesian information criterion (BIC)
468 (Schwarz, 1978) as an approximation to the posterior model probability, con-
469 ditional on the latent locations when performing model selection with the
470 latent position cluster model.

Here the AICM (Akaike’s information criterion - Monte (Carlo)) (Raftery
et al., 2007) is employed as a model assessment tool. The AICM is a sim-
ulation based estimate of the AIC (Akaike’s information criterion) (Akaike,
1973). The AICM is a penalized version of the posterior mean of the loglike-
lihoods and is defined to be

$$\text{AICM} = 2(\bar{l} - s_l^2)$$

471 where \bar{l} and s_l^2 are the mean and variance respectively of the log likelihoods
472 of a posterior sample. Clearly the AICM is easily computed once a posterior
473 sample has been generated. The AICM was successfully employed in Raftery
474 et al. (2007) and in Gormley and Murphy (2007b) to select the dimensionality
475 of latent spaces.

476 Even though AICM is easily computed, model selection in the mixture of
477 experts latent position cluster model is difficult due to the large model space
478 induced by considering a range of dimensions of the latent space, a range of
479 numbers of clusters, a range of covariates and the range of possible ways in
480 which the covariates may enter the model. Strategic searching of a reduced
481 model space is therefore necessary.

482 **5. Applying the mixture of experts latent position cluster model**

483 Data on the relations between 71 lawyers in a northeastern American
484 law firm were collected in Lazega (2001). A number of network data sets
485 were recorded including a friendship network and a co-workers network. The
486 friendship network was constructed by asking the lawyers to list the names of
487 those lawyers in the firm that they socialize with outside of work. For the co-
488 workers network, lawyers were asked to list all the lawyers in the firm with

489 whom they had worked during the previous year. Additionally covariates
 490 associated with the lawyers were recorded; these are detailed in Table 1.

Table 1: The seven covariates and their respective levels (if categorical variables) recorded on 71 lawyers in an northeastern American law firm. The last category in each categorical covariate is treated as the baseline category in all future analyzes.

Covariate	Levels
Seniority	1 = partner 2 = associate
Gender	1 = male 2 = female
Office	1 = Boston 2 = Hartford 3 = Providence
Practice	1 = litigation 2 = corporate
Law school	1 = Harvard or Yale 2 = University of Connecticut 3 = other
Years with the firm	—
Age	—

491 Interest lies in examining the underlying social structure and processes
 492 within the law firm from both the recorded network data sets and the asso-
 493 ciated covariates. The data has previously been examined in Snijders et al.
 494 (2006) and is available as part of the SIENA software (Snijders et al., 2005).

495 5.1. Exploring the model space

496 The number of possible mixture of experts latent position cluster models
 497 is very large. In addition to choosing the number of clusters and the dimen-
 498 sion of the latent space, the covariates can enter the model in several ways,
 499 leading to a large model space.

500 Initially, the latent position cluster model was fitted for various numbers
 501 of clusters and latent space dimensions. Selection of the optimal number of
 502 clusters and dimension of the latent space was performed using the methods
 503 employed by Handcock et al. (2007) — the posterior model probability of each

504 of the competing models was considered and that with the largest posterior
505 probability selected. In practice, each model (i.e. each specific combination
506 of G and d) was assigned equal prior probability and conditional (on the
507 estimated latent locations) posterior model probabilities were approximated
508 by the Bayesian Information Criterion (Schwarz, 1978). The model with
509 largest BIC was selected as optimal. Once an optimal choice of model was
510 found within the class of models considered, yielding a value for G and d ,
511 the selection of covariates was addressed.

512 To select a candidate set of covariates for the link probabilities (2), a
513 logistic regression model of the same form was fitted with the z_i values fixed
514 to be their posterior mean values from the latent position cluster model
515 analysis. For categorical covariates, a variable indicating that the values of
516 the actor covariates are equal was used as the dyadic specific edge covariate
517 (ie. $x_{i,j,k} = 1$ if $w_{ik} = w_{jk}$). For continuous covariates, the absolute difference
518 and squared differences in covariate values were included as dyadic specific
519 covariates. A backwards stepwise selection procedure was used, with model
520 selection being determined using AIC (Akaike, 1973). This yielded a subset
521 of the candidate variables for further analysis in the mixture of experts latent
522 position cluster model.

523 In order to find a candidate set of covariates for the cluster membership
524 model (3), a multinomial logistic regression model was fitted with the modal
525 cluster membership values from the latent cluster position model analysis
526 used as the response. A backwards stepwise variable selection procedure was
527 used to select a candidate set of covariates for further analysis in the mixture
528 of experts latent position cluster model.

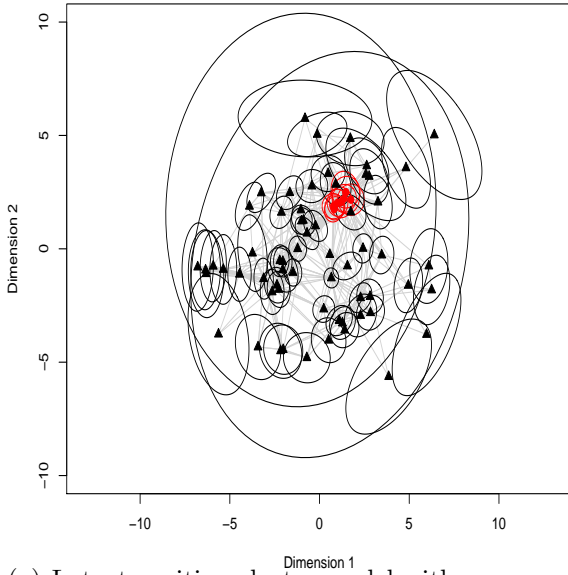
529 Four versions of the mixture of experts latent position cluster model were
530 fitted for the choices of G , d and covariate selections by considering whether
531 (i) the covariates should enter the link probabilities, (ii) the mixing propor-
532 tions, (iii) both or (iv) neither. These models were assessed using AICM and
533 the posterior summaries for the model parameters.

