

Analysing Exponential Random Graph (p-star) Models with Missing Data Using Bayesian Data Augmentation*

Johan Koskinen †, Garry Robins, and Philippa Pattison
MelNet Social Networks Laboratory Technical Report 08–04
Department of Psychology, School of Behavioural Science
University of Melbourne, Parkville Victoria 3010, Australia

November 28, 2008

*Previous versions were presented at the Sunbelt XXVII, International Sunbelt Social Network Conference, 8th, on Corfu Island, Greece; and at The 8th Asia-Pacific Complex Systems Conference, July 2-5, 2007, Surfers Paradise, Gold Coast, QLD, Australia; the authors have benefitted from numerous comments and suggestions received at these occasions.

†The work of Johan Koskinen was funded by the Defence Science and Technology Organisation.
E-mail: johank@unimelb.edu.au.

Abstract

Missing data, such as non-response, is often problematic in social network analysis since what is missing may potentially alter the conclusions about what we have observed. This in the sense that individual tie-variables typically need to be interpreted in relation to their local neighbourhood and the global structure. Some ad-hoc methods for dealing with missing data in social networks have been proposed but here we consider a model-based approach. We discuss various aspects of fitting exponential family random graph (or p-star) models (ERGMs) to networks with missing data and present a Bayesian data augmentation algorithm for the purpose of estimation. This involves drawing from the fully conditional posterior distribution of the parameters, something which is made possible by a recently developed algorithm. With ERGMs already having complicated interdependencies, we argue that it is particularly important to provide inference that adequately describes the uncertainty, something that the Bayesian approach caters for. To the extent that we wish to explore the missing parts of the network, the posterior predictive distributions, immediately available at the termination of the algorithm, are at our disposal, which allows us to explore the distribution of what is missing unconditionally on any particular parameter values. Some important features of treating missing data and of the implementation of the algorithm are illustrated using a well known collaboration network.

Keywords: Missing data; Data augmentation; ERGM (p-star); Auxiliary variable; Social network analysis.

1 Introduction

When studying social structure in the social network paradigm the social interaction is typically interpreted with respect to a network defined by a fixed set of actors (Wasserman and Faust, 1994). Thus the ontological issues of social interaction are intimately related to the definition of the boundary of the network (Laumann et al.,

1983). While determining a meaningful definition of what constitutes the network (not only the set of actors but also relations, etc) is ultimately a theoretical question, the general issue of boundary definition also has practical consequences for empirical investigation. As pointed out by Stork and Richards (1992), ignoring missing data (more particularly in their case non-respondents) amounts to redefining the boundary. The theoretical and practical implications of this are further discussed in Kossinets (2006) and Burt (1987).

Thus far relatively scant attention has been paid to the treatment of missing data. Stork and Richards (1992) discusses and reviews the problem and proposes an ad-hoc scheme for complementing data in the face of non-respondents. This has the advantage that the network may be analysed on its intended scale but this advantage comes with several disadvantages. Kossinets (2006) give an extensive review of the topic of missing data and investigates the performance of some common graph indices under different missing data scenarios. Other accounts of the effect on indices of missing data are given in Costenbader and Valente (2003) and in Huisman (2007). Huisman (2007) translates several imputation strategies that are common in the statistical literature to the context of social network data but it appears that the answer as to which imputation scheme is most effective is inconclusive. Hence, while researchers commonly acknowledge the dangers of ignoring missing data, practical advice appears difficult to give in the general case since while features of the graph are influenced by the unknown values, it is not clear how the knowledge about the observed features of the graph may inform us about the unknown values. This is the case, at least, if we do not have a model for how the observed relates to the unobserved available.

The family of exponential random graph (or p-star) models (ERGM) has been proposed as a family of models that is able to model social network data, with the type of complex interdependencies this typically entails (Holland and Leinhardt, 1981; Fienberg and Wasserman, 1981; Frank and Strauss, 1986; Wasserman & Pattison, 1996; Pattison & Wasserman, 1999; Robins, Pattison, and Wasserman, 1999; Snijders et al., 2006; Hunter and Handcock, 2006). Representing statistical models for networks,

the ERGMs have the potential for allowing us to use the model and observed data to inform us about what is missing. ERGMs in the context of missing data have earlier been treated by Robins, Pattison, and Woolcock (2004), whereby non-respondents were treated as a particular category of actors. Their approach consequently allowed estimation of the model only under very specific assumptions. Recently Handcock and Gile (2007) (see also Gile and Handcock, 2006) outlined an approach to do likelihood-based inference for ERGMs with missing data under the assumption of data missing at random. A particular merit of that paper is that it highlights the connections between ERGMs with missing data and inference for various sampling designs, the latter a field in which considerable amounts of statistical research has been made (see e.g. Thompson and Frank, 2000).

Since it may be hard to unambiguously define what constitutes an observation, our interpretation of what we have observed may be radically changed in the light of knowing what is missing. There are analogues in non-network data, e.g. in event history analysis missing time-varying covariates might mean that some observed values on the dependent variable are undefined (Ghilagaber and Koskinen, 2008). Here this problem is further confounded by the fact that due to the dynamic nature of data, the response variable also acts as a covariate.

Here we present a Bayesian inference scheme for fitting ERGMs to data that have missing observations. We argue that a Bayesian approach is better suited than a non-Bayesian approach on principled grounds since uncertainty is more clearly defined and handled in the former: we want to take the added uncertainty due to unknown data into account when interpreting the model in terms of its parameters and the predictive, goodness of fit, distributions. Additionally, though not explicitly treated here, a Bayesian approach allows for incorporating prior information, something that may prove essential when data is scarce but decisions have to be made. An example of a case that warrants the use of informative prior distributions is given in Koskinen, Robins, and Pattison (2008).

The rest of the paper is structured as follows. In the next Section the model is

defined along with some necessary definitions and notation for missing data. This is followed by Section 3 in which the inference scheme is presented with explanations of the principal inferential goals. Next, in Section 4 some analytical bounds on the degree to which data may be degraded are presented. A well-known data set is used to illustrate various aspects of fitting ERGMs to data with missing observations. An approximate inference scheme is presented in Section 6 followed by a concluding discussion.

2 The model and missing data structure

Throughout we assume that we are interested in modelling graphs or digraphs on a fixed set $V = \{1, \dots, n\}$ of vertices. To this end we define the model for binary adjacency matrices x that take values in \mathcal{X} . The proposed methods do not vary greatly for valued graphs, multigraphs, bipartite graphs or other graph objects for which ERGMs have been defined but for the purpose of making the presentation clear the approach is presented assuming an undirected, univariate network.

Let $\mathcal{N} = \binom{V}{2}$ denote the index set of the non-redundant elements of $x = (x_e : e \in \mathcal{N})$, so that for the stochastic edge set $E \subseteq \mathcal{N}$, $x_e = 1\{e \in E\}$ is the usual tie variable, and $\mathcal{X} = \{0, 1\}^{\mathcal{N}}$. When clear from the context x denotes the conventional square, symmetric adjacency matrix. Define the probability mass function (pmf) by

$$p(x|\theta) = \exp\{q(x; \theta) - \psi(\theta)\},$$

where q is a real-valued function of x , a $p \times 1$ vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^p$, and possibly some fixed set of covariates, and where $\psi(\theta) = \log \sum_{x \in \mathcal{X}} \exp\{q(x; \theta)\}$, is a normalising constant ensuring that p sums to one over \mathcal{X} . For exponential family random graph distributions $q(x; \theta)$ may be expressed in terms of $\theta^T z(x)$ the inner product of a natural parameter and a $p \times 1$ vector valued function $z(x)$. For curved exponential family random graph model (Hunter, 2007; Hunter and Handcock, 2006), there exists a function $\eta : \theta \rightarrow \mathbb{R}^{p'}$, for $p' > p$, such that $q(x; \theta) = \eta(\theta)^T z(x)$, for some

function z of the adjacency matrix x and covariates.

There are many forms that missingness can take in social network analysis and while most real applications are likely to have elements of several forms we focus here on the case that satisfies the set of assumptions that Koskinen, Robins, and Pattison (2008) refer to as “the standard model”. Firstly, we assume the particular case where we unambiguously may determine whether a tie-variable x_e is observed or not. Note that while this covers the case of non-respondents for directed networks, it might not automatically cover the case of non-respondents in non-directed networks or other instances where the observations are laden with measurement error (such as in the case of cognitive social structures, Krackhardt, 1987). The standard model is further described in Koskinen, Robins, and Pattison (2008) in the context of common social network issues.

Let $\xi = (\xi_e : e \in \mathcal{N})$ be a collection of indicator variables such that ξ_e is equal to one if the tie-variable x_e is observed, and zero otherwise. For a given ξ , we shall write $x = (u, v)$ using the notational convention that u denotes the collection of tie-variables that are observed and v denotes the collection of tie-variables than are unobserved. Sometimes it shall be convenient to refer to the missing elements using the index set $M = M(\xi) = \{e \in \mathcal{N} : \xi_e = 1\}$, and the complement $M^c = \mathcal{N} \setminus M$, and the conditional range-space $\mathcal{X}_M(u) = \{x \in \mathcal{X} : x_e = u_e, e \in M\}$. Central to the proposed methodology is that the pmf of ξ has the property

$$p(\xi|u, v) = p(\xi|u),$$

where we have suppressed the notational dependency of the pmf on some parameter distinct from θ . That is, tie-variables are missing at random (MAR) in the sense of Rubin (1976) and consequently, with prior distribution $\pi(\theta)$, for the posterior we

have

$$\begin{aligned}
\pi(\theta|u, \xi) &\propto p(u, \xi|\theta)\pi(\theta) \\
&= \pi(\theta) \sum_v p(u, v, \xi|\theta) \\
&= \pi(\theta) \sum_v p(\xi|\theta, u, v)p(u, v|\theta) \\
&= \pi(\theta)p(\xi|u) \sum_v p(u, v|\theta) \\
&\propto \pi(\theta) \sum_v p(u, v|\theta).
\end{aligned}$$

Henceforth summation over v is taken to mean summation over elements in $\mathcal{X}_M(u)$. Consequently, if tie-variables are missing at random then we do not have to take into account what cause observations to be missing when doing inference for the parameters. The missing data mechanism can to all intents and purposes be ignored (hence the term “ignorable”; Handcock and Gile, 2007; note that “ignorable” in addition to MAR requires “distinct” parameters, Rubin, 1976:587). Thompson and Frank (2000) showed that most link-tracing designs are missing at random, and Handcock and Gile (2007) demonstrates this for the particular case of ERGMs, as well as give several examples of sampling designs that satisfy this criterion.

A simple form of missing data mechanism is when individual tie variables are missing independently with a probability that does not depend on their value. That individual tie-variables are missing is what Huisman (2007) called item non-response. If the missing data is accumulated within actors, assuming all ties pertaining to an actor to be missing, which Huisman (2007) calls unit non-response, roughly translates to respondent non-response (for self-reported ties), this may not in itself violate MAR. For directed networks where actors report on their out-going ties, $\xi_{ij} = 0$ if i is a non-respondent and 1, otherwise, which under suitable assumptions is ignorable (for technical definitions see Handcock and Gile, 2007). In Huisman’s (2007) terminology, unit and item non-response both encompass the case when information may be missing for exogenous covariates as well as tie-variables, which seems a realistic

scenario, but which is left out of the present presentation.

