CHAPTER 1

BAYESIAN ANALYSIS OF DYNAMIC NETWORK REGRESSION WITH JOINT EDGE/VERTEX DYNAMICS

ZACK W. ALMQUIST[1] AND CARTER T. BUTTS[2]

[1]University of Minnesota, USA.
[2]University of California, Irvine, USA.

1.1 Introduction

Change in network structure and composition has been a topic of extensive theoretical and methodological interest over the last two decades; however, the effects of endogenous group change on interaction dynamics within the context of social networks is a surprisingly understudied area. Network dynamics may be viewed as a process of change in the edge structure of a network, in the vertex set on which edges are defined, or in both simultaneously. Recently, Almquist and Butts (2014) introduced a simple family of models for network panel data with vertex dynamics—referred to here as *dynamic network logistic regression* (DNR)—expanding on a subfamily of temporal exponential-family random graph models (TERGM) (see Robins and Pattison, 2001; Hanneke

et al., 2010). Here, we further elaborate this existing approach by exploring Bayesian methods for parameter estimation and model assessment. We propose and implement techniques for Bayesian inference via both maximum *a posteriori* probability (MAP) and Markov chain Monte Carlo (MCMC) under several different priors, with an emphasis on minimally informative priors that can be employed in a range of empirical settings. These different approaches are compared in terms of model fit and predictive model assessment using several reference data sets.

This chapter is laid out as follows: (1) We introduce the standard (exponential family) framework for modeling static social network data, including both MLE and Bayesian estimation methodology; (2) we introduce network panel data models, discussing both MLE and Bayesian estimation procedures; (3) we introduce a subfamily of the more general panel data models (dynamic network logistic regression)—which allows for vertex dynamics—and expand standard MLE procedures to include Bayesian estimation; (4) through simulation and empirical examples we explore the effect of different prior specifications on both parameter estimation/hypothesis tests and predictive adequacy; (5) finally, we conclude with a summary and discussion of our findings.

1.2   Statistical Models for Social Network Data

The literature on statistical models for network analysis has grown substantially over the last two decades (for a brief review see Butts, 2008b). Further, the literature on dynamic networks has expanded extensively in this last decade – a good overview can be found in Almquist and Butts (2014). In this chapter we use a combination of commonly used statistical and graph theoretic notation. First, we briefly introduce necessary notation and literature for the current state of the art in network panel data models, then we review these panel data models in their general form, including their Bayesian representation. Last, we discuss a specific model family (DNR) which reduces to an easily employed regression-like structure, and formalize it to the Bayesian context.

1.2.1   Network Data and Nomenclature

For purposes of this chapter, we will focus on networks (social or otherwise) that can be represented in terms of dichotomous (i.e., unvalued) ties among pairs of discrete entities. [For more general discussion of network representation, see, e.g., Wasserman and Faust (1994); Butts (2009).] We represent the set of potentially interacting entities via a *vertex set* (V), with the set of interacting pairs (or ordered pairs, for directed relationships) represented by an *edge set* (E). In combination, these two sets are referred to as a *graph*, $G = (V, E)$. (Here, we will use the term "graph" generically to refer to either directed or undirected structures, except as indicated otherwise.) Networks

may be static, e.g., representing relationships at a single time point or aggregated over a period of time, or dynamic, e.g., representing relationships appearing and disappearing in continuous time or relationship status at particular discrete intervals.

For many purposes, it is useful to represent a graph in terms of its *adjacency matrix*: for a graph G of order $N = |V|$, the adjacency matrix $Y \in \{0, 1\}^{N \times N}$ is a matrix of indicator variables such that $Y_{ij} = 1$ iff the ith vertex of G is adjacent (i.e., sends a tie to) the jth vertex of G. Following convention in the social network (but not graph theoretic) literature, we will refer to N as the *size* of G.

The above extends naturally to the case of dynamic networks in discrete time. Let us consider the time series $\ldots, G_{t-1}, G_t, G_{t+1}, \ldots$, where $G_t = (V_t, E_t)$ represents the state of a system of interest at time t. This corresponds in turn to the adjacency matrix series $\ldots, Y_{..t-1}, Y_{..t}, Y_{..t+1}, \ldots$, with $N_t = |V_t|$ being the size of the network at time t and $Y_{..t} \in \{0, 1\}^{N_t \times N_t}$ such that $Y_{ijt} = 1$ iff the ith vertex of G_t is adjacent to the jth vertex of G_t at time t. As this notation implies, the vertex set of an evolving network is not necessarily fixed; we shall be particularly interested here in the case in which V_t is drawn from some larger risk set, such that vertices may enter and leave the network over time.

1.2.2 Exponential Family Random Graph Models

When modeling social or other networks, it is often helpful to represent their distributions via random graphs in discrete exponential family form. Graph distributions expressed in this way are called *exponential family random graph models* or ERGMs. Holland and Leinhardt (1981) are generally credited with the first explicit use of statistical exponential families to represent random graph models for social networks, with important extensions by Frank and Strauss (1986) and subsequent elaboration by Wasserman and Pattison (1996), Pattison and Wasserman (1999), Pattison and Robins (2002), Snijders et al. (2006), Butts (2007), and others. The power of this framework lies in the extensive body of inferential, computational, and stochastic process theory [borrowed from the general theory of discrete exponential families, see, e.g., Barndorff-Nielsen (1978); Brown (1986)] that can be brought to bear on models specified in its terms.

We begin with the "static" case in which we have a single random graph, G, with support $\mathcal{G}$. It is convenient to model G via its adjacency matrix Y, with $\mathcal{Y}$ representing the associated support (i.e., the set of adjacency matrices corresponding to all elements in $\mathcal{G}$). In ERGM form, we express the pmf of Y as follows:

$$\Pr(Y = y \mid S, \theta, X) = \frac{\exp\left(\theta^T S(y, X)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\theta^T S(y', X)\right)} \mathbb{I}_{\mathcal{Y}}(y), \qquad (1.1)$$

where $S : \mathcal{Y}, \mathcal{X} \to \mathbb{R}^s$ is a vector of sufficient statistics, $\theta \in \mathbb{R}^s$ is a vector of natural parameters, $X \in \mathcal{X}$ is a collection of covariates, and $\mathbb{I}_\mathcal{Y}$ is the indicator function (i.e., 1 if its argument is in the support of $\mathcal{Y}$, 0 otherwise).[1] If $|\mathcal{G}|$ is finite, then the pmf for any G can obviously be written with finite-dimensional S, θ (e.g., by letting S be a vector of indicator variables for elements of $\mathcal{Y}$); this is not necessarily true in the more general case, although a representation with S, θ of countable dimension still exists. In practice, it is generally assumed that S is of low dimension, or that at least that the vector of natural parameters can be mapped to a low-dimensional vector of "curved" parameters [see, e.g., Hunter and Handcock (2006)].

While the extreme generality of this framework has made it attractive, model selection and parameter estimation are often difficult due to the normalizing factor $(\kappa(\theta, S, X) = \sum_{y' \in \mathcal{Y}} \exp(\theta^T S(y', X)))$ in the denominator of equation (1.1). This normalizing factor is analytically intractable and difficult to compute, except in special cases such as the Bernoulli and dyad-multinomial random graph families (Holland and Leinhardt, 1981); the first applications of this family (stemming from Holland and Leinhardt's seminal 1981 paper) focused on these special cases. Later, Frank and Strauss (1986) introduced a more general estimation procedure based on cumulant methods, but this proved too unstable for practical use. This, in turn, led to an emphasis on approximate inference using maximum pseudo-likelihood (MPLE) estimation (Besag, 1974), as popularized in this application by Strauss and Ikeda (1990) and later Wasserman and Pattison (1996). Although MPLE coincides with maximum likelihood estimation (MLE) in the limiting case of edgewise independence, the former was found to be a poor approximation to the MLE in many practical settings, thus leading to a consensus against its general use [see, e.g., Besag (2001) and van Duijn et al. (2009)]. The late 1990s saw the development of effective Markov chain Monte Carlo strategies for simulating draws from ERG models (Anderson et al., 1999; Snijders, 2002) which led to the current focus on MLE methods based either on first order method of moments (which coincides with MLE for regular ERGMs) or on importance sampling (Geyer and Thompson, 1992).[2]

Theoretical developments in the ERGM literature have arguably lagged inferential and computational advances, although this has become an increasingly active area of research. A major concern of the theoretical literature on ERGMs is the problem of *degeneracy*, defined differently by different authors but generally involving an inappropriately large concentration of probability mass on a small set of (generally unrealistic) structures. This issue was

---

[1]$\mathbb{I}_\mathcal{Y}$ can also be interpreted here as the counting measure on $\mathcal{Y}$. Although this is adequate for all cases considered here, see Krivitsky (2012) for the importance of reference measures in valued ERGM settings.

