Efficient Bayesian inference for exponential random graph models by correcting the pseudo-posterior distribution

Lampros Bouranis*, Nial Friel, Florian Maire

School of Mathematics and Statistics & Insight Centre for Data Analytics, University College Dublin, Ireland

Abstract

Exponential random graph models are an important tool in the statistical analysis of data. However, Bayesian parameter estimation for these models is extremely challenging, since evaluation of the posterior distribution typically involves the calculation of an intractable normalizing constant. This barrier motivates the consideration of tractable approximations to the likelihood function, such as the pseudolikelihood function, which offers an approach to constructing such an approximation. Naive implementation of what we term a pseudo-posterior resulting from replacing the likelihood function in the posterior distribution by the pseudolikelihood is likely to give misleading inferences. We provide practical guidelines to correct a sample from such a pseudo-posterior distribution so that it is approximately distributed from the target posterior distribution and discuss the computational and statistical efficiency that result from this approach. We illustrate our methodology through the analysis of real-world graphs. Comparisons against the approximate exchange algorithm of Caimo and Friel (2011) are provided, followed by concluding remarks.

Keywords: Exponential random graph models, Intractable normalizing constants, Large networks, Logistic regression, Pseudolikelihood, Tractable approximation.

1. Introduction

The study of networks is central to a broad range of applications including epidemiology (dynamics of disease spread), genetics (protein interactions), telecommunications (worldwide web connectivity, phone calls) and social science (Facebook, Twitter, LinkedIn), among others. The high–dimensionality and complexity of such structures poses a real challenge to modern statistical computing methods.

Exponential random graph (ERG) models play an important role in network analysis since they allow for complex correlation patterns between the nodes of the network. However this model presents several difficulties in practice, mainly due to the fact that likelihood function can only be specified up to a parameter dependent normalizing constant. This impacts upon maximum likelihood estimation which is difficult to perform for larger networks, where the full likelihood

*Corresponding author. Address: Insight, O'Brien Centre for Science, Science Centre East, University College Dublin, Belfield, Dublin 4, Ireland

Email addresses: lampros.bouranis@insight-centre.org (Lampros Bouranis), nial.friel@ucd.ie (Nial Friel), florian.maire@ucd.ie (Florian Maire)

function is available but it is just too complex to be evaluated. Robins et al. (2007b) presented various approaches for overcoming this problem.

The Bayesian paradigm has been used to infer Exponential random graph models. The challenge of carrying out Bayesian estimation for these models has received attention from the statistical community in the recent past (Caimo and Friel, 2011). The main challenge encountered by such a Bayesian setting is the evaluation of the posterior that typically involves the calculation of an intractable normalizing constant. Sampling from distributions with intractable normalization can be done via Markov chain Monte Carlo methods (MCMC) and especially with the celebrated Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970).

Nevertheless, the normalizing term in the ERG probability distribution is a function of the model parameters and thus does not cancel as usual in the standard Metropolis–Hastings acceptance probability. This gives rise to a target distribution $\pi(\theta \mid y)$ which is termed *doubly–intractable* (Murray et al., 2006), where at each iteration of the Markov chain Monte Carlo scheme intractability of the normalizing term of the likelihood model within the posterior and intractability of the posterior normalizing term must be handled. More sophisticated Markov chains, such as the Exchange algorithm (Møller et al., 2006; Murray et al., 2006) have been proposed to sample from those doubly–intractable targets. Here again, these methods are not directly applicable in the ERG context as they require independent and identically distributed (iid) draws from the likelihood, which is not feasible for this type of models.

This motivated the Approximate Exchange algorithm of Caimo and Friel (2011) which substitutes iid draws from the likelihood with draws from an auxiliary Markov chain admitting the ERG likelihood as limiting distribution. Previous studies have shown that convergence of sampling from the ERG likelihood through Markov chain is likely to be exponentially slow (Bhamidi et al., 2011). This is likely to lead to increased computational burden when analyzing graphs with complex dependencies.

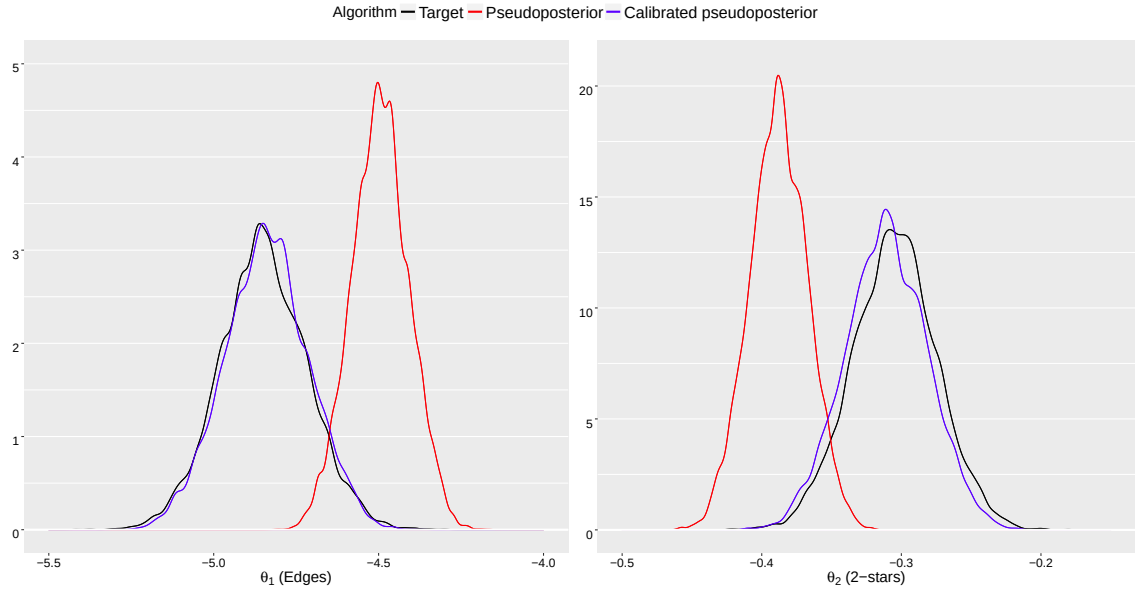

Figure 1: International E–road network: marginal density estimates of the posterior distribution (based on a long and computationally expensive MCMC run) (black curve); misspecified pseudo–posterior density estimates (where the pseudolikelihood has replaced the likelihood) (red curve); calibrated pseudo–posterior density estimates (blue curve). One can see that the calibration step has resulted in density estimates which are very close to the target posterior.

Increasing computational complexity has motivated the development of misspecified but tractable and computationally affordable models. From this perspective, the computational tractability of the pseudolikelihood function and the simplicity in defining the objective function seem to make it a tempting alternative to the full likelihood function when dealing with data with such a complex structure (Robins et al., 2007a; Handcock et al., 2007). The use of such an approximation to the likelihood should be treated with caution, though, as discussed by van Duijn et al. (2009); their work studied the quality of the pseudolikelihood approximation to the true likelihood, concluding that it may under-estimate endogenous network formation processes. Despite its use by practitioners in the frequentist setting to fit ERG models, little is known about the pseudolikelihood approximation efficiency when embedded in a Bayesian posterior distribution. Our empirical analysis shows that the Bayesian estimators resulting from using the pseudolikelihood function as a plug-in for the true likelihood function are biased and their variance can be underestimated. However, the calibration procedure which we develop shows how to correct a sample from this so-called pseudo-posterior distribution, so that it is approximately distributed from the true posterior distribution.

We propose a novel methodology that falls into the area of MCMC targeting a misspecified posterior: a Metropolis–Hastings sampler whose stationary distribution we refer to as a *pseudo-posterior* (a posterior distribution where the likelihood is replaced by a pseudolikelihood approximation). Such a method is fast and overcomes double intractability but is also *noisy*, in the sense that it samples from an approximation of the true posterior distribution, π. A graphical illustration of marginal posterior densities under misspecification can be seen in Fig. 1. This example involves a two–dimensional model which we provide details of in Section 5.2. Here the misspecification of the actual posterior yields a disastrous approximation; compare the black and red curves. It is evident that a sampler which targets an approximate posterior resulting from replacing the intractable likelihood with a pseudolikelihood approximation leads to biased posterior mean estimates and considerably underestimated posterior variances. Nevertheless, we will present a correction method that allows to calibrate the sample to the true density; warping the red to the blue curve which is now a sensible approximation to the black one.

