THE UNIVERSITY OF MELBOURNE

Missing data in social network analysis: Inferences from incomplete missing data.

Interim Report for DSTO

MelNet

Johan Koskinen
Garry Robins
Philippa Pattison

The University of Melbourne

Missing data in social network analysis: Inferences from incomplete missing data. Interim Report for DSTO

Johan Koskinen
Garry Robins
Philippa Pattison

Social Networks Laboratory,
University of Melbourne.

Executive summary

The tools and theories of social network analysis (SNA) are important to understand how social interactions influence the behaviour of social actors and how in turn their actions shape the social and organisational environment in which they operate. SNA has been employed to explain as diverse social phenomena as the diffusion of innovations; disease spread; team performance; coordination of corporate board decisions; balance of trade between countries; etc. Understanding the relational structures of groups and categories of social actors is essential to understanding their goals, motivations and actions.

More recently, SNA has been used to inform intelligence decisions, to investigate the structure of terrorist networks, and to examine exchanges within criminal and fraud networks. Although general social science suffers from issues of data inadequacy, it still enjoys the benefits of being able to gather open information from surveys, experiments and interviews. Needless to say these options are not always available to intelligence analysts who crucially must deal with data inadequacies when data sources are neither open nor complete.

For such reasons, recent scientific attention has been addressed to how data inadequacies affect SNA conclusions (e.g. Kossinets, 2006). For SNA data, the presence of missing information can have major effects, e.g. in determining connectivity and other system-level properties. Classical statistical treatments for missing data are simply not appropriate for relational data. The recent literature provides a somewhat better understanding of missing data effects but does <u>not</u> offer principled methods for drawing sound inferences.

This paper summarises our approaches to establish principled inference based on incomplete SNA data. These approaches are all under development, but are extremely promising. We propose to deal with five distinct yet interrelated data problems:

- fitting statistical models to determine social network properties when particular links in the social network are not known;
- methods for pooling incomplete relational data information from different sources;
- methods for dealing with network data involving *covert actors* (i.e where the network relations of some actors are not known);
- methods for dealing with network data where some actors may be operating with multiple aliases (*network doppelgangers*), and identifying who those actors might be;
- methods for dealing with network data where personal information about particular actors is not known.

We have made good progress on the first of these (in an illustrative example, we have managed to deal successfully with 20% missing ties.) We can generalise our first approach to develop methods in the other four areas. Each method provides principled inference in the face of incomplete data, and a measure of uncertainty. Moreover, each method may be used to direct further data collection, identifying areas where additional information would best diminish uncertainty. For each method, we envisage an iterative exchange of data analysis and collection to optimise the chances of obtaining valuable information where data gathering resources are limited.

Table of contents

A. Fitting exponential random graph models to data with missing information on dyads

A.I.1. Introduction

The importance of dealing with missing data in social network analysis (SNA) becomes clear when we consider the fundamental differences between the nature of missing data in the typical standard statistical analysis with the typical scenario in SNA.

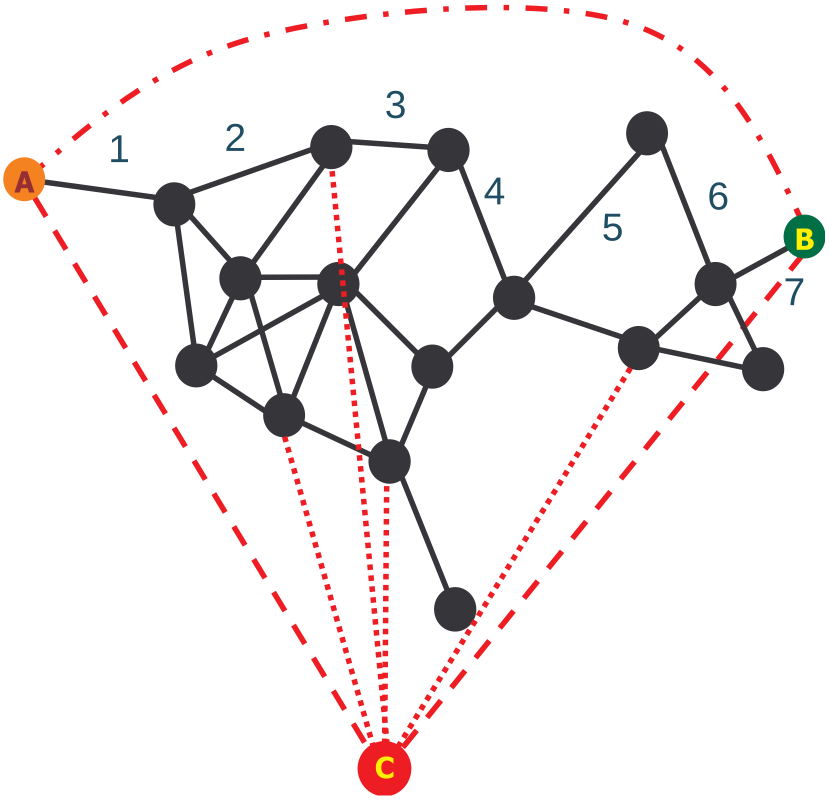

Figure 1 Illustration of missingness in SNA

It seems plausible to assume that we have an intuitive understanding of the network topology in terms of distances between nodes. If we for example consider the two vertices A and B in Figure 1, and assumed that only the solid lines were present, it would be natural for us to presume that whatever network processes that are going on in B's part of the network, these are unlikely to affect the network processes in A's part of the network. Had there on the other hand existed a tie (dash-dot) between A and B that went undiscovered, the picture would have been quite different. All of a sudden A and B are directly connected and when thinking about the layout of the network, the "pseudo-spatial" arrangement of nodes according to their proximity, we would be inclined to redraw the entire network. Similarly, had a vertex C went undiscovered, it may well have been the case that this vertex constituted an indirect link (via dashed edges) between A and B. Again, our intuitive understanding of vertices A and B's position in the network in relation to one another, as well as their overall position, would alter drastically with this new piece of information. Should, on top of everything else, C be connected to what we previously considered to be the centre of the network (dotted lines) many of our initial conclusions regarding the workings of the network would be altered radically.

Very generally put, what distinguishes missingness in standard statistical setups from those in SNA are the facts that in statistical analysis with independent observations:

1. we only really worry about whether missing observations differ from the obtained observations
2. interpretation of observed observations do not change with the knowledge (of the values) of missing observations

In SNA, because of the inherent dynamics, the self-organisation, and emergent qualities, other observations may be affected even if missing observations similar to observed observations.

As was pointed out by Stork and Richards (1992), ignoring missing data (e.g. in the form of non-respondents), and only treating the available data (c.f. ‘‘available-case’’ analysis, Little and Rubin, 1987, sec. 3.3), may be problematic since this is akin to redefining the boundary of the network. The importance and the difficulties surrounding the issue of the so-called boundary problem (Laumann et al., 1983), may be highlighted by considering the two basic ontological principles for social networks: the nominalist approach and the realist approach.

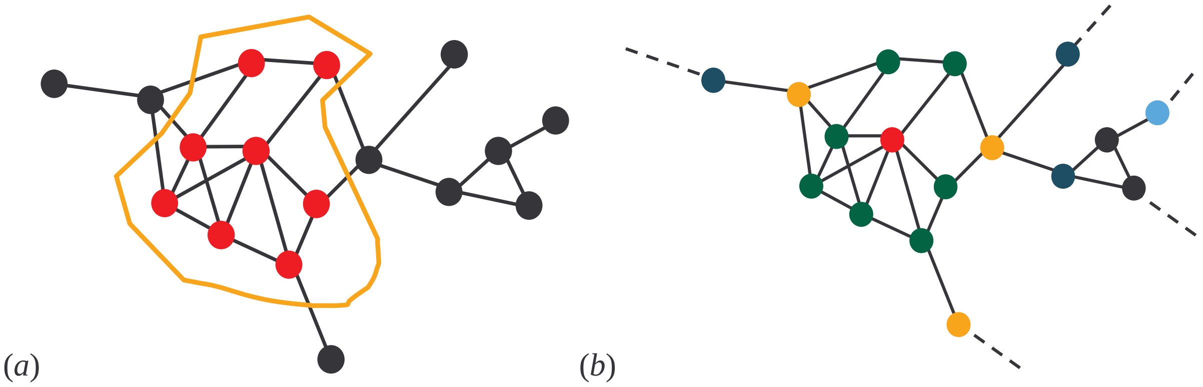

Figure 2 Sketch of principle ontological definitions of networks: (*a*) the nominalist approach, (*b*) the realist approach

The nominalist approach to defining the network is in terms of a predefined set of vertices that is defined in terms of group membership, actor attributes, relations, events, etc, that is meant to capture the relevant social neighbourhood of the actors for the particular relation that is studied. A relation may also be self-defining in terms of whom it relates to such as when work difficulties are studied in an organisation (workplace), in which case, say, work difficulties are likely not to have sensible interpretation across different settings. It is however hard to say how generally this may be applied to relations and, as remarked by Kossinets (2006), when an substantial number of ties extend beyond the predefine set of vertices (as in the case of Bearman et al., 2004, where 60% of choices made by student in the school studied extended outside the school) it may be hard to motivate their exclusion. Not only the quantity of ties extending outside of the prespecified set of vertices may be a problem but we may not always know whether or not the most important ties are those that cross over or not.

Even when we may rule out errors associated with when the same type of tie crosses the boundary, there may be different types of ties (that are not so easily confined to the set of vertices) that cross over. An example of this could be when work place advice is contingent on friendship (see e.g. Lazega and Pattison, 1999, for an elaboration on the interdependency of different relations). Similar issues may result from overlap of settings (Pattison and Robins, 2002).

The realist approach assumes that the relevant boundary of the network is that which the actors themselves consider to be the boundary. Roughly this would translate into defining the network

boundary from successively sampling waves in a snowball sampling approach. Conceptually this might beg questions of how we determine the initial sample, where and how do we stop without violating the realist approach and, what is more - what if the theory (hypothesis) of six degrees of separation is indeed true?
The above issues serve to illustrate that it is not only difficult to deal with missing data in SNA but also that the conceptualisation of missingness in SNA highlights the manner in which assumptions of ontology and epistemology are intertwined in SNA. (A fairly comprehensive survey of network measurement is given by Marsden, 2005)

A frequently used model for the ties in a social network, the ERGM, may be derived from how ties from one actor to another may depend on ties between other actors (Robins and Pattison, 2005). One interdependence assumption gives rise to the class of Markov graphs (Frank and Strauss, 1986) with the accompanying dependence graph in the left hand panel of **Error! Reference source not found.**, for 4 vertices (*i*, *j*, *k*, and *l*). Interestingly, if we were to remove the variable corresponding to the edge indicator for e.g. {*k*,*l*}, and wish to estimate a Markov model to the remaining indicators we see that the dependence structure is ‘‘distorted’’, that some of the indicators that were not previously tied are so when we marginalise with respect to the missing indicator.

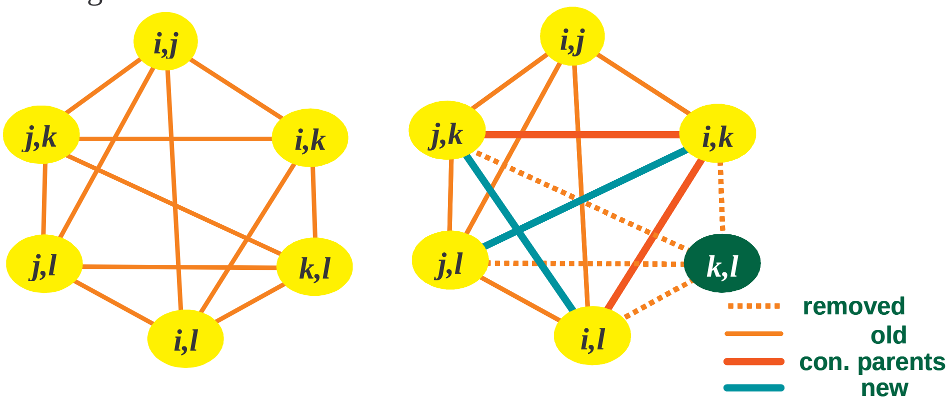

Figure 3 Dependence structure for Markov model on 4 vertices with complete data (left) and with dyad {*k*,*l*} missing (right)

A.I.2. Missing data and ERGMs

The first category, A, assumed that our primary interest is in fitting an exponential random graph model (ERGM) (Frank and Strauss, 1986; Pattison & Wasserman, 1999; Robins, Pattison, and Wasserman, 1999; Wasserman & Pattison, 1996; Snijders et al., in press) to social network data in the form of standard sociometric data represented by an adjacency matrix (Wasserman and Faust, 1994). The ERGMs have proved to be superior to many competing models (such as various scale-free based models) in that ERGMs (and their recent extensions, c.f. Snijders et al., in press) are capable of reproducing real social networks to a greater extent (Robins, Woolcock, and Pattison, 2005). Furthermore, we assume that information is missing as to the interactional status for some dyads (pairs of actors), either as a result of insufficient monitoring of some dyads or through non-response or lack of knowledge of the ties to or from specific actors (in which case information would be missing row-wise or column-wise from the adjacency matrix). Researchers have pointed to the difficulty in dealing with this and similar types of missing data (Burt, 1987) and the pitfalls of not dealing with it (Kossinets, 2003) but there are few suggestions as to how we should deal with this type of missing data. Robins et al. (2004) define a model for the ties in the network that allowed for some actors being respondents and others being non-respondents but apart from not being flexible enough (it requires that non-respondents are uniquely defined) their estimation relied on an approximation, the pseudo likelihood estimate (MPLE) (Besag, 1975; was elaborated for random digraph models by Strauss and Ikeda, 1990;

Frank, 1991; and Wasserman and Pattison, 1996), of the maximum likelihood estimate (MLE), that is known to be unreliable (Crouch, Wasserman, and Trachtenberg, 1998; Dahmström and Dahmström, 1993; Corander, Dahmström, and Dahmström, 1998, 2002; Snijders, 2002; Handcock, 2002, 2003). Other approaches to handling missing data rely chiefly on ad-hoc methods for imputing missing data (Stork and Richards, 1992; Huisman, 2007; Gile and Handcock, 2006).

Taking a Bayesian approach[1] we propose a Markov chain Monte Carlo (MCMC) algorithm that given a few assumptions regarding what causes observations on dyads to be missing allows us to fit an (curved) exponential random graph model to social network data with partially missing information. Parameter inference conditional on complete data is performed using the Linked Importance Sampler Auxiliary MCMC algorithm (Koskinen, 2006) and for missing data the conditional distribution is given straightforwardly by the (curved) exponential random graph model. Note that the success of this model-based missing data scheme relies crucially on the Bayesian approach, since that is the only statistical paradigm capable of treating missing data in a consistent way (see the seminal paper on data augmentation by Tanner and Wong, 1987). There is also a more general motivation for favouring the Bayesian approach in that it is unclear whether the asymptotic results that are the main motivations for using ML estimation hold. Tentatively it looks as if the normal approximations of the distribution of MLEs and standard errors are reasonably good (as judged by the similarity of the point estimates and estimated s.e.s to the posterior distributions in Koskinen, 2004; for possible pitfalls when using MLEs for binary data see e.g. Mantel, 1987). That said, the MLE is typically used because it is consistent (i.e. given enough data the MLE will be arbitrarily close to the true parameter values) but since we do not really know how the ERGM scales up we don't know how to make use of this asymptotic result (see Anderson, Butts, and Carley, 1999; Robins, Woolcock, & Pattison, 2005, for scaling up; some asymptotic results for exponential family distributions with interdependent observations are given in Strauss, 1986).

Being able to accommodate missing data is important to social networks in general, since it provides a way of dealing with scaling effects. It is well known that ERGMs defined for different size networks are not readily comparable because of how different graph statistics ‘‘scale-up’’ differently (again, Anderson, Butts, and Carley, 1999; Robins, Woolcock, & Pattison, 2005). Hence, if you disregard the missing portion of a network when fitting an ERGM or a multivariate ERGM, you will end up with a model that is not (in principle) comparable on the size level of the network you really want to investigate.

The notion of missing data on the level of the dyads extends the set-up where actors may be unambiguously classified as respondent or non-respondents (Robins et al., 2004) to data collection schemes where dyads are observed one at a time. Other interesting extensions include the cases of missing unknown actors; Multivariate ERGMs (Pattison and Wasserman, 1999; Koehly & Pattison, 2005) where different relations may be defined on different subsets of actors; Similarities of different networks of different sizes.

We briefly discuss performance of the suggested approach and its limitations. Of particular interest is the question of how little data is needed for the inference scheme to be practically feasible. We provide some conditions that have to be met in order for inference to be possible when we have missing data.

[1] For an accessible and non-technical introduction to Bayesian inference in a general behavioural and social science context is given in for example the special issue of Sociological Methods and Research (Western, 1999)

Central to solving this inferential problem is the development of the LISA algorithm and hence it must be consider a first priority for the project to write up and submit Koskinen (2006). Some initial progress has been made and the algorithm seems to compare favourably with the only other available algorithm (Møller et al., 2005) for performing Bayesian inference for this class of models (models for which the normalising constant in the likelihood is not analytically tractable; Note however, that there is an approximate Bayesian inference scheme proposed in Koskinen, 2004a). The second step is to apply LISA to the problem of missing data as framed here and subject Koskinen (2007a) to public scrutiny. Some progress is reported here in the research plan and tentatively we can conclude that in a case study, the removal of up to two actors from a the collaboration network of 36 lawyers (Lazega, 2001) has little influence on the parameter estimates.

Future extensions consist primarily of extending the model and inference scheme to include more complicated missing data mechanisms than that used in Koskinen (2007a). In the long run it would be desirable to have more realistic assumptions about what causes data to be missing. Some extensions that increase the plausibility of missing data scheme are easy to incorporate in theory but to asses whether these are practically implementable is an empirical issue as well as matter of technical experimentation.

A.II. The Linked Importance Sampler Auxiliary Variable (LISA) Metropolis Hastings for Distributions with Intractable Normalising Constants.

In most instances of Bayesian analysis (for a comprehensive treatment of Bayesian inference see for example Bernardo and Smith, 1994, Lindley, 1965, and Box and Tiao, 1973) of empirical data one has to rely on numerical methods for estimation. The most common set of tools for performing numerical Bayesian inference is Markov chain Monte Carlo (MCMC) methods (Gilks et al., 1996). Rather than calculating point estimates, measures of uncertainty, interval estimates, etc, analytically from the posterior distribution of the parameters given observed data, MCMC is a methodology for drawing samples from the posterior distribution. Since all information and the extent of the uncertainty regarding the parameters is captured by the posterior distribution, all relevant quantities needed for drawing conclusions about the model can be obtained from the posterior sample. For example we may calculate the posterior expected values for a parameter given data using the ergodic mean, i.e. taking the sample average for a parameter in the posterior sample. The main advantage of MCMC is that the precision of the results obtained depend in principle only on the number of sample points drawn from the posterior of the parameters given data (Tierney, 1994).

MCMC for posterior sampling typically only requires that the posterior distribution is known up to a normalising constant. This means that we only need to be able to evaluate the likelihood function and the prior distribution for any given parameter point. Lenient though this requirement is, there are several important models in statistical mechanics and in the social sciences where the likelihood function cannot be evaluated because the normalising constant in the likelihood is analytically intractable. To the extent that Bayesian analysis has at all been conducted for this class of models it has relied on approximate numerical methods with unknown properties (Møller et al., 2005). Having to rely on approximate procedure is unsatisfactory if not because the extremely general results for MCMC that hold under very generous assumption do not apply and the appropriateness of any given approximation is likely to be very sensitive to what model is used and what data analysed and hence the appropriateness has to be decided on a case to case basis. This has an inbuilt contradiction in that it may prove difficult to assess the appropriateness when there is no procedure to evaluate the approximation against.

Here we summarise work in progress building on Koskinen (2006) to propose and investigate the properties of an ‘‘exact’’ or ‘‘pure’’ MCMC algorithm for performing Bayesian inference for models with intractable normalising constants. Because of the performance deficiency of the auxiliary variable MCMC algorithm (SISA) proposed by Møller et al. (2005) when applied to complex models we propose instead the use of The Linked Importance Sampler Auxiliary Variable (LISA) Metropolis Hastings for Distributions with Intractable Normalising Constants. We show how the poor mixing of the SISA can be understood if SISA is formulated as a regular MCMC with an embedded importance sampler that estimates the normalising constant in each step using only one sample point. We proceed to suggest that this makes it natural to replace the one-sample simple importance sample by more elaborate and proved more efficient variations of the importance sampler. There is a host of different very efficient importance sampler and Gelman and Meng (1998) brought recently brought to the attention of the statistical community the similarities and communalities between traditional importance samplers and methods that have long been used in the Physics literature. Unfortunately the ergodic theorem (Tierney, 1994) that ergodic estimator of the normalising constant converges to its mean almost surely as the number of sample points tends to infinity is of little use to use since we firstly have to take an estimate in each iteration of the MCMC and secondly because while the estimate of the normalising constant is simulation consistent the estimate of the acceptance ratio is not. It turns out however, that a specific importance sampler, the linked importance sampler (LIS), can be incorporate in the MCMC as an auxiliary variable when the space on which the importance distribution is defined is considered an extended sample space. This extended sample space is discrete but is of a quite complicated nature. However, we need to consider the variable defined on the extended state space explicitly in the sense that we need to save memory-consuming variables, the part of the LISA that concerns the auxiliary variable reduced to taking an importance sample.

A.II.1. Model definition and notational preliminaries

We consider inference for a family of models for a variable **X**, taking values in a finite space $\mathscr{X}$. These models are assumed to be indexed by a *p* × 1 vector $\theta \in \Theta \subseteq \Re$ of real-valued parameters and that a model may be written with a probability mass function (pmf) of the form

$$P(\mathbf{X} \mid \theta) = \frac{1}{Z_\theta} q_\theta(\mathbf{X}) \text{, (E-1)}$$

where $q_\theta(\mathbf{X})$ is a real valued function of both the parameter vector and the variable **X**, and

$$Z_\theta = \sum_{\mathbf{U} \in \mathscr{X}} q_\theta(\mathbf{U})$$

is the normalising constant (or partition function according to these models' use in statistical mechanics, Strauss, 1986) that is only a function of the parameter vector θ.

A.II.2. Maximum likelihood estimation

For all but the trivial parameterisations of the models considered here, the main obstacle to performing statistical inference is the fact that Z_θ is analytically intractable. Although the function *q* may itself be relatively easy to evaluate, given θ and **X**, Z_θ is a sum over a set whose cardinality quickly becomes very big as a function of the number of coordinates or elements of

X. The sample space for an Ising model on an $n \times m$ grid, for example, has a cardinality of 2^{nm}. An exponential random graph model (Frank and Strauss, 1986; Pattison & Wasserman, 1999; Robins, Pattison, and Wasserman, 1999; Wasserman & Pattison, 1996; Wasserman and Robins, 2005, Snijders et al., in press; Hunter and Handcock, 2006) for the directed graph on *n* vertices has a sample space of cardinality $2^{n(n-1)}$. Many of the models considered here may be seen as special cases of the Gibbs distribution defined for different applications and sample spaces. Examples include Markov random fields (Besag, 1975; Cressie, 1993); Markov point processes (Ripley and Kelly, 1977; Møller and Waagepetersen, 2003b); and metric random graphs (Banks and Constantine, 1998). Non-Bayesian inference first relied on pseudo likelihood estimation (Besag 1974; Frank and Strauss, 1986; Strauss and Ikeda, 1990; Geyer and Thompson 1992) but because the pseudo likelihood estimates (MPLEs) thus obtained are suspect in certain circumstances and because of the generally higher efficiency of the maximum likelihood estimator (MLE) Geyer and Thompson (1992) proposed a Markov chain Monte Carlo (MCMC) scheme for performing approximate MLE inference. The MLE can be obtained using MCMC to get an approximation of the normalising constant (e.g. Geyer and Thompson, 1992; Gelman and Meng, 1998; Gu and Zhu, 2001) and for a few special cases the normalising constant can even be calculated exactly using an iterative scheme (Reeves and Pettitt, 2003). For a subset of distributions in the exponential family of distributions Lindsey (1974) proposed a method for fitting distributions to data in such a was so as the normalising constant does not have to be evaluated (Aitkin, 1995). The properties of the exponential family of distributions have also been utilised for MLE algorithms based on cumulants (Dahmström and Dahmström, 1993; Corander, Dahmström, and Dahmström, 1998, 2002) and the moment equation using either stochastic approximation (Snijders, 2002) or importance sampling (Crouch, Wasserman, and Trachtenberg, 1998; Handcock, 2002; for the extension to the curved exponential family of distributions see Hunter and Handcock, 2005).

Here we propose a Bayesian approach primarily because it gives us more nuanced information regarding the parameters than the ML estimation that typically only provides us with points estimates and standard errors. In addition to the wealth of information provided about the parameters by the posterior distribution, a Bayesian inference scheme also opens for ways of performing model selection (using posterior predictive p-values, Meng, 1994, or Bayes factors), handling missing data, etc. Another reason is that the Bayesian inference has somewhat more favourable properties. Although it is likely to have a small impact on the actual analysis, it is unclear whether the asymptotic results that are the main motivations for using ML estimation hold. Tentatively it looks as if the normal approximations of the distribution of MLEs and standard errors are reasonably good (as judged by the similarity of the point estimates and estimated s.e.s to the posterior distributions in Koskinen, 2004; for possible pitfalls when using MLEs for binary data see e.g. Mantel, 1987). That said, the MLE is typically used because it is consistent (i.e. given enough data the MLE will be arbitrarily close to the true parameter values) but with interdependent observations the asymptotic results are different (Strauss, 1986).

A.II.3. Previous Bayesian approaches

To be able to evaluate the likelihood function is as central to Bayesian analysis as it is to non-Bayesian analysis. Given that non-Bayesian estimation previously relied on maximisation of the pseudo likelihood it is perhaps a natural approach to perform Bayesian inference using the pseudo likelihood rather than the true likelihood function as was done in Heikkinen and Högmander (1994). This transforms the problem into a regular inference issue and standard MCMC methods may be used but Heikkinen and Högmander (1994) acknowledged that it is unclear what distribution one samples from. Another way of avoiding having to evaluate the normalising constant is for example by clever use of prior distributions (Besag et al., 1991) or limiting the analysis to finding a point estimate (Heikkinen and Penttinen, 1999).

