MODEL SELECTION FOR LONGITUDINAL SOCIAL NETWORKS

JOHAN KOSKINEN

ABSTRACT. This paper concerns model selection for a class of continuous-time Markov chains for modeling longitudinal social networks. Many models of this kind have been proposed in the literature (Holland and Leinhardt, 1977a,b; Wasserman, 1977, 1980b,a; Snijders, 1996, 2001) but until recently likelihood-based inference has only been explored under the assumption of dyad independence. Using data augmentation it was shown in Koskinen (2004b) how the class of continuous-time Markov chains open to likelihood-based inference can be extended to entail more complex dependence structures. Arguably, the main theoretical motivation behind models for longitudinal social networks is to infer what components are important in the dynamics of social interaction. This calls for statistical procedures for testing hypothesis, something which in the absence of procedures for conducting model selection, is limited to inspection of posterior credibility regions. The Bayesian paradigm is well suited for model selection but the relative complexity of this class of models prevents the use of any standard techniques for calculating the relevant quantities (the evaluated likelihood and marginal likelihood respectively). Although an analytically tractable form for the likelihood function is not strictly necessary for performing parameter inference (c.f. Koskinen, 2004b), most model selection techniques rely heavily on the assumption that the likelihood can easily be evaluated. We identify a family of models, with the property that they have the reciprocity model Wasserman (1977) as special case, for which the scheme of Chib and Jeliazkov (2001) for estimating the marginal likelihood can be adapted, thus providing the posterior distribution over a set of models. If the analysis is restricted to comparisons between nested models, the likelihood function does not have to be evaluated and model selection need not be restricted to models with the reciprocity model as a special case. The procedure is illustrated using van de Bunt's (1999) freshmen students, a stochastic actor oriented model (Snijders, 1996, 2001, 2004) and partial Bayes factors.

1. INTRODUCTION

A class of stochastic actor-oriented models was proposed by Snijders (1996, 2001), and later extended by Koskinen (2004b), that built upon the notion that change in social networks is driven by sequences of incremental changes. Most realistic models for change in social networks would incorporate an element of structural evaluation of the network by the actors. To be more precise, it is reasonable to assume that whenever an aspect of a social network changes it is likely that the present structure of the network is a major influence on what direction

Key words and phrases. Longitudinal social networks; Bayesian model selection; Partial Bayes factor.

the change is taking. When the structure itself changes, the basic data for decision making is in constant flux, each change redefining the environment. As a consequence, although the order of changes in between two observation on a social network is unknown it cannot be neglected. These changes are explicitly modeled in the inference procedure proposed in Koskinen (2004b) turning parameter estimation into a conventional Bayesian inference procedure. Comparisons across models are however complicated by the implicit dependence on the latent evolution and the typically high levels of posterior correlation between parameters. Although it in Koskinen (2004b) was suggested that predictive odds (of various types of transitions) can be used for interpreting the results, the nature of the interdependencies of the statistics corresponding to the parameters complicates matters.

2. Continuous-time Markov Chains for Networks

Here follows a brief characterization of the general class of models. We restrict attention to binary social network data. Denote by $V = \{1, \dots, n\}$ a fixed set of actors with a relation $\mathscr{N} \subseteq V \times V$. A (di-) graph on V with (arc) edge set $E \subseteq \mathscr{N}$, can be described by a collection $X = (X_e : e \in \mathscr{N})$ of (arc) edge indicators, $X_e = \mathbf{1}\{e \in E\}$. Observe that X is only defined for the non-redundant pairs of actors and does not, as is standard practice in social network analysis for un-directed networks, include structural zeros for reflexive ties or other un-defined ties. Let $\mathscr{X} = \{0,1\}^N$, $N = |\mathscr{N}|$, be the space of all binary adjacency matrices.

We consider a continuous-time Markov chain $\{X(t)\}_{t\in\mathbb{R}}$, with outcome space $\mathscr{X}$. We have that for any element $x \in \mathscr{X}$, and any pair of time points $t_a < t_b$,

$$\Pr\left(X\left(t_b\right) = x \middle| X\left(t\right) = y\left(t\right), \text{ for all } t \leqslant t_a \right) = \Pr\left(X\left(t_b\right) = x \middle| X\left(t_a\right) = y\left(t_a\right) \right). \tag{2.1}$$

The infinitesimal generator is a function $q : \mathscr{X}^2 {\rightarrow} \mathbb{R}$ such that

$$\begin{aligned} q\left(x,y\right) &= \lim_{dt \searrow 0} \frac{\Pr\left(X\left(t+dt\right) = y \middle| X\left(t\right) = x\right)}{dt}, \quad \text{for } x \neq y \\ q\left(x,x\right) &= \lim_{dt \searrow 0} \frac{1 - \Pr\left(X\left(t+dt\right) = x \middle| X\left(t\right) = x\right)}{dt}. \end{aligned}$$

For $s < t$, and $x, y \in \mathscr{X}$, denote the transition probability

$$P_{xy}^{t-s} = \Pr\left(X\left(t\right) = y \middle| X\left(s\right) = x\right).$$

The transition matrix $P^t = (P_{xy}^t)_{x,y\in\mathscr{X}}$, is completely determined by the intensity matrix $Q = \left(q\left(x,y\right)\right)_{x,y\in\mathscr{X}}$ through the identity $P^t = e^{tQ}$.

