

# Bayesian model selection for exponential random graph models

Alberto Caimo & Nial Friel  
School of Mathematical Sciences,  
University College Dublin, Ireland  
{alberto.caimo,nial.friel}@ucd.ie

January 13, 2012

## Abstract

Exponential random graph models are a class of widely used exponential family models for social networks. The topological structure of an observed network is modeled by the relative prevalence of a set of local sub-graph configurations termed network statistics. One of the key tasks in the application of these models is which network statistics to include in the model. This can be thought of as statistical model selection problem. This is a very challenging problem—the posterior distribution for each model is often termed “doubly intractable” since computation of the likelihood is rarely available, but also, the evidence of the posterior is, as usual, also intractable. We present a fully Bayesian model selection method based on a Markov chain Monte Carlo algorithm of [Caimo and Friel \(2011\)](#) which estimates the posterior probability for each competing model as well as a possible approach for computing the model evidence.

## 1 Introduction

Exponential random graph models are a powerful and flexible family of statistical models for networks which allows us to model network topologies without requiring any independence assumption between dyads (pairs of nodes). These models have been utilized extensively in the social science literature since they allow to statistically account for the complexity inherent in many network data. The basic assumption of these models is that the topological structure in an observed network  $\mathbf{y}$  can be explained by the relative prevalence of a set of overlapping sub-graph configurations  $s(\mathbf{y})$  also called graph or network statistics (see [Figure 1](#)).

Formally a random network  $\mathbf{Y}$  consists of a set of  $n$  nodes and  $m$  dyads  $\{Y_{ij} : i = 1, \dots, n; j = 1, \dots, n\}$  where  $Y_{ij} = 1$  if the pair  $(i, j)$  is connected (full dyad), and

$Y_{ij} = 0$  otherwise (empty dyad). Edges connecting a node to itself are not allowed so  $Y_{ii} = 0$ . The graph  $\mathbf{Y}$  may be directed (digraph) or undirected depending on the nature of the relationships between the nodes.

Exponential random graph models (ERGMs) are a particular class of discrete linear exponential families which represent the probability distribution of  $\mathbf{Y}$  as

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{q_{\boldsymbol{\theta}}(\mathbf{y})}{z(\boldsymbol{\theta})} = \frac{\exp\{\boldsymbol{\theta}^T s(\mathbf{y})\}}{\sum_{\mathbf{y} \in \mathcal{Y}} \exp\{\boldsymbol{\theta}^T s(\mathbf{y})\}} \quad (1)$$

where  $s(\mathbf{y})$  is a known vector of sufficient statistics computed on the network (or graph) (see [Snijders et al. \(2006\)](#) and [Robins et al. \(2007\)](#)) and  $\boldsymbol{\theta}$  are model parameters describing the dependence of  $p(\mathbf{y}|\boldsymbol{\theta})$  on the observed statistics  $s(\mathbf{y})$ . Estimating ERGM parameters is a challenging task due to the intractability of the normalising constant  $z(\boldsymbol{\theta})$  and the issue of model degeneracy (see [Handcock \(2003\)](#) and [Rinaldo et al. \(2009\)](#)).

An important problem in many applications is the choice of the most appropriate set of explanatory statistics network statistics  $s(\mathbf{y})$  to include in the model from a set of *a priori* plausible ones. In fact in many applications there is a need to classify different types of networks based on the relevance of a set of configurations with respect to others.

From a Bayesian point of view, the model choice problem is transformed into a parameter estimation problem aiming at estimating the posterior probability of all models within the considered class of competing models. In order to account for the uncertainty concerning the model selection process, Bayesian Model Averaging ([Hoeting et al., 1999](#)) offers a coherent methodology which consists in averaging over many different competing models.

In the ERGM context, the intractability of the likelihood makes the use of standard techniques quite challenging. The purpose of this paper is to present two new methods for Bayesian model selection for ERGMs. This article is structured as follows. A brief overview of Bayesian model selection theory is given in Section 2. An across-model approach based on a trans-dimensional extension of the exchange algorithm of [Caimo and Friel \(2011\)](#) is presented in Section 3. The issue of the choice of suitable jump proposals is addressed by presenting an automatic reversible jump exchange algorithm involving an independence sampler based on a distribution fitting a parametric density approximation to the within-model posterior. This algorithm bears some similarity to that presented in Chapter 6 of [Green et al. \(2003\)](#). The second novel method is a within-model approach for computing the model evidence. This approach is based on the path sampling approximation for estimating the likelihood normalizing constant and it makes use of nonparametric density estimation procedures for approximating the posterior density of each competing model (Section 4). Three illustrations of how these new method perform in practice are give in Section 5. Some conclusions are outlined in Section 6. The `Bergm` package for R, provided some of the functions used in this paper and it is available on the CRAN package repository at <http://cran.r-project.org/web/packages/Bergm>.

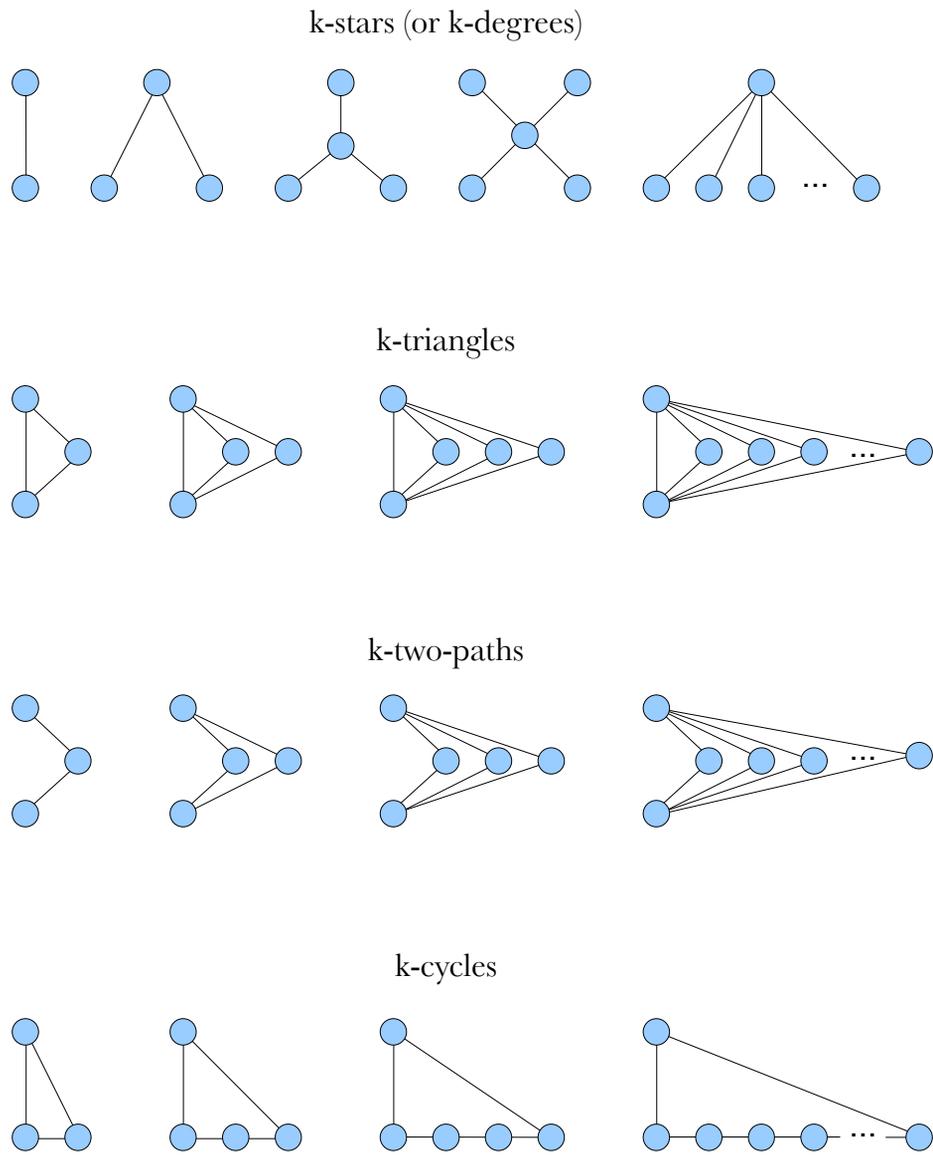


Figure 1: Some of the most used sub-graph configurations for undirected graphs (analogous directed versions can be used for digraphs).