534 5.2. *Lawyers friendship network*

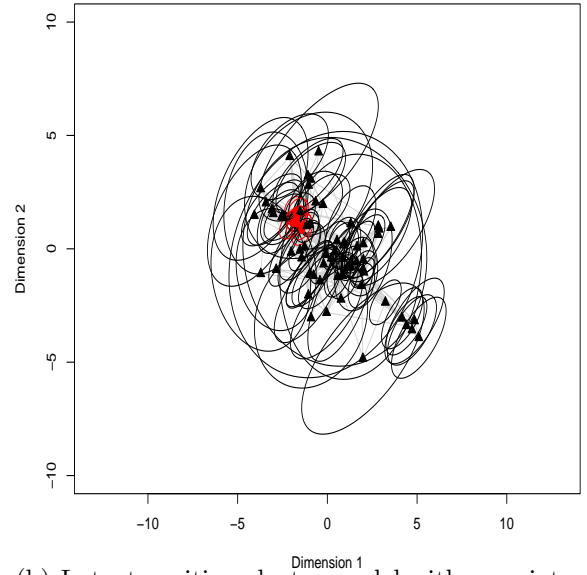
535 Initially, a latent position cluster model was fitted and a model with $G = 2$
536 and $d = 2$ was deemed optimal. A number of candidate covariates were then
537 selected using the procedure detailed in Section 5.1. The covariates in the
538 cluster membership model were their office and years working in the firm.
539 The link probability covariates proposed were seniority, gender, office, prac-
540 tice, school and the difference in years worked in the firm. The four cases of

541 the mixture of experts model were fitted to the data; the estimated latent
542 positions and their uncertainties are shown in Figure 2. The resulting param-
543 eter estimates and AICM values are reported in Table 2. The AICM criterion
544 suggests that covariates enter the model through the link probabilities only,
545 however we will explore and compare all of the models.

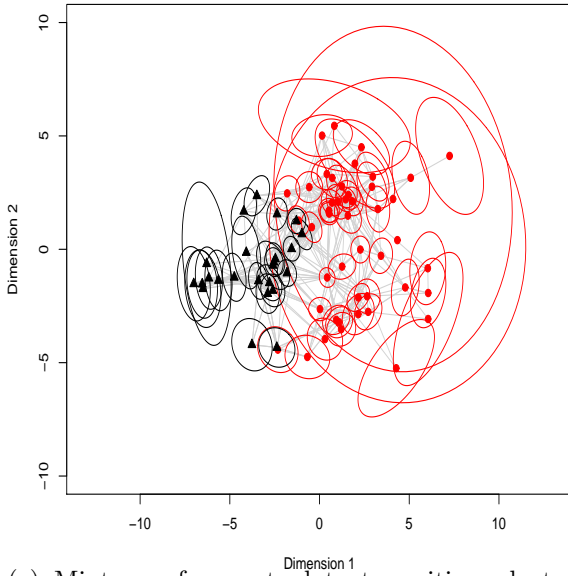
546 The latent positions of the actors and the cluster memberships vary con-
547 siderably when comparing the models that do not have covariates that enter
548 the link probabilities to those that do (Figures 2(a) and 2(c) versus Figures
549 2(b) and 2(d)). This is due to the fact that the latent positions of the actors
550 account for any remaining network structure which has not been explained
551 by the covariates in the link probabilities. So, the latent positions in the
552 models without covariates in the link probabilities are explaining more of the
553 network structure than in the other models. The clustering is quite different
554 in the four models and the clustering structure is clearer in the models where
555 the covariates enter the cluster membership model; this can be explained by
556 the different role that the latent locations play in the two models.



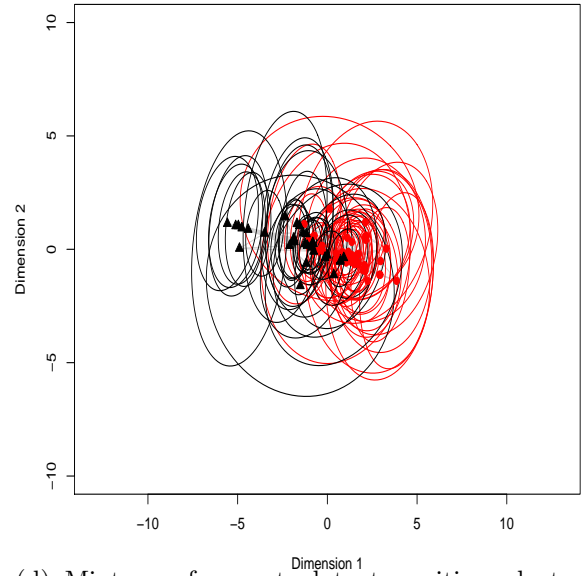
(a) Latent position cluster model with no covariates.



(b) Latent position cluster model with covariates in the link probabilities.



(c) Mixture of experts latent position cluster model with covariates in the cluster memberships.



(d) Mixture of experts latent position cluster model with covariates in both the cluster memberships and the link probabilities.

Figure 2: Estimates of clusters and latent positions of the lawyers from the friendship network data. The ellipses are 50% posterior sets illustrating the uncertainty in the latent locations. Lawyers who are members of the same cluster are illustrated using the same colour and symbol. Observed links between lawyers are also illustrated.

557 The coefficients in the link probability function (Table 2) are consistent
558 across model types. The coefficients suggest that lawyers with similar co-
559 variates are more likely to have relations, thus demonstrating homophily by
560 attributes in the lawyer friendship data.

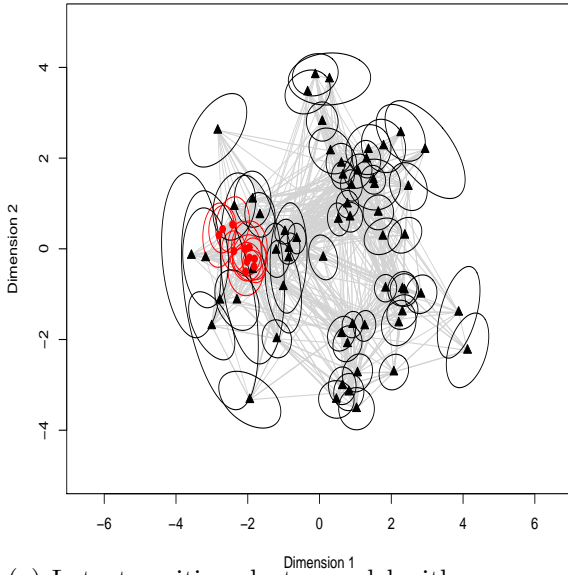
561 There is a considerable (Boston) office effect in determining the cluster
562 memberships in model (c), but this does not persist in model (d). The num-
563 ber of years that the lawyers have been in the firm is influential in determining
564 cluster membership in model (c) but again this does not persist in (d). The
565 smaller cluster in model (c) appears to be a group of lawyers in the Boston
566 office who have only spent a few years in the firm. Thus the models have
567 given us insight to the characteristics which cause the lawyers to be friends,
568 and have provided an view of the clustering structure within the firm.