For the elements of $\mathscr{X}$ define the Hamming metric

$$|x - y| = \sum_{e\in\mathscr{N}} |x_e - y_e|,$$

and define $x(\widetilde{e})$ as the matrix that differs from x in exactly the element $e \in \mathscr{N}$. In the following we will only consider processes such that

$$q\,(x, y) = \begin{cases} q_e(x) & \text{if } y = x(\widetilde{e}) \\ 0 & \text{if } |x - y| > 1 \\ -\sum_{e \in \mathscr{N}} q_e(x) & \text{if } x = y. \end{cases}$$

This can be interpreted as a process on the vertices of the binary N-cube, i.e. a graph $\mathscr{G} = (\mathscr{X}, \mathscr{E})$, $\mathscr{E} = \{(x, y) \in \mathscr{X} \times \mathscr{X} : |x - y| = 1\}$, that jumps between adjacent vertices. If we set $q_e(x) = N^{-1}$, we obtain the random walk on $\mathscr{X}$, and we have (see e.g. Aldous, 1983) that

$$P^t_{xy} = 2^{-N}(1 + e^{-2t/N})^{N-|x-y|}(1 - e^{-2t/N})^{|x-y|}, \tag{2.2}$$

for all $x, y \in \mathscr{X}$. Consider now discriminating between changing an element of x from zero to one and changing from one to zero, the first corresponding to adding an edge and the second to removing an edge. To obtain the so called independent arcs model (Wasserman, 1977), we set $q_e(x)$ equal to λ_0 if $x_{ij} = 0$, and λ_1 if $x_e = 1$. For $x, y \in \mathscr{X}$, define

$$N_{hk} = N_{hk}(x, y) = \sharp\,\{e|\, x_e = h, y_e = k\}\,,$$

and the transition probabilities are given by

$$P^t_{xy} = \binom{N_{00} + N_{01}}{N_{01}}\binom{N_{11} + N_{10}}{N_{11}}\xi_0(t)^{N_{01}}(1 - \xi_0(t))^{N_{00}}\xi_1(t)^{N_{11}}(1 - \xi_1(t))^{N_{10}},$$

where

$$\xi_a(t) = \frac{\lambda_a}{\lambda_a + \lambda_{1-a}} + \frac{(-1)^{1-a}}{\lambda_{1-a}} \exp\left\{-\left(\lambda_a + \lambda_{1-a}\right) t\right\},$$

for $a = 0, 1$.

A third model with an explicit form for P^t_{xy} is the reciprocity model (Wasserman, 1977, 1980b,a; Leenders, 1995b,a), defined for di-graphs by

$$q_{ij}(x) = \lambda_0 + \mu_0 x_{ji} \qquad (x_{ij} = 0), \tag{2.3}$$

and

$$q_{ij}(x) = \lambda_1 + \mu_1 x_{ji} \qquad (x_{ij} = 1), \tag{2.4}$$

subject to the constraint that q_{ij} is positive for all $j \neq i$. The formula for P^t_{xy} requires a little more space than is available here (we refer to Leenders 1995b or Snijders 1999 for the exact expression). Note, in equations (2.3) and (2.4), the distinction that is made between reciprocated ties, $x_{ij} = x_{ji} = 1$, and non-reciprocated ties $x_{ij} \neq x_{ji}$.

3. Model Specification

We define the class of models considered in this paper from the point of view of the embedded chain of a continuous-time process on $\mathscr{X}$. The transition probabilities in the embedded chain are

$$\pi(\theta, x, y), \tag{3.1}$$

and the time spent in $x \in \mathscr{X}$ exponentially distributed with rate

$$\lambda(\theta, x). \tag{3.2}$$

It is assumed that $\pi(\theta, x, y) > 0$ for $x, y \in \mathscr{X}$ such that $|x-y| = 1$ and 0 otherwise. The $p \times 1$ parameter vector $\theta \in \Theta$ includes all unknown parameters. These two functions determine the rate functions

$$q(\theta, x, y) = \lambda(\theta, x)\, \pi(\theta, x, y)$$

which defines the generator of the continuous-time process on $\mathscr{X}$.

Assume that we have observations on the network $X(t)$ for fixed time points, $t_0 < t_1 < \cdots < t_{M-1}$. The analysis is throughout made conditional on the first observation at t_0. Because of the Markov property we can drop the notational dependency on the observation points, in order to make the notation more lucid. In the sequel we refer by t_0 and t_1 to two generic consecutive observation points, $t_0 < t_1$, and $(t_1 - t_0) = T$.

For t_1 and t_0, denote the distance between these two observations by

$$H = |X(t_1) - X(t_0)|.$$

From the construction of the mini-step, the number of mini-steps used to transform $X(t_0)$ into $X(t_1)$ must equal $m = H + 2k$, for some $k \in \{0, 1, 2, \ldots\}$.

For given m let $\mathscr{W}_m(x, y)$ be the set of all sequences $v_0, v_1, \ldots, v_m$, such that $v_0 = x$ and $v_m = y$, and $|v_{i-1} - v_i| = 1$, $v_i \in \mathscr{X}$ for $i = 1, \ldots, m$. Equivalently, $\mathscr{W}_m(x, y)$ is the set of all m-walks in the N-cube $\mathscr{G} = (\mathscr{X}, \mathscr{E})$, from x to y. Unless otherwise stated $\mathscr{W}_m = \mathscr{W}_m(x(t_0), x(t_1))$. Define also $\mathscr{S}_m = \{(u_1, \ldots, u_m) \in (0, T)^m : u_1 + \cdots + u_m < T\}$. For a fixed m, let the latent variable $w = (y, u)$, comprise $y = (y_h)_{h=0}^m \in \mathscr{W}_m$, and $u = (u_h)_{h=1}^m \in \mathscr{S}_m$, with the interpretation that y constitutes a walk from the observation $x(t_0)$ to the observation $x(t_1)$, for which the time in-between consecutive changes y_h and y_{h+1} is u_h, for $h = 1, \ldots, m$. For a given $w \in \mathscr{C}_m = \mathscr{W}_m \times \mathscr{S}_m$, we define the augmented likelihood

$$L(\theta; w, x(t_0)) = \exp\left\{-\sum\nolimits_{h=1}^{m+1} u_h \lambda(\theta, y_{h-1})\right\} \prod\nolimits_{h=1}^{m} \pi(\theta, y_{h-1}, y_h) \lambda(\theta, y_{h-1}). \tag{3.3}$$

