Model selection for Cognitive Social Structures

MODEL SELECTION FOR COGNITIVE SOCIAL STRUCTURES

JOHAN KOSKINEN

Abstract. Measurement accuracy is an inherent problem in social network analysis. The issue of actor accuracy in the reporting of their interactions with others, was raised by Bernard, Killworth and Sailer (e.g. Bernard et al., 1980) and provoked extensive debate. Krackhardt (1987) later introduced the concept of Cognitive Social Structures and several methods for aggregating different actor reports on the network into a single graph, with the aid of which for example the congruence of reports could be gaged. Often when this data collecting paradigm is used the interest is in correlating bias on the part of the perceivers with exogenous attributes of the perceivers (e.g. Bondonio, 1998; Casciaro, 1998; Casciaro et al., 1999). A statistical model for aggregating separate reports into a single consensus network, with the additional benefit of allowing estimates of actor accuracy to be obtained in the process, was proposed by Batchelder et al. (1997). Using an extension of this model and a Bayesian approach we are able to incorporate effects of known covariates and network effects on perceptional biases. In Koskinen (2002a) it was suggested that the conditional probability of reporting a tie as present when a tie is really present (or absent) be modeled using a probit link function. This is further elaborated here with a special focus on finding standard reference priors that enables model selection. The main obstacle is that the model is not fully identified, something which can not be solved in any obvious way through restrictions or highly informative priors. The proposed solution is to asses a posteriori which are the main determinants of identifying conditions. We present a procedure for choosing prior distributions and provide the necessary adjustments to the original sampling scheme.

1. Introduction

Since analysis of social interaction among actors often is based on information provided by the actors themselves the mechanisms of *perception of social interaction* are crucial to understanding the results. The views on the nature of the relationship between the reports given by actors and the actual network of social interaction can roughly be divided in two, whom we may call a "relativist" perspective and an "essentialist" perspective. Naturally, we are not suggesting that there are any studies conforming to either extreme, the distinction is purely conceptual. In the sequel we assume that the study of perception of social interaction is formalized according to the data collection paradigm first introduced by Moreno (1934), further elaborated by Newcomb (1961), and subsequently given a formal structure by Krackhardt (1987). The latter introduced the term Cognitive Social Structures (CSS), referring to the collection of data obtained when each actor in

Key words and phrases. Bayesian statistical modeling; Consensus analysis; Cognitive social structures (CSS); Measurement reliability; Social network analysis; Binary probit; Bayesian model selection; Reference priors; Bayesian analysis of social networks.

a given network gave a report as to the network structure in its entirety. Hence, for each pair of actors A and B in the network, all actors are assumed to have reported on the relationship between A and B (for an extensive review on CSS see Pattison, 1994).

The paper by Krackhardt (1987) can be seen as a direct reaction on the BKS studies (e.g. Bernard et al. 1980, and Killworth and Bernard 1979; acronym due to Krackhardt 1987). This identifies two main perspectives on the relation between reports and what the reports relate to. One perspective focuses on the relation between the reports and some external reality. The main objective is identifying discrepancies between the reports and the true network structure represented by, for example, actual behaviour. The premise is that these reports are laden with bias that should be explained. A second perspective would be focusing entirely on the cognitive reconstructions themselves. This represents a shift of interest from comparing the reports to some external truth to comparing the various perceptions of the network with each other. These differences are interesting in their own right as measures of similarity of actors, but the actor reports can also be seen as the only true representations of the network. The latter point rests upon the assumption that actors act upon what they consider real (c.p. Krackhardt, 1987, p 112, on W.I. Thomas).

The essentialist view would state that there is a set network of social interaction which is true, every representation not concordant with this is bias. This leaves us the problem of defining and finding the true representation of the network. Reasoning from relativist perspective this formulation of the problem would not make sense. Should we want to go beyond mere descriptions of data, each perspective poses different problems. For an essentialist point of view, defining a truth is problematic since any rule based on the actual reports compromises the modeling of reports. To see this consider using the Locally aggregated structure from the union rule (Krackhardt, 1987) as a representation of the true network structure, i.e. a tie between actors A and B is assumed to exist if either A or B reports it as present. Now, this does not allow us to fully study biases such as over representing own ties (prestige bias in some settings). If A and B are in agreement about the existence of a tie between them, this could mean that "there exists a tie between them" as well as "there does not exist a tie between them but they both would like you to think so". Studies have been conducted in which the true interaction pattern is measured in advance and the actor reports are evaluated against the previously measured interaction (e.g. Bernard et al., 1980; Killworth and Bernard, 1979). It can be argued however that measured frequencies of interaction and actor reports does not relate to the same phenomenon. From the relativist perspective the main problem is to relate the reports to each other. Using for example multi-dimensional scaling or correspondence analysis, the spatial structure of the reports can be studied (Kumbasar et al., 1994) but the reports cannot then be understood in terms of social interaction, i.e. the reports are not directly related to the network structure.

Batchelder et al. (1997) specified a statistical model that incorporates elements from both perspectives. The existence of a true structure is assumed and a model

for informant reports is specified conditional on the true structure. The true structure is not known however, wherefore it can be thought of as either true and unknown *or* as merely a conceptual truth that serves as a reference point for interpreting the conditional model, i.e. "had this been the true network...". While their model in a way reconciles the above mentioned different perspectives on perceptions of social interaction, inference is not at all straightforward with standard frequentist tools. Firstly, their proposed inference scheme excludes certain configurations of candidate true graphs and data sets, and secondly, their model is only identifiable if restrictions are imposed on the parameter space. A fully Bayesian analysis of their model that solved these both problems as well as provide a procedure for model selection was presented in Koskinen (2001) and further elaborated in Koskinen (2002b). Some of these results have independently been given by Butts (2003). As it were, this Bayesian adoption only incorporated network effects (e.g. is "accuracy" effected by whether the informant is involved in the dyad that he or she is reporting on) in a slightly static manner, requiring many parameters. A more flexible way of approaching the model of Batchelder et al. (1997), is to specify the conditional model given the true structure as a probit model (Koskinen, 2002a). Network effects as well as observable attributes are easily incorporated into the study of informant accuracy and, subsequently, model inference that involves only little additional programming (as compared to parameter inference conditional on a model specification) made possible (Koskinen et al., 2002).

Although one can assume that prior information in some form is always available and hence that the formulation of prior distributions of the parameters is possible, the influence of the prior distributions on model selection is an issue that needs to be dealt with on a case to case basis. More specifically, the precision of the prior distributions and their relation to the precision of the posterior distribution over candidate models is an area for further investigation and most certainly something for which a universal rule does not apply. The model of Batchelder et al. (1997) is not fully identified and we argue in this paper that it is in general difficult to determine *a priori* what restrictions lead to identifiability. The proposed solution is to explore the posterior distribution of the parameters to locate modal regions. In this paper we propose an inference scheme where reference priors are used allowing some degree of automation in the model selection. Further, we introduce the necessary modifications to the original algorithm needed to carry out the inference scheme with reference priors. The concept of "reference" prior carries some of the notions of "non-informative" priors, i.e. priors that do not favor any particular regions of the parameter space. For model inference or model selection to be possible all prior distributions have to be proper which rules out using constant priors (c.f. O'Hagan, 1995).

2. Preliminaries

Let the actors be represented by the fixed set of nodes $V = \{1, \ldots, n\}$, in a (di-) graph $G = (V, E)$. The set of arcs $E \subseteq V^{(2)} = \{(u, v) \in V \times V : u \neq v\}$, represents the relation among the actors. In the sequel we will treat this arc set as

if it were the true structure of the network and let $\mathbf{Z} = (z_{jk} : (j,k) \in V^{(2)})$ be the corresponding adjacency matrix, i.e. the indicator function such that $z_{jk} = 1$ if $(j,k) \in E$ and nought otherwise (this is not strictly speaking an adjacency *matrix* since the diagonal is excluded). $\mathbf{Z}$ takes values in $\mathscr{Z} = \{0,1\}^{n(n-1)}$, i.e. the space of adjacency matrices for G.