## 2 Overview of Bayesian model selection

Bayesian model comparison is commonly performed by estimating posterior model probabilities. More precisely, suppose that the competing models can be enumerated and represented by the set  $\{m_h : h = 1, \dots, H\}$ . Suppose data  $\mathbf{y}$  is assumed to have been generated by model  $m_h$ , the posterior distribution is:

$$p(\boldsymbol{\theta}_h | \mathbf{y}, m_h) = \frac{p(\mathbf{y} | \boldsymbol{\theta}_h, m_h) p(\boldsymbol{\theta}_h | m_h)}{p(\mathbf{y} | m_h)} \quad (2)$$

where  $p(\mathbf{y} | \boldsymbol{\theta}_h, m_h)$  is the likelihood and  $p(\boldsymbol{\theta}_h | m_h)$  represents the prior distribution of the parameters of model  $m_h$ . The model evidence (or marginal likelihood) for model  $m_h$

$$p(\mathbf{y} | m_h) = \int_{\boldsymbol{\theta}_h} p(\mathbf{y} | \boldsymbol{\theta}_h, m_h) p(\boldsymbol{\theta}_h | m_h) d\boldsymbol{\theta}_h \quad (3)$$

represents the probability of the data  $\mathbf{y}$  given a certain model  $m_h$  and is typically impossible to carry out analytically. However, the model evidence is crucial for Bayesian model selection since it allows us to make statements about posterior model probabilities. Bayes' theorem can be written as

$$p(m_h | \mathbf{y}) = \frac{p(\mathbf{y} | m_h) p(m_h)}{\sum_1^H p(\mathbf{y} | m_h) p(m_h)}. \quad (4)$$

Based on these posterior probabilities, pairwise comparison of models,  $m_h$  and  $m_k$  say, can be summarised by the posterior odds:

$$\frac{p(m_h | \mathbf{y})}{p(m_k | \mathbf{y})} = \frac{p(\mathbf{y} | m_h)}{p(\mathbf{y} | m_k)} \times \frac{p(m_h)}{p(m_k)}. \quad (5)$$

This equation reveals how the data  $\mathbf{y}$  through the Bayes factor

$$BF_{hk} = \frac{p(\mathbf{y} | m_h)}{p(\mathbf{y} | m_k)} \quad (6)$$

updates the prior odds

$$\frac{p(m_h)}{p(m_k)} \quad (7)$$

to yield the posterior odds.

By treating  $p(m_h | \mathbf{y})$  as a measure of the uncertainty around of model  $m_h$  a natural approach for model selection is to choose the most likely  $m_h$  a posteriori, i.e. the model for which  $p(m_h | \mathbf{y})$  is the largest.

Bayesian model averaging ([Hoeting et al., 1999](#)) provides a way of summarising model uncertainty in inference and prediction. After observing the data  $\mathbf{y}$  is possible to predict a possible future outcome  $\mathbf{y}^*$  by calculating an average of the posterior

distributions under each of the models considered, weighted by their posterior model probability.:

$$p(\mathbf{y}^*|\mathbf{y}) = \sum_{h=1}^H p(\mathbf{y}^*|m_h, \mathbf{y})p(m_h|\mathbf{y}) \quad (8)$$

where  $p(\mathbf{y}^*|m_h, \mathbf{y})$  represents the posterior prediction of  $\mathbf{y}^*$  according to model  $m_h$  and data  $\mathbf{y}$ .

As we said above model evidence is generally difficult to compute and exact solutions are known for a small class of distributions. Numerical integration methods are usually needed, either general methods such as Gaussian integration or a Monte Carlo method, or methods specialized to statistical problems such as the Laplace approximation, Gibbs sampling or the EM algorithm.

## 2.1 Computing Bayes factors

Generally speaking there are two approaches for computing Bayes factors: across-model and within-model estimation. The former strategies involve the use of an MCMC algorithm generating a single Markov chain which crosses the joint model and parameter space so as to sample from

$$p(\boldsymbol{\theta}_h, m_h|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}_h, m_h) p(\boldsymbol{\theta}_h|m_h) p(m_h). \quad (9)$$

One of the most popular approach used in this context of the reversible jump MCMC algorithm of [Green \(1995\)](#) which is briefly reviewed in [Section 2.1.1](#).

Within-model strategies focus on the posterior distribution (2) of each competing model  $m_h$  separately aiming at estimating their model evidence (3) which can then be used to calculate Bayes factors (see for example [Chib \(1995\)](#), [Chib and Jeliazkov \(2001\)](#), [Neal \(2001\)](#), [Friel and Pettitt \(2008\)](#), and [Friel and Wyse \(2012\)](#), who present a review of these methods).

### 2.1.1 Reversible jump MCMC

The Reversible Jump MCMC (RJMCMC) algorithm is a flexible technique for model selection introduced by [Green \(1995\)](#) which allows simulation from target distributions on spaces of varying dimension. In the reversible jump algorithm, the Markov chain “jumps” between parameter subspaces (models) of differing dimensionality, thereby generating samples from the joint distribution of parameters and model indices.

To implement the algorithm we consider a countable collection of candidate models,  $\{m_l : l = 1, \dots, K\}$ , each having an associated vector of parameters  $\boldsymbol{\theta}_l$  of dimension  $D_l$  which typically varies across models. We would like use MCMC to sample from the joint posterior (9).

In order to jump from  $(\boldsymbol{\theta}_k, m_k)$  to  $(\boldsymbol{\theta}_h, m_h)$ , one may proceed by generating a random vector  $\mathbf{u}$  from a distribution  $g$  and setting  $(\boldsymbol{\theta}_h, m_h) = f_{hj}((\boldsymbol{\theta}_k, m_k), \mathbf{u})$ . Similarly to jump from  $(\boldsymbol{\theta}_h, m_h)$  to  $(\boldsymbol{\theta}_k, m_k)$  we have  $(\boldsymbol{\theta}_k, m_k) = f_{jh}((\boldsymbol{\theta}_h, m_h), \mathbf{u}^*)$  where  $\mathbf{u}^*$  is a

random vector from  $g^*$  and  $f_{hk}$  is some deterministic function. However reversibility is only guaranteed when the parameter transition function  $f_{hk}$  is a diffeomorphism, that is, both a bijection and its differential invertible. A necessary condition for this to apply is the so-called “dimension matching”:  $\dim(\boldsymbol{\theta}_k) + \dim(\mathbf{u}) = \dim(\boldsymbol{\theta}_h) + \dim(\mathbf{u}^*)$  (where  $\dim(\cdot)$  stands for “dimension of”). In this case the acceptance probability can be written as:

$$\min \left\{ 1, \frac{p(\boldsymbol{\theta}_h, m_h | \mathbf{y}) p(m_h \rightarrow m_k) g^*(\mathbf{u}^*)}{p(\boldsymbol{\theta}_k, m_k | \mathbf{y}) p(m_k \rightarrow m_h) g(\mathbf{u})} |J| \right\} \quad (10)$$

where  $p(m_h \rightarrow m_k)$  is the probability of jumping from model  $m_h$  to model  $m_k$ , and  $|J|$  is the Jacobian resulting from the transformation from  $((\boldsymbol{\theta}_k, m_k), \mathbf{u})$  to  $((\boldsymbol{\theta}_h, m_h), \mathbf{u}^*)$ .

Mixing is crucially affected by the choice of the parameter of the jump proposal distribution  $g$  and this is one of the fundamental difficulties that makes RJMCMC often hard to use in practice (Brooks et al., 2003).

### 3 Reversible jump exchange algorithm

In the ERGM context, RJMCMC techniques cannot be used straightforwardly because both the likelihood normalizing constant  $z(\boldsymbol{\theta})$  in (1) cannot be computed analytically. Here we present an implementation of an RJMCMC approach for ERGMs based on the extension of the exchange algorithm for exponential random graph models (Caimo and Friel, 2011).

For each model  $m_h$ , this algorithm allows sampling from the following augmented distribution:

$$p(\boldsymbol{\theta}'_h, \mathbf{y}', \boldsymbol{\theta}_h | \mathbf{y}, m_h) \propto p(\mathbf{y} | \boldsymbol{\theta}_h, m_h) p(\boldsymbol{\theta}_h | m_h) h(\boldsymbol{\theta}'_h | \boldsymbol{\theta}_h, m_h) p(\mathbf{y}' | \boldsymbol{\theta}'_h, m_h) \quad (11)$$

where  $p(\mathbf{y} | \boldsymbol{\theta}_h, m_h)$  and  $p(\mathbf{y}' | \boldsymbol{\theta}'_h, m_h)$  are respectively the original likelihood defined on the observed data  $\mathbf{y}$  and the augmented likelihood defined on simulated data  $\mathbf{y}'$ ,  $p(\boldsymbol{\theta}_h | m_h)$  is the parameter prior and  $h(\boldsymbol{\theta}'_h | \boldsymbol{\theta}_h, m_h)$  is any arbitrary proposal distribution for  $\boldsymbol{\theta}'_h$ . Marginalising (11) over  $\boldsymbol{\theta}'_h$  and  $\mathbf{y}'$  yields the posterior of interest  $p(\boldsymbol{\theta}_h | \mathbf{y}, m_h)$ .