In a snowball sampling design (see Thompson and Frank, 2000), what is observed depends on an initial sampling mechanism as well as the realised graph, x . Despite this, as Handcock and Gile (2007) observe, the sampling mechanism is ignorable. If the initial sample is denoted by $s_0 \subset V$, and iteratively, a node $j \in s_k$, where s_k is wave k , if there exists an $i \in \cup_{r=0}^{k-1} s_r \setminus \{j\}$, $x_{ij} = 1$. For two-waves, the inclusion variable ξ_{ij} thus indicates whether $i \in s_0 \cup s_1$ or $j \in s_0 \cup s_1$.

It is tempting to perform inference only with what you have observed. The function $L(\theta; u) = \sum_v p(u, v | \theta)$ is called the face likelihood and is a function of θ given u that does not depend on any particular v . To say that ignorability implies that we may ignore the sampling mechanism does, while here allowing us to ignore any particular v , not translate into saying that we may ignore the fact that data is missing. The case when you ignore the fact that data is missing we refer to as available case analysis (cf Little and Rubin, 1987), and corresponds to assuming the likelihood to be $p(u | \theta) = \exp\{q(u; \theta) - \psi(\theta)\}$. While available case analysis may lead to serious bias in non-network data (Little and Rubin, 1987), Koskinen, Robins, and Pattison (2008) discuss how this may be especially detrimental in social network analysis. In particular one can use the graphical modelling framework (Lauritzen, 1996) to show that removing tie-variables means adding ties between tie-variables in the dependence graph, from which the non-zero interactions (Frank and Strauss, 1986) are derived. Removing missing data thus corresponds to adding non-zero interactions, increasing the complexity of the configurations we need to include in $z(x)$. Thus a subgraph of an ERGM graph does not follow ERGM (Koskinen, Robins, and Pattison, 2008).

3 The algorithm

The standard approach of data augmentation (Tanner and Wong, 1987) as applied to the present missing data set-up, consists of constructing a Markov chain Monte Carlo (MCMC) sampler (Gilks, Richardson and Spiegelhalter, 1996) that produces

a sequence $(\theta^{(g)}, v^{(g)})$ of variates from the joint posterior distribution $\pi(\theta, v|x)$, by alternating between draws from the fully conditional posteriors $\pi(\theta|u, v)$ and $\pi(v|u, \theta)$. The sequences $(\theta^{(g)})$ and $(v^{(g)})$ considered separately may be considered samples from the marginal posterior distributions $\pi(\theta|u)$ and $\pi(v|\theta)$.

As a simple example, consider the homogeneous Bernoulli model, with pmf $p(x|\theta) = \exp\{\theta L(x) - \psi(\theta)\}$, where $L(x) = \sum_{e \in \mathcal{N}} x_e$ is the number of edges, and assume an improper constant prior distribution $\pi(\theta)$. Now, clearly the fully conditional posterior of θ given the rest is proportional to $p(u, v|\theta)$ and for each missing element v_e , the probability the edge is present is $(1 + e^{-\theta})^{-1}$. Perhaps more clear, is to consider the reparametrisation in terms of the probability $\mu = (1 + e^{-\theta})^{-1}$. The constant prior on θ corresponds loosely to the Haldane prior (Zellner, 1996) $\pi(\mu) \propto 1/(\mu(1 - \mu))$ which is functionally identical to a beta(0, 0). The algorithm thus re-expressed, would circle through the steps

Algorithm 1 For $g = 1, \dots$ iterate

- (a) Draw μ from $\text{beta}(u_1 + \dots + u_{n_u} + v_1 + \dots + v_{n_v}, |\mathcal{N}| - (u_1 + \dots + u_{n_u} + v_1 + \dots + v_{n_v}))$
- (b.1) Draw v_1 from a Bernoulli distribution with probability of success μ
- ⋮
- (b.k) Draw v_k from a Bernoulli distribution with probability of success μ
- ⋮
- (b. n_v) Draw v_{n_v} from a Bernoulli distribution with probability of success μ

In other words, given the current number of successes, draw a likely value of μ , then from each of the missing tie-variables flip a μ -coin to determine whether the missing tie should be set to one or zero. Naturally, for this trivial example we may obtain the distributions exactly as $\pi(\theta|u) = B(L(u), n_u - L(u))^{-1} (e^{-\theta} + 1)^{-n_u} e^{-\theta(n_u - L(u))}$,

$B(a, b) = (a - 1)!(b - 1)!/(a + b - 1)!$, and $\pi(v|u) = L(u)/n_u$, for $L(u) = \sum_{e \in M} x_e$ and $n_u = |M|$.

The homogeneous Bernoulli model is an inadequate model for social networks in most instances and more sophisticated dependence assumptions are required. Implementing the data augmentation algorithm for general forms of ERGMs is however not straightforward with the updating step that corresponds to (a) above being the main hurdle. The next section gives a brief description of how to up-date θ , perform step (a), in the general case using a form of the auxiliary variable MCMC (Møller et al., 2006) called the linked importance sampler auxiliary Metropolis-Hastings algorithm (LISA) (Koskinen, 2008).

3.1 Drawing from the fully conditional posterior of the parameters

When drawing from the posterior $\pi(\theta|x)$, one may use the fact that most MCMC samplers only require that the posterior is known up to a constant of proportionality. The only part of $\pi(\theta|x)$ we need to be able to evaluate is $p(x|\theta)\pi(\theta)$. In the case of ERGMs however, the likelihood $p(x|\theta) = \exp\{q(x; \theta) - \psi(\theta)\}$ contains the function $c(\theta) = \exp \psi(\theta)$ which is analytically intractable. The linked importance sampler auxiliary Metropolis-Hastings (LISA) solves the problem by defining a Markov chain $(\theta^{(g)}, \omega^{(g)})$, on an extended state space $\omega \in \Omega \subseteq \prod_{j=1}^m \mathcal{X}^K \times \{1, \dots, K\} \times \{1, \dots, K\}$. Defining a distribution $P_{\zeta, \theta}^B(\omega) = Q_{\zeta, \theta}^B(\omega)/c(\zeta)$ on Ω , with the property that it has a reverse distribution $P_{\theta, \zeta}^F(\omega) = Q_{\theta, \zeta}^F(\omega)/c(\theta)$, we may draw from

$$\pi(\omega, \theta|x) = P_{\zeta, \theta}^B(\omega)\pi(\theta|x) \propto Q_{\zeta, \theta}^B(\omega) \frac{e^{q(x; \theta)}}{c(\theta)} \pi(\theta),$$

using Metropolis-Hastings for a particular choice of proposal distributions. For proposal distributions, let a candidate θ^* be proposed from a multivariate normal distribution $N(\theta^{(g)}, \Psi)$, centred over the previous state, and conditional on the proposed θ^* ,

propose ω^* from the distribution $P_{\theta^*, \zeta}^F(\omega^*)$. The Hastings acceptance ratio becomes

$$\begin{aligned} & \frac{\pi(\omega^*, \theta^* | x) P_{\theta, \zeta}^F(\omega) \varphi(\theta^{(g)}; \theta^*, \Psi) \pi(\theta^*)}{\pi(\omega, \theta | x) P_{\theta^*, \zeta}^F(\omega^*) \varphi(\theta^*; \theta^{(g)}, \Psi) \pi(\theta)} \\ = & \frac{e^{q(x; \theta^*)/c(\theta^*)} P_{\zeta, \theta^*}^B(\omega^*) P_{\theta, \zeta}^F(\omega) \varphi(\theta^{(g)}; \theta^*, \Psi) \pi(\theta^*)}{e^{q(x; \theta)/c(\theta)} P_{\zeta, \theta}^B(\omega) P_{\theta^*, \zeta}^F(\omega^*) \varphi(\theta^*; \theta^{(g)}, \Psi) \pi(\theta)}. \end{aligned}$$

Upon writing out the distributions $P_{\theta, \zeta}^F$ and $P_{\zeta, \theta}^B$, and rearranging, the acceptance ratio reduces to

$$\frac{e^{q(x; \theta^*)} Q_{\zeta, \theta^*}^B(\omega^*) Q_{\theta, \zeta}^F(\omega) \varphi(\theta^{(g)}; \theta^*, \Psi) \pi(\theta^*)}{e^{q(x; \theta)} Q_{\theta^*, \zeta}^F(\omega^*) Q_{\zeta, \theta}^B(\omega) \varphi(\theta^*; \theta^{(g)}, \Psi) \pi(\theta)},$$

which only contains analytically tractable expressions. The details of how to construct $P_{\theta, \zeta}^F$, and $P_{\zeta, \theta}^B$, are described in Koskinen (2008), but the algorithm has an equivalent interpretation in terms of the linked importance sampler (LIS) of Neal (2005). The LIS is an importance sampler, where $\lambda(\theta, \zeta; \omega)$ is an estimator of $\lambda(\theta, \zeta) = c(\zeta)/c(\theta)$, in the sense that $E_{P_{\theta, \zeta}^F}(\lambda(\theta, \zeta; \omega)) = \lambda(\theta, \zeta)$. Hence, running LISA amounts to proposing a move to θ^* , estimating $\lambda(\theta^*, \zeta)$ using LIS, and accepting the move with an acceptance ratio that is equal to a ratio of normal densities times $\pi(\theta^* | x)/\pi(\theta^{(g)} | x)$ bar for $\lambda(\theta^*, \theta^{(g)})$ that is substituted by $\lambda(\theta^*, \zeta; \omega^*)/\lambda(\theta^{(g)}, \zeta; \omega)$. The denominator in the latter expression is the estimate of $\lambda(\theta^{(g)}, \zeta)$, from when $\theta^{(g)}$ was accepted for the first time. The accuracy of $\lambda(\theta, \zeta; \omega)$ and consequently the mixing of the MCMC, depends on and can be manipulated by adjusting the dimensions of the state space Ω , through K and m . It remains to explain the role and properties of the constant ζ .

3.1.1 The choice of ζ and Ψ

The role of ζ is essentially to define a reference distribution against which the pmf defined by θ is evaluated. The simplest form of the auxiliary variable algorithm of Møller et al. (2006) may be considered the special case of LISA when $K = m = 1$. In this case $\Omega = \mathcal{X}$, $P_{\zeta, \theta}^B(\omega) = p(\omega | \zeta)$, and $P_{\theta, \zeta}^F(\omega) = p(\omega | \theta)$, with the interpretation that $\exp\{q(\omega; \zeta) - q(\omega; \theta)\}$ is used to estimate $\lambda(\theta, \zeta)$ using (one) ω drawn from the importance distribution $p(\omega | \theta)$. This illustrates why the choice of a “good” ζ is

essential for the performance of LISA (for an example of adverse effects of less good choices see further Koskinen, 2008; Møller et al., 2006, discuss the choice of ζ for the auxiliary variable MCMC at length and the accuracy of LIS as a function of the difference between ζ and θ is discussed in Neal, 2005).