In many applications the set E of arcs is taken to be the self-reports of the actors in the network but in accordance with the above mentioned data collecting paradigm, for m informants, we assume that we have for $i \in \mathscr{I} = \{1,\ldots,m\}$ reports $\mathbf{X}_i = (x_{ijk})$ on $\mathbf{Z}$, where $x_{ijk} = 1$ if informant i states that $z_{jk} = 1$, and 0 otherwise. In the regular Cognitive Social Structure the set of informants and the set of actors is one and the same but we retain the notational distinction. We collect these reports in $\mathbf{X} = (\mathbf{X}_i : i \in \mathscr{I})$ giving an array of $n \times (n-1) \times m$ elements in three dimensions. Assume that conditional on $\mathbf{Z}$, $\mathbf{X}_1,\ldots,\mathbf{X}_m$ are independent and their elements are independent and satisfy

$$\Pr\left(X_{ijk} = x_{ijk} \middle|\, \mathbf{Z} = \mathbf{z}\right) = \Pr\left(X_{ijk} = x_{ijk} \middle|\, Z_{jk} = z_{jk}\right). \tag{2.1}$$

The model of Batchelder et al. (1997) is obtained by dividing the conditional probabilities in Eq (2.1) into two types, *hit* and *false alarm* probabilities respectively

$$\Pr\left\{X_{ijk} = 1 \middle|\, Z_{jk} = z_{jk}\right\} = \begin{cases} H_{ijk} & if \quad z_{jk} = 1 \\ F_{ijk} & if \quad z_{jk} = 0 \end{cases}. \tag{2.2}$$

Now, for obvious reasons (compared to the dimensions of $\mathbf{X}$) the $n\,(n-1)\,(2m+1)$ number of parameters of (2.2) need to be reduced if we wish to make inference about hit and false alarm probabilities as well as the true structure. Ways for doing this are presented in Batchelder et al. (1997) and Koskinen (2002b). In the latter $\mathscr{I} \times V^{(2)}$ was partitioned into classes $C = \{0,1,\ldots,c-1\}$, and it was assumed that $H_{ijk} = H_\ell$, and $F_{ijk} = F_\ell$ for (i,j,k) belonging to class ℓ in C. An alternative (Koskinen, 2002a) is to assume that these conditional probabilities can be modeled using a normal link function. Let us assume that we have observed, for each combination of informant and pair of actors in the network, $(i,j,k) \in \mathscr{I} \times V^{(2)}$, a covariate $p \times 1$ vector $\mathbf{w}_{ijk}$ and that we have two $p \times 1$ vectors of unknown coefficients $\beta_1 \in \Theta_1$ and $\beta_0 \in \Theta_0$, $\Theta_1, \Theta_0 \subseteq \mathbb{R}^p$. We could then let $H_{ijk} = \Phi(\mathbf{w}'_{ijk}\beta_1)$, and $F_{ijk} = \Phi(\mathbf{w}'_{ijk}\beta_0)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf)[1]. Thus the likelihood given data and an observed set of covariates can be written

$$\begin{aligned} \mathcal{L}(\beta_1,\beta_0,\mathbf{z};\mathbf{x}) \;=\; & \prod_{(i,j,k)\in\mathscr{I}\times V^{(2)}} \left\{\Phi(\mathbf{w}'_{ijk}\beta_1)^{x_{ijk}} + (1-\Phi(\mathbf{w}'_{ijk}\beta_1))^{(1-x_{ijk})}\right\}^{z_{jk}} \\ & \times \left\{\Phi(\mathbf{w}'_{ijk}\beta_0)^{x_{ijk}} + (1-\Phi(\mathbf{w}'_{ijk}\beta_0))^{(1-x_{ijk})}\right\}^{1-z_{jk}} \end{aligned} \tag{2.3}$$

Recall that the choice of a standard normal link function rather than a normal link function with another scale is arbitrary because of the scale invariance of the probit model (see McCulloch and Rossi, 1994, for a fully Bayesian approach to the

[1] In order to interpret these probabilities in the *theory of signal detection* paradigm, Batchelder et al. (1997) used standard normal cdf's to infer *signal perceptability* and *response bias*.

identification problem in the multinomial probit model). The choice of a probit link function rather than a logit link function is motivated by the convenient form of the MCMC sampling scheme (see further Section 4).

Note that if we do not hypothesize or stipulate the direction of influence that covariates have on the reporting we cannot infer the structure $\mathbf{Z}$. Put more plainly, if, say all actors reports a certain tie to be present, they could all either lie or speak the truth.

Observation 1. *The model as defined by the likelihood 2.3 is not identified, i.e. given data* $\mathbf{x}$ *for every* $\beta_1 \in \Theta_1$, $\beta_0 \in \Theta_0$, *and* $\mathbf{z} \in \mathscr{Z}$

$$L(\beta_1, \beta_0, \mathbf{z}; \mathbf{x}) = L(\beta_0, \beta_1, \widetilde{\mathbf{z}}; \mathbf{x}),$$

where $\widetilde{z}_{jk} = 1 - z_{jk}$ *for the elements of* $\widetilde{\mathbf{z}}$.

To counter this, one can for example impart restrictions on the parameters (see Batchelder et al., 1997, in the context of CSS; and Salabasis and Villani, 2000, in a related model). It is not straightforward to decide what restrictions to use. Firstly, it is very hard a priori to evaluate what effects these restrictions might have on the analysis. What seemed sensible a priori might turn out to be unrealistic (cp the low correspondence between reported communication and "actual" communication in the BKS studies Bernard et al., 1980; Killworth and Bernard, 1979). Secondly, note that it is not necessarily sufficient to put restrictions on one coordinate, since the model might still be conditionally un-identified in some sub-space of $\Theta_1 \times \Theta_0 \times \mathscr{Z}$. We favour and consider it a more sensible idea to quantify prior information and ideas into proper prior distributions, which do not exclude, unlike restrictions, regions in the parameter space (Koskinen, 2002b,a).

The use of proper prior distributions paves the way for a rich analysis of data ranging from the analysis of residuals to model selection. It is however a definite advantage if there exists a family of reference priors. These provide reference for testing the sensitivity of parameter estimation and model selection to different prior specifications. When selecting one model from a large class of models it is also convenient to have a procedure for setting priors "automatically". In the present context there is the additional problem of identification. Observation 1 suggests that the posterior distribution is multimodal if the prior density is vague. The regions with high posterior density are separated by regions with low posterior density. One might add prior information that remedies this to a certain extent but the effectiveness of this in alleviating the problem is unclear. Our suggestion is that symmetric priors should be used that allows us to fully explore the multimodality of the posterior distribution since the location of the high density regions in relation to each other is essential information. By symmetry, we refer to the indeterminacy following Observation 1.

3. Prior distributions

In this section we derive a class of reference prior distributions for the probit coefficients β_1, β_0, and $\mathbf{Z}$. For the probit coefficient β_s denote the prior mean

$\mu_s = (\mu_{s,v} : v = 1, \ldots, p)$, and the corresponding (marginal) variance covariance matrices Σ_s, for $s = 0, 1$. Below we will set $\mu_s = \mathbf{0}$ and $\Sigma_s = \Sigma$ for $s = 0, 1$.

The problem of determining proper subjective prior distributions can be viewed from two perspectives. On the one hand, the dilemma is not so much about determining the direction of the effects of the covariates (mean level) as much as determining the joint influence (covariance) and our prior uncertainty about these directions (variances). On the other hand, we do not know which covariates have an effect at all and hence if tight priors for these coefficients are enough to assure that the posteriors are unimodal. For example, by setting a tight prior on the intercept coefficients $\beta_{1,0}$ and $\beta_{0,0}$, so that we a priori believe actors to be accurate "on average" when reporting ties (i.e. strong positive and strong negative means respectively) would seem a reasonable measure for achieving uni-modality through the implicit "exclusion" of certain structures in $\mathscr{Z}$. Should these mean levels of accuracy turn out to be marginal (with posteriors centered over 0) in comparison with the rest of the effects, we could still end up with multimodal posteriors. A second point relating to the former is *just how much* precision is needed to counter multimodality?

We consider only multivariate normal reference priors since the normal distribution is conjugate with respect to the latent variable to be defined in Section 4.

Definition 1. *For the model with likelihood (2.3) the joint prior on β_0, β_1, and $\mathbf{Z}$ that satisfies*

(1) *β_0, β_1, and $\mathbf{Z}$ are independent*
(2) *$\mathbf{Z}$ is uniformly distributed on $\mathscr{Z}$*
(3) *$\beta_0, \beta_1 \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a diagonal matrix*

$$\mathbf{\Sigma} = \frac{\xi^2}{p} diag(w_1^{-2}, \ldots, w_p^{-2}),$$

for $\xi > 0$, is a reference prior with respect to the typical covariate vector $\mathbf{w} = (w_1, \ldots, w_p)'$, $w_v \neq 0$, for all v.

Although it is not strictly necessary for a reference prior to be either sensible or interpretable, the idea of the prior in Definition 1 is to provide a measure of (little) information with can be understood in terms of data. Here follows some characterizations of the reference prior and its interpretations.

When using the probit model in general one has to take the scale of the covariates into consideration when constructing prior distributions. In analogy to the Imaginary Minimal Experiment of Spiegelhalter and Smith (1982) the reference prior is constructed working from the notion of a *typical observation*, a kind of thought experiment. For each model we assume that there is a typical report x_{ijk}, corresponding to a triple (i, j, k) and covariate vector $\mathbf{w} = (w_v)$, to be constructed later. The use of one typical observation captures the "minimal" part of the thought-experiment, since we construct our priors in relation to the influence of a single observation.