Auxiliary variable methods for intractable likelihood models, such as the exchange algorithm, have not been used in a trans-dimensional setting before. In order to propose to move from  $(\boldsymbol{\theta}_k, m_k)$  to  $(\boldsymbol{\theta}'_h, m'_h)$ , the algorithm (11) can be extended to sample from:

$$p(\boldsymbol{\theta}'_h, \boldsymbol{\theta}_k, m'_h, m_k, \mathbf{y}' | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}_k, m_k) p(\boldsymbol{\theta}_k | m_k) p(m_k) h(\boldsymbol{\theta}'_h, m'_h | \boldsymbol{\theta}_k, m_k) p(\mathbf{y}' | \boldsymbol{\theta}'_h, m'_h) \quad (12)$$

where  $p(\mathbf{y} | \boldsymbol{\theta}_k, m_k)$  and  $p(\mathbf{y}' | \boldsymbol{\theta}'_h, m'_h)$  are the two likelihood distributions for the data  $\mathbf{y}$  under model  $m_k$  and the auxiliary data  $\mathbf{y}'$  under the competing model  $m'_h$  respectively,  $p(\boldsymbol{\theta}_k | m_k)$  and  $p(m_k)$  are the priors for the parameter  $\boldsymbol{\theta}_k$  and the respective model  $m_k$  and  $h(\boldsymbol{\theta}'_h, m'_h | \boldsymbol{\theta}_k, m_k)$  is some jump proposal distribution. Analogously as before, the marginal of (12) for  $\boldsymbol{\theta}'_h$  and  $m'_h$  is the distribution of interest (9).

The issue with this method is that tuning the jump proposals  $h(\cdot)$  in a sensible way so as to get a reasonable mixing can be difficult and automatic choice of jump parameters (Brooks et al., 2003) does not apply in this context due to the double intractability of the likelihood distribution.

### 3.1 Pilot-tuned RJ exchange algorithm

We now consider nested models or models differing by at most one variable. In this case, the move from  $(\boldsymbol{\theta}_k, m_k)$  to a larger model  $(\boldsymbol{\theta}'_{k+1}, m'_{k+1})$  such that  $\dim(m'_{k+1}) = \dim(m_k) + 1$  can be done by proposing the transformation  $(\boldsymbol{\theta}'_{k+1}, m'_{k+1}) = ((\boldsymbol{\theta}_k, \theta'_{k+1}), m_{k+1})$  where the  $(k + 1)$ -th parameter value  $\theta'_{k+1}$  is generated from some distribution  $g_{k+1}$  and then accepting the move with the following probability:

$$\alpha = \min \left\{ 1, \frac{q_{\boldsymbol{\theta}_k, m_k}(\mathbf{y}')}{q_{\boldsymbol{\theta}_k, m_k}(\mathbf{y})} \frac{q_{\boldsymbol{\theta}'_{k+1}, m'_{k+1}}(\mathbf{y})}{q_{\boldsymbol{\theta}'_{k+1}, m'_{k+1}}(\mathbf{y}')} \frac{p(\boldsymbol{\theta}'_{k+1} | m'_{k+1})}{p(\boldsymbol{\theta}_k | m_k)} \frac{p(m'_{k+1})}{p(m_k)} \frac{1}{g_{k+1}(\theta'_{k+1})} \frac{h(m'_{k+1} | m_k)}{h(m_k | m'_{k+1})} \right\} \quad (13)$$

where  $q_{\boldsymbol{\theta}_k, m_k}(\mathbf{y})$  indicates the unnormalized likelihood of  $p(\mathbf{y} | \boldsymbol{\theta}_k, m_k)$  (and so forth for the other functions  $q(\cdot)$ ). The reverse move is similar and is accepted with probability  $\alpha^{-1}$ .

The jump within the same model  $m_k$  is accepted with the following probability:

$$\alpha = \min \left\{ 1, \frac{q_{\boldsymbol{\theta}_k, m_k}(\mathbf{y}')}{q_{\boldsymbol{\theta}_k, m_k}(\mathbf{y})} \frac{q_{\boldsymbol{\theta}'_k, m'_k}(\mathbf{y})}{q_{\boldsymbol{\theta}'_k, m'_k}(\mathbf{y}')} \frac{p(\boldsymbol{\theta}'_k | m'_k)}{p(\boldsymbol{\theta}_k | m_k)} \frac{p(m'_k)}{p(m_k)} \frac{g(\boldsymbol{\theta}_k)}{g(\boldsymbol{\theta}'_k)} \right\}. \quad (14)$$

### 3.2 Auto-RJ exchange algorithm

Finding suitable proposals for the jump between models is an very challenging task and is vital in order to ensure adequate mixing of the trans-dimensional Markov chain. In practice, tuning the jump proposals of the pilot-tuned algorithm is very difficult without any information about the posterior density covariance structure. A possible approach would be to use an independence sampler which does not depend on the current state of the MCMC chain but fits a parametric density approximation to the within-model posterior distribution so as to have an acceptance rate as high as possible.

In this spirit, we can propose to jump from  $(\boldsymbol{\theta}_k, m_k)$  to  $(\boldsymbol{\theta}'_h, m'_h)$  using the following jump proposals:

$$h(\boldsymbol{\theta}'_h, m'_h | \boldsymbol{\theta}_k, m_k) = w(\boldsymbol{\theta}'_h | m'_h) h(m'_h | m_k) \quad (15)$$

where  $h(m'_h | m_k)$  represents between-model jump proposal from model  $m_k$  to model  $m'_h$  and  $w(\boldsymbol{\theta}'_h | m'_h)$  is the within-model jump proposal for model  $m'_h$ . As remarked above, the within-model proposals have to be tuned in a sensible way. Posterior density approximations such as standard distributions with parameters determined by the moments of a sample drawn from (12) can be used as within model proposals for each competing model. For example,  $w(\boldsymbol{\theta}_l | m_l)$  can be a normal distribution  $\mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}_l)$

where  $\hat{\boldsymbol{\mu}}_l$  and  $\hat{\boldsymbol{\Sigma}}_l$  are the posterior mean and covariance estimates for each model  $m_l$ . In our experience the choice of normal proposals appear to fit quite well in most of the examples we looked at.

The algorithm can be therefore summarized in two steps: the first step (offline) is used to sample from the posterior (11) of each model  $m_l$  and to estimate the parameters  $\hat{\boldsymbol{\mu}}_l$  and  $\hat{\boldsymbol{\Sigma}}_l$  of the within-model jump proposal; the second step (online) carries out the MCMC computation of (12).

The algorithm can be written in the following concise way:

#### OFFLINE RUN

- (0) ESTIMATION OF  $p(\boldsymbol{\theta}_l|\mathbf{y}, m_l)$  FOR  $l = 1, \dots, H$   
*i* SET  $\hat{\boldsymbol{\mu}}_l = \mathbb{E}(\boldsymbol{\theta}_l|\mathbf{y}, m_l)$  AND  $\hat{\boldsymbol{\Sigma}}_l = \text{Cov}(\boldsymbol{\theta}_l|\mathbf{y}, m_l)$   
*ii* USE  $w(\boldsymbol{\theta}_l|m_l) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_l, \hat{\boldsymbol{\Sigma}}_l)$  AS WITHIN-MODEL JUMP PROPOSALS, WHEN PROPOSING TO JUMP TO MODEL  $m_l$

#### ONLINE RUN

- (1.1) GIBBS UPDATE OF  $(m'_h, \boldsymbol{\theta}'_h, \mathbf{y}')$   
*i* PROPOSE  $m'_h$  FROM THE PRIOR  $p(\cdot)$   
*ii* PROPOSE  $\boldsymbol{\theta}'_h$  WITH PROBABILITY  $w(\cdot|\hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h)$   
*iii* DRAW  $\mathbf{y}'$  FROM  $p(\cdot|\boldsymbol{\theta}'_h, m'_h)$   
(1.2) ACCEPT THE JUMP FROM  $(\boldsymbol{\theta}_k, m_k)$  TO  $(\boldsymbol{\theta}'_h, m'_h)$  WITH PROBABILITY:

$$\min \left\{ 1, \frac{q_{\boldsymbol{\theta}_k, m_k}(\mathbf{y}') q_{\boldsymbol{\theta}'_h, m'_h}(\mathbf{y}) p(\boldsymbol{\theta}'_h|m'_h) p(m'_h) w(\boldsymbol{\theta}_k|\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}{q_{\boldsymbol{\theta}_k, m_k}(\mathbf{y}) q_{\boldsymbol{\theta}'_h, m'_h}(\mathbf{y}') p(\boldsymbol{\theta}_k|m_k) p(m_k) w(\boldsymbol{\theta}'_h|\hat{\boldsymbol{\mu}}_h, \hat{\boldsymbol{\Sigma}}_h)} \right\}. \quad (16)$$

## 4 Estimating model evidence

In this section we present a within-model approach for estimating the evidence  $p(\mathbf{y})$  (For ease of reading, we will omit the conditioning on the model indicator  $m_l$ ). The aim is to provide a useful method for low-dimensional models to use as a “ground-truth” reference to compare with the reversible jump exchange algorithm. The method follows from noticing that for any parameter  $\boldsymbol{\theta}^*$ , equation (2) implies that:

$$p(\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta}^*) \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^*|\mathbf{y})} = \frac{q_{\boldsymbol{\theta}^*}(\mathbf{y})}{z(\boldsymbol{\theta}^*)} \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^*|\mathbf{y})}. \quad (17)$$

This is also the starting point for Chib’s method for estimating the evidence (Chib, 1995). Typically  $\boldsymbol{\theta}^*$  is chosen as a point falling in the high posterior probability region so as to increase the accuracy of the estimate. To estimate (17), the calculation of the intractable likelihood normalizing constant  $z(\boldsymbol{\theta}^*)$  and an estimate of the posterior density  $p(\boldsymbol{\theta}^*|\mathbf{y})$  are required.