[2]Algorithms for parameter estimation and model selection using these approaches are implemented in a number of software packages (see, e.g., Snijders et al., 2007; Handcock et al., 2003; Wang et al., 2009), and empirical applications are increasingly common (e.g., Goodreau et al., 2009; Robins and Pattison, 2001, etc.).

recognized as early as Strauss (1986), who showed asymptotic concentration of probability mass on graphs of high density for models based on triangle statistics. [This motivated the use of *local* triangulation by Strauss and Ikeda (1990), a recommendation that went unheeded in later work.]   More general treatments of the degeneracy problem can be found in Handcock (2003), Schweinberger (2011), and Chatterjee and Diaconis (2011). Butts (2011) introduced analytical methods that can be used to bound the behavior of general ERGMs by Bernoulli graphs (i.e., ERGMs with independent edge variables), and used these to show sufficient conditions for ERGMs to avoid certain forms of degeneracy as $N \to \infty$. One area of relatively rich theoretical development in the ERGM literature has been the derivation of sufficient statistics from first principles (particularly dependence conditions). Following the early work of Frank and Strauss (1986), many papers in this area employ Hammersley-Clifford constructions (Besag, 1974) in which initially posited axioms for conditional dependence among edge variables (usually based on substantive theory) are used to generate sets of statistics sufficient to represent all pmfs with the posited dependence structure. Examples of such work for single-graph ERGMs include Wasserman and Pattison (1996), Pattison and Robins (2002), and Snijders et al. (2006), with multi-relational examples including Pattison and Wasserman (1999) and Koehly and Pattison (2005). Snijders (2010) has showed that statistics based on certain forms of dependence allow for models that allow conditional marginalization across components (i.e., graph components are conditionally independent); this suggests statistics that may be appropriate for social processes in which edges can only influence each other "through" the network itself, and provides insight into circumstances which facilitate inference for population network parameters from data sampled at the component level (see also Shalizi and Rinaldo, 2013). An alternative way to motivate model statistics is via generative models that treat the observed network as arising from a stochastic choice process. Examples of such developments include Snijders (2001) and Almquist and Butts (2013).

1.2.2.1 Bayesian Inference for ERGM Parameters   Given the likelihood of equation (1.1), Bayesian inference follows in the usual fashion by application of Bayes' Theorem, i.e.,

$$p(\theta|Y = y, S, X) = \frac{ERG(y|\theta, S, X)p(\theta|S, X)}{\int_{\mathbb{R}^s} ERG(y|\theta', S, X)p(\theta'|S, X)d\theta'}$$
$$\propto ERG(y|\theta, S, X)p(\theta|S, X),$$

where $p(\theta|Y = y, S, X)$ is the posterior density of θ given the observed state of Y, statistic vector S, and covariate set X, $p(\theta|S, X)$ is the corresponding prior density of theta on $\mathbb{R}^s$, and $ERG(y|\theta, S, X)$ represents the ERGM likelihood for $\Pr(Y = y|\theta, S, X)$ from equation (1.1). In the case of ERGMs belonging to regular exponential families (e.g., non-curved), we can immediately obtain

a conjugate prior for θ using known properties of exponential families:

$$p(\theta|S, X) = \frac{\exp\left(\theta^T \phi - \nu \ln \kappa(\theta, S, X)\right)}{\int_{\mathbb{R}^s} \exp\left(\theta'^T \phi - \nu \ln \kappa(\theta', S, X)\right) d\theta'},$$

where $\phi \in \mathbb{R}^s$ and $\nu > 0$ are hyperparameters and κ is the ERGM normalizing factor (as defined above). Note that ϕ and ν have natural interpretations in terms of "prior pseudo-data" and "prior pseudo-sample size," as is clear from the joint posterior:

$$
\begin{aligned}
p(\theta|Y = y, S, X) &\propto ERG(y|\theta, S, X) \frac{\exp\left(\theta^T \phi - \nu \ln \kappa(\theta, S, X)\right)}{\int_{\mathbb{R}^s} \exp\left(\theta'^T \phi - \nu \ln \kappa(\theta', S, X)\right) d\theta'} \\
&= \frac{\exp\left(\theta^T S(y, X)\right)}{\kappa(\theta, S, X)} \frac{\exp\left(\theta^T \phi - \nu \ln \kappa(\theta, S, X)\right)}{\int_{\mathbb{R}^s} \exp\left(\theta'^T \phi - \nu \ln \kappa(\theta', S, X)\right) d\theta'} \\
&= \frac{\exp\left(\theta^T \left(\phi + S(y, X)\right) - (\nu + 1) \ln \kappa(\theta, S, X)\right)}{\int_{\mathbb{R}^s} \exp\left(\theta'^T \phi - \nu \ln \kappa(\theta', S, X)\right) d\theta'} \\
&\propto \frac{\exp\left(\theta^T \left(\phi + S(y, X)\right) - (\nu + 1) \ln \kappa(\theta, S, X)\right)}{\int_{\mathbb{R}^s} \exp\left(\theta'^T \left(\phi + S(y, X)\right) - (\nu + 1) \ln \kappa(\theta', S, X)\right) d\theta'}, \quad (1.2)
\end{aligned}
$$

with equation (1.2) giving the (re)normalized form.

Despite the attractiveness of the conjugate prior, it is less helpful than it might be due to the intractability of the ERGM normalizing factor. While standard MCMC methods (e.g., the Metropolis-Hastings algorithm) can often manage intractable normalizing constants of a posterior density when the posterior density in question is known up to a constant, the kernel of equation (1.2) also involves the (usually intractable) normalizing factor κ from the ERGM likelihood. Such posteriors have been described as "doubly intractable" (Murray et al., 2012), and pose significant computational challenges in practice. In the more general case for which $p(\theta)$ does not necessarily include κ (i.e., non-conjugate priors), MCMC or related approaches must generally deal with posterior odds of the form

$$
\begin{aligned}
\frac{p(\theta|Y = y, S, X)}{p(\theta'|Y = y, S, X)} &= \frac{ERG(y|\theta, S, X)}{ERG(y|\theta', S, X)} \frac{p(\theta|S, X)}{p(\theta'|S, X)} \\
&= \exp\left(S(y, X)^T (\theta - \theta')\right) \frac{\kappa(\theta', S, X)}{\kappa(\theta, S, X)} \frac{p(\theta|S, X)}{p(\theta'|S, X)},
\end{aligned}
$$

which still require evaluation of normalizing factor ratios at each step. Provided that the prior ratio can be easily calculated, the complexity of this calculation is no worse than the associated ratios required for likelihood maximization, and indeed MAP can be performed in such cases using MCMC-MLE methods (see e.g. Hunter et al., 2008, 2012, for the MLE case) via the addition of prior odds as a penalty function. Approaches to direct posterior simulation in this regime include the use of exchange algorithms (Caimo and Friel, 2011)

and other approximate MCMC methods (see Hunter et al., 2012, for a review). To date these latter methods have proven too computationally expensive for routine use, but the area is one of active research.

An alternative (if less generically satisfying) approach to the problem arises by observing that there are some classes of models for which κ is directly computable, and hence for which Bayesian analysis is more readily performed. An early example of this work is that of Wong (1987), who provided a fully Bayesian treatment of the p_1 family of Holland and Leinhardt (1981). Because the likelihood for this family factors as a product of categorical pmfs (the four edge variable states associated with each dyad), κ is easily calculated and Bayesian inference is greatly simplified. This intuition was subsequently elaborated by van Duijn et al. (2004), who used it as a basis for a much richer family of effects. Although we are focused here on models in ERGM form, it should also be noted that many latent variable models for networks can be viewed as positing that Y is drawn from an ERGM with strong conditional independence properties (leading to a tractable normalizing factor), given a (possibly very complex) set of latent covariates on which a prior structure is placed. Models such as those of Hoff et al. (2002), Handcock et al. (2007), Nowicki and Snijders (2001) and Airoldi et al. (2008) can be viewed in this light. While the simultaneous dependence in cross-sectional data tends to limit the utility of simplified ERGMs (or to require a shifting of computational burden into a complexly specified parameter structure), this problem is sometimes reduced in dynamic data due to the ability to condition on past observations (i.e., replacing simultaneous dependence in the present with dependence on the past) (Almquist and Butts, 2014). It is to this setting that we now turn.

1.2.3   Temporal Models for Network Data

Temporal models for social network data can be generally classified into two broad categories: (1) continuous time models; and (2) panel data models. Here we will focus only on panel data models – for examples of models for continuous time interaction data see Butts (2008a), DuBois, Butts, McFarland, and Smyth (2013), and DuBois, Butts, and Smyth (2013). Current theory and software are focused on statistical inference for panel data models based on four general approaches. The first is the family of *actor oriented models,* which assumes an underlying continuous-time model of network dynamics, where each observed event represents a single actor altering his or her outgoing links to optimize a function based on sufficient statistics (for details, see Snijders, 1996; Snijders and Van Duijn, 1997; Snijders, 2001, 2005). The second is the family of *latent dynamic structure models,* which treat network dynamics as emerging from a simple network process influenced by the evolution of set of latent covariates; for example, see Sarkar and Moore (2005), Sarkar et al. (2007), and Foulds et al. (2011). The third is the family of *temporal exponential family random graph models* (TERGMs), which attempt to

directly parameterize the joint pmf of a graph sequence using discrete exponential families (Hanneke and Xing, 2007a; Hanneke et al., 2010; Hanneke and Xing, 2007b; Cranmer and Desmarais, 2011; Desmarais and Cranmer, 2011, 2012; Almquist and Butts, 2012, 2013, 2014). Finally, the fourth approach is the *separable temporal ERGM* family (or STERGM), which assumes each panel observation is a cross-sectional observation from a latent continuous time process in which edges evolve via two separable processes of edge formation and edge dissolution (Krivitsky and Handcock, 2010). Here, we focus on the TERGM case.