With constant prior distributions a good choice of ζ is one that is equal to the modal point in the posterior distribution, i.e. the MLE. Assuming this to hold true also in the case of missing data, where $x = (u, v)$ will vary between iterations, ζ should be set to the maximiser of $\ell(\theta; u) = \log L(\theta; u)$. The gradient of the posterior is with a constant prior the differentiated log likelihood,

$$\frac{\partial}{\partial \theta} \ell(\theta; u) = \frac{\frac{\partial}{\partial \theta} \sum_v p(u, v | \theta)}{\sum_v p(u, v | \theta)} - \frac{\partial}{\partial \theta} \psi(\theta)$$

which we may refer to as the score function, $S(\theta; u)$. In the general case when $q(x; \theta) = \eta(\theta)^T z(x)$,

$$S(\theta; u) = \dot{\eta}(\theta)^T [E_{\eta(\theta)}(z(u, v) | u) - E_{\eta(\theta)}(z(x))],$$

where $\dot{\eta}(\theta_0) = ((\partial/\partial \theta)\eta(\theta))^T|_{\theta=\theta_0}$, and $E_{\eta}(\cdot | u)$ is the conditional expectancy on $\mathcal{X}_M(u)$. The Hessian equals

$$\begin{aligned} H(\theta) &= \frac{\partial}{\partial \theta} S(\theta; u)^T \\ &= [E_{\eta(\theta)}(z(u, v) | u) - E_{\eta(\theta)}(z(x))]^T \ddot{\eta}(\theta) - \dot{\eta}(\theta) \{I(\eta)\} \dot{\eta}(\theta)^T, \end{aligned}$$

where $\ddot{\eta}(\theta_0) = ((\partial^2/[\partial \theta \partial \theta^T])\eta(\theta))^T|_{\theta=\theta_0}$, and

$$-I(\eta) = \frac{\partial}{\partial \eta} [E_{\eta}(z(u, v) | u) - E_{\eta}(z(x))] \Big|_{\theta=\theta_0} = Cov_{\eta}(z(u, v) | u) - Cov_{\eta}(z(x)).$$

For the regular exponential family case $H(\theta)$ simplifies to $-I(\theta)$, and $S(\theta; u) = [E_{\theta}(z(u, v) | u) - E_{\theta}(z(x))]$. In this case the MLE may be found using the Robbins-Monro procedure (Snijders, 2007) or by approximating the expectancies by their MCMC analogues as in Handcock and Gile's (2007) adoption of the approach of Geyer and Thompson (1992). For the curved exponential form one may not make

use of the standard likelihood equation to find the MLE but an approximate Fisher scoring algorithm may still be implemented for the case of no missing data as shown in Hunter and Handcock (2006). Similarly in the case when we have missing data, we may find the modal point in the posterior distribution by iteratively updating according to $\zeta^{(t)} = \zeta^{(t-1)} - J(\zeta^{(t-1)})^{-1}S(\zeta^{(t-1)}; u)$, where $J(\zeta) = \dot{\eta}(\zeta) \{I(\eta)\} \dot{\eta}(\zeta)^T$, and where the expectancies are approximated using importance samples. We have found that performance of this approximate Fisher scoring heavily depends on the choice of importance sample and in order for this algorithm to work a number of approximate Robins-Monro steps are required to find a suitable initial value, $\zeta^{(0)}$. It is important to note however that for the purposes of implementing LISA we do not require the exact solution but only a ζ which is good enough. In practice, however, using the missing data analogues of the convergence criteria of Snijders (2002) is more straightforward than assessing whether ζ is “close enough”.

In the cases where an available case MLE may be obtained, it would be quicker to set ζ to the available case MLE. In the light of Section 2.1. the available case MLE does appear to be a poor choice and indeed, the missing data MLE and the available case MLE may in some instances be quite different (Koskinen, Pattison, Robins, and Wang, 2008).

In addition to the choice of ζ , the mixing of the chain $(\theta^{(g)}, \omega^{(g)})$ depends on the choice of proposal distribution in standard fashion (see e.g. Chib and Greenberg, 1995). Settling on the choice of $N(\theta^{(g-1)}, \Psi)$ for θ^* in iteration g , the efficiency of the proposal density is guided by the choice of Ψ . Tierney (1994), and Roberts, Gelman and Giles (1997) show that $\Psi = c/\sqrt{1+p}\Sigma$ is an efficient choice, where $c > 0$ is a constant and Σ is the variance covariance matrix of the target distribution. For LISA with missing data we use the inverse of the Fisher information matrix as an approximation of Σ and allow c to be rather lower than in the ideal case considered in Tierney (1994). The reason for the latter is the added uncertainty introduced by the need to also draw ω and v .

When approximating the expected value $E_{\eta}(z(x))$ and $Cov_{\eta}(z(x))$, we may use

the standard Metropolis-Hastings approach (Corander, Dahmström, and Dahmström, 1998; Snijders, 2002; Handcock, 2003). Approximating the conditional expectancies $E_\eta(z(u, v)|u)$ and $Cov_\eta(z(u, v)|u)$ needs only the minor modification of only proposing to update elements v . The Metropolis-Hasting algorithm is more closely described in the next section.

3.2 Drawing from the fully conditional distribution of the missing tie variables

The ease with which Metropolis-Hasting allows us to simulate from $p(\theta|x)$ is arguably what makes this algorithm attractive. Snijders (2002) proposes alternative MCMC schemes for drawing x but the computational costs involved have to be balanced against the simplicity (and therefore speed and small memory requirements) of the Metropolis-Hastings algorithm. An updating step is constructed as follows. Assuming we wish to draw $x \in \mathcal{X}_A(u)$, for $A \subseteq \mathcal{N}$, and consequently only the elements in $A^c = \mathcal{N} \setminus A$ are the one allowed to change. Writing $\Delta_e x$ for the matrix that is equal to x for all elements but has element e set to $1 - x_e$, we may propose to change the current state x to $\Delta_e x$, where e is chosen uniformly at random from among the elements of A^c , and accept this proposed move with probability the minimum of one and

$$\frac{p(\Delta_e x|\theta)}{p(x|\theta)} = \exp\{q(\Delta_e x; \theta) - q(x; \theta)\}.$$

If the move is not accepted, the chain stays in its current state. A commonly used fact is that for ERGMs $q(\Delta_e x; \theta) - q(x; \theta)$ may be written in terms of the so called change statistics, $\theta^T \delta_e z(x) = \theta^T (z(\Delta_e x) - z(x))$, something which leads to considerable computational advantages (Frank and Strauss, 1987; Corander, Dahmström, and Dahmström, 1998; Snijders, 2002; Handcock, 2003; Snijders et al., 2006). Note that for the curved exponential family for fixed θ , $\eta(\theta)$ and z still constitute a non-curved ERGM, wherefore the ratio of pmf's may still be written in terms of the change statistics $\eta(\theta)^T \delta z(x) = \eta(\theta)^T (z(\Delta_e x) - z(x))$ (Hunter, 2007; Hunter and Handcock,

2006).

For the purposes of drawing v conditional on u and a parameter vector θ , it is more convenient to update the elements of \mathcal{M} systematically than chosen at random. In the case of missing data, the chain $(\theta^{(g)}, \omega^{(g)})$ is thus augmented by, for each $e_1, \dots, e_{n_v} \in \mathcal{M}$, corresponding to the tie variables v_1, \dots, v_{n_v} , performing the updating step: set $v_k^{(g)} = 1 - v_k^{(g-1)}$ with probability $\min(1, H)$ for

$$H = \frac{p(1 - v_k^{(g-1)} | \theta^{(g)}, \omega^{(g)}, u, v_1^{(g)}, \dots, v_{k-1}^{(g)}, v_{k+1}^{(g-1)}, \dots, v_{n_v}^{(g-1)})}{p(v_k^{(g-1)} | \theta^{(g)}, \omega^{(g)}, u, v_1^{(g)}, \dots, v_{k-1}^{(g)}, v_{k+1}^{(g-1)}, \dots, v_{n_v}^{(g-1)})}.$$

Denoting the current state by x , the proposed move is to $\Delta_{e_k} x$, the log acceptance ratio expressed in terms of change statistics is $\log H = \eta(\theta)^T \delta_{e_k} z(x)$.

3.2.1 Exploring the unobserved structure and link prediction

On the one hand the sample $(v^{(g)})$ may be considered a mere by-product of the algorithm for producing $(\theta^{(g)}, \omega^{(g)})$. Considering the properties of the distribution of $(v^{(g)})$ highlights some of its potential uses and benefits over and above what would be available had the goal merely been to estimate the MLE. From the properties of the MCMC algorithm it follows that marginally, $(v^{(g)})$ is a sample from the distribution

$$\pi(v|u) = \int_{\Theta} \pi(\theta, v|u) d\theta = \int_{\Theta} p(v|\theta, u) \pi(\theta|u) d\theta$$

(a sample in the conventional MCMC sense with the usual caveats regarding burnin, thinning, etc, see e.g. Gilks, Richardson, and Spiegelhalter, 1996). Integrating out the parameter we effectively have a probability distribution for what is missing, v , given only what we have observed, u . Let us make two observations regarding this. Assume that we were ultimately interested in studying some features that required information on the entire graph, either because these were graph level measures or simply because the measured properties would change with different configurations of v . If v is missing we cannot calculate these measures (or they might be highly misleading) with anything less than filling in v . Since, and this is the second point,

we have an entire distribution of v 's at our disposal we may nonetheless obtain the probabilities of different values of the measures given what we have observed.

The use of simulation to understand model behaviour (Handcock, 2003) and making sense of the model has been a frequently employed tool (Robins, Pattison, Woolcock, 2005). Lately simulation from a fitted model has become a commonly employed diagnostics tool for assessing to what extent a model is able to capture important features of the data that were not explicitly modelled (Hunter, Goodreau and Handcock, 2008). The fact that a portion of data is missing does not prevent us from simulating from the model for any given θ but it does mean that we do not have any data to compare with, we only have u . For assessing whether a fitted model is plausible, we need a principled procedure for filling in the elements of v .

More formally, let $T(x)$ be some function of graphs $x \in \mathcal{X}$, then the pmf of T given that we have only observed u is given by $\pi(t|u) = \Pr(T(u, v) = t|u) = \sum_{v: T(u, v)=t} \pi(v|u)$. From $(\theta^{(g)}, \omega^{(g)}, v^{(g)})$, we may obtain a sample from the distribution $\pi(t|u)$, simply as $t^{(g)} = T(u, v^{(g)})$. We shall see some examples of quantities whose distribution may be obtained in this fashion later on. The MCMC estimator of $\Pr(T(u, v) \in A|u)$ is straightforward to obtain as the relative frequency $\frac{1}{G} \sum 1\{T(u, v^{(g)}) \in A\}$. When feasible it is however more efficient to use the fact that some of the fully conditional distributions are available to us in closed form. As an example we could approximate the predictive probability that a tie is present given the data we have observed as $\hat{\pi}(v_k|u) = \frac{1}{G} \sum v_k^{(g)}$. More efficient is however to estimate this quantity by

$$\hat{\pi}(v_k|u) = \frac{1}{G} \sum \Pr(v_k = 1 | \theta^{(g)}, \omega^{(g)}, u, v_1^{(g)}, \dots, v_{k-1}^{(g)}, v_{k+1}^{(g-1)}, \dots, v_{n_v}^{(g-1)}),$$

a procedure that may be referred to as Rao-Blackwellisation (Gelfand and Smith, 1990; Tanner and Wong, 1987).