569 5.3. *Lawyers co-workers network*

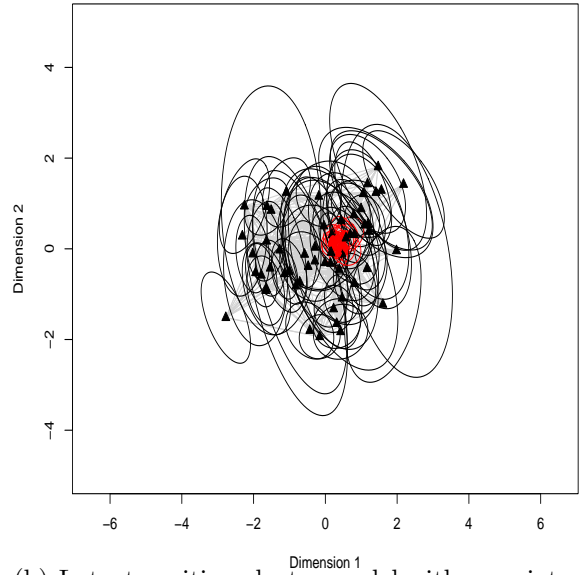
570 For the co-workers network the initial selection of covariates detailed in
571 Section 5.1 suggested that a lawyer’s seniority, their office, their years with
572 the firm and their type of practice were important covariates in the cluster
573 membership model. For the dyadic covariates in the link probabilities senior-
574 ity, gender, office, (absolute difference in) age and their practice types were
575 deemed influential. Hence, four mixture of experts latent position cluster
576 models with $G = 2$, $d = 2$ (as selected by fitting a latent position cluster
577 model and choosing the optimal model) and these selected covariates were
578 fitted to the co-workers network data set. The posterior mean estimates of
579 the latent locations and the clusters of lawyers are illustrated in Figure 3,
580 for each of the four mixture of experts latent position cluster models. The
581 resulting parameter estimates and AICM values for each model are detailed
582 in Table 3.

Table 2: Posterior mean parameter estimates for the four mixture of experts models fitted to the lawyers friendship data as detailed in Figure 2. Standard deviations are given in parentheses. Note that cluster 1 was used as the baseline cluster in the case of the cluster membership parameters. Baseline categories for the covariates are detailed in Table 1.

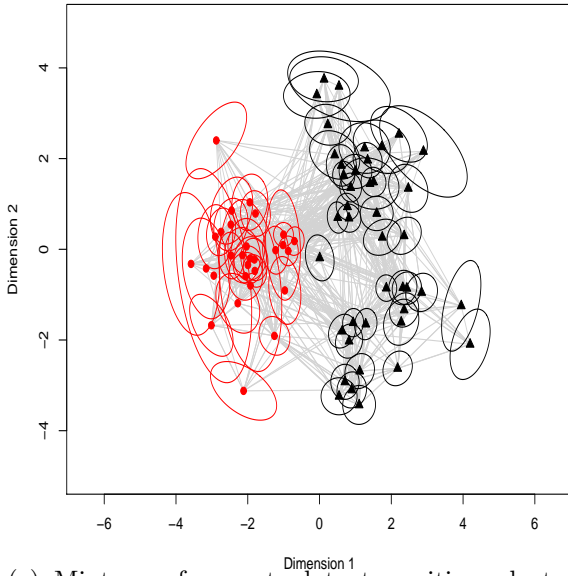
	(a)	(b)	(c)	(d)
Link Probabilities				
Intercept	1.71 (0.14)	-2.20 (0.23)	1.73 (0.13)	-2.18 (0.26)
Seniority		1.15 (0.19)		1.07 (0.20)
Gender		0.66 (0.12)		0.65 (0.12)
Office		2.38 (0.14)		2.41 (0.13)
Practice		0.78 (0.10)		0.78 (0.11)
School		0.37 (0.10)		0.35 (0.11)
Years		-0.01 (0.00)		-0.01 (0.00)
Cluster Memberships				
Intercept	1.63 (1.40)	1.43 (1.67)	1.13 (1.02)	2.11 (1.59)
Office (=1)			2.93 (1.06)	1.00 (1.21)
Office (=2)			-1.18 (1.17)	0.70 (1.48)
Years			-0.77 (0.27)	-0.57 (0.37)
Latent Space Model				
Group 1 Mean	-0.24 (0.60)	0.33 (0.64)	-3.40 (0.53)	-1.71 (0.89)
	-0.21 (0.65)	-0.27 (0.54)	-0.85 (0.66)	0.09 (1.34)
Group 1 Variance	10.59 (3.74)	6.66 (2.49)	3.95 (1.17)	5.03 (1.35)
Group 2 Mean	1.06 (0.53)	-1.35 (0.76)	1.72 (0.47)	1.36 (0.66)
	1.86 (0.79)	0.98 (0.63)	0.44 (0.50)	-0.03 (1.20)
Group 2 Variance	1.40 (3.89)	1.44 (2.53)	8.69 (1.66)	4.17 (1.06)
AICM	-2359	-2325	-2341	-2390



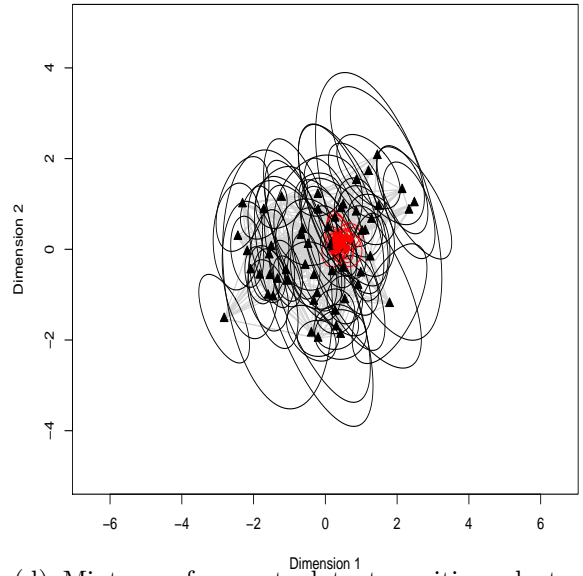
(a) Latent position cluster model with no covariates.