Given observations $x(t_0)$ and $x(t_1)$ the data likelihood is given by

$$L_D(\theta; x(t_1), x(t_0)) = \int_{\mathscr{C}} L(\theta; w, x(t_0)) dw$$

where $\mathscr{C} = \cup_{k=0}^{\infty} \mathscr{C}_{H+2k}$.

A set $\mathscr{M}$ of models for x is characterized by a parameter space Θ_M and data likelihoods $L_{D,M}$. For the rest of this paper we limit our attention to a family $\mathscr{M}_{rec}$ of models "containing" the reciprocity model, and a special subclass of models. The former is defined by

Definition 1. *A process with $\pi(\theta, x, y)$ and $\lambda(\theta, x)$ defined as above with $\mathscr{N} = V^{(2)}$, belongs to $\mathscr{M}_{rec}$ if there exists a subspace $\widetilde{\Theta} \subseteq \Theta$, and a known transformation $f(\theta) = (\lambda_{00}, \lambda_{01}, \lambda_{10}, \lambda_{11})$ with the property that*

$$\lambda(\theta, x)\, \pi(\theta, x, y) = \lambda_{x_{ij}, x_{ji}},$$

and

$$\lambda_{k\ell} > 0, \textit{ for } k = 0, 1, \ \ell = 0, 1$$

for all x and y that differs from x only in the element (i, j), and $\theta \in \widetilde{\Theta}$.

(To get the definition in the same notational form as Wasserman (1980a), set $\lambda_{k0} = \lambda_k$, and $\lambda_{k1} = \lambda_k + \mu_k$, for $k = 0, 1$.) The purpose of this limitation is to assure that we can choose parameter values in a way so that there exists parameter values for which we can evaluate the data likelihood. Since the data likelihood is equivalent to the transition probability of going from one observed network to another and the transition probabilities are uniquely determined by the intensity matrix, we can for a model in $\mathscr{M}_{rec}$ evaluate the data likelihood for $\theta \in \widetilde{\Theta}$ using the explicit formula for the transition probabilities in the reciprocity model. Note that we do not require $\widetilde{\Theta}$ to contain a vector that is the maximizer of the data likelihood for given data. The usefulness of this definition will become apparent in Section 4. For social networks where x_{ij} is not defined for all ordered pairs (i, j), for example un-directed networks or bi-partite networks, we may make a similar definition but requiring instead that the models have to be expressable in terms of an independent arcs model.

Another important property of a subset of models is

Definition 2. *A subset $\mathscr{M}' \subset \mathscr{M}$ of models is said to be a pairwise nested set of models if it is possible to construct a sequence $M_1, M_2, \ldots, M_r$, containing all models in $\mathscr{M}'$, with the property that*

$$L_{D,M_i}(\theta; x(t_1), x(t_0)) = c_i(\theta, \psi) L_{D,M_{i+1}}(\psi; x(t_1), x(t_0)),$$

for some $\theta \in \Theta_{M_i}$, $\psi \in \Theta_{M_{i+1}}$, with $\pi_{M_i}(\theta), \pi_{M_{i+1}}(\psi) > 0$, and a known constant $c_i > 0$, for $i = 1, \ldots, r - 1$.

This is a weaker condition than that implied by the first definition and we see that $\mathscr{M}' \subset \mathscr{M}_{rec}$ is a sufficient condition for $\mathscr{M}'$ to be pairwise nested. By choosing for each model in $\mathscr{M}'$, to evaluate the data likelihoods for parameter values according to Definition 1, the constants (in Definition 2) for each compared pair are simply the ratio of (evaluated) likelihood functions. Note that the random walk on $\mathscr{G}$ as well as the independent arcs model are nested within the reciprocity model. The latter is obtained by setting $\mu_1 = \mu_0 = 0$, and the former by additionally setting $\lambda_1 = \lambda_0 = N^{-1}$, in the reciprocity model.

4. Parameter inference

Under the assumption that the augmented likelihood $L(\theta; w, x(t_0))$ is cheap to evaluate for all θ and $w \in \mathscr{C}$, a sample from the joint posterior of w and θ can be obtain using the adaption of the reversible jump MCMC (Green, 1995) proposed in Koskinen (2004b). This sampling scheme rests on the fact that the full conditional posteriors

$$\pi(w|\theta, x(t_1), x(t_0)) \propto L(\theta; w, x(t_0))$$
$$\pi(\theta|w, x(t_1), x(t_0)) \propto L(\theta; w, x(t_0))\pi(\theta),$$

only depend on the data likelihood through multiplicative constants.