TERGMs can be viewed as the natural analog of time series (e.g., VAR) models for the random graph case. Typically, we assume a time series of adjacency matrices $\ldots, Y_{t-1}, Y_t, \ldots$ and parameterize the conditional pmf of $Y_t | Y_{t-1}, Y_{t-2}, \ldots$ in ERGM form. As with classical time series models, it is typical to introduce a temporal Markov assumption of limited dependence on past states; specifically, we assume the existence of some $k \geq 0$ such that that Y_t is independent of $Y_{t-k-1}, Y_{t-k-2}, \ldots$ given $Y_{t-1}, \ldots, Y_{t-k} \equiv Y_{t-k}^{t-1}$. Under this assumption, the standard TERGM likelihood for a single observation is written as

$$\Pr(Y_t = y_t \mid Y_{t-k}^{t-1} = y_{t-k}^{t-1}, \theta, S, X_t) = \frac{\exp\left(\theta^T S(y_t, y_{t-k}^{t-1}, X_t)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\theta^T S(y_t', y_{t-k}^{t-1}, X_t)\right)} \mathbb{I}_{\mathcal{Y}}(y_t).$$

(1.3)

As before, S is an s-vector of real-valued sufficient statistics, but for the TERGMs $S : \mathcal{Y}^{k+1}, \mathcal{X} \mapsto \mathbb{R}^s$ (i.e., each function may involve observations at the k time points prior to t instead of a single graph). Otherwise, nothing is intrinsically different from the cross-sectional case. (In particular, note that from the point of view of Y_t, y_{t-k}^{t-1} is a fully observed covariate. This is useful for the development that follows.) The denominator of (1.3) is again intractable in the general case, as it is for ERGMs.

For a complete TERGM series, the joint likelihood of the sequence $Y_1, \ldots, Y_t$ is given by $\prod_{i=1}^{t} TERG(y_i | Y_{i-k}^{i-1} = y_{i-k}^{i-1}, \theta, S, X_i)$, where $TERG$ refers to the single-observation TERGM likelihood of equation (1.3). MCMC-based maximum likelihood estimation for θ is feasible for very short series, but becomes costly as sequence length grows. Cranmer and Desmarais (2011) propose estimation via MPLE combined with a bootstrapping procedure to estimate standard errors as a computationally cheaper alternative. Alternately, scalable estimation is greatly simplified for TERGMs with no simultaneous dependence terms; i.e., models such that Y_{ijt} is conditionally independent of Y_{klt} given Y_{t-k}^{t-1} for all distinct $(i, j), (k, l)$. The TERGM likelihood for such models reduces to a product of Bernoulli graph pmfs, and hence the corresponding inference problem is equivalent to (dynamic) logistic regression. Although by no means novel, these conditional Bernoulli families have recently been advocated by Almquist and Butts (2014) as viable alternatives for network time series in which the time period between observations is on or faster than the time scale of network evolution, or whether it is for other reasons possible to

capture much of the simultaneous dependence among edges by conditioning on the past history of the network. Almquist and Butts (2014) also show how this family can be easily extended to incorporate endogenous vertex dynamics (a feature not currently treated in other dynamic network families). In the remainder of this chapter, we focus on this case, with a particular emphasis on Bayesian inference for joint vertex/edge dynamics.

1.2.3.1 TERGM with Vertex Dynamics

The TERG model in Section 1.2.3 can be further extended to handle vertex dynamics by employing a separable parameterization between the vertex set and edge set as proposed by Almquist and Butts (2014). Here we take the vertex set V_t to arise at each point in time from a fixed support of possible vertex sets, $\mathcal{V}$, with the associated pmf parameterized via a discrete exponential family. Y_t then arises from an ERG distribution conditional on V_t. To clarify notation, let $Z_t = (V_t, Y_t)$ be a representation for graph G_t, and as before let Z_a^b be the network time series $Z_a, \ldots, Z_b$. The pmf for a single observation under vertex dynamics is then

$$
\begin{aligned}
\Pr(Z_t = z_t \mid & Z_{t-k}^{t-1} = z_{t-k}^{t-1}, \psi, \theta, W, S, X_t) \\
= \Pr(V_t &= v_t \mid Z_{t-k}^{t-1} = z_{t-k}^{t-1}, \psi, W, X_t) \\
\times \Pr(Y_t &= y_t \mid V_t = v_t, Z_{t-k}^{t-1} = z_{t-k}^{t-1}, \theta, S, X_t) \\
= \frac{\exp\left(\psi^T W(v_t, z_{t-k}^{t-1}, X_t)\right)}{\sum_{v' \in \mathcal{V}} \exp\left(\psi^T W(v'_t, z_{t-k}^{t-1}, X_t)\right)} &\times \frac{\exp\left(\theta^T S(y_t, v_t, z_{t-k}^{t-1}, X_t)\right)}{\sum_{y' \in \mathcal{Y}_{v_t}} \exp\left(\theta^T S(y'_t, v_t, z_{t-k}^{t-1}, X_t)\right)},
\end{aligned}
\tag{1.4}
$$

where $\mathcal{Y}_{v_t}$ is the set of possible adjacency matrices compatible with vertex set v_t, W is a w-vector of sufficient statistics on the vertex set, and ψ is a w-vector of vertex set parameters. The joint TERGM likelihood for a time series is then the product of the likelihoods for each observation. We refer to the conditional likelihood of a single observation in equation (1.4) as $TERGV$ (i.e., temporal exponential family random graph with vertex processes) in the discussion that follows.

The likelihood of equation (1.4) is inferentially "separable" in the sense that it factorizes into terms respectively dealing with ψ (and the vertex set) and with θ (and the edge set). These may be estimated separately, even when both depend on the same data (i.e., the edge history and vertex history may both enter into S and W). On the other hand, inferential separability does not imply *predictive* separability: the vertex model will strongly impact the edge structure of graphs drawn from the model, and in some cases vice versa. [See Almquist and Butts (2014) for a discussion.]

1.2.3.2 Bayesian Estimation of TERGMs

As before, Bayesian inference for the full TERGM family (with vertex dynamics) is based on the posterior

distribution of θ, ψ given $Z_1, \ldots, Z_t$:

$$p(\psi, \theta | Z_1^t = z_1^t, S, W, X) \propto p(\psi, \theta | S, W, X)$$

$$\times \prod_{i=1}^t TERGV(z_i | Z_{i-k}^{i-1} = z_{i-k}^{i-1}, \psi, \theta, W, S, X_i).$$

It is frequently reasonable to treat the parameters of the edge and vertex processes as *a priori* independent given X. In that case, the above factors as

$$p(\psi, \theta | Z_1^t = z_1^t, S, W, X) \propto p(\psi | W, X) p(\theta | S, X)$$

$$\times \prod_{i=1}^t TERGV(z_i | Z_{i-k}^{i-1} = z_{i-k}^{i-1}, \psi, \theta, W, S, X_i)$$

$$= p(\psi | W, X) \frac{\exp\left(\psi^T W(v_t, z_{t-k}^{t-1}, X_t)\right)}{\sum_{v' \in \mathcal{V}} \exp\left(\psi^T W(v_t', z_{t-k}^{t-1}, X_t)\right)}$$

$$\times p(\theta | S, X) \frac{\exp\left(\theta^T S(y_t, v_t, z_{t-k}^{t-1}, X_t)\right)}{\sum_{y' \in \mathcal{Y}_{v_t}} \exp\left(\theta^T S(y_t', v_t, z_{t-k}^{t-1}, X_t)\right)},$$

which implies that the joint posterior itself factors as

$$p(\psi | Z_1^t = z_1^t, W, X) \propto p(\psi | W, X) \frac{\exp\left(\psi^T W(v_t, z_{t-k}^{t-1}, X_t)\right)}{\sum_{v' \in \mathcal{V}} \exp\left(\psi^T W(v_t', z_{t-k}^{t-1}, X_t)\right)} \quad \text{and}$$

$$p(\theta | Z_1^t = z_1^t, S, X) \propto p(\theta | S, X) \frac{\exp\left(\theta^T S(y_t, v_t, z_{t-k}^{t-1}, X_t)\right)}{\sum_{y' \in \mathcal{Y}_{v_t}} \exp\left(\theta^T S(y_t', v_t, z_{t-k}^{t-1}, X_t)\right)}. \quad (1.5)$$

This is a manifestation of the inferential separability remarked on previously. Although ψ and θ are jointly dependent on both the covariates and the observed data, the two may be analyzed independently. In the special case where no vertex dynamics are present, or where such dynamics are exogenous, the joint posterior simplifies to that of equation (1.5).

As with the ERGM case, posterior estimation for TERGMs inherits the normalizing factor problem (exacerbated in the case of vertex dynamics by the presence of two distinct exponential families, each with a normalizing factor!). Because of these technical complications there has been very little work in applying Bayesian analysis to the more general TERGM framework.[3] In the special case in which all observations in the present are independent conditional on the past, however, the normalizing factor becomes tractable and analysis is greatly simplified. As noted above, the similarity of the resulting inference problem to logistic regression (and the direct analogy with existing

---

[3]There has been some work into modeling the generative process of TERGM models through Bayesian networks (not to be confused with Bayesian models *of* networks) by Guo et al. (2007).

network regression models) has led to these being dubbed "dynamic network logistic regression" (DNR) families (Almquist and Butts, 2014). In the following section we will discuss Bayesian estimation in the DNR case, with and without vertex dynamics.