## Estimating $z(\boldsymbol{\theta}^*)$ via path sampling

The first problem can be tackled using a path sampling approach. Consider introducing an auxiliary variable  $t \in [0, 1]$ . We consider the following distribution:

$$p_t(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})^t = p(\mathbf{y}|\boldsymbol{\theta}t) = \frac{q_{\boldsymbol{\theta}t}(\mathbf{y})}{z(\boldsymbol{\theta}t)} = \frac{\exp\{(\boldsymbol{\theta}t)^T s(\mathbf{y})\}}{\sum_{\mathbf{y} \in \mathcal{Y}} \exp\{(\boldsymbol{\theta}t)^T s(\mathbf{y})\}} \quad (18)$$

Taking logarithm and differentiating  $z(\boldsymbol{\theta}^*t)$  with respect to  $t$  yields basic identity:

$$\begin{aligned} \frac{d}{dt} \log [z(\boldsymbol{\theta}^*t)] &= \frac{1}{z(\boldsymbol{\theta}^*t)} \frac{d}{dt} z(\boldsymbol{\theta}^*t) \\ &= \frac{1}{z(\boldsymbol{\theta}^*t)} \frac{d}{dt} \sum_{\mathbf{y} \in \mathcal{Y}} \exp\{(\boldsymbol{\theta}^*t)^T s(\mathbf{y})\} \\ &= \frac{1}{z(\boldsymbol{\theta}^*t)} \sum_{\mathbf{y} \in \mathcal{Y}} [\boldsymbol{\theta}^{*T} s(\mathbf{y})] \exp\{(\boldsymbol{\theta}^*t)^T s(\mathbf{y})\} \\ &= \sum_{\mathbf{y} \in \mathcal{Y}} [\boldsymbol{\theta}^{*T} s(\mathbf{y})] p(\mathbf{y}|\boldsymbol{\theta}^*t) \\ &= \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t} [\boldsymbol{\theta}^{*T} s(\mathbf{y})]. \end{aligned} \quad (19)$$

where  $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t}$  denotes the expectation with respect to the sampling distribution  $p(\mathbf{y}|\boldsymbol{\theta}^*t)$ . Therefore integrating (19) from 0 to 1 we have that:

$$\log \left\{ \frac{z(\boldsymbol{\theta}^*)}{z(\mathbf{0})} \right\} = \int_0^1 \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t} [\boldsymbol{\theta}^{*T} s(\mathbf{y})] dt.$$

Now if we choose a discretization of the variable  $t$  such that  $t_0 = 0 < \dots < t_i < \dots < t_I = 1$ , this leads to the following approximation:

$$\log \left\{ \frac{z(\boldsymbol{\theta}^*)}{z(\mathbf{0})} \right\} \approx \sum_{i=0}^{I-1} (t_{i+1} - t_i) \left( \frac{\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t_i} [\boldsymbol{\theta}^{*T} s(\mathbf{y})] + \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t_{i+1}} [\boldsymbol{\theta}^{*T} s(\mathbf{y})]}{2} \right), \quad (20)$$

where  $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t_i} [\boldsymbol{\theta}^{*T} s(\mathbf{y})]$  is equal to  $\boldsymbol{\theta}^{*T} \mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t_i} [s(\mathbf{y})]$  i.e. the expected network statistic counts simulated from  $\boldsymbol{\theta}^*t_i$  multiplied by the constant  $\boldsymbol{\theta}^*$ . Remember that  $z(\mathbf{0})$  is analytically available and it is equal to  $2^{\binom{n}{2}}$  i.e. the number of possible graphs on the  $n$  nodes of the observed network. In terms of computation,  $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t_i} [\boldsymbol{\theta}^{*T} s(\mathbf{y})]$  can be easily estimated using the same procedures used for simulating auxiliary data from the ERGM likelihood. Hence in (20) two types of error emerge: discretization of (4) and Monte Carlo error due to the simulation approximation of  $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}^*t_i} [\boldsymbol{\theta}^{*T} s(\mathbf{y})]$ .

The path of  $t_i$ 's is important for the efficiency of the evidence estimate. For example, we can choose a path of the type  $t_i = (1/I)^c$  where  $c$  is some tuning constant: for  $c = 1$  we have equal spacing of the  $I$  points in the interval  $[0, 1]$ , for  $c > 1$  we have that the  $t_i$ 's are chosen with high frequency close to 0 and for  $0 < c < 1$  we have that the  $t_i$ 's are chosen with high frequency close to 1.

## Estimating $p(\boldsymbol{\theta}^*|\mathbf{y})$

A sample from the posterior  $p(\boldsymbol{\theta}|\mathbf{y})$  can be gathered (via the exchange algorithm, for example) and used to calculate a kernel density estimate of the posterior probability at the point  $\boldsymbol{\theta}^*$ . In practice, because of the curse of dimensionality, this implies that the method cannot be used, for models with greater than 5 parameters. In this paper we used the fast and easy to use `np` package for R (Hayfield and Racine, 2008) to perform a nonparametric density estimation of the posterior  $p(\boldsymbol{\theta}^*|\mathbf{y})$ .

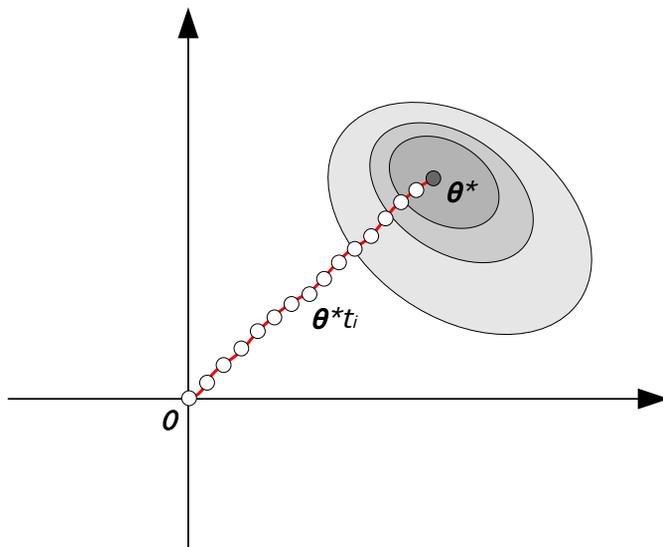


Figure 2: Path sampling: for each  $\boldsymbol{\theta}^*$  we estimate  $z(\boldsymbol{\theta}^*)$  via path sampling using the expected network statistics simulated from some points  $\boldsymbol{\theta}^{*t_i}$  along the line connecting  $\mathbf{0}$  to  $\boldsymbol{\theta}^*$ .

## 5 Applications

### 5.1 Gahuku-Gama system

The Gahuku-Gama system (Read, 1954) of the Eastern Central Highlands of New Guinea was used by Hage and Harary (1984) to describe an alliance structure among 16 sub-tribes of Eastern Central Highlands of New Guinea (Figure 3). The system has been split into two network: the “Gamaneg” graph for antagonistic (“hina”) relations and the “Gamapos” for alliance (“rova”) relations. An important feature of these structures is the fact that the enemy of an enemy can be either a friend or an enemy.