4 Limits on degradation

It seems unlikely that general results may be obtained for the effect on the posterior distributions of missing data. When it comes to the question of how much data may be removed while still retaining estimability, the answer in part depends on what data are removed and what we mean by estimability. For estimability we may distinguish between the question of whether the observed data provides practically useful information about the model and whether the posterior distribution exists at all. If a proper prior distribution is used for θ then the posterior will always exist but in extreme cases this posterior may be completely determined by the prior distribution if the data does not add any information.

How do we know that the posterior even exists when we use a constant, improper prior distribution? Some insight into this is given by Proposition 5, below. First we present a Lemma needed for Proposition 5, as well as a couple of useful corollaries. Let us first assume that we have a model $p(x|\theta) = \exp\{\theta^T z(x) - \psi(\theta)\}$, the image of \mathcal{X} under z is denoted by \mathcal{Z} , and C is the relative interior of the convex hull on \mathcal{Z} . Recall that the posterior with improper prior is proper if $\exp\{\theta^T z(x) - \psi(\theta)\}$ is integrable.

Lemma 2 *For an observation x , the posterior is proper if and only if $z(x) \in C$*

Proof. A proof follows directly from the proof of Theorem 1 in Diaconis and Ylvisaker (1979) upon observing that $\exp\{\theta^T z(x) - \psi(\theta)\}$ is functionally equivalent to their conjugate prior for the special case $n_0 = 1$. ■

From the known properties of exponential family distributions we have that $z(x) \in C$ is a necessary and sufficient condition for the maximum likelihood estimate to exist and consequently we may state that the posterior exists for improper prior if and only if the MLE exists. All the results regarding the existence of posteriors presented here may therefore be equivalently expressed in terms of existence of MLEs. In particular Hancock (2003) elaborates on a non-Bayesian version of Lemma

1 that draws on general results for exponential family distributions (Barndorff-Nielsen, 1978).

From Lemma 1 we may derive the next Corollary, but first we need some additional notation and another Lemma. Let $z_I(x) = (z_i(x))_{i \in I}$ for some index set I of distinct functions on \mathcal{X} and let \mathcal{Z}_I and C_I denote the image under z_I and the relative interior of the convex hull of this image respectively. Let the model associated with each I be defined by $p_{\theta, I}(x) = \exp(\theta_I^T z(x) - \psi_I(\theta_I))$ for $z \in \mathcal{Z}_I$, $x \in \mathcal{X}$.

Lemma 3 *For a fixed I , if for an observation $x \in \mathcal{X}$ there exists $a \subseteq I$ such that $z_a(x) \notin C_a$, then $z_b(x) \notin C_b$ for all $b \supseteq I \cap a$.*

Proof. Assume $z_a(x) \notin C_a$ and let b be the disjoint union in I of a and a singleton set $\{k\}$ such that $z_b(x) = (z_a(x)^T, z_k(x)^T)^T$. If $z_b(x) \in C_b$, then we could find $x_0, x_1 \in \mathcal{X}$ and $\alpha \in (0, 1]$, so that $z_b(x_0) \in C_b$, $z_b(x_1) \in \text{cl}(C_b)$ and $z_b(x) = \alpha z_b(x_0) + (1 - \alpha) z_b(x_1)$. If this were true however, $z_a(x) = \alpha z_a(x_0) + (1 - \alpha) z_a(x_1)$ which contradicts the assumption $z_a(x) \notin C_a$. Having established that the proposition holds for singleton set $\{k\}$ we may proceed by induction to show that $z_b(x) \notin C_b$ for all $b \supseteq I \cap a$ and indeed that $z_I(x) \notin C_I$. ■

Even though this result is self evident from the definition of the convex hull, it is nonetheless important since it implies that we can never fit a model if it has an inferentially degenerate model (Handcock, 2003) nested in it and that we can never alleviate a degeneracy issue for a model by merely including extra parameters. The practical implications of this may be that if we find one sub model that is degenerate then we need not fit the model in which it is nested and, it does not matter in what order we fit the parameters of nested models (that is to say, the order is not important but combinations are). This however holds true for the case when model fitting difficulties stem from degeneracy but not necessarily numerical difficulties, something which goes for all the results stated in terms of degeneracy - even when the posterior is finite and the MLE exists we may not be able to fit a model in practice because of for example “multi-modality” of the distribution of statistics (for an in-

depth analysis see Handcock, 2003). The analytical implications are summed up in the following Corollary.

Corollary 4 *If for an observation $x \in \mathcal{X}$ and a model $p_{\theta, I}$, there exists $a \subseteq I$ such that the posterior under the model $p_{\theta, a}$ is improper, then the posterior under the model $p_{\theta, I}$ is improper.*

Proof. As $z_a(x) \notin C_a$ is a necessary and sufficient condition for the posterior under $p_{\theta, a}$ to exist, we must have that $z_a(x) \notin C_a$ and thus by Lemma 2 $z_I(x) \notin C_I$ and consequently the posterior under model $p_{\theta, I}$ is improper ■

Now we may state the main proposition

Proposition 5 *For observed data u , with missing data assumptions as above, the posterior distribution exists if and only if it would exist for every potential realisation $x = (u, v)$.*

Proof. By definition, the posterior proportional to the face likelihood is integrable if

$$\int \sum_v p(u, v|\theta) d\theta = \sum_v \int p(u, v|\theta) d\theta < \infty.$$

From the positivity of the pmf and that the sum is finite for the posterior given u to exist we must have $\int p(u, v|\theta) d\theta < \infty$ for all v , in other words that the posteriors given all the potential observations $x = (u, v)$ must exist. ■

Naturally this proposition may be equivalently stated in terms of the relative interiors of the convex hull, i.e. the posterior given u exists iff $z(u, v) \in C$ for all v . In other words, if there is even one “extreme” graph among the possible graphs that we could have then the posterior does not exist (nor does the MLE).

For curved exponential family models it might be hard to devise similar results. For curved exponential models where the model defined by $\exp\{\theta_{-p}^T z(x; \theta_p) - \psi_{\theta_p}(\theta_{-p})\}$, where $\theta_{-p} = (\theta_1, \dots, \theta_{p-1})^T$, for θ_p considered a fixed, known constant, defines an ERGM, we may draw on the above results. Let \mathcal{Z}_{θ_p} be the image of \mathcal{X}

under $z(\cdot; \theta_p)$, and C_{θ_p} denote the relative interior of the convex hull on \mathcal{Z}_{θ_p} . We say that these models have a *simple ERGM projection*, and for these models we may state the following Corollary.

Corollary 6 *The posterior $\pi(\theta|x)$ given x for the model curved exponential family random graph $p(x|\theta) = \exp\{\eta(\theta)^\top z(x) - \psi(\theta)\}$ with a simple ERGM projection is improper if there exists a non-null set B such that $z(x; \theta_p) \notin C_{\theta_p}$ for all $\theta_p \in B$*

Proof. The normalising constant of $\pi(\theta|x)$ is

$$\begin{aligned} & \int_{\Theta} \exp\{\eta(\theta)^\top z(x) - \psi(\theta)\} d\theta \\ &= \int_{\Theta_p} \left[\int_{\theta_{-p} \in \Theta_{-p}:\theta_p} \exp\{\eta(\theta)^\top z(x) - \psi(\theta)\} d\theta_{-p} \right] d\theta_p \\ &\geq \int_B \left[\int_{\theta_{-p} \in \Theta_{-p}:\theta_p} \exp\{\eta(\theta)^\top z(x) - \psi(\theta)\} d\theta_{-p} \right] d\theta_p \\ &\geq \int_B \left[\min_{\theta_p \in B} g(\theta_p) \right] d\theta_p \end{aligned}$$

where $g(\theta_p) = \int_{\theta_{-p} \in \Theta_{-p}:\theta_p} \exp\{\eta(\theta)^\top z(x; \theta_p) - \psi_{\theta_p}(\theta_{-p})\} d\theta_{-p}$, is a function of θ_p which is not finite since the integrand is not integrable anywhere on B by Lemma 1 since $z(x; \theta_p) \notin C_{\theta_p}$, by assumption. ■

For models that include several “new specifications” statistics (Snijders et al., 2006), note that for each such statistic there exists a simple ERGM projection. If the posterior is undefined for any of these simple ERGM projections in the sense of Corollary 6, then by Corollary 3 the curved exponential family model containing all of the new specifications statistics is inferentially degenerate in the sense that the posterior is undefined.

Combining Proposition 5 and Corollary 6 the following Corollary follows.

Corollary 7 *For observed data u , with missing data assumptions as above, and a curved exponential family random graph model with simple ERGM projection, if there*

exists a $v = v_0$, θ_p , and θ_p^* , $\theta_p < \theta_p^*$, such that $z((u, v_o); \alpha) \notin C_{\theta_p}$, $\alpha \in (\theta_p, \theta_p^*)$, then the posterior is improper.

4.1 The edges and alternating k -triangle model

What the analytical results imply depend crucially on whether the model allows us to ascertain whether the posterior exists according to Proposition 5. The degeneracy problems are known to be more of a problem for homogeneous Markov models than for the so called new specifications ERGMs (Snijders et al., 2006; for an in-depth treatment of degeneracy for ERGMs, see Handcock, 2003). Surprisingly, for the model that contains only the number of edges and the alternating k -triangles statistic as statistics, establishing whether the posterior exists according to Proposition 5 is simple. We offer the following without a detailed proof.

Proposition 8 *For observed data u , and a model with $z(x) = (L(x), AKT(x; \lambda))^T$, and $L(u) + n_v \leq \lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil$, the posterior exists if the subgraph induced by the edges of u has at least one triangle.*

Proof. By Proposition 5 we need to show that $z(u, v) \in C$ for all v . If the subgraph induced by the edges of u has at least one triangle, $x = (u, v)$ must have at least one triangle for all v while $L(u, v) \leq \lfloor \frac{n}{2} \rfloor \lceil \frac{n}{2} \rceil$, upon which the proof of Proposition 2 in Koskinen, Robins, and Pattison (2008) follows. ■

The remarkable thing is that the model is inferentially non-degenerate if the observation has any triangles as long as it does not have all triangles! This is a clear departure from some Markov models, say, e.g. the edges and two-path model, which has plenty of graphs on the relative boundary. Additionally, this is valid for all values of λ and thus, by Corollary 10, the posteriors may exist even for the curved exponential family model.