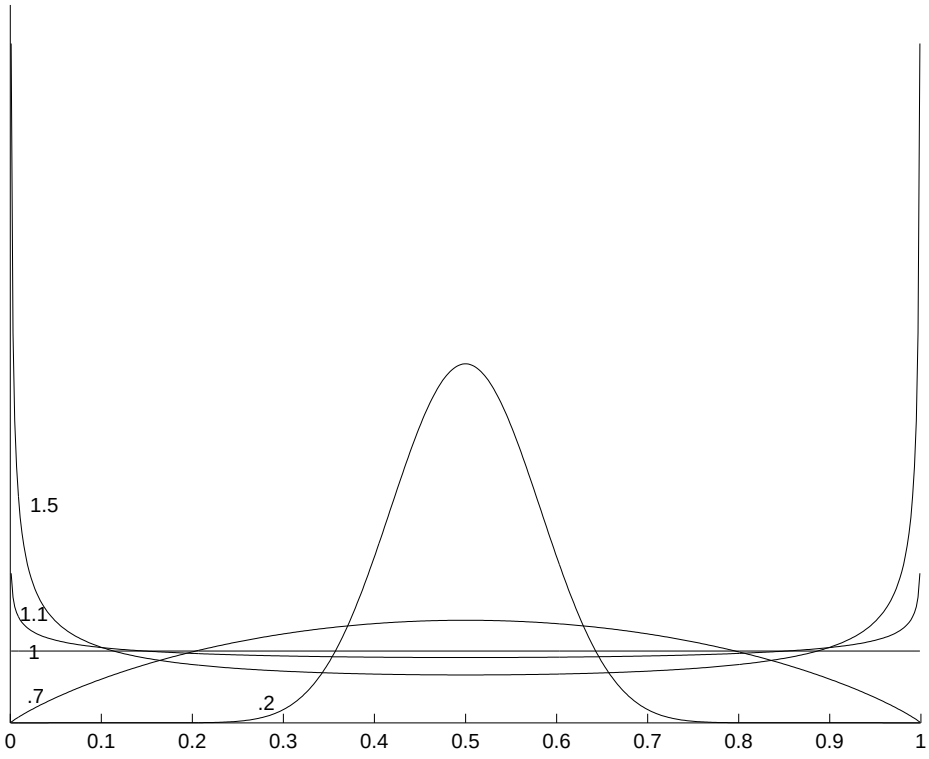

FIGURE 1. Densities for the cdf transform for $\xi = .2, .7, 1, 1.1, 1.5$

The prior independence of parameters is motivated by the fact that there is no immediately apparent way of defining dependencies. In addition, prior dependencies would possibly disturb the symmetry of posteriors following from Observation 1. Similarly, with Σ a diagonal matrix the prior does not favor any types of interdependencies between the effects of covariates. Although this neglects the fact that the scale of a covariate coordinate needs to be understood in relation to the other covariate values, i.e. the covariation among covariates. The impact of these covariances on the posteriors of the coefficients is however dependent on what observations are grouped together (for example the case when, say for two reports, if one is considered a false alarm, the other is likely to be as well).

We now take a look at the induced prior on H_{ijk} for the typical observation (the derivation is given in the appendix). Since by assumption $\mu_1 = \mathbf{0}$ it is clear that the expected value of H_{ijk} with respect to the reference prior is $1/2$. For $\xi = 1$, we have that

$$H_{ijk} \sim \text{Rectangular}(0, 1),$$

which is a common way of modeling "ignorance" in Bernoulli trials. It is straightforward to show (using a little algebra; see the appendix) that the probability density function for general ξ of H_{ijk} corresponding to a typical covariate vector is given by

$$\xi^{-1} \exp\left\{ -\frac{1}{2}\left(\xi^{-2} - 1\right)\left\{\Phi^{-1}(H_{ijk})\right\}^2 \right\},$$

which is well approximated by a Beta$(\alpha(\xi), \alpha(\xi))$ density, $\alpha(\xi) = (\pi/4)\,(\xi^{-2} - 1) + 1$. How the scaling factor ξ affects the distribution of H_{ijk} is illustrated in Figure 1. Using the Beta approximation, Jeffreys prior is for example obtained as the induced prior for $\xi = \sqrt{\pi/(\pi - 2)} \approx 1.66$. Since one typical observation serves as the reference point, we now turn to the induced priors on H for observations whose covariate vectors differ from the typical covariate vector (the index of H is dropped since its function here is replaced by the covariate vector). Assume that the "atypical" observation has covariate vector $\mathbf{w}^*$ and the typical covariate vector is $\mathbf{w}$. Now for given ξ if $|w_v^*| < |w_v|$ for all $v = 1, \ldots, p$, the induced prior on H for the covariate vector $\mathbf{w}^*$ will, using the beta approximation, be Beta(a^*, a^*) with $a^* > a(\xi)$ (follows from the fact that $v = \mathbf{w}^{*\prime}\beta$ is a linear combination of independent normal variates, and the relation between the variance of v and the beta approximation is given in the appendix). Conversely, if $|w_v^*| > |w_v|$ for all $v = 1, \ldots, p$, the induced prior on H for the covariate vector $\mathbf{w}^*$ will be Beta(a^*, a^*) with $a^* < a(\xi)$.The interpretation being that , loosely speaking, the less "informative" a covariate vector is in relation to the typical covariate vector, the more concentrated around zero the induced prior will become. The same argument as for less "informative" can be stated for more "informative". The more "informative", the more u-shaped the induce prior on H and F will become. The reason for putting informative in quotes is that there is no simple relation between how much information an observation contribute and the absolute values of the individual elements of the covariates. Recall also that the posterior distributions are not in general uni-modal. Hence, a certain amount of u-shapedness in the induced priors lends support to a model for which the posteriors of the coefficients are not centered over the origin.

Dellaportas and Forster (1999) consider the induced prior on the logarithm of the mean in log-linear models, for different choices of prior distribution for the parameters. They do not however consider the explicit form of the induced prior on the mean, but use the first and second moments for matching the prior mean and covariance structure to that of a gamma conjugate prior distribution. The relationship between the Dirichlet and gamma distribution then allows them to study the prior expected cell counts. In the context of generalised linear models, Raftery (1996a) developed a methodology for setting reference priors that for normal regression corresponds to giving all model parameters, except the intercept, independent normal prior distributions with a zero mean and common variance. The intercept is given a special status, and is modeled with a separate independent normal distribution. For generalised linear models, these priors are however transformed using observed data. It is important to keep in mind that we want to retain some notion of the "minimal" in our priors, and in contrast with setting priors in generalised linear models, the information provided by data about the coefficients is mediated through the latent structure, $\mathbf{z}$.

Further properties of the reference prior is that, in analogy with the Imaginary Minimal Experiment, when comparing two models, both of which using the same typical observation as their point of reference for setting the prior distributions, the Bayes factor is unity for given ξ. This follows from the fact that the induced

priors on H and F are identically distributed for the (thought) experiment of obtaining a realization on the typical observation. Hence every realization of the typical observation has the same probability in both models.

It remains only to construct the typical actor. One way is to arbitrarily, say at random, select an actual covariate vector and using it if all elements are non-zero. Another way is setting the covariate vector equal to the standard deviations of the covariates. In the concluding empirical section, we tentatively investigate the effects of different specifications of the typical observation.

4. Posterior distributions

To facilitate estimation procedures, following Albert and Chib (1993), we introduce the $n \times (n-1)$ vector of latent variables $\mathbf{Y}_i = (Y_{ijk})$, that for each triple (i, j, k), are independently distributed

$$Y_{ijk} \sim N\left(\mathbf{w}'_{ijk}\beta_1, 1\right) \text{ if } Z_{jk} = 1,$$

and

$$Y_{ijk} \sim N\left(\mathbf{w}'_{ijk}\beta_0, 1\right) \text{ if } Z_{jk} = 0,$$

and let $X_{ijk} = 1$ if $Y_{ijk} > 0$ and $X_{ijk} = 0$ if $Y_{ijk} \leqslant 0$. Consequently we have the standard regression form

$$Y_{ijk} = \sum_{v=1}^{p} w_{ijk,v}\beta_{s,v} + \varepsilon_{ijk}$$

for $s = 0, 1$. One important implication of this formulation is that the underlying structures behind the observable graphs $\mathbf{X}_1, \ldots, \mathbf{X}_m$ can be seen as a collection of valued graphs $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$, and if in $\mathbf{Y}_i$, (j, k) has a value or intensity greater than zero we get an edge (j, k) in $\mathbf{X}_i$. Additionally, if we have a valued graph with continuous strengths of the relationships it is natural to model it with ordinary regression techniques.

It is straightforward to check that the introduction of the latent variables ,$\mathbf{Y}_i$, retains the structure of the likelihood function in (2.3). For example, for given triple (i, j, k), $H_{ijk} = \Pr(Y_{ijk} > 1 | Z_{jk} = 1)$, which in terms of a standard normal variate is $\Phi(\mathbf{w}'_{ijk}\beta_1)$. Also, with conjugate priors, sampling from the exact posteriors of the parameters is fairly straightforward. Full details about the Gibbs sampling algorithm can for example be found in Gelfand and Smith (1990).