#### 5.1.1 Gamaneg

We first focus on the Gamaneg network by using the 3 competing models specified in Table 1 using the following network statistics:

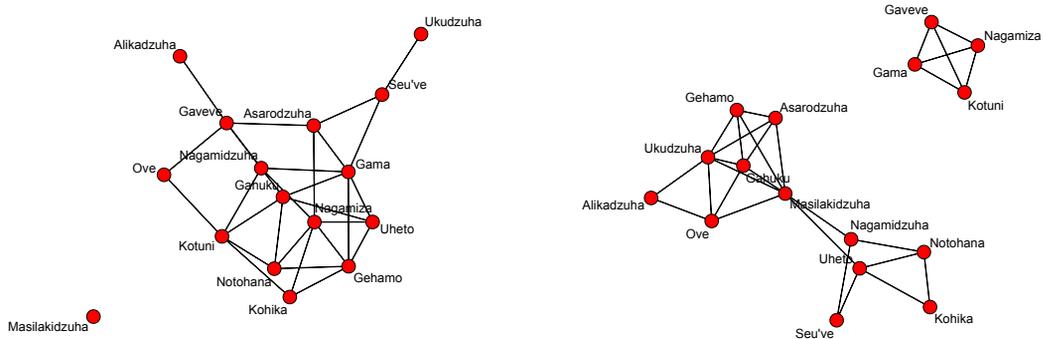


Figure 3: Gahuku-Gama system graphs: Gamaneg (left) and Gamapos (right).

$$\begin{aligned}
 \text{edges} & \quad \sum_{i < j} y_{ij} \\
 \text{triangles} & \quad \sum_{i < j < k} y_{jk} y_{ik} y_{ij} \\
 \text{4-cycle} & \quad \sum_{i < j < l < k} y_{ij} y_{jl} y_{lk} y_{ki}
 \end{aligned}$$

We are interested to understand if the transitivity effect expressed by triad closure (triangle) and 4-cycle is a closed structure that can sustain mutual social monitoring and influence (Pattison and Robins, 2002).

Model $m_1$	$\mathbf{y} \sim$ edges
Model $m_2$	$\mathbf{y} \sim$ edges + triangles
Model $m_3$	$\mathbf{y} \sim$ edges + triangles + 4-cycle

Table 1: Competing models.

Both the pilot-tuned RJ and auto-RJ exchange algorithms were run for 100,000 iterations using very flat normal parameter priors  $p(\boldsymbol{\theta}_l | m_l) \sim \mathcal{N}(0, 100 \mathbf{I}_l)$  for each model  $m_l$  where  $\mathbf{I}_l$  is the identity matrix of size equal to the number of dimensions of model  $m_l$ , uniform model priors and 3,000 iterations for the auxiliary network simulation. The proposal distributions of the pilot-tuned RJ has been empirically tuned so as to get reasonable acceptance rates for each competing model. The offline step of the auto-RJ consisted of gathering an approximate sample from  $p(\boldsymbol{\theta} | \mathbf{y})$  and then estimating the posterior moments  $\hat{\boldsymbol{\mu}}_l$  and  $\hat{\boldsymbol{\Sigma}}_l$  for each of the three models using the parallel ADS update procedure. The exchange algorithm was run for  $1,000 \times D_l$  iterations (discarding the first  $100 \times D_l$  iterations as burn-in) where  $D_l$  is the dimension of the  $l$ -th model using the population MCMC approach described in Caimo and Friel (2011). We tuned the parallel ADS move factors  $\gamma$  so as to get a reasonable acceptance rate during the offline step of the estimation. The accuracy of the estimates  $\hat{\boldsymbol{\mu}}_l$  and  $\hat{\boldsymbol{\Sigma}}_l$  depends on the number of iterations of the auto-RJ offline run. In this example, the above number of iterations  $1,000 \times D_l$  of has been empirically shown to be sufficient for each competing model  $m_l$ . Tables 2 and 3 report the posterior parameter estimates of the model selected for the pilot-tuned RJ and auto-RJ. Figure 4 shows the results

from the pilot-tuned RJ: model posterior diagnostics plot and the parameter posterior diagnostics plot. Figure 5 shows the same plots from auto-RJ. Between-model and within-model acceptance rates (reported in Table 3) are calculated as the proportions of accepted moves from  $(\theta_k, m_k)$  to model  $(\theta'_h, m'_h)$  for each  $k : k \neq h$  and when  $k = h$  respectively. The mixing of the auto-RJ algorithm within each model is faster than the pilot-tuned RJ algorithm due to the good approximation to the posterior distribution. The pilot-tuned algorithm took about 24 minutes to complete the estimation and the auto-RJ took about 31 minutes (including the offline step).

Parameter	Pilot-tuned RJ		Auto-RJ	
	Post. Mean	Post. Sd.	Post. Mean	Post. Sd.
Model $m_1$				
$\theta_1$ (edge)	-1.15	0.21	-1.15	0.21
Model $m_2$				
$\theta_1$ (edge)	-0.97	0.36	-0.96	0.37
$\theta_2$ (triangle)	-0.31	0.41	-0.29	0.37
Model $m_3$				
$\theta_1$ (edge)	-0.98	0.51	-1.15	0.37
$\theta_2$ (triangle)	-0.76	0.47	-0.31	0.42
$\theta_3$ (4-cycle)	-0.05	0.12	0.02	0.17

Table 2: Summary of posterior parameter estimates.

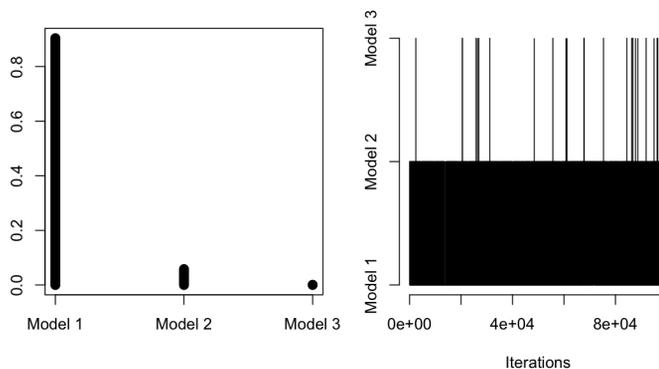
Within-model	Pilot-tuned RJ	Auto-RJ
Model $m_1$	0.14	0.62
Model $m_2$	0.11	0.42
Model $m_3$	0.00	0.48
Between-model	0.07	0.04

Table 3: Acceptance rates.

	Pilot-tuned RJ	Auto-RJ
$\widehat{BF}_{1,2}$	14.46	21.68
$\widehat{BF}_{1,3}$	1506.43	1425.77
$\widehat{p}(m_1 \mathbf{y})$	0.93	0.95
$\widehat{p}(m_2 \mathbf{y})$	0.06	0.04
$\widehat{p}(m_3 \mathbf{y})$	0.01	0.01

Table 4: Bayes factor and posterior model probability estimates.

**MCMC output : posterior model probabilities**



**MCMC output for Model 1 : posterior parameter probabilities**

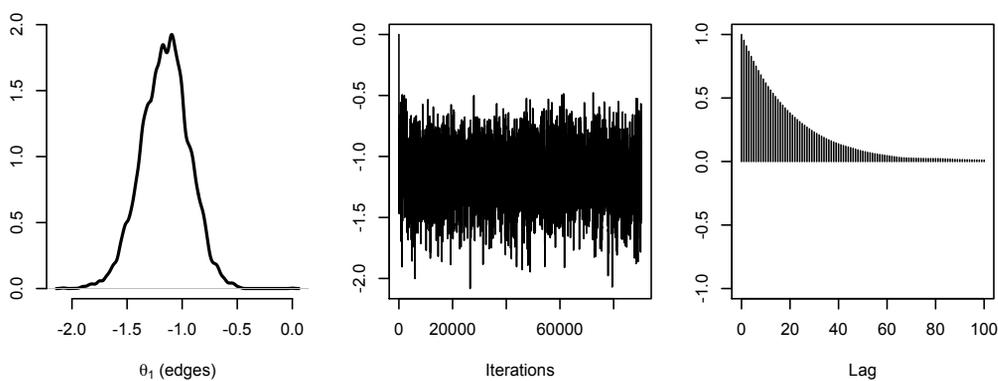
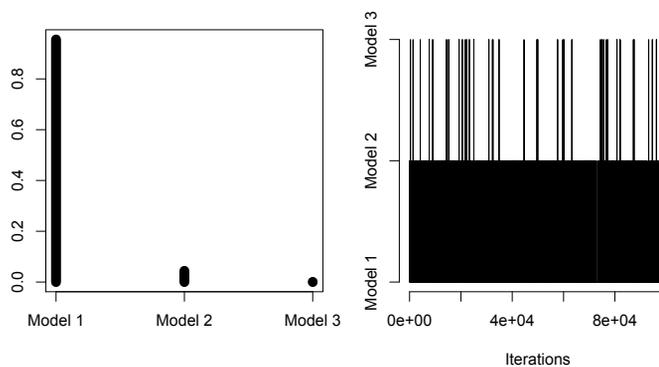


Figure 4: Pilot-tuned RJ exchange algorithm output: posterior model probabilities (top) and posterior parameter probabilities (bottom).