To obtain a sample from the marginal posterior of θ given data, we sample from the joint distribution of θ and w given data giving a sequence $(\theta^{(g)}, w^{(g)})_{g=1}^{G}$, where $(\theta^{(g)})_{g=1}^{G}$ is a sample from the marginal distribution of θ given data. In each iteration g we successively draw

$$w^{(g)} \sim w|\theta^{(g-1)}, x(t_1), x(t_0), \text{ and}$$

$$\theta^{(g)} \sim \theta|w^{(g)}, x(t_1), x(t_0).$$

How to make draws of the first type is more closely described in Koskinen (2004b). To make draws from $\cdot|w^{(g)}, x(t_1), x(t_0)$, we first propose a move to θ^* sampled from a distribution $q(\theta^{(g-1)}, \theta^*)$, and then set $\theta^{(g)} := \theta^*$, with probability

$$\alpha(\theta^{(g-1)}, \theta^*|w^{(g)}) = \min\left\{1, \frac{L(\theta^*; w, x(t_0))\pi(\theta^*)}{L(\theta^{(g-1)}; w, x(t_0))\pi(\theta^{(g-1)})} \frac{q(\theta^*, \theta^{(g-1)})}{q(\theta^{(g-1)}, \theta^*)}\right\}. \tag{4.1}$$

Note that the first fraction on the RHS is equal to the ratio of the full conditional posterior of θ^* to $\theta^{(g-1)}$, conditional on everything else.

5. Prior Distributions

The interpretation in terms of the models of different prior specifications is hard to asses because of the model complexity. The statistics corresponding to the parameters are of different magnitudes which makes it difficult to determine the a priori spread of the parameters. In addition, these statistics are highly interdependent making the usually convenient choice of a priori independent parameters unsuitable. Here we suggest two procedures for finding reference priors. We demand of these reference priors that they do not give undue support to any particular effects, that they are to a degree "non-informative". The reason for this is not so much because of their influence on parameter estimation. Rather, we are concerned with the unintended consequences different scales in the prior distributions might have on model selection. We also wish our priors to have a closed analytical form or at least that they be cheap to evaluate. Here we briefly sketch to possibilities.

Since all inference is made conditional on the first observation, the prior distributions may well be set dependent on $x(t_0)$. For stochastic actor-oriented models it was shown in (Koskinen, 2004b) how, when a probit link function rather than a logit link function was used, the full conditional posterior of two blocks of parameters followed standard distributions (Gamma and Normal). Using this as an approximation, and choosing two time points t' and t'', $t_0 < t' < t'' < t_1$, the full conditional posteriors conditional on the first observation $x(t_0)$ and a latent walk consisting of a single step to $x(t'')$, taken at time t', could be averaged over all one-step walks. For convenience, the resulting normal mixture could then be approximated by a normal model.

When we have observations for time points $t_0, \dots, t_{M-1}$ for $M - 1 > 2$, an initial analysis can be carried out on $x(t_i)$, for $i = 0, \dots, M' < M - 1$, with a vague prior to obtain a proper posterior. The posterior is then used as the prior distribution when analysing the model for $x(t_k)$ for $k > M'$, conditional on the previous observations. In the second part of the procedure, model selection can be carried out in a standard fashion. Although standard result regarding the asymptotic properties of Bayesian model selection are not (automatically) altered by the use of training samples (O'Hagan, 1995), it not self-evident what type of information these priors contribute with. This procedure will be treated in greater detail in Section 9.1. after we have described model selection and its implementation.

6. Model Selection

In this and the following section x is used to refer to data $x(t_1)$ and the implicit conditioning on $x(t_0)$. For a set $\mathscr{M}$ of models for x characterized by parameters $\theta_M \in \Theta_M$, data likelihoods $L_{D,M}$ and prior distributions π_M, the marginal likelihood of model M is given by

$$m_M(x) = \int_{\Theta_M} L_{D,M}\left(\theta_M; x\right) \pi_M\left(\theta_M\right) d\theta_M,$$

for $M \in \mathscr{M}$. With prior probability $\pi(M)$, the posterior probability of model M is

$$\pi(M|x) = \frac{m_M(x)\pi(M)}{\sum_{M\in\mathscr{M}} m_M(x)\pi(M)}.$$

Bayes rule for model selection is to choose the model which maximizes the posterior probability over $\mathscr{M}$. Model selection can be carried out by pairwise comparison of the models in $\mathscr{M}$, through their posterior odds ratio

$$\frac{\pi(M_i|x)}{\pi(M_j|x)} = \frac{\pi(M_i)}{\pi(M_j)}\frac{m_{M_i}(x)}{m_{M_j}(x)},$$

for $M_i, M_j \in \mathscr{M}$. The ratio of prior probabilities is known as the prior odds and the ration of marginal likelihoods as the Bayes factor.

Since the only quantity involved in model selection to cause any trouble is the marginal likelihood, considerable attention has been devoted to estimation procedures for this (see Raftery, 1996b, for a review). Another issues arising in connection with model selection is the need for proper prior distributions. As mentioned above, minor difference for two sets of priors might have little impact on the posterior distribution of the parameters but a large impact on the posterior distribution over the class of models (Raftery, 1996a). It is important to point out that this problem is in no way eliminated by using say information criteria or goodness of fit methods (rather the opposite since model complexity typically is reduced to some function of the dimensions of the parameters).

7. Implementation

This section presents the main result of this paper, which in principle deals with the two major problems involved in model selection. The first problem is of a general nature whereas the second is specific to the family of models treated here.

Firstly, the marginal likelihood is an expectation with respect to the prior distributions which rarely lends itself to direct calculation. Calculations of the marginal likelihood based upon importance sampling, such as the harmonic mean estimator (see e.g. Raftery, 1996b), typically suffer from stability problems. The solution adapted here, suggested by Chib and Jeliazkov (2001), takes full advantage of the information about the parameters contained in the MCMC algorithm. Their method does however require that we evaluate the data likelihood, which brings us to problem number two: the data likelihood is not analytically tractable. This problem often arises in model selection for social network models but can usually be solved by finding a point in the parameter space for which it is possible to evaluated the likelihood (Koskinen, 2002, 2004a). The harmonic mean estimators, in addition to being very unstable, require that the likelihood can be evaluated for every point in the parameter space.