4.1. Gibbs sampler.

To implement the Gibbs sampler we need the full conditional posteriors of each of the parameters. When the parameters are a priori

independent the joint conditional distribution of the parameters and latent variables given data is

$$\begin{aligned}\pi\left(\beta_1,\beta_0,\mathbf{z},\mathbf{y}|\,\mathbf{x}\right) \;\propto\; & \pi\left(\beta_1\right)\pi\left(\beta_0\right)\pi\left(\mathbf{z}\right)\\ & \times\prod_{(i,j,k)\in\mathscr{I}\times V^2}\left\{\mathbf{1}\left(Y_{ijk}>0\right)\mathbf{1}\left(x_{ijk}=1\right)\right.\\ & \left.+\mathbf{1}\left(Y_{ijk}\leqslant 0\right)\mathbf{1}\left(x_{ijk}=0\right)\right\}\varphi\left(Y_{ijk}-\mathbf{w}_{ijk}'\beta_1\right)^{z_{jk}}\\ & \times\left\{\mathbf{1}\left(Y_{ijk}>0\right)\mathbf{1}\left(x_{ijk}=1\right)\right.\\ & \left.+\mathbf{1}\left(Y_{ijk}\leqslant 0\right)\mathbf{1}\left(x_{ijk}=0\right)\right\}\varphi\left(Y_{ijk}-\mathbf{w}_{ijk}'\beta_0\right)^{1-z_{jk}},\end{aligned}$$

where $\varphi(\quad)$ is the $N(0,1)$ probability density function (pdf). Here we have used π to denote a generic prior or, alternatively, posterior distribution, which is meant to be interpreted according to its argument. With reference priors of the kind we have proposed earlier, the prior densities of the coefficients are multivariate normal whereas the prior probability mass function $\pi(\mathbf{z})$ is incorporated into the normalizing constant.

The forms of the full conditional posteriors deviates from how they are described in Albert and Chib (1993) only in the dependence on $\mathbf{z}$. With independent conjugate priors of the form $\beta_s\sim N_p(\mathbf{0},\mathbf{\Sigma}_s)$, we have independently for $s=0,1$

$$\left(\beta_s|\,\mathbf{z},\mathbf{y}\right)\sim N_p\left(\beta_s^*,\mathbf{B}_s^*\right)$$

where $\beta_s^*=\mathbf{B}_1^*\mathbf{S}_s$, $\mathbf{B}_s^*=\left(\mathbf{\Sigma}_s^{-1}+\mathbf{C}_s\right)^{-1}$, and

$$\mathbf{S}_s=\sum_{(j,k)\in V^{(2)}}\mathbf{1}\left\{z_{jk}=s\right\}\sum_{i\in\mathscr{I}}y_{ijk}\mathbf{w}_{ijk},\qquad \mathbf{C}_s=\sum_{(j,k)\in V^{(2)}}\mathbf{1}\left\{z_{jk}=s\right\}\sum_{i\in\mathscr{I}}\mathbf{w}_{ijk}\mathbf{w}_{ijk}'.$$

The full conditional posterior of the coefficients does not depend on data given the latent variable $\mathbf{y}$. For the latent variables $\mathbf{Y}$ we have that for each element if $Z_{jk}=s$

$$\left(Y_{ijk}|\,\mathbf{x},\beta_1,\beta_0,\mathbf{z}\right)\sim N\left(\mathbf{w}_{ijk}'\beta_s,1\right),$$

truncated to the left at 0 if $x_{ijk}=1$ and truncated to the right at 0 if $x_{ijk}=0$, for $s=0,1$. We can write the full conditional posterior of Z_{jk} independently for each j,k

$$\left(Z_{jk}|\,\mathbf{x},\beta_1,\beta_0,\mathbf{y}\right)\sim\left(Z_{jk}|\,x_{1jk},\ldots,x_{mjk},\beta_1,\beta_0\right)\sim Bernoulli\left(\frac{1}{1+q_{jk}}\right)$$

where

$$q_{jk}=\prod_i\left(\frac{\Phi\left(\mathbf{w}_{ijk}'\beta_0\right)}{\Phi\left(\mathbf{w}_{ijk}'\beta_1\right)}\right)^{x_{ijk}}\left(\frac{1-\Phi\left(\mathbf{w}_{ijk}'\beta_0\right)}{1-\Phi\left(\mathbf{w}_{ijk}'\beta_1\right)}\right)^{1-x_{ijk}}.$$

Conditioning on both data and the latent variable is superfluous since the only relevant information for Z_{jk} is whether the informants have reported the arc (i,j) as present or not, given everything else.

By cycling through these conditional posteriors for a certain number of steps, after a certain burn-in period the joint output is a sample from the exact joint

posterior of the parameters given data ("exact" here means arbitrarily close to the desired distribution in total variation distance).

4.2. **A switching device.** As a consequence of Observation 1, the posterior mass will be concentrated on two disjoint regions of the parameter space. To characterize these regions, let $\mathscr{P}$ be the class of all partitions $P : \mathscr{Z} \to \{0,1\}$, such that $P(\mathbf{z}) = a$ implies that $P(\widetilde{\mathbf{z}}) = 1 - a$. For $P \in \mathscr{P}$ define the "inverse" $P^{-1}(a) = \{\mathbf{z} \in \mathscr{Z} : P(\mathbf{z}) = a\}$. Now, from observation 1, when we have symmetric prior distributions, there exist a $P \in \mathscr{P}$, such that

$$\pi(\beta_1, \beta_0 | \mathbf{x}) = \frac{1}{2}\pi_1(\beta_1, \beta_0 | \mathbf{x}) + \frac{1}{2}\pi_0(\beta_1, \beta_0 | \mathbf{x}) \tag{4.1}$$

where

$$\pi_a(\beta_1, \beta_0 | \mathbf{x}) = \sum_{\mathbf{z} \in P^{-1}(a)} \pi(\beta_1, \beta_0, \mathbf{z} | \mathbf{x}).$$

Furthermore $\pi_a(\beta_1, \beta_0 | \mathbf{x}) = \pi_{1-a}(\beta_0, \beta_1 | \mathbf{x})$, for all $\beta_0 \in \Theta_0$, $\beta_1 \in \Theta_1$. Typically, when the modal regions are far apart, the ordinary Gibbs algorithm will get stuck in one of the regions and produce a sample from $\pi_a(\beta_1, \beta_0 | \mathbf{x})$ rather than $\pi(\beta_1, \beta_0 | \mathbf{x})$, for some $P \in \mathscr{P}$. Even if there is the occasional jump from one modal region to another, the proportions in the mixture 4.1 are likely to be bad.

There are various possibilities for handling these problems. One solution would be to force the Markov chain to stay in one of the modal regions and adjust the weight afterwards. Another solution would be using a stepping stone (as described in e.g. Gilks and Roberts, 1996). Both of these approaches would require a presampling scheme, to identify the modal regions and a quite elaborate analysis for adjusting the Gibbs algorithm.

Consider instead the following coupling scheme. The Gibbs sampler can be put in a form similar to the Metropolis-Hastings algorithm. Let $\{U_{ij}\}$, $j = 1, \ldots, p$, $i = 1, \ldots, N$ be an array of independent draws from the uniform distribution, and let θ_{ji} be the jth component of the parameter vector in the ith iteration. Denote by $F(\cdot|\cdot)$ a generic conditional distribution function derived from the posterior density $\pi(\cdot)$, and $F^{-1}(\cdot|\cdot)$ the corresponding quantile function.

The algorithm is initiated by drawing θ_{j0} from some, possibly degenerate, distribution. In every iteration $i = 1, \ldots, N$, the following up-dating step is used: for $j = 1, \ldots, p$, equate

$$\theta_{ji} = F^{-1}(u_{ij} | \theta_{1,i}, \ldots, \theta_{j-1,i}, \theta_{j+1,i-1}, \ldots, \theta_{p,i-1}),$$

for continuous components, and for discrete components set θ_{ji} to the minimum value satisfying

$$u_{ij} \leq F(\theta_{ji} | \theta_{1,i}, \ldots, \theta_{j-1,i}, \theta_{j+1,i-1}, \ldots, \theta_{p,i-1}).$$

Now consider the case when we have several copies of the Markov chain $\theta_{ji}^1, \ldots, \theta_{ji}^K$, each obeying the updating rules described above, starting with different initial configurations $\theta_{j0}^1, \ldots, \theta_{j0}^K$, but using the *same* sequence of random numbers $\{U_{ij}\}$. This technique is called coupling and can be used to study convergence of MCMC samples to the stationary distribution (in particular, see Johnson, 1996, on coupling schemes for the Gibbs sampler). Let N denote the first iteration such that

$\theta^k_{ji} = \theta^h_{ji}$, for $1 \leq k < h \leq K$, for all discrete components and $|\theta^k_{ji} - \theta^h_{ji}| < \varepsilon$, $\varepsilon > 0$, for $1 \leq k < h \leq K$, and for all continuous components. N is said to be the time at which all the copies have "coupled", or coalesced, and once all copies have coupled they stay together. In particular, if $\theta^k_{jt} = \theta^h_{jt}$, for $1 \leq k < h \leq K$, for all components at iteration $t = N$, $\theta^k_{ji} = \theta^h_{ji}$,, for $1 \leq k < h \leq K$ for $i \geq t$.