1.3   Dynamic Network Logistic Regression with Vertex Dynamics

Dynamic Network Logistic Regression with vertex dynamics was introduced by Almquist and Butts (2014), who modeled $\Pr(V_t = v_t \mid Z_{t-k}, X_t)$ and $\Pr(Y_t = y_t \mid V_t, Z_{t-k}, X_t)$ as separable processes whose elementwise realizations are independent conditional on covariates and past history. Under the necessary conditional independence, homogeneity, and temporal Markov assumptions one can derive the likelihood function for DNR from the general $TERGV$ likelihood, with the vertex likelihood given by

$$\Pr(V_t = v_t \mid Z_{t-k}^{t-1} = z_{t-k}^{t-1}, \psi, W, X_t)$$
$$= \prod_{i=1}^{n} B\left(\mathbb{I}(v^{(i)} \in v_t) \left| \mathrm{logit}^{-1}\left(\psi^T W(i, z_{i-k}^{t-1}, X_t)\right)\right.\right)$$

and the edge likelihood by

$$\Pr(Y_t = y_t \mid V_t = v_t, Z_{t-k}^{t-1} = z_{t-k}^{t-1}, \theta, S, X_t)$$
$$= \prod_{(v^{(i)}, v^{(j)}) \in v_t \times v_t} B\left(Y_{ijt} = y_{ijt} \left| \mathrm{logit}^{-1}\left(\theta^T S(i, j, v_t, z_{t-k}^{t-1}, X_t)\right)\right.\right), \quad (1.6)$$

where B is understood to be the Bernoulli pmf, $\mathbb{I}$ is the indicator function, and $v^{(i)}$ indicates the ith vertex from a known total risk set V^{max}. (Thus, the support $\mathcal{V}$ of V_t is the power set of V^{max}.) The analogy of this model family with logistic regression is clear from the form of the joint likelihood, which is equivalent to a (relatively complex) logistic regression of indicator variables for edge and vertex set memberships on a set of statistics associated with the network history and/or covariates. In the special case when V_t is exogenously varying, the joint likelihood of the data reduces to the edge process in equation (1.6); when it is fixed, the likelihood reduces to the "classic" dynamic network logistic regression model. Model specification, maximum likelihood based inference, and adequacy checking for this family are discussed in Almquist and Butts (2013, 2014).

1.3.1   Bayesian Inference for DNR Parameters

Because the DNR family reduces to a logistic regression structure, Bayesian inference is nominally straightforward. However, choice of prior structure for

DNR families has not been explored to date. Justifiably or otherwise, researchers typically seek to employ a default prior specification if they do not have a strong rationale for endorsing a specific prior. There is an extensive literature on noninformative, default, and reference prior distributions within the Bayesian statistical field (see Jeffreys, 1998; Hartigan, 1964; Bernardo, 1979; Spiegelhalter and Smith, 1982; Yang and Berger, 1994; Kass and Wasserman, 1996). More recent work has continued the traditions of both research on informative prior distributions using application-specific information and on minimally informative priors (often motivated by invariance principles) (for a review see, Gelman et al., 2008). One increasingly widely used approach to evaluating default priors (particularly in the machine learning literature) is the use of *predictive assessment*, i.e., examination of the extent to which a given prior structure reliably leads to accurate predictions on test data for a given body of training data. While arguably less principled than priors derived from other considerations, priors found to give good predictive performance on past data may be attractive on pragmatic grounds; by turns, such priors can also be justified more substantively as representing distributions *compatible with past observations* on similar data, and hence plausible at least as a crude starting point. Likewise, priors that consistently lead to poor predictive performance on test data should be suspect, whatever the principles used to construct them. The balance of this chapter is thus concerned with the predictive evaluation of various candidate priors in the context of DNR models.

While both MCMC and MAP are feasible for inference in DNR families, our focus here will be on posterior simulation via MCMC. In addition to giving us a more complete view of the posterior distribution, posterior simulation is particularly well-adapted to predictive model adequacy checking (Gelman et al., 2004). Specifically, simulation of future observations conditional on a point estimate (e.g., the posterior mean or mode) can greatly underestimate the uncertainty associated with the posterior distribution, and by extension can fail to reveal the benefits to be gained by, e.g., prior specifications that reign in extreme parameter values without greatly changing the central tendency of the posterior distribution.

Given the above, there are many reasonable choices of prior specifications for DNR families that may be applicable in one or another context. Given that our focus here is on evaluating simple, default priors, we will focus our attention on four prior specifications suggested as default priors for logistic regression in the Bayesian statistical literature (e.g., Gelman et al., 2008). Our core questions are as follows. First, what are the inferential consequences of employing these reference priors versus maximum likelihood estimation for DNR families in typical social network settings? Second, to what extent do various reasonable default priors lead to differences in either point estimation or posterior uncertainty in such settings? Finally, what differences (if any) does selection of one or another default prior make to prediction, in the specific sense of forecasting properties of an evolving network? If, in typical

settings, inferential and predictive outcomes are fairly insensitive to choice of prior, then selection based on computational or other factors may be a reasonable practice. If, by contrast, we find substantial differences in inferential and/or predictive performance among default priors, then these choices must be scrutinized far more carefully. The balance of this chapter is intended as a first step towards assessing these questions.

Although simulation and inference for ERGMs is in general a highly specialized art (see, e.g., Snijders, 2002; Hunter et al., 2008; Wang et al., 2009), the logistic form of DNR families facilitates parameter estimation (if not network simulation) using more standardized tools and techniques. Examples of off-the-shelf toolkits suitable for posterior simulation in the cases studied here include Winbugs, JAGS or MCMCpack in R (Spiegelhalter et al., 2003; Plummer, 2003; Martin et al., 2011). Here, we employ the Metropolis-Hastings algorithm as implemented Martin et al. (2011) for posterior simulation, with sufficient statistics computed in the same manner as Almquist and Butts (2014, 2013) via custom `statnet`-based tools (Handcock et al., 2003). The latter were also used for posterior predictive simulation of graph-theoretic quantities. MCMC convergence was assessed using both Geweke's convergence diagnostic (Geweke et al., 1991) and Raftery and Lewis's diagnostic (Raftery and Lewis, 1992).

1.3.2   Bayesian Estimation of DNR with Vertex Dynamics

As noted above, inferential separability of the edge and vertex processes in the DNR context allows both to be treated as independent logistic regression problems (so long as the associated parameters are *a priori* independent). As this implies, there is no reason that the same prior structure must be used for both; for simplicity and practicality, however we here consider the case where we assume both the edge and vertex parameters have the same prior distributions.

As noted above, we here consider a number of typical priors for logistic regression (which have been recommended in the literature) for use as priors in DNR with and without vertex dynamics. The baseline point of comparison for all Bayesian results will be the ML estimate (and its sampling distribution), reflecting the dominant practice within the ERGM literature. In addition to this baseline, we consider five prior specifications within three general classes: (1) an improper uniform prior (i.e., the fully-Bayesian counterpart to maximum likelihood estimation); (2-3) independent normally distributed priors (one centered at 0, and one offset to inflate prior density); and (4-5) weakly informative t families of proper priors recommended by Gelman et al. (2008) (a Cauchy distribution centered at 0 with a scale parameter of 2.5, and a t_7 prior centered at 0). Empirical experiments on the effects of these priors on the analysis and interpretation can be found in Section 1.4.

1.4   Empirical Examples and Simulation Analysis

Given the six specifications described above (including the MLE), we seek to evaluate the practical consequences of these priors for Bayesian inference in typical settings. To that end, we consider a comparative analysis of inference under our proposed priors for two empirical cases. The first is a dynamic network of citations among bloggers during the 2004 US presidential election ("*Blog*"), and the second is a dynamic network of face-to-face communication ties among windsurfers on a southern California beach (*"Beach"*). Both networks are typical of social network data sets in that each involves multiple, complex mechanisms of interaction as well as actor-level heterogeneity. These networks have also been studied in the literature, making them useful reference cases for our present analysis.

1.4.1   Blog Data

The first data set considered in this chapter is a dynamic inter- and intra-group blog citation network collected by Butts and Cross (2009) and previously analyzed with DNR in Almquist and Butts (2013). This dynamic network consists of interactions among all blogs credentialed by the Democratic National Committee (DNC) or Republican National Committee (RNC) for their respective 2004 conventions. The vertex set is comprised of actors representing 34 DNC and 14 RNC credentialed blogs, as well as the one blog credentialed by both groups, and is static over the observation period. The edge set is comprised of directed hyperlinks on the front pages of the aforementioned blogs as observed over a 121 day period. The data was obtained by automatically querying the main page of each blog at six hour intervals starting at midnight, Pacific time. The period of study begins on 7/22/04 (shortly before the DNC convention) and ends 11/19/04 (shortly after the Presidential election), leading to a total of 484 time points. There exists an edge from blog i to blog j at time t iff a link to blog j appears on the main page of i at time t. Specifically, this data is represented as an adjacency array, Y, such that $Y_{ijt} = 1$ if i cites (i.e., links to) j at time t, and 0 otherwise.

1.4.2   Beach Data

The second data set considered in this chapter was originally collected by Freeman et al. (1988). It is a dynamic interpersonal communication network of windsurfers in the late 1980s which Freeman et al. (1988) subsequently analyzed in aggregate. These data have been used in a number of influential articles (see Cornwell, 2009; Hummon and Doreian, 2003; Zeggelink et al., 1996, among others), and later analyzed dynamically using DNR with vertex dynamics in Almquist and Butts (2014). The network was originally collected daily (aggregated over a morning and an afternoon observation period) for 31

days from August 28, 1986 to September 27, 1986.[4] The data set consists of 95 individuals which were originally classified ethnographically by Freeman et al. (1988) into *"regulars"* ($N = 54$)—frequent attendees who were well-integrated into the social life of the beach community—and *"irregulars"* ($N = 41$). Further, the *regulars* were categorized by the researchers into two groups, *group 1* ($N = 22$) and *group 2* ($N = 21$), with 11 individuals not classified as belonging to either group 1 or group 2. The number of windsurfers appearing on the beach ranged from 3 to 37 individuals over the observation period. The number of communication ties per day ranging from 0 to 96 in the network.

1.4.3   Case Analysis: Static Vertex Set

Given a collection of alternative priors, what are the consequences of choosing one or another for inference in the case of a DNR family with a fixed vertex set? To assess this, we consider parameter estimation and prediction on the Blog data. While the entire time series consists of 484 time points, we here restrict ourselves to a much smaller series of 32 time points, in order to form a reasonable comparison with the Beach data. As noted below, the likelihood specification (i.e., sufficient statistics or effects) for models employed here is based on prior work by Almquist and Butts (2013); we hold the effects constant across all models to isolate the impact of prior specifications.