The aim of this paper is to exploit the use of pseudolikelihoods in Bayesian inference of Exponential random graph models. Our work explores the computational efficiency of the resulting Markov chains, and the trade–off between computational and statistical efficiency in estimates derived from such pseudo–posteriors in comparison to the approximate exchange algorithm of Caimo and Friel (2011). We present the reader with a viable approach to calibrate the posterior distribution resulting from using a misspecified likelihood function in an efficient manner (Fig. 1), while providing the theoretical framework for this approach.

The outline of the article is as follows. A basic description of Exponential random graph models is given in Section 2. The pseudolikelihood function as a surrogate for the true likelihood is introduced in Section 3. In Section 4 we formulate the Bayesian model in the presence of likelihood misspecification, and discuss the theoretical properties and practical aspects of the calibration of the posterior distribution. In Section 5, we illustrate our methods through numerical examples involving real networks. We conclude the paper in Section 6 with final remarks.

2. Exponential Random Graph Models

Networks are relational data represented as graphs, consisting of nodes and edges. Many probability models have been proposed in order to understand, summarize and forecast the general

structure of graphs by utilizing their local properties. Among those, Exponential random graph models play an important role in network analysis since they can represent transitivity and other structural features in network data that define complicated dependence patterns not easily modeled by more basic probability models (Wasserman and Pattison (1996), see also Robins et al. (2007b) for a review and the references therein for more details).

Let $\mathcal{Y}$ denote the set of all possible graphs on n nodes. The network topology structure is measured by a $n \times n$ random adjacency matrix Y of a graph on n nodes (actors) and a set of edges (relationships). The presence or absence of a tie between node i and j is coded as $Y_{ij} = 1$ if the dyad (i, j) is connected, $Y_{ij} = 0$ otherwise. An edge connecting a node to itself is not permitted so $Y_{ii} = 0$.

ERG models are a particular class of discrete linear exponential families which represent the probability distribution of a random network graph using a likelihood function, expressed as:

$$p(y \mid \theta) = \frac{q(y \mid \theta)}{z(\theta)} = \frac{\exp\left\{\theta^{\mathrm{T}} s(y)\right\}}{\sum_{y \in \mathcal{Y}} \exp\left\{\theta^{\mathrm{T}} s(y)\right\}}, \qquad \theta^{\mathrm{T}} s(y) = \sum_{j=1}^{d} \theta_j s_j(y), \tag{1}$$

where $q(\mathrm{y} \mid \theta)$ is the unnormalized likelihood. In this paper we deal with ERG models that are edge-dependent, in which case the likelihood is intractable and the pseudolikelihood provides an inaccurate approximation. For such models, $s(y) = \{s_1(y), \ldots, s_d(y)\}$ is a known vector of overlapping sub–graph configurations/ sufficient statistics (e.g. the number of edges, degree statistics, triangles, etc.) and $\theta \in \Theta \subseteq \mathbb{R}^{\mathrm{d}}$ are the model parameters. More examples can be found at Snijders et al. (2006) and Hunter and Handcock (2006). A positive parameter value for $\theta_i \in \theta$ results in a tendency for the certain configuration corresponding to $s_i(y)$ to be observed in the data than would otherwise be expected by chance.

Despite their popularity, Bayesian parameter estimation for these models is challenging, since evaluation of the posterior typically involves the calculation of an intractable normalizing constant $z(\theta)$. Its evaluation is extremely difficult for all but trivially small graphs since this sum involves $2^{\binom{n}{2}}$ terms for undirected graphs. The intractable normalizing constant makes inference difficult for both frequentist and Bayesian approaches.

3. Pseudolikelihood Approximation of Likelihood Function

Complex dependencies in several applications lead to full likelihoods that may be difficult, or even impractical, to compute. In such situations it may be useful to resort to approximate likelihoods, if it is possible to compute the likelihood for some subsets of the data.

The pseudolikelihood of Besag (1975, 1977) was the earliest method of parameter estimation that was proposed for models with complicated dependence structure for which the likelihood function could not be calculated exactly, or even approximated well in a reasonable amount of time. The composite likelihood of Lindsay (1988) is a generalization of the pseudolikelihood (see Varin et al. (2011) for a recent review).

Strauss and Ikeda (1990) have applied the idea of pseudolikelihood to social networks. The pseudolikelihood method defines the approximation to the full joint distribution as the product of the full conditionals for individual observations/ dyads:

$$p_{\mathrm{PL}}(y \mid \theta) = \prod_{i \neq j} p(y_{ij} \mid y_{-ij}, \theta) = \prod_{i \neq j} \frac{p(y_{ij} = 1 \mid y_{-ij}, \theta)^{y_{ij}}}{\{1 - p(y_{ij} = 1 \mid y_{-ij}, \theta)\}^{y_{ij}-1}}, \tag{2}$$

where y_{-ij} denotes $\mathrm{y}\backslash\{\mathrm{y}_{ij}\}$. The distribution of the Bernoulli variable Y_{ij}, conditional on the rest of the network, can be easily calculated under an alternative specification of model (1). Let us define the vector of change statistics as

$$\delta_s(y)_{ij} = s(y_{ij}^+) - s(y_{ij}^-).$$

This vector is associated with a particular pair (ordered or unordered, depending on whether the network is directed or undirected) of nodes and it equals the change in the vector of network sufficient statistics when that pair is toggled from a 0 (no edge, y_{ij}^-) to a 1 (edge, y_{ij}^+), holding the rest of the network, $y_{-ij} = y\backslash\{y_{ij}\}$, fixed. Using such a reparameterization allows to express the distribution of the Bernoulli variable Y_{ij}, conditional on the rest of the network, as a function of the change statistics vector:

$$p(y_{ij} = 1 \mid y_{-ij}, \theta) = \mathrm{logit}^{-1}\left\{\theta^{\mathrm{T}}\delta_s(y)_{ij}\right\}. \tag{3}$$

This expression is extremely convenient as the values $\{p(y_{ij} = 1 \mid y_{-ij}, \theta)\}_{i \neq j}$, referred to as the predictor matrix, are tractable, fast to compute and require the computation of the change statistics only once up front. The predictor matrix for a new parameter can be updated just by computing the scalar product $\theta^{\mathrm{T}}\delta_s(y)$ for the new parameter. As a consequence, estimating the pseudolikelihood $p_{\mathrm{PL}}(y \mid \theta)$ in Eq. (2) for any $\theta \in \Theta$ is effortless.

However, the method implicitly relies on the strong and often unrealistic assumption of independent dyads, conditionally on their neighbours. Its properties are poorly understood (van Duijn et al., 2009): it does not directly maximize the likelihood and in empirical comparisons (Corander et al., 2002) has larger variability than the MLE. In the context of the autologistic distribution in spatial statistics, Friel et al. (2009) showed that the pseudolikelihood estimator can lead to inefficient estimation.

From Eq. (2) and (3), the pseudolikelihood for model (1) is identical to the likelihood for a logistic regression model. A detailed discussion on the use of the pseudo-likelihood and its relationship with logistic regression is provided by Wasserman and Pattison (1996). In that setting, the binary response data consist of the off–diagonal elements of the observed adjacency matrix (Cranmer and Desmarais, 2011). Each row of the matrix of change statistics is associated with the vector of change statistics for a particular pair of nodes and can also accommodate for individual attributes and dyadic covariates. This kind of formulation leads to efficient computational algorithms.

4. Bayesian Inference for ERGMs

In Bayesian statistics, the uncertainty about the unknown parameters is modeled by assigning a probability distribution to those parameters of interest that are therefore regarded as random variables. During the last decade there has been an increasing interest for modeling networks under the Bayesian framework. Recently, Bayesian approaches for inferring ERG models have been proposed by Koskinen et al. (2010), Caimo and Friel (2013) and Caimo and Mira (2015). The focus of interest in Bayesian inference is the posterior distribution:

$$\pi(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\pi(y)}, \tag{4}$$

where $p(\theta)$ is a prior distribution on θ. Throughout the article we consider a prior distribution that has a simple form, to ensure that it can be simulated using standard techniques and that the posterior always exists. Markov chain Monte Carlo can be used to infer $\pi(\theta \mid y)$.

Consider the use of a Metropolis-Hastings (MH) update (Metropolis et al., 1953; Hastings, 1970). Given a proposal $h(\theta' \mid \theta)$ the algorithm proposes the move from θ to θ' with probability

$$A(\theta, \theta') = \min \left\{ 1, \frac{q(y \mid \theta')}{q(y \mid \theta)} \frac{p(\theta')}{p(\theta)} \frac{h(\theta \mid \theta')}{h(\theta' \mid \theta)} \frac{z(\theta)}{z(\theta')} \right\}$$

and has a stationary distribution of $\pi(\theta \mid y)$. When the likelihood is known in closed form, an exact MCMC targeting π can be implemented. However, when a likelihood with a very large number of observations or a likelihood with unknown normalizing constant is present, as is the case for ERGMs, standard MCMC algorithms cannot be implemented.