As mentioned above, many non-Bayesian methods exist that depend on MCMC approximation of the normalising constant. Since there are numerous efficient algorithms for numerically calculating (approximating) the normalising constant (Gelman and Meng, 1998), many MCMC schemes have been proposed for models with intractable normalising constants where a MCMC approximation to the normalising constant in the likelihood is substituted for the exact value. Normalising constants can be evaluated on a grid of parameter values and stored (Berthelsen and Møller, 2003) or estimated repeatedly in the course of the MCMC, using a sample from an importance distribution that is stored off-line (Koskinen and Robins, 2007) or regenerated on-line (Koskinen, 2004a).

Common to the previously employed Bayesian inference schemes are that it is hard to establish what properties MCMC that relies on approximations to distributions rather than the exact expressions has. The estimators of the normalising constant that are currently available are mostly constructed to estimate individual constants (or ratios of constants) and are not suited to repeated estimation. If one wishes to have one estimate or approximation of a normalising constant one is willing to allow for more iterations, to sacrifice efficiency for precision.

Møller et al. (2005) recently proposed the first generally applicable ‘‘exact’’ MCMC algorithm for distributions with intractable normalising constants. By introducing an auxiliary variable defined on the same state space as data, they avoid having to deal with the normalising constant in evaluation of the likelihood explicitly.

A.II.4. Obtaining the posterior distribution from posterior simulation

If we for the model (E-1) have a prior distribution $\pi(\theta)$, the posterior distribution of θ given the we have observed data **X**, is given by

$$\pi(\theta \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \theta)\pi(\theta)}{m(\mathbf{X})} \propto \frac{1}{Z_\theta} q_\theta(\mathbf{X})\pi(\theta),$$

where

$$m(\mathbf{X}) = \int \frac{1}{Z_\theta} q_\theta(\mathbf{X})\pi(\theta)\,\mathrm{d}\,\theta$$

is the marginal likelihood. The Metropolis-Hastings (MH) algorithm produces an MCMC sample $(\theta^{(k)})_{k=0}^{N}$ from the posterior distribution that can be used for exploring the posterior distribution. An iteration in the MH algorithm consists of proposing a move from the present point $\theta^{(k)}$ to a new point $\theta *$ draw from a proposal density $q(\theta * \mid \theta^{(k)})$ and accepting this move, setting $\theta^{(k+1)} = \theta *$ with a probability min(1,H), where H is the Hastings ratio

$$H = \frac{\pi(\theta * \mid \mathbf{X})}{\pi(\theta^{(k)} \mid \mathbf{X})} \frac{q(\theta^{(k)} \mid \theta *)}{q(\theta * \mid \theta^{(k)})}.$$

The marginal likelihood, that is typically very hard to evaluate, cancels in the Hastings ratio since it is only a function of data. Hence, drawing parameters from the posterior distribution in the MH reduces to a sequence of evaluations of the likelihood function (and prior distribution), something that in the standard case is easily done. Here we cannot evaluate the likelihood by assumption wherefore the cancellation of the marginal likelihood in the Hastings ratio is of little comfort to us.

A.II.5. SISA, the auxiliary variable MCMC

The problem in creating a MH (or indeed any type of MCMC) for the models considered here is that in the Hastings ratio

$$H = \frac{q_{\theta^*}(\mathbf{X}) / Z_{\theta^*}}{q_{\theta^{(k)}}(\mathbf{X}) / Z_{\theta^{(k)}}} \frac{q(\theta^{(k)} \mid \theta^*)}{q(\theta^* \mid \theta^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})}$$

is that we cannot evaluate the ratio $\Lambda(\theta^*, \theta^{(k)}) = Z_{\theta^{(k)}} / Z_{\theta^*}$. To circumvent the need to evaluate $\Lambda(\theta^*, \theta^{(k)})$ while retaining the properties of the MCMC scheme, Møller et al. (2005) proposed to introduce an auxiliary variable **Y**, that has the same state space $\mathscr{X}$ as **X**, and to set up the MH to produce a sample $(\theta^{(k)}, \mathbf{Y}^{(k)})_{k=0}^{N}$ from the joint posterior of **Y** and θ. By letting **Y** have the pmf $q_{\theta_0}(\mathbf{Y}) / Z_{\theta_0}$ for θ_0 fixed, the Hastings ratio for the joint acceptance of $(\theta^*, \mathbf{Y}^*)$ becomes

$$H = \frac{q_{\theta^*}(\mathbf{X}) / Z_{\theta^*}}{q_{\theta^{(k)}}(\mathbf{X}) / Z_{\theta^{(k)}}} \frac{q_{\theta_0}(\mathbf{Y}^*) / Z_{\theta_0}}{q_{\theta_0}(\mathbf{Y}^{(k)}) / Z_{\theta_0}} \frac{q(\theta^{(k)}, \mathbf{Y}^{(k)} \mid \theta^*, \mathbf{Y}^*)}{q(\theta^*, \mathbf{Y}^* \mid \theta^{(k)}, \mathbf{Y}^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})}.$$

While we see that the normalising constant Z_{θ_0} in the pmf of **Y** cancel out, the problem of evaluating $\Lambda(\theta^*, \theta^{(k)})$ remains. The trick employed in Møller et al. (2005) was to firstly factorise the proposal density $q(\theta^*, \mathbf{Y}^* \mid \theta^{(k)}, \mathbf{Y}^{(k)}) = q(\mathbf{Y}^* \mid \theta^*) q(\theta^* \mid \theta^{(k)})$ so that **Y** is drawn conditional on the proposed new value of θ. Secondly, the conditional proposal distribution for **Y** is set to $q_{\theta^*}(\mathbf{Y}) / Z_{\theta^*}$. Doing this, the Hastings ratio simplifies to

$$\begin{aligned} H &= \frac{q_{\theta^*}(\mathbf{X}) / Z_{\theta^*}}{q_{\theta^{(k)}}(\mathbf{X}) / Z_{\theta^{(k)}}} \frac{q_{\theta_0}(\mathbf{Y}^*)}{q_{\theta_0}(\mathbf{Y}^{(k)})} \frac{q(\mathbf{Y}^{(k)} \mid \theta^{(k)})}{q(\mathbf{Y}^* \mid \theta^*)} \frac{q(\theta^{(k)} \mid \theta^*)}{q(\theta^* \mid \theta^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})} \\ &= \frac{q_{\theta^*}(\mathbf{X}) / Z_{\theta^*}}{q_{\theta^{(k)}}(\mathbf{X}) / Z_{\theta^{(k)}}} \frac{q_{\theta_0}(\mathbf{Y}^*)}{q_{\theta_0}(\mathbf{Y}^{(k)})} \frac{q_{\theta^{(k)}}(\mathbf{Y}^{(k)}) / Z_{\theta^{(k)}}}{q_{\theta^*}(\mathbf{Y}^*) / Z_{\theta^*}} \frac{q(\theta^{(k)} \mid \theta^*)}{q(\theta^* \mid \theta^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})} \\ &= \frac{q_{\theta^*}(\mathbf{X})}{q_{\theta^{(k)}}(\mathbf{X})} \frac{q_{\theta_0}(\mathbf{Y}^*)}{q_{\theta_0}(\mathbf{Y}^{(k)})} \frac{q_{\theta^{(k)}}(\mathbf{Y}^{(k)})}{q_{\theta^*}(\mathbf{Y}^*)} \frac{q(\theta^{(k)} \mid \theta^*)}{q(\theta^* \mid \theta^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})} \end{aligned}$$

and $\Lambda(\theta^*, \theta^{(k)})$ dissapears. Hence, with only a bit of algebra we have done away with the need to evaluate the normalising constant. We will call this algorithm SISA because of its relation to simple importance sampling, to be more closely described in section A.1.vi.

A.II.6. Why SISA? Mixing of the auxiliary variable MCMC

In the paper Møller et al. (2005) make the remark that for some models and parameter specifications the SISA has a tendency ‘‘get stuck’’ for long times. In the Illustrations section of this paper we intend to show further examples of this when the SISA is applied to models with more complicated variable interaction than the autologistic model. In general however, there is cause for suspicion or at least some reasons to be cautious as to the performance of the SISA.

Firstly, note that in order to accept a proposed move in SISA, we need to accept the update to both **Y** and θ. This suggests that the acceptance rate of SISA should be considerably lower than a MH using the true Hastings ratio. Clearly the choice of distribution for **Y** is crucial to retain an acceptable acceptance rate but it is not immediately clear how the Hastings ratio in the SISA relates to the true Hastings ratio. Møller et al. (2005) only give some heuristic motivations for the choice of auxiliary density. Secondly, for the algorithm to work θ has to be independent of **Y** given **X** and θ_0, which would seem to have the interpretation that the posterior distribution is ‘‘diffused’’ or spread out.

For understanding the performance and (loosely speaking) efficiency of the SISA it is helpful to consider SISA in terms of importance sampling. If we inspect the part in the Hastings ratio in the SISA algorithm that pertains to the auxiliary variable and write

$$\hat{\Lambda}(\theta^*,\theta^{(k)};\mathbf{Y}^*,\mathbf{Y}^{(k)})=\frac{\hat{\Lambda}(\theta^*,\theta_0;\mathbf{Y}^*)}{\hat{\Lambda}(\theta^{(k)},\theta_0;\mathbf{Y}^{(k)})}$$

where

$$\hat{\Lambda}(\theta,\theta_0;\mathbf{Y})=\frac{q_{\theta_0}(\mathbf{Y})}{q_\theta(\mathbf{Y})},$$

we see that $\hat{\Lambda}(\theta,\theta_0;\mathbf{Y})$ is an estimator of $\Lambda(\theta,\theta_0)$ in the sense that

$$E_{\mathbf{Y}|\theta}\left[\hat{\Lambda}(\theta,\theta_0;\mathbf{Y})\right]=\sum_{\mathbf{Y}\in\mathscr{X}}\left[\frac{q_{\theta_0}(\mathbf{Y})}{q_\theta(\mathbf{Y})}\right]\frac{1}{Z_\theta}q_\theta(\mathbf{Y})=\frac{Z_{\theta_0}}{Z_\theta}.$$

In other words, if $\mathbf{Y}_1,\ldots,\mathbf{Y}_M$ is a sample from the importance distribution $q_\theta(\mathbf{Y})/Z_\theta$, the ergodic average $\overline{\Lambda}(\theta,\theta_0)=\frac{1}{M}\sum\hat{\Lambda}(\theta,\theta_0;\mathbf{Y}_m)$ is the simple importance sampler (SIS) estimator of $\Lambda(\theta,\theta_0)$. The SISA may then be seen as a Metropolis-Hastings algorithm where a SIS is run in each iteration to approximate the true Hastings ratio. Given some regularity assumptions (and assuming that the sample points $\mathbf{Y}_1,\ldots,\mathbf{Y}_M$ are approximately independent) $\overline{\Lambda}(\theta,\theta_0)$ is a simulation consistent estimator of $\Lambda(\theta,\theta_0)$ with variance $Var_{\mathbf{Y}|\theta}\left[\hat{\Lambda}(\theta,\theta_0;\mathbf{Y})\right]/M$. Without going into too much detail (some of which is treated in the following section), we may note that there is reason to be concerned about the fact that SISA employs SIS with only *M* = 1 sample point.

A.II.7. Importance sampling for calculating ratios of normalising constants

Here we are going to briefly recap different kinds of importance samplers. We are not making any claims at this being an exhaustive account since there are many good reviews and introductions to importance sampling (e.g. Gilks et al., 1996, and in particular Gelman and Meng, 1998).

A.II.7.a. Simple importance sampling

As mentioned above, the simple importance sampler (SIS) estimator of the ratio of normalising constants $\Lambda(\theta,\theta_0)$, is $\overline{\Lambda}(\theta,\theta_0)=\frac{1}{M}\sum\hat{\Lambda}(\theta,\theta_0;\mathbf{Y}_m)$ for $\mathbf{Y}_1,\ldots,\mathbf{Y}_M$ is a sample from the importance distribution $q_\theta(\mathbf{Y})/Z_\theta$. Independence is not needed for the estimator to be simulation consistent but the variance will be larger due to autocorrelation. SIS is an intuitive and straightforward way of estimating ratios of normalising constants and we only really require that the support of **Y** under the different distributions defined by θ and θ_0 is the same, that q_{θ_0} dominates q_θ. In most applications for finite supports this condition is met but it is common for the supports of q_{θ_0} and q_θ to be well separated in the sense that there is a region in $\mathscr{X}$ which has a very low probability under both q_{θ_0} and q_θ that separates the regions of high probability under the respective distributions as illustrated in Figure 4.

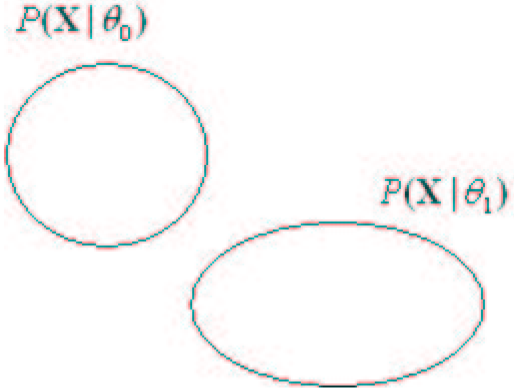

Figure 4 Two supports that are well separated

This situation means that we are rarely going to get **Y** that have high probability under q_{θ_0} when q_θ is used as the importance distribution. This typically manifests itself as high or infinite variance for $\overline{\Lambda}(\theta,\theta_0)$. Note however that this also applies to less extreme cases as long as the ‘‘overlap’’ between distributions is too small.

In addition to the high variability and instability of SISA due to the fact that *M* = 1, in the course of running SISA we have to perform many SIS for many different values of θ. Consequently there is nothing to assure us that q_{θ_0} when q_θ are close to eachother other than that θ_0 is chosen so that most proposed values of θ are close to θ_0.

A.II.7.b. Bridged importance sampling

To remedy the deficiency of SIS when the supports are separated we may introduce a bridging distribution between q_{θ_0} when q_θ connecting their respective supports. We do this by expanding $\Lambda(\theta,\theta_0)$ using a bridging distribution indexed by a parameter $\theta_{1/2}$

$$\Lambda(\theta,\theta_0)=\frac{Z_{\theta_0}}{Z_\theta}=\frac{Z_{\theta_0}}{Z_{\theta_{1/2}}}\times\frac{Z_{\theta_{1/2}}}{Z_\theta}\,.$$

With a bridging distribution $q_{\theta_{1/2}}$ we write the estimator $\overline{\Lambda}(\theta,\theta_0)=\overline{\Lambda}(\theta,\theta_{1/2})\overline{\Lambda}(\theta_{1/2},\theta_0)$. Hence, even if the supports of q_{θ_0} and q_θ are disjoint, there is some overlap between the supports of q_{θ_0} and $q_{\theta_{1/2}}$, and between $q_{\theta_{1/2}}$ and q_θ . as illustrated in Figure 5.

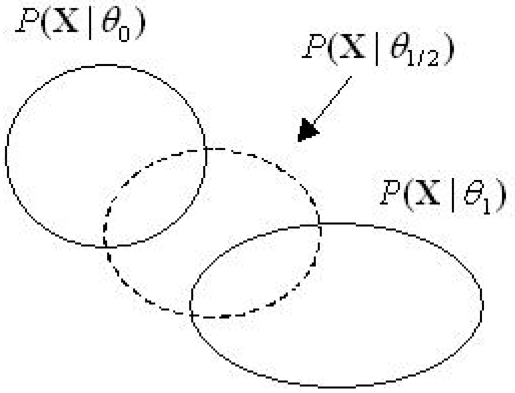
$P(\mathbf{X}\,|\,\theta_0)$
$P(\mathbf{X}\,|\,\theta_{1/2})$
$P(\mathbf{X}\,|\,\theta_1)$

Figure 5 Two disjoint supports with the support of a bridging distribution linking them

In practice we may be required to have more than one bridging distribution but the principle remains unchanged.

A.II.7.c. Path sampling

Gelman and Meng (1998) alerted the statistical community to the affinities between existing importance samplers in the statistical literature and various methods used in physics for calculating ratios of normalising constants. In particular path sampling is an elegant generalisation of bridged importance sampling. Consider extending the number of bridging distributions to ‘‘uncountably many’’ bridging distributions. We could for example have bridging distributions indexed by parameters $\theta(t)$ for a smooth mapping $\theta:[0,1]\to\Theta$ that is linear $\theta(t)=t\theta_0+(1-t)\theta_1$. The distributions given by *t* would then connect $\theta(0)=\theta_0$ with $\theta(1)=\theta_1$ in a continuous fashion. The estimator of the logarithm of $\Lambda(\theta_1,\theta_0)$ may then be derived from the path sampling identity:

$$\log\Lambda(\theta_1,\theta_0)=\int_0^1 E_{\mathbf{Y}|\theta(t)}\left\{\frac{\mathrm{d}}{\mathrm{d}\,\theta(t)}\log q_{\theta(t)}(\mathbf{Y})\right\}^T\frac{\mathrm{d}\,\theta(t)}{\mathrm{d}\,t}\mathrm{d}\,t\ .$$

The most straightforward estimator is suggested by the fact that the RHS of the path sampling identity looks like the expectation of the quantity in the integrand with respect to a random variable *t* ∈R(0,1). Hence, we may take a sample $t_1,\ldots,t_K$ from a uniform distribution and for each $\theta(t_u)$ we draw **Y** and calculate the quantity in the integrand and in the end we average these quantities.

A.II.7.d. Linked importance sampling (LIS)

Neal (2005) propose a method he called linked importance sampling (LIS) that combines the merits of the SIS (being unbiased) with the advantages of using bridging distributions while not requiring more than one independent realisation from an importance distribution. The path sampler, though being very efficient, requires that each sample point used is independent of the other given the parameters. When we use MCMC to generate sample points in the data space this translates into having to wait for the MCMC to burn in between each sample point. LIS is best describes a sequence of MCMC samples each from different distributions but that are linked (as in share a realisation) with each other. The principle is illustrated in Figure 6. In order for the estimator calculated in LIS to have the right properties we need to choose the starting points of the MCMC samples in a specific way. In addition we also have perform MCMC sampling forward in time as well as backwards in time.

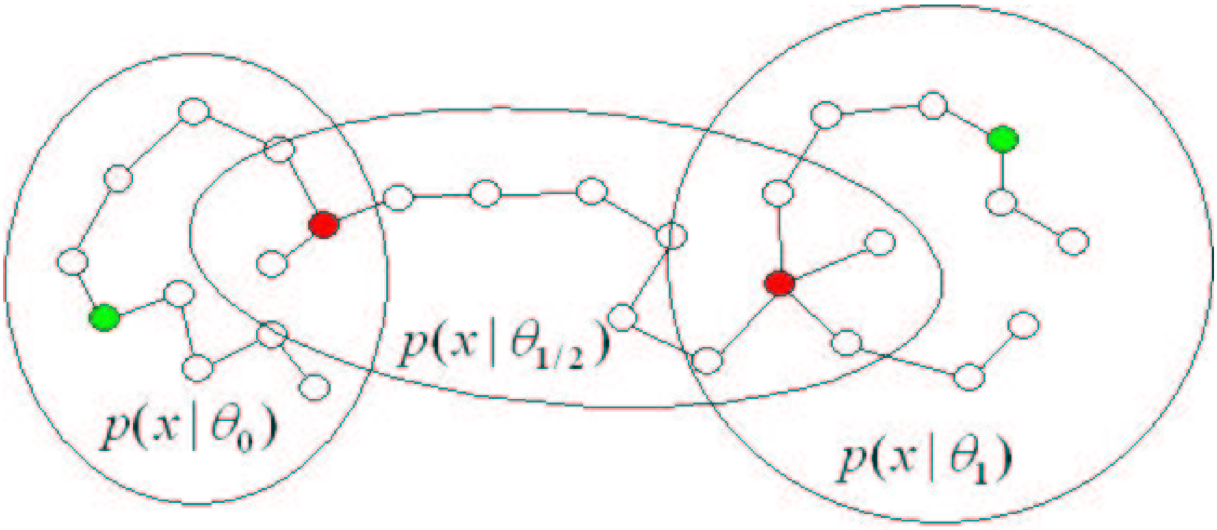

Figure 6 An illustration (based on Figure 1 in Neal, 2005) of LIS that starts in the green vertex on the left and ends in the green vertex on the right

*A.II.7.d.i. **Simulating forwards and backwards***

When we draw sample points $\mathbf{Y}$ from $q_\theta(\mathbf{Y})/Z_\theta$ using MCMC we usually simulate forwards with Markov chain transition probabilities T_θ, which may be schematically represented as

$$\mathbf{Y}^{(t)} \to T_\theta(\mathbf{Y}^{(t)},\mathbf{Y}^{(t+1)}) \to \mathbf{Y}^{(t+1)}$$

but we may also simulate backwards

$$\mathbf{Y}^{(t-1)} \leftarrow \underline{T}_\theta(\mathbf{Y}^{(t)},\mathbf{Y}^{(t-1)}) \leftarrow \mathbf{Y}^{(t)},$$

using the reverse transition probabilities $\underline{T}_\theta$. Most of the time we are dealing with reversible MCMC in which case $T_\theta(\mathbf{Y},\mathbf{X}) = \underline{T}_\theta(\mathbf{Y},\mathbf{X})$.

A.II.7.d.ii. The sample

The estimator is based on *K* sample points from *m* different distributions

$$\begin{array}{c}\mathbf{Y}_1^{(1)},\dots,\mathbf{Y}_1^{(K)}\\ \mathbf{Y}_2^{(1)},\dots,\mathbf{Y}_2^{(K)}\\ \vdots\\ \mathbf{Y}_m^{(1)},\dots,\mathbf{Y}_m^{(K)}\end{array}$$

drawn using Metropolis-Hasting transition probabilities $T_{\theta(t)}$ and $\underline{T}_{\theta(t)}$, for

$$\theta(0)=\theta_0,\theta(1),\ldots,\theta(m)=\theta_1 .$$

A.II.7.d.iii. Connecting the samples

The *m* samples are connected in points

$$\mu_1,\ldots,\mu_m \text{ and } \nu_1,\ldots,\nu_m$$

such that given μ_j and $\mathbf{Y}_j^{(1)},\ldots,\mathbf{Y}_j^{(K)}$,

$$\mathbf{Y}_{j+1}^{(\nu_{j+1})} := \mathbf{Y}_j^{(\mu_j)} .$$

Given ν_j and $\mathbf{Y}_j^{(\nu_j)}$ we create the chain $\mathbf{Y}_j^{(1)},\ldots,\mathbf{Y}_j^{(K)}$ by simulating forward from $\mathbf{Y}_j^{(\nu_j)}$ using $T_{\theta(j)}(\mathbf{Y}_j^{(\nu_j)},\mathbf{Y}_j^{(\nu_j+1)})$, $T_{\theta(j)}(\mathbf{Y}_j^{(\nu_j+1)},\mathbf{Y}_j^{(\nu_j+2)})$, etc., until we have produced $\mathbf{Y}_j^{(K)}$. We also simulate backwards from $\mathbf{Y}_j^{(\nu_j)}$ using $\underline{T}_{\theta(j)}(\mathbf{Y}_j^{(\nu_j)},\mathbf{Y}_j^{(\nu_j-1)})$, $\underline{T}_{\theta(j)}(\mathbf{Y}_j^{(\nu_j-1)},\mathbf{Y}_j^{(\nu_j-2)})$, etc, until we have produced $\mathbf{Y}_j^{(1)}$. The implied pmf of a chain conditional on the insertion point and the linking state is

$$\Pi(\mathbf{Y}_j \mid \nu_j,\mathbf{Y}_j^{(\nu_j)}) = \prod_{i=1}^{\nu_j-1}\underline{T}_{\theta(j)}(\mathbf{Y}_j^{(i+1)},\mathbf{Y}_j^{(i)})\prod_{i=\nu_j}^{K}T_{\theta(j)}(\mathbf{Y}_j^{(i)},\mathbf{Y}_j^{(i+1)}) .$$

A.II.7.d.iv. Choosing connection points

We have now explained how to produce the *m* sample chains and how to link them to eachother. We now proceed to explain how to choose the connection points. After ν_1 is drawn uniformly at random, the starting state $\mathbf{Y}_1^{(\nu_1)}$ is chosen according to $q_{\theta_0}(\mathbf{Y})/Z_{\theta_0}$. The first chain is simulated as described above. To choose which of the *K* sample points that should provide the link to the next chain, we choose μ_j with probabilities

$$p(\mu_j \mid (\mathbf{Y}_j^{(i)})_{i=1}^{K}) = \frac{w_{\theta(j),\theta(j+1)}(\mathbf{Y}_j^{(\mu_j)})}{\sum_{i=1}^{K} w_{\theta(j),\theta(j+1)}(\mathbf{Y}_j^{(i)})},$$

where

$$w_{\theta(j),\theta(j+1)}(\mathbf{Y}_j^{(i)}) = \sqrt{q_{\theta(j)}(\mathbf{Y}_j^{(\mu_j)})q_{\theta(j+1)}(\mathbf{Y}_j^{(\mu_j)})}\Big/ q_{\theta(j)}(\mathbf{Y}_j^{(\mu_j)})$$

and insertion points ν_j uniformly on {1,...,*K*}.

A.II.7.d.v. The estimator

Given a sample $(\mathbf{Y},\mu,\nu)$, an estimate of $\Lambda(\theta_0,\theta_1)$ is given by

$$\hat{r}_{LIS}(\mathbf{Y},\mu,\nu)=\prod_{j=1}^{m}\frac{\sum_{i=1}^{K} w_{\theta(j),\theta(j+1)}(\mathbf{Y}_j^{(i)})}{\sum_{i=1}^{K} w_{\theta(j+1),\theta(j)}(\mathbf{Y}_{j+1}^{(i)})}.$$

A.II.7.d.vi. Why LIS works - a tentative proof

Here we outline a proof of the unbiasedness (simulation consistency) of LIS, the chief aim of which will be for understanding how LIS is incorporated in the MCMC algorithm LISA. Details of the proof are given in Neal (2005).