Because V_t is fixed here, we need only consider priors for the edge parameters (θ). Posterior distributions based on these models are compared the standard ML estimate derived from iterative weighted least squares methods (McCullagh and Nelder, 1999), noting that the ML estimate can also be thought of as a MAP estimator under an improper uniform prior. The posteriors evaluated are as follows: (1) an improper uniform prior (i.e., a fully Bayesian analog to the MLE); (2) independent $N(0, .1)$ priors on each parameter (a simple default choice with relatively strong shrinkage towards 0); (3) independent $N(5, 1)$ priors on each parameter (included as a non-normative test to evaluate the consequences of poor prior specification); (4) independent standard t_7 priors on each parameter, emulating the example used by Gelman et al. (2008); and (5) independent scaled Cauchy priors $C(0, 2.5)$ as recommended by Gelman et al. (2008) for logistic regression in another context. All Bayesian DNR models were be estimated via posterior simulation by MCMC, thus allowing for improved predictive modeling. This is discussed in the following section.

Our specification of the Blog network model follows Almquist and Butts (2013); for expository purposes, we have simplified the original model slightly by removing less important effects. Those retained fall into two main classes: (1) density effects, e.g., DNC/RNC mixing and daily seasonality; and (2)

---

[4]Unfortunately, one day (September 21st) is missing due to a race on a different beach, which precluded data collection. Thus, complete data is available for 30 days during the observation period.

lagged network effects, e.g., interaction or preferential attachment effects (Wasserman and Faust, 1994). In Table 1.1 we consider first mixing effects for DNC and RNC internal ties (respectively), followed by terms for interaction between DNC and RNC blogs; next we consider daily effects, *Tuesday* to *Sunday* effects with *Monday* as the reference category; a lag effect (Y_{t-1}); and a lagged indegree effect $\log(Deg(Y_{t-1}))$, which can be thought of as preferential attachment effect. The table is laid out to illustrate the differences and similarities of the parameter estimates for different minimally informative priors on estimation. Table 1.1 demonstrates that most reasonable minimally informative priors produce results comparable to each other and the MLE, as suggested by Gelman et al. (2008). It is worth pointing out that the prior *can* have a large effect on the parameter estimate (and thus the resulting predictions, see Section 1.4.3.1): the $N(5,1)$ parameter is noticeably farther from the other estimates (although the bias is not always in a positive direction, e.g., for the lag and lagged degree terms). That said, it is perhaps reassuring that even a fairly strong prior still leads to generally comparable estimates for a data set of only moderate size (44 nodes over 32 time points). In general, the MLE and posterior mean estimates for the standard default priors differ by an amount that is approximately an order of magnitude smaller than the statistical uncertainty associated with the estimates themselves; for the "extreme" $N(5,1)$ prior, the differences are on the scale of the posterior standard deviation. Likewise, the posterior standard deviations themselves are generally quite similar to each other, and to the standard error of the MLE (although, again, the "extreme" prior deviates somewhat).

To further illustrate the similarities and differences among estimates, we consider Figure 1.1 which contains the 95% posterior marginals for each parameter (where the ML case shows the asymptotic Gaussian approximation to the sampling distribution of the MLE). Here we see again that the typical minimally informative priors result in relatively similar posterior distributions, both in the central mass and in the tails; all are likewise quite similar to the MLE. The deviations associated with the "extreme" prior are clearer here, and demonstrate visually the fact that (for models of this sort) a bias in the prior may not manifest within the posterior in a clear way. Specifically, for this case the inflated prior mean on the baseline interaction rate parameters tends to lead to a compensating deflation of the lagged terms; such effects suggest that it is important to be careful when using highly informative priors, since influence on one parameter may propagate to the posterior marginals for other parameters in a non-obvious manner.

The fact that all reasonable prior specifications lead to posterior marginals that are close to each other and to the sampling distribution of the MLE suggests that we are effectively operating within the regime in which the posterior
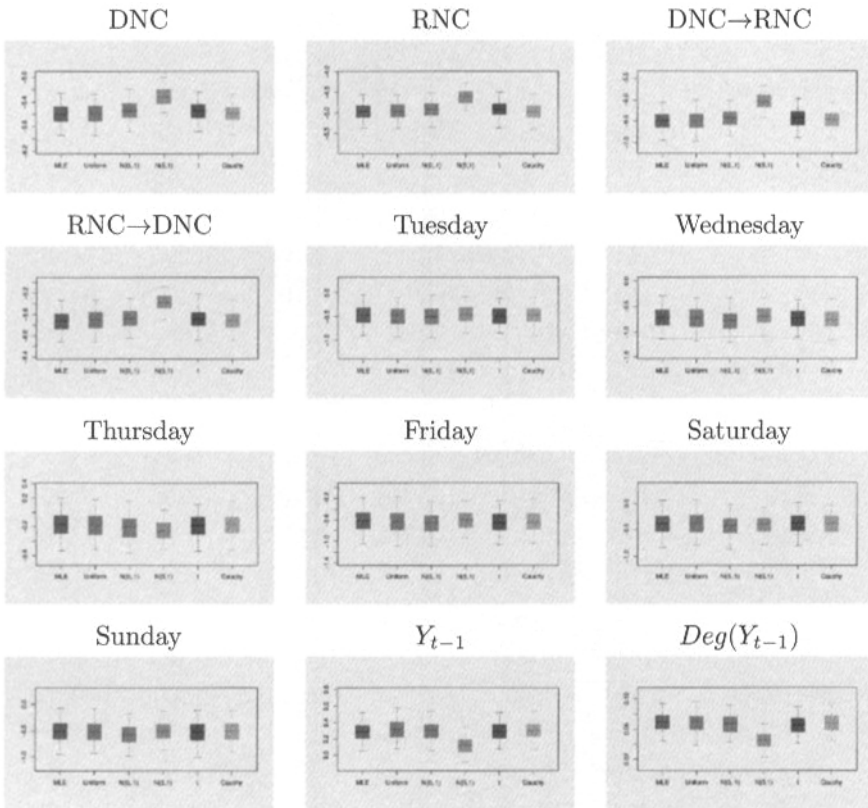
Figure 1.1: Posterior marginals for the five prior specifications discussed in Section 1.4.3; MLE sampling distribution based on asymptotic Gaussian approximation

Table 1.1: Parameter estimates: MLE and posterior means under the five prior specifications discussed in Section 1.4.3 (standard errors and posterior standard deviations are given in parentheses).

	MLE	U	$N(0, .1)$	$N(5, 1)$	t_7	$C(0, 2.5)$
DNC	−5.598	−5.599	−5.543	−5.308	−5.547	−5.589
	(0.176)	(0.184)	(0.176)	(0.152)	(0.164)	(0.164)
RNC	−4.959	−4.956	−4.919	−4.61	−4.917	−4.966
	(0.208)	(0.219)	(0.217)	(0.178)	(0.201)	(0.216)
DNC→RNC	−6.486	−6.494	−6.432	−6.027	−6.425	−6.474
	(0.225)	(0.243)	(0.221)	(0.196)	(0.238)	(0.218)
RNC→DNC	−5.728	−5.711	−5.672	−5.37	−5.681	−5.714
	(0.199)	(0.212)	(0.196)	(0.164)	(0.198)	(0.194)
Y_{t-1}	9.284	9.317	9.302	9.121	9.296	9.302
	(0.119)	(0.129)	(0.122)	(0.108)	(0.117)	(0.119)
$Deg(Y_{i-1})$	0.069	0.068	0.067	0.045	0.065	0.068
	(0.013)	(0.013)	(0.013)	(0.011)	(0.013)	(0.013)
Monday	−	−	−	−	−	−
Tuesday	−0.469	−0.495	−0.509	−0.461	−0.501	−0.482
	(0.213)	(0.219)	(0.219)	(0.197)	(0.198)	(0.205)
Wednesday	−0.717	−0.728	−0.781	−0.683	−0.734	−0.745
	(0.214)	(0.222)	(0.221)	(0.193)	(0.198)	(0.202)
Thursday	−0.167	−0.187	−0.216	−0.257	−0.197	−0.179
	(0.184)	(0.182)	(0.183)	(0.147)	(0.174)	(0.169)
Friday	−0.619	−0.636	−0.667	−0.601	−0.646	−0.635
	(0.219)	(0.232)	(0.222)	(0.186)	(0.216)	(0.219)
Saturday	−0.388	−0.388	−0.428	−0.412	−0.388	−0.393
	(0.216)	(0.231)	(0.206)	(0.183)	(0.211)	(0.208)
Sunday	−0.51	−0.517	−0.571	−0.505	−0.529	−0.514
	(0.217)	(0.213)	(0.210)	(0.195)	(0.217)	(0.210)

is close to its asymptotic Gaussian limit (Gelman et al., 2004). Although this phenomenon is data dependent, our case suggests that, for models of this kind, data sets of moderate size may prove adequate for the asymptotic approximation to hold. As such, any long-tailed, symmetric prior distribution is likely to work well from an inferential point of view, and choice of distribution can be reasonably made for computational or other reasons.