In the following we drop the notational dependency on the models, since all calculations are carried out separately for each model. Note that the marginal likelihood is the normalizing constant of the posterior distribution. By solving for the marginal likelihood in Bayes theorem we obtain what is commonly called the *basic marginal likelihood identity*

$$m(x) = \frac{L_D(\theta; x)\pi(\theta)}{\pi(\theta|x)}. \tag{7.1}$$

Since this equality holds for all $\theta \in \Theta$, if the model under investigation belongs to $\mathscr{M}_{rec}$ we can evaluate the numerator for an arbitrarily chosen $\theta^* \in \widetilde{\Theta}$. If the models considered constitute a pairwise nested set of models, we can compute Bayes factors in such an order that the data likelihood cancels out. This leaves the problem of evaluating the posterior ordinate $\pi(\theta^*|x)$. Following Chib and Jeliazkov (2001) we can write

$$\pi(\theta^*|x) = \frac{E_1\left\{\alpha(\theta, \theta^*|w)q(\theta, \theta^*)\right\}}{E_2\left\{\alpha(\theta^*, \theta|w)\right\}}, \tag{7.2}$$

where E_1 is the expectation with respect to the joint posterior distribution $\pi(\theta, w|x)$ and E_2 is the expectation with respect to $\pi(w|x, \theta^*)q(\theta^*, \theta)$.

Since expectancies can be simulation consistently estimated taking samples from the appropriate distributions and averaging the desired quantities, and we have a procedure for sampling from the desired distributions, we are more or less done. To describe the estimation process in a little more detail, consider first obtaining a sample $(\theta^{(g)}, w^{(g)})_{g=1}^{G}$ as described in Section 4. An estimate of the numerator in 7.2, is given by averaging

$$\frac{1}{G}\sum_{g=1}^{G} \alpha(\theta^{(g)}, \theta^*|w^{(g)})q(\theta^{(g)}, \theta^*).$$

For the denominator, run the algorithm for another J iterations giving a sample $\{w^{(j)}\}_{j=1}^{J}$, from the full conditional posterior $\pi(w|\theta^*, x)$. For each $j = 1, \ldots, J$, make a draw

$$\theta^{(j)} \sim q(\theta^*, \theta^{(j)}),$$

and thus we have pairs $(\theta^{(j)}, w^{(j)})$ drawn from the distribution

$$\pi(w|\theta^*, x)q(\theta^*, \theta).$$

The estimate of the denominator in 7.2 is given by

$$\frac{1}{J}\sum_{j=1}^{J}\alpha(\theta^{*},\theta^{(j)}|w^{(j)}).$$

8. Valued data

We have only dealt with binary data, i.e. $\mathscr{X} = \{0,1\}^N$, but the extension of these model selection procedures to valued data is straightforward as long as we keep to a set of models that is pairwise nested. By valued data we mean networks where we at different points in time record the strength of the relationship between i and j for each dyad $(i,j) \in \mathscr{N}$. If the strength take values $\{0,1,\ldots,R-1\}$, the evolution of the network is described by a process on $\mathscr{X} = \{0,1,\ldots,R-1\}^N$, defined in a way equivalent to section 3. The inference scheme as described in Koskinen (2004b) still applies as does the procedure for estimating the posterior ordinate. Of course, the definition $\mathscr{M}_{rec}$ does not immediately apply, and some modifications are needed.

9. Example

To illustrate the procedures we fit two models to data on 32 freshmen students collected by van de Bunt (1999). The observations are made at seven points in time, $t_0,\ldots,t_6$, the time span between consecutive observations is three weeks for t_0 through t_4, and six weeks between t_4 and t_5, t_5 and t_6. We have focused on the "friendly relation", more closely described in van de Bunt (1999). Missing data has been imputed with the last observed value (or 0 for missing values at t_0), which is a rather conservative choice with respect to the evolution model to be presented next.

The models fitted are so called *actor-oriented models* and for details of different specifications and interpretations we refer to Snijders (2004). Effects considered are

(1) Density effect $s_{i1}(x) = \sum_{j\in V\setminus\{i\}} x_{ij}$
(2) Reciprocity effect $s_{i2}(x) = \sum_{j\in V\setminus\{i\}} x_{ij}x_{ji}$
(3) Distance 2 effect $s_{i3}(x) = |\{j : x_{ij} = 0, \quad \max_k(x_{ik}x_{kj}) > 0\}|$
(4) Balance $s_{i4}(x) = \sum_{j\in V\setminus\{i\}} x_{ij} \sum_{h\in V\setminus\{i,j\}} |x_{ik} - x_{jk}|$. (Note that we choose not to include the constant b_0 in the balance statistic c.p. Snijders, 2001).
(5) Transitivity effect $s_{i5}(x) = \sum_{j,k\in V\setminus\{i\}} x_{ij}x_{ik}x_{jk}$.
In addition for covariates sex (female/male), smoking (yes/no) and program (length in years of program participation, 2,3,4), denoted by v_{1i}, v_{2i} and v_{3i} we have the following effects
(6) Popularity with respect to sex $s_{i6}(x) = \sum_{j\in V\setminus\{i\}} x_{ij}v_{1j}$
(7) Dissimilarity with respect to sex $s_{i7}(x) = \sum_{j\in V\setminus\{i\}} x_{ij}|v_{1i} - v_{1j}|$
(8) Dissimilarity with respect to program $s_{i8}(x) = \sum_{j\in V\setminus\{i\}} x_{ij}|v_{2i} - v_{2j}|$
(9) Dissimilarity with respect to smoking $s_{i9}(x) = \sum_{j\in V\setminus\{i\}} x_{ij}|v_{3i} - v_{3j}|$.