(b) Latent position cluster model with covariates in the link probabilities.



(c) Mixture of experts latent position cluster model with covariates in the cluster memberships.



(d) Mixture of experts latent position cluster model with covariates in both the cluster memberships and the link probabilities.

Figure 3: Estimates of clusters and latent positions of the lawyers from the co-workers network data. The ellipses are 50% posterior sets illustrating the uncertainty in the latent locations. Lawyers who are members of the same cluster are illustrated using the same colour and symbol. Observed links between lawyers are also illustrated.

583 In the case of the co-workers network data, AICM suggests the mixture
584 of experts latent position cluster model is optimal, where covariates enter the
585 cluster membership probabilities. All models will be discussed however since,
586 as indicated by Figure 3, all provide interesting insights to both the network
587 of lawyers, and to the mixture of experts latent position cluster model itself.

588 As in the case of the friendship network, the posterior estimates of the
589 latent locations are relatively consistent within models with covariates in the
590 link probabilities (Figures 3(b) and 3(d)) and within models without covari-
591 ates in the link probabilities (Figures 3(a) and 3(c)). Again it is clear that
592 the cluster memberships do differ between the models. The two clusters are
593 particularly distinct in the latent space (see Figure 3(c)) where the covariates
594 enter the cluster memberships; this appears to be intuitive in that the co-
595 variates of a lawyer influence their cluster membership, and then given their
596 cluster membership, the lawyer forms relations.

597 The posterior means detailed in Table 3 also provide insight to the net-
598 work structure and to the mixture of experts latent position cluster model.
599 Of note are the consistent link probability regression parameter estimates.
600 The link probability regression parameter estimates also have intuitive in-
601 terpretations; for example, the negative seniority coefficient suggests lawyers
602 in the firm appear less likely to work together if they have the same level of
603 seniority. This is conceivable as it is likely that a partner and an associate
604 would work together on a case, rather than two partners working together
605 on a case.

606 Also of note is that the practice covariate appears to be influential in
607 the co-workers network cluster membership coefficients, but office does not
608 appear to be, as it did in the case of the friendship network. As expected the
609 practice effect has a positive influence on link probability, but it also has a
610 strong influence on cluster membership in model (c) where covariates enter
611 the cluster memberships. The smaller cluster in model (c) is a cluster of liti-
612 gation lawyers. Thus the models have given us insight to the characteristics
613 which cause the lawyers to be co-workers, and have provided an view of
614 the clustering structure within the firm.

615 **6. Discussion**

616 The mixture of experts latent position cluster model offers an important
617 extension to the latent position cluster model of Handcock et al. (2007).

Table 3: Posterior mean parameter estimates for the four mixture of experts models fitted to the lawyers co-worker data as detailed in Figure 3. Standard deviations are given in parentheses. Note that cluster 1 was used as the baseline cluster in the case of the cluster membership parameters. Baseline categories for the covariates are detailed in Table 1.

	(a)	(b)	(c)	(d)
Link Probabilities				
Intercept	1.6 (0.1)	-1.46 (0.16)	1.56 (0.09)	-1.42 (0.16)
Seniority		-0.45 (0.09)		-0.49 (0.09)
Gender		0.40 (0.10)		0.41 (0.10)
Office		1.91 (0.10)		1.91 (0.10)
Practice		1.64 (0.10)		1.64 (0.09)
Age		-0.01 (0.00)		-0.00 (0.00)
Cluster Memberships				
Intercept	-1.2 (1.7)	1.41 (0.66)	-2.07 (1.17)	1.29 (0.88)
Seniority			-0.47 (0.97)	-3.20 (1.33)
Office (=1)			0.52 (0.94)	1.49 (0.87)
Office (=2)			-2.34 (1.50)	0.17 (0.84)
Practice			3.30 (0.78)	0.72 (0.59)
Years			-0.04 (0.08)	0.06 (0.05)
Latent Space Model				
Cluster 1 mean	0.47 (0.48)	-0.09 (0.19)	1.19 (0.31)	-0.11 (0.20)
	0.21 (0.69)	-0.03 (0.21)	0.26 (0.52)	-0.03 (0.20)
Cluster 1 variance	3.14 (1.45)	1.98 (0.45)	2.93 (0.76)	2.15 (0.46)
Cluster 2 mean	-1.65 (0.92)	0.40 (0.21)	-1.88 (0.48)	0.40 (0.13)
	-0.12 (0.66)	0.12 (0.21)	-0.25 (0.28)	0.13 (0.18)
Cluster 2 variance	0.92 (1.40)	0.09 (0.32)	1.29 (0.73)	0.07 (0.28)
AICM	-4074	-4095	-4066	-4117

618 The model has the flexibility to allow actor covariates to influence network
619 structure in a number of ways, yielding a rich family of network models.

620 A major issue with this model is searching the potential models, so that
621 an optimal model can be found. An efficient procedure is proposed in this
622 work, but it does not guarantee finding an optimal model. In addition to
623 this, the choice of model needs to be guided using a criterion like AICM but
624 also the interpretation of the model outputs. The manner in which covariates
625 enter the mixture of experts model has a major impact on the interpretation
626 of the latent positions of the actors. If the aim of the analysis is to capture
627 much of the network structure using the latent positions, then the covariates
628 should enter the cluster membership model. However, if the latent positions
629 are being used to find network structure beyond that which can be explained
630 by the covariates, then it is important that the covariates enter the link
631 probabilities.

632 The latent position cluster model in its original and mixture of experts
633 extended form can accommodate alternative link types. In order to do this,
634 the form of (2) would need to be changed to match the link type. This
635 shows that this modeling framework has potential for applications beyond
636 the analysis of binary link data.

637 The method of fitting the model using Markov Chain Monte Carlo is
638 computationally expensive, but is much improved by employing surrogate
639 proposal distributions. However, as the number of actors increases compu-
640 tational issues pose problems. There is considerable scope for finding more
641 efficient model fitting procedures for these models and the development of
642 these would greatly extend the scope of the latent position cluster method-
643 ology to the analysis of larger networks.