Given a fixed starting point $\mathbf{Y}_1^{(\nu_1)}$, the sampling scheme outlined above defines a distribution

$$\frac{\Pi^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)}{q_{\theta_0}(\mathbf{Y}_1^{(\nu_1)})/Z_{\theta_0}}$$

on $\prod_{j=1}^{m}\mathscr{X}^K\times\{1,\ldots,K\}\times\{1,\ldots,K\}$, where $\Pi^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)=P^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)/Z_{\theta_0}$. The distribution $P^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)$ is simply that which is implied by drawing linking states, insertion points and simulating forwards and backwards

$$P^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)=\prod_{j=1}^{m}\tfrac{1}{K}\Pi(\mathbf{Y}_j\mid\nu_j,\mathbf{Y}_j^{(\nu_j)})\,p(\mu_j\mid(\mathbf{Y}_j^{(i)})_{i=1}^{K}).$$

For each $(\mathbf{Y},\mu,\nu)$ we may also define the algorithm in reverse, i.e. starting in $\mathbf{Y}_m^{(\mu_m)}$, treating this as $\mathbf{Y}_1^{(\nu_1)}$ and proceeding as above but swapping roles for ν and μ. This analogously defines a pmf $P^B_{\theta_1,\theta_0}(\mathbf{Y},\mu,\nu)$.

It can be shown using a little algebra that the LIS estimator can be written

$$\hat{r}_{LIS}(\mathbf{Y},\mu,\nu;\theta_0,\theta_1)=\frac{P^B_{\theta_1,\theta_0}(\mathbf{Y},\mu,\nu)}{P^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)}.$$

Now, the joint distribution $\Pi^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)$ of $(\mathbf{Y},\mu,\nu)$ and $\mathbf{Y}_1^{(\nu_1)}$ simulated according to the forward algorithm is simply

$$\left(\frac{\Pi^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)}{q_{\theta_0}(\mathbf{Y}_1^{(\nu_1)})/Z_{\theta_0}}\right)\left(q_{\theta_0}(\mathbf{Y}_1^{(\nu_1)})/Z_{\theta_0}\right)=\frac{P^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)}{Z_{\theta_0}}.$$

Hence, if we take $\Pi^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)$ to be the importance distribution for drawing a sample $(\mathbf{Y},\mu,\nu)$ we see that $\hat{r}_{LIS}(\mathbf{Y},\mu,\nu;\theta_0,\theta_1)$ is an estimator of $\Lambda(\theta_0,\theta_1)$ in the sense that

$$E_{\Pi^F_{\theta_0,\theta_1}}\{\hat{r}_{LIS}(\mathbf{Y},\mu,\nu;\theta_0,\theta_1)\}=\sum_{\mathbf{Y},\mu,\nu}\frac{P^B_{\theta_1,\theta_0}(\mathbf{Y},\mu,\nu)}{P^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)}\frac{P^F_{\theta_0,\theta_1}(\mathbf{Y},\mu,\nu)}{Z_{\theta_0}}=\Lambda(\theta_0,\theta_1).$$

A.II.8. Combining Importance sampling and auxiliary variable (LISA)

The question is now whether we can improve on the performance of SISA by getting a better estimate of $\Lambda(\theta*,\theta_0)$ than the SIS with M = 1? There are a few aspects of the importance samplers presented that prevents an immediate incorporation of them in the SISA. For example, here $\overline{\Lambda}(\theta,\theta_0)\rightarrow\Lambda(\theta,\theta_0)$ only as M gets large and we have to get an estimate in every iteration. If the distributions indexed by θ are close to θ_0 are separated there could be a severe bias or infinite variance. As we have seen this can be remedied by introducing bridging distributions but in general for the importance sampler, while for the Hastings ratio

$$E_{\mathbf{Y}|\theta*}\left[\hat{H}\right]=E_{\mathbf{Y}|\theta*}\left[\frac{q_{\theta*}(\mathbf{X})}{q_{\theta^{(t)}}(\mathbf{X})}\overline{\Lambda}(\theta*,\theta^{(t)})\right]=\frac{q_{\theta*}(\mathbf{X})}{q_{\theta^{(t)}}(\mathbf{X})}\Lambda(\theta*,\theta^{(t)})=\frac{q_{\theta*}(\mathbf{X})/Z_{\theta*}}{q_{\theta^{(t)}}(\mathbf{X})/Z_{\theta^{(t)}}}$$

we typically have that

$$E_{\mathbf{Y}|\theta*}\left[\hat{H}\right]\neq E_{\mathbf{Y}|\theta*}\left[\min\left\{1,\hat{H}\right\}\right].$$

Consequently, if we use importance samplers indiscriminately we may accept updates in the Metropolis-Hastings with on average wrong probabilities.

A.II.8.a. LISA - extended state space

In SISA we performed draws from the joint distribution of the parameters and the auxiliary variable $\mathbf{Y}\in\mathscr{X}$. Consider now as an auxiliary variable $(\mathbf{Y},\mu,\nu)\in$ $\prod_{j=1}^{m}\mathscr{X}^{K}\times\left\{1,\ldots,K\right\}\times\left\{1,\ldots,K\right\}$ and a distribution $\Pi^{B}_{\theta_0,\theta}(\mathbf{Y},\mu,\nu)$ that depends on both θ and θ_0. The linked importance sampler (LISA) MCMC is a Metropolis-Hastings algorithm that performs draws from the joint distribution

$$\begin{aligned}\pi(\mathbf{Y},\mu,\nu,\theta\mid\mathbf{X})&=\Pi^{B}_{\theta_0,\theta}(\mathbf{Y},\mu,\nu)\pi(\theta\mid\mathbf{X})\\&\propto\frac{P^{B}_{\theta_0,\theta}(\mathbf{Y},\mu,\nu)}{Z_{\theta_0}}\frac{q_{\theta}(\mathbf{X})}{Z_{\theta}}\pi(\theta)\end{aligned}$$

It is straightforward to show that θ has the desired marginal distribution

$$\sum_{\mathbf{Y},\mu,\nu}\pi(\mathbf{Y},\mu,\nu,\theta\mid\mathbf{X})=\pi(\theta\mid\mathbf{X}).$$

The Hastings ratio still contains the ratio $\Lambda(\theta*,\theta^{(t)})$

$$H=\frac{q_{\theta*}(\mathbf{X})/Z_{\theta*}}{q_{\theta^{(k)}}(\mathbf{X})/Z_{\theta^{(k)}}}\frac{P^{B}_{\theta_0,\theta*}(\mathbf{Y}*,\mu*,\nu*)}{P^{B}_{\theta_0,\theta^{(t)}}(\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)})}\frac{g(\theta^{(k)},\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)}\mid\theta*,\mathbf{Y}*,\mu*,\nu*)}{g(\theta*,\mathbf{Y}*,\mu*,\nu*\mid\theta^{(k)},\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)})}\frac{\pi(\theta*)}{\pi(\theta^{(k)})}$$

where we for now denote by g a generic proposal distribution. Assume now that we conditional on $\theta*$ propose

$$(\mathbf{Y}^*,\mu^*,\nu^*) \sim \Pi^F_{\theta^*,\theta_0}(\mathbf{Y}^*,\mu^*,\nu^*) = \frac{P^F_{\theta^*,\theta_0}(\mathbf{Y}^*,\mu^*,\nu^*)}{Z_{\theta^*}},$$

the ratio $\Lambda(\theta^*,\theta^{(t)})$ cancel in H

$$\begin{aligned}
H &= \frac{q_{\theta^*}(\mathbf{X})/Z_{\theta^*}}{q_{\theta^{(k)}}(\mathbf{X})/Z_{\theta^{(k)}}} \frac{P^B_{\theta_0,\theta^*}(\mathbf{Y}^*,\mu^*,\nu^*)}{P^B_{\theta_0,\theta^{(t)}}(\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)})} \frac{\Pi^F_{\theta^{(t)},\theta_0}(\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)})g(\theta^{(k)}\,|\,\theta^*)}{\Pi^F_{\theta^*,\theta_0}(\mathbf{Y}^*,\mu^*,\nu^*)g(\theta^*|\,\theta^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})} \\
&= \frac{q_{\theta^*}(\mathbf{X})/Z_{\theta^*}}{q_{\theta^{(k)}}(\mathbf{X})/Z_{\theta^{(k)}}} \frac{P^B_{\theta_0,\theta^*}(\mathbf{Y}^*,\mu^*,\nu^*)}{P^B_{\theta_0,\theta^{(t)}}(\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)})} \frac{P^F_{\theta^{(t)},\theta_0}(\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)})/Z_{\theta^{(k)}}g(\theta^{(k)}\,|\,\theta^*)}{P^F_{\theta^*,\theta_0}(\mathbf{Y}^*,\mu^*,\nu^*)/Z_{\theta^*}g(\theta^*|\,\theta^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})} \\
&= \frac{q_{\theta^*}(\mathbf{X})}{q_{\theta^{(k)}}(\mathbf{X})} \frac{P^B_{\theta_0,\theta^*}(\mathbf{Y}^*,\mu^*,\nu^*)}{P^F_{\theta^*,\theta_0}(\mathbf{Y}^*,\mu^*,\nu^*)} \frac{P^F_{\theta^{(t)},\theta_0}(\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)})g(\theta^{(k)}\,|\,\theta^*)}{P^B_{\theta_0,\theta^{(t)}}(\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)})g(\theta^*|\,\theta^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})}
\end{aligned}$$

and furthermore, by the definition of $\hat{r}_{LIS}(\mathbf{Y},\mu,\nu;\theta^*,\theta_0)$

$$\hat{r}_{LIS}(\mathbf{Y},\mu,\nu;\theta^*,\theta_0) = \frac{P^B_{\theta_0,\theta^*}(\mathbf{Y},\mu,\nu)}{P^F_{\theta^*,\theta_0}(\mathbf{Y},\mu,\nu)}$$

so that the Hastings ratio reduces to

$$H = \frac{q_{\theta^*}(\mathbf{X})}{q_{\theta^{(k)}}(\mathbf{X})} \frac{\hat{r}_{LIS}(\mathbf{Y}^*,\mu^*,\nu^*;\theta^*,\theta_0)}{\hat{r}_{LIS}(\mathbf{Y}^{(t)},\mu^{(t)},\nu^{(t)};\theta^{(t)},\theta_0)} \frac{g(\theta^{(k)}\,|\,\theta^*)}{g(\theta^*|\,\theta^{(k)})} \frac{\pi(\theta^*)}{\pi(\theta^{(k)})}.$$

Note that SISA can be seen as a special case of LISA with K = 1 and m =1, i.e. when we only produce the initial state $\mathbf{Y}_1^{(\nu_1)}$ and when we have no bridging states.

A.II.8.b. LISA in summary

We have proposed an exact or ‘‘pure’’ Metropolis-Hastings algorithm for drawing paramters from the posterior given data that follows a distribution with an intractable normalising constant. The MCMC is ‘‘pure’’ as long as we are able to draw (an approximate) a sample point from our data model.

The algorithm LISA is tunable having two constants K and m that may be set by the researcher to tune the mixing of the Markov chain. This is a notable improvement on SISA, for which there was very little scope for improving mixing. Since LISA employs bridging distributions it is not as sensitive to the choice of θ_0 as SISA.

Although LISA may give the impression of being complicated to implement and require a lot of extra computational time as compared to SISA, LISA only requires sampling in the data space using M-H and the evaluation of discrete variable probabilities. LISA only requires $K \times m$ extra M-H updating step (for the auxiliary variable) as compared to SISA. Note that a procedure for sampling in the data space is almost always required for the models considered here in order to make any sort of inference.

A.II.9. Illustrations

In order to illustrate the performance of LISA, we compare LISA and SISA in inference for simulated data from the Ising model as well as employ LISA to the inference for models and data with more elaborate interdependence structure.

The Ising model is used for illustration since it is a well known model that is reasonably well understood. In addition, since the Ising model was used for illustration in Møller et al. (2005) it makes for an intuitive point of departure. Having established that LISA compares favourably with SISA in the case of the Ising model we proceed to investigate its performance for models with more intricate dependence structure. The social influence model proposed by Robins, et al. (2001) is very similar to the Ising model but the dependence structure is more complicated and less regular. Because of this the influence of auxiliary distribution and the choice of *K* and *m* on the mixing of LISA are more accentuated in the case of the social influence model. Finally, we shall illustrate the algorithm for inference for the (Curved) exponential family of distributions.

A.II.9.a. Ising model on 50×50 grid

A classic case of an autologistic model is the Ising model (Besag, 1972; Cressie, 1993). It is assumed that you have points on grid with binary marks. The Ising model on a binary $m \times n$ lattice has been used to model how the charges of particles interact an in the simplest case it is assumed that the particles can have either of two spins, up or down. The spin of a given particle depends on the general tendency towards spin up and the spins of its neighbours on the lattice. We define the model for $\mathbf{X} = (x_{ij} : i = 1, \ldots, m, \text{and} , j = 1, \ldots, n)$, where for the elements $x_{ij} \in \{-1,1\}$. The pmf is defined as in (E-1) with

$$q_\theta(\mathbf{X}) = \exp(\theta_0 V_0 + \theta_1 V_1),$$

where

$$V_0 = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \text{ , and } V_1 = \sum_{i=1}^{m-1} \sum_{j=1}^{n} x_{ij} x_{(i+1)j} + \sum_{i=1}^{m} \sum_{j=1}^{n-1} x_{ij} x_{i(j+1)} .$$

When generating data according to the model we have to rely on MCMC since there is no direct way drawing data from an Ising model. In the case of some autologistic models (and the Ising model in particular), it is possible to ‘‘sample perfectly’’ from the model, to take as an output a state that we know have been produced after the Markov chain has converged to the target distribution. Here we have used Wilson's (2000) modification to the Propp and Wilson (1996) algorithm.

A realisation for a 50×50 lattice with $\theta = (0,0.3)^{\mathrm{T}}$ is given in Figure 7.

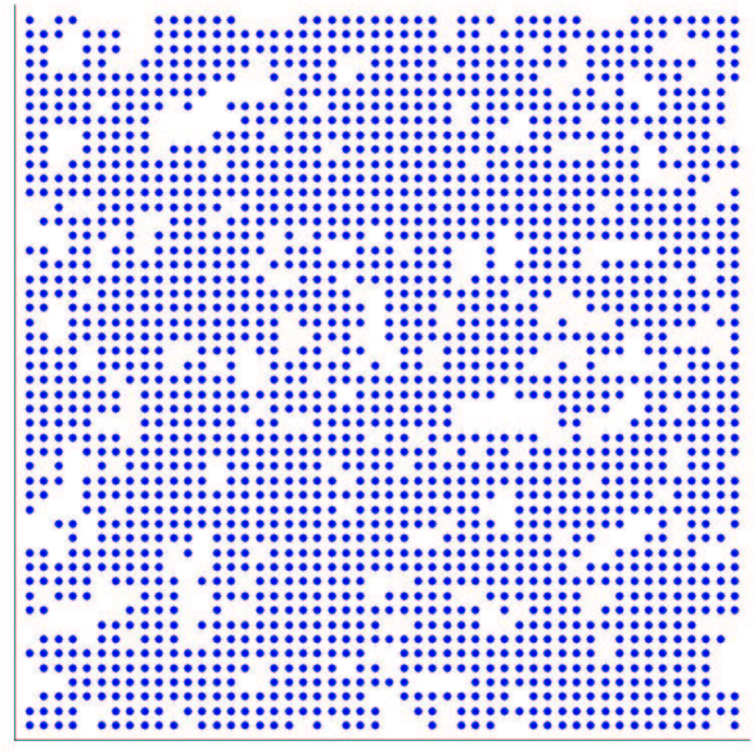

Figure 7 A realisation for an Ising model on a binary 50×50 lattice with $\theta = (0,0.3)^{\mathrm{T}}$. ''Spin-ups'' are indicated by dots.

To illustrate how the chains $\mathbf{Y}_j^{(1)},\ldots,\mathbf{Y}_j^{(K)}$ are connected for $j = 1,\ldots,m$, we have plotted the sufficient statistics V_0 and V_1 for m = 5 chains in Figure 8. In the right-hand panel we see that $\theta(1)$ and $\theta(5)$ produce radically different numbers of same-site (V_1) pairs. The bridging distributions make it possible for the sampler to incorporate values of V_1 that are probable under $\theta(5)$ but highly unlikely under $\theta(1)$. The starting point $\mathbf{Y}_1^{(\nu_1)}$ is generated (using the Prop Wilson algorithm) from an Ising model defined by $\theta = (0,0.3)^{\mathrm{T}}$. For the right hand panel, showing the traces of (V_1), we expect the chains to progressively move downwards since the parameter $\theta(1)$ corresponding to the number of same spinn sites is gradually lowered. The state connecting the first chain with the second is $\mathbf{Y}_1^{(\mu_1=550)}$, with V_1=1716. This state is then set as the starting state $\mathbf{Y}_2^{(\nu_2=10)}$, in the second chain etc until the last chain is started in $\mathbf{Y}_5^{(\nu_5=714)}$, whose V_1=1450 is considerably lower than the overall level of the number of same spinn sites in the first chains. Thus the bridging chains have managed to link the supports of the two extreme distributions defined by $\theta(1)$ and $\theta(5)$.

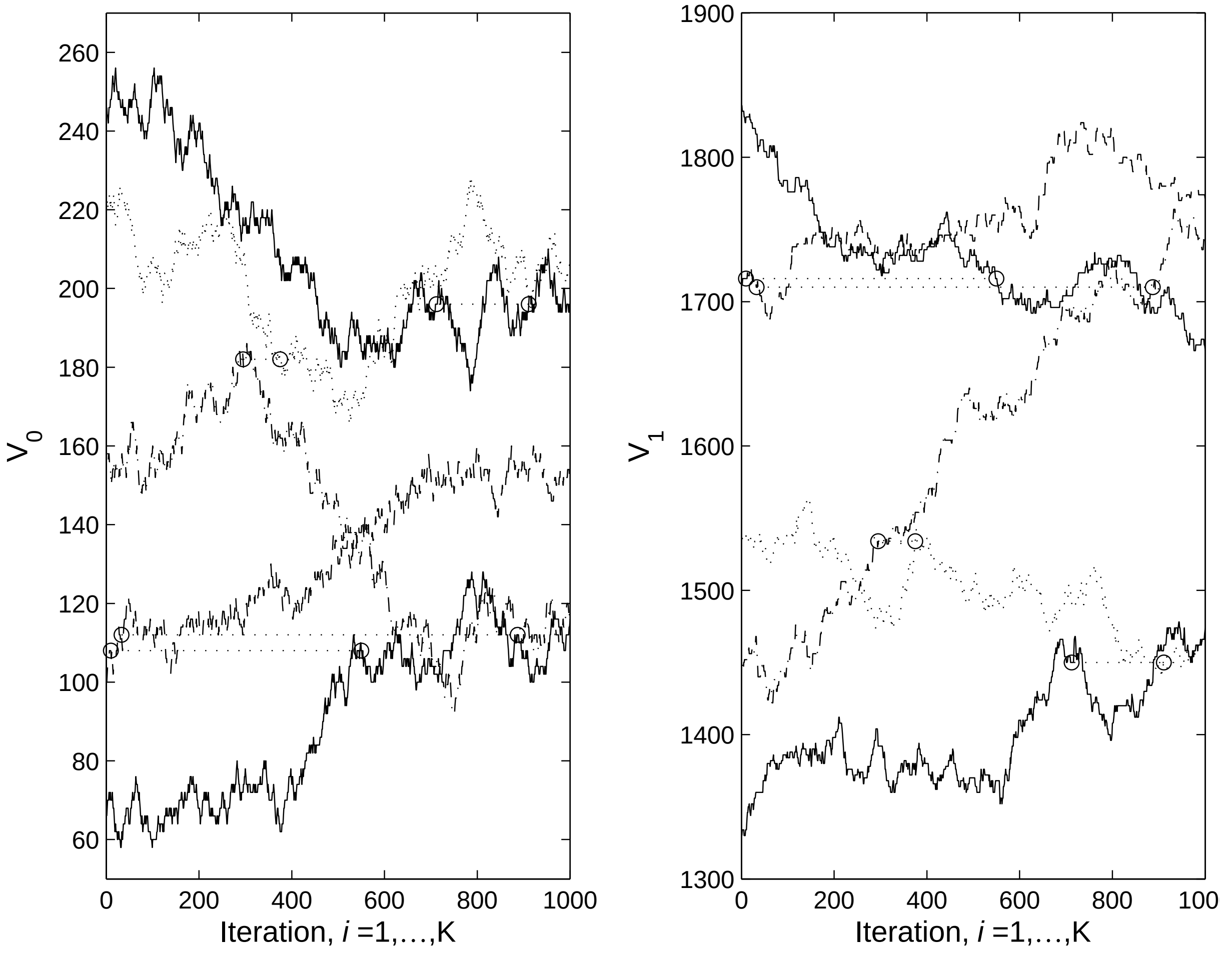

Figure 8 A linked importance sample for an Ising model on a binary 50×50 lattice where $\theta(1) = (0,0.3)^{\mathrm{T}}$ and $\theta(5) = (0.1,0.2)^{\mathrm{T}}$. Circles mark the linking states.

Since $\hat{r}_{LIS}(\mathbf{Y}, \mu, \nu; \theta^*, \theta_0)$ has lower variance than the SIS estimate, we expect the variation in LISA to be smaller than in SISA. Additionally, since SIS is biased when the proposed parameter value is far from θ_0, we expect SISA to get stuck in bad estimates of the ratio of normalising constants $\Lambda(\theta^*, \theta_0)$ occasionally. When these two issues combine we expect them to manifest themselves in a low acceptance rate and hence large sample autocorrelations even for big lags.

To investigate differences in performance we simulated data using $\theta = (0.2,0.1)^{\mathrm{T}}$ and $\theta = (0,0.3)^{\mathrm{T}}$ and binary 50×50 lattice. In Table 1 some summaries for LISA with different choice *K* and *m* are given. In all algorithms the MPLEs have been used as θ_0. Increasing *K* and *m* drastically reduces the sample autocorrelations of the Markov chain. The lag 50/100 SACF efficiency is a measure of the gain in efficiency scaled by the number of extra iterations required to calculate the LIS estimate. Note that this is conservative in favour of SISA since one Metropolis-Hastings updating step in the LIS algorithm corresponds to one function evaluation (the change in sufficient statistics as one element is change into its oposite) but one iteration in the burn-in phase required in the perfect sampling scheme for drawing $\mathbf{Y}_1^{(\nu_1)}$ requires 5 function evaluations (the maximal and minimal chain in the main chain and the criterion chain respectively plus the update of the present state). The number of extreme proposals as judged by Prop. min(1,H) < e^{-10} deserves closer attention. This reflects, not that some proposed parameter values hade a very low posterior probability, rather it reflects the bias and variation of the SIS in SISA. In the applications to come low acceptance probabilities sometimes cause the SISA to become stuck in some states for a long time.

	K = 1		*K* = 3000		*K* = 7000	
	m = 1		*m* = 5		*m* = 9	
	θ_0.	θ_1.	θ_0.	θ_1.	θ_0.	θ_1.
True	0.2	0.1	0.2	0.1	0.2	0.1
MPLE.	0.196	0.109	0.196	0.109	0.196	0.109
Proposal std	0.005	0.005	0.005	0.005	0.005	0.005
Posterior mean	0.2006	0.1053	0.1994	0.1054	0.1994	0.1051
Posterior std	0.0214	0.0142	0.0217	0.0144	0.0221	0.0149
Lag 50 SACF	0.6464	0.5124	0.5262	0.3619	0.4596	0.3126
Lag 100 SACF	0.4336	0.3288	0.2923	0.1945	0.2011	0.1088
Mean acc. prob	0.3931		0.5890		0.7249	
Prop. min(1,H) $< e^{-10}$	0.0178		0.0006		0.0000	
Ave. iter	5.6×10^4		5.6×10^4		5.6×10^4	
Lag 50 SACF efficiency	1	1	1.0558	1.0310	0.7185	0.6627
Lag 100 SACF efficiency	1	1	0.9844	0.9455	0.6631	0.6241

Table 1 Comparison of performance for different choices of *K* and *m* in LISA for estimating parameters for the Ising model (Lag 50/100 efficiency is 1-SACF/(Ave.iter+*Km*) relative to *K* =1 and *m* =1)

In Figure 9 the autocorrelations of the entire MCMC samples for SISA and LISA (*K* = 3000, *m* = 5 and *K* = 7000, *m* = 9) are compared. The increase in efficiency when *K* = 3000 and *m* = 5, as compared to SISA is substantial but the marginal increase when we increase *K* to 7000 and *m* to 9 is smaller.

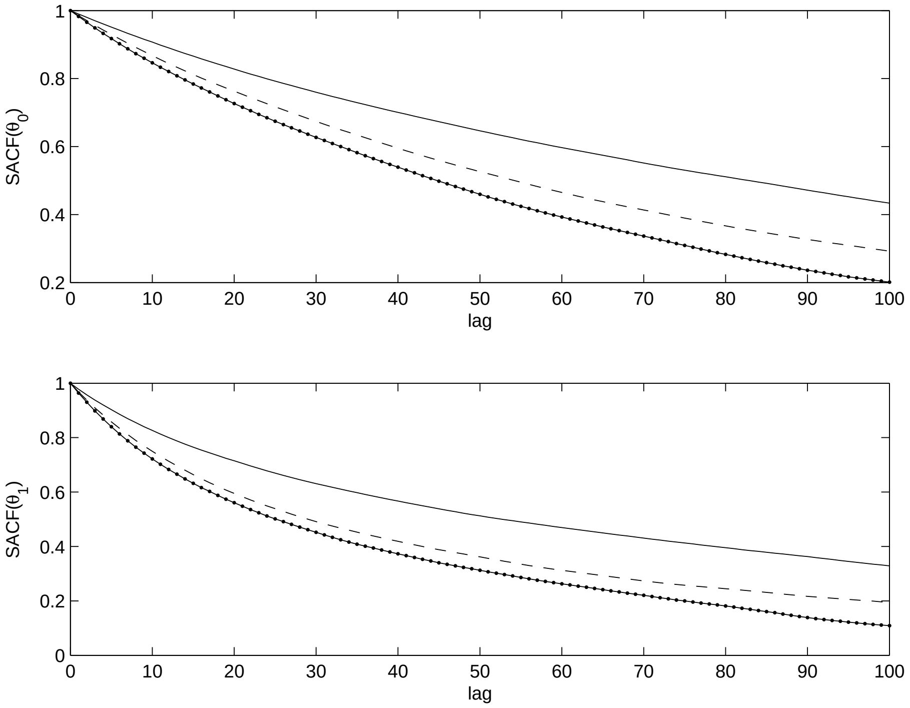

Figure 9 Comparing the sample autocorrelation functions (SACF) for SISA and LISA(dashed *K* = 3000, *m* =5; dotted *K* = 7000, *m* = 9) for data simulated from an Ising model on a binary 50×50 lattice with $\theta = (0.2, 0.1)^{\mathrm{T}}$.