Being a function of the parameters, the normalizing term $z(\theta)$ cannot be ignored. The acceptance probability is dependent on a ratio of intractable normalizing constants; therefore, a naive application of the Metropolis-Hastings algorithm is not feasible here. This has generated research to find Markov chains that are exact and can be implemented is these situations, such as the exchange algorithm, presented below.

4.1. The Exchange Algorithm

A popular approach for dealing with intractable likelihoods, such as Markov random fields, is the exchange algorithm. Murray et al. (2006) extended the work of Møller et al. (2006) to allow inference on doubly intractable distributions using the exchange algorithm. The exchange samples from the following augmented distribution

$$\pi(\theta', y', \theta \mid y) \propto p(y \mid \theta) p(\theta) h(\theta' \mid \theta) p(y' \mid \theta'), \tag{5}$$

where $p(y' \mid \theta')$ is the same distribution as the original distribution on which the data y is defined.

This algorithm deals with the intractability of the normalizing constant by introducing an auxiliary variable, hence sampling on an extended state space, that allows the cancellation of all of the normalizing constants in the acceptance probability. This gives rise to a Metropolis–Hastings acceptance probability which is tractable even for target distributions whose normalizing constant depends on the parameter of interest (doubly intractable problems).

The exchange algorithm requires exact simulation of the auxiliary variable y' from the likelihood; perfect sampling from the likelihood is feasible for other MRFs like the Ising and Potts models (Propp and Wilson, 1996; Huber, 2004), but this is not the case for ERG models. The *approximate exchange algorithm* (AEA) of Caimo and Friel (2011) modifies the original exchange algorithm and makes it applicable also in settings where sampling from the auxiliary likelihood (the ERG likelihood in our case) is not feasible. It is possible to carry out inference for graphs of larger size (eg. 1000 nodes), but at the cost of an increased computational time.

The approximate exchange resorts to the tie-no-tie (TNT) sampler (Hunter et al., 2008) to get approximate draws. The TNT sampler simulates a Markov chain whose transition kernel admits $p(y' \mid \theta')$ as limiting distribution. The Markov kernel is iterated a large number of times, M, so that the final point is approximately distributed under $p(y' \mid \theta')$. At this stage it can be anticipated that the number of TNT iterations will be proportional to the number of dyads of the graphs, n^2, for a graph with n nodes. This fact has been theoretically supported by the two following recent

studies:

- Everitt (2012) has proved that when the MCMC kernel for the exact exchange algorithm is uniformly ergodic, the invariant distribution (when it exists) of the corresponding approximate exchange algorithm becomes closer to the "true" target (that of the exact exchange algorithm) as the number of auxiliary iterations, M, increases.

- Bhamidi et al. (2011) have shown that convergence of sampling from an ERG model through Markov chain Monte Carlo (typically the TNT sampler) is likely to be exponentially slow. Therefore, this suggests that one should take a conservative approach and choose a large number of auxiliary iterations for a graph with 100 nodes. The resulting exchange algorithm may be infeasible for large graphs due to the exponentially long mixing time for the auxiliary draw from the likelihood.

4.2. Bayesian Pseudo-posteriors

The aforementioned issues relating to the exchange and approximate exchange algorithms have motivated tractable approximations to the likelihood function which we consider here in the context of Bayesian inference. More specifically, we propose to replace the true likelihood $p(y \mid \theta)$ with a misspecified likelihood model, leading us to focus on the approximated posterior distribution (Pauli et al., 2011), or "pseudo–posterior":

$$\pi_{\mathrm{PL}}(\theta \mid y) \propto p_{\mathrm{PL}}(y \mid \theta) p(\theta). \tag{6}$$

To conduct Bayesian inference using Eq. (6), we adopt the full-update Metropolis-Hastings sampler. The acceptance probability is now tractable but approximated, due to the use of a misspecified model. Algorithm 1 provides the pseudocode for the Metropolis–Hastings algorithm that samples from a pseudo-posterior and forms the basis for our correction method.

Algorithm 1 Pseudolikelihood-based Metropolis–Hastings sampler

1: **Input:** Initial setting: θ;
   A proposal distribution, $h(\cdot \mid \theta)$;
   Number of iterations, T;
   Matrix of change statistics.
2: **Output:** A realization of length T from a Markov chain.
3: **for** $\mathrm{t} = 0, \ldots, \mathrm{T}$ **do**
4: Propose $\theta' \sim h(\cdot \mid \theta)$;
5: $\widetilde{A}(\theta^{(t)}, \theta') = \min\left\{1, \frac{p_{\mathrm{PL}}(y|\theta')}{p_{\mathrm{PL}}(y|\theta^{(t)})} \frac{p(\theta')}{p(\theta^{(t)})} \frac{h(\theta^{(t)}|\theta')}{h(\theta'|\theta^{(t)})}\right\}$;
6: Draw $\mathrm{u} \sim \mathrm{Uniform}[0,1]$;
7: **if** $\mathrm{u} \leq \widetilde{A}(\theta^{(t)}, \theta')$ **then** $\theta^{(t+1)} \leftarrow \theta'$;
8: **end for**
9: **return** $\{\theta_t\}_{\mathrm{t}=1,\ldots,\mathrm{T}}$;

4.3. Adjustment of the pseudo–posterior distribution

Recent work by Stoehr and Friel (2015) focused on calibrating conditional composite likelihoods with overlapping components/ blocks. They observed that a non–calibrated composite

likelihood (the generalization of the pseudolikelihood misspecification) leads to an approximated posterior distribution with substantially lower variability than the true posterior distribution, leading in turn to overprecision about posterior parameters.

Calibration of the unadjusted posterior sample to obtain appropriate inference, as described in their work, is executed with two operations: a "mode adjustment" to ensure that the true and the approximated posterior distributions have the same mode and a "curvature adjustment" that modifies the geometry of the approximated posterior at the mode.

We will use the following notation:

$$\arg\max_{\theta} \pi(\theta \mid y) = \theta^*,\ \mathcal{H} \log \pi(\theta \mid y)|_{\theta^*} = H^*,$$
$$\arg\max_{\theta} \pi_{\mathrm{PL}}(\theta \mid y) = \hat{\theta}_{\mathrm{PL}},\ \mathcal{H} \log \pi_{\mathrm{PL}}(\theta \mid y)|_{\hat{\theta}_{\mathrm{PL}}} = \hat{H}_{\mathrm{PL}},$$

where for any function f taking values in the set of strictly positive real numbers and assuming that $\log f$ is twice differentiable at θ_0, $\mathcal{H} \log f(\theta)|_{\theta_0}$ is the Hessian matrix of $\log f$ at θ_0. In the ERGM context the parameters θ^*, H^*, $\hat{\theta}_{\mathrm{PL}}$ and $\hat{H}_{\mathrm{PL}}$ are either available in closed form or can be estimated using Monte Carlo. In the second half of this section we provide practical guidelines on how to obtain these. We denote by $\widetilde{\pi}(\theta \mid y)$ the fully calibrated target whose mode is located at θ^* and whose second order derivative at the mode is equal to H^*:

$$\arg\max_{\theta} \widetilde{\pi}(\theta \mid y) = \theta^*,\ \mathcal{H} \log \widetilde{\pi}(\theta \mid y)|_{\theta^*} = H^*. \tag{7}$$

In this paper, we choose to adjust the pseudo–posterior through an affine transformation $g(\theta) = W\theta + \lambda$, where $W \in \mathcal{M}(\mathbb{R}^d)$ is an invertible matrix. More precisely, the calibrated pseudo–posterior has the following probability density function:

$$\widetilde{\pi}(\theta \mid y) \propto \pi_{\mathrm{PL}}[g(\theta) \mid y]\,. \tag{8}$$

To satisfy the two equations in (7), we identify W and λ as follows:

Condition 1:

$$\nabla_\theta \log \widetilde{\pi}(\theta \mid y)|_{\theta^*} = 0\,,$$

which, irrespective of the choice of W, gives

$$\begin{aligned} \nabla_\theta \log \pi_{\mathrm{PL}}(\theta \mid y)|_{\hat{\theta}_{\mathrm{PL}}} &= 0 \\ \iff \nabla_\theta \log \widetilde{\pi}(\theta \mid y) &= W^T \nabla_\theta \log \pi_{\mathrm{PL}}(\theta \mid y)\big|_{W\theta+\lambda} \\ \Rightarrow \lambda &= \hat{\theta}_{\mathrm{PL}} - W\theta^*. \end{aligned}$$