644 **Acknowledgements**

645 We would like to thank the reviewers who suggested many changes that
646 have greatly improved this work. We would like to acknowledge Prof. Adrian
647 Raftery for his discussions on this research. This work is partially sup-
648 ported by Science Foundation Ireland Research Frontiers Programme grant
649 (06/RFP/M040) and Science Foundation Ireland Strategic Research Cluster
650 grant (08/SRC/I1407).

651 **A. Derivation of surrogate proposal distributions**

652 As detailed in Section 3.3, surrogate proposal distributions are employed
 653 in the Metropolis-within-Gibbs sampler when sampling the regression pa-
 654 rameters $\underline{\beta}$ and $\underline{\tau}_g$ for $g = 2, \dots, G$. Such proposal distributions are de-
 655 rived by approximating terms within the full conditional distributions. A
 656 quadratic approximation to terms within the full conditional distributions
 657 are employed.

658 For clarity in the following derivations, the covariate vector associated
 659 with actor i is now defined to be the $p + 1$ vector $\underline{w}_i^T = (1, w_{i1}, \dots, w_{ip})$.
 660 Similarly, the dyadic specific covariate vector is defined to be the $p + 1$ vector
 661 $\underline{x}_{i,j}^T = (1, x_{i,j,1}, \dots, x_{i,j,p})$.

662 *A.1. Derivation of the surrogate proposal distribution for the link probability*
 663 *regression parameters $\underline{\beta}$.*

664 The full conditional distribution of the link probability regression param-
 665 eters $\underline{\beta}$ is

$$\underline{\beta} | \dots \propto \mathbf{P}\{\mathbf{Y} | \mathbf{Z}, \mathbf{X}, \underline{\beta}\} \text{MVN}(\underline{\beta} | \underline{\xi}, \Psi). \quad (7)$$

666 Let $d_{ij} = \|\underline{z}_i - \underline{z}_j\|$. Then

$$\begin{aligned} \mathbf{P}\{\mathbf{Y} | \mathbf{Z}, \mathbf{X}, \underline{\beta}\} &= \prod_{i=1}^n \prod_{i \neq j} \left\{ \frac{\exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})}{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})} \right\}^{y_{i,j}} \left\{ \frac{1}{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})} \right\}^{1-y_{i,j}} \\ \log \mathbf{P}\{\mathbf{Y} | \mathbf{Z}, \mathbf{X}, \underline{\beta}\} &= \sum_{i=1}^n \sum_{j \neq i} y_{i,j} (\underline{\beta}^T \underline{x}_{i,j} - d_{ij}) - \sum_{i=1}^n \sum_{j \neq i} \log \{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})\}. \end{aligned}$$

667 Examination of the form of the (log) likelihood combined with the form of
 668 the multivariate normal prior does not clearly suggest the distributional form
 669 of a good proposal distribution, or its associated parameters for $\underline{\beta}$.

Within the log likelihood, we can approximate the term

$$g(\underline{\beta}) = - \sum_{i=1}^n \sum_{j \neq i} \log \{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})\}$$

670 via a quadratic multivariate Taylor expansion (6) in $\underline{\beta}$ about $\bar{\underline{\beta}}$:

$$g(\underline{\beta}) \approx g(\bar{\underline{\beta}}) + (\underline{\beta} - \bar{\underline{\beta}})^T \{\nabla g(\bar{\underline{\beta}})\} + 0.5(\underline{\beta} - \bar{\underline{\beta}})^T \{\nabla^2 g(\bar{\underline{\beta}})\}(\underline{\beta} - \bar{\underline{\beta}}).$$

671 It follows that

$$\nabla g(\underline{\beta}) = - \sum_{i=1}^n \sum_{j \neq i} \frac{\underline{x}_{i,j} \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})}{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})}$$

672 and

$$\nabla^2 g(\underline{\beta}) = - \sum_{i=1}^n \sum_{j \neq i} \underline{x}_{i,j} \frac{\exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})}{[1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})]^2} \underline{x}_{i,j}^T$$

673 Substituting the quadratic approximation of $g(\underline{\beta})$ into the log likelihood gives

$$\begin{aligned} \log \mathbf{P}\{\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}\} &\approx \sum_{i=1}^n \sum_{j \neq i} y_{i,j} (\underline{\beta}^T \underline{x}_{i,j} - d_{ij}) - \sum_{i=1}^n \sum_{j \neq i} \log\{1 + \exp(\underline{\beta}^T \underline{x}_{i,j} - d_{ij})\} \\ &\quad - (\underline{\beta} - \bar{\underline{\beta}})^T \left[\sum_{i=1}^n \sum_{j \neq i} \frac{\underline{x}_{i,j} \exp(\bar{\underline{\beta}}^T \underline{x}_{i,j} - d_{ij})}{1 + \exp(\bar{\underline{\beta}}^T \underline{x}_{i,j} - d_{ij})} \right] \\ &\quad - 0.5(\underline{\beta} - \bar{\underline{\beta}})^T \left[\sum_{i=1}^n \sum_{j \neq i} \underline{x}_{i,j} \frac{\exp(\bar{\underline{\beta}}^T \underline{x}_{i,j} - d_{ij})}{[1 + \exp(\bar{\underline{\beta}}^T \underline{x}_{i,j} - d_{ij})]^2} \underline{x}_{i,j}^T \right] (\underline{\beta} - \bar{\underline{\beta}}) \end{aligned}$$

674 when in practice $\bar{\underline{\beta}}$ is the previously sampled value of $\underline{\beta}$. Ignoring terms con-
 675 stant in $\underline{\beta}$ and introducing the term $-\sum_{i=1}^n \sum_{j \neq i} y_{i,j} \bar{\underline{\beta}}^T \underline{x}_{i,j}$ to facilitate algebraic
 676 completeness gives

$$\begin{aligned} \log \mathbf{P}\{\mathbf{Y}|\mathbf{Z}, \mathbf{X}, \underline{\beta}\} &\approx \sum_{i=1}^n \sum_{j \neq i} y_{i,j} (\underline{\beta} - \bar{\underline{\beta}})^T \underline{x}_{i,j} \\ &\quad - (\underline{\beta} - \bar{\underline{\beta}})^T \left[\sum_{i=1}^n \sum_{j \neq i} \frac{\underline{x}_{i,j} \exp(\bar{\underline{\beta}}^T \underline{x}_{i,j} - d_{ij})}{1 + \exp(\bar{\underline{\beta}}^T \underline{x}_{i,j} - d_{ij})} \right] \\ &\quad - 0.5(\underline{\beta} - \bar{\underline{\beta}})^T \left[\sum_{i=1}^n \sum_{j \neq i} \underline{x}_{i,j} \frac{\exp(\bar{\underline{\beta}}^T \underline{x}_{i,j} - d_{ij})}{[1 + \exp(\bar{\underline{\beta}}^T \underline{x}_{i,j} - d_{ij})]^2} \underline{x}_{i,j}^T \right] (\underline{\beta} - \bar{\underline{\beta}}) \end{aligned}$$