A.II.9.b. The social influence model

When studying for example binary educational outcomes it is common to take interdependence between response variables as a consequence of respondents sharing teachers, schools, etc, into account using random effects. Many researchers have however pointed to the importance of taking peer influence into account. Robins, et al. (2001) proposed a model that uses the empirically collected interaction structure to model how outcomes for (for example) pupils may depend on the outcomes for their friends. There are some obvious similarities to Ising but the interdependence structure is given by empirical observations, not homogeneous and usually quite complicated.

For a set of actors $N = \{1, \dots, n\}$, we let the binarised masculinity variable (0=gender equity attitudes, 1=male dominance attitudes; for description of data and substantive motivation see Lusher, 2006) be the response variable $\mathbf{X} = (x_i)_{i \in N}$ in a model for which the pmf is defined as in (E-1) with

$$q_\theta(\mathbf{X}) = \exp\left\{\theta_1 \sum_{i \in N} x_i + \theta_2 \sum_{j \in N} \sum_{i \in N} x_i y_{ij} + \theta_3 \sum_{j \in N} \sum_{i \in N} x_i x_j y_{ij} + \sum_{k=4}^{7} \theta_k \sum_{i \in N} x_i z_{ik-3}\right\}$$

where z_{i1} is a binary covariate capturing whether actor *i* belongs to the dominant culture (1=dominant Anglo-Australian ethno-cultural background; 0=marginal ethno-cultural background); z_{i2} is the SES of actor *i* based on postcode (here standardised); z_{i3} is fathers's occupational status of actor *i* (standardised);); z_{i4} is mother's occupational status of actor *i* (standardised); x_{ij} is 1 if either *i* nominates *j* as a friend or *j* nominates *i* as a friend.

Trace plots for the parameters in this model are given in Figure 10 and summaries and comparisons with the MLEs and MPLEs are given in Table 2. The importance of the choice of θ_0 and *K* and *m* is visible in how freely the (3 first) parameters move in the state space varies. It is clear from the marginal histograms that the sample using MPLE, *K* = 1 and *m* =1 is not very useful. As *K* and *m* increase, the mixing gradually improves and for *K* = 2000 and *m* = 7 the chain moves well in the state space (possibly with the exception of a run around iteration 40,000). When the MLE is used in the auxiliary distribution there is a marked improvement for all algorithms but SISA displays the characteristic ‘‘freezing’’ - SISA seems to be mixing perfectly well for the first 20,000 or so iterations, after which it gets stuck for more than 50,000 iterations.

For the influence model, it is not straightforward to construct a monotonic chain such that we may implement a perfect sampling scheme. Instead we have relied on the rule of thumb 100*n* for the length of the burn as is suggested by Snijders (2002). A number of post-hoc tests (based on simulation in the data space with relatively ‘‘extreme’’ values of θ) confirm that this burnin is sufficient.

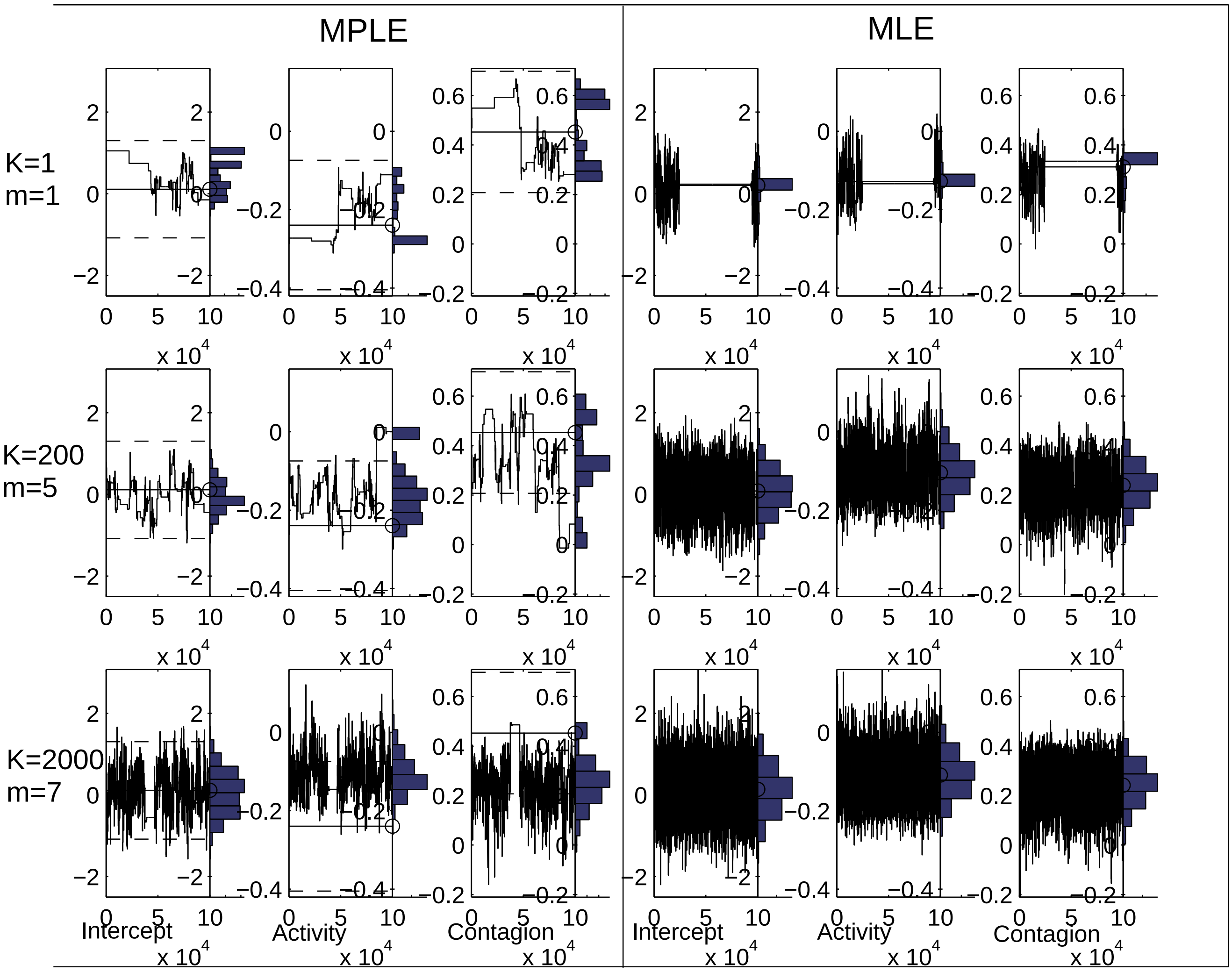

Figure 10 Comparing effect of choice of auxiliary distribution and tuning parameters in LISA: Trace plots (and histograms) for 3 of the parameters in an influence model fitted to Lusher's (2006) 106 school data when auxiliary distribution in LISA uses MPLE (left hand panels; value indicated by solid line and approximate 95 confidence interval by dashed lines) and MLE (right hand panels; value indicated by solid line).

The (approximate) confidence interval for θ_3 and θ_2 given by the pseudo likelihood analysis and maximum likelihood suggests that the corresponding two effects are significant. The (exact) Bayesian analysis is however less conclusive (activity does not seem to have an effect) but still lending some support to contagion: the posterior probability that $\theta_3 > 0$ given data, i.e. that there is a contagion effect, is 0.9932. No contaigion effect is not incluđced in the 95 highest posterior density region (95 HPD) but in the 99 HPD. Point estimates and measure of uncertainty are very similar for the ML approach and the Bayesian approach and the differences are largely to be attributed to the skewness of the posteriors.

		MPLE		MCMCMLE				Posterior			
		EST	SE	EST	SE	MEAN	STD	95 HPD		99 HPD	
Intercept	θ_1	0.11	0.596	0.12	0.504	0.14	0.560	-1.15	1.45	-1.42	1.81
Activity	θ_2	-0.24	0.083	-0.13	0.046	-0.11	0.055	-0.23	0.02	-0.26	0.06
Contagion	θ_3	0.45	0.123	0.29	0.067	0.24	0.082	0.04	0.42	-0.03	0.47
Dominant culture	θ_4	-0.03	0.472	-0.39	0.427	-0.44	0.451	-1.44	0.57	-1.74	0.82
SES	θ_5	0.10	0.227	0.20	0.215	0.23	0.223	-0.27	0.72	-0.38	0.83
Dad	θ_6	-0.16	0.220	-0.17	0.210	-0.19	0.219	-0.69	0.30	-0.83	0.42
Mum	θ_7	-0.05	0.224	0.08	0.206	0.08	0.215	-0.41	0.57	-0.54	0.71

Table 2 Point estimates for influence model fitted to Lusher's (2006) 106 school data

A.II.9.c. An ERGM

While the interaction pattern of the sites in an Ising model is described by a binary $m \times n$ lattice, that is a regular graph (with a difference in the degrees of boundary vertices), which implies certain convenient conditional independencies, there are models that have a much more complicated dependence structure. An example of this is the exponential family random graph (ERGM) distributions for social networks introduced by Frank and Strauss (1986) and further extended in for example : Pattison & Wasserman (1999); Robins, Pattison, and Wasserman (1999); Wasserman & Pattison (1996); Snijders et al. (in press).

We begin by fitting a special case of an ERGM where we model a collaboration network on a set of 36 actors. We let the set of actors be represented by a set $N = \{1, \ldots, n\}$ of vertices and let the colaboration network be represented by a random edge set on *N* with adjacency matrix **X**, the elements of which are

$$x_{ij} = \begin{cases} 1 & \text{if actor } i \text{ collaborates with actor } j \\ 0 & \text{otherwise} \end{cases}.$$

We assume that it is not meaningful to speak of actors collaborating with themselves wherefore the main diagonal is all zeros. The pmf is defined as in (E-1) with

$$q_\theta(\mathbf{X}) = \exp(\sum\nolimits_{k=1}^{p} \theta_k z_k(\mathbf{X}))\,,$$

where the z_k´s are functions of the adjacency matrix and a set of fixed vertex level attributes. In this particular example we follow the specification in Hunter and Handcock (2005) and for the attributes a_i (seniority of actor i in terms of rank), b_i (binary indicator of role of actor i), c_i (sex of actor i), d_i (office location of actor i), (for more details about the model and the data set see e.g. Lazega, 2001; Lazega and Pattison, 1999; and Snijders et al., in press):

k	Effect	$z_k(\mathbf{X})$
1	Activity/popularity	$\sum_{i<j} x_{ij}$
2	Main effect of seniority	$\sum_{i<j} x_{ij}(a_i + a_j)$
3	Main effect of practice	$\sum_{i<j} x_{ij}(b_i + b_j)$
4	Homophily practice	$\sum_{i<j} x_{ij}\mathbf{1}(b_i = b_j)$
5	Homophily sex	$\sum_{i<j} x_{ij}\mathbf{1}(c_i = c_j)$
6	Homophily office	$\sum_{i<j} x_{ij}\mathbf{1}(d_i = d_j)$

Here **1** denotes the indicator function.
The sociogram of the data is depicted in Figure 11.

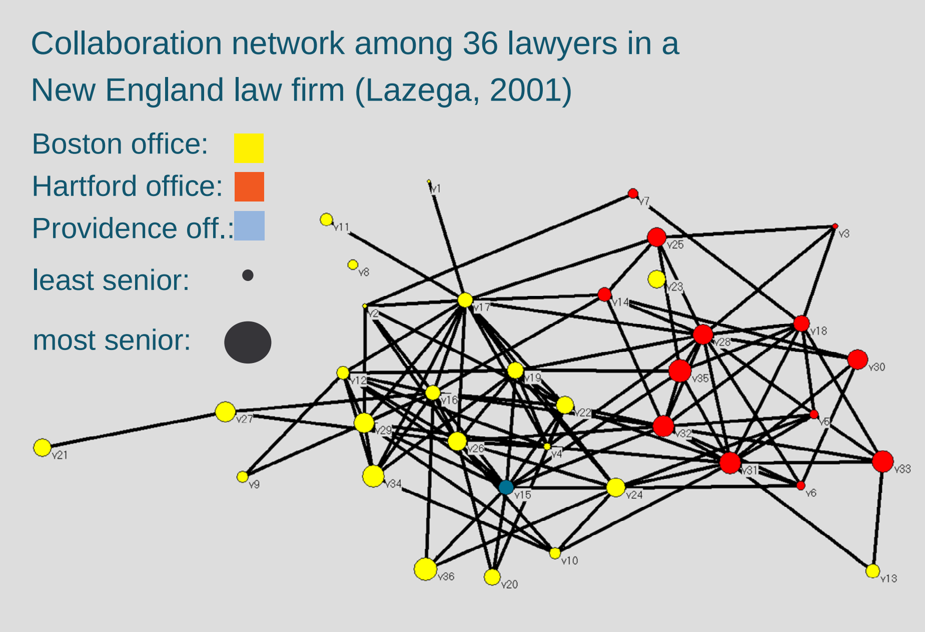

Figure 11 Lazega's lawyers

Since the edges are conditionally independent conditional on the attributes and parameters, it is easy calculate the normalising constant analytically as

$$Z_\theta = \prod_{i<j} \left\{1 + \exp\left[\sum\nolimits_{k=1}^{p} \theta_k (z_k(\mathbf{X}_{ij}^{+}) - z_k(\mathbf{X}_{ij}^{-}))\right]\right\}$$

for this model. In this expession $\mathbf{X}_{ij}^{+}$ is the adjacency matrix that is identical to $\mathbf{X}$ for all elements but may differ in (i, j) that is set to 1. Analogously $\mathbf{X}_{ij}^{-}$ is the adjacency matrix that is identical to $\mathbf{X}$ for all elements but may differ in (i, j) that is set to 0.

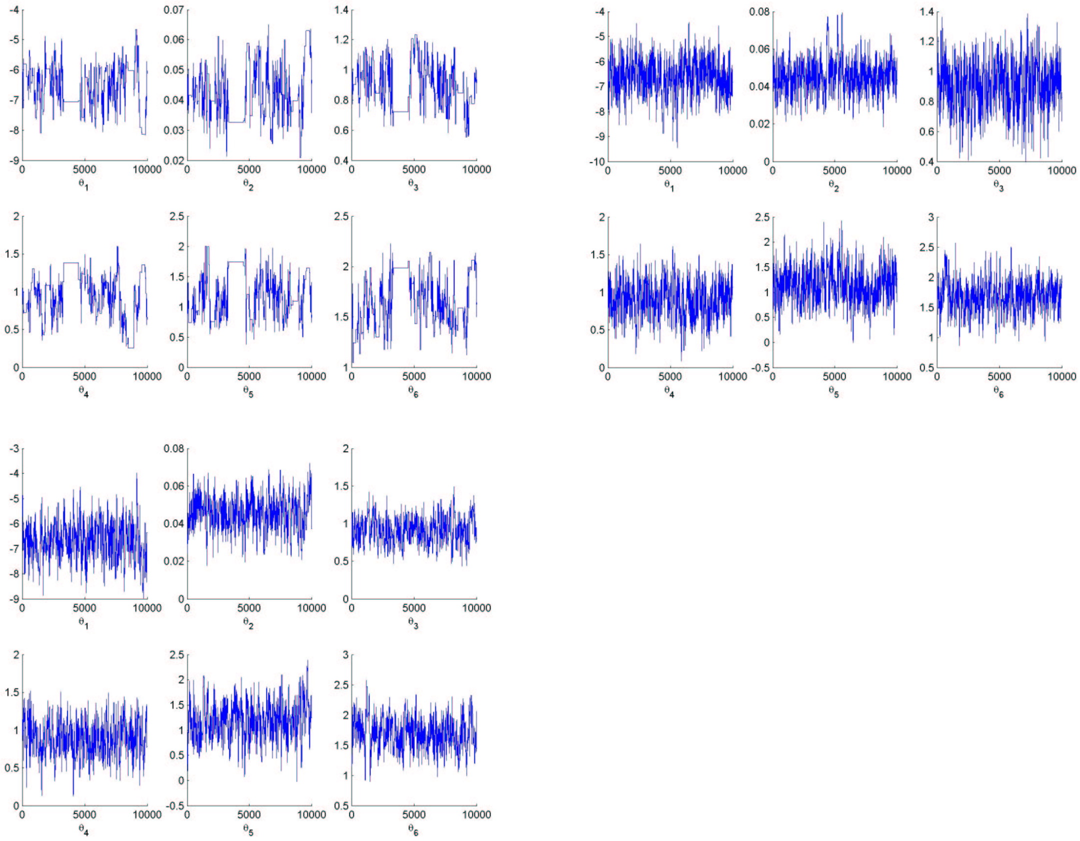

Figure 12 Trace plots for (from left to right, top to bottom) SISA, LISA(*K* = 1000, *m* =5), LISA(*K* = 1000, *m* = 10) for a dyad independent ERGM fittend to Lazega's (2001) New England Lawyers collaboration network.

As seen in Figure 12 the mixing improves markedly with LISA (here we will include a table with estimates, SACF, etc). What is more interesting is that since can evaluate the normalising constant analytically, we are able to study the bias in SISA stemming from the biasedness of SIS.

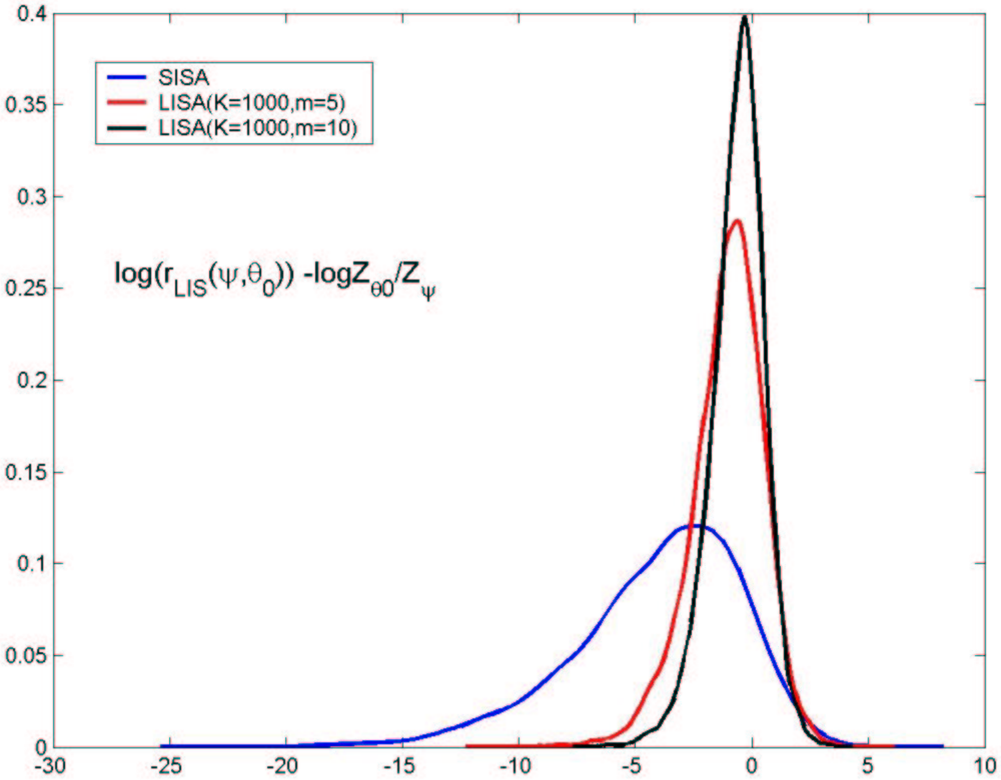

Figure 13 Distribution of bias for SISA, LISA(*K* = 1000, *m* =5), and LISA(*K* = 1000, *m* = 10) for a dyad independent ERGM fittend to Lazega's (2001) New England Lawyers collaboration network.

Figure 13 is an illustration of the difference between the true $\Lambda(\theta^*,\theta_0)$ and that estimated by SIS and LIS. The distributions are those of $\log \overline{\Lambda}(\theta^*,\theta_0)/\Lambda(\theta^*,\theta_0)$ under the marginal distribution of θ^*

$$\int g(\theta^* \mid \theta)\pi(\theta \mid \mathbf{X})\mathrm{d}\theta \, .$$

The worrying feature about SISA is the difference often is extremely small, meaning that in this particular example the ratio $\Lambda(\theta^*,\theta_0)$ is underestimated. When a θ^* is accepted by SISA because $\Lambda(\theta^*,\theta_0)$ is underestimated is when the algorithm gets stuck.

A.II.9.d. A CERGM

We now proceed to show how LISA performs when we increase the complexity of the dependence structure on data. Following Hunter and Handcock (2005) we introduce the geometrically weighted shared partner statistic (GWEPS), which was derived as the alternating triangle statistic

$$3t_1(\mathbf{X}) - \frac{t_2(\mathbf{X})}{\lambda^1} + \cdots + (-1)^{n-3}\frac{t_{n-2}(\mathbf{X})}{\lambda^{n-3}} \, .$$

from the partial dependence assumption by Snijders et al., (in press). When λ in the alternating triangle statistic is allowed to be a free parameter the model containing both GWEPS and

$\theta_8 = \log\lambda$ belongs to the curved exponential family of distributions. The results for fitting the model to Lazega's (2001) New England Lawyers collaboration network using LISA are illustrated in Figure 14. The visual impression is one of good mixing.

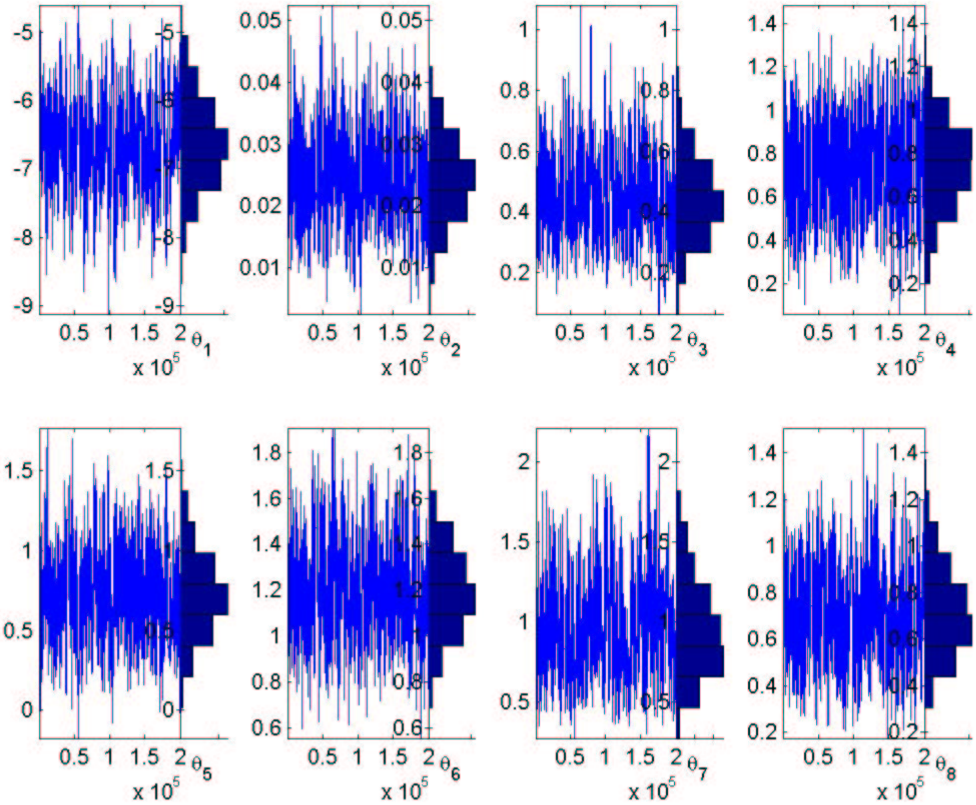

Figure 14 Trace plots with histograms for a model with GWEPS fitted to Lazega's (2001) New England Lawyers collaboration network with LISA(K = 2000, m = 7)

A.III. Bayesian analysis of (curved) exponential family distributions for graph

We consider here a probability model for the edge set of a graph that is commonly referred to as the exponential random graph model (ERGM), and its extension, the curved ERGM. Although some issues remains to be resolved when it comes to how to specify the ERGM, this class of models holds some promise when it comes to capturing network processes. Currently the favoured methods for statistical inference are Markov chain Monte Carlo (MCMC) Maximum likelihood estimate (MLE) and an MCMC implementation of the Robbins-Monroe algorithm, both of which rely on the properties of the method of moments for exponential family distributions. We propose instead to take a Bayesian approach that (i) yields clearly defined answers in terms of probabilities (the asymptotic properties of the MLE are not fully understood in the case of the ERGM); (ii) offers a rich picture of uncertainty (the MLEs and approx. s.e.'s do not adequately reflect the uncertainty stemming from the pronounced dependencies between observations); (iii) makes allowances for penalising "degenerate parts" of the parameter space using proper subjective prior distributions; (iv) provides us with a natural and probabilistic approach for handling missing data; (v) offers a principled and probabilistic procedure for performing model selection; (vi) provides us with posterior predictive distributions; etc. How to implement a Bayesian inference scheme for the ERGM is, however, far from straightforward.