Condition 2:

$$\mathcal{H} \log \widetilde{\pi}(\theta \mid y)\big|_{\theta^*} = H^*, \tag{9}$$

where the left–hand side can be written as

$$\mathcal{H} \log \widetilde{\pi}(\theta \mid y)\big|_{\theta^*} = W^T\ \mathcal{H} \log \pi_{\mathrm{PL}}(\theta \mid y)\big|_{W(\theta^*-\theta^*)+\hat{\theta}_{\mathrm{PL}}} W = W^T \hat{H}_{\mathrm{PL}} W.$$

Since H^* and $\hat{H}_{\mathrm{PL}}$ are both Hessians at the mode of their respective distribution, they are negative-

definite matrices and therefore admit a Cholesky decomposition: $-H^* = N^T N$ and $-\hat{H}_{\mathrm{PL}} = M^T M$. Based on this observation, it is straightforward to check that the choice $W = M^{-1}N$ satisfies Eq. (9). $\widetilde{\pi}$ is now a proper density entirely specified by

$$\widetilde{\pi}(\theta \mid y) = |\det(W)|\, \pi_{\mathrm{PL}}[W(\theta - \theta^*) + \hat{\theta}_{\mathrm{PL}} \mid y] \tag{10}$$

and satisfies the two conditions $\arg\max_\theta \widetilde{\pi}(\theta \mid y) = \theta^*$ and $\mathcal{H} \log \widetilde{\pi}(\theta \mid y)|_{\theta^*} = H^*$.

Section 4.2 shows how one can gather a sample $(\theta_1, \ldots, \theta_T)$ from π_{PL}. The correction step then consists of applying the mapping $g^{-1} : \Theta \to \Theta$ so that the corrected sample

$$\zeta_i = g^{-1}(\theta_i) = V\theta_i + (\theta^* - V\hat{\theta}_{\mathrm{PL}}), \text{ where } V = W^{-1},$$

for $i = 1, \ldots, T$, is distributed under $\widetilde{\pi}$.

4.4. Optimization algorithms

It is clear that implementing such a transformation is feasible provided that the quantities V, $\hat{\theta}^*$ and $\hat{\theta}_{\mathrm{PL}}$ are available. Equivalently, once those quantities are estimated from the data and without having performed any MCMC sampling in advance as described in section 4.2, one can opt for sampling directly from the fully calibrated pseudo-posterior distribution, whose analytical form is known (Eq. 10), with the Metropolis–Hastings algorithm. We now provide practical guidelines for estimation of the parameters θ^*, $\hat{\theta}_{\mathrm{PL}}$, H^* and $\hat{H}_{\mathrm{PL}}$.

1. To estimate θ^*_{PL} one can use any gradient-based optimiser. Here we used a BFGS algorithm (Nocedal and Wright, 2006) which was based on available exact gradient evaluations. In practice, the gradient evaluated in the BFGS algorithm is calculated using standard logistic regression theory (Hosmer et al., 2013).
2. The Robbins–Monro stochastic approximation algorithm (Robbins and Monro, 1951) can be used to estimate the maximum a posteriori θ^*, when the gradient of the (log)posterior is not analytically tractable. The Robbins–Monro algorithm takes steps in the direction of the slope of the distribution and follows the stochastic process

$$\theta_{i+1} = \theta_i + \varepsilon_i\, \hat{\nabla}_\theta \log \pi(\theta_i \mid y)\ , \qquad \text{where} \quad \sum_{i=0}^{\mathrm{N}} \varepsilon_i = \infty \text{ and } \sum_{\mathrm{i}=0}^{\mathrm{N}} \varepsilon_i^2 < \infty,$$

   where $\hat{\nabla}_\theta \log \pi(\theta_i \mid y)$ denotes a noisy version of the gradient of the log-posterior at θ_i and $\{\varepsilon_i\}_i$ is a non-increasing sequence of positive numbers. In our analysis we use a sequence of steps which satisfy these conditions, having the form $\varepsilon_i = \alpha/i,\ \alpha > 0$. Once the difference between successive values of this process is less than a specified tolerance level, the algorithm is deemed to have converged to the MAP. The noisy gradient of the log-posterior at θ is derived as follows. First, the logarithm of the posterior can be written as:

$$\log \pi(\theta \mid y) = \theta^{\mathrm{T}} s(y) - \log z(\theta) + \log p(\theta) - \log \pi(y).$$

   The first derivative of the log-posterior with respect to θ yields:

$$\nabla_\theta \log \pi(\theta \mid y) = s(y) - \sum_{y \in \mathcal{Y}} s(y) p(y \mid \theta) + \nabla_\theta \log p(\theta) \tag{11}$$

$$= s(y) - \mathbb{E}_{y|\theta}[s(y)] + \nabla_\theta \log p(\theta).$$

The exact evaluation of $\mathbb{E}_{y|\theta}[s(y)]$ is intractable in practice due to the presence of the normalizing term. However it is possible to estimate $\mathbb{E}_{y|\theta}[s(y)]$ using Monte Carlo sampling. We can simulate $y_\theta = (y'_1, y'_2, \ldots, y'_N) \sim p(\cdot \mid \theta)$ and then estimate $\mathbb{E}_{y|\theta}[s(y)]$ as $\frac{1}{N}\sum_{i=1}^{N} s(y'_i)$. Hence the noisy version of the gradient at θ will be given by:

$$\hat{\nabla}_\theta \log \pi(\theta \mid y) = s(y) - \frac{1}{N}\sum_{i=1}^{N} s(y'_i) + \nabla_\theta \log p(\theta). \tag{12}$$

In all our examples we started the algorithm from the corresponding maximum pseudolikelihood estimate, $\hat{\theta}_{\text{MPLE}}$ and as we explain in Section 5.1 it will be important to monitor that $\hat{\theta}_{\text{MPLE}}$ does not lie in a degenerate region. The Robbins–Monro algorithm is highly sensitive to the starting point in terms of time to convergence; in our context we set the sequence $\{\varepsilon_i\}_i$ small enough by setting $\alpha = 0.001$ and N appropriately large to try to avoid this problem. More precisely, the simulation step (TNT sampler) to draw y' was run for 1,000 iterations followed by an extra 12,000 iterations thinned by a factor of 30, yielding $N = 400$ graphs.

3. The Hessian matrix $\hat{H}_{\text{PL}}$ of the approximated posterior distribution at the mode $\hat{\theta}_{\text{PL}}$ is analytically available.
4. The curvature of the true posterior distribution H^* at the MAP θ^* can be calculated by taking the derivative of Eq. (11):

$$\begin{aligned}
\mathcal{H} \log \pi(\theta \mid y) &= -\frac{z''(\theta)z(\theta) - z'(\theta)z'(\theta)}{z^2(\theta)} + \mathcal{H} \log p(\theta) \\
&= -\left[\frac{z''(\theta)}{z(\theta)} - \left(\frac{z'(\theta)}{z(\theta)}\right)^2\right] + \mathcal{H} \log p(\theta) \\
&= -\left\{\mathbb{E}^2_{y|\theta}[s(y)] - \left[\mathbb{E}_{y|\theta}[s(y)]\right]^2\right\} + \mathcal{H} \log p(\theta) \\
&= -\mathbb{V}_{y|\theta}[s(y)] + \mathcal{H} \log p(\theta).
\end{aligned} \tag{13}$$

The Monte Carlo estimators of the expected value $\mathbb{E}_{y|\hat{\theta}^*}[s(y)]$ and the covariance matrix $\mathbb{V}_{y|\hat{\theta}^*}[s(y)]$ are based on a number of random samples of networks drawn from the specified model. With the aforementioned procedure Monte Carlo error emerges due to the simulation approximation of $\mathbb{E}_{y|\theta}[s(y)]$ and $\mathbb{V}_{y|\theta}[s(y)]$.

Algorithm 2 summarizes all the actions that need to be taken to calibrate the posterior draws from the misspecified target distribution.

5. Applications

This section demonstrates the performance of Bayesian parameter estimation of Exponential random graph models with the pseudolikelihood–based Metropolis–Hastings algorithm pre– and post– calibration of the respective posterior sample and provides comparisons to the approximate exchange algorithm. A toy example and two real networks of increased complexity drawn from

disparate fields will be illustrated. In particular, we considered undirected and unweighted graphs. All computations in this paper were carried out with the statistical environment R (R Development Core Team, 2014) on a laptop computer with an Intel ®Core$^{\text{TM}}$ i7-4500U CPU (1.80GHz) and 16GB RAM.