1.4.3.1 One-step Prediction for Model Assessment   Inference is vital for hypothesis testing, model understanding, and the like; however, one can obtain parameters that are qualitatively reasonable from models that nevertheless predict poorly. In this section we will focus on one type of predictive model assessment procedure, described by Almquist and Butts (2014) in the DNR context as *inhomogeneous Bernoulli prediction*. The algorithm as suggested by Almquist and Butts (2014) is as follows: for each time point t we predict $Y_t | Y_{t-k}^{t-1} = y_{t-k}^{t-1}$ n times (in other words, we take n draws from the posterior predictive distribution of the network at time t given the previous k observed time steps) to predict the edge structure. We then summarize the resulting network via a suite of Graph Level Indices (GLI; Anderson et al., 1999), yielding a GLI distribution for each time point.

Here we consider two important graph level indices: density (Wasserman and Faust, 1994) and the fraction of triads that form 3-cliques (i.e., triangles) (Wasserman and Faust, 1994). The first of these is an extremely fundamental GLI that describes the fraction of all possible ties that are present in the network; the second is a simple index related to clustering. For purposes of assessment, we examine the one-step inhomogeneous Bernoulli predictions of graph density and 3-clique formation rates (respectively) under the ML and Bayesian posterior predictive distributions. The results are summarized in Figures 1.2 and 1.3. An important advantage of using posterior predictive distributions in a temporal modeling context (versus point estimates) is that this allows for a realistic propagation of posterior uncertainty in parameter estimates into the resulting predictive distribution for prediction in temporal modeling is that it allows for more realistic prediction through sampling of the prior distribution. We can see this quite clearly in Figures 1.2 and 1.3, where the prediction intervals for the MCMC Bayesian estimates are typically much larger than those of the MLE (blue dotted lines). Further, the we see that the "extreme" $N(5,1)$ prior has surprisingly minimal effect on the prediction estimates and prediction intervals. Overall, it appears that the increased variance in the model is a more realistic portrayal of the data as wider prediction interval often covers the observed data, while the "pure" ML estimated prediction often does not.

1.4.4 Bayesian DNR with Vertex Dynamics

To experiment with the effects of different minimally informative priors on models with vertex dynamics, we again look to an empirical case; here we begin by considering the beach data set collected by Freeman et al. (1988), which is comprised of windsurfers engaging in interpersonal communication on a beach in South California in early fall of 1986. This is a temporally
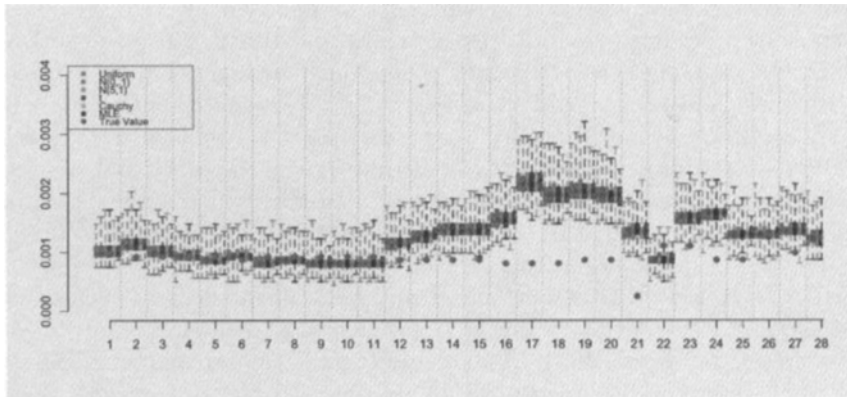
Figure 1.2: One-step prediction of graph density; red dots indicate observed values.
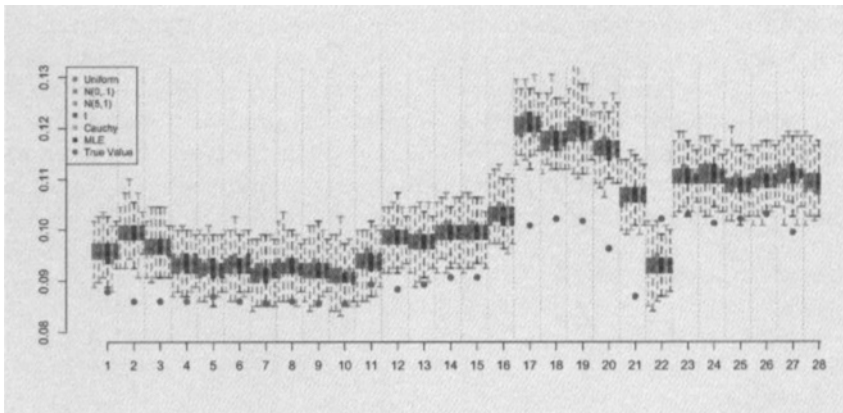

Figure 1.3: One-step prediction of fraction of triads forming 3-cliques; red dots indicate observed values.

evolving network that includes endogenous vertex dynamics, i.e., change in windsurfers who show up over time. Thus, our priors here involve two classes of parameters: θ (the edge parameters), and ψ (the vertex parameters). For the present study, we use the same priors as in the Blog case (along with the MLE), with the same prior specification employed in each instance for θ and for ψ.

Our specification of the likelihood for the Beach data is a simplified version of the model employed by Almquist and Butts (2014). Because the model includes vertex dynamics, note that effects must be specified for both the edge set (given the vertices present each day) and the vertex set (given the past history up to the day in question). We include two main classes of effects: (1) density effects, e.g., weekday/weekend effects and log of the vertex set size[5]; and (2) lagged network effects, e.g., interaction or appearance on the day before and the number of interactions one engaged in the day before (i.e., degree effects). Degree effects may be particularly important in this context as they capture general tendency towards engagement in the network the day before. In Table 1.2 we consider *weekend* and *weekday* effects for both the vertex and edge set dynamics (in a beach setting we expect this to be particularly important as it captures the natural rhythm of activity of the work week in the United States), a $\log(n_t)$ effect for the edge set, a lag effect for both the edge (Y_{t-1}) and vertex set (V_{t-1}), and last a lagged degree effect for both the edge $\log(Deg(Y_{t-1}))$ and vertex set $Deg(V_{t-1})$. The table is laid out to illustrate the differences and similarities of the parameter estimates for different minimally informative priors on estimation (as well as the "extreme" $N(5,1)$ prior as a point of comparison).

Table 1.2 demonstrates that (as with the Blog model), we obtain comparable point estimates using either the MLE or standard default priors, with variation across prior specifications generally at or below the level of statistical uncertainty associated with the estimates themselves. As before, the "extreme" prior has a noteworthy effect, although for many parameters the posterior means given this prior are quite close to those obtained via other specifications. The primary exceptions in this regard are the density related effects, which are more weakly estimated and subject to considerable influence by a strongly informative prior. On the whole, however, our results suggest that reasonable default priors would lead to qualitatively (and quantitatively similar conclusions) for the processes shaping the Beach network.

Turning to the posterior marginal distributions, Figure 1.4 shows results that closely parallel the above (and our findings for the Blog data). The posteriors under reasonable default priors closely mirror each other and the MLE

---

[5]This helps to stabilize mean degree in the presence of changing vertex set sizes; see Krivitsky et al. (2011) and Butts (2011).

Table 1.2: Posterior means and MLE estimates for the Beach data under the prior specifications discussed in Section 1.4.4 (posterior standard deviations and MLE standard errors given in parentheses).

	MLE	U	$N(0,.1)$	$N(5,1)$	t_7	$C(0,2.5)$
Edge Parameters						
Weekend (E)	−0.966	−0.932	−0.922	0.240	−0.687	−0.875
(s.e./p.s.d.)	(0.379)	(0.390)	(0.375)	(0.318)	(0.337)	(0.367)
Weekday (E)	−1.679	−1.635	−1.635	−0.318	−1.365	−1.583
(s.e./p.s.d.)	(0.428)	(0.434)	(0.423)	(0.356)	(0.380)	(0.415)
$\log(n_t)$	−0.096	−0.108	−0.110	−0.491	−0.188	−0.126
(s.e./p.s.d.)	(0.125)	(0.127)	(0.123)	(0.105)	(0.111)	(0.121)
Y_{t-1}	0.995	0.989	0.991	1.072	0.956	0.981
(s.e./p.s.d.)	(0.174)	(0.176)	(0.175)	(0.173)	(0.171)	(0.175)
$Deg(Y_{t-1})$	0.037	0.036	0.037	0.026	0.036	0.037
(s.e./p.s.d.)	(0.015)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)
Vertex Parameters						
	MLE	U	$N(0,.1)$	$N(5,1)$	t_7	$C(0,2.5)$
Weekday (V)	−1.858	−1.86	−1.858	−1.828	−1.846	−1.858
(s.e./p.s.d.)	(0.076)	(0.073)	(0.077)	(0.075)	(0.076)	(0.076)
Weekend (V)	−1.935	−1.93	−1.937	−1.850	−1.910	−1.930
(s.e./p.s.d.)	(0.128)	(0.129)	(0.128)	(0.122)	(0.126)	(0.128)
V_{t-1}	0.737	0.725	0.727	0.768	0.699	0.723
(s.e./p.s.d.)	(0.230)	(0.232)	(0.235)	(0.221)	(0.223)	(0.231)
$Deg(V_{t-1})$	0.596	0.598	0.602	0.57	0.604	0.603
(s.e./p.s.d.)	(0.144)	(0.142)	(0.146)	(0.141)	(0.139)	(0.146)

sampling distribution, suggesting that we are once again within the Gaussian asymptotic regime. The use of an extreme informative prior does clearly disrupt inference, with a stronger effect on some parameters than others; as before, the sign of the effect on a given parameter may not reflect the direction of prior bias, due to relationships among the parameters within the likelihood.