which is a quadratic function in $(\underline{\beta} - \bar{\underline{\beta}})$. The quadratic form of the approximated log likelihood suggests a multivariate normal distribution with

covariance matrix

$$\Sigma_L = \left[\sum_{i=1}^n \sum_{j \neq i} \underline{\mathbf{x}}_{i,j} \frac{\exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})}{[1 + \exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})]^2} \underline{\mathbf{x}}_{i,j}^T \right]^{-1}$$

and mean

$$\underline{\mu}_L = \Sigma_L \left[\sum_{i=1}^n \sum_{j \neq i} \underline{\mathbf{x}}_{i,j} \left\{ y_{i,j} - \frac{\exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})}{1 + \exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})} \right\} \right].$$

677 The prior on $\underline{\beta}$ is $\text{MVN}(\underline{\xi}, \Psi)$, meaning that $(\underline{\beta} - \underline{\bar{\beta}}) \sim \text{MVN}(\underline{\xi} - \underline{\bar{\beta}}, \Psi)$.
 678 Hence the sum of the approximated log likelihood and the log of the prior
 679 distribution on $(\underline{\beta} - \underline{\bar{\beta}})$ is the sum of (the log of) two multivariate normal
 680 distributions, suggesting that the distribution of $(\underline{\beta} - \underline{\bar{\beta}})$ is approximately
 681 $\text{MVN}(\underline{\delta}, \Delta_{\underline{\bar{\beta}}})$ where

$$\Delta_{\underline{\bar{\beta}}} = \left[\left\{ \sum_{i=1}^n \sum_{j \neq i} \underline{\mathbf{x}}_{i,j} \frac{\exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})}{[1 + \exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})]^2} \underline{\mathbf{x}}_{i,j}^T \right\} + \Psi^{-1} \right]^{-1} \quad (8)$$

and

$$\underline{\delta} = \Delta_{\underline{\bar{\beta}}} \left[\sum_{i=1}^n \sum_{j \neq i} \underline{\mathbf{x}}_{i,j} \left\{ y_{i,j} - \frac{\exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})}{1 + \exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})} \right\} + \Psi^{-1}(\underline{\xi} - \underline{\bar{\beta}}) \right].$$

682 Thus the full conditional distribution of $\underline{\beta}$ (7) can be approximated by a
 683 multivariate normal distribution with covariance matrix $\Delta_{\underline{\bar{\beta}}}$ and mean

$$\underline{\delta}_{\underline{\beta}} = \underline{\bar{\beta}} + \Delta_{\underline{\bar{\beta}}} \left[\sum_{i=1}^n \sum_{j \neq i} \underline{\mathbf{x}}_{i,j} \left\{ y_{i,j} - \frac{\exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})}{1 + \exp(\underline{\bar{\beta}}^T \underline{\mathbf{x}}_{i,j} - d_{ij})} \right\} + \Psi^{-1}(\underline{\xi} - \underline{\bar{\beta}}) \right] \quad (9)$$

684 A multivariate normal distribution with mean $\underline{\delta}_{\underline{\beta}}$ and covariance matrix $\Delta_{\underline{\bar{\beta}}}$
 685 is therefore employed as a surrogate proposal distribution in the Metropolis-
 686 Hastings step to sample $\underline{\beta}$ of the Metropolis-within-Gibbs sampler.

687 *A.2. Derivation of the surrogate proposal distribution for the cluster mem-*
 688 *bership regression parameters $\underline{\tau}_g$ for $g = 2, \dots, G$.*

689 The full conditional distribution of the mixing proportions regression pa-
 690 rameters $\underline{\tau}_g$ is

$$\underline{\tau}_g | \dots \propto \mathbf{P}\{\underline{K} = (g, \dots, g) | \underline{\tau}_g, \mathbf{W}\} \text{MVN}(\underline{\tau}_g | \underline{\gamma}, \Phi). \quad (10)$$

691 Then

$$\log \mathbf{P}\{\underline{K} = (g, \dots, g) | \underline{\tau}_g, \mathbf{W}\} = \sum_{i=1}^n \mathbf{1}\{K_i = g\} (\underline{\tau}_g^T \underline{w}_i) - \sum_{i=1}^n \log \left(\sum_{g'=1}^G \exp(\underline{\tau}_{g'}^T \underline{w}_i) \right)$$

692 Examination of the form of the (log) likelihood combined with the form of
 693 the multivariate normal prior does not clearly suggest the distributional form
 694 of a good proposal distribution, or its associated parameters for $\underline{\tau}_g$.

Within the log likelihood, we can approximate the term

$$g(\underline{\tau}_g) = - \sum_{i=1}^n \log \left(\sum_{g'=1}^G \exp(\underline{\tau}_{g'}^T \underline{w}_i) \right)$$

695 via a quadratic multivariate Taylor expansion (6) in $\underline{\tau}_g$ about $\bar{\underline{\tau}}_g$:

$$g(\underline{\tau}_g) \approx g(\bar{\underline{\tau}}_g) + (\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \{\nabla g(\bar{\underline{\tau}}_g)\} + 0.5(\underline{\tau}_g - \bar{\underline{\tau}}_g)^T \{\nabla^2 g(\bar{\underline{\tau}}_g)\} (\underline{\tau}_g - \bar{\underline{\tau}}_g).$$

696 It follows that

$$\nabla g(\underline{\tau}_g) = - \sum_{i=1}^n \frac{\underline{w}_i \exp(\underline{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\underline{\tau}_{g'}^T \underline{w}_i)}$$

697 and

$$\nabla^2 g(\underline{\tau}_g) = - \sum_{i=1}^n \underline{w}_i \frac{\exp(\underline{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\underline{\tau}_{g'}^T \underline{w}_i)} \left[1 - \frac{\exp(\underline{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\underline{\tau}_{g'}^T \underline{w}_i)} \right] \underline{w}_i^T$$