The MCMC algorithms in this paper sample from the corresponding target distributions with the use of a full–update Metropolis-Hastings sampler. The main MCMC chain consists of 40,000 iterations with a burn–in period of 10,000 iterations. Throughout our analysis we assumed a diffuse Multivariate Normal prior distribution for the model parameters, $\mathcal{MVN}(0_d, 30 \times I_d)$, where 0_d is the null vector and I_d is the identity matrix of size equal to the number of model dimensions d, unless stated otherwise. A conjugate prior is unavailable in the context of an ERG likelihood because of the intractable normalising constant. The **Bergm** package for R (Caimo and Friel, 2014) implements the methodology developed in this paper.

Algorithm 2 Calibration of pseudo–posterior draws

Input: Unadjusted pseudo–posterior draws, $\theta_t,\ \mathrm{t} = 1, \ldots, \mathrm{T}$.
Output: Mode and curvature–adjusted pseudo–posterior sample, $\zeta_t,\ \mathrm{t} = 1, \ldots, \mathrm{T}$.

MAP estimation

Estimate $\hat{\theta}_{\mathrm{PL}}$ (BFGS algorithm) based on exact gradient evaluations (logistic regression theory);
Estimate $\hat{\theta}^*$ (Robbins–Monro algorithm) based on a Monte Carlo estimator of $\nabla_\theta \log \pi(\theta \mid y)$;

Curvature Adjustment

Estimate $\hat{H}_{\mathrm{PL}}$ using logistic regression theory;
Estimate H^* based on a Monte Carlo estimator of $\mathcal{H} \log \pi(\theta \mid y)$;
Perform Cholesky decompositions of H^*, $\hat{H}_{\mathrm{PL}}$: $-H^* = N^T N$, $-\hat{H}_{\mathrm{PL}} = M^T M$;
Calculate $W = M^{-1} N$;
Calculate the transformation matrix V by inverting W, $V = W^{-1}$;
return Adjusted sample $\zeta_t = V(\theta_t - \hat{\theta}_{\mathrm{PL}}) + \theta^*,\ \mathrm{t} = 1, \ldots, \mathrm{T}$.

5.1. Toy example with degeneracy check

Here we illustrate using a simulated toy example that the correction procedure embedded in our algorithm does not inadvertently mask degeneracy in the model specification. To this end, we considered a Markov model with edge and triangle parameters which is known to be degenerate (Snijders et al., 2006).

We simulated an undirected graph with 30 nodes with the edge parameter set to -3.0 and triangle parameter set to 1.2 (Fig. 2). Following the experiments of Snijders et al. (2006), we expect higher density graphs to arise from simulations from the likelihood under this parameter configuration. We first note that the MPLE was initially used as a starting point for the Robbins–Monro stochastic algorithm. However, since $\hat{\theta}_{MPLE} = (-3.08, 0.95)$ lies in the degenerate region (Fig. 3), this causes the algorithm not to converge, as networks simulated from the likelihood around this parameter vector are typically fully connected and this indeed gives a first indication that this model is inappropriate. As an alternative, we used the 0_d vector as a starting point and we ran the

algorithm for 5,000 iterations, where at each iteration 400 graphs were randomly drawn to ensure convergence, see Eq. (12).

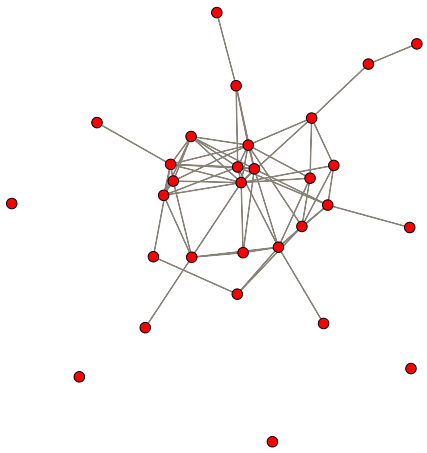

Figure 2: Toy example network with 30 nodes and 65 edges.

To check for issues of degeneracy we used the Bayesian procedure of Caimo and Friel (2014). In particular, we gathered a sample from the corrected pseudo-posterior distribution and simulated networks from the likelihood for each θ parameter in the sample. Note that the mode of the corrected pseudo-posterior distribution was found to be $(\theta_{Edges}, \theta_{Triangle}) = (-1.895, 0.335)$, illustrating that the prior distribution, centered at 0_d, has the effect of shifting the high posterior density region away from the degenerate region. This mode is consistent with the estimated mode of the true posterior based on a long approximate exchange MCMC run. It is therefore important to check if the posterior distribution still supports parameters in the degenerate region.

To examine this, we generated a sample of size 600 from $\widetilde{\pi}(\theta \mid y)$, yielding 600 networks simulated from each parameter in the sample. Fig. 3 illustrates the distribution of the number of edges based on those simulated networks. While a reasonable proportion of the simulated networks has a density close to the observed network density, there is a large number of networks which are highly connected. This illustrates that the posterior distribution supports some parameter values which lie in the degenerate region, illustrating the unsuitability of this model and reinforcing the need for a model degeneracy check.

5.2. International E–road Network

The International E–road Network constructed by Šubelj and Bajec (2011) and displayed in Fig. 4 represents all roads included in the International E–road Network. Nodes correspond to European cities and edges represent direct (class A, B) road connections among them. The network has 1177 nodes and 1417 edges. To demonstrate our methodology we fit a 2–dimensional model with edge and 2-star terms, the posterior distribution of which is:

$$\pi(\theta \mid y) \propto \frac{1}{z(\theta)} \exp\left[\theta_1 \sum_{i<j} y_{ij} + \theta_2 \sum_{i<j<k} y_{ik} y_{jk}\right] p(\theta), \tag{14}$$

where $p(\theta)$ is a $\mathcal{MVN}(0_d, 30 \times I_d)$ distribution. In further simulation experiments not shown here, changing the prior distribution to $\mathcal{MVN}(0_d, 300 \times I_d)$ did not lead to different results. We

note that the purpose of this example is mainly to illustrate the gains that can result from using our methodology. Such Markov models do not provide a good specification for many real human social networks, where degeneracy and phase transition issues are likely to occur (see Handcock (2002) and Park and Newman (2004) for detailed studies).

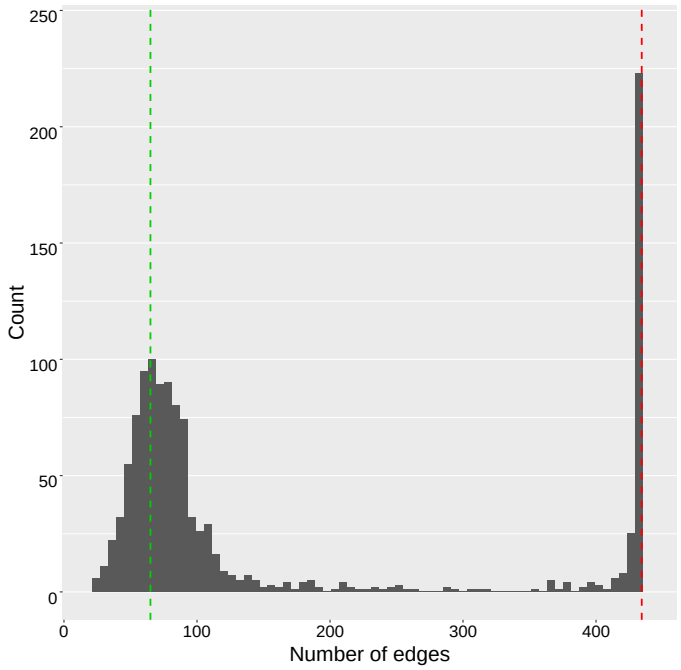

Figure 3: Toy example network - Distribution of the number of edges based on 600 simulated graphs from the corrected pseudo-posterior distribution $\widetilde{\pi}(\theta \mid y)$. The green dashed vertical line corresponds to the observed number of edges, while the red dashed vertical line corresponds to the average number of edges based on 20 simulated graphs under the parameter configuration $(\theta_{Edges}, \theta_{Triangle}) = (-3.0, 1.2)$.

5.2.1. Sampling from the Approximate Exchange

The *approximate exchange algorithm* (AEA) was first used to generate draws from the target posterior. The results from this served as a ground truth against which the posterior estimates of θ using the various approximated likelihood estimators were compared. The **Bergm** package (Caimo and Friel, 2014) for R allows to carry out inference with the approximate exchange algorithm described above.

Bergm uses a multivariate Gaussian proposal distribution. To account for possible correlations between the model parameters, all samplers in this paper assumed the same proposal distribution with a proposal variance-covariance in the form $\Sigma = \mathrm{T}(\mathrm{B}_0 + \mathrm{C}^{-1})^{-1}\mathrm{T}$. T denotes the diagonal positive definite matrix formed from a Metropolis tuning parameter, chosen by the user to reach a reasonable mixing rate, B_0 is the prior precision, and C is the large sample variance–covariance matrix of the MPLEs. The precision matrix C^{-1} is the same as the negative Hessian, $-H(\hat{\theta})$.