It is clear that in all but trivial cases we have to rely on numerical methods. It is probably fair to say that as far as numerical methods go, MCMC is the gold standard. Thus far, however, efforts at designing an MCMC algorithm for the ERGM has been hampered by the fact that it is typically not possible to evaluate the normalising constant (the partition function) in the likelihood function. Although the (pure) MCMC does not require that we can evaluate the normalising constant in the posterior distribution it usually requires that we can evaluate the likelihood function. Recently an auxiliary variable MCMC (SISA; our acronym) was proposed that circumvented the need to evaluate the partition function. The key being to introduce an auxiliary variable defined on the same state space as data. However, while SISA performs sufficiently well in order for it to be useful for "simpler" models like the Ising model, it seems as if it runs into serious problems when applied to the ERGM. It is not only a question of whether the mixing is good or not, rather it is a question of whether it mixes at all. The reasons for this being so are easily understood when the SISA is understood in terms of the Simple Importance Sampler (SIS). We propose a solution (LISA) where the (single) auxiliary variable is replaced by an auxiliary variable defined on an extended state space. Whereas SISA may be seen as an algorithm that performs a one-sample point SIS in each iteration of the Metropolis-Hastings sampler, LISA performs a bridged (linked) importance sampling (LIS) estimation in each iteration, with the number of bridging distributions and sample points chosen to tune mixing. The extra number of calculations necessary to perform LISA as compared to the SISA is negligible. We illustrate LISA when applied to the analysis of the Ising on a 50x50 grid and a network for a New England law firm.

A.III.1. Scope of the paper

This paper will use the results on LISA in Koskinen (2006; and paper outlined in previous section) and present them to in a less technical way with an emphasis on the application to social network analysis. We will go into more detail regarding specific research issues that arise in social network analysis and in particular when exponential and curved exponential family distributions are fitted to sociometric data. Much effort will be put on interpretation of results and in doing this we will treat in some detail posterior predictive distributions in order to interpret models in terms of observables. This approach also extends the alternative goodness of fit that is proposed in Koskinen et al. (2007b).

A.III.2. Comparison of different approaches

A.III.2.a. An MCMC importance sampler

Koskinen, (2004a) proposes to use a Bayesian version of the MCMC scheme of Geyer and Thompson (1992). For multivariate ERGMs, Koskinen and Robins (2007) suggested a similar approach. For drawing a sample $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(h)},\ldots,\boldsymbol{\theta}^{(H)}$ with "non-informative" prior $\pi(\boldsymbol{\theta})$=1, from the posterior distribution

$$\pi(\boldsymbol{\theta} \mid \mathbf{X}) \propto p(\mathbf{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \frac{\exp\left[\boldsymbol{\theta}^{\mathrm{T}}\mathbf{z}(\mathbf{X})\right]}{\sum_{\mathbf{U}\in\mathscr{X}}\exp\left[\boldsymbol{\theta}^{\mathrm{T}}\mathbf{z}(\mathbf{U})\right]},$$

they used the Metropolis-Hastings sampler suggested for inference for exponential random graphs in Koskinen (2004a) but with an alternative method for calculating the Hastings ratio in the updating steps for the parameters. Similar to the Maximum likelihood inference schemes suggested in Geyer and Thompson (1992) as applied to exponential random graphs (and curved exponential family models for networks, Hunter and Handcock, 2006), a single importance sample $\mathbf{Y}_1,\ldots,\mathbf{Y}_N$ is drawn from an exponential random graph model conditional on a provisional point estimate $\boldsymbol{\theta}_0$, that is used for approximating the Hastings ratio

$$\begin{aligned}
&\exp\left[(\boldsymbol{\theta}^* - \boldsymbol{\theta}^{(h)})^{\mathrm{T}}\mathbf{z}(\mathbf{X})\right]\frac{\sum_{g=1}^{N}\exp\left[(\boldsymbol{\theta}^{(h)} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{z}(\mathbf{Y}_g)\right]}{\sum_{g=1}^{N}\exp\left[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{z}(\mathbf{Y}_g)\right]} \\
&\doteq \frac{\exp\left[\boldsymbol{\theta}^{*\mathrm{T}}\,\mathbf{z}(\mathbf{X})\right]}{\exp\left[\boldsymbol{\theta}^{(h)\mathrm{T}}\mathbf{z}(\mathbf{X})\right]}\left[\frac{\sum_{\mathbf{U}\in\mathscr{X}}\exp\left[\boldsymbol{\theta}^{(h)\mathrm{T}}\mathbf{z}(\mathbf{U})\right]}{\sum_{\mathbf{U}\in\mathscr{X}}\exp\left[\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{z}(\mathbf{U})\right]}\right]\Bigg/\left[\frac{\sum_{\mathbf{U}\in\mathscr{X}}\exp\left[\boldsymbol{\theta}^{*\mathrm{T}}\,\mathbf{z}(\mathbf{U})\right]}{\sum_{\mathbf{U}\in\mathscr{X}}\exp\left[\boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{z}(\mathbf{U})\right]}\right] \\
&= \frac{\pi(\boldsymbol{\theta}^* \mid \mathbf{X})}{\pi(\boldsymbol{\theta}^{(h)} \mid \mathbf{X})},
\end{aligned}$$

where the equality $\doteq$ holds in the limit as N gets large (given some regularity conditions and caveats, Handcock, 2003). The importance sampler parameter $\boldsymbol{\theta}_0$ may have to be updated a few times and the algorithm repeated with the updated value of $\boldsymbol{\theta}_0$. An alternative approach is suggested in Berthelsen and Møller (2003) where estimates of the normalising constant for different parameter values on a grid are used in the Metropolis-Hastings. This does however require that the grid points are chosen beforehand which could prove tricky when the number of parameters is large.

In Table 3 three models with varying degrees of interdependence are fitted to the marriage and business networks of Padgett's (Padgett and Ansell, 1993) 15'th century Florentine families (multirelational stars with parameters σ_{b}, σ_{d} and σ_{c} corresponds to the structures b, d and c in Figure 15 respectively and analogously for triangles *e* and *f*; Marriage ties and Business ties taking the role of dashed and full lines respectively)

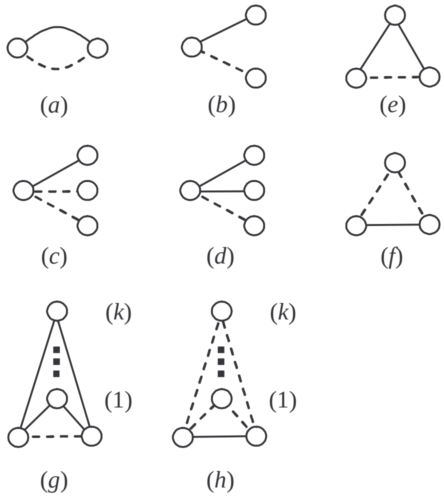

Figure 15 Some multirelational graph statistics

	Conditional		Multivariate dyad		Multivariate higher order	
	$\hat{\theta}_{Bayes}$	SD(θ)	$\hat{\theta}_{Bayes}$	SD(θ)	$\hat{\theta}_{Bayes}$	SD(θ)
Edges business, θ_1	-3.69	0.581	-3.31	0.673	-4.37	1.020
Alt. tri business, θ_2	0.87	0.298	0.68	0.361	1.42	0.535
Edges marriage, θ_3			-2.03	0.450	-2.18	0.532
Alt. tri marriage, θ_2			-0.11	0.299	-0.003	0.353
(a) Edge sim τ	2.32	0.655	2.05	0.738	3.01	0.880
Stars σ_b					0.17	0.235
Stars σ_d					-0.05	0.073
Stars σ_c					-0.04	0.081
Triangles τ_e					-0.63	0.722
Triangles τ_f					-0.08	0.666

Table 3 Summaries of posterior distributions of parameters in the ERGMs fitted to the business and marriage networks of Padgett's Florentine families. The lables (b) through (f) refer to respective subgraph counts of Figure 3. Point estimates are the MCMC estimators of the posterior expected value of the corresponding parameter given data.

The appropriateness of this approximation to the posterior distribution is something that arguable has to be decided on a case to case basis. We have seen in the treatment of LISA that the choice of importance distribution for ERGMs may have a large impact on the estimate of the estimated value of the likelihood function.

A.III.2.b. Laplace approximation

Koskinen, Wang, Lusher, and Robbins (2007) approximate the posterior distribution by a multivariate normal distribution $\boldsymbol{\theta} \mid \mathbf{X} = N_p(\hat{\boldsymbol{\theta}}_{MLE}, I(\hat{\boldsymbol{\theta}}_{MLE})^{-1})$, where the MLE and information matrix are readily available from standard SNA computational packages. The chief aim of the approximation was to supply a quick and easy procedure for drawing adjacency matrices from the (approximate) posterior predictive distribution (to be used for goodness of fit).

Some preliminary results suggest that the normal approximation might not be bad contrary to what one would expect from the high degree of interdependency and the dichotomous data (c.p. Hauck and Donner, 1977; Zellner. and Rossi, 1984). However, there are instances when the departure from normality may be small (in, say, covariance norm) but have great impact on the properties of the distribution. One property of the distribution that lends itself to intuitive interpretation is in terms of the posterior predictive distribution and, more particularly, the types of graphs that the distribution produces. In the left hand panel of Figure 16 the probability of a degenerate graph (Handcock, 2003) is plotted as a function of the edge parameter and the alternating k-triangle parameter for a network with 7 nodes. It is clear that there is a ‘‘stable’’ region in the parameter space roughly centred over the origin. The two right hand panels of Figure 16 superimpose the approximate and exact posterior on the picture of degeneracy for two different realisations of the sufficient statistics. The departure from normality (in both cases) that causes problems is the ‘‘tongue’’ that protrudes from the lower right hand parts of the contours - whereas the exact posterior is mostly contained within the non-degenerate region, the ‘‘tongue’’ extends into the degenerate region. This is most visible in the bottom panel.

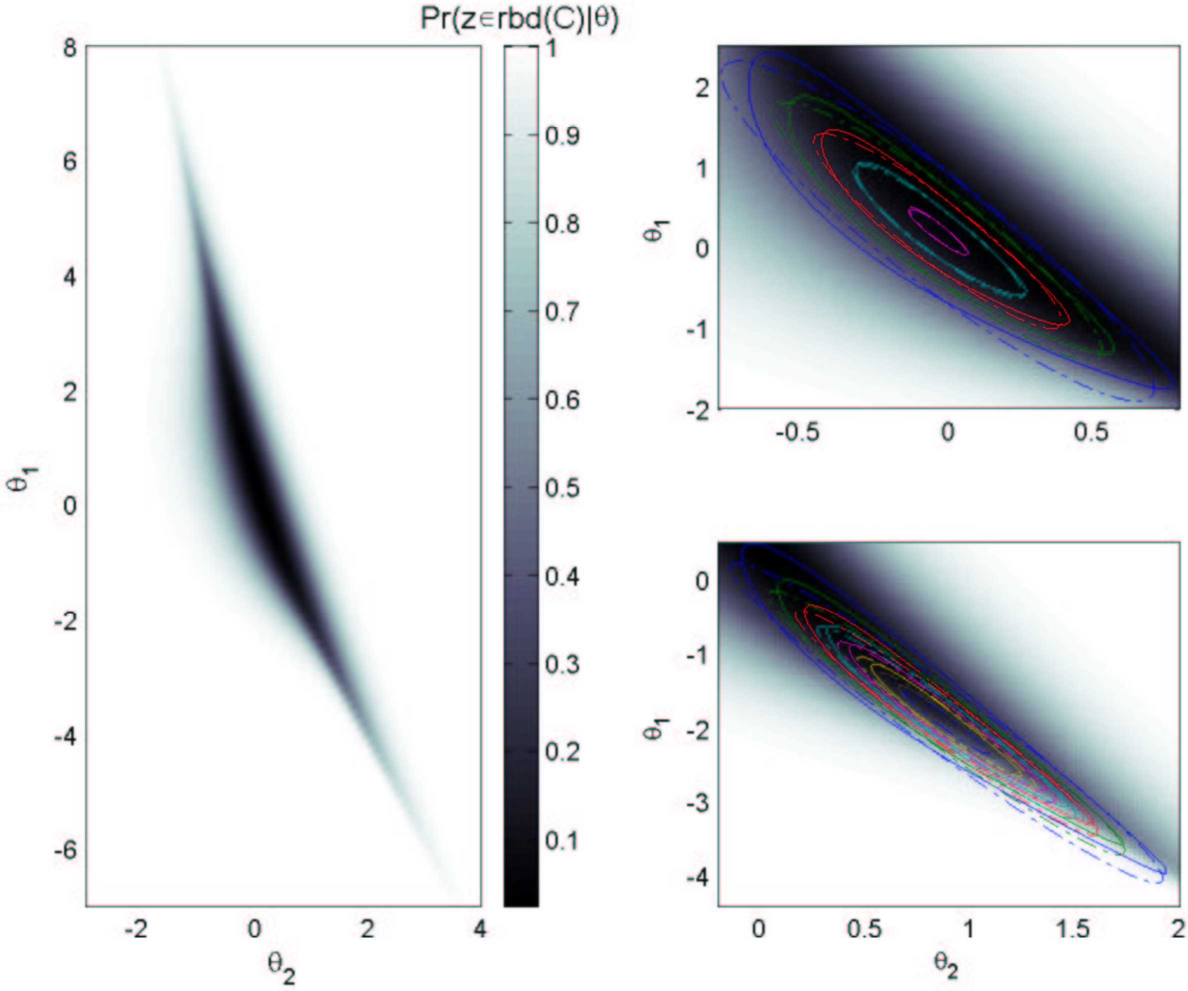

Figure 16 Degenerate edges and k-triangle graphs for 7 nodes: Probability of degenerate graph as a function of parameters (left panel); with superimposed posterior distribution (solid) and approximate posterior distribution (-·) given 11 edges and 11.25 k-triangels (λ=2) (top right); with superimposed posterior distribution (solid) and approximate posterior distribution (-·) given 11 edges and 15.5 k-triangels (λ=2) (bottom right) (Figure reproduced from Koskinen, Wang, Lusher, and Robbins, 2007).

As in the case of the MCMC importance sampler approximation, the appropriateness of the approximation is hard to establish on a general basis. In any case, in order to evaluate the approximations we need an exact inference scheme as our criterion.

A.III.3. The work-around non-convergence of MH for ERGMs

Many authors have testified to the difficulty of sampling graphs from exponential family of distributions for graphs (c.f. Snijders, 2002; Handcock, 2003). When implementing LISA, however, we assume that we may produce and unbiased draw $\mathbf{Y}_1^{(\nu_1)}$ of from an arbitrary ERGM. Typically the rule of thumb $100n^2$ (Snijders, 2002) is a sufficiently long burnin time for a Metropolis-Hastings sampler to ‘‘converge’’. For certain parameterisations and parameter values the MH may take very loch to settle and hence the produced $\mathbf{Y}_1^{(\nu_1)}$ may depend heavily on the choice of the initial state. An intuitive way of studying the time to convergence is by inspecting trace plots with parallel chains with overdispersed initial states. Using this logic, an approximately, or a pseudo, perfect sampling scheme may be implemented for ERGMs. The basic ingredients are the same as in Wilson's (2000) modification to the Propp and Wilson (1996) algorithm with the primary difference being that the stopping rules are based on approximate coalescence of chains projected onto the space of sufficient statistics instead of exact coalescence of states. A brief illustration of this approximate coalescence is given in Figure 17. For 4 different models with varying degrees of dependency are depicted the 6 dyadic statistics of the model fitted to Lazega's collaboration network as well as the # edge-wise shared partners 1 through 14. The interpretation is that when the 3 chains have ‘‘merged’’ then we know that we could have started in any state in-between (used here in a loose sense) and we still would have produced a very similar graph. To get rid of the bias induced by the deterministic stopping rule (approximate coalescence) we need to run 2 independent copies of the (approximately) coupled chains as well as do several restarts (departures from Perfect simulation are not treated here).

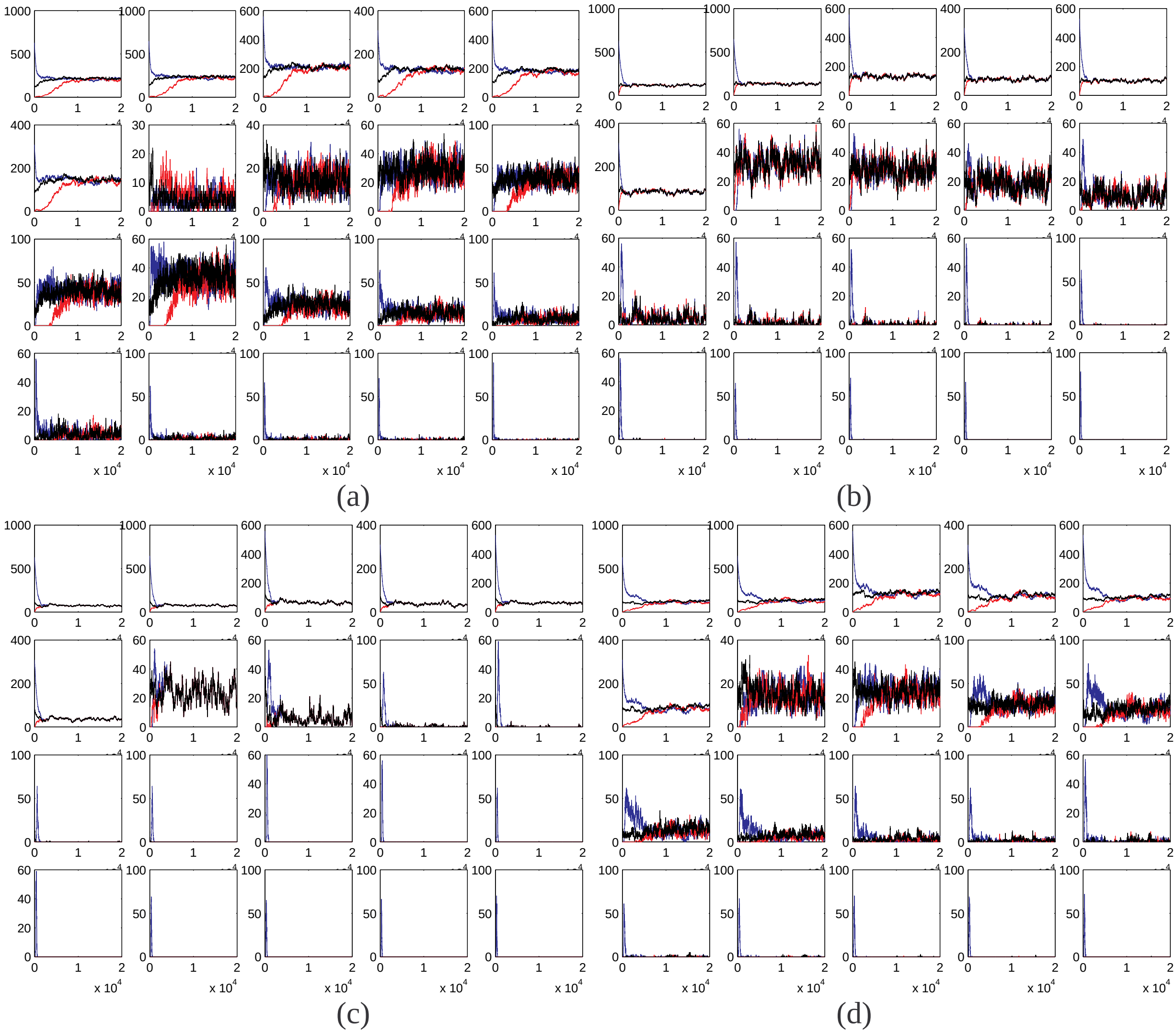

Figure 17 Approximate perfect sampling of (C)ERGM: 4 realisations of approximately coupled processes with different parameter values for HH specification for Lazega's collaboration network

A.IV. Fitting models to social networks with missing data

Here we assume that we have data of the type illustrated in Figure 18, i.e. we have one or several dyads for which we have no information as to whether an edge is present or not.

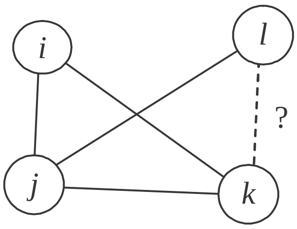

Figure 18 Observed network with missing information for dyad {*k*,*l*}

The elements of the adjacency matrix **X** thus are

$$x_{ij} = \begin{cases} 1 & \text{if actor } i \text{ relates to actor } j \\ 0 & \text{if actor } i \text{ doe not relate to actor } j \\ ? & \text{otherwise} \end{cases}.$$

In the following we only treat missing data models for graphs explicitly but the extension to directed graphs is straightforward.

Previous work on missing data in SNA has primarily focused on missing edge indicators stemming from non-respondents. The seriousness of the issue was acknowledged by Stork and Richards (1992) but they only offered an ad hoc solution (for digraphs), “complementation” of data, which has some obvious shortcomings. The impact of non-respondents has received some attention, largely concerned with the impact of missing data on various indices (Kossinets, 2006; Costenbader & Valente, 2003; & Huisman, 2007). Principled approaches are thus far relatively scarce with Robins et al. (2004) treating non-respondents as a special breed of actors and Gile and Handcock (2006) using the missing data principle, both of which in the ERGM framework.

A.IV.1. The missing data structure and ERGM

Assume that we for the complete data (i.e. assuming that we have no missing information) have an adjacency matrix **X**, taking values in $\mathscr{X}$. The model is assumed to be indexed by a $p \times 1$ vector $\theta \in \Theta \subseteq \Re$ of real-valued parameters and that a model may be written with a probability mass function (pmf) of the form

$$P(\mathbf{X} \mid \theta) = \frac{1}{Z_\theta} q_\theta(\mathbf{X}) \text{, (E-2)}$$

where $q_\theta(\mathbf{X})$ is a real vector valued function of both the parameter vector and the variable **X**, and

$$Z_\theta = \sum_{\mathbf{U} \in \mathscr{X}} q_\theta(\mathbf{U})$$

is the normalising constant that is only a function of the parameter vector θ. Furthermore, for the exponential random graph model (ERGM) the pmf is defined as in (E-2) with

$$q_\theta(\mathbf{X}) = \exp(\sum\nolimits_{k=1}^{p} \theta_k z_k(\mathbf{X})) \,,$$

where the z_k´s are functions of the adjacency matrix and a set of fixed vertex level attributes. For the simplest form of missing data mechanism we assume that edge indicators are missing independently at random. We introduce the missing data indicators

$$\xi_{ij} = \begin{cases} 1 & \text{if } x_{\mathrm{ij}} \text{ is observed} \\ 0 & \text{if } x_{\mathrm{ij}} \text{ is missing} \end{cases},$$

and throughout we assume that we can factor the joint distribution of $\xi = (\xi_{ij}; 1 \le i < j \le n)$ and **X**

$$P(\mathbf{X}, \xi \mid \theta) = P(\mathbf{X} \mid \theta) P(\xi) \,.$$

Initially we also assume that $\xi_{ij} \overset{i.i.d.}{\sim} \mathrm{Bern}(\eta)$ for $1 \le i < j \le n$ but this can relatively easy be elaborated on. A convenient yet flexible way of incorporating exogenous information about what edge indicators are missing is through e.g. a logistic regression on ξ.

When we are not interested in studying the missing data mechanisms per se, since the inclusion variables ξ are observed and therefore fixed throughout the estimation process, we define the inference procedure conditional on ξ. Hence, given ξ we define a partition of data into an observed fixed part **U** and a latent unobserved part **V**. For notational convenience let it be understood from the context that the structure on **U** and **V** is retained and unambiguous.

Missing data is straightforwardly handled in the Bayesian framework using the principle of data augmentation (Tanner and Wong, 1987). This consist of setting up a Markov chain Monte Carlo (MCMC) sampler that samples from the joint posterior of the missing data and the parameters given observed data by alternating between performing draws from

(a) The fully conditional posterior of the parameters given data and a realisation of the missing data
(b) The fully conditional posterior of the missing data given data and a realisation of the parameters

A.IV.2. The fully conditional posterior distribution of the parameters given U and V

Given a realisation of the complete data **X** = (**U, V**), and a prior distribution $\pi(\theta)$, the posterior distribution of θ given the we have observed data **X**, is given by

$$\pi(\theta \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \theta)\pi(\theta)}{m(\mathbf{X})} \propto \frac{1}{Z_\theta} q_\theta(\mathbf{X}) \pi(\theta) \,,$$

where

$$m(\mathbf{X}) = \int \frac{1}{Z_\theta} q_\theta(\mathbf{X}) \pi(\theta) \,\mathrm{d}\,\theta$$

is the marginal likelihood. In Koskinen (2006; and above) it was shown how a Metropolis-Hastings (MH) algorithm, LISA, can be used to produce an MCMC sample $(\theta^{(k)})_{k=0}^{N}$ from the posterior distribution that can be used for exploring the posterior distribution.