The coefficients β_1 through β_9 together with two rate-related parameters, α and ρ, were collected in a parameter vector $\theta = (\rho, \alpha, \beta_1, \dots, \beta_9)$. We define a full model with rate

$$\lambda(\theta, x) = \sum\nolimits_{i \in V} \lambda_i(\rho, \alpha, x),$$

where

$$\lambda_i(\rho, \alpha, x) = \frac{\rho}{n-1}\left[(n-1-x_{i+})\, e^{\alpha} + x_{i+}e^{-\alpha}\right],$$

in which $+$ in place of an index means that the variable should be summed over that index. The jump probabilities are of the form

$$\pi(\theta, x, y) = \frac{\lambda_i(\rho, \alpha, x)}{\lambda(\theta, x)} \frac{e^{r(\theta,x,y)}}{\sum_{z \in \mathscr{X}_i(x)} e^{r(\theta,x,z)}},$$

where y differs from x only in the element (i, j), and

$$r(\theta, x, y) = \sum_{k=1}^{9} \beta_k s_{ik}(y)$$

for the full model. The statistics 3, 4, and 5, are related and represent various aspects of network closure. The local maxima of 4 and 5 and the local minimum of 3 are all achieved for networks with several disconnected subgraphs (Snijders, 2004). A way of removing effects from the model that can be theoretically motivated would be to exclude distance 2 and transitivity. In Snijders (2004) it was concluded that the distance 2 effect and the transitivity effect should be omitted. This was motivated by the approximate standard errors of the method of moments estimates for the corresponding parameters. The purpose here is to illustrate how these two models, the full and reduced, can be compared using Bayes methodology. The reduced model does not contain the effects transitivity and balance but is defined equivalently for the other components in the full model (note that the model fitted to van de Bunt's freshmen students in Snijders 2004 had a slightly different parameterization for the balance effect, included an effect corresponding to differences between sexes with respect to friendship formation activity, and assumed different ρ_m for all $m = 0, \dots, M-2$).

Both models belong to $\mathscr{M}_{rec}$ since for any θ^* with $\alpha = \beta_1$, and $\beta_k = 0$ for $k > 1$, we have that

$$\lambda(\theta^*, x)\pi(\theta^*, x, y) = \lambda_{x_{ij}},$$

where y differs from x only in the element (i, j), and

$$\lambda_0 = \frac{\rho}{n-1}e^{\alpha}, \text{ and } \lambda_1 = \frac{\rho}{n-1}e^{-\alpha}.$$

Since the reduced model is nested within the full model, these two models also constitutes a pairwise nested set of models. Hence, for appropriate values of the parameters, when we compute the Bayes factor using the expression (7.1), the data likelihoods cancel each other out. More specifically, we can evaluate the posterior ordinate in any θ^* for the reduced model if we at the same time evaluate the posterior ordinate for the full model for the same parameter values except for the extra parameters β_4^* and β_5^*, both of which are set to 0. In the calculations to follow we illustrate this procedure only.

For updating the parameter vector in the algorithm as described in section 4, the following proposal distribution was used. Given the current parameter vector $\theta = (\rho, \alpha, \beta_1, \dots, \beta_p)'$, a candidate vector θ^* was proposed from

$$\left(\alpha^*, \beta_1^*, \dots, \beta_p^*\right)' \sim N_{p+1}\left((\alpha, \beta_1, \dots, \beta_p)', \Omega\right),$$

and independently thereof

$$\rho^* \sim Gamma(\rho k, k^{-1}).$$

The jumping scales Ω and k were set to $\frac{\gamma}{\sqrt{p+1}}\hat{\Sigma}$, where $\hat{\Sigma}$ is the posterior covariance estimated from a test run, and λ is chosen so that the acceptance rates are appropriate.

9.1. **Training sample.** When model selection is carried out using proper prior distributions with convenient analytical forms, once we have obtained the estimates of the relevant posterior ordinates, the marginal likelihoods can be computed in the manner described. Using priors obtained from a training sample there are some additional complications that need to be dealt with. Consider first the case when using the first two observations for training priors. When training a prior on the first two observations, we obtain a sample from the posterior distribution of θ given $x(t_1)$ and $x(t_0)$. In the algorithm of Koskinen (2004b) it is assumed that the ratio of posterior distributions in 4.1 is cheap to evaluate, which is not the case if we need to estimate the prior density in a new point in each iteration. The sequential nature of Bayesian inference however allows us to obtain a sample from the posterior of θ given data $x(t_1), \dots, x(t_{M-1})$ with prior $\pi(\theta|x(t_1), x(t_0))$ by conducting the analysis for the entire data set with a vague (constant) prior, giving a sample from θ given $x(t_0), \dots, x(t_{M-1})$. Using $\pi(\theta|x(t_1), x(t_0))$ as the prior distribution, the posterior distribution is written

$$\pi(\theta|x(t_1), \dots, x(t_{M-1})) = \frac{L_D(\theta; x(t_1), \dots, x(t_{M-1}))\pi(\theta|x(t_1), x(t_0))}{\int L_D(\theta; x(t_1), \dots, x(t_{M-1}))\pi(\theta|x(t_1), x(t_0))d\theta}$$