The approximate exchange algorithm requires a number of iterations, M, to simulate approximately from the likelihood. A "conservative" approach would be to choose a large number of auxiliary iterations, eg. 500,000, to ensure that the invariant distribution of the approximate exchange algorithm will be considered as the "true" target. Table 1 suggests that 10,000 is a practical number of auxiliary iterations for this two-dimensional model.

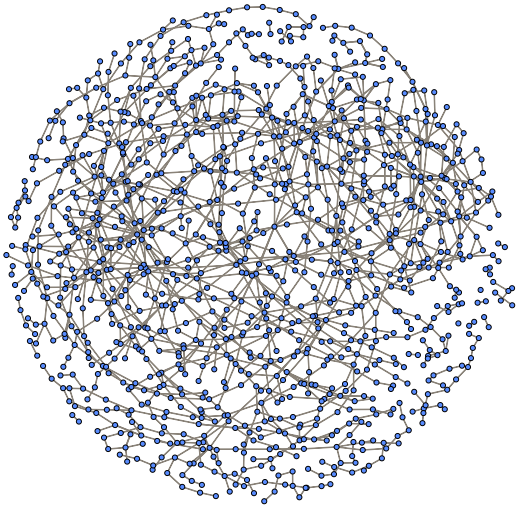

Figure 4: International E–road network graph.

The total variation distance metric (Dudley, 1968) between a given Markov chain and the conservative AEA sample provides a numerical tool to assess the improvement at each step of the calibration. It gives a measure of how well the posterior densities match the "ground truth" posterior density. Recall that the TV–distance of two densities $f(\theta)$ and $g(\theta)$ is given by

$$TV(f,g) = \frac{1}{2}\int |f(\theta) - g(\theta)|\,\mathrm{d}\theta.$$

It is equal to 1 when f and g have disjoint supports and it vanishes when the functions are identical. For the two–dimensional example considered here, the total variation distance was approximated by splitting the lattice into bins with a pre–defined window size. Within each grid the absolute difference of the frequencies was calculated, such that TV takes values in [0,1].

Table 1: International E–road network: average total variation distance on 20 simulations for Model (14), using the approximate exchange algorithm under an increasing number of auxiliary iterations. The total variation distance was calculated between a given Markov chain and the conservative AEA sample using 500,000 auxiliary iterations.

Auxiliary Iterations	50	100	300	500	10^3	3×10^3
Average TV	0.744	0.665	0.603	0.598	0.545	0.234
SE of mean ($\times10^{-4}$)	25	20	28	30	27	21
Auxiliary Iterations	5×10^3	10^4	2×10^4	4×10^4	10^5	
Average TV	0.075	0.029	0.029	0.030	0.028	
SE of mean ($\times10^{-4}$)	19	16	11	16	15	

At this point it is worth stressing the fact that when a practitioner opts for the approximate exchange, they will not have a–priori knowledge of the number of auxiliary iterations needed to sample from the true target distribution. This means that they have two options: the first is to

perform an experiment similar to the one for this example, where the total variation distance is calculated with respect to a conservative MCMC run. The second option would be to proceed by choosing a large number of auxiliary iterations, eg. 500,000, and run the approximate exchange algorithm. However it is demonstrated that the latter can take up to 1.75 hours to run (Table 3), which turns out to be considerably more expensive from a computational point of view than the algorithm we propose.

5.2.2. Pseudolikelihood-based Posterior Inference

Although fast (Table 3), sampling from the non–calibrated pseudo-posterior via the Metropolis-Hastings leads to considerably underestimated posterior standard deviations (Table 2).

The next step is to consider correcting the sample coming from the pseudo–posterior distribution. Monte Carlo estimation of the true posterior maximum can be performed in a quick manner. The Robbins–Monro stochastic algorithm converged after 50 iterations, where at each iteration 400 graphs were randomly drawn. Once the mode adjustment and the curvature adjustment were performed, a very good approximation of the true posterior with efficient correction of the posterior variance was obtained (Fig. 5), while achieving a five–fold speedup relative to the approximate exchange with 10,000 auxiliary iterations.

Table 2: International E–road network: Pseudo–posterior parameter estimates (mean and standard deviation) at each phase of the calibration procedure. The total variation distance was calculated between each joint posterior distribution and the joint posterior that was generated by the approximate exchange with 500,000 auxiliary iterations.

	θ_1 (Edges)	θ_2 (2-stars)	TV
Pseudo–posterior	-4.496 (0.089)	-0.388 (0.021)	0.753
Mode + Curvature calibrated	-4.840 (0.127)	-0.311 (0.029)	0.028
AEA (10^4 aux. iters)	-4.846 (0.133)	-0.305 (0.030)	0.029

The results of Table 1 indicate that in the best case scenario, the average total variation distance will be 0.03 with respect to the long MCMC run. If one decides to use less iterations (eg. 5×10^3), this will result in a worse approximation with a higher total variation distance. On the other hand, after calibrating the pseudo–posterior sample, the resulting total variation distance equals 0.028, showing a very good approximation to the true posterior. Any (slight) differences between the two distributions arise because of the Monte Carlo error from the Robbins–Monro approximation. Hence, the overall calibration approach of the pseudo–posterior distribution provides a good trade–off between computational time and efficiency.

For an MCMC run of length T with lag k autocorrelation ρ_k we measure the efficiency of the sampler by using the Effective Sample Size, $\text{ESS} = \text{T}/(1 + 2\sum_{k=1}^{\infty} \rho_k)$ (Liu, 2001). This Effective Sample Size (the larger the better) gives an estimate of the equivalent number of independent iterations that the chain represents. The algorithms considered in this paper have different running times; a fair comparison can be made by standardizing the ESS by the CPU run time, defining the efficiency ratio

$$\text{ER(Sampler)} = \frac{\text{Min ESS(Sampler)}}{\text{CPU(Sampler)}}$$

as a performance metric. To examine the performance of each algorithm relative to the ground truth approximate exchange sampler with 500,000 auxiliary iterations, we define the relative effi-

ciency by

$$\text{RE} = \frac{\text{ER(Sampler)}}{\text{ER(AEA}_{500\text{K}})}.$$

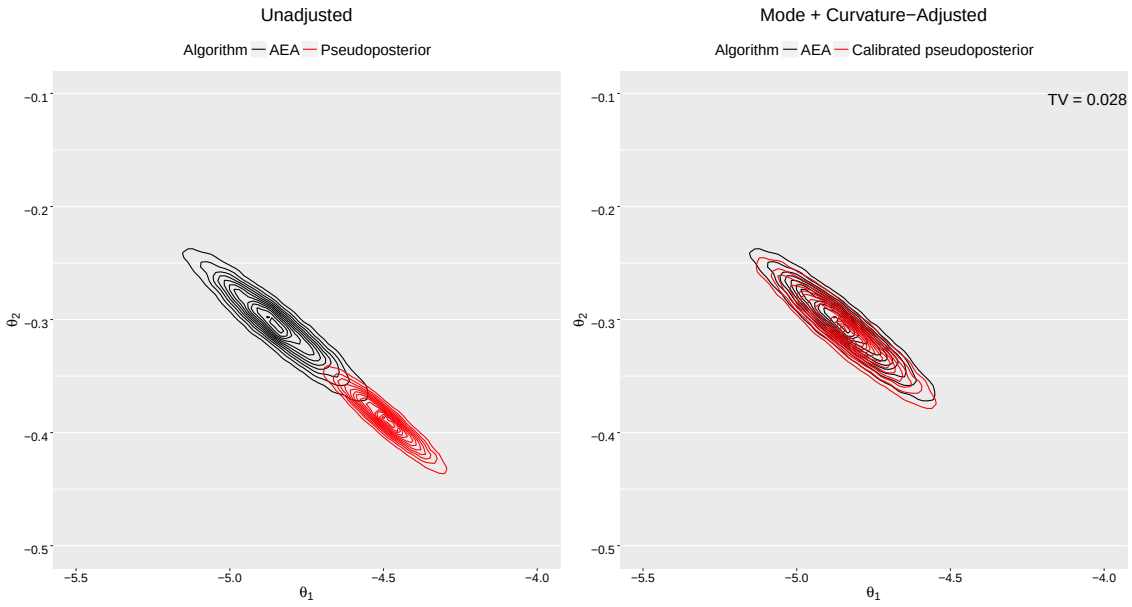

Figure 5: International E–road network: Phases of calibration of the misspecified posterior distribution using a pseudolikelihood approximation.