A.IV.3. The fully conditional posterior distribution of V given the parameters and U

Using Bayes theorem we can write the fully conditional posterior distribution of **V** given the parameters and **U** as

$$\pi(\mathbf{V} \mid \mathbf{U}, \theta) = \frac{P(\mathbf{V}, \mathbf{U} \mid \theta)}{\sum_{\mathbf{V}} P(\mathbf{V}, \mathbf{U} \mid \theta)} \propto q_\theta(\mathbf{V}, \mathbf{U}) .$$

Now label the elements in **V** using the index set $\Im$. In order to further simplify the MCMC we may now use the property of the Metropolis-Hastings sampler that in order to draw **V** we may draw elements v_i sequentially from their respective fully conditional posteriors for $i \in \Im$

$$\pi(v_i \mid \mathbf{U}, (v_j)_{j \in \Im_{-i}} \theta) = \frac{P(\mathbf{X} \mid \theta)}{P(\mathbf{X} \mid \theta) + P(\Delta_i \mathbf{X} \mid \theta)} ,$$

where $\Im_{-i} = \{ j \in \Im : j \neq i \}$, **X** = (**U, V**) and $\Delta_i \mathbf{X}$ is the matrix **X** in which v_i has been set to $1 - v_i$. Now let the change statistic $\delta_i \mathbf{X} = z(\Delta_i \mathbf{X}) - z(\mathbf{X})$ and

$$\pi(v_i \mid \mathbf{U}, (v_j)_{j \in \Im_{-i}} \theta) = \frac{1}{1 + P(\Delta_i \mathbf{X} \mid \theta) / P(\mathbf{X} \mid \theta)} = \left\{ 1 + \exp(\theta^{\mathrm{T}} \delta_i \mathbf{X}) \right\}^{-1} . \quad \text{(E-3)}$$

We may note that (E-3) is exactly the quantity we use when we use the nearest neighbour Metropolis-Hastings algorithm for drawing graphs according to ERGM. In other words the missing data algorithm can be summarised circling through the steps

(a) draw a parameter vector from the posterior using one step of LISA and the complete data **X** = (**U, V**)

(b) for each of the elements in $i \in \Im$, use (E-3) to perform a single Metropolis-Hastings updating step

A.IV.4. The simple case of the Bernoulli graph

For the Bernoulli(θ) graph the edges indicators are all identically independent Bernoulli trials, each with the probability of success (edge present) θ. The model may hence be written

$$P(\mathbf{X} \mid \theta) = \theta^{y}(1-\theta)^{n-y}$$

where $y = \sum_{i<j} x_{ij}$ is the density of the graph. With a conjugate Beta prior distribution for θ, $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, the posterior of θ given complete data **X** = (**U, V**) is hence $\theta \mid \mathbf{X} \sim \text{Beta}(y+a, n+b-y)$. When drawing the missing edges we may note that since the edges are independent by assumption, the only information regarding the values of the missing edge indicators provided by data is that which is mediated through the parameter. More succinctly put, if we have a lot of edges in the observed portion of data, the draw from the full conditional posterior of θ is likely to have produces a large value. If the value of θ is large the probability that the missing edge indicator is a success (1) is large. The MCMC circles through the steps (a) through (b.K):

(a) $\pi(\theta \mid \mathbf{X}) \propto \theta^{y+a-1}(1-\theta)^{b+n-y-1}$

(b.1.) $\pi(v_{i_1} \mid \mathbf{U}, (v_j)_{j \in \mathfrak{I}_{-i_1}} \theta) = \pi(v_{i_1} \mid \theta) = \theta^{v_{i_1}}(1-\theta)^{1-v_{i_1}}$

…

(b.K.) $\pi(v_{i_K} \mid \mathbf{U}, (v_j)_{j \in \mathfrak{I}_{-i_K}} \theta) = \pi(v_{i_K} \mid \theta) = \theta^{v_{i_K}}(1-\theta)^{1-v_{i_K}}$

where $K = |\mathfrak{I}|$.

A.IV.5. Illustration of the general case

The principle of the missing data algorithm is illustrated in Figure 19. We have data of the kind in the left of the picture, where there is a dyad {*k*,*l*} for which we have no information regarding the status of the corresponding edge. In principle either of two scenarios are true, either there is an edge between *k* and *l*, or there isn't. Each of these scenarios gives two different conditional posteriors, represented by the red and the blue curve respectively. The repeated draws of different scenarios in the MCMC assures the red and blue curve are mixed with proportions that are in proportion to the likelihood (actually probability) of their respective scenarios.

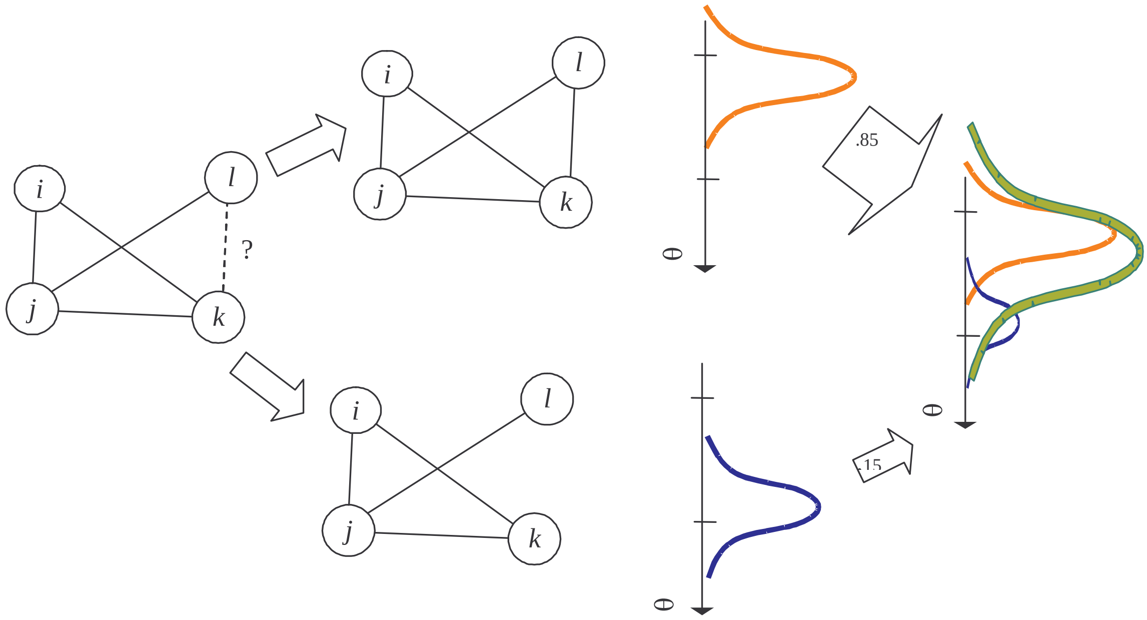

Figure 19 The principle of the missing data algorithm for social networks

In the course of fitting the ERGM to data we are also given probabilities for different true structures, in fact we are giving a whole distribution of realisations of the part of the graph that is missing. This distribution can be summarised by for example a graph with weighted edges - a weight of 1 is assigned to an edge that is certain or observed, a value close to 0 indicates a highly unlikely edge or and observed absence of an edge. It is however important to recall that we have made quite simplistic assumptions regarding the missing data mechanism and hence this algorithm should be used for ‘‘coping’’ with missing data and performing inference for the model when faced with missing data rather than inferring the missing data intself.

A.IV.6. Illustration using an empirical data set

We fit an ERGM where we model a collaboration network on a set of 36 actors (Lazega, 2001). We let the set of actors be represented by a set $N = \{1, \ldots, n\}$ of vertices and let the colaboration network be represented by a random edge set on N with adjacency matrix **X**, the elements of which are

$$x_{ij} = \begin{cases} 1 & \text{if actor } i \text{ collaborates with actor } j \\ 0 & \text{otherwise} \end{cases} .$$

In this particular example we follow the specification in Hunter and Handcock (2005) and for the attributes a_i (seniority of actor i in terms of rank), b_i (binary indicator of role of actor i), c_i (sex of actor i), d_i (office location of actor i), (for more details about the model and the data set see e.g. Lazega, 2001; Lazega and Pattison, 1999; and Snijders et al., in press):

k	Effect	$z_k(\mathbf{X})$
1	Activity/popularity	$\sum_{i<j} x_{ij}$
2	Main effect of seniority	$\sum_{i<j} x_{ij}(a_i + a_j)$
3	Main effect of practice	$\sum_{i<j} x_{ij}(b_i + b_j)$
4	Homophily practice	$\sum_{i<j} x_{ij}\mathbf{1}(b_i = b_j)$
5	Homophily sex	$\sum_{i<j} x_{ij}\mathbf{1}(c_i = c_j)$
6	Homophily office	$\sum_{i<j} x_{ij}\mathbf{1}(d_i = d_j)$

Here **1** denotes the indicator function. Following Hunter and Handcock (2005) we introduce the geometrically weighted shared partner statistic (GWEPS) which was derived as the alternating triangle statistic

$$3t_1(\mathbf{X}) - \frac{t_2(\mathbf{X})}{\lambda^1} + \cdots + (-1)^{n-3}\frac{t_{n-2}(\mathbf{X})}{\lambda^{n-3}} .$$

from the partial dependence assumption by Snijders et al., (in press). When λ in the alternating triangle statistic is allowed to be a free parameter the model containing both GWEPS and $\theta_8 = \log \lambda$ belongs to the curved exponential family of distributions.

Now, let us pretend that observations are missing for different hypothetical scenarios for missing data.

A: $\varnothing$
B:{1,2},{17,26}
C: $\{1\}\times N \setminus \{1\}$, $\{2\}\times N \setminus \{2\}$.
D: $\{1\}\times N \setminus \{1\}$, $\{2\}\times N \setminus \{2\}$, $\{3\}\times N \setminus \{3\}$, $\{4\}\times N \setminus \{4\}$

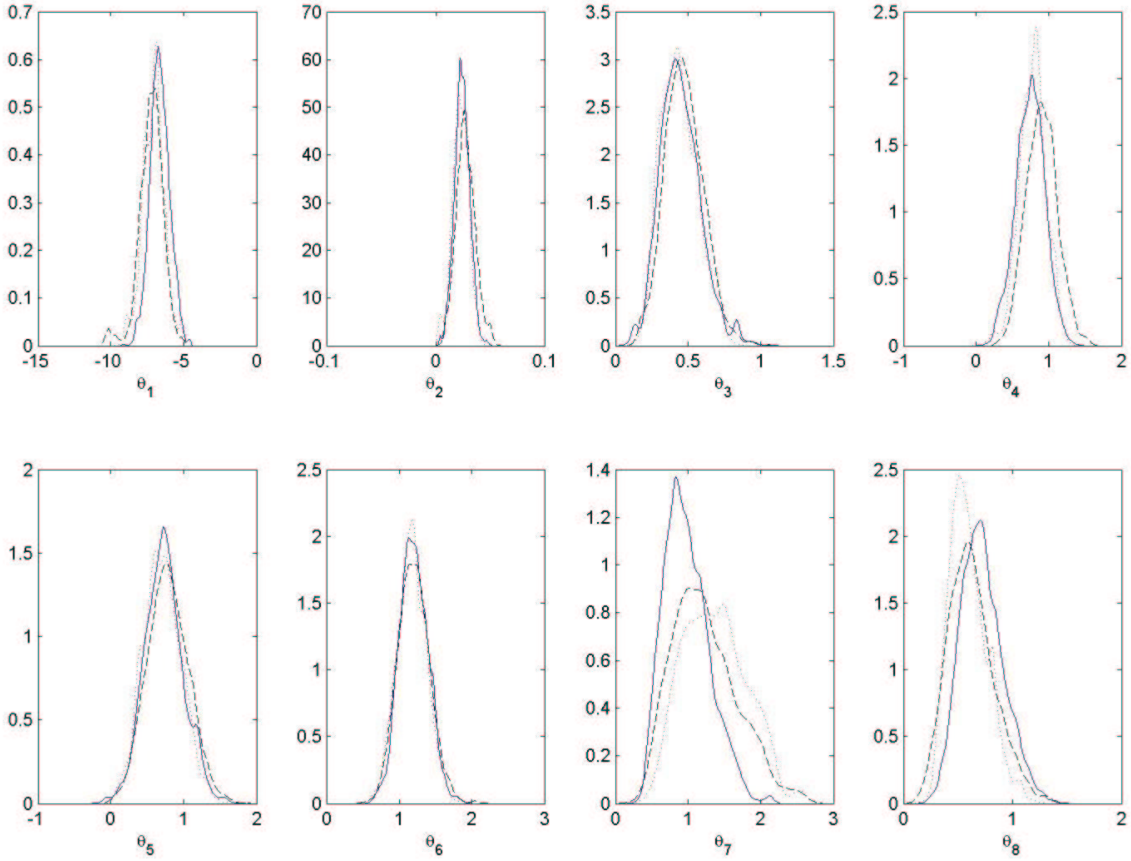

Figure 20 Density estimates for parameters is a model with GWEPS fitted to Lazega's (2001) New England Lawyers collaboration network with LISA(*K* = 2000, *m* = 7). Missing data scenario is A for solid line, C for red dotted line, D for black dashed line

For the dyadic parameters we should not expect much change other than the increase uncertainty associated with the loss of precision stemming from the reduction of the number of observations. On the whole the only noticeable difference is for the posterior of the shared partner parameter, θ_7 , where the removal of data has increased the amount of uncertainty.

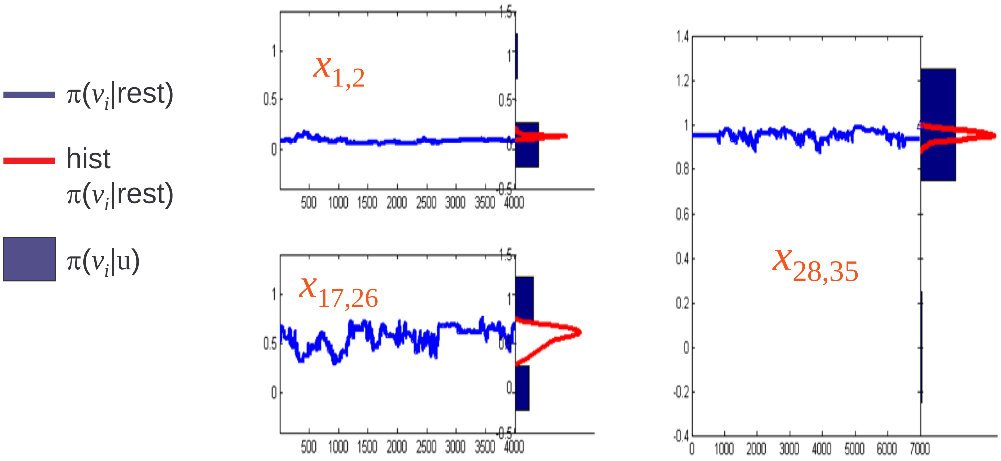

Figure 21 Posterior predictive distributions for missing edge indicators

The posterior predictive distributions for the missing edge indicators ought to give us a general idea of whether the missing data mechanism and the augmentation scheme manages to plausible capture basic features of the data. The trace plots for the missing edge indicators in scenario B are give in Figure 21. Looking back to e.g. Figure 11 we know that $x_{1,2}$ should be 0 and $x_{17,26}$ should be 1. Whereas the algorithm seems to pick up on the former the latter, at first, appears to be less convincing. When interpreting these results one has, however, to keep in mind that the overall density is fairly low in the data set and thus, given no other information, our best guess for a missing edge indicator would always be 0. Accordingly, a posterior predictive probability that an edge is present of around .5 is fairly strong evidence. In some instances the evidence is even stronger. When 10 dyads are removed at random and the missing data algorithm is run, one of the dyads {28,35} stands out as evidenced in the right hand side of Figure 21. There seems to be very strong evidence for the corresponding ties to be present. Looking again at Figure 21 we see that 28 and 35 are two senior people in the same office who are situated in a dense and highly triangulated and clustered part of the network.

Making these remarks regarding the ability of recovering missing data we do have to keep in mind that this approach is primarily designed to facilitate estimation of structural properties of a network in the face of missing data and not designed to predict missing values.

A.IV.6.a. Improvement on naive, available observations approach

As illustrated in A.I.1, a subgraph of an ERGM need not necessarily be an ERGM. From a modelling perspective, the data augmentation scheme outlined above is to be preferred to an analysis that only uses the available, what way be termed the naive approach (c.p. c.f. ‘‘available-case’’ analysis, Little and Rubin, 1987, sec. 3.3). For Markov graphs we know that we introduce non-Markov dependencies when information is removed. For other types of dependence structures the marginal dependence structures probably requires a bit more work. In the case of the exercises above, the difference between the principled approach and the naiv approach turned out to be fairly small. There are several possible explanations. One of these might be that when all the covariate effects are taken into account then the remaining interdependence between dyads is not strong enough to distort the dependence structure to such an extent that it is visible as a difference between the naive and the principled approach to missing data. Some evidence to suggest that this is indeed the case is provided in Figure 22.

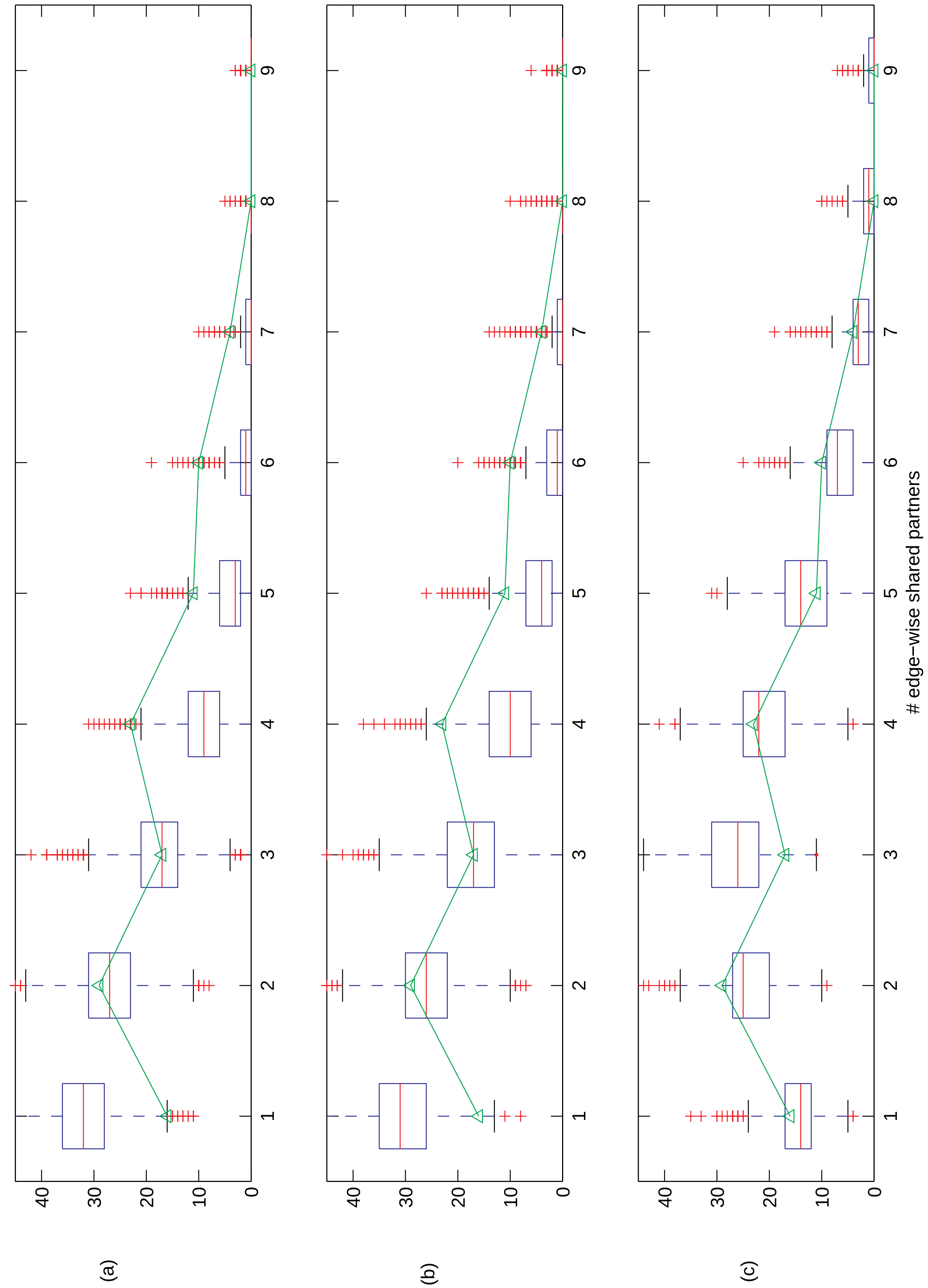

Figure 22 Goodness of fit distributions for Lazega's collaboration network: the predictive distribution for a dyad-independent model where distribution of shared partners is conditional on MLE (a); posterior predictive distribution for dyad-independent distribution only conditional on observed data (b); posterior predictive distribution for full curved exponential family specification only conditional on observed data (c).

A.IV.7. Conditions for existence of posterior distribution

As is the case in general in Bayesian inference the posterior distribution is always proper whenever the prior distribution is proper. For the most part in this presentation however we have employed flat, improper prior distributions and it is important to investigate the conditions under which the posteriors are proper. For exponential random graphs this translates into checking that

$$0 < m(\mathbf{X}) = \int \frac{1}{Z_\theta} q_\theta(\mathbf{X}) \pi(\theta) \,\mathrm{d}\,\theta < \infty ,$$

when $\pi(\theta) = 1$.

Define first the convex hull of the vector of statistics $C = \{z(\mathbf{X}) : \mathbf{X} \in \mathscr{X}\}$ and define the relative interior rint(*C*) of *C*, and the relative boundary rbd(*C*)= cl(*C*)\rint(*C*). For the Bernoulli(θ) random graph model it is obvious that the density of the complete graph belongs to rbd(*C*) and it is well known that the MLE does not exist when the complete graph is observed (Handcock, 2002). With an improper prior it is easy to check that the normalising constant in the posterior

$$\int_{-\infty}^{\infty} \left(\frac{e^\theta}{1+e^\theta} \right)^{\sum_{i<j} x_{ij}} \mathrm{d}\,\theta = \infty .$$

This suggests the following **Theorem**:
When we use an improper prior and have $\mathbf{X}_{obs}$ from an ERGM, $z(\mathbf{X}_{obs}) \in \mathrm{rbd}(C)$ imply that the posterior is improper.

For a proof see Theorem 1 in Diaconis and Ylvisaker (1979) from which we see that the above theorem may in fact be strengthened.

A.IV.7.a. Existence for network data with missing information

When the adjacency matrix is completely observed we may in principle ascertain whether $z(\mathbf{X}_{obs}) \in \mathrm{rbd}(C)$ or not (though it may be difficult in practice). Given the previous partition of data **X** = (**U, V**) but assuming that we only have completely missing vertices so that we may speak of the subgraph induced by the observed vertices that has an adjacency matrix $\mathbf{U} \in \mathscr{X}_{\mathbf{U}}$, denote by $C_{\mathbf{U}}$ the convex hull on the sufficient statistics of graphs in $\mathscr{X}_{\mathbf{U}}$ and by $C = \{z(\mathbf{X}) : \mathbf{X} \in \mathscr{X}\}$.

We put forth the conjecture that the marginal posterior of θ is improper if a vague prior is used and $z(\mathbf{U}) \in \mathrm{rbd}(C_{\mathbf{U}})$.

An interesting question is whether $z(\mathbf{X}_{obs}) \in \mathrm{rbd}(C)$ even if $z(\mathbf{U}) \in \mathrm{rint}(C_{\mathbf{U}})$.

A.V. Using a logistic link function to model the missing data mechanism

As mentioned in the previous approach, as long as the inclusion variables of the dyads are assumed independent the incorporation of more sophisticated missing data mechanisms does not alter the analysis of the network at all. Inferring determinants of missingness may be interesting in its own right though.

A.VI. Network dependent missingness

A possible extension is to allow for dependence within dyads when we have directed networks. Should we include even more elaborate interdependence assumptions the analysis quickly become complicated. It would be worthwhile pursuing the possibilities for have, at first, a limited dependence on the edge-indicators leading to probability structures such as

$$\Pr(\xi_{ij} = 1 \mid x_{ij} = 1) \neq \Pr(\xi_{ij} = 1 \mid x_{ij} = 0).$$

Tentatively we may write a conditional model as

$$\Pr(\xi_{ij} = 1 \mid \mathbf{X}, \xi_{-(ij)}) = \Pr(\xi_{ij} = 1 \mid x_{ij}) = \frac{\exp\{\eta_1 + \eta_2 x_{ij} + \gamma^{\mathrm{T}} y_{ij}\}}{1 + \exp\{\eta_1 + \eta_2 x_{ij} + \gamma^{\mathrm{T}} y_{ij}\}},$$

where $\gamma^{\mathrm{T}} y_{ij}$ is a linear combination of exogenous attributes of the dyad. When drawing the missing edge-indicators according to (E-3) conditional on everything else, we would have

$$\pi(\nu_i \mid \mathbf{U}, (\nu_j)_{j \in \Im_{-i}}, \theta, \eta, (\xi_j)_{j \in \Im_{-i}}, \gamma) = \left\{ 1 + \exp(\theta^{\mathrm{T}} \delta_i \mathbf{X}) \frac{1 + e^{\eta_1 + \eta_2 (1 - \nu_i) + \gamma^{\mathrm{T}} y_i}}{1 + e^{\eta_1 + \eta_2 \nu_i + \gamma^{\mathrm{T}} y_i}} \right\}^{-1}.$$

Hence, if η_i is negative the fact that ν_i is missing ($\xi_i = 0$) contributes negatively to the probability that the tie is present.

B. A principled approach to the data gathering process

Whereas approach A only assumes that we have data in the standard adjacency matrix form, possibly with some entries in the adjacency matrix missing, the approach B takes a principled view on the entire data analysis process. This means, in principle, that we set out policy for how the data gathering process be integrated in the statistical mechanism to produce clearly interpretable results than lend themselves to re-evaluation of the process. More explicitly our goal should be to set up rules for data gathering and data format that retains as much information as possible about the data generating process. This is to tailor the data to a specific statistical model for data. Important is also to enable and incorporate the possibility of making subjective judgements explicit. Under the Bayesian paradigm no conclusions can ever be drawn from data analysed by a model that does not explicate and quantify subjective information. If the subjective judgements are not retained at each stage of the analysis process we are unable to parse out what data tells us and what is prior knowledge and thus limiting our ability to learn from data.

The scope of this approach is quite ambitious and a full implementation is far away in the future. Apart from recommending aims and goals for the data gathering process, we may develop some minor aspects of this approach. One such approach is the partial cognitive social structures (PCSS) approach. The PCSS data collecting paradigm assumes that we have multiple sources of information on one and the same relation for a set of actors. In the original cognitive social structures approach (CSS)(Krackhardt, 1987), it was assumed that each member of a set of actors gave their version of the entire network, resulting in as many adjacency matrices as there where actors. These different versions of the network could then be used to asses different actors accuracy in judging who interacted with whom or to ‘‘estimate’’ the commonly agreed on network structure (by some criteria). The fact that this approach to CSS was completely ad-hoc made impossible statistical inference and an informed, principled analysis of actor accuracy and the true underlying structure of interaction. A Bayesian approach for analysing CSS was proposed in Koskinen, 2001, 2002 and 2004b, and it was argued that the nature of data was such that one could only really perform inference using a Bayesian methodology. There is a growing body of research on similar Bayesian approaches to CSS-like data (see Dombroski and Carley, 2002; Dombroski, Fischbeck, and Carley, 2003; Butts, 2003; Karabatsos and Batchelder, 2003).