For our purposes, assume that we run two parallel chains θ^1_{ji} and θ^2_{ji}, in the way prescribed above with a common sequence of random numbers $\{U_{ij}\}$, but with an alteration to the up-dating order of the components. For θ^1_{ji} we successively up-date

(1) $\theta^1_{1i} \sim \mathbf{z}|\cdot$
(2) $\theta^1_{2i} \sim \mathbf{y}|\cdot$
(3) $\theta^1_{3i} \sim \beta_0|\cdot$
(4) $\theta^1_{4i} \sim \beta_1|\cdot$

whereas for θ^2_{ji} we use the up-dating order

(1) $\theta^2_{1i} \sim \mathbf{z}|\cdot$
(2) $\theta^2_{2i} \sim \mathbf{y}|\cdot$
(3) $\theta^2_{3i} \sim \beta_1|\cdot$
(4) $\theta^2_{4i} \sim \beta_0|\cdot$

Now if θ^1_0 is set to an arbitrarily chosen point $\theta^1_0 = (\mathbf{z}, \mathbf{y}, \beta_0, \beta_1)$, and $\theta^2_0 = (\widetilde{\mathbf{z}}, \mathbf{y}, \beta_1, \beta_0)$, note that θ^1_0 and θ^2_0 will belong to different modal regions for all partitions $P \in \mathscr{P}$. Furthermore, the coupling mechanism will make sure that $\theta^2_i = (\widetilde{\mathbf{z}}^{(i)}, \mathbf{y}^{(i)}, \beta_1^{(i)}, \beta_0^{(i)})$ when $\theta^1_i = (\mathbf{z}^{(i)}, \mathbf{y}^{(i)}, \beta_0^{(i)}, \beta_1^{(i)})$, for every iteration $i = 1, \ldots$, and hence θ^1_0 and θ^2_0 will belong to different modal regions for all partitions $P \in \mathscr{P}$ in every iteration. The combined sample consisting of $\{\theta^1_i\}$ and $\{\theta^2_i\}$ will clearly be sampled with the right proportions from each modal region but since there is a functional dependence between θ^1_i and θ^2_i for each i, there will also be unwanted dependencies in the sample. The solution we employ is to taking a sub sample form the combined sample. More specifically, we construct a sample $\{\theta^*_i\}$, where for each i, we perform a Bernoulli experiment with probability $1/2$. If this is a success, we set $\theta^*_i = \theta^1_i$, and otherwise $\theta^*_i = \theta^2_i$.

Of course, there is no need to actually run two copies and to take the sub-sample afterwards. The sample $\{\theta^*_i\}$, can be obtained by running a single chain and after each iteration we swap places for θ^*_{3i} and θ^*_{4i}, and set θ^*_{1i} equal to $\widetilde{\theta}^*_{1i}$, with probability $1/2$. This is what we mean by switching devise. To keep track of the instances in which a jump from one region to another has taken place we introduce an sequence of indicator functions $\{I_i\}$, where $I_i = 1$ if a jump was made in iteration i and 0 otherwise. If the regions are distinct, an indicator of modal region can be constructed. Set $A^{(0)} = 0$, and let $A^{(i)} = A^{(i-1)} + I_i$ modulo 2, and if the modal regions are sufficiently distinct there will exist a partition $P \in \mathscr{P}$, such that $P(\theta^*_{1i}) = A^{(i)}$, for all i.

We might point out that when implementing the Gibbs sampler, there is no need to put it explicitly into the form described in this section, i.e. we do not need to use the form with inverted cdf and the uniform random variates.

5. Model selection

Let $\mathscr{M}$ be a collection of models reflecting different hypothesis regarding the data generation process. Each model $M \in \mathscr{M}$ is characterised by a sampling probability mass function $p(\mathbf{x}|\theta,M)$, and a model specific set of parameters $\theta \in \mathbf{\Theta}_M$. In our case we limit the class of models to a subset of the models whose sampling probability mass function can be put in the form (2.3). We are however allowed to make comparisons with other models for the same data structure with other sampling distributions, a subject that we return to in the next section. The parameter space of the models here considered is $\mathbb{R}^{p_M} \times \mathbb{R}^{p_M} \times \{0,1\}^{n(n-1)}$, where p_M may varies over models. When making model selection we have from standard statistical theory that the marginal likelihood, $m(\mathbf{x}|M) = \sum_{\mathbf{z}\in\mathscr{Z}} \int_\Theta p(\mathbf{x}|\beta,\mathbf{z},M)\pi(\beta,\mathbf{z},|M)d\beta$, of a model M is proportional to the posterior probability of model M given data when all models considered are given equal probability a priori. Typically the model with highest marginal likelihood is chosen as the best model (c.f. e.g. Raftery, 1996b). To obtain the marginal likelihood is not always straightforward but Chib (1995) showed how an estimate of the posterior ordinate can be calculated from the Gibbs output and inserted in the basic marginal likelihood identity to obtain an estimate of the marginal likelihood.

The basic marginal likelihood identity (BMI) is given by

$$m(\mathbf{x}|M) = \frac{L(\beta_1^*,\beta_0^*,\mathbf{z}^*;\mathbf{x})\pi(\beta_1^*,\beta_0^*,\mathbf{z}^*)}{\pi(\beta_1^*,\beta_0^*,\mathbf{z}^*|\,\mathbf{x})},$$

which follows from solving Bayes Theorem for the marginal likelihood. Following Chib (1995), we first note that for any arbitrary point $\beta_1^*,\beta_0^*,\mathbf{z}^*$ in the parameter space

$$\pi(\beta_1^*,\beta_0^*,\mathbf{z}^*|\,\mathbf{x}) = \pi(\mathbf{z}^*|\,\beta_1^*,\beta_0^*,\mathbf{x})\pi(\beta_1^*,\beta_0^*|\,\mathbf{x}). \tag{5.1}$$

The first factor on the right hand side is immediately available as noted above. The MCMC estimate of the second factor is

$$\widehat{\pi}(\beta_1^*,\beta_0^*|\,\mathbf{x}) = G^{-1}\sum_{g=1}^{G}\pi(\beta_1^*,\beta_0^*|\,\mathbf{x},\mathbf{z}^{(g)},\mathbf{y}^{(g)}), \tag{5.2}$$

where $\mathbf{z}^{(g)},\mathbf{y}^{(g)}$ are sample points from the Gibbs sample.

Note that this procedure requires little extra programming in addition to what is needed for implementing the Gibbs sampler.

6. Analysis

To illustrate the model selection procedure, we fit a number of models to Krackhardt's (1987) high-tech managers. The 21 manages each gave their version of who went to whom for advice. For each individual the age (AGE), position in the hierarchy (HIE), length of tenure (TEN), and department were recorded (DEP). There were four different departments and hierarchal position was coded as one if

the individual was a vice president or CEO, and zero otherwise. Out of the individual attributes we construct covariates corresponding to different combinations of attributes of perceiver (i), sender (j) and receiver (k).

Two possibly important factors in the perception of advice-giving, could be structural proximity and similarity. By the former is meant, what role does the perceiver have in relation to the dyad in terms of structure, e.g. perceiving own ties as opposed to ties of others. When perceiving own ties there could be some form of prestige bias - you either inflate your own position in the network and/or over report/under report your activity. Similarity could be influential on two levels. Firstly you could be biased towards over-reporting ties between people that are similar with respect to certain characteristics and secondly, and more important, you know more about people that are similar to you. We consider similarity with respect to the categories: age, tenure, position in the corporation hierarchy and department.

We list a few potentially interesting covariates that can be include in the models, and give commonsensical motivations meant to capture the ambiguities of their "effects". All are various combinations of informant, sender and receiver attributes and indices are only used when we need to distinguish between informant-sender and informant-receiver. Additionally, all listed effects, bar the intercept, depend on the informant.

INT**:** Intercept, representing the average level hits and false alarms

AGE_i**:** There is no specific motivation behind including age except perhaps seniority

$DA_j = AGE_i - AGE_j$**:** (Dis-) Similarity between perceiver and advice seeker with respect to age

$DA_k = AGE_i - AGE_k$**:** The effect of (dis-) similarity in age on the perception of receiving advice

TEN_i**:** The longer you have been in the corporation the better your knowledge of it or, alternatively the more likely you are to be stuck in old hierarchies

$DT_j = TEN_i - TEN_j$**:** The effect of similarity with respect to length of tenure on perception of advice seeking

$DT_k = TEN_i - TEN_k$**:** The effect of similarity with respect to length of tenure on perception of advice giving

HIE_i**:** Actors in high positions know their corporation or they are too far removed

$DE_j = HIE_i * HIE_j$**:** Do people know who seeks advice on the same level of the hierarchy (upper)

$DE_k = (1 - HIE_i)(1 - HIE_k)$**:** Do people know who is sought for advice on the same level of the hierarchy (lower)

$DP_j = 1(DEP_i = DEP_j)$**:** You know who people in your own department consult, or you exaggerate the networking

$DP_k = 1(DEP_i = DEP_k)$**:** You know what people in your department are consulted

$DAT = (DA_j^2 + DT_j^2)/c$**:** The combined effect of age and tenure dissimilarity (with a suitably chosen constant c)

$EGO_j = 1(i = j)$**:** You are more accurate/biased when perceiving own advice structure

$EGO_k = 1(i = k)$**:** You are more accurate/biased when perceiving who receives advice from you