The pseudolikelihood–based MH sampler can achieve a higher effective sample size compared to the approximate exchange (Table 3). Combined with the small overall computational cost of the calibration procedure, there is a huge gain in efficiency with respect to the ground truth approximate exchange algorithm, as well as the sampler with the 10,000 auxiliary iterations which were found to be practical for this example.

Table 3: International E–road network: CPU time (in seconds) of the different steps of the inference algorithms, minimum ESS, Efficiency Ratio and relative efficiency ratio. The total CPU time for the calibration procedure represents the sum of the CPU times of each of the individual stages.

Calibration phase	CPU (s)	ESS	ER	Relative ER
Pseudo–posterior	14.36	3638	253.34	816.23
Robbins–Monro (50 iters)	11.86	–	–	–
Mode + Curvature calibration	8.87	–	–	–
Total	35.09	3638	103.67	333.41
AEA (10^4 aux. iters)	174.63	1826	10.45	32.71
AEA (5×10^5 aux. iters)	6297.96	1927	0.31	1

5.3. Faux Mesa High School Network

The undirected graph displayed in Fig. 6 represents friendship relations in a school community of 205 students (Handcock et al., 2007). The Faux Mesa High School network is a well known network in social science and consists of 203 undirected edges (mutual friendships). Similarly

to Caimo and Mira (2015), we are interested in the vertex attributes x for the "grade" (values 7 through 12 indicating each student's grade in school) of each student.

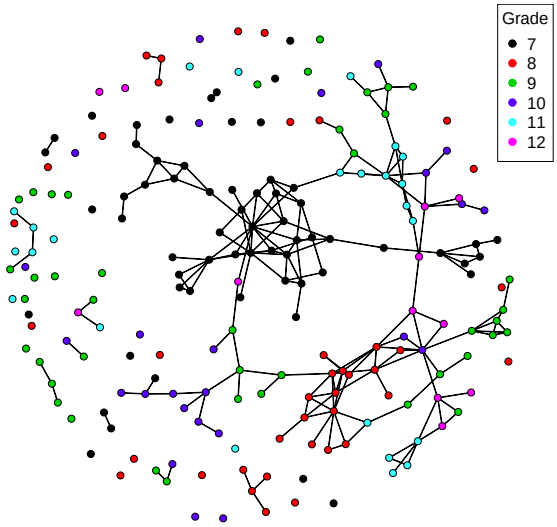

Figure 6: Faux Mesa High School friendship network graph.

The main focus lies in the factor attribute effects (which give information about the tendency of a node with a specific attribute to form an edge in the network) and on the transitivity effect expressed by the GWESP statistics. We consider a model defined by the following 8 network statistics:

$$
\begin{aligned}
s_1(y) &= \textstyle\sum_{i<j} y_{ij} & s_5(y,x) &= \textstyle\sum_{i<j} y_{ij}\{1_{(grade_i=10)} + 1_{(grade_j=10)}\} \\
s_2(y,x) &= \textstyle\sum_{i<j} y_{ij}\{1_{(grade_i=7)} + 1_{(grade_j=7)}\} & s_6(y,x) &= \textstyle\sum_{i<j} y_{ij}\{1_{(grade_i=11)} + 1_{(grade_j=11)}\} \\
s_3(y,x) &= \textstyle\sum_{i<j} y_{ij}\{1_{(grade_i=8)} + 1_{(grade_j=8)}\} & s_7(y,x) &= \textstyle\sum_{i<j} y_{ij}\{1_{(grade_i=12)} + 1_{(grade_j=12)}\} \\
s_4(y,x) &= \textstyle\sum_{i<j} y_{ij}\{1_{(grade_i=9)} + 1_{(grade_j=9)}\} & s_8(y) &= v(y,\phi_v),
\end{aligned}
$$

where $1_{(\cdot)}$ is the indicator function. GWESP, the geometrically weighted edgewise shared partners, are given by the formula:

$$
v(y,\phi_v) = e^{\phi_v} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\phi_v}\right)^i \right\} EP_i(y),
$$

where $EP_i(y)$ are the edgewise shared partners. To get a model which is a non–curved ERG model, we set $\phi_v = 1$. The pseudolikelihood–based Metropolis–Hastings sampler was tuned in the same way as described in Section 5.2 to obtain an overall acceptance rate of 24.3%. Accordingly, each of the approximate exchange samplers for the model of this example had an overall acceptance rate 23-30%.

The experiment for this example were nodal attributes are accounted for consists of running the approximate exchange algorithm long enough (500,000 auxiliary iterations) to draw from the target distribution and then compare the posterior summary statistics with the corresponding quantities from the pseudo–posterior distribution (pre– and post–calibration). We additionally

compare the samplers in terms of the effective sample size and their respective efficiencies, as described below.

Table 4: Faux Mesa High School network: Posterior parameter estimates (mean and standard deviation) obtained by the approximate exchange sampler under an increasing number of auxiliary iterations.

Auxiliary Iterations	θ_1	θ_2	θ_3	θ_4
50	-6.641 (0.757)	2.613 (1.477)	2.122 (2.411)	2.495 (2.260)
100	-6.561 (0.500)	2.365 (0.863)	2.076 (1.424)	2.430 (1.347)
500	-6.505 (0.252)	2.251 (0.401)	2.026 (0.563)	2.308 (0.527)
1×10^3	-6.471 (0.204)	2.156 (0.298)	1.971 (0.447)	2.204 (0.400)
5×10^3	-6.267 (0.166)	1.875 (0.223)	1.902 (0.288)	1.944 (0.298)
1×10^4	-6.209 (0.162)	1.855 (0.203)	1.951 (0.247)	1.921 (0.267)
2×10^4	-6.192 (0.166)	1.897 (0.212)	2.057 (0.239)	2.000 (0.243)
4×10^4	-6.166 (0.171)	1.926 (0.212)	2.117 (0.233)	2.045 (0.243)
1×10^5	-6.149 (0.196)	2.012 (0.221)	2.195 (0.239)	2.058 (0.277)
5×10^5	-6.103 (0.177)	2.052 (0.202)	2.225 (0.221)	2.051 (0.259)
Auxiliary Iterations	θ_5	θ_6	θ_7	θ_8
50	3.886 (3.479)	2.827 (3.526)	3.497 (4.960)	2.265 (0.989)
100	3.731 (2.115)	2.713 (2.446)	4.793 (3.884)	1.572 (0.539)
500	3.015 (0.805)	2.414 (0.897)	4.258 (1.551)	1.403 (0.214)
1×10^3	2.913 (0.601)	2.406 (0.637)	4.082 (1.133)	1.367 (0.158)
5×10^3	2.544 (0.432)	2.271 (0.409)	3.625 (0.625)	1.221 (0.106)
1×10^4	2.371 (0.397)	2.303 (0.337)	3.479 (0.524)	1.152 (0.092)
2×10^4	2.285 (0.379)	2.413 (0.284)	3.295 (0.485)	1.074 (0.078)
4×10^4	2.218 (0.388)	2.446 (0.281)	3.119 (0.421)	1.024 (0.073)
1×10^5	2.210 (0.385)	2.500 (0.274)	2.979 (0.406)	0.951 (0.065)
5×10^5	2.213 (0.353)	2.506 (0.251)	2.839 (0.373)	0.885 (0.059)

The results of Table 5 suggest that the network is very sparse (negative parameter θ_1). Homophily is expressed by the positive posterior parameter means for the main effect of student grade and the positive GWESP parameter (θ_8) expresses the transitivity effect. There is a clear tendency for the posterior standard deviations from the approximate exchange algorithm to decrease with an increasing number of auxiliary iterations (Table 4). Overall, the posterior summary statistic values appear to stabilize when a few hundreds thousand auxiliary iterations are used.

The calibration procedure corrects the posterior standard deviations to a great extent (Table 5) and the respective posterior means are close to those obtained by the approximate exchange algorithm. A high–dimensional model such as the one examined in this example leads to increased computational time for the approximate exchange algorithm, if thousands of iterations are to be used in the auxiliary chain.

Table 5: Faux Mesa High School network: Posterior parameter estimates (mean and standard deviation).