5 Examples

The application of the data augmentation scheme is illustrated and some of the above issues highlighted using the collaboration network of Lazega’s (2001) $n = 36$ law firm partners and a model specification that frequently has been used for this data to illustrate various aspects of ERGMs (Snijders et al., 2006; Hunter and Handcock, 2006; Handcock and Gile, 2007; Hunter, 2007; van Duijn et al., 2007). More specifically, this model has the statistics: edges $z_1(x) = \sum_{i < j} x_{ij}$; main effect of seniority $z_2(x) = \sum_{i < j} x_{ij}(\text{SEN}_i + \text{SEN}_j)$, SEN_i reversed rank order of seniority divided by n ; main effect of practice $z_3(x) = \sum_{i < j} x_{ij}(\text{PRA}_i + \text{PRA}_j)$, PRA_i equal to one or zero according to whether partner i practices corporate law or not; homophily of practice $z_4(x) = \sum_{i < j} x_{ij} \mathbf{1}\{\text{PRA}_i = \text{PRA}_j\}$; homophily of sex $z_5(x) = \sum_{i < j} x_{ij} \mathbf{1}\{\text{MAN}_i = \text{MAN}_j\}$; homophily of office $z_6(x) = \sum_{i < j} x_{ij} \mathbf{1}\{\text{OFF}_i = \text{OFF}_j\}$, for OFF_i being a categorical variable denoting whether i works in the Boston, Hartford or Providence office.

In addition we assume the curved exponential form of Hunter and Handcock (2006) whose simple ERGM projection has $z(x; \theta_{p-1}) = (z_1(x), \dots, z_6(x), \text{AKT}(x; e^{\theta_p}))^T$ as its vector of statistics, i.e. dyadic covariates one through six, and the alternating k -triangle statistic (Snijders et al., 2006), or geometrically shared partners (Hunter and Handcock, 2006; Hunter, 2007). For the function $\eta(\theta)$ and associated vector of statistics we refer to Hunter and Handcock (2006) and Hunter (2007). The sociogram of the network is provided in Figure 1, and for numerical summaries of the model fitted to the complete network see Table 4 of Koskinen (2008).

We stress that these empirical examples are used to illustrate various aspects of fitting models to data with missing observations. Recall that in the Bayesian framework frequentist properties of e.g. point estimates are largely irrelevant.

The effect on the missing data on inference may be view from the perspective of the posterior of θ and summary measures thereof, or the degree of structural information retained. This division is clearly an artifice (since all relevant information

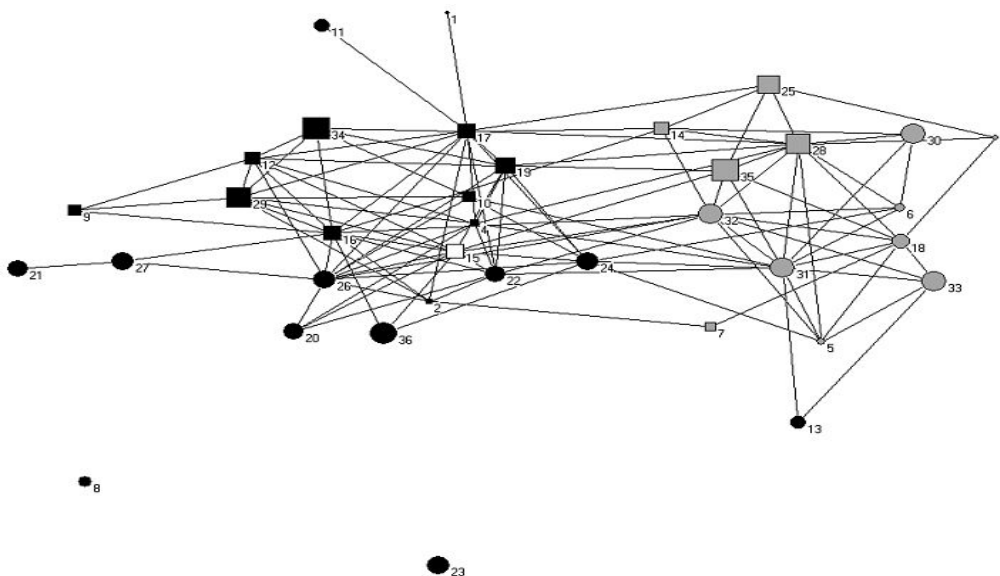


Figure 1: Collaboration network for Lazega’s (2001) 36 partners. Office Boston (black), Hartford (grey), and Providence (white); size proportional to seniority; practice corporate (ellipse) practice litigation (box). Actors 27, 29, and 34 are women, all other actors men

about the model is captured by θ) but one perspective is useful for understanding the other. Loosely speaking we may say that some regions of the graph provides more information or at least different information than others. What exactly would constitute a region of the graph may not be straightforward to define but in the particular case of non-respondents the regions may simply be couched in terms of the actors. Clearly the tie-variables of some actors are going to have a greater impact on the inference than others. Koskinen, Pattison, Robins, and Wang (2008) propose a range of influence measures to investigate the role of different actors in the network from a model-based perspective. According to these criteria removing actor 15 from the Lazega data set would result in a greater change in estimates for the model investigated here than the removal of any other actor. By the same criterion actor

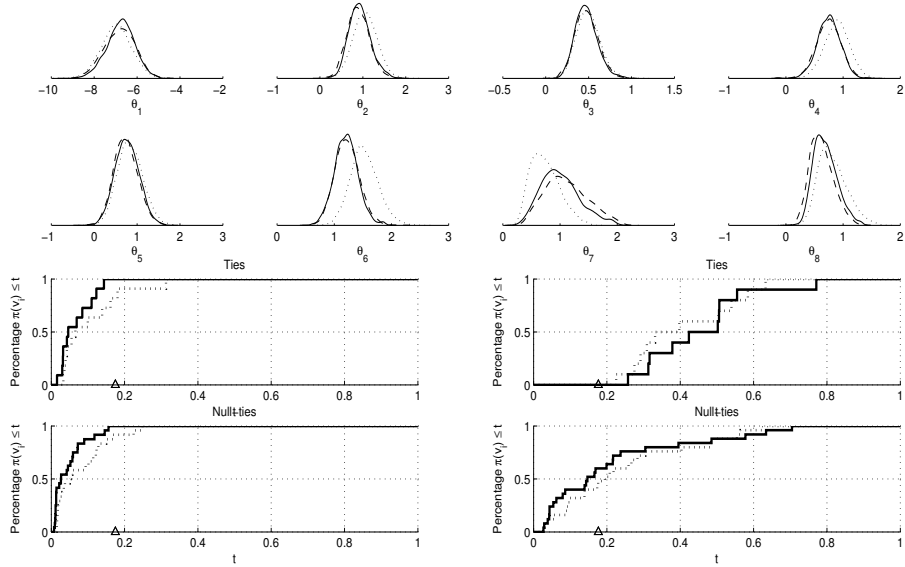


Figure 2: Posteriors for θ (top 8 panels) for complete data (solid line), actor 15 assumed non-responsive (dotted line), and actor 19 assumed non-responsive (dashed line). Posterior predictive ordered tie-probabilities grouped by known ties and null-ties for non-responsive 15 (left) and non-responsive 19 (right) for full curved (solid), dyad-independent (dotted), and Bernoulli (triangle) models.

19 would be one of the less influential actors. All estimations are based on thinned samples of a total of 100,000 iterations with $K = 7000$ and $m = 9$, and ζ and Ψ set as described in Section 3.1.1.

If we consider the case of having one non-responsive, actor 15 and 19 thus represent two extremes and the differences in posteriors are illustrated in Figure 2. For the posteriors of the parameters, removing 15 has the most visible effect on θ_6 (homophily office) and θ_7 (alternating k -triangles). The effects on the former are due to the fact that actor 15 is the only partner in the Providence office (in Figure 1 indicated by

white nodes) but still has many ties. The actor 15 also sits in a highly triangulated zone and his contribution to the alternating k -triangle statistic is high, and hence removing 15 we tend to underestimate the corresponding parameter θ_7 (see Figure 2). The removal of actor 19, on the other hand, barely changes the posteriors. with the exception of slightly higher uncertainty regarding θ_7 . Note that actor 19 has a well above average number of ties.

The differences between removing 15 and 19 in the posteriors for the θ_k 's are mirrored in the predictive probabilities for the tie-variables associated with these two actors (see lower two rows of Figure 2). The model does not manage to discriminate between the null-ties and actual ties of actor 15 - there is little difference between the two lower left panels of of Figure 2. This not merely to do with the dyadic attributes as may be seen by comparing the dyad-independent with the full model in Figure 2). Furthermore, if the model containing only the density parameter (indicated by a triangle in Figure 2) was to be used as reference for when a tie is deemed to be present or not (cut-off probabilities of .17 and .18 for non-respondents 15 and 19 respectively), all null-ties of 15 would be correctly classified but none of the present ties would be correctly classified (for a homogeneous Bernoulli model, our best guess for a missing tie-variable would always be one, if observed density is lower than .5, and zero if the observed density is greater than .5; for other models, if the posterior predictive probability that a missing tie-variable is one is greater than the observed density, then this is evidence for that likelihood of the tie being present is greater than compared to the base-line probability of a tie). The dyad independent model here does better for ties (82% misclassified) but worse for null-ties (92% correctly classified). The tie-variables of actor 19 are much easier to predict; the full model gets 60 per cent of the null-ties right but 100 per cent of the ties correct. Note here that the full model does a slightly better job of discriminating ties from null-ties than the dyad-independent model (that gives 48% null-ties correctly classified but also under predicts tie-probabilities for ties in the interval roughly between .2 and .5).

Snowball sampling offers another aspect of degradation of the network informa-

tion, in which the regions included and left out are dependent on the network structure itself. Here we consider snowball samples with a seed set of two nodes and two waves - the alters of ego and the alters of the alters (cf Figure 1 in Handcock and Gile, 2007). Snowball sampling was also used in Handcock and Giles (2007) for illustrating the effect of missing data and is convenient since the sample is deterministic given the network and the seed set.

To illustrate the results in Section 4, we may take a moment to reflect on what those result would imply for snowball samples taken on the Lazega (2001) data set. In the following sociograms of samples the actors not reported on are left out, however, since the population is known, the fact that i is an alter of a seed node means that we have observed x_{ik} for *all* $k \in V$. Naturally, if the isolate nodes of Figure 1, 8 and 23, were taken as a two-node seed set, we would not be able to fit any model at all. Figure 3 provides some more subtle examples of snowball samples that are inferentially degenerate in the sense that the posterior $\pi(\theta|u)$ does not exist for the model considered here. The sample in Figure 3 (a) with $\{3,5\}$ as seed set does not include any observed ties involving women and hence a model containing θ_5 may not be estimated (recall that actors 27, 29, and 34, are women, the rest men). The sample in Figure 3 (b), with seed set $\{29,34\}$, captures 25 out of the 36 nodes and 330 out of the 630 dyads but main practice (θ_3) and homophily practice (θ_4) cannot be fitted at the same time.

Samples with seed sets $\{3,5\}$ and $\{29,34\}$ do not thus provide enough information to fit a dyad-independent model (referred to as “separation”; Handcock, 2003; Albert and Anderson, 1984; Santner and Duffy, 1986) but also, according to Corollary 3, this means that the model that also includes θ_7 and θ_8 is inferentially degenerate.