To test the sensitivity of the results to misspecification of typical observation, typical covariate vectors $\mathbf{w}_{max}$, $\mathbf{w}_{mean}$, and $\mathbf{w}_{min}$, are used. The first type, $\mathbf{w}_{max}$, models ignorance from the point of view of an (imaginary) informant-dyad combination whose corresponding covariates are maximally informative (the amount and nature of information contributed by an observation is naturally a more complicated matter, strictly speaking; "informative" is only meant to be interpreted in the sense of the induced priors as described in Section 3). Setting $\xi = 1$, the induced priors of H_{max} and F_{max} are uniform. Observations with less informative covariate vectors will have their induced priors on H and F more concentrated around 1/2 (c.p. going form $\xi = 1$ to $\xi = .2$ in Figure 1). The minimally informative typical observation is set such that H_{min} and F_{min} are uniform for the corresponding covariate vector, $\mathbf{w}_{min}$. The more informative the covariates are for other observations relative to the minimal covariate vector, the more u-shaped they will be (since higher absolute covariate values lead to higher variance, again c.p. Figure 1). For the "mean" prior, an average covariate vector $\mathbf{w}_{mean}$ was used for reference, with uniform induced priors on the hit and false alarm probabilities. The covariate vectors used were

$$\mathbf{w}_{max} = (1, \max_{ijk}\{|w_{2,ijk}|\}, \ldots, \max_{ijk}\{|w_{p,ijk}|\})',$$

$$\mathbf{w}_{min} = (1, \min_{ijk}\{|w_{2,ijk}| : |w_{2,ijk}| > 0\}, \ldots, \min_{ijk}\{|w_{p,ijk}| : |w_{p,ijk}| > 0\})',$$

and

$$\mathbf{w}_{mean} = \left(1, \frac{\sum_{ijk}|w_{2,ijk}|}{n(n-1)m}, \ldots, \frac{\sum_{ijk}|w_{p,ijk}|}{n(n-1)m}\right)'.$$

One might criticize these choices of typical observations covariates seeing as many covariates are dichotomous, and hence the priors on the coefficients based on $\mathbf{w}_{max}$ and $\mathbf{w}_{min}$, will be equal for some coordinates. To test the sensitivity of prior variance, three scaled up priors were also used, where the mean vector was the reference covariate vector but with $\xi = p\lambda$, for $\lambda = 1, 10, 20$.

An arbitrary subset $M_1, \ldots, M_{27} \subset \mathscr{M}$ of models were fitted. The resulting marginal log-likelihoods are given in Table 1. For comparisons across models Kass and Raftery (1995) review a few rules of thumb for Bayes factors on the logarithmic scale. The marginal likelihoods are surprisingly stable over different prior specifications (as can be seen in Figure 2 there are quite drastic differences in scale between them). In Table 1, the highest values are in bold face and the choice seems to be between model 24 and 25. The former includes the effect of working in the same department for perceiver and sender. It is interesting to compare the evaluated marginal likelihoods from the models fitted here to those of the models fitted in Koskinen (2002b) to the same material. The marginal log-likelihoods of

M	DA_j	DA_k	TEN_i	DT_j	DT_k	HIE_i	DE_j	DE_k	DP_j	DP_k	DAT	EGO_j	EGO_k	max	mean	min	$\lambda = 1$	$\lambda = 10$	$\lambda = 20$
1														-4731	-4731	-4731	-4731	-4733	-4734
2						*					*			-4709	-4712	-4712	-4714	-4721	-4723
3											*			-4723	-4726	-4726	-4726	-4730	-4732
4						*								-4718	-4719	-4718	-4720	-4724	-4725
5												*	*	-4658	-4662	-4658	-4664	-4670	-4672
6									*	*				-4710	-4712	-4710	-4714	-4720	-4722
7				*	*									-4713	-4717	-4726	-4719	-4725	-4727
8			*											-4726	-4727	-4732	-4728	-4732	-4734
9	*	*												-4716	-4719	-4729	-4721	-4728	-4730
10							*	*						-4591	-4591	-4594	-4590	-4595	-4597
11											*	*	*	-4655	-4660	-4658	-4664	-4672	-4675
12				*			*					*	*	-4625	-4630	-4635	-4635	-4645	-4649
13				*								*	*	-4649	-4654	-4656	-4657	-4666	-4669
14							*					*	*	-4640	-4644	-4644	-4647	-4657	-4741
15								*				*	*	-4526	-4528	-4526	-4529	-4537	-4540
16								*	*			*	*	-4525	-4527	-4524	-4530	-4540	-4543
17								*		*		*	*	-4514	-4517	-4512	-4519	-4529	-4532
18								*	*	*		*	*	-4516	-4519	-4514	-4522	-4535	-4539
19				*				*		*		*	*	-4490	-4492	-4492	-4495	-4507	-4511
20				*				*	*			*	*	-4498	-4499	-4500	-4503	-4514	-4519
21				*			*	*	*	*		*	*	-4470	-4465	-4459	-4465	-4481	-4487
22				*			*	*		*		*	*	-4467	-4463	-4458	-4462	-4474	-4479
23							*	*		*		*	*	-4498	-4498	-4495	-4499	-4511	-4516
24	*			*			*	*	*	*		*	*	-4459	**-4454**	**-4453**	-4454	-4473	-4480
25	*			*			*	*		*		*	*	**-4454**	**-4454**	-4454	**-4451**	**-4465**	**-4473**
26	*						*	*	*	*		*	*	-4466	-4464	-4458	-4464	-4479	-4484
27	*			*			*	*		*		*	*	-4469	-4464	-4464	-4462	-4478	-4483

TABLE 1. Predictors (indicated by *) for different models fitted to Krackhardt's high-tech managers with marginal likelihood's for different prior specifications. The intercept is included in all models

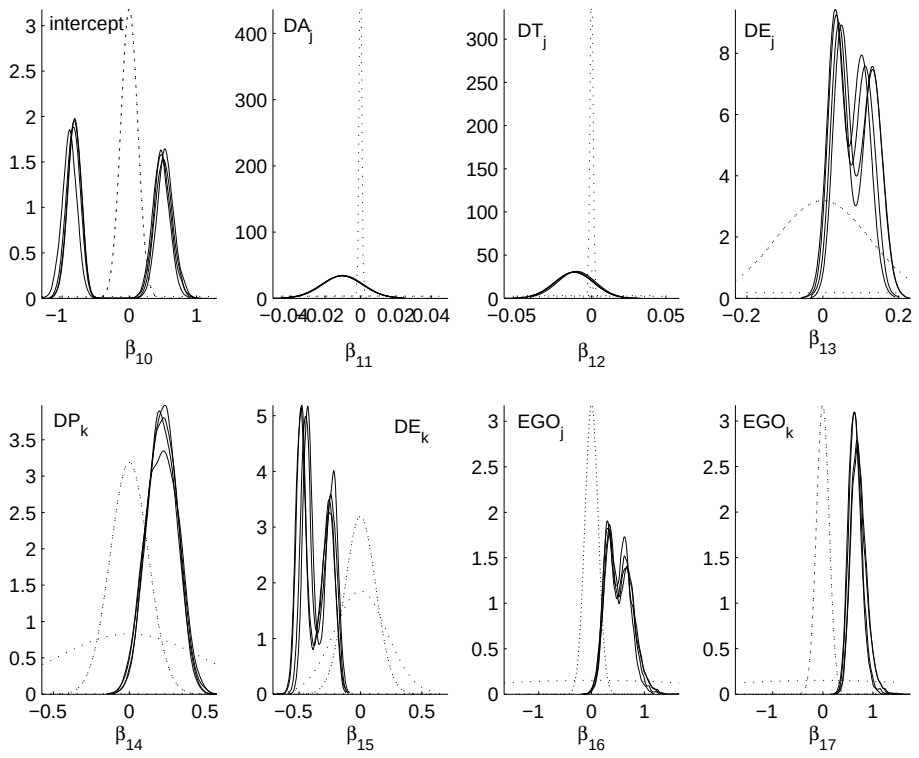

FIGURE 2. Comparison of marginal prior $(\cdots)$ and posterior $(-)$ densities of $\beta_{1,v}$ for model specification 25

the six models fitted there ranged from -4506 to -3928. The model with highest posterior probability had parameters corresponding to those of model 5 in Table 1. Although model 5 is the second best among the models with 6 parameters (models 1 through 10; not counting the additional $n(n-1)$ parameters corresponding to $\mathbf{z}$), the marginal log-likelihood of between -4672 and -4658 is small in comparison to that of the model selected in Koskinen (Ibid), which was -3928. The latter however, had $3 \times 2 \times m = 126$ parameters, one parameter for each perceiver $i \in \mathscr{I}$ and case $EGO_j = 1$, $EGO_k = 1$, $EGO_j \times EGO_k = 0$, conditionally upon whether $Z_{jk} = 1$ or not. With this in mind, the model (-s) selected here, 25 (24), with 8×2 parameters does remarkably well in comparison. That the model with 126 parameters is better supported by data would indicate that there are extreme observations if only the ego-effects are considered (which indeed seems to be the case judging by the analysis in Koskinen, Ibid). Also, the list of models fitted here is in no way exhaustive and the combination of effects were chosen arbitrarily for the sole purpose of illustration. Thus, it is quite possible that some combination of effects might explain away the extreme observations that were not explained by model 5.