1.4.4.1 One-step Prediction for Model Assessment   For the Blog models, we saw that while the MLE was inferentially comparable to the posterior distributions obtained by reasonable default priors, the resulting predictions (MLE-based simulations versus the full posterior predictive) were quite different.
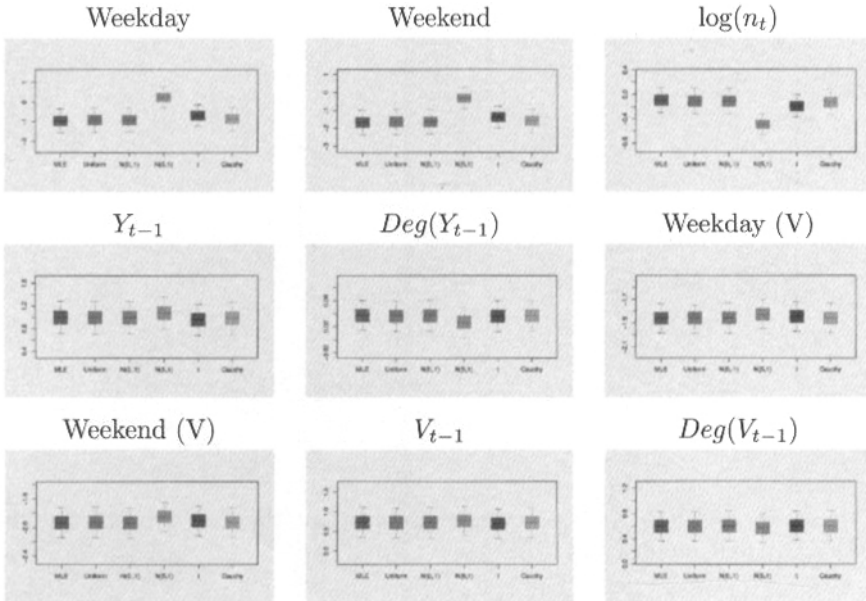
Figure 1.4: Posterior marginals and MLE sampling distributions for the Beach data.

Here, we repeat this analysis for the Beach models, bearing in mind that we now sample vertices as well as edges. (I.e., at each step we first draw the vertex set from the associated model, next drawing the edge set conditional on the realization of the vertex set.) As before, we employ the density and fraction of triads forming 3-cliques in the next time period as our outcome measures.

The results of our simulation experiment are summarized in Figures 1.5 and 1.6. Broadly, the results are similar to those of the Blog data: incorporating posterior uncertainty into our predictions leads to much wider prediction intervals, and these are often necessary to capture the observed data. In general, predicting clique structure is hard for this model (it has no simultaneous dependence, and many other potential predictors have been removed), but the uncertainty associated with the model predictions is clearer for the Bayesian models than for the MLE. There is not an obvious "winner" from the default priors, although the extreme tails of their prediction intervals do vary somewhat from time point to time point. In general, the differences between them appear small and unsystematic.
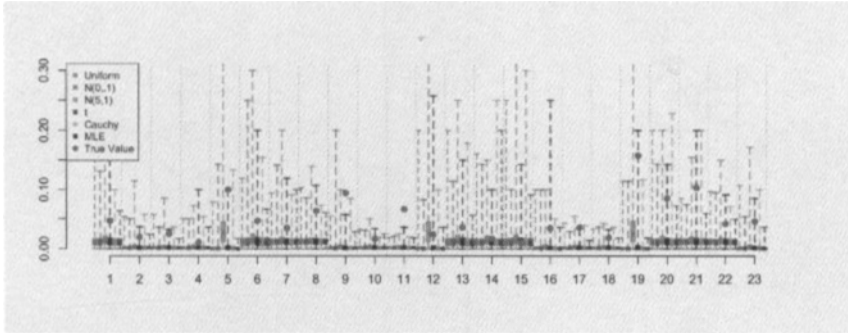
Figure 1.5: One-step prediction of graph density on the Beach data for all five Bayesian DNR models and MLE.
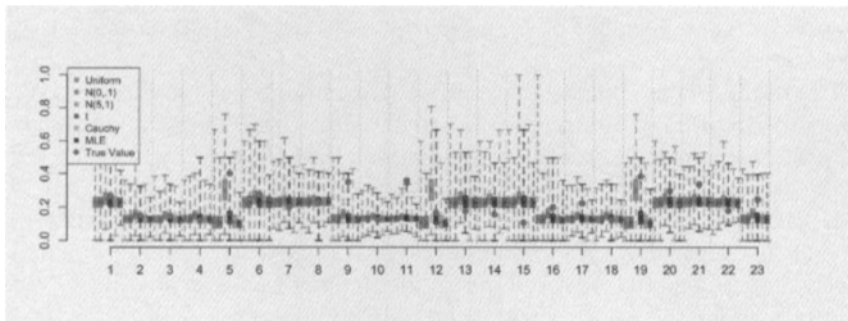

Figure 1.6: One-step prediction of graph 3-cliques on the Beach data for all five Bayesian DNR models and MLE.

1.5   Discussion

In our test cases for Bayesian DNR with and without vertex dynamics, we see that several standard minimally informative priors produce both inferential and predictive results that are comparable to each other and—in the case of inference—to maximum likelihood estimation. Differences between point estimates for the default priors tested here are typically smaller (often by an order of magnitude) than the statistical uncertainty associated with the estimates themselves, and any are hence likely to lead to very similar interpretation in practice. Since these results were achieved with two very different types of dynamic social networks of relatively moderate size, we are led to the preliminary conclusion that selection of default priors for DNR families is unlikely to strongly impact results in typical settings (and the choice of prior can thus be made on computational or other grounds).

It is noteworthy that there is a general advantage for the use of Bayesian posterior predictive distributions versus predictive distributions based on the MLE, in terms of reduced overconfidence. This is a general and well-known phenomenon, and not particular to network models; however, it is perhaps worth reinforcing the point that predictions should incorporate uncertainty regarding parameters, and the Bayesian approach greatly facilitates this practice.

While we saw that inference in our test cases was generally fairly robust to prior specification, it is of course possible to "break" a data set by employing a sufficiently informative prior. Here, we observed that use of an "extreme" prior biased in a single direction strongly affected some (but not all) parameters, and that the direction of bias in the posterior did not necessarily correspond to the direction of bias in the prior (due to correlations among the associated statistics). As would be expected, parameters that are more poorly estimated are more subject to influence from the prior, and these (along with others closely related to them via the likelihood) are the ones that are most vulnerable to poor prior specifications. As in other contexts, our results suggest that use of strongly informative priors in DNR must be undertaken with caution, and in particular with an awareness of the degree to which parameters are entangled via the likelihood. Robustness tests are strongly recommended.

The strong concordance between posteriors under various choices of prior with each other and the sampling distribution of the MLE strongly suggests that our test cases place us in the asymptotic Gaussian regime, an encouraging development given that our networks are of relatively modest size. That said, it is important to bear in mind the large number of degrees of freedom inherent in dynamic network data. In general, data size for such problems grows as $\mathcal{O}(N^2 T)$, where N is the vertex set size and T is the number of time points. For graphs of even moderate size—and even for small numbers of time points—this can easily result in an extremely large number of edge variables. Of course, the asymptotic limit depends on more than simply the number

of degrees of freedom (the sparsity of the data is also important), but this heuristic provides a reasonable intuition for why the Gaussian approximation is likely to work well here. This is a property that is potentially exploitable for e.g. approximate Bayesian computation in large N,T settings.

1.6   Conclusion

In this chapter, we have reviewed the problem of modeling static and dynamic networks in exponential family form. As we have shown, Bayesian analysis of both problems in the general case is made difficult via the central role of the incomputable ERGM normalizing factor, which enters into the likelihood (and sometimes the prior) in a manner that makes traditional MCMC-based sampling schemes slow and/or impractical. For some model families, however, this problem does not apply (due to the presence of a tractable normalizing factor); such families are especially useful in the case of dynamic network modeling, where conditioning on the past can in some cases allow us to model edges as conditionally independent in the present. For such cases, the temporal ERGM form reduces to a simple product of inhomogeneous Bernoulli graph likelihoods, dubbed "dynamic network regression" because of the similarity of the resulting model to logistic regression on time series data. Because of the simplicity of this family, and its similarity to logistic regression, it would be desirable to be able to employ standard "default" priors for analysis in routine settings. Our experiments with two different data sets suggest that this is a reasonable approach: alternative default priors lead to very similar conclusions, with all being similar to inferences resulting from maximum likelihood estimation. One reason for this concordance is that even fairly modest dynamic network data sets supply enough data degrees of freedom to—for DNR families—place the posterior within the asymptotic Gaussian regime. While it is possible to obtain poor results by selection of an especially inappropriate prior, reasonable choices thus lead to reasonable outcomes. Given this, and given the advantages of the Bayesian approach for problems such as prediction, there seems little reason not to recommend this as a standard technique for analyzing network dynamics with TERGM DNR families.

Bibliography

Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing (2008, June). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research 9*, 1981–2014.

Almquist, Z. W. and C. T. Butts (2012). Evolving context: Evidence from temporal change in organizational collaboration over the course of the 2005 katrina disaster. Working paper, University of California, Irvine.

Almquist, Z. W. and C. T. Butts (2013). Dynamic network logistic regression: A logistic choice analysis of inter- and intra-group blog citation dynamics in the 2004 US presidential election. *Political Analysis 21*(4), 430–448.

Almquist, Z. W. and C. T. Butts (2014). Logistic network regression for scalable analysis of networks with joint edge/vertex dynamics. *Sociological Methodology* (forthcoming).

Anderson, B. S., C. Butts, and K. Carley (1999). The interaction of size and density with graph-level indices. *Social Networks 21*(3), 239–267.

Anderson, C. J., S. Wasserman, and B. Crouch (1999). A p* primer: Logit models for social networks. *Social Networks 21*(1), 37–66.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. New York: John Wiley and Sons.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 113–147.