The idea in approach B that is of highest priority is to develop a simple model for pooling information sources that builds on the extension of the Bayesian CSS given in Koskinen, Jansson and Spreen (2002). In an analogously fashion to Koskinen, Jansson and Spreen (2002), PCSS requires that we have multiple reports on one underlying (but unobserved) relation on a set of actors but in contrast to the traditional CSS, we do not require all the reports to cover the whole set of actors. Hence, PCSS may be considered as CSS with partially missing data. In the early version of developing this model we focus on the reporting mechanism of the reports (or sources of information). In order to do this we introduce two so called detection mechanisms that are specific to the perceiver (reporter, rater, or equivalently defined source). These are α, that relates to accurate detection of an interactional tie, and γ, that represents the propensity or probability of false detection of an interactional tie. To capture that some actors in the network may be more difficult to monitor of less visible, two actor specific parameters are introduced. These are collectively called ‘‘visibility parameters’’ and individually they are ϕ and ψ. The former, ϕ, is called the accurate visibility parameter and represents the probability that interaction involving this actor is seen and detected by a generalised perceiver. The latter, ψ, is the false visibility parameter of an actor. Because an actor with a high visibility parameter is likely to be seen as having relations to others when in fact this actor does not, an alternative name for ψ is the ‘‘celebrity parameter’’, something which immediately suggest the alternative name the ‘‘talent’’ parameter for ϕ.

In a full model α, γ, ϕ, and ψ are all assumed to be individual, that is, specific to each perceiver and perceive. To reduce the number of parameters and in order to make the model m ore parsimonious we need to impose various homogeneity restrictions on these parameters, assuming e.g. that groups of say actors have the same level of visibility. The success of the model depends in part on how well we are able to handle the nature of the unobserved latent network that we are interested in. We suggest that the most fruit-full approach is to treat the unobserved network as an unobserved latent variable. The aim is the to include in the posterior analysis the posterior predictive distribution for the latent network. Hence, instead of giving one point estimate for the true unobserved structure we propose (as in e.g. Koskinen, 2002) that the analyst may explore the entire distribution of networks that is implied by the model and the reports that we have observed. If we for example we are interested in knowing how many steps separates actors A and B in the true network given the reports that we have, we may calculate the posterior probability that there are 1 step, 2 steps, 3 steps, and so on. This could for example inform us of the risk that A is able to pass a parcel to B through the ties in the network. At this stage we do however suspect that the model used for the unobserved structure in the inference procedures in previous approaches (Koskinen, 2001, 2002 and 2004b; Koskinen, Jansson and Spreen, 2002) is far to simplistic and that the analysis would benefit from incorporating a more sophisticated network model such as the ERGM. This suggests that the work on LISA takes priority.

B.I. The principal data structure

We assume that there is a true network representing some generic form of interaction amongst actors in a set N= {1,...,n}. We assume this set of actors to be specified exogenous to any modelling assumption we make and that the set is considered fixed. The true network is represented by the adjacency matrix **Z**, which by assumption is unobserved. What we observe is a collection of reports $i \in \mathscr{I} = \{1,\ldots,I\}$ on the unobserved **Z**. In the classic cognitive social structures approach (CSS) of Krackhardt (1987) **Z** is the adjacency matrix of a directed graph and each report i produces a report $\mathbf{X}_{ijk}$ on the entire network with elements

$$x_{ijk} = \begin{cases} 1 & \text{if } i \text{ reports that } z_{jk} = 1 \\ 0 & \text{o.w.} \end{cases}.$$

In the partial cognitive social networks (PCSS) setup we allow reporters to report on different subsets of the dyads so that $\mathbf{X}_i = (x_{ijk} : (j,k) \in M_i \subseteq N^{(2)})$. Hence if $m_i = | M_i |$, we have

$$m = \sum_{i \in \mathscr{I}} m_i ,$$

observations as opposed to the $In(n-1) = n^2(n-1)$ observations for the traditional CSS setup.

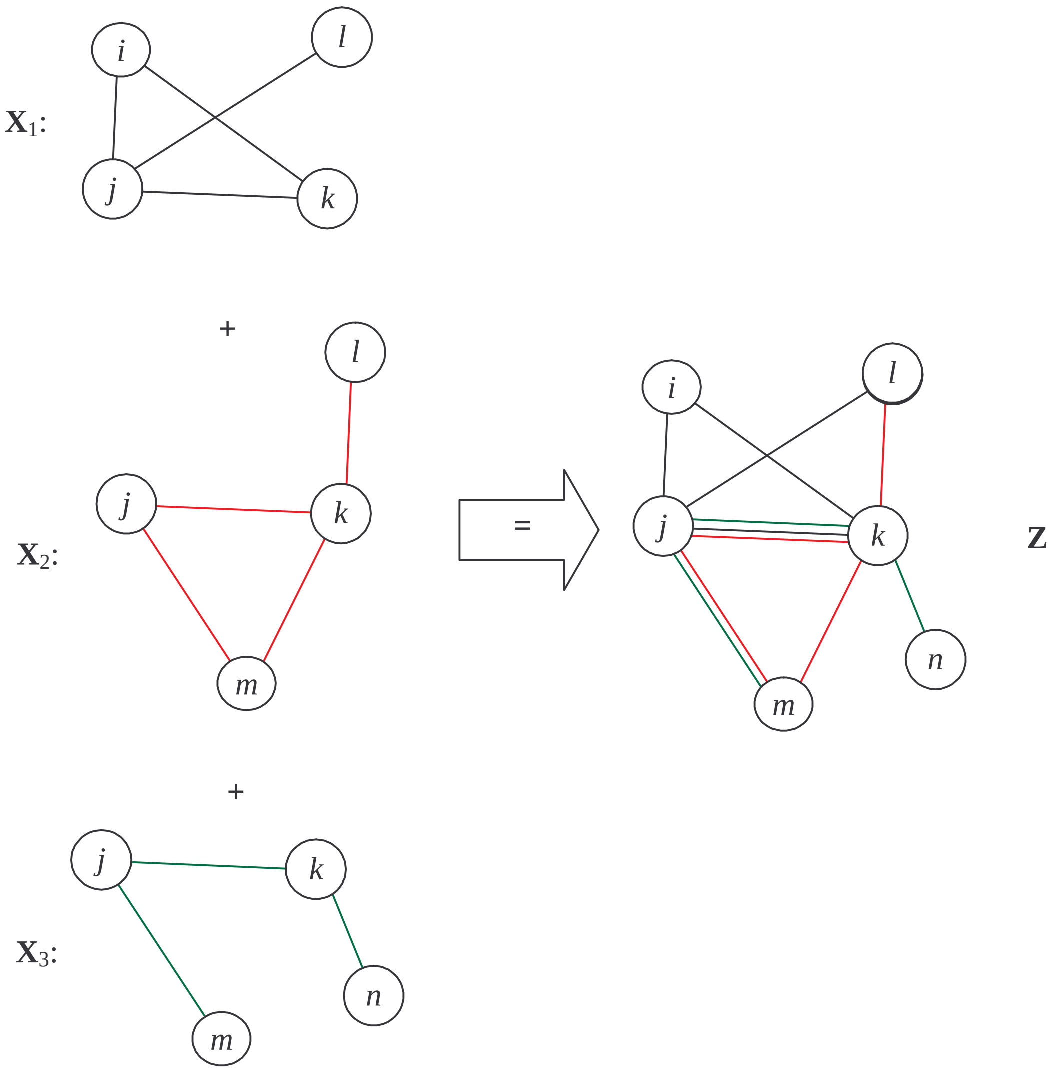

Figure 23 The data structure of PCSS.

In Figure 23 we illustrate the data principle of PCSS. Note that we do not propose to simply aggregate the reports to create a true graph with elements $z_{jk} = \min_{i \in \mathscr{I}}(x_{ijk})$ since we take inconsistencies in the reports to mean something substantial. We might for example ascribe great significance to the fact that in Figure 23, reporter 1 has reported $\{l,k\}$ as absent whereas reporter 2 has reported it as present. Note also the reference that is made here to the distinction between a tie reported as absent $x_{ijk} = 0$ by *i* and the tie not reported on at all by *i*, $\{j,k\} \notin M_i$. This is a crucial observation (admittedly not fully acknowledged in Koskinen, Janson and Spreen, 2002) as we do very rarely have explicit reports on null-dyads - typically actors are asked to name their alters, reports are made on what pairs of actors interact, etc.

Butts (2007) proposes an extension to the CSS approach in Butts (2003) that explicitly models the reporting mechanism using a kind of biased nets. This approach could prove extremely useful in dealing with the potentially systematic errors introduce by viewing the M_i's as fixed and exogenous. As shall become clear from the descriptions below, the CSS approach in Butts (2003) is likely to be a little too simplistic to adequately represent the PCSS.

B.I.1. Partial CSS with independent arcs and uniform prior on latent network

In line with Batchelder et al. (1997), we introduce the probabilities

$$H_{ijk} = \Pr(X_{ijk} = 1 \mid Z_{jk} = 1)$$

and

$$F_{ijk} = \Pr(X_{ijk} = 1 \mid Z_{jk} = 1)\,.$$

These are generally referred to as hit and false alarm probabilities respectively. In a first model we make the assumption that for the reports $\mathbf{X}_i$ is independent of $\mathbf{X}_u$ for all $\{i,u\} \in \binom{\mathscr{I}}{2}$ conditional on **Z**, and their elements are independent and satisfy

$$\Pr(X_{ijk} = x_{ijk} \mid \mathbf{Z}) = \Pr(X_{ijk} = x_{ijk} \mid Z_{jk} = z_{jk})\,.$$

If each element of **Z** is considered a parameter to be estimated, a saturated model with distinct hit and false alarm parameters for each triple $(i,j,k) \in \bigcup_{u \in \mathscr{I}} \{u\} \times M_u$, would mean $2m + n(n-1)$ parameters. Considering that we have only m observations we need a way of reducing the number of parameters. A convenient way of doing this is to introduce homogeneity restrictions on the hit and false alarm parameters. One of the simplest forms of homogeneity restrictions is to assume that the accurate or false report only relates to who the informant is

$$H_{ijk} = \alpha_i \text{ and } F_{ijk} = \gamma_i \text{ for all } i \in \mathscr{I}\,.$$

For this model the likelihood function would be

$$L(\mathbf{Z},\alpha,\gamma;(\mathbf{X}_i)_{i\in\mathscr{I}}) = \prod_{i\in\mathscr{I}} \prod_{(j,k)\in M_i} \alpha_i^{z_{jk}x_{ijk}} (1-\alpha_i)^{z_{jk}(1-x_{ijk})} \gamma_i^{(1-z_{jk})x_{ijk}} (1-\gamma_i)^{(1-z_{jk})(1-x_{ijk})}, \text{ (E-4)}$$

which is basically the likelihood for a collection of independent Bernoulli trials.

B.I.1.a. Posterior distributions of detection parameters

With a likelihood of the form (E-4), obtaining the posterior distributions for the edge indicators and the detection parameters is straightforward. Because of the large cardinality of the space of **Z,** $\mathscr{Z}$, we cannot however obtain the posterior distribution in an analytically tractable form. Assume that let **Z** and the detection parameters be independent a priori with product prior $\pi(\mathbf{Z})\pi(\alpha,\gamma)$, giving a joint posterior distribution

$$\pi(\mathbf{Z},\alpha,\gamma \mid (\mathbf{X}_i)_{i\in\mathscr{I}}) \propto L(\mathbf{Z},\alpha,\gamma;(\mathbf{X}_i)_{i\in\mathscr{I}})\pi(\mathbf{Z})\pi(\alpha,\gamma)\,.$$

Assuming independent Beta($a_{\alpha_i}, b_{\alpha_i}$) and Beta($a_{\gamma_i}, b_{\gamma_i}$) priors for the detection parameters, the fully conditional posterior of α_i given the rest is Beta($\bar{a}_{\alpha_i}, \bar{b}_{\alpha_i}$), where

$$\bar{a}_{\alpha_i} = a_{\alpha_i} + \sum_{(j,k)\in M_i} z_{jk} x_{ijk} \text{ and } \bar{b}_{\alpha_i} = b_{\alpha_i} + \sum_{(j,k)\in M_i} z_{jk}(1-x_{ijk}) .$$

Similarly for γ_i given the rest the posterior is Beta($\bar{a}_{\gamma_i}, \bar{b}_{\gamma_i}$), with

$$\bar{a}_{\gamma_i} = a_{\gamma_i} + \sum_{(j,k)\in M_i} z_{jk} x_{ijk} \text{ and } \bar{b}_{\gamma_i} = b_{\gamma_i} + \sum_{(j,k)\in M_i} (1-z_{jk})(1-x_{ijk}) .$$

Hence for a given realisation of **Z** we may easily draw detection parameters from the posterior distribution or calculate moments for the detection parameters. The problem is that **Z** is unobserved and has to be inferred somehow.

B.I.1.b. Posterior distributions of latent network - Bernoulli graph

Consider using as a prior distribution for **Z** a Bernoulli(*p*) random graph model. Investigating the likelihood function (E-4) we see that the fully conditional posterior of **Z** given everything else is

$$\pi(\mathbf{Z} \mid \alpha, \gamma, (\mathbf{X}_i)_{i\in\mathscr{I}}) \propto L(\mathbf{Z}, \alpha, \gamma; (\mathbf{X}_i)_{i\in\mathscr{I}})\pi(\mathbf{Z})$$
$$= \left(\prod_{i\in\mathscr{I}} \prod_{(j,k)\in M_i} \alpha_i^{z_{jk}x_{ijk}} (1-\alpha_i)^{z_{jk}(1-x_{ijk})} \gamma_i^{(1-z_{jk})x_{ijk}} (1-\gamma_i)^{(1-z_{jk})(1-x_{ijk})} \right) \prod_{(j,k)\in N^{(2)}} p^{z_{jk}} (1-p)^{(1-z_{jk})} .$$

This may be further simplified so that we recognise this to be the pmf of a Bernoulli graph with inhomogeneous arc probabilities p_{jk}

$$p_{jk} = \left\{ 1 + \frac{(1-p) \prod_{i\in\mathscr{I}_{(j,k)}} \gamma_i^{x_{ijk}} (1-\gamma_i)^{(1-x_{ijk})}}{p \prod_{i\in\mathscr{I}_{(j,k)}} \alpha_i^{x_{ijk}} (1-\alpha_i)^{(1-x_{ijk})}} \right\}^{-1} \quad \text{(E-5)}$$

where $\mathscr{I}_{(j,k)} = (i \in \mathscr{I} : (j,k) \in M_i)$.

B.I.1.c. Gibbs sampler

Based on these fully conditional posteriors we may set up a Gibbs sampler that in each iteration circles through the following updating steps:

(a) For each $(j,k) \in N^{(2)}$ draw

$$Z_{jk} \mid \alpha, \gamma, (\mathbf{X}_i)_{i\in\mathscr{I}}, \mathbf{Z}_{-(j,k)} \sim \text{Bernoulli}(p_{jk}),$$

where p_{jk} is calculated according to (E-5).

(b) Conditional on the drawn **Z**, for each $i \in \mathscr{I}$ draw

$$\alpha_i \mid \alpha_{-i}, \gamma, (\mathbf{X}_i)_{i \in \mathscr{I}}, \mathbf{Z} \sim \text{Beta}(\overline{a}_{\alpha_i}, \overline{b}_{\alpha_i})$$

(c) Conditional on the drawn **Z**, for each $i \in \mathscr{I}$ draw

$$\gamma_i \mid \alpha, \gamma_{-i}, (\mathbf{X}_i)_{i \in \mathscr{I}}, \mathbf{Z} \sim \text{Beta}(\overline{a}_{\gamma_i}, \overline{b}_{\gamma_i})$$

The resulting sequence $(\alpha^{(g)}, \gamma^{(g)}, \mathbf{Z}^{(g)})_{h=1}^{G}$ thus produced will be an approximate sample from the joint posterior distribution of the detection parameters and unobserved arcs given the reports.

B.I.1.d. Use of prior distributions

The hyper parameter *p* for **Z**, is used as a tool for us to adding prior information regarding our belief in the density of the latent network. It is a very blunt tool and has to be used with some caution. The effect of using different values of *p* is seen in (E-5) where it enters multiplicatively in the second term of the denominator as (1 - *p*)/*p*. Note that (1 - *p*)/*p* = 1 for *p* = .5, i.e. with a uniform distribution on all digraphs a prior there is ‘‘no’’ contribution of the prior distribution to the posterior distribution.

The prior distributions for the detection parameters are meant to capture our prior belief regarding the accuracy of the different reporters (records, sources or whatever is considered a report). The Beta distributions are fairly easy to understand since the support is bounded and the shape of the distributions allocate probability mass to different sub-intervals in the interval (0,1). It is tempting to use the ‘‘vague’’ prior distribution for the detection paramters, i.e. Beta(1,1), which is equivalent to the rectangular distribution on the unit interval. There is a danger in doing this however since the model is not fully identified.

B.I.1.e. Identifiability

The model with likelihood function (E-4) is not fully identified since

$$L(\mathbf{Z}, \alpha, \gamma; (\mathbf{X}_i)_{i \in \mathscr{I}}) = L(\mathbf{Z}^c, \alpha^*, \gamma^*; (\mathbf{X}_i)_{i \in \mathscr{I}}),$$

where $\mathbf{Z}^c$ is the adjacency matrix of the complement graph to that for which **Z** is the adjacency matrix, $\alpha^* = \gamma$, and $\gamma^* = \alpha$. Hence with uniform priors everywhere the posterior distribution is multimodal in the sense that each point in the parameter space has a complementary point, that in a sense has the opposite interpretation, with the same posterior probability (really, the same value of the posterior ordinate). What is needed to achieve identification and near identification is discussed in Koskinen (2004b). As reported there it is far from a trivial problem and it is also argued against using restrictions on the parameter space. For our purposes it seems most appropriate to adopt a fully subjective Bayesian analysis and employing meaningful proper prior distributions. Naturally, a fully subjective Bayesian analysis ought to be checked for sensitivity to prior specifications.

B.I.1.f. Improving the prior for the unobserved network

Since the Bernoulli graph model seldom is a useful model for real social network data (c.f. Robins, et al., 2006) we would do well to have a more realistic prior distribution for the latent network that gives us more freedom in defining likely latent structures. There are many candidates for improving on the Bernoulli graph model but instead of listing different options we propose to use an exponential random graph model (Frank and Strauss, 1986; Pattison & Wasserman, 1999; Robins, Pattison, and Wasserman, 1999; Wasserman & Pattison, 1996; Wasserman and Robins, 2005, Snijders et al., in press; Hunter and Handcock, 2006) since many of the models proposed in the literature are special cases of this model (for example the Bernoulli graph model). To recap the exponential random graph model (ERGM), we define the pmf for **Z** a priori to be

$$\pi(\mathbf{Z}) = \exp(\textstyle\sum_{k=1}^{p} \theta_k u_k(\mathbf{Z}) - \vartheta(\theta)),$$

where $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ is a vector of real-valued hyper parameters, the u_k´s are functions of the adjacency matrix and a set of fixed vertex level attributes, and $\vartheta(\theta)$ is a normalising constant that is only a function of the hyper parameters.

The Gibbs sampler updating steps for the detection parameters remain the same as before since the prior $\pi(\mathbf{Z})$ cancel in the full conditional posterior distributions for these. Since the ERGM allows for dependencies between arcs the updating step for **Z** differs from the previously presented Gibbs sampler however. To begin with the interdependencies mean that we have to draw the entire **Z** in one step. For most ERGMs performing draws cannot be made directly but requires an MCMC approach. If, say, a Metropolis-Hastings step is used to update **Z**, the algorithm is not strictly speaking a Gibbs sampler anymore (it is important to keep in mind however that the Gibbs sampler is a special case of the Metropolis-Hastings algorithm).

To write up the fully conditional posterior of **Z** given the rest we begin by inspecting the posterior as given by Bayes theorem

$$\pi(\mathbf{Z} \mid \alpha, \gamma, (\mathbf{X}_i)_{i \in \mathscr{I}}) \propto L(\mathbf{Z}, \alpha, \gamma; (\mathbf{X}_i)_{i \in \mathscr{I}}) \pi(\mathbf{Z})$$
$$= \left(\prod_{i \in \mathscr{I}} \prod_{(j,k) \in M_i} \alpha_i^{z_{jk} x_{ijk}} (1-\alpha_i)^{z_{jk}(1-x_{ijk})} \gamma_i^{(1-z_{jk}) x_{ijk}} (1-\gamma_i)^{(1-z_{jk})(1-x_{ijk})} \right) \exp(\textstyle\sum_{k=1}^{p} \theta_k u_k(\mathbf{Z}) - \vartheta(\theta)).$$

By taking the logarithm of the likelihood function and rearanging we see that we may write the full conditional posterior as an ERGM

$$\pi(\mathbf{Z} \mid \alpha, \gamma, (\mathbf{X}_i)_{i \in \mathscr{I}}) \propto \textstyle\sum_{k=1}^{p} \theta_k u_k(\mathbf{Z}) + \displaystyle\sum_{i \in \mathscr{I}} \sum_{(j,k) \in M_i} z_{jk} f(x_{ijk}, \alpha_i, \gamma_i),$$

where

$$f(x, a, b) = x \log \frac{a(1-b)}{b(1-a)} + \log \frac{(1-a)}{(1-b)}.$$

In order to get this expression in a more convenient form we write $\tilde{u}_i(\mathbf{Z}) = \sum_{(j,k) \in M_i} z_{jk} x_{ijk}$, $\tilde{\theta}_i = \tilde{\theta}_i(\alpha_i, \gamma_i) = \log \alpha_i (1-\gamma_i) - \log \gamma_i (1-\alpha_i)$, $\tilde{\tilde{u}}_i(\mathbf{Z}) = \sum_{(j,k) \in M_i} z_{jk}$, and

$\tilde{\tilde{\theta}}_i = \tilde{\tilde{\theta}}_i(\alpha_i, \gamma_i) = \log(1-\alpha_i) - \log(1-\gamma_i)$, and we may put the fully conditional posterior in the familiar form

$$\pi(\mathbf{Z} \mid \alpha, \gamma, (\mathbf{X}_i)_{i \in \mathscr{I}}) \propto \sum\nolimits_{k=1}^{p} \theta_k u_k(\mathbf{Z}) + \sum_{i \in \mathscr{I}} \tilde{\theta}_i \tilde{u}_i(\mathbf{Z}) + \sum_{i \in \mathscr{I}} \tilde{\tilde{\theta}}_i \tilde{\tilde{u}}_i(\mathbf{Z}) \,. \text{ (E-6)}$$

The updating step (a) in the Gibbs sampler can now be replaced by a Metropolis-Hastings updating step. Given the current adjacency matrix **Z**, we propose a move to **Z*** which has a randomly chosen element (*j**,*k**) set to its opposite, $z^{*}_{j,k} = 1 - z_{jk}$. This move is accepted with probability min{1,H}, for

$$H = \frac{\pi(\mathbf{Z}^* \mid \alpha, \gamma, (\mathbf{X}_i)_{i \in \mathscr{I}})}{\pi(\mathbf{Z} \mid \alpha, \gamma, (\mathbf{X}_i)_{i \in \mathscr{I}})}$$

$$= \exp\left\{ \sum\nolimits_{k=1}^{p} \theta_k (u_k(\mathbf{Z}^*) - u_k(\mathbf{Z})) + \sum_{i \in \mathscr{I}} \tilde{\theta}_i (\tilde{u}_i(\mathbf{Z}^*) - \tilde{u}_i(\mathbf{Z})) + \sum_{i \in \mathscr{I}} \tilde{\tilde{\theta}}_i (\tilde{\tilde{u}}_i(\mathbf{Z}^*) - \tilde{\tilde{u}}_i(\mathbf{Z})) \right\}.$$

B.I.1.g. Drawbacks of ERGM prior

Fixing the parameters in an ERGM prior, in other words fixing the hyper parameters, may prove both hard and overly restrictive. Hard because it may be hard to know a priori what type of data a specific model produces. Overly restrictive because the chosen model may place very little probability mass on graphs in the likely region of data. There is no obvious way of scaling the ERGM to diffuse the pmf while still retaining the ‘‘location’’ as is possible in location-scale models. Even if we do have good prior information, say parameters estimated from similar data, incorporating this information through hyper parameters for **Z** may not be the best strategy.

B.I.2. Treating unobserved network as a latent variable

An approach to PCSS that is subtly different from the one outlined above is to treat **Z** not as a parameter to be estimate but rather a latent unobserved variable. This reduces the number of parameters as well as being more flexible. The procedure (like that with ERGM prior) produces a distribution of graphs but also a distribution for the structural parameters. We may point out that this is better that using a point estimate (of **Z**) since a point estimate does not include the uncertainty in data collection - examples of measures to calculate from the distribution of graphs is connectivity; the probability that A is able to send a package to B through his network using at most k steps; etc.. The posterior distribution for the ERGM parameters allow for further and future use of this inference in analyses of social network data in accordance with the sequential nature of Bayesian inference

To implement this addition to the PCSS we may keep to the details above for updating **Z**, α, and γ with only one extra additional step. Let $\pi(\theta)$ denote the prior distribution for the parameters in the model for the latent variable **Z**. Since **Z**, **X,** α, and γ are all fixed in the updating step for θ we can easily see that updating θ may be done as in the case of inference for a regular ERGM. Hence we can perform the updating step for θ using one LISA updating step.

B.I.3. Introducing visibility parameters in PCSS

Introducing visibility parameters in the inference scheme outline above is straightforward if these are introduced multiplicatively:

$$H_{ijk} = \alpha_i \phi_j \phi_k \text{ and } F_{ijk} = \gamma_i \psi_j \psi_k \text{ for all } i \in \mathscr{I} \text{ and } j,k \in N .$$

The MCMC inference scheme basically remains the same with some minor differences.

B.I.4. A probit formulation of the PCSS

When we start relaxing the homogeneity assumptions one by one, we eventually run out of free parameters. In addition interpretation of the results become more cumbersome as the number of parameters increase, especially seeing as we need to assign prior distributions to all the detection and visibility parameters. A way forward is modelling the probabilities using probit link functions as in Koskinen, Jansson and Spreen (2002).