In Figure 2 the posteriors for model 25 are plotted along with the priors. What is interesting is the distributions for DA_j and DT_j. The priors that are tightly concentrated around 0, the spike-like curves, are the ones set with $\mathbf{w}_{max}$ and $\mathbf{w}_{mean}$

		$s=0$		$s=1$	
predictor	par.	mean	std	mean	std
average level	β_{s1}	$-.896$	.109	.495	.129
informant-sender age dissimilarity	β_{s2}	$-.010$	.008	$-.011$	.005
informant-sender tenure dissimilarity	β_{s3}	$-.009$	.009	$-.015$	.008
informant-sender hierarchy similarity	β_{s4}	.049	.020	.131	.025
informant-receiver dep. similarity	β_{s5}	.288	.082	.144	.078
informant-receiver hie. similarity	β_{s6}	$-.273$	.068	$-.488$	.039
informant is sender	β_{s7}	.732	.165	.330	.112
informant is receiver	β_{s8}	.691	.163	.756	.147

TABLE 2. Posterior means and standard deviations for parameters in model 25 ($\lambda = 20$) for Krackhardt's high-tech managers conditional on partition induced by the switching mechanism

as reference covariates. For all eight parameters, the marginal posterior distributions are not very different across models. It is further instructive to study the marginal posterior of $\beta_{1,1}$. The marked bimodal shape suggests that the modal regions are distinct. Indeed, if we consider the partition that is induced by the switching mechanism, the sign of $\beta_{0,1}$ is always the opposite of the sign of $\beta_{1,1}$, conditional on the partition. A good indication of whether the switching mechanism induces a true partition of $\mathscr{Z}$ is given by the marginal posterior edge probabilities in Figure 3. The posterior of the parameter corresponding to EGO_k, illustrates to some extent the dangers of relying on restrictions or highly informative priors for achieving uni-modality, since the marginal posterior does not distinguish between the case when there is an edge present and when there is not, conditional on all other parameters. Hence it is unclear whether a restriction on or a highly informative prior for this parameter would have been enough for achieving joint uni-modality.

The marginal posterior distribution of the parameters in model 25 conditional on the partition induced by the switching mechanism are given in Figure 4. The corresponding posterior means and standard deviations are given in Table 2. Looking at Figure 4 it seems as if the model could be improved upon by removing the effects DA_j and DT_j. However, when we inspect the bivariate posterior of these parameters in Figure 5, it turns out the origin is just included in a 99% highest posterior density region (for $\beta_{1,2}$ and $\beta_{1,3}$ the bivariate is even further removed from the origin).

Assuming the perspective of accuracy versus in-accuracy when interpreting Table 2 and Figure 4, the partition of the parameter space corresponds to a scenario when the actors are in general accurate. It seems as if actors are not likely to report other actors who are junior to them as givers of advice, regardless of of the "true state of affairs". The interpretation for the effect of seniority with respect to length of tenure is similar. In terms of the model, the effect of spatial proximity is reflected in that actors recognizes who in their department seeks advice. There is however a tendency towards false positives, i.e. reporting co-workers in the same

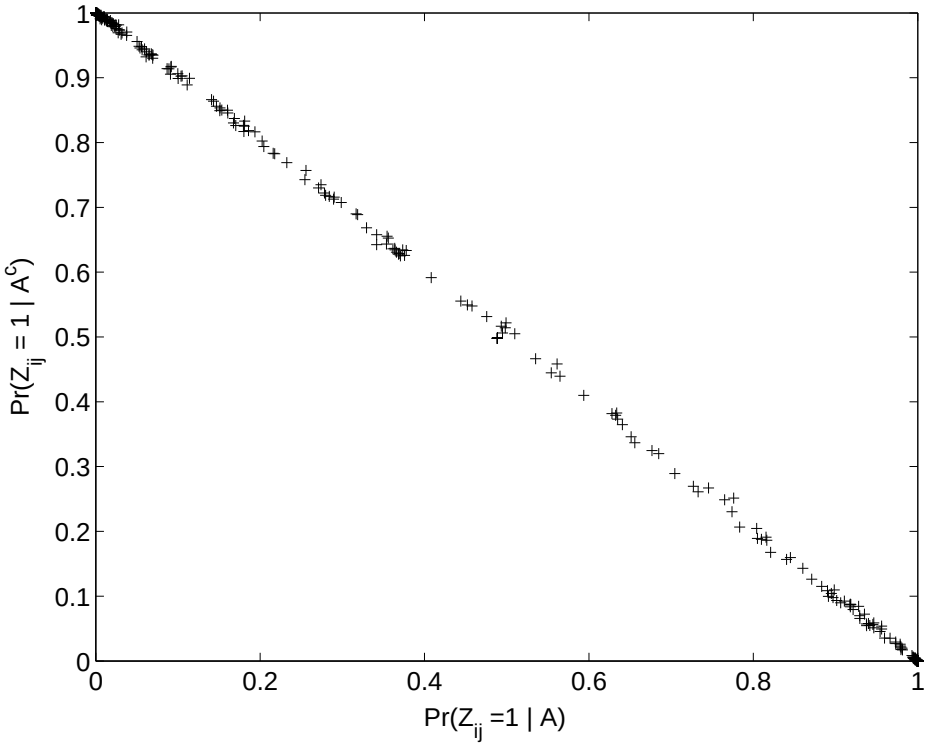

FIGURE 3. Marginal posterior probability of and edge (i, j), for $(i, j) \in V^{(2)}$, conditional on partition induced by the switching mechanism for model $25(\lambda = 20)$

department as seekers of advice when they in fact are not. Considering that there are correlations such as those between DA_j and DT_j, interpretations are perhaps best done through analysing the predictive probabilities of false positives and hits for different configurations.

APPENDIX A. THE INDUCED PRIOR ON H AND F

For a fixed covariate vector $\mathbf{w}$, make the transformation $V = \mathbf{w}'\beta_1$. With reference priors with respect to $\mathbf{w}$, we have from the properties of the normal distribution that the induced prior on V is $N(0, \xi^2)$. Denote by $\varphi(\cdot)$ and $\Phi(\cdot)$ the probability distribution function and cumulative distribution function, respectively, of a standard normal variate. The pdf of $H = \Phi(V)$ has the form

$$f_H(h) = \varphi\left(\Phi^{-1}(h)/\xi\right)\left|\frac{d}{dh}\Phi^{-1}(h)\right|, \tag{A.1}$$

where the derivative equals $\varphi(\Phi^{-1}(h))$, and insertion in (A.1) yields

$$f_H(h) = \xi^{-1}\exp\left\{-\frac{1}{2}\left(\xi^{-2}-1\right)\left[\Phi^{-1}(h)\right]^2\right\}. \tag{A.2}$$

For general $V \sim N(\mu, \xi^2)$, with $H = \Phi(V)$, the expected value of H is straightforward to calculate. Firstly write $V = \mu + \xi W$, where W is a standard normal

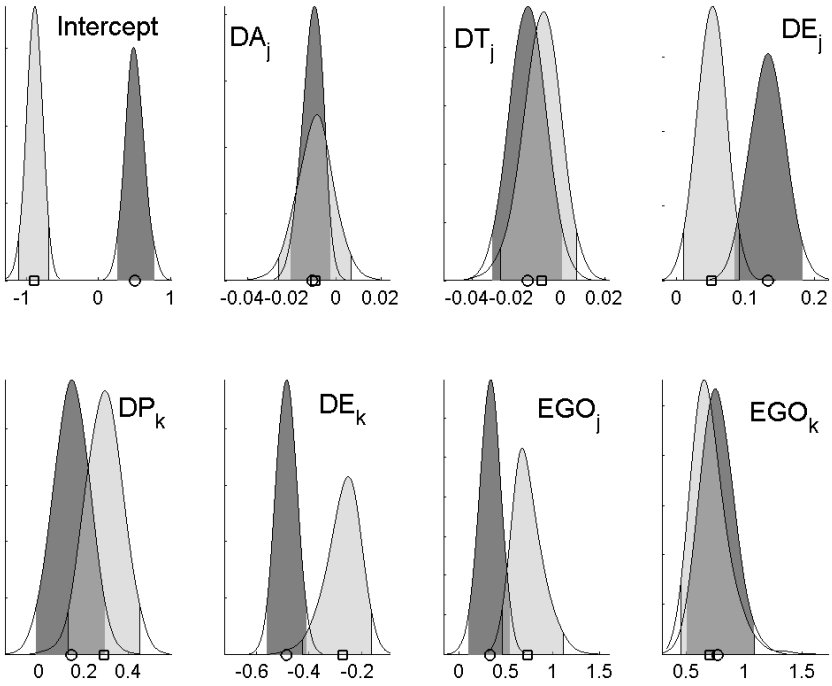

FIGURE 4. Marginal posterior densities for model 25 ($\lambda = 20$) conditional on a region in $\mathscr{Z}$ induced by the switching mechanism, with 95% Credibility intervals for β_{1v} (β_{0v}), dark (pale) shades, and posterior means $\circ$ ($\square$)

variate. Secondly, note that

$$H = \Phi(\mu + \xi W) = \Pr(U \leq \mu + \xi W | W),$$

in which U is a standard normal variate which is independent of W. From this follows that the expected value of H is given by

$$E(H) = \Pr(U - \xi W \leq \mu) = \Phi\left(\frac{\mu}{\sqrt{1+\xi^2}}\right).$$

Note that $E(H) = 0$ if and only if $\mu = 0$.