For the sample in Figure 3 (c), with seed set $\{7,21\}$, the dyad-independent model may be fully estimated but the observed sample is bi-partite and hence according to Proposition 7 the model including alternating k -triangles is degenerate and consequently, by Corollary 6 (and the fact that Proposition 7 holds for all λ), this sample is inferentially degenerate. Had we obtained the sample in Figure 3 (d), we would

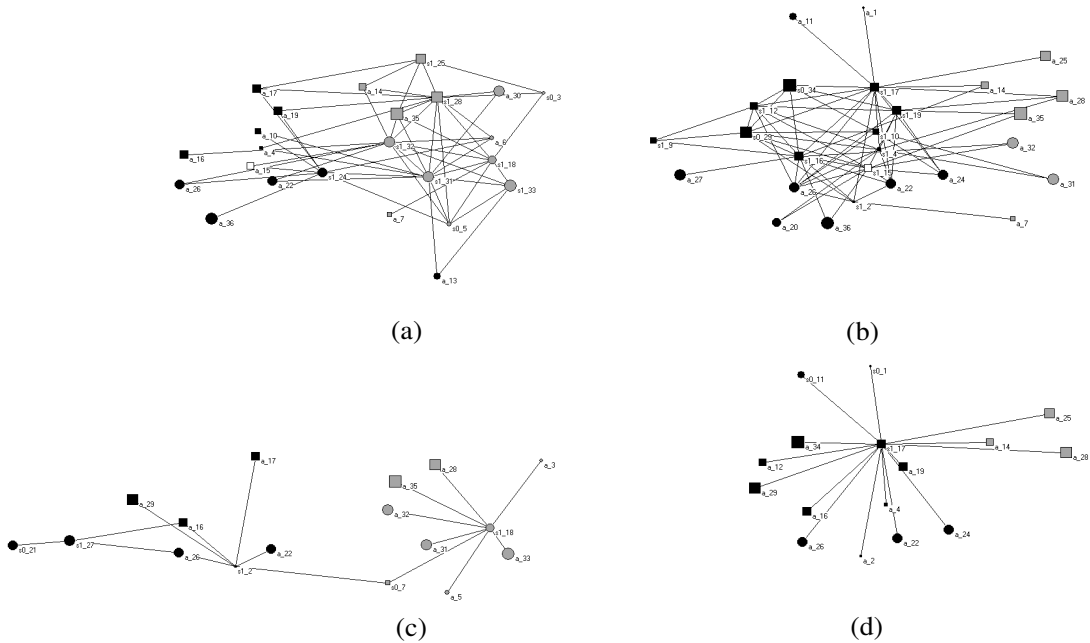


Figure 3: Inferentially degenerate snowball samples from Lazega's (2001) 36 partners with seed sets $\{3, 5\}$ (a); $\{29, 34\}$ (b); $\{7, 21\}$ (c); $\{1, 11\}$ (d);

neither be able to fit a dyad independent model, since there are no combinations of main effect practice and main practice, nor an edges and alternating k -triangle model since this sample is a star graph without triangles.

Now let us consider some examples where the missing data is given by the snowball sampling mechanism in a way that allows inference. In addition to prediction of ties, here it might be interesting to look at other more global characteristics of the graph that require increasingly detailed knowledge of the graph structure: the degrees of the nodes; betweenness; and single figure summaries of the concept of structural hole (Burt, 1992). Note that we could calculate the posterior predictive probability that the particular node occupies a structural hole, for each node, however, even with complete data "exact" structural holiness has proved too stringent a measure, mostly the degree of conformity to the theoretical ideal used. (These measures are here employed primarily to give numerical summaries of the predictive structure, if one

was truly interested in predicting what actors would be likely to occupy structural holes one would most likely have to take into consideration that these people differ from other actors and, in covert networks in particular, we would have to relax MAR assumption).

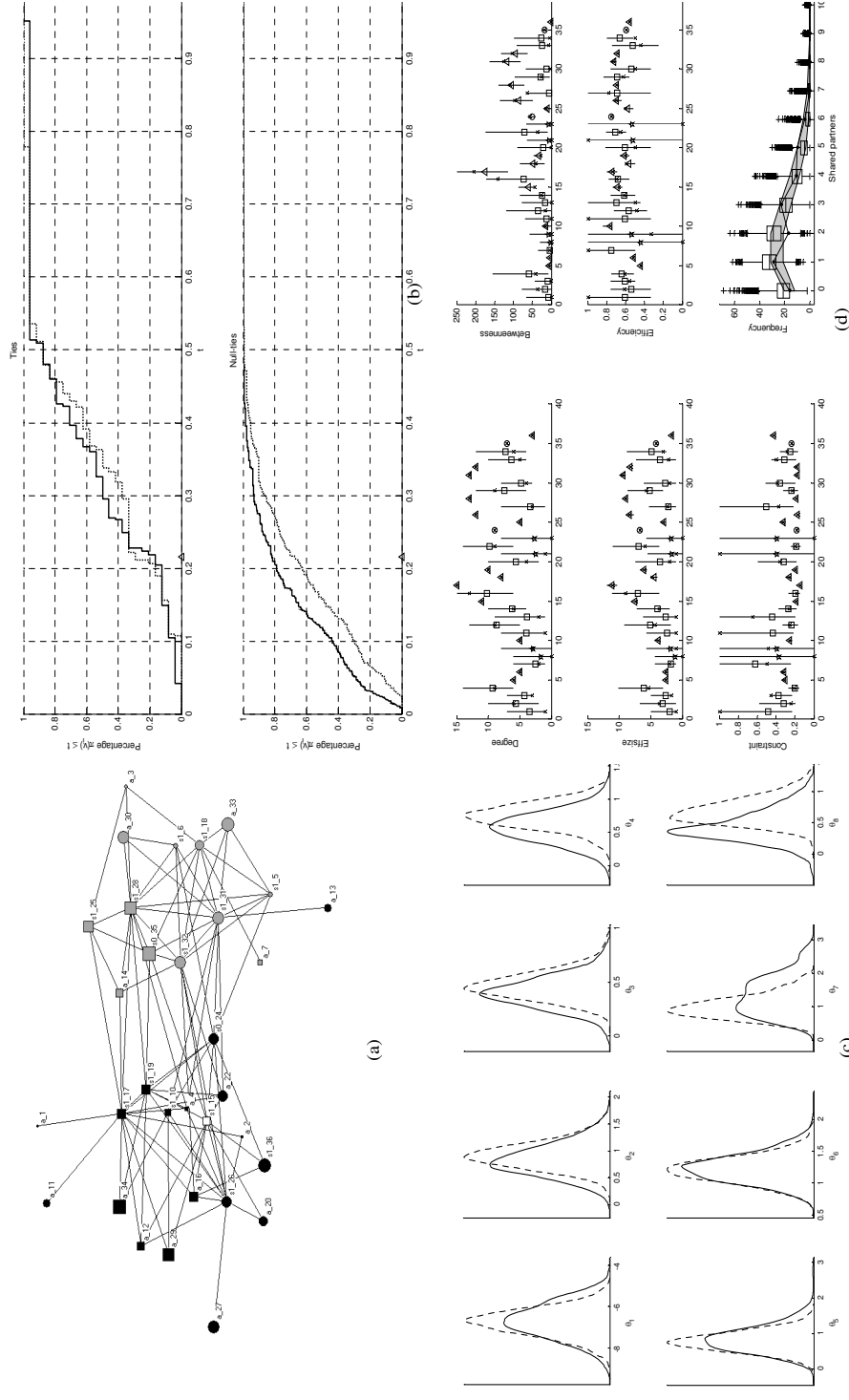


Figure 4: Snowball sample from Lazega's (2001) 36 partners with seed set $\{24, 35\}$ (a); posterior predictive marginal tie probabilities grouped by true value (b); posterior distributions (true dashed line) for parameters (c); posterior predictive nodal measures (d) (.95 intervals; means given by: true, x; initial sample, circle; wave 1, triangle; alters of wave 1 square; the rest, star) and the posterior predictive shared partner distribution (sample, boxes; complete data, shaded area; observed complete data, connected points)

The sample with seed set $\{24, 35\}$ in Figure 4 (a) covers most of the network, 32 out of the 36 vertices are reported on and 420 out of the 630 dyads are observed. The marginal posterior distributions of the θ_k 's Figure 4 (c) are very close to those for the complete data. Discriminating between ties and null-ties is relatively easy for the model. If we let the cut-off criterion for a tie be equal to the posterior predictive probability of the homogeneous Bernoulli model (triangle in Figure 4 (b)), 17% of the ties are misclassified according to the full model and 29% according to the dyad-independent model; 82% of the null-ties are correctly classified by the full model whereas only 68% of the null-ties are correctly classified by the dyad-independent model.

Since all the tie-variables associated with the seed set actors and the alters of the seed set are known, the degrees of these nodes are known (there is no uncertainty about the circles and triangles in the first panel of Figure 4 (d)). As only some of the tie-variables of the alters of the alters are observed there is some uncertainty regarding their degrees which is reflected in the prediction intervals, more specifically the intervals of the squares in the first panel of Figure 4 (d). For calculating the betweenness we need a little more information which is reflected in the fact that there is extra uncertainty as compared to the degrees in the second panel of Figure 4 (d). As an example, while we have observed that node 17 has a degree of 15, we do not know whether 17 is on a geodesic between 1 and 11, since the tie-variable $x_{1,11}$ is unobserved. On the whole, the true betweenness scores are well predicted by the model.

For the three structural hole indicators (Burt, 1992), Effsize (effective size) looks and performs like the predictions of the degrees, and Efficiency (effective size divided by degree) and Constraint perform well if we disregard the four non-sampled actors. While the predictions for the Effsize of the four missing vertices looks reasonable, the predictions for Efficiency and Constraint cover the entire range from 0 to 1. The reason for this is evident when the missing actors 9,21,8, and 23 are studied in the context of the sociogram in Figure 1 - they are all low degree or isolate fringe

nodes for which the normalisation becomes very sensitive. The posterior predictive distribution for the number of shared partners does not appear to differ much from the corresponding posterior predictive distribution for the completely observed data. The greater uncertainty associated with having missing data is reflected however in the many outliers (the corresponding whiskers and out-liers for complete data analysis are not given here) which is also mirrored in the spread of the posterior for θ_7 (the “bumps” in the distribution are artefacts of the sampling procedure).

With the seed set $\{1, 3\}$ 26 out of the 36 actors are observed and only 195 out of the 620 dyads are observed with a resulting graph whose left side is dominated by the star with actor 17 in the middle and the right side with a more triangulated region of actors of the Hartford office (Figure 5 (a)). The increased uncertainty associated with having this much missing data is reflected in the greater variation in the marginal posteriors of the θ_k 's (Figure 5 (c)). Some of the posteriors are slightly shifted. The main effect of practice is higher than for the complete data, something which is accounted for by the fact that the two hubs nodes 17 and 28 both practice litigation. Since the two seed nodes are at opposite ends of the graph the way it has its layout dominated by office, the waves just barely manage to bridge the gap between the Boston and Hartford components. Consequently the homophily effect of office location appears to be stronger than in the complete data.