and noting that $\pi(\theta|x(t_1), x(t_0)) \propto L_D(\theta; x(t_0), x(t_1))$, since the prior used in the training set was vague, and by the Markov property

$$L_D(\theta; x(t_0), x(t_1))L_D(\theta; x(t_1), \dots, x(t_{M-1})) = L_D(\theta; x(t_0), \dots, x(t_{M-1})),$$

we have

$$\pi(\theta|x(t_1), \dots, x(t_{M-1})) = \frac{L_D(\theta; x(t_0), \dots, x(t_{M-1}))}{\int L_D(\theta; x(t_0), \dots, x(t_{M-1}))d\theta},$$

in which the RHS is the posterior of θ given data $x(t_0), \dots, x(t_{M-1})$, with a vague prior. Thus, when estimating the posterior ordinate $\pi(\theta|x(t_1), \dots, x(t_{M-1}))$, we may apply the technique described in section 7 on the sampling algorithm for θ given $x(t_0), \dots, x(t_{M-1})$ with a vague prior. The data likelihood in the numerator of the basic marginal likelihood identity is evaluated for $x(t_1), \dots, x(t_{M-1})$, and the prior ordinate in the numerator is again estimated using the technique in section 7 for a separate analysis of $x(t_0)$ and $x(t_1)$.

The number of observation points used for training prior distributions need not be limited to one or two, but can be any number of observations. The procedure described still applies as long as at least one observation is spared for model evaluation. Recommendations for the proportion of sample points that should be used for training ranges from as few as possible to half of the sample (see O'Hagan, 1995). If we let y denote the part of the sample that is used for training, the *training sample*, and let z denote the part of the sample used for model comparison,

$$B(z|y) = \frac{m_1(z|y)}{m_2(z|y)} \tag{9.1}$$

is usually called the partial Bayes factor (O'Hagan, 1995).

9.2. **Posteriors.** For each model the first three observations, $x(t_0)$, $x(t_1)$ and $x(t_2)$ were used as a training sample, and the marginal trained prior distributions along with the resulting posteriors are given in Figures 1, and 2 for the reduced and full model respectively. These distributions are summarized in Table 1. The partial log-Bayes factor for the reduced model relative to the full model was well over 200, which is rather strong evidence that balance and transitivity should be excluded from the model. One explanation of this is that the full model is penalized for including the transitivity effect, β_4. Looking at Table 1 and Figure 2, notice that whereas the (95%) probability interval of β_4 includes the origin, the trained prior suggests that β_4 has an effect. In addition the posterior of β_1 is centered (more or less) over zero for the full model but not for the reduced model. Comparing the number of intermediate changes between t_1 and t_2, the uncertainty is greater for the reduced model than for the full model, which together with the location of the trained priors for the full model suggest over fitting. This might also contribute to the low posterior probability of the full model. While the reduced model is better, neither model seems to capture the differences in early stages of the network evolution (which could be termed a "getting to know" period) as compared to the evolution of the later phase. This inadequacy manifests itself in the differences between trained priors and the posteriors of the rate-related parameters, α and ρ, corresponding to the density effect on the change rate and the constant factor in the rate function. Whether this could have been taken care of within the framework of stochastic actor-oriented models or not is a question for future research.

References

Aldous, D. (1983). "Minimization algorithms and random walk on the d-cube." *Ann. Probab.*, 11, 2, 403–413.

Chib, S. and Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis-Hastings output." *J. Amer. Statist. Assoc.*, 96, 453, 270–281.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computaion and Bayesian model determination." *Biometrika*, 82, 711–732.

Holland, P. and Leinhardt, S. (1977a). "A dynamic model for social networks." *Journal of Mathematical Sociology*, 5, 5–20.

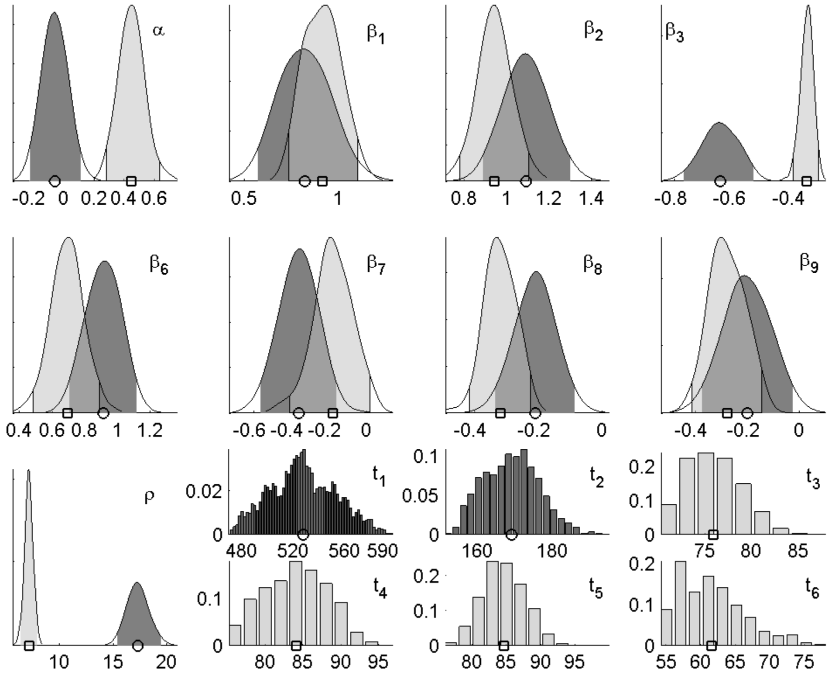

FIGURE 1. Marginal posterior densities for reduced model, with 95% Credibility intervals (pale), and posterior means (○). Trained prior 95% Credibility intervals (dark) and means given by (□)