Besag, J. (1974). Spatial interactions and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological) 36*(2), 192–236.

Besag, J. (2001). Markov chain Monte Carlo for statistical inference. *Center for Statistics and the Social Sciences Working Papers*(9).

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*. Hayward, CA: Institute of Mathematical Statistics.

Butts, C. T. (2007). Permutation models for relational data. *Sociological Methodology 37*(1), 257–281.

Butts, C. T. (2008a). A relational event framework for social action. *Sciological Methodology 38*(1), 155–200.

Butts, C. T. (2008b). Social networks: A methodological introduction. *Asian Journal of Social Psychology 11*(1), 13–41.

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science 325*, 414–416.

Butts, C. T. (2011). Bernoulli graph bounds for general random graphs. *Sociological Methodology 41*, 299–345.

Butts, C. T. and B. R. Cross (2009). Change and external events in computer-mediated citation networks: English language weblogs and the 2004 U.S. electoral cycle. *The Journal of Social Structure 10*(3), 1–29.

Caimo, A. and N. Friel (2011). Bayesian inference for exponential random graph models. *Social Networks 33*(1), 41–55.

Chatterjee, S. and P. Diaconis (2011). Estimating and understanding exponential random graph models. arXiv:1102.2650.

Cornwell, B. (2009). Good health and the bridging of structural holes. *Social Networks 31*(1), 92–103.

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

Desmarais, B. and S. Cranmer (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications 391*(4), 1865–1876.

Desmarais, B. A. and S. J. Cranmer (2011). Forecasting the locational dynamics of transnational terrorism: A network analytic approach. In *Proceedings of the European Intelligence and Security Informatics Conference (EISIC)*. IEEE Computer Society.

DuBois, C., C. T. Butts, D. McFarland, and P. Smyth (2013). Hierarchical models for relational event sequences. *Journal of Mathematical Psychology, 57*(6), 297–309.

DuBois, C., C. T. Butts, and P. Smyth (2013). Stochastic blockmodeling of relational event dynamics. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, (AISTATS).

Foulds, J. R., C. DuBois, A. U. Asuncion, C. T. Butts, and P. Smyth (2011). A dynamic relational infinite feature model for longitudinal social networks. *Journal of Machine Learning Research - Proceedings Track 15*, 287–295.

Frank, O. and D. Strauss (1986). Markov graphs. *Journal of the American Statistical Association 81*(395), 832–842.

Freeman, L. C., S. C. Freeman, and A. G. Michaelson (1988). On human social intelligence. *Journal of Social Biological Structure 11*, 415–425.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis* (2nd ed.). New York, NY: Chapman & Hall/CRC.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 1360–1383.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. Federal Reserve Bank of Minneapolis, Research Department.

Geyer, C. J. and E. A. Thompson (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B(Methodological) 54*(3), 657–699.

Goodreau, S. M., J. A. Kitts, and M. Morris (2009). Birds of a feather, or friend of a friend?: Using exponential random graph models to investigate adolescent social networks. *Demography 46*(1), 103–125.

Guo, F., S. Hanneke, W. Fu, and E. P. Xing (2007). Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 321–328. ACM.

Handcock, M. S. (2003). Statistical models for social networks: Inference and degeneracy. In R. Breiger, K. M. Carley, and P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis*, pp. 229–240. Washington, DC: National Academies Press.

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris (2003). `statnet`: A suite of R packages for the statistical modeling of social networks. Electronic.

Handcock, M. S., A. E. Raftery, and J. M. Tantrum (2007). Model based clustering for social networks. *Journal of the Royal Statistical Society, Series A 170*, 301–354.

Hanneke, S., W. Fu, and E. P. Xing (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics 4*, 585–605.

Hanneke, S. and E. P. Xing (2007a). Discrete temporal models of social networks. In *Statistical Network Analysis: Models, Issues, and New Directions*, Volume 4503 of *Lecture Notes in Computer Science*, pp. 115–125. Berlin, Germany: Springer.

Hanneke, S. and E. P. Xing (2007b). Statistical network analysis: Models, issues, and new directions: ICML 2006 workshop on statistical network analysis, Pittsburgh, PA, USA, June 29, 2006, revised selected papers. *TBA 4503*, 115–125.

Hartigan, J. (1964). Invariant prior distributions. *The Annals of Mathematical Statistics 35*(2), 836–845.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association 97*(460), 1090–1098.

Holland, P. W. and S. Leinhardt (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association 76*(373), 33–50.

Hummon, N. and P. Doreian (2003). Some dynamics of social balance processes: bringing Heider back into balance theory. *Social Networks 25*(1), 17–49.

Hunter, D. R. and M. S. Handcock (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics 15*, 565–583.

Hunter, D. R., M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software 24*(3).

Hunter, D. R., P. N. Krivitsky, and M. Schweinberger (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics 21*(4), 856–882.

Jeffreys, H. (1998). *The Theory of Probability.* Oxford University Press.

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association 91*(435), 1343–1370.

Koehly, L. M. and P. Pattison (2005). Random graph models for social networks: Multiple relations or multiple raters. In P. J. Carrington, J. Scott, and S. Wasserman (Eds.), *Models and Methods in Social Network Analysis,* Chapter 9, pp. 162–191. Cambridge: Cambridge University Press.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electronic Journal of Statistics 6*, 1100–1128.

Krivitsky, P. N. and M. S. Handcock (2010). A separable model for dynamic networks. Technical Report arXiv:1011.1937, arXive e-print.

Krivitsky, P. N., M. S. Handcock, and M. Morris (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology 8*, 319–339.

Martin, A. D., K. M. Quinn, and J. H. Park (2011). MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software 42*(9), 1–21.

McCullagh, P. and J. A. Nelder (1999). *Generalized Linear Models* (2nd ed.). New York: Chapman & Hall/CRC.

Murray, I., Z. Ghahramani, and D. MacKay (2012). MCMC for doubly-intractable distributions. arXiv preprint, arXiv:1206.6848.

Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96*, 1077–1087.

Pattison, P. E. and G. L. Robins (2002). Neighborhood-based models for social networks. *Sociological Methodology 32*, 301–337.

Pattison, P. E. and S. Wasserman (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology 52*, 169–193.

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March,* pp. 20–22.

Raftery, A. and S. Lewis (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science 7*, 493–497.

Robins, G. and P. Pattison (2001). Random graph models for temporal processes in social networks. *Journal of Mathematical Sociology 25*(1), 5–41.

Sarkar, P. and A. W. Moore (2005, December). Dynamic social network analysis using latent space models. *SIGKDD Explor. Newsl. 7*(2), 31–40.

Sarkar, P., S. M. Siddiqi, and G. J. Gordon (2007). A latent space approach to dynamic embedding of co-occurrence data. In M. Meila and X. Shen (Eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*.

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association 106*(496), 1361–1370.

Shalizi, C. R. and A. Rinaldo (2013). Consistency under sampling of exponential random graph models. *The Annals of Statistics 41*(2), 508–535.

Snijders, T. A. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology 21*(1–2), 149–172.

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. In M. Sobel and M. B. Boston (Eds.), *Sociological Methodology*, pp. 361–395. London: Basil Blackwell.

Snijders, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure 3*(2).

Snijders, T. A. (2005). Models for longitudinal network data. In P. Carrington, J. Scott, and S. Wasserman (Eds.), *Models and Methods in Social Network Analysis*, Chapter 11. New York: Cambridge University Press.

Snijders, T. A., C. E. Steglich, M. Schweinberger, and M. Huisman (2007). *Manual for SIENA Version 3.1*. University of Groningen: ICS Department of Sociology; University of Oxford: Department of Statistics.

Snijders, T. A. and M. Van Duijn (1997). Simulation for statistical inference in dynamic network models. In R. Conte, R. Hegselmann, and P. Terna (Eds.), *Simulating Social Phenomena*, pp. 493–512. Berlin: Springer.

Snijders, T. A. B. (2010). Conditional marginalization for exponential random graph models. *Journal of Mathematical Sociology 34*, 239–252.

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

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn (2003). Winbugs user manual. *Cambridge: MRC Biostatistics Unit.*

Spiegelhalter, D. J. and A. F. Smith (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 377–387.

Strauss, D. (1986). On a general class of models for interaction. *SIAM Review 28*(4), 513–527.

Strauss, D. and M. Ikeda (1990). Psuedolikelihood estimation for social networks. *Journal of the American Statistical Association 85*(409), 204–212.

van Duijn, M. A. J., K. Gile, and M. S. Handcock (2009). Comparison of maximum pseudo likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks 31*(1), 52–62.

van Duijn, M. A. J., T. A. B. Snijders, and B. H. Zijlstra (2004). p_2: A random effects model with covariates for directed graphs. *Statistica Neerlandica 58*, 234–254.

Wang, P., G. Robins, and P. Pattison (2009, September). Pnet: Program for the simulation and estimation of exponential random graph (p*) models. http://www.sna.unimelb.edu.au/pnet/pnet.html.

Wasserman, S. and K. Faust (1994). *Social Network Analysis: Methods and Applications.* New York: Cambridge University Press.

Wasserman, S. and P. Pattison (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika 61*(3), 401–425.

Wong, G. Y. (1987). Bayesian models for directed graphs. *Journal of the American Statistical Association 82*(397), 140–148.

Yang, R. and J. O. Berger (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 1195–1211.

Zeggelink, E. P., F. N. Stokman, and G. G. van der Bunt (1996). The emergence of groups in the evolution of friendship networks. *Journal of Mathematical Sociology 21*(1-2), 29–55.