A simple (and not very sophisticaded) way for finding an approximation to the density (A.2) by a beta density is by approximating the inverse standard normal cdf. An upper bound for the standard normal c.d.f. is given by Zelen and Severo (1984)

$$\Phi(x) \leqslant \frac{1}{2} + \frac{1}{2}\sqrt{1 - e^{2x^2/\pi}}, \qquad (x > 0)$$

and solving for x

$$x^2 \geqslant -\frac{\pi}{4}\ln\left(-4\Phi(x)^2 + 4\Phi(x)\right).$$

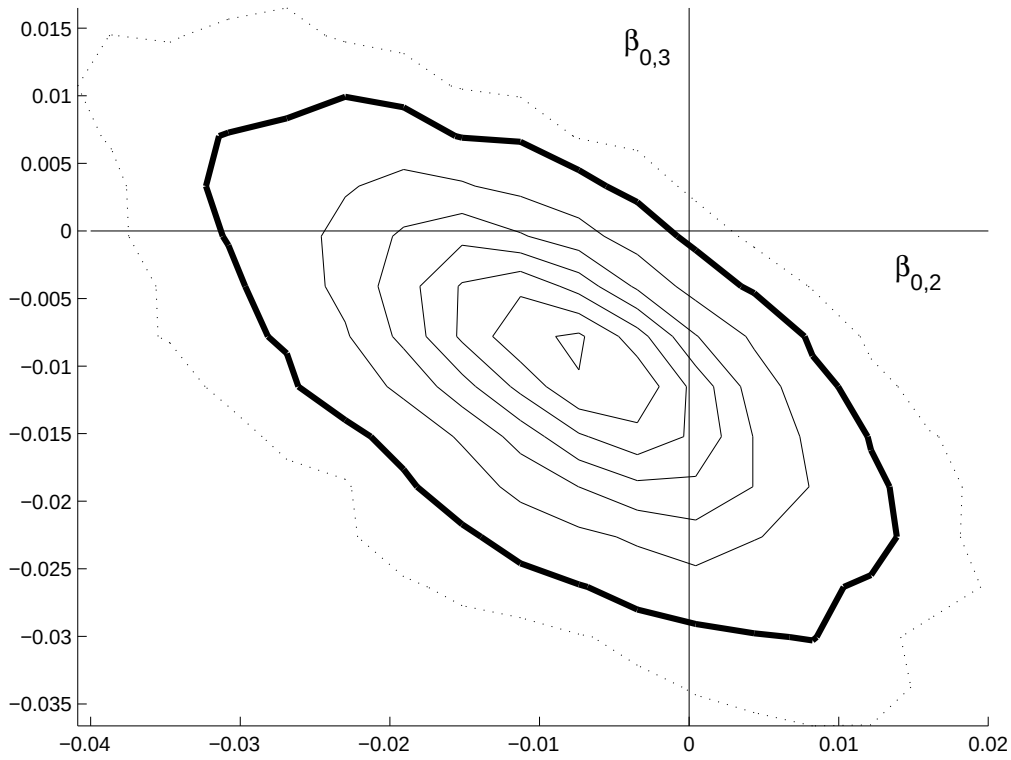

FIGURE 5. Level curves of bivariate posterior for parameters corresponding to DA_j and DT_j in model 25 ($\lambda = 20$), conditional on partition induced by switching mechanism with 95% (thick line) and 99% (dotted) HPD regions

We now have an upper bound for the p.d.f.

$$\begin{aligned} f_H(h) &= \xi^{-1} \exp\left\{-\frac{1}{2}\left(\xi^{-2}-1\right)\left\{\Phi^{-1}(h)\right\}^2\right\} \\ &\leqslant \xi^{-1} \exp\left\{\frac{1}{2}\left(\xi^{-2}-1\right)\frac{\pi}{4}\ln\left(-4h^2+4h\right)\right\} \\ &= \xi^{-1} \exp\left\{a\ln\left(-4h^2+4h\right)\right\} \\ &= \xi^{-1} 4^{a-1} h^{a-1}(1-h)^{a-1}, \qquad a = \frac{\pi}{4}\left(\xi^{-2}-1\right)+1, \end{aligned}$$

with equality for $\xi = 1$, and $h = 1/2$ for all ξ. We recognize $h^{a-1}(1-h)^{a-1}$ as proportional to a Beta(a, a) density.

Tentative simulation results show that the first 9 moments are almost perfectly matched for the beta approximation and $\xi \leq 1.5$. To calculate the moments exactly, one may apply the same technique as was used for calculating the expected value.

References

Albert, J. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88, 669–679.

Batchelder, W., Kumbasar, E., and Boyd, J. (1997). "Consensus analysis of three-way social network data." *Journal of Mathematical Sociology*, 22, 29–58.

Bernard, H., Killworth, P., and Sailer, L. (1980). "Information accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data." *Social Networks*, 2, 191–218.

Bondonio, D. (1998). "Predictors of accuracy in perceiving informal social networks." *Social Networks*, 20, 301–330.

Butts, C. T. (2003). "Network inference, error, and informant (in)accuracy: a Bayesian approach." *Social Networks*, 25, 103–140.

Casciaro, T. (1998). "Seeing things clearly: social structure, personality, and accuracy in social network perception." *Social Networks*, 20, 331–351.

Casciaro, T., Carley, K., and Krackhardt, D. (1999). "Positive affectivity and accuracy in social network perception." *Motivation and Emotion*, 23, 285–306.

Chib, S. (1995). "Marginal likelihood from the Gibbs output." *Journal of the American Statistical Association*, 90, 1313–1321.

Dellaportas, P. and Forster, J. J. (1999). "Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models." *Biometrika*, 86, 3, 615–633.

Gelfand, A. and Smith, A. F. (1990). "Sampling-based approaches to calculating marginal densities." *Journal Of The American Statistical Association*, 85, 398–409.

Gilks, W. and Roberts, G. O. (1996). "Strategies for improving MCMC." In *Markov Chain Monte Carlo in Practice*, eds. W. Gilks, S. Richardson, and D. J. Spiegelhalter. London: Chapman & Hall.

Johnson, V. E. (1996). "Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths." *J. Amer. Statist. Assoc.*, 91, 433, 154–166.

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

Killworth, P. and Bernard, H. (1979). "Information accuracy in social network data III, or: A comparison of triadic structure in behavioural and cognitive data." *Social Networks*, 2, 19–46.

Koskinen, J. (2001). "Aggregation of perceived social networks." Paper presented at the Sunbelt XXI conference, Budapest (Hungary).

— (2002a). "Bayesian analysis of cognitive social structures with covariates." Research Report 2002:3, Department of Statistics, Stockholm University.

— (2002b). "Bayesian analysis of perceived social networks." Research Report 2002:2, Department of Statistics, Stockholm University.

Koskinen, J. H., Jansson, I., and Spreen, M. (2002). "The role of perceptional data in sampling large networks from hidden populations." In *Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics;*

A Festschrift in honour of Ove Frank, ed. J. Hagberg, 56–73. Stockholm: Dept. of Statistics, Stockholm University.

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

Kumbasar, E., Romney, A. K., and Batchelder, W. H. (1994). "Systematic biases in social perception." *American Journal of Sociology*, 100, 477–505.

McCulloch, R. and Rossi, P. E. (1994). "An exact likelihood analysis of the multinomial probit model." *Journal of Econometrics*, 64, 207–240.

Moreno, J. (1934). *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama*. Washington, D.C.: Nervous and Mental Disease Publishing Co. Reprinted in 1978 (Third Edition) by Becon House, inc., Beacon, NY.

Newcomb, T. (1961). *The Acquaintance Process*. New York: Holt, Rinehart and Winston.

O'Hagan, A. (1995). "Fractional Bayes factors for model comparison." *J. Roy. Statist. Soc. Ser. B*, 57, 1, 99–138. With discussion and a reply by the author.

Pattison, P. (1994). "Social cognition in context – some applications of social network analysis." In *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*, eds. S. Wasserman and J. Galaskiewicz, 79–109. Thousand Oaks: Sage.

Raftery, A. E. (1996a). "Approximate Bayes factors and accounting for model uncertainty in generalised Linear models." *Biometrika*, 83, 2, 251–266.

— (1996b). "Hypothesis testing and model selection." In *Markov Chain Monte Carlo in Practice*, eds. W. Gilks, S. Richardson, and D. J. Spiegelhalter. London: Chapman & Hall.

Salabasis, M. and Villani, M. (2000). "Panel regression with unobserved classes." SSE/EFI Working Paper Series in Economics and Finance 353, Stockholm School of Economics.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). "Bayes factors for linear and log-linear models with vague prior information." *J. Roy. Statist. Soc. Ser. B*, 44, 3, 377–387.

Zelen, M. and Severo, N. C. (1984). "Probability Functions." In *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, eds. M. Abramowitz and I. A. Stegun, A Wiley-Interscience Publication, xiv+1046. New York: John Wiley & Sons Inc. Reprint of the 1972 edition, Selected Government Publications.

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden
E-mail address: johan.koskinen@stat.su.se