698 Substituting the quadratic approximation of $g(\underline{\tau}_g)$ into the log likelihood
 699 gives that $\log \mathbf{P}\{\underline{K} = (g, \dots, g) | \underline{\tau}_g, \mathbf{W}\}$ is

$$\approx \sum_{i=1}^n \mathbf{1}\{K_i = g\} (\underline{\tau}_g^T \underline{w}_i) - \sum_{i=1}^n \log \left(\sum_{g'=1}^G \exp(\bar{\underline{\tau}}_{g'}^T \underline{w}_i) \right)$$

$$\begin{aligned}
& -(\underline{\tau}_g - \bar{\tau}_g)^T \left\{ \sum_{i=1}^n \frac{\underline{w}_i \exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right\} \\
& -0.5(\underline{\tau}_g - \bar{\tau}_g)^T \left\{ \sum_{i=1}^n \underline{w}_i \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \left[1 - \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right] \underline{w}_i^T \right\} (\underline{\tau} - \bar{\tau}_g)
\end{aligned}$$

700 when in practice $\bar{\tau}_g$ is the previously sampled value of $\underline{\tau}_g$. Ignoring terms
701 constant in $\underline{\tau}_g$ and introducing the term $-\sum_{i=1}^n \mathbf{1}\{K_i = g\}(\bar{\tau}_g^T \underline{w}_i)$ to facilitate
702 algebraic completeness gives that $\log \mathbf{P}\{\underline{K} = (g, \dots, g) | \underline{\tau}_g, \mathbf{W}\}$ is

$$\begin{aligned}
& \approx \sum_{i=1}^n \mathbf{1}\{K_i = g\} (\underline{\tau}_g - \bar{\tau}_g)^T \underline{w}_i \\
& -(\underline{\tau}_g - \bar{\tau}_g)^T \left\{ \sum_{i=1}^n \frac{\underline{w}_i \exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right\} \\
& -0.5(\underline{\tau}_g - \bar{\tau}_g)^T \left\{ \sum_{i=1}^n \underline{w}_i \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \left[1 - \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right] \underline{w}_i^T \right\} (\underline{\tau} - \bar{\tau}_g)
\end{aligned}$$

which is a quadratic function in $(\underline{\tau}_g - \bar{\tau}_g)$. The quadratic form of the approximated log likelihood suggests a multivariate normal distribution with covariance matrix

$$\Sigma_L = \left[\sum_{i=1}^n \underline{w}_i \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \left[1 - \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right] \underline{w}_i^T \right]^{-1}$$

and mean

$$\mu_L = \Sigma_L \left[\sum_{i=1}^n \underline{w}_i \left\{ \mathbf{1}\{K_i = g\} - \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right\} \right].$$

703 The prior on $\underline{\tau}_g$ is $\text{MVN}(\underline{\gamma}, \Phi)$, meaning $(\underline{\tau}_g - \bar{\tau}_g) \sim \text{MVN}(\underline{\gamma} - \bar{\tau}_g, \Phi)$.
704 Hence the sum of the approximated log likelihood and the log of the prior
705 distribution on $(\underline{\tau}_g - \bar{\tau}_g)$ is the sum of (the log of) two multivariate normal
706 distributions, suggesting that the distribution of $(\underline{\tau}_g - \bar{\tau}_g)$ is also approxi-
707 mately $\text{MVN}(\underline{\delta}, \Delta_{\bar{\tau}_g})$ where

$$\Delta_{\bar{\tau}_g} = \left[\sum_{i=1}^n \left\{ \underline{w}_i \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \left(1 - \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right) \underline{w}_i^T \right\} + \Phi^{-1} \right]^{-1} \quad (11)$$

708 and

$$\underline{\delta} = \Delta_{\bar{\tau}_g} \left[\sum_{i=1}^n \underline{w}_i \left\{ \mathbf{1}\{K_i = g\} - \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right\} + \Phi^{-1}(\underline{\gamma} - \bar{\tau}_g) \right].$$

709 Thus the full conditional distribution of $\underline{\tau}_g$ (10) can be approximated by a
710 multivariate normal distribution with covariance matrix $\Delta_{\bar{\tau}_g}$ and mean

$$\underline{\delta}_{\bar{\tau}_g} = \bar{\tau}_g + \Delta_{\bar{\tau}_g} \left[\sum_{i=1}^n \underline{w}_i \left\{ \mathbf{1}\{K_i = g\} - \frac{\exp(\bar{\tau}_g^T \underline{w}_i)}{\sum_{g'=1}^G \exp(\bar{\tau}_{g'}^T \underline{w}_i)} \right\} + \Phi^{-1}(\underline{\gamma} - \bar{\tau}_g) \right] \quad (12)$$

711 Thus a multivariate normal distribution with mean $\underline{\delta}_{\bar{\tau}_g}$ and covariance ma-
712 trix $\Delta_{\bar{\tau}_g}$ is employed as a surrogate proposal distribution in the Metropolis-
713 Hastings step of the Metropolis-within-Gibbs sampler to sample $\underline{\tau}_g$.

714 References

- 715 Akaike, H., Information theory and an extension to the maximum likeli-
716 hood principle. Second International Symposium on on Information The-
717 ory. (1973) 267–281
- 718 Banfield, J. D. and Raftery, A. E., Model-based Gaussian and non-Gaussian
719 clustering. *Biometrics*. 49 (1993) 803–821.
- 720 Carlin, B. P. and Louis, T. A., Bayes and empirical bayes methods for data
721 analysis. Chapman & Hall, New York, 2000.
- 722 Celeux, G. and Hurn, M. and Robert, C.P., Computational and inferential
723 difficulties with mixture posterior distributions. *Journal of the American*
724 *Statistical Association.*, 95 (2000) 957–970
- 725 Chib, S. and Greenberg, E. Understanding the Metropolis-Hastings Algo-
726 rithm. *The American Statistician*. 49 (1995) 327–335.