		Prior		Post.		Prior		Post.	
effect	par.	mean	std	mean	std	mean	std	mean	std
den.-rate	α	-0.06	0.084	0.44	0.087	0.34	0.133	0.56	0.085
activity	β_1	0.82	0.136	0.91	0.098	-0.32	0.149	0.12	0.152
mutuality	β_2	1.09	0.104	0.94	0.08	0.65	0.111	0.53	0.093
no dist.-2	β_3	-0.65	0.061	-0.36	0.022	-0.29	0.029	-0.24	0.022
balance	β_4					0.027	0.0076	0.007	0.0048
trans.	β_5					0.117	0.0131	0.056	0.0055
pop. sex	β_6	0.91	0.107	0.69	0.102	0.84	0.106	0.69	0.098
dis. sex	β_7	-0.36	0.104	-0.18	0.104	-0.4	0.103	-0.29	0.086
dis. prog.	β_8	-0.21	0.064	-0.32	0.05	-0.19	0.057	-0.29	0.046
dis. smok.	β_9	-0.21	0.091	-0.28	0.069	-0.07	0.075	-0.27	0.06
rate	ρ	17.25	1.008	7.2	0.396	13.95	1.234	6.86	0.447

TABLE 1. Posterior and trained prior means and standard deviations for van de Bunt's freshmen students, where priors are based on a training set consisting of $x(t_0)$, $x(t_1)$, $x(t_2)$

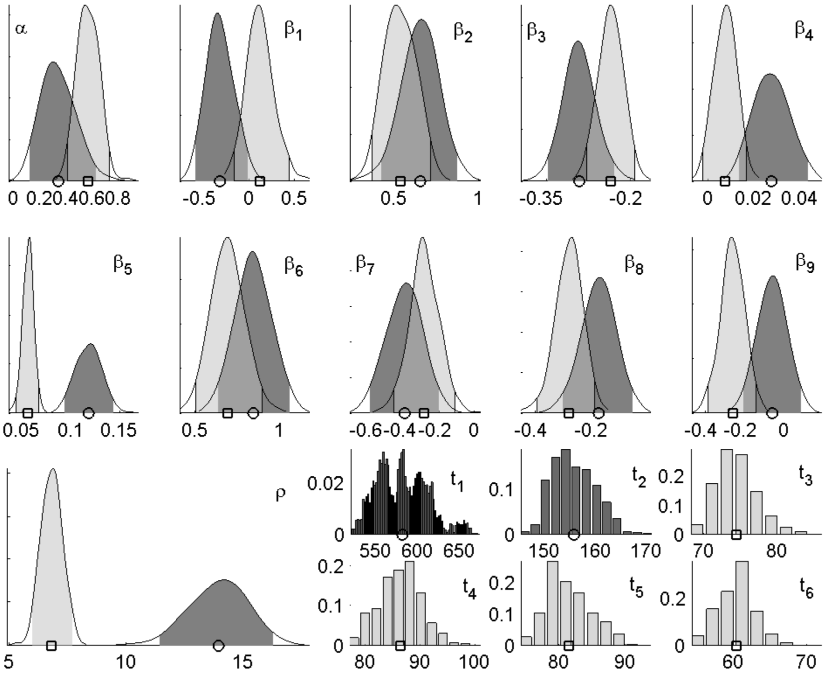

FIGURE 2. Marginal posterior densities for saturated model, with 95% Credibility intervals (pale), and posterior means (○). Trained prior 95% Credibility intervals (dark) and means given by (□)

— (1977b). “Social structure as a network process.” *Zeitschrift für Soziologie*, 6, 386–402.

Koskinen, J. (2002). “Bayesian analysis of perceived social networks.” Research Report 2002:2, Department of Statistics, Stockholm University.

— (2004a). “Bayesian analysis of exponential random graphs – estimation of parameters and model selection.” Research Report 2004:2, Department of Statistics, Stockholm University.

— (2004b). “Bayesian inference for longitudinal social networks.” Research Report 2004:4, Department of Statistics, Stockholm University.

Leenders, R. T. A. J. (1995a). “Models for network dynamics: a Markovian framework.” *Journal of Mathematical Sociology*, 20, 1–21.

— (1995b). “Structure and Influence. Statistical Models for the Dynamics of Actor Attributes, Network Structure and their Interdependence.” Ph.D. thesis, Amsterdam.

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

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

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

Snijders, T. A. B. (1996). "Stochastic actor-oriented models for network change." *Journal of Mathematical Sociology*, 21, 149–172. Also published in Doreian and Stokman (1997).

— (1999). "The transition probabilities of the reciprocity model." *Journal of Mathematical Sociology*, 23, 241–253.

— (2001). "The statistical evaluation of social network dynamics." *Sociological Methodology*, 30, 361–395.

— (2004). "Models for longitudinal network data." To appear as Chapter 11 in P. Carrington, J. Scott, and S. Wasserman (Eds.), Models and methods in social network analysis. New York: Cambridge University Press, 2004.

van de Bunt, G. G. (1999). "Friends by Choice. An Actor-Oriented Statistical Network Model for Friendship Networks through Time." Ph.D. thesis, Amsterdam.

Wasserman, S. (1977). "Stochastic Models for Directed Graphs." Ph.D. thesis, University of Harvard, Department of Statistics.

— (1980a). "Analyzing social networks as stochastic processes." *Journal of the American Statistical Association*, 75, 280–294.

— (1980b). "A stochastic model for directed graphs with transition rates determined by reciprocity." *Sociological Methodology*, 11, 392–412.

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