**MCMC output : posterior model probabilities**



**MCMC output for Model 1 : posterior parameter probabilities**

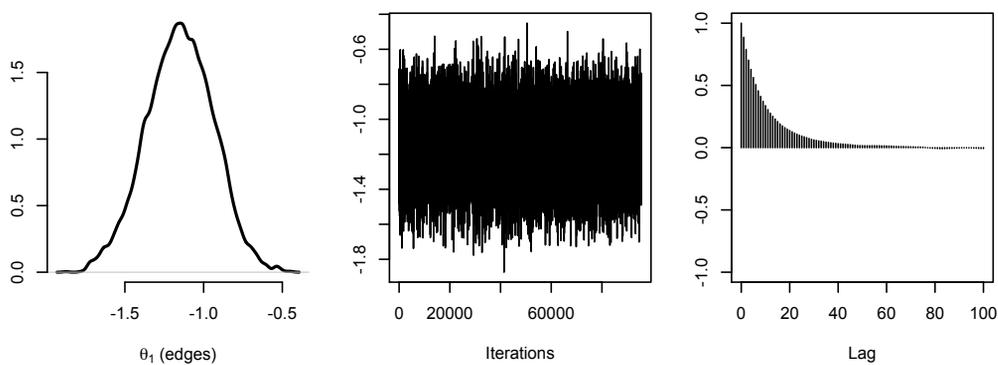


Figure 5: Auto-RJ exchange algorithm output: posterior model probabilities (top) and posterior parameter probabilities (bottom).

In terms of the calculating the evidence based on path sampling, Figure 6 shows the behaviour of  $\mathbb{E}_{\mathbf{y}|\theta^*t} [\theta^{*T} s(\mathbf{y})]$  for 50 equally-spaced path points  $t_i$  from 0 to 1. The larger the number of temperatures  $I$  and the number of simulated networks, the more precise the estimate of the likelihood normalizing constant and the longer the computing effort. In this example we estimated (19) using 100 path points and sampling 500 network statistics for each of them. In this case, this setup has been empirically shown to be sufficiently accurate. We set  $c$  to be equal to 1 for all the models. However different choices for  $c$  do not seem to have a big influence on the estimation results if  $I$  is large enough.

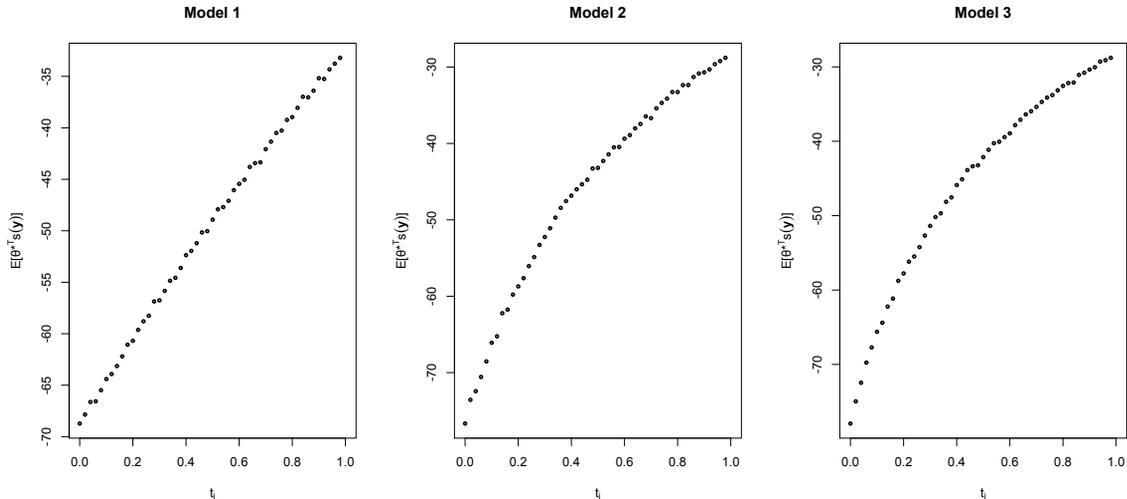


Figure 6:  $\mathbb{E}[\theta^{*T} s(\mathbf{y})]$  estimated from a ladder of 50 equally-spaced path points.

A nonparametric density estimation of  $p(\theta|\mathbf{y})$  for each competing model was implemented using approximate posterior samples gathered from the output of the exchange algorithm. Bayes Factor estimates for different sample sizes (which are increasing with the number of model dimension) are reported in Table 5. The results are consistent with the ones obtained by RJ exchange algorithms displayed in Table 4. The evidence-based approach took about a few seconds to estimate model evidence for  $m_1$  and  $m_2$  and about 6 minutes for model  $m_3$  using the biggest sample sizes displayed in Table 4.

According to the scale of Kass and Raftery (1995) there is positive/strong evidence in favor of model  $m_1$  which is the one including the number of edges. Thus in this case the only strong effect of the antagonistic structure of the Gahuku-Gama tribes is represented by the low edge density.

### 5.1.2 Gamapos

In this example, we considered the same competing models of Table 1. In this case the pilot-tuned RJ exchange algorithm was infeasible to be used effectively as it turned out to be very sensitive to the choice of the jump proposal. We used the auto-RJ exchange

	Sample sizes			
Model $m_1$	100	500	1,000	5,000
Model $m_2$	150	750	1,500	7,500
Model $m_3$	200	1,000	2,000	10,000
$\widehat{BF}_{1,2}$	18.83	18.72	17.75	19.09
$\widehat{BF}_{1,3}$	1029.67	1483.52	1363.91	1390.08

Table 5: Bayes Factor estimates for increasing values of sample sizes used for the posterior density estimation.

algorithm with the same setup of the previous example. The output from auto-RJ exchange algorithm is displayed in Figure 8 and the parameter posterior estimates in Table 6.

Parameter	Post. Mean	Post. Sd.
Model $m_3$ (within-model acc. rate: 0.3)		
$\theta_1$ (edge)	-2.41	0.45
$\theta_2$ (triangle)	2.91	0.71
$\theta_3$ (4-cycle)	-0.66	0.22
Model $m_1$ (within-model acc. rate: 0.64)		
$\theta_1$ (edge)	-1.15	0.20
Model $m_2$ (within-model acc. rate: 0.3)		
$\theta_1$ (edge)	-1.69	0.35
$\theta_2$ (triangle)	0.48	0.20
Between-model acc. rate: 0.03		

Table 6: Summary of posterior parameter estimates and acceptance rates.

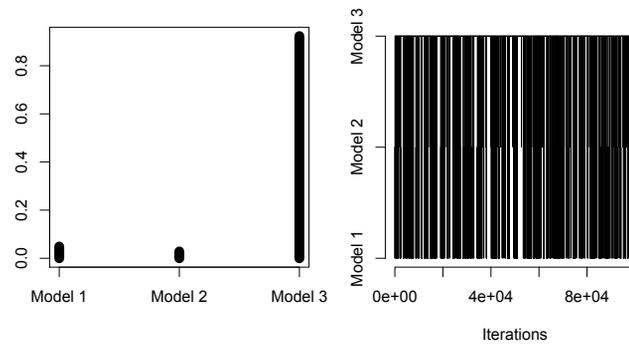
We also calculated the evidence for each models following the same setup of the Gamaneg example. In Table 7 are reported the Bayes Factor estimates of the auto-RJ exchange algorithm and evidence-based method using the biggest sample sizes for the posterior density estimation of the previous example.

	Auto-RJ algorithm	Evidence-based method
$BF_{3,1}$	17.83	19.31
$BF_{3,2}$	34.81	32.82

Table 7: Bayes factors estimates.

In the Gamapos network the transitivity and the 4-cycle structure are important features of the network. The tendency to a low density of edges and 4-cycles expressed by the negative posterior mean of the first and third parameters is balanced by a