- 727 Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum Likelihood From
728 Incomplete Data via the EM Algorithm (with discussion). *Journal of the*
729 *Royal Statistical Society, Ser. B.*, 39 (1977) 1–38.
- 730 Diebolt, J. and Robert, C. P., Estimation of finite mixture distributions
731 through Bayesian sampling. *Journal of the Royal Statistical Society, Series*
732 *B.* 56 (1994) 363–375.
- 733 Fraley, C. and Raftery, A. E., Model-Based Clustering, Discriminant Analy-
734 sis, and Density Estimation. *Journal of the American Statistical Associa-*
735 *tion.* 97 (2002) 611-631
- 736 Geman, S. and Geman, D., Stochastic relaxation, Gibbs Distributions and
737 the Bayesian Restoration of Images. *IEEE Transactions on Pattern Anal-*
738 *ysis and Machine Intelligence.* 6 (1984) 721–741.
- 739 Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., Markov chain Monte
740 Carlo in practice. Chapman & Hall, London, 1996
- 741 Gormley, I. C. and Murphy, T. B., Discussion of Handcock et al. ‘Model-based
742 clustering for social networks’. *Journal of the Royal Statistical Society,*
743 *Series A.* 170 (2007) 327–327.
- 744 Gormley, I. C. and Murphy, T. B., Discussion of Raftery et al. ‘Estimating the
745 Integrated Likelihood via Posterior Simulation Using the Harmonic Mean
746 Identity’. In: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
747 D. Heckerman, A. F. M. Smith and M. West (Eds.), *Bayesian Statistics 8,*
748 *Oxford University Press,* 2007, pp. 38–40.
- 749 Gormley, I. C. and Murphy, T. B., A Mixture of Experts Model for Rank Data
750 with Applications in Election Studies. *The Annals of Applied Statistics.* 2
751 (2008) 1452–1477.
- 752 Gormley, I. C. and Murphy, T. B., Clustering ranked preference data using
753 sociodemographic covariates. In: *Choice Modelling: The State-of-the-Art*
754 *and the State-of-Practice.* S. Hess and A. Daly (Eds). Emerald (2009) In
755 Press.
- 756 Gormley, I. C. and Murphy, T. B., A grade of membership model for rank
757 data. *Bayesian Analysis.* 4 (2009b) 265–296.

- 758 Handcock, M. S., Raftery, A. E. and Tantrum, J. M., Model-based clustering
759 for social networks. *Journal of the Royal Statistical Society, Series A.* 170
760 (2007) 301–354.
- 761 Hoff, P. D., Raftery, A. E. and Handcock, M. S., Latent Space Approaches to
762 Social Network Analysis. *Journal of the American Statistical Association.*
763 97 (2002) 1090-1098.
- 764 Hunter, D. R. and Lange, K., A tutorial on MM algorithms. *The American*
765 *Statistician*, 58 (2004) 30–37.
- 766 Hunter, D. R., MM algorithms for generalized Bradley-Terry models, *The*
767 *Annals of Statistics*, 32, (2004) 384–406.
- 768 Hurn, M., Justel, A. and Robert, C. P., Estimating Mixtures of Regressions.
769 *Journal of Computational and Graphical Statistics.* 12 (2003) 55–79.
- 770 Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E., Adaptive
771 mixture of local experts. *Neural Computation.* 3 (1991) 79–87.
- 772 Jordan, M. I. and Jacobs, R. A., Hierarchical mixtures of experts and the
773 EM algorithm. *Neural Computation.* 6 (1994) 181–214.
- 774 Kass, R. E. and Raftery, A. E., Bayes factors. *Journal of the American*
775 *Statistical Association.* 90 (1995) 773–795
- 776 Krivitsky, P. N. and Handcock, M. S., Fitting Position Latent Cluster Mod-
777 els for Social Networks with latentnet. *Journal of Statistical Software.* 24
778 (2008) 1–23.
- 779 Krzanowski, W. J., *Principles of Multivariate Analysis: A User’s Perspec-*
780 *tive.*, Clarendon Press, 1988.
- 781 Lange, K., Hunter, D. R. and Yang, I., Optimization transfer using surrogate
782 objective functions. *Journal of Computational and Graphical Statistics.* 9
783 (2000) 1–59.
- 784 Lazega, E., *The Collegial Phenomenon: The Social Mechanisms of Coop-*
785 *eration Among Peers in a Corporate Law Partnership.* Oxford University
786 Press. Oxford, England, 2001

- 787 McCullagh, P. and Nelder, J.A., Generalized Linear Models. Chapman and
788 Hall, London, 1983.
- 789 Metropolis, N., Rosenbluth A. W., Rosenbluth, M. N., Teller, A. H. and
790 Teller, E. Equations of state calculations by fast computing machine. Jour-
791 nal of Chemical Physics. 21 (1953) 1087 – 1091.
- 792 O’Hagan, A. and Forster, J. Kendall’s Advanced Theory of Statistics: Vol-
793 ume 2B Bayesian Inference. Arnold, London, UK, 2004.
- 794 Peng, F. and Jacobs, R. A. and Tanner, M. A., Bayesian Inference in
795 Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With
796 and Application to Speech Recognition. Journal of the American Statisti-
797 cal Association. 91 (1996) 953–960.
- 798 Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N., Esti-
799 mating the Integrated Likelihood via Posterior Simulation Using the Har-
800 monic Mean Identity. In: *Bayesian Statistics 8*, J. M. Bernardo, M. J.
801 Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and
802 M. West (Eds.), Oxford University Press (2007), pp. 1–45.
- 803 Richardson, S. and Green, P. J., On Bayesian Analysis of Mixtures With
804 An Unknown Number of Components., Journal of the Royal Statistical
805 Society, Series B. 59 (1997) 731–758.
- 806 Salter-Townshend, M. and Murphy, T. B., Variational Bayesian inference for
807 the Latent Position Cluster Model. (2009) *Submitted*.
- 808 Schwarz, G., Estimating the dimension of a model. The Annals of Statistics.,
809 6 (1978) 461–464
- 810 Snijders, T. A. B., Christian E. G. S., Schweinberger, M. and Huisman, M.,
811 Manual for SIENA version 2.1. Groningen, Netherlands (2005)
- 812 Snijders, T.A., Pattison, P.E., Robins, G.L. and Handcock, M.S., New speci-
813 fications for exponential random graph models. Sociological Methodology.
814 (2006) 99–153
- 815 Stephens, M., Dealing with label-switching in mixture models. Journal of the
816 Royal Statistical Society, Series B., 62 (2000) 795–810.

817 Wainwright, M. J. and Jordan, M. I., Graphical Models, Exponential Fam-
818 ilies, and Variational Inference. Foundations and Trends® in Machine
819 Learning., 1 (2008) 1–305.