Using the posterior predictive tie-probability of the homogeneous Bernoulli model as the cut-off point for determining a tie as present or absent, the full model misclassifies 35% of ties and correctly classifies 62% of the null-ties. The corresponding figures for the dyad-independent model are 32% and 59% respectively.

The different nodal measures are naturally not as good as for the seed set 24,35 but are overall not far off the mark. An example of where the model “misleads” us is the case of nodes 31 and 32. The degrees of these two nodes are somewhat underestimated as the model would fail to predict that 4 out of 12 ties for 32, and 5 out of 11 ties for 31 are to actors in another office.

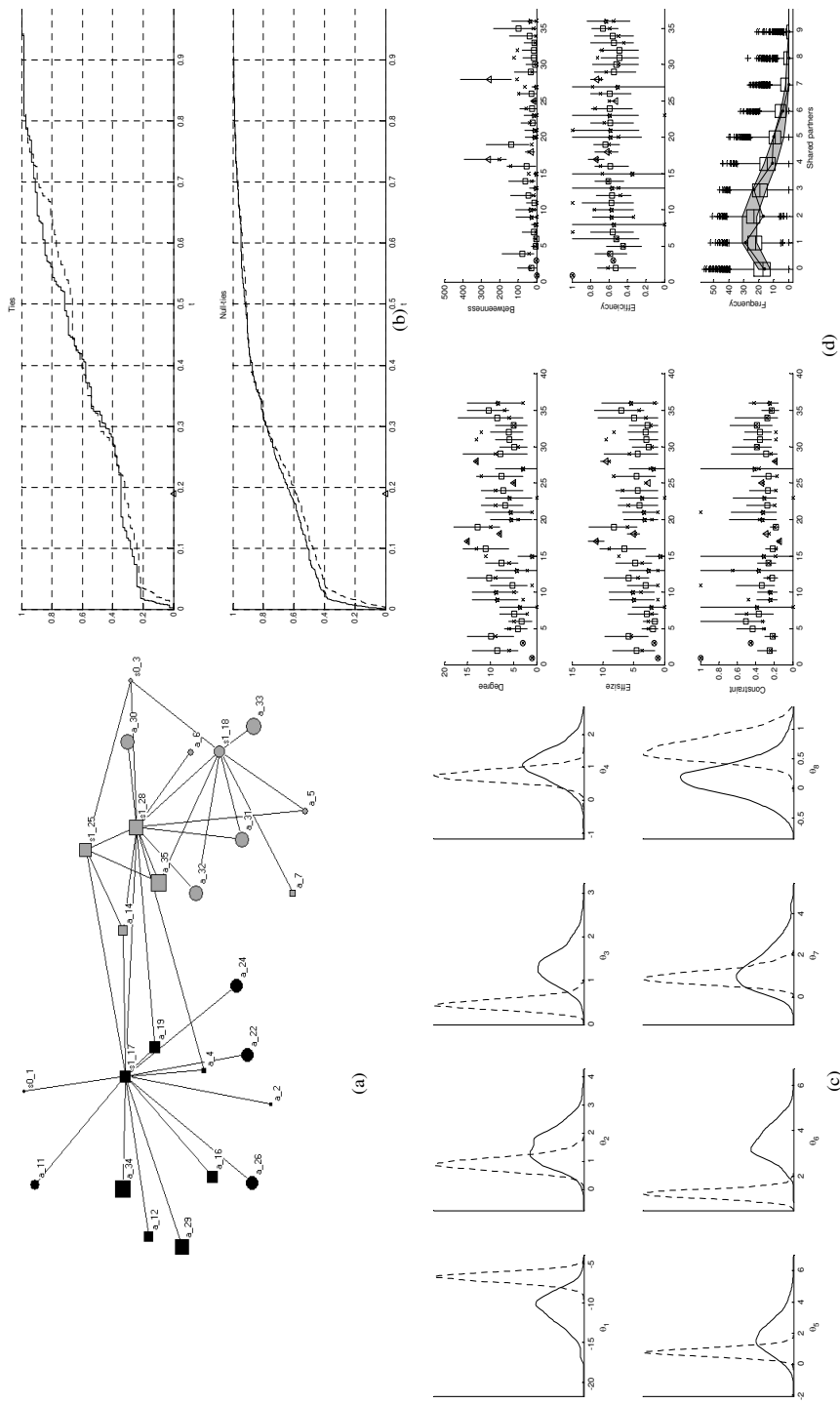


Figure 5: Snowball sample from Lazega's (2001) 36 partners with seed set $\{1, 3\}$ (a); posterior predictive marginal tie probabilities grouped by true value (b); posterior distributions (true dashed line) for parameters (c); posterior predictive nodal measures (d) (.95 intervals; means given by: true, x; initial sample, circle; wave 1, triangle; alters of wave 1 square; the rest, star) and the posterior predictive shared partner distribution (sample, boxes; complete data, shaded area; observed complete data, connected points)

The posterior predicted shared partner distribution is very close to that of the complete data model. The extra uncertainty is however reflected in the many outlier observations in the predictive distribution. This extra uncertainty in the shared partner distribution naturally must resonate in the clustering parameters θ_7 and θ_8 . As noted above, the added uncertainty makes the corresponding marginal posterior densities appear more spread out than those of the complete data analysis. Inspecting the joint posterior of these parameters in Figure 6 we see that it not only the spread but also the shape that has changed relative to the complete data analysis.

6 Approximate Bayesian inference

As the up-dating step of θ , step (a), is time consuming, it would be useful to be able to approximate this step for routine applications. Two main approaches suggest themselves: approximate the log-ratio $\psi(\theta^{(g)})/\psi(\theta^*)$ numerically using importance sampling (see Gelman and Meng, 1998, for a review); and, employ a series expansion of the posterior about its modal point. For the first approach, we could consider regenerating importance samples in each iteration or to store one sample off line. Regenerating samples on-line would not constitute a major computational advantage over LISA and since we would have no way of telling how accurate the approximations were it would be hard to assess what distribution we were actually sampling from. In storing a sample off-line, we are allowed a greater precision in that the off-line sample may be larger. The accuracy of estimates are known to be highly sensitive to the choice of the importance distribution so that again, ascertaining what distribution we were sampling from would be hard.

For constant prior, we may approximate the posterior by expanding the log likelihood around the modal point, the missing data MLE. Doing this we obtain the familiar approximation $\theta|u \sim N_p(\hat{\theta}, J(\hat{\theta})^{-1})$. For the curved form we assume that the Fisher information matrix is sufficiently close to the negative Hessian. The approximate inference scheme would hence follow roughly the same procedure as for

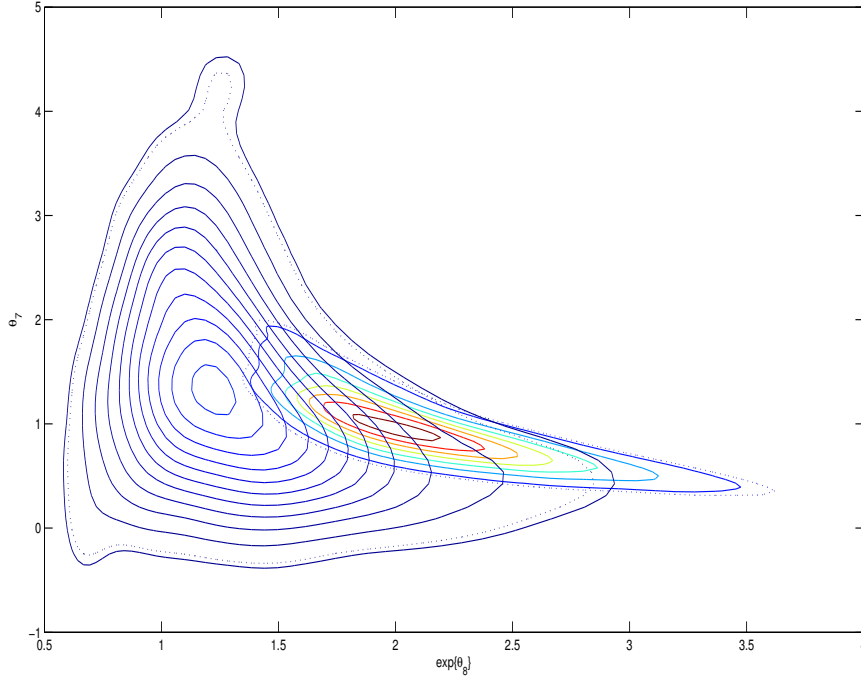


Figure 6: Bivariate posterior distribution of clustering parameters compared for the complete data analysis and missing data analysis based on snowball sampling with seed set $\{1, 2\}$. .95 HPD regions indicated by dashed line

approximate Bayesian goodness of fit (Koskinen, Wang, Robins, and, Lusher, 2008). Here, as before, G is the total sample size and T is now the number of iterations the ERGM is allowed to burn in (in the Bayesian data augmentation scheme using LISA, $T = 1$).

Algorithm 9 For positive integers G and T

Step 0 Initialize by setting $\theta^{(0)} := \hat{\theta}$

Step 1 if $g < G$, proceed to step 2, else terminate

Step 2 draw θ^* from $N_p(\hat{\theta}, J(\hat{\theta})^{-1})$, and set $\theta^{(g+1)} := \theta^*$, set $t = 0$

Step 3 if $t < T$, set $t := t + 1$, for $k \in \{1, \dots, n_v\}$

Step 3a draw a number s uniformly at random on the interval $(0, 1)$, if

$$s < \frac{p(1 - v_k | \theta^{(g+1)}, u, v_1, \dots, v_{k-1}, v_{k+1}, \dots, v_{n_v})}{p(v_k | \theta^{(g+1)}, u, v_1, \dots, v_{k-1}, v_{k+1}, \dots, v_{n_v})},$$

set $v_k := 1 - v_k$, else do nothing.

Step 3b set $v_1^{(g+1)} = v_1, \dots, v_{n_v}^{(g+1)} = v_{n_v}$

Step 4 return to Step 1.

For setting T it is convenient to use the heuristic burnin $T = \kappa \binom{n}{2}$ of Snijders (2002). If the parameters $(\theta^{(g)})$ can be ordered this could increase the autocorrelation of consecutive $v^{(g)}$ and thus reduce the T . For θ of higher dimensions that one ordering the parameters may however be difficult. As the $\theta^{(g)}$'s are independent draws, no thinning or burnin is needed with respect to G , which may be set to obtain adequate precision. This approximate algorithm is straightforward and easy to implement. However, it is hard to see how the adequacy of the approximation could be ascertained by anything other than applying an MCMC scheme as that of LISA. In particular, in some instances the approximate posterior density will lead us to produce parameter vectors that lead to unrealistic models with unwanted behaviour (see Koskinen, Wang, Robins, and, Lusher, 2008). For the example with seed set $\{1, 3\}$ considered in Section 5, the joint posterior of θ_7 and θ_8 in Figure 6 appear to differ substantially in form from normal densities.

The third step, updating the missing variables, is superfluous for the purpose of obtaining an approximation to the marginal posterior $\pi(\theta|u)$. Having a sample from the joint distribution $(\theta^{(g)}, v^{(g)})$ is however essential for performing goodness of fit and for exploring the missing data unconditional on the parameters.