MCMC output : posterior model probabilities



MCMC output for Model 3 : posterior parameter probabilities

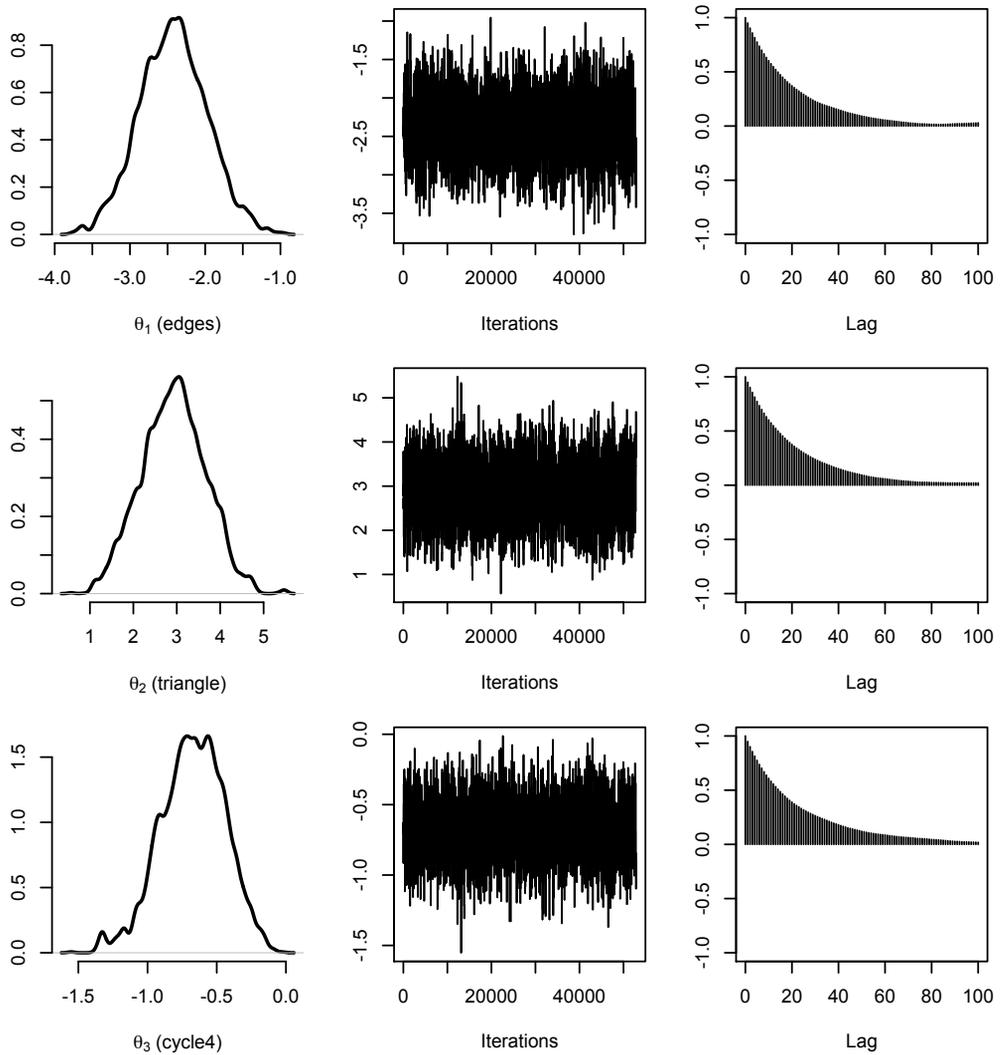


Figure 7: Auto-RJ exchange algorithm output: posterior model probabilities (top) and posterior parameter probabilities (bottom)

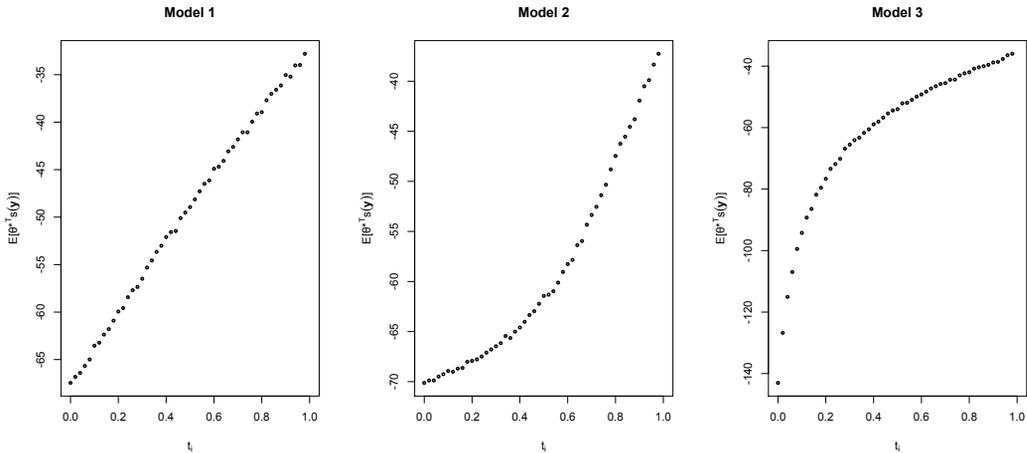


Figure 8:  $\mathbb{E}[\boldsymbol{\theta}^{*T} \mathbf{s}(\mathbf{y})]$  estimated from a ladder of 50 equally-spaced path points.

propensity for local triangles which gives rise to the formation of small well-defined alliances.

## 5.2 Collaboration between Lazega's lawyers

The Lazega network data collected by Lazega (2001) and displayed in Figure 9 represents the symmetrized collaboration relations between the 36 partners in a New England law firm, where the presence of an edge between two nodes indicates that both partners collaborate with the other.

In this example we want to compare 4 models (Table 8) using the edges, geometrically weighted degrees and geometrically weighted edgewise shared partners (Snijders et al., 2006):

edges	$\sum_{i < j} y_{ij}$
geometrically weighted degree (gwd)	$e^{\phi_u} \sum_{k=1}^{n-1} \left\{ 1 - (1 - e^{-\phi_u})^k \right\} D_k(\mathbf{y})$
geometrically weighted edgewise shared partner (gwesp)	$e^{\phi_v} \sum_{k=1}^{n-2} \left\{ 1 - (1 - e^{-\phi_v})^k \right\} EP_k(\mathbf{y})$

where  $\phi_u = \log(2)$ ,  $\phi_v = \log(2)$ ,  $D_k(\mathbf{y})$  is the number of pairs that have exactly  $k$  common neighbors and  $EP_k(\mathbf{y})$  is the number of connected pairs with exactly  $k$  common neighbors.

As happened in the previous example, the pilot-tuned RJ exchange algorithm proved to be ineffective due to the difficulty of the tuning problem. The auto-RJ exchange algorithm was run for 100,000 iterations using the same flat normal priors of the previous examples and 20,000 auxiliary iterations for network simulation. The offline run consisted of estimating  $\hat{\boldsymbol{\mu}}_l$  and  $\hat{\boldsymbol{\Sigma}}_l$  for each of the 4 models by using  $6,000 \times D_l$  main iterations (discarding the first  $1,000 \times D_l$  iterations as burnin). The algorithm

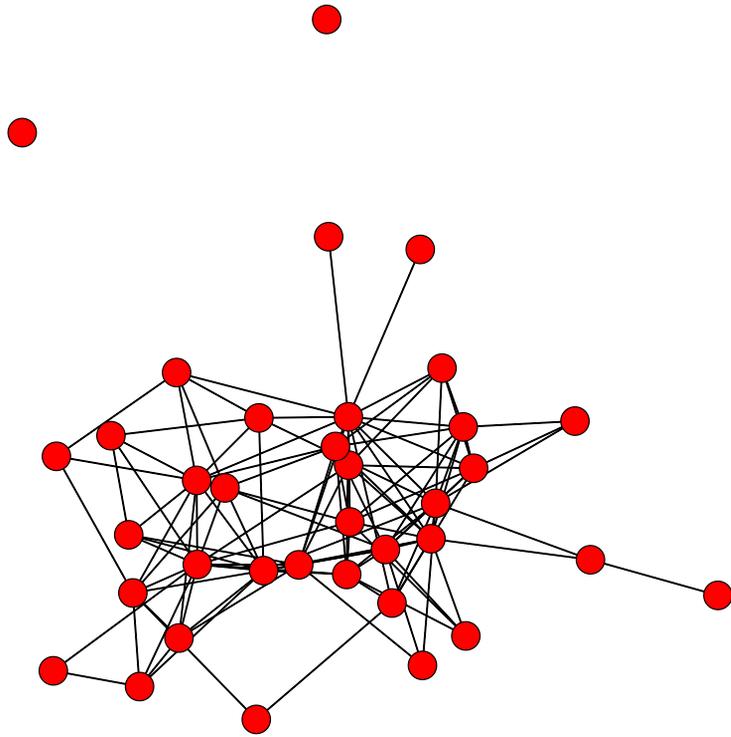


Figure 9: Lazega's lawyers graph.

Model $m_1$	$\mathbf{y} \sim$ edges
Model $m_2$	$\mathbf{y} \sim$ edges + gwesp( $\log(2)$ )
Model $m_3$	$\mathbf{y} \sim$ edges + gwesp( $\log(2)$ ) + gwd( $\log(2)$ )
Model $m_4$	$\mathbf{y} \sim$ edges + gwd( $\log(2)$ )

Table 8: Competing models.

took about 1 hour and 50 minutes to complete the estimation, the results of which are displayed in Figure 10 and Table 9.

Parameter	Post. Mean	Post. Sd.
Model $m_2$ (within-model acc. rate: 0.24)		
$\theta_1$ (edge)	-3.93	0.33
$\theta_2$ (gwesp( $\log(2)$ ))	1.15	0.16
Model $m_3$ (within-model acc. rate: 0.26)		
$\theta_1$ (edge)	-4.54	0.56
$\theta_2$ (gwesp( $\log(2)$ ))	-1.39	0.23
$\theta_3$ (gwd( $\log(2)$ ))	0.79	0.62
Between-model acc. rate: 0.03		

Table 9: Summary of posterior parameter estimates and acceptance rates.

The evidence-based algorithm was carried out using 200 path points from each of which we sampled 500 networks. The results are reported in Table 10. The algorithm took 25 seconds to estimate the evidence for model  $m_1$ , 8 minutes for model  $m_2$ , 9 minutes for model  $m_3$ , 1 minute for model  $m_4$ .

	Auto-RJ algorithm	Evidence-based method
$BF_{2,1}$	$> 10^6$	$> 10^6$
$BF_{2,3}$	5.72	4.65
$BF_{2,4}$	$> 10^6$	$> 10^6$

Table 10: Bayes Factor estimates.