	θ_1	θ_2	θ_3	θ_4
Pseudo–posterior	-6.250 (0.163)	1.805 (0.223)	1.821 (0.281)	2.090 (0.290)
Mode + Curvature calibrated	-6.104 (0.150)	2.051 (0.189)	2.238 (0.219)	2.061 (0.244)
AEA (5×10^5 aux. iters)	-6.103 (0.177)	2.052 (0.202)	2.225 (0.221)	2.051 (0.259)
	θ_5	θ_6	θ_7	θ_8
Pseudo–posterior	2.353 (0.395)	2.487 (0.331)	2.827 (0.539)	1.136 (0.053)
Mode + Curvature calibrated	2.208 (0.356)	2.501 (0.218)	2.859 (0.510)	0.889 (0.082)
AEA (5×10^5 aux. iters)	2.213 (0.353)	2.506 (0.251)	2.839 (0.373)	0.885 (0.059)

While we cannot provide a measure like the total variation distance to calculate the similarity between the joint posterior distributions under examination, we can compare the efficiency of the samplers. Table 6 illustrates that the analyst can benefit from a huge gain in terms of "information size" if the pseudolikelihood–based MH sampler is used. This, in combination with the short runtime of the calibration procedure, leads to increased efficiency relative to the long–run approximate exchange algorithm.

Table 6: Faux Mesa High School network: CPU time (in seconds) of the calibration procedure of the uncalibrated pseudo–posterior sample, minimum ESS and efficiency ratio. The total CPU time for the calibration procedure represents the sum of the CPU times of each of the individual stages.

Calibration stage	CPU (s)	ESS	ER	Relative ER
Pseudo–posterior	11.71	1554	132.78	44,259
Robbins–Monro (50 iters)	12.73	–	–	–
Mode + Curvature calibration	4.42	–	–	–
Total	28.86	1554	53.85	17,949
AEA (1×10^5 aux. iters)	4,607.82	130	0.028	8.33
AEA (5×10^5 aux. iters)	40,212.43	124	0.003	1

6. Discussion

In this paper we explored Bayesian inference of ERG models with tractable approximations to the true likelihood and we applied our methodology in real networks of increased complexity. The computational tractability of the pseudolikelihood function, which is algebraically identical to the likelihood for a logistic regression, has made it an attractive alternative to the full likelihood function. Bayesian logistic regression based on the pseudo-posterior distribution that results from replacing the true likelihood function with the pseudo-likelihood function can be carried relatively quickly. However the drawback of using such an approach is that it ignores strong dependencies in the data, since it involves the pseudolikelihood function and therefore can result in biased estimation. Our main contribution has been to demonstrate how to successfully correct a sample from the pseudo-posterior distribution so that it is approximately distributed from the target posterior distribution.

Our results showed that calibrating the pseudo-posterior distribution is a viable approach that outperforms the approximate exchange algorithm in terms of computational time and scales well

to realistic-size problems, e.g networks with around $1,000$ nodes. In comparison with the approximate exchange algorithm which is routinely used for the Bayesian analysis of ERGMs and is feasible for networks up to 1,000 nodes, our algorithm gave a dramatic reduction in CPU time and we anticipate it to scale well for even larger networks. With a two-step post-processing task we were able to perform a fast and efficient correction using the posterior mean and covariance. The immediate advantage of the proposed framework is evident when analyzing even larger networks and/ or more complex ERG models are fitted, where algorithms like the approximate exchange algorithm might struggle due to the size of the graph or the higher–dimensionality of the parameter vectors. In such cases where a comparison with the approximate exchange would be infeasible, we recommend applying our calibration algorithm.

Additional experiments were conducted using conditional composite likelihoods, which are higher-order generalizations of the pseudolikelihood function. In particular, we considered conditional composite likelihood components, or "blocks", consisting of three dyads each, conditioned on the remainder of the network. However the number of possible such blocks grows very quickly with n, and so one immediate difficulty was the issue of how to choose a subset of these in order to achieve a reasonable computational cost. Overall the results were not so promising. While the pseudolikelihood approach is a fast and viable method that uses all possible dyads, using the composite likelihood did not lead to a competitive sampler. Due to the complex nature of the random graph, difficulties arise in the systematic partition of observations into blocks and the selection of the subsets. Additionally, the overall computational cost can be prohibitive as composite likelihoods do not enjoy the fast point-wise estimation of pseudolikelihoods allowed by the change statistics reparameterization.

We hope that the arguments and findings presented here will serve as a set of guidelines for practitioners to draw meaningful inferences from relational data through the Bayesian paradigm, thus broadening the scope of network modeling beyond its current limits.

Acknowledgments

The authors are grateful for the constructive feedback they received from the Editors and Reviewers at Social Networks. The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289. Nial Friel's research was also supported by a Science Foundation Ireland grant: 12/IP/1424.

References

Besag, J. (1975). Statistical analysis of non-lattice data. *Statistician*, 24:179–195.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64:616–618.

Bhamidi, S., Bresler, G., and Sly, A. (2011). Mixing time of exponential random graphs. *Ann. Appl. Prob.*, 21(6):2146–2170.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Netw.*, 33:41–55.

Caimo, A. and Friel, N. (2013). Bayesian model selection for exponential random graph models. *Social Netw.*, 35(1):11–24.

Caimo, A. and Friel, N. (2014). Bergm: Bayesian Exponential Random Graphs in R. *J. Stat. Softw.*, 61(2):1–25.

Caimo, A. and Mira, A. (2015). Efficient computational strategies for doubly intractable problems with applications to bayesian social networks. *Statist. Comput.*, 25(1):113–125.

Corander, J., Dahmström, K., and Dahmström, P. (2002). Maximum likelihood estimation for exponential random graph models. In *Contributions to Social Network Analysis, Information Theory and Other Topics in Statistics: A Festschrift in Honour of Ove Frank*, pages 1–17, Univ. Stockholm. (J. Hagberg, ed.).

Cranmer, S. J. and Desmarais, B. A. (2011). Inferential network analysis with exponential random graph models. *Polit. Anal.*, 19(1):66–86.

Dudley, R. M. (1968). Distances of probability measures and random variables. *Ann. Math. Stat.*, 39(5):1563–1572.

Everitt, R. G. (2012). Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks. *J. Comput. Graph Stat.*, 24(4):940–960.

Friel, N., Pettitt, N., Reeves, R., and Wit, E. (2009). Bayesian inference in hidden markov random fields for binary data defined on large lattices. *J. Comput. Graph Stat.*, 18(2):243–261.

Handcock, M. (2002). Statistical models for social networks: Degeneracy and inference. In Breiger, R., Carley, K., and Pattison, P., editors, *Dynamic social network modeling and analysis*, pages 229–240. National Academies Press, Washington DC.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. (2007). statnet: Software tools for the statistical modeling of network data. *J. Comput. Graph Stat.*, 24(1):1–11.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hosmer, D., Lemeshow, S., and Sturdivant, R. (2013). *Applied Logistic Regression (Wiley Series in Probability and Statistics)*. Wiley, 3rd edition.

Huber, M. (2004). Perfect sampling using bounding chains. *Ann. Appl. Prob.*, 14(2):734–753.

Hunter, D. and Handcock, M. (2006). Inference in Curved Exponential Family Models for Networks. *J. Computnl Graph Statist.*, 15(3):565–583.

Hunter, D., Handcock, M., Butts, C., Goodreau, S., Morris, and Martina (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *J. Comput. Graph Stat.*, 24(3):1–29.

Koskinen, J., Robins, G., and Pattison, P. (2010). Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. *Stat. Methodol.*, 7(3):366–384.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.*, 80:221–239.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092.

Møller, J., Pettit, A., Reeves, R., and Bertheksen, K. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93:451–458.

Murray, I., Ghahramani, Z., and MacKay, C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer, New York, 2nd edition.

van Duijn, M., Gile, K., and Handcock, M. S. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Netw.*, 31(1):52–62.

Park, J. and Newman, M. (2004). Solution of the two-star model of a network. *Phys. Rev. E*, 70:66–146.

Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Stat. Sin.*, 21(1):149–164.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Struct. Alg.*, 9(1-2):223–252.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Stat.*, 23(3):400–407.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007a). An introduction to exponential random graph (p*) models for social networks. *Social Netw.*, 29(2):173–191.

Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007b). Recent developments in exponential random graph (p*) models for social networks. *Social Netw.*, 29:192–215.

Snijders, T., Pattison, P., Robins, G., and Handcock, M. (2006). New specifications for exponential random graph models. *Sociol. Methodol.*, 36(1):99–153.

Stoehr, J. and Friel, N. (2015). Calibration of conditional composite likelihood for bayesian inference on gibbs random fields. In *AISTATS, Journal of Machine Learning Research: W & CP*, volume 38, pages 921–929.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *J. Am. Stat. Assoc.*, 85:204–212.

Šubelj, L. and Bajec, M. (2011). Robust network community detection using balanced propagation. *Eur. Phys. J. B*, 81(3):353–362.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Stat. Sin.*, 21(1):5–42.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p*. *Psycometrica*, 61:401–425.