7 Conclusions and future directions

We have presented and demonstrated the use of a Bayesian data augmentation scheme for fitting (curved) exponential family distribution models to social network data that has missing data. Although it is hard to characterise the nature of missing data and its effect on inference in general, we have demonstrated the consequence of missing data in a few fairly clear-cut examples.

Among the future directions that the presented results suggest are both straightforward elaborations as well as hard-to-solve issues. Among the former are the investigation of the behaviour of the approximate algorithm and the use of the posterior predictive distributions to assess model fit. An example of the latter is the incorporation of missing covariates, something that would seem natural since we cannot reasonably assume that only the response variable suffers from missing data. Some results for missing covariates are presented in Koskinen, Robins, and Pattison (2008). What role proper subjective prior distributions might play becomes a hard problem when you consider how little we know about what kind of data an ERGM might produce for different values of the parameters. Though it might not be common in scholarly contexts, there may be situations where a prior distribution may be trained on a network one has reasons to assume is similar to the one one wishes to analyse. Hence, setting a prior might be a difficult problem in the first place. Secondly, as the algorithm relies on being able to bridge the gaps between proposed values of θ and the ζ of the auxiliary distribution, the performance of the algorithm is affected by choice of prior distribution relative to the likelihood and auxiliary distribution.

The inference scheme here presented relies crucially on the MAR assumption. If for example link-prediction were the primary aim, then this approach may not be optimal. The MAR assumption may be relaxed to some extent or prior information may inform us on the likelihood of a tie-variable being missing as a function of its value and some other covariates (Huisman, 2007, defines several such “not MAR”, missing data mechanisms). In many instances however, the issue may be one of measurement error

rather than missing tie-variables (as discussed in Koskinen, Robins, and Pattison, 2008).

As we noted above, the snowball sampling design is a convenient way of investigating missing data. Here though is another case of degrees of missing data and the proposed schemes appropriateness. The approach, as we have seen, may be applied for snowball-sampled data, but computationally this amounts to fitting models to the entire population from which the sample was taken. In the empirical illustrations the population only consisted of 36 nodes but in most applications where snowball sampling is used the population is greater than this by far. Fitting ERGMs to large data sets still constitutes major computational challenges (even more so for such a computationally intensive approach as LISA) but there is also the question of whether we may plausibly assume that we may fit a *homogeneous* graph model to the entire population.

References

- Albert, A., and Anderson, J. A., (1984), “On the existence of maximum likelihood estimates in logistic regression,” *Biometrika*, 71, 1–10.
- Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, New York: Wiley.
- Burt, R. S., (1992), *Structural Holes: The Social Structure of Competition*, Cambridge, MA: Harvard University Press.
- Burt, R. S., (1987), “A Note on Missing Social Network Data in the General Social Survey,” *Social Networks*, 9, 63–73.
- Chib, S., and Greenberg, E. (1995), “Understanding the Metropolis Algorithm,” *The American Statistician*, 49, 327–335.

- Corander, J. and Dahmström, K. and Dahmström, P. (1998), “Maximum likelihood estimation for Markov graphs,” Research report, 1998:8, Stockholm University, Department of Statistics.
- Corander, J., and Dahmström, K., and Dahmström, P. (2002), “Maximum likelihood estimation for exponential random graph model,” pp:1-17 in Jan Hagberg (ed.), *Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in honour of Ove Frank*, University of Stockholm: Department of Statistics.
- Costenbader, E., and Valente, T. W. (2003), “The Stability of Centrality Measures when Networks are Sampled,” *Social Networks*, 25, 283?-307.
- Crouch, B., Wasserman, S., and Trachtenberg, F. (1998), “Markov Chain Monte Carlo maximum likelihood estimation for p^* social network models,” Paper presented at the Sunbelt XVIII and Fifth European International Social Networks Conference, Sitges (Spain), May 28–31, 1998.
- Diaconis, P., and Ylvisaker, D., (1979), “Conjugate Priors for Exponential Families,” *Ann. Stat.*, 7, 269–281.
- Frank, O., and Strauss, D. (1986), “Markov Graphs,” *Journal of the American Statistical Association*, 81, 832–842.
- Fienberg, S., and Wasserman, S. S. (1981), “Categorical Data Analysis of Single Sociometric Relations,” in *Sociological Methodology*, ed. S. Leinhardt, San Francisco: Jossey-Bass, pp. 156–192.
- Gelfand, A. E., and Smith, A. F. M., (1990), “Sampling based approaches to calculating marginal densities,” *Journal of the American Statistical Association*, 85, 398–409.

- Gelman, A., and Meng, X. L. (1998), “Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling,” *Statistical Science*, 13, 163–185.
- Geyer, C. J., and Thompson, E. (1992), “Constrained Monte Carlo maximum likelihood for dependent data,” *Journal of the Royal Statistical Society, Series B*, 54, 657–699.
- Ghilagaber, G., and Koskinen, J. (2008), “Bayesian Adjustment of Anticipatory Covariates in Analyzing Retrospective Data,” *Mathematical Population Studies* (forthcoming).
- Gile, K., and Handcock, M. S. (2006), “Model-based Assessment of the Impact of Missing Data on Inference for Networks,” Working Paper no. 66, Center for Statistics and the Social Sciences, University of Washington.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.
- Handcock, M. S. (2003). “Assessing degeneracy in statistical models of social networks,” Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. Obtainable from <http://www.csss.washington.edu/Papers/wp39.pdf>.
- Handcock, M, and Gile, K. (2007). “Modeling social networks with sampled or missing data,” CSSS working paper 75, University of Washington. Obtainable from: <http://www.csss.washington.edu/Papers/wp75.pdf>.
- Holland, P., and Leinhardt, S. (1981), “An exponential family of probability distributions for directed graphs” (with discussion), *Journal of the American Statistical Association*, 76, 33–65.
- Hunter, D. R. (2007), “Curved exponential family models for social networks,” *Social Networks*, 29, 216–230.

- Hunter, D. R., Goodreau, S. M., and Handcock, M. S., (2008), “Goodness of Fit of Social Network Models,” *Journal of the American Statistical Association*, 103, 248-258.
- Hunter, D. R., and Handcock, M. S. (2006), “Inference in Curved Exponential Family Models for Networks,” *Journal of Computational and Graphical Statistics*, 15, 565–583.
- Huisman, M. (2007), “Imputation of missing network data: Some simple procedures,” *Journal of Social Structure*. Forthcoming.
- Kossinets, G. (2006), “Effects of Missing Data in Networks,” *Social Networks*, 28, 247–268.
- Koskinen, J. (2008) “The Linked Importance Sampler Auxiliary Variable Metropolis Hastings Algorithm for Distributions with Intractable Normalising Constants,” MelNet Social Networks Laboratory Technical Report 08–01, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia. (available from http://www.sna.unimelb.edu.au/publications/MelNet_Techreport_08_01)
- Koskinen, J., Pattison, P. E., Robins, G., and Wang, P., (2008) “Extreme Actors - Outliers and Influential Observations in exponential random graph (p-star) models,” MelNet Social Networks Laboratory Technical Report 08–05, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia. (available from http://www.sna.unimelb.edu.au/publications/MelNet_Techreport_08_05)
- Koskinen, J., Robins, G., Pattison, P. E., (2008) “Missing data in social networks: Model-based inference,” MelNet Social Networks Laboratory Technical Report 08–03, Department of Psychology, School of Be-

- havioural Science, University of Melbourne, Australia. (available from http://www.sna.unimelb.edu.au/publications/MelNet_Techreport_08_03)
- Koskinen, J., Wang, P., Robins, G., and Lusher, D., (2008) “Approximate Bayesian Analysis for Assessing Goodness of Fit for Exponential Family Random Graph (p-star) Models,” Social Networks Laboratory Working Paper, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia.
- Krackhardt, D. (1987), “Cognitive social structures,” *Social Networks*, 9, 109–134.
- Laumann, E. O., Marsden, P. V., Prensky, D., (1983), “The boundary specification problem in network analysis,” In: Burt, R. S., Minor, M. J. (Eds.), *Applied Network Analysis*, Sage Publications, London, pp. 18-34.
- Lauritzen, S. L. (1996), *Graphical models*, Oxford: Oxford University Press.
- Lazega, E. (2001), *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*, Oxford: Oxford University Press.
- Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.
- Neal, R. M. (2005), “Estimating Ratios of Normalizing Constants Using Linked Importance Sampling,” Technical Report No. 0511, Department of Statistics, University of Toronto. (available from <http://arxiv.org/abs/math.ST/0511216>)
- Pattison, P. E., Wasserman, S. (1999), “Logit Models and Logistic Regressions for Social Networks: II. Multivariate Relations,” *British Journal of Mathematical and Statistical Psychology*, 52, 169–193.
- Robins, G. L., and Morris, M. (2007), “Advances in Exponential Random Graph (p*) Models,” *Social Networks*, 29, 169–172 .

- Robins, G. L., Pattison, P., and Woolcock, J., (2004), “Models for Social Networks with Missing Data,” *Social Networks*, 26, 257–283.
- Robins, G. L., Pattison, P. E., and Woolcock, J. (2005), “Small and other worlds: Global network structures from local processes,” *American Journal of Sociology*, 110, 894-936.
- Rubin, D. B. (1976), “Inference and Missing Data (with discussion),” *Biometrika*, 63, 581-592.
- Santner, Thomas, J. and Duffy, D. E. (1986), “A Note on A. Albert and J. A. Andersons conditions for the existence of maximum likelihood estimates in logistic regression models,” *Biometrika*, 73, 755–758.
- Snijders, T. A. B. (2007), “The Robbins-Monro algorithm for estimation of ERGMs with missing or sampled data,” Working paper: Nuffield College, University of Oxford.
- Snijders, T. A. B., (2002), “Markov chain Monte Carlo estimation of exponential random graph models,” *Journal of Social Structure*, 3(2), April.
- Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), “New Specifications for Exponential Random Graph Models,” *Sociological Methodology*, 36, 99–153.
- Stork, D., and Richards, W. D., (1992), “Nonrespondents in Communication Network Studies: Problems and Possibilities,” *Group and Organization Management*, 17, 193–209.
- Tanner, M. A., and Wong, W. H., (1987), “The calculation of posterior distributions by data augmentation (with discussion),” *Journal of the American Statistical Association*, 82, 528–550.

- Thompson, S. K., and Frank, O., (2000), “Model-based estimation with link-tracing sampling designs,” *Survey Methodology*, 26, 87–98.
- Tierney, L. (1994), “Markov Chains for Exploring Posterior Distributions” (with discussion and a rejoinder by the author), *Annals of Statistics*, 22, 1701–1762.
- van Duijn, M., Gile, K., and Handcock, M., (2007), “Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models,” CSSS working paper 74, University of Washington. Obtainable from: <http://www.csss.washington.edu/Papers/wp74.pdf>.
- Wasserman, S., Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.
- Wasserman, S., and Pattison, P. E. (1996), “Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^* ,” *Psychometrika*, 61, 401–425.