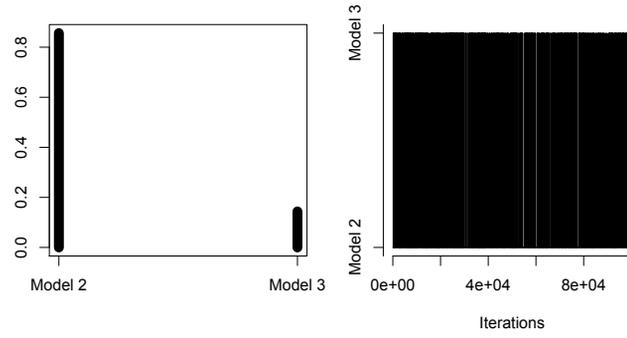
Table 10 displays the Bayes Factor for the comparison between model  $m_2$  (best model) against the others. There is positive evidence to reject model  $m_3$  and very strong evidence to models  $m_1$  and  $m_4$ .

We can therefore conclude that the low density effect expressed by the negative edge parameter combined with the positive transitivity effect expressed by the geometrically weighted edgewise partners parameter are strong structural features not depending on popularity effect expressed by the weighted degrees. These results are in agreement with the findings reported in the literature (see Snijders et al. (2006) and Hunter and Handcock (2006)).

## 6 Discussion

In this paper, we have presented two novel methods to Bayesian model selection for exponential random graph models. The first one is an across-model approach based on a trans-dimensional extension of the exchange algorithm for exponential random graph models of Caimo and Friel (2011). An independence sampler making use of a

MCMC output : posterior model probabilities



MCMC output for Model 2 : posterior parameter probabilities

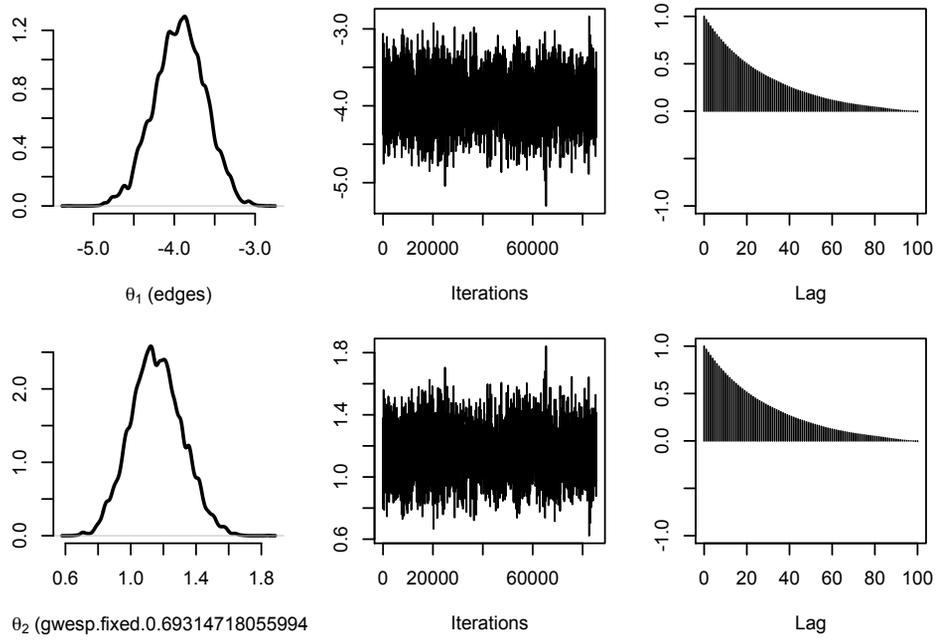


Figure 10: Auto-RJ exchange algorithm output: posterior model probabilities (top) and posterior parameter probabilities (bottom).

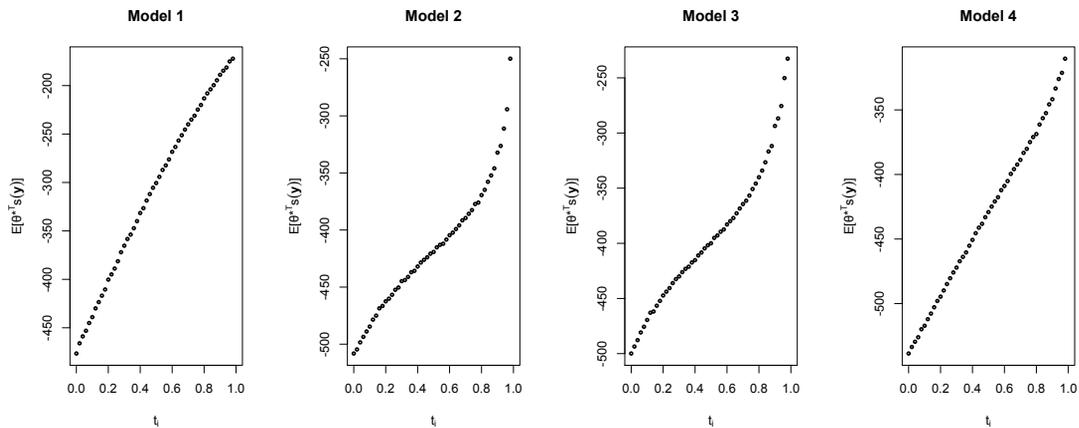


Figure 11:  $\mathbb{E}[\boldsymbol{\theta}^{*T} s(\mathbf{y})]$  estimated from a ladder of 50 equally-spaced path points.

parametric approximation of the posterior is proposed in order to overcome the issue of tuning of the jump proposal distributions and increase within-model acceptance rates. A within-model approach for estimating the model evidence relying on the path sampling approximation of the likelihood normalizing constant and posterior density estimation via nonparametric techniques is also proposed. Both methods have been illustrated by three examples.

**Acknowledgement** Alberto Caimo was supported by an IRCSET Embark Initiative award and Nial Friel’s research was supported by a Science Foundation Ireland Research Frontiers Program grant, 09/RFP/MTH2199.

## References

- Brooks, S. P., Giudici, P., and O., R. G. (2003), “Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion),” *Journal of the Royal Statistical Society, Series B*, 65, 3–57.
- Caimo, A. and Friel, N. (2011), “Bayesian inference for exponential random graph models,” *Social Networks*, 33, 41 – 55.
- Chib, S. (1995), “Marginal Likelihood from the Gibbs Output,” *Journal of the American Statistical Association*, 90, 1313–1321.
- Chib, S. and Jeliazkov, I. (2001), “Marginal Likelihood From the MetropolisHastings Output,” *Journal of the American Statistical Association*, 96, 270–281.
- Friel, N. and Pettitt, A. N. (2008), “Marginal likelihood estimation via power posteriors,” *Journal of the Royal Statistical Society, Series B*, 70, 589–607.

- Friel, N. and Wyse, J. (2012), “Estimating the statistical evidence – a review,” *Statistica Neerlandica*, (to appear).
- Green, P. J. (1995), “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,” *Biometrika*, 82, 711–732.
- Green, P. J., Hjort, N. L., and Richardson, S. (eds.) (2003), *Highly Structured Stochastic Systems*, Oxford University Press, chap. Trans-dimensional Markov chain Monte Carlo.
- Hage, P. and Harary, F. (1984), *Structural Models in Anthropology*, Cambridge University Press.
- Handcock, M. S. (2003), “Assessing Degeneracy in Statistical Models of Social Networks,” *Working Paper no.39*, Center for Statistics and the Social Sciences, University of Washington.
- Hayfield, T. and Racine, J. S. (2008), “Nonparametric Econometrics: The np Package,” *Journal of Statistical Software*, 27.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), “Bayesian Model Averaging: A Tutorial,” *Statistical Science*, 14, 382–401.
- Hunter, D. R. and Handcock, M. S. (2006), “Inference in curved exponential family models for networks,” *Journal of Computational and Graphical Statistics*, 15, 565–583.
- Kass, R. E. and Raftery, A. E. (1995), “Bayes factors,” *Journal of the American Statistical Association*, 90, 773–795.
- Lazega, E. (2001), *The collegial phenomenon : the social mechanisms of cooperation among peers in a corporate law partnership*, Oxford University Press.
- Neal, R. M. (2001), “Annealed importance sampling,” *Statistics and Computing*, 11, 125–139.
- Pattison, P. and Robins, G. L. (2002), “Neighbourhood-based models for social networks,” *Sociological Methodology*, 32, 301–337.
- Read, K. E. (1954), “Cultures of the Central Highlands, New Guinea,” *Southwestern Journal of Anthropology*, 10, 1–43.
- Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), “On the geometry of discrete exponential random families with application to exponential random graph models,” *Electronic Journal of Statistics*, 3, 446–484.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007), “An introduction to exponential random graph models for social networks,” *Social Networks*, 29, 169–348.

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