C. Missing vertices

Dealing with network data where vertices, possibly an unknown number, are missing presents us with a host of interrelated problems of both a conceptual nature as well as technical. Although we may sometimes think that we have a well-defined notion of what constitutes a network, typically our intuition is based on an ill-founded vague notion of social interaction. We tend to confuse the units of measurement with what is actually measured. Naturally some definitions of what constitutes a network are more susceptible to falling pray to this fallacy than others. Say for example that we subscribe to the hypothesis of six degrees of separation for a relation such as that defined in Milgram (1967), in which case it would be nonsensical to speak of estimating the size of the network. Rudimentary statistical methods are available for the case when the underlying population is well defined. These methods typically rely on very simplistic assumptions regarding social interaction but the simplicity of these models may also promote their use in preliminary analyses, in which case we - because of the high degree of uncertainty - recommend using the Bayesian approach to fitting the model of Frank and Snijders (1994) presented in Tallberg (2004).

We think that in the context of dealing with missing vertices one need to consider the subtle issues underlying social network theory more thoroughly. Kossinets (2003) touch on some of the issues surrounding missing vertices, there are however much more fundamental ontological and epistemological questions that have to be developed and explored. One of these questions is generally know by the name of the boundary specification problem (Laumann et al 1983; White, 1992). Another issue is that of scaling and how to large networks relate to smaller networks (Anderson, Butts, and Carley, 1999; Robins, Woolcock, & Pattison, 2005).

The source of the most frequent discrepancies between actually collected data and the data of the ideal research setup must arguably be the way inclusion of vertices depend on their ties to other actors. When trying to analyse social network data from a statistical point of view the fact that the inclusion probability of the units of analysis (the dyads, triads, etc) are dependent of the values of the same make things complicated. However, it makes intuitive sense to sample the network using a link tracing mechanism or a snowball sampling scheme. This ties into the problem of missing vertices since one may argue that a plausible missing vertex mechanism is that of stopped sampling. In combination with tool for handling edge likelihood assessment, a statistical model that handles snowball sampled data could prove very useful for dealing with this type of missing data mechanism. Thus far however, there only seems to exist one approach that has the potential for modelling snowball sampled data in a non-trivial way and that is work in progress by Pattison (2007). This approach holds a lot of promise but tailoring it to our needs may be far off into the future.

If we have reason to believe that there is only a small but unknown number of missing vertices an extension of the approach in A may be applicable. This consists of transforming the missing node problem into a set of ERGMs with missing dyads where each conditional model is conditional on a fixed number of missing vertices. In practice we do a series of estimations where we start by assuming that there are no missing vertices. We the proceed to introduce one ‘‘imaginary’’ ghost vertex an do estimation in the manner detailed in approach A. We then continue to introduce successively more ghost vertices. The goal is to produce posterior predictive distributions on the edges or arcs that pertain to the missing vertices for each scenario separately. If there is some structure on the network the posterior predictive distributions will inform us as to where the missing vertices are in the structure of the network effectively informing us of where (in the network) to look for further actors. A potential source of problems is the homogeneity assumption of the ERGM and unless we include actor attributes and other exogenous sources of information there is a risk that (observed) actors are likely to be related to the missing vertices in proportion to their activity. At a higher order of interdependency, this would be expressed by mirror-scenarios by which we mean that certain structures relating to the missing actor may be invariant under partially homeomorphic transformations. Naturally we do not have any covariates for the ghost vertices and hence we also have a problem of missing covariates that has to be dealt with using approach E.

The next step is to develop a scheme for obtaining a posterior distribution on the number of missing vertices. This problem is equivalent to a model selection problem. To obtain the posterior we hence need proper prior distributions for the model parameters, the determination of which is a delicate problem already in the case of an ERGM without missing data. Here it is rendered even more difficult since we want to compare models defined for networks of different sizes (again see networks Anderson, Butts, and Carley, 1999; Robins, Woolcock, & Pattison, 2005). Some of these problems can be alleviated by reporting result for different prior distributions in order to get a handle on how the posterior is influenced by the prior distribution. There is also the additional technical problem of how to calculate the so called marginal likelihood that is needed for Bayesian model selection. A procedure that works for the ERGM for an approximate MCMC algorithm was presented in Koskinen (2004a) but it is yet to be determined how this can be modified to work in the LISA algorithm and in the presence of missing data. There are several interesting aspects that may be studied in this framework such as how scaling works; whether increasing the number of ghost vertices improves fit enough to offset the increase in complexity; etc.

Fitting the ERGMs sequentially, conditional on the number of vertices can in theory be done using conditional inference (relatively straightforward once A solved). For similar types of problems in other areas of statistics, one usually prefers to fit the model and the model dimensionality simultaneously. The inference procedure would then answer several questions at the same time by producing a joint posterior for the parameters and the number of vertices. Technically though, it seems very hard to implement a joint analysis because of (yet again) scaling issues.

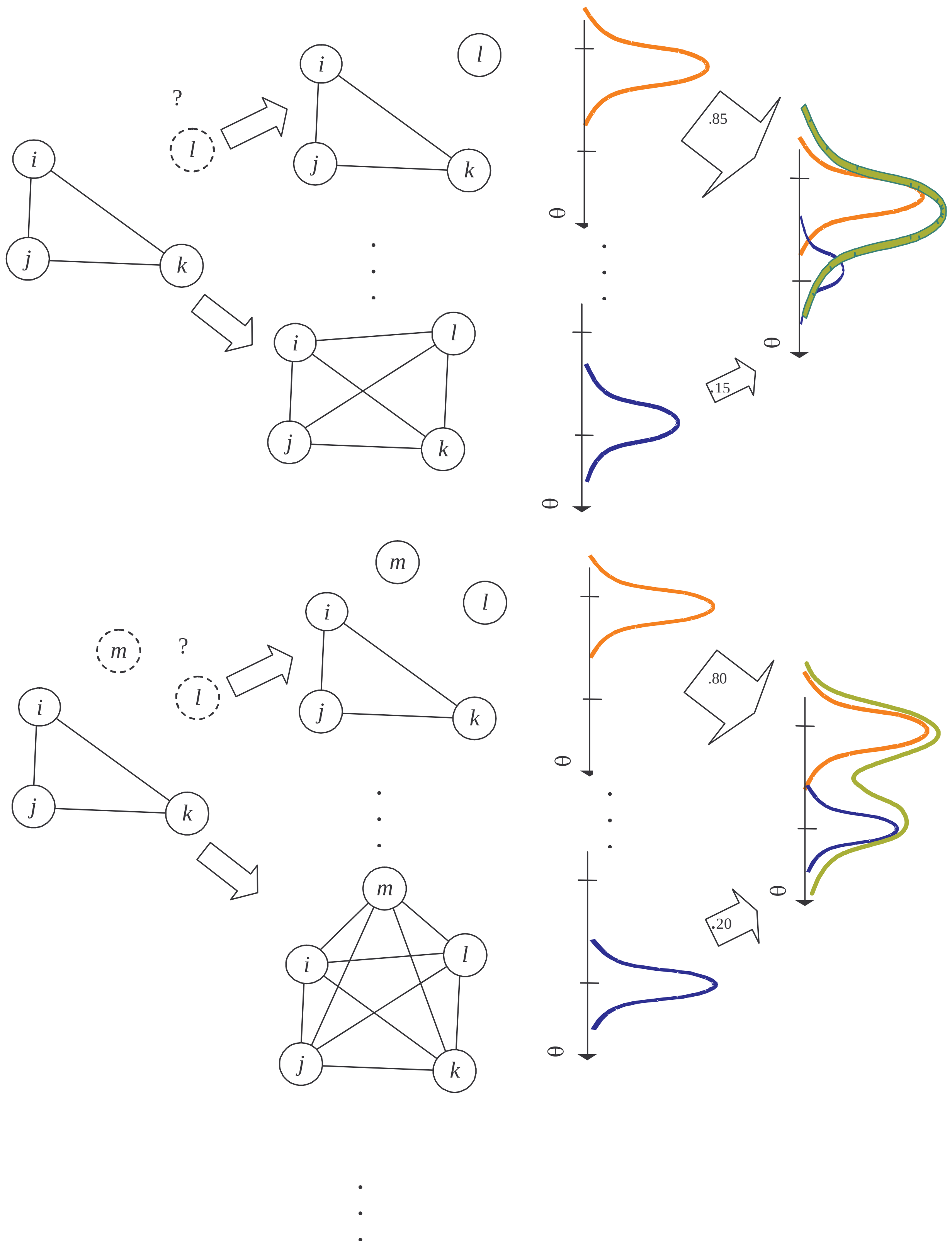

Figure 24 Illustration of sequential fitting of ERGMs to networks with ghost vertices.

D. Vertex doppelgangers - Node mapping between networks

Sometimes when using different data collecting schemes for different relations we may end up with ambiguously defined vertices. As an example we may collect interaction data from exchanges of emails where we only know the identities of the senders and receivers up to their aliases or email addresses. In each network considered separately, there may be confusion as what unique individuals map to what email addresses and if there are individuals with multiple email addresses the size of the network is undefined. When the networks are considered jointly there is the additional problem of non-overlapping email addresses as addresses that are not in the intersection of address-sets may be pertain to either unique individuals as well as individuals that are in the intersect of (true) vertex sets.

We propose a solution for the case when we wish to fit a multivariate ERGM to two (or more) relations on a set of vertices where small proportion of the vertices may be doppelgangers (triplegangers, and so on). The solution is primarily of a technical nature and as in the case of approach A it relies on simulating missing data in the course of estimating parameters. If we denote by ‘‘id’’ the identities of vertices as these are presented in the data and denote by ‘‘true vertices’’ the true but unknown set of unique vertices, an observed network with doppelgangers maps to several different graphs depending on what ids are joined together to true vertices and what ids are considered unique, separate vertices. In the course of implementing the LISA algorithm vertices die or are born in each iteration so that the size of the network fluctuates during the estimation process. This has the interpretation that inferences are made for all possible mappings of ids onto true vertices and that these inferences are weighted together to create a cumulative inference that is unconditional on any specific mapping. The proportions used in this weighting, the weights as it were, are given automatically in the course of running the algorithm through standard probability calculus.

Since the fitting of the MERGM to data with doppelgangers involves (probabilistically) comparing networks of different sizes (see A on scaling up and different size networks) there are possibly several implementation issues that need to be tackled in order to make the inference procedure practically feasible. It is conceivable that one in certain circumstances need to invoke a degree of the notion of structural equivalence for ids in different vertex sets for two different relation in order to make use of data. By this we refer to cases when we may have reason to believe that the structural position of ids in the network is a source of information as to the true identity of the id - if two ids have ties to almost the same set of others we may suspect that these two ids belong to the same true vertex. This line of thinking applied across networks may require that we assign high prior probability to multiple ties (in the sense that u relates to v on relation 1 implies that u relates to v on relation 2).

As in approach C there are many conceptual issues to be resolved and investigated. There may also be a scope for looking into the literature on matching using relative automorphism.

E. Missing covariates

Although treated here as a separate issue, the issue of missing covariates when doing social network analysis is likely to be an integral part of any missing data scheme. It is for example hard to imagine cases where all covariates are known for all non-respondents. Nevertheless, since attributes and covariates are known to improve model fit for most social network problems (see e.g. McPherson, Smith-Lovin, and Cook, 2001, on the importance of homophily; see Robins, 1998, and Robins and Pattison, 2005,for details of how attributes are incorporated into ERGMs) detailing the treatment of missing covariates is essential for effective modelling. The principal difference from the treatment of missing data for interaction is that there are no obvious models for what is missing. For approach A for example, observed data informs us about missing data through the model for data. Since covariates are typically considered fixed there is no covariate model in the standard case. We propose to use a version of the general location model (Olkin & Tate,1961; Krzanowski, 1980; see also Cho et al, 2001) to allow for both continuous and categorical missing covariates.

F. References

Aitkin, M. (1995). Probability model choice in single samples from exponential families using Poisson log-linear modelling, and model comparison using Bayes and posterior Bayes factors. Statistics and Computing, 5:113-120.

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

Banks, B. D., and Constantine, G. (1998). Metric Models for Random Graphs. Journal of Classification, 15: 199-223.

Batchelder, W.H., Kumbasar, E. and Boyd, J.P. (1997). Consensus analysis of three-way social network data. Journal of Mathematical Sociology, 22:29-58.

Bearman, Peter S., Moody, James, and Stovel, Katherine (2004). Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks, American Journal of Sociology, 110: 44-99.

Bernard, H. R., Killworth, P., Kronenfeld, D., Sailer, L., 1984. The problem of informant accuracy:the validity of retrospective data. Annual Review of Anthropology 13, 495-517.

Bernardo, J. M. and Smith, A. F. M. (1994). Bayesian Theory. New York: Wiley.

Berthelsen, K. K. and Møller, J. (2003). Likelihood and Non-parametric Bayesian MCMC Inference for Spatial Point Processes Based on Perfect Simulation and Path Sampling. Scandinavian Journal of Statistics, 30, 549–564.

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

Besag, J., York, J., and Molli´e, A. (1991). Bayesian Image Restoration, with two Applications in Spatial Statistics, Annals of the Institute of Statistical Mathematics, 43, 1–59.

Box, G. P. and Tiao, G. C. (1973). Bayesian Inference in Statistical Analysis. Reading, Massachusetts: Addison–Wesley.

Burt, R. S., (1987). A note on missing social network data in the General Social Survey. Social Networks, 9:63-73.

Butts, C.T. (2003). Network inference, error, and informant (in)accuracy: a Bayesian approach. Social Networks, 25:103-140.

Butts, C.T. (2007). Network Inference from Unstructured Sources. Presented at Sunbelt XXVII, International Sunbelt Social Network Conference, 8th European Social Network Conference, Chandris Hotel Complex at Dassia Bay, Corfu Island, Greece, May 1-6, 2007.

Cho, M., Schenker, N., Taylor, J. M. G. and Zhuang D. (2001). Survival analysis with longterm survivors and partially observed covariates. The Canadian Journal of Statistics, 29:421-436.

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.

Cressie, N. A. C. (1993), Statistics for Spatial Data, Wiley, New York, 2nd ed.

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.

Dahmström, K. and Dahmström, P. (1993). ML-estimation of the clustering parameter in a Markov Graph model, Research report, 1993:4, Stockholm University, Department of Statistics.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate Priors for Exponential Families, Ann. Stat. 7:269–281.

Dombroski, M.J. & K.M. Carley, (2002). NETEST: Estimating a Terrorist Network's Structure, Graduate Student Best Paper Award, CASOS 2002 Conference, Computational & Mathematical Organization Theory, 8, pp. 235-241.

Dombroski, M., P. Fischbeck & K.M. Carley, (2003). Estimating the Shape of Covert Networks In Proceedings of the 8th International Command and Control Research and Technology Symposium. Conference held at the National Defense War College, Washington DC. Evidence Based Research, Vienna, VA.

Frank, O. 1991. Statistical analysis of change in networks. Statistica Neerlandica, 45:283 -293.

Frank, O. and Snijders, T.A.B. (1994). Estimating the Size of Hidden Populations Using Snowball Sampling. Journal of Official Statistics, 10: 53-67.

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

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 E. Thompson (1992), Constrained Monte Carlo maximum likelihood for dependent data, Journal of the Royal Statistical Society, Series B, 54: 657–699.

Gile, Krista, and Handcock, Mark 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., Richardson, S., and Spiegelhalter, D. J., eds. (1996). Markov Chain Monte Carlo in Practice. London: Chapman & Hall.

Gu, M. G. and Zhu, H.-T. (2001), Maximum Likelihood Estimation for Spatial Models by Markov Chain Monte Carlo Stochastic Approximation, Journal of the Royal Statistical Society, Series B, 63, 339–355.

Handcock, M. S. (2002) Statistical Models for Social Networks: Inference and Degeneracy. Pp. 229 –240 in Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, edited by Ronald Breiger, Kathleen Carley, and Philippa E. Pattison. National Research Council of the National Academies. Washington, DC: The National Academies Press.

Handcock, M. S. (2003), Assessing degeneracy in statistical models of social networks, Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. Available from http://www.csss.washington.edu/Papers/

Hauck, W.W.; and Donner, A. (1977), Wald's Test as Applied to Hypotheses in Logit Analysis. Journal of the American Statistical Association, Vol. 72, No. 360. pp. 851-853.

Heikkinen, J. and Penttinen, A. (1999). Bayesian Smoothing in the Estimation of the Pair Potential Function of Gibbs Point Processes. Bernoulli, 5, 1119–1136.

Huisman, Mark (2007). Imputation of missing data in social networks. To be presented at Sunbelt XXVII, International Sunbelt Social Network Conference, 8th European Social Network Conference, Chandris Hotel Complex at Dassia Bay, Corfu Island, Greece, May 1-6, 2007.

Hunter, D.R., and Handcock, M. (2005). Inference in curved exponential family models for networks. To appear the Journal of Graphical and Computational Statistics.

Karabatsos, George and Batchelder, William H. (2003). Markov chain estimation for test theory without an answer key. Psychometrika, 68:373-389.

Koehly, L., & Pattison, P. (2005). Random graph models for social networks: Multiple relations or multiple raters. In P. Carrington, J. Scott, & S.Wasserman (Eds.), Models and Methods in Social Network Analysis. Canbridge University Press.

Koskinen, J.H. (2001). Aggregation of Perceived Social Networks. Sunbelt XXI, Budapest, Hungary. (http://wwwold.sztaki.hu/conferences/sunbelt21/22.html)

Koskinen, J. (2002). Bayesian analysis of perceived social networks. Research Report 2002:2, Department of Statistics, Stockholm University. (available from http://gauss.stat.su.se/site/pdfer/rr2002-2.pdf)

Koskinen, J. H. (2004a). Bayesian Analysis of Exponential Random Graphs - Estimation of Parameters and Model Selection. Research Report 2004:2, Department of Statistics, Stockholm University. (http://gauss.stat.su.se/site/pdfer/RR2004_2.pdf)

Koskinen, J. (2004b). Model selection for Cognitive Social Structures. Research Report 2004:3, Department of Statistics, Stockholm University. (available from http://gauss.stat.su.se/site/pdfer/RR2004_3.pdf)

Koskinen, J.H. (2006). Aspects of Bayesian Inference for (Curved) Exponential Families of distributions for Graphs and Digraphs - The Linked Importance Sampler Auxiliary Variable

(LISA) Metropolis Hastings for Distributions with Intractable Normalising Constants. SNAC, Department of Psychology, Melbourne University, December 2006 (http://www.sna.unimelb.edu.au/publications/SNAC_JHK_dec6_handout.pdf).

Koskinen, J.H: (2007a). Fitting models to social networks with missing data. Sunbelt XXVII, International Sunbelt Social Network Conference, 8th European Social Network Conference, Chandris Hotel Complex at Dassia Bay, Corfu Island, Greece, May 1-6, 2007.

Koskinen, J.H. (2007b). Inferring structural dynamics from insufficiently observed social networks. The fourth workshop on Network Tomography, May 23-24th in Adelaide, Australia.

Koskinen, J.H. (2007c). Missing data in social network analysis. The 8th Asia-Pacific Complex Systems Conference, July 2-5, 2007, Surfers Paradise, Gold Coast, QLD, Australia.

Koskinen, J.H. and Robins, G. (2007). Comparing networks: Associations and interrelations, non-parametric and Bayesian approaches (submitted for publication).

Koskinen, J.H., Jansson, I., and Spreen, M. (2002). The Role of Perceptional Data in Sampling Large Networks from Hidden Populations, in Jan Hagberg (ed.), Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in honour of Ove Frank. Stockholm: Dept. of Statistics, Stockholm University.

Koskinen, J.H., Wang, P., Lusher, D., and Robbins, G. (2007). Approximate Bayesian Analysis for Assessing Goodness of Fit in Exponential Random Graph Models. In preparation, Department of Psychology, Melbourne University.

Kossinets, G. (2006). Effects of missing data in networks. Social Networks, 28, 247-268.

Krackhardt, David (1987). Cognitive Social Structures. Social Networks, 9:109-134.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. Biometrics, 36: 493-499.

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.

Lazega, E. (2001). The collegial phenomenon: A structural theory of collective action among peers. Oxford University Press.

Lazega, E. and P. E. Pattison (1999), Multiplexity, generalized exchange and cooperation in organizations: a case study. Social Networks, 21: 6790.

Lindley, D. V. (1965). Introduction to Probability and Statistics from a Bayesian Viewpoint. Cambridge: Cambridge University Press. Vol. 1 and 2.

Lindsey, J.K. (1974). Comparison of probability distributions. Journal of the Royal Statistical Society, B, 36:38-47.

Little, R.J.A., and Rubin, D.B. (1987). Statistical Analysis with Missing Data, New York: Wiley.

Lusher, Dean Stewart (2006). Masculinities in local contexts: Structural, individual and cultural interdependencies. PhD thesis, Department of Psychology, University of Melbourne.
Marsden, P. (2005). Recent Developments in Network Measurement, pp 8-30 in Peter J. Carrington, John Scott, and Stanley Wasserman (eds.) Models and Methods in Social Network Analysis. New York: Cambridge University Press.

McPherson, Miller, Smith-Lovin, Lynn, and Cook, James M (2001). Birds Of A Feather: Homophily in Social Networks. Annual Review of Sociology, 27:415-444.

Meng, Xiao-Li (1994). Posterior Predictive p-values, The Annals of Statistics, 22:1142-1160.

Milgram, S. (1967). The small world problem. Psychology Today, 2, 60-67.

Møller, J., Pettitt, A.N. , Berthelsen, K. K. and Reeves, R.W. (2005). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants

Møller, J. and Waagepetersen, R. P. (2003a), “An Introduction to Simulation-based Inference for Spatial Point Processes,” in Spatial Statistics and Computational Methods, ed. Møller, J., Springer-Verlag, New York, Lecture Notes in Statistics 173, pp. 143–198.

Møller, J. and Waagepetersen, R. P. (2003b), Statistical Inference and Simulation for Spatial Point Processes, Chapman and Hall/CRC.

Neal, R. M. (2005). Estimating ratios of normalizing constants using Linked Importance Sampling, Technical Report No. 0511, Dept. of Statistics, University of Toronto. (available from http://arxiv.org/abs/math.ST/0511216)

Olkin, I. & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. The Annals of Mathematical Statistics, 32: 448-465.

Pattison, P.E. (2007). Estimating exponential random graph models from snowball sampled network data. Social Networks Workshop, University of Melbourne, January 12, 2006

Pattison, P. and G. Robins (2002). Neighborhood-based models for social networks. In R. Stolzenberg (Ed.), Sociological Methodology, Volume 32, Chapter 9, pp. 301–337. Boston: Blackwell Publishing.

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

Propp, J., and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statitical mechanics, Random Structures and Algorithms, 9:232-252.

Reeves, R. and Pettitt, A. N. (2003), Efficient Recursions for General Factorisable Models, manuscript submitted for publication.

Ripley, B. D. and Kelly, F. P. (1977). Markov Point Processes, Journal of the London Mathematical Society, 15, 188–192.

Robins, G. (1998). Personal Attributes in Interpersonal Contexts: Statistical Models for Individual Characteristics and Social Relationships. Ph.d. dissertation, University of Melbourne, Department of Psychology.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2006). An Introduction to Exponential Random Graph (p*) Models for Social Networks. Social Networks. Forthcoming. (preprint available from http://www.sna.unimelb.edu.au/publications/ERGM11.1.pdf)

Robins, G. L., Pattison, P. E., & Wasserman, S. (1999). Logit models and logistic regressions for social networks, III. Valued relations. Psychometrika, 64, 371-394.¿

Robins, G.L., & Pattison, P. (2005). Interdependencies and social processes: Generalized dependence structures. In Carrington, Scott & Wasserman (Eds) Models and Methods in Social Network Analysis (pp. 192-214). Cambridge University Press.

Robins, G.L., Pattison, P, & Elliott, P. (2001). Network models for social influence processes. Psychometrika, 66, 161-190

Robins, G.L., Woolcock, J., & Pattison, P. (2005). Small and other worlds: Global network structures from local processes. American Journal of Sociology, 110, 894-936.

Robins, G.L., Woolcock, J., & Pattison, P. (2004). Models for social networks with missing data. Social Networks, 26, 257-283.

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

Snijders, Tom A.B., Pattison, Philippa E., Robins, Garry L., and Handcock, Mark S. (in press). New specifications for exponential random graph models. Sociological Methodology.

Stork, D., Richards, W. D., (1992). Nonrespondents in communication network studies:Problems and possibilities. Group and Organization Management 17 (2), 193-209.

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

Strauss, D., and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. Journal of the American Statistical Association, 85, 204- 212.

Tallberg, C. (2003). Estimating the Size of Hidden populations. Research Report 2003:3, Department of Statistics, Stockholm University, http://gauss.stat.su.se/site/pdfer/rr2003-3.pdf.

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.

Tierney, L. (1994). Markov chains for exploring posterior distributions. Ann. Statist., 22:1701–1762. With discussion and a rejoinder by the author.

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

Wasserman, S., & Robins, G.L. (2005). An Introduction to Random Graphs, Dependence Graphs, and p*. In Carrington, Scott & Wasserman (Eds) Models and Methods in Social Network Analysis (pp. 148-161). Cambridge University Press.

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

Western, B. (1999). Special Issue on Bayesian Methods in the Social Sciences, (guest editor Western) Sociological Methods and Research, 28.

White, H. C., 1992. Identity and Control:A Structural Theory of Social Action. Princeton University Press, Princeton, NJ.

Wilson, D. (2000). How to couple from the past using a read-once source of randomness, Random Structures and Algorithms, 16:85-113.

Zellner, A. and Rossi, P. E. (1984). Bayesian analysis of dichotous quantal response models. Journal of Econometrics, 25:365-393.

G. Pseudo-code for fittting ERGMs to data with missing information on dyads

Routine LISAforMissingDyads

This routine produces a sequence ($\theta^{(t)}$, $\mathbf{V}^{(t)}$) from the posterior distribution given **U**. Here model specification, number of parameters, covariates, etc., are suppressed. Note that in implementation a lot of efficiency may be gained from tailoring the algorithm to the model considered. As an example a lot of efficiency can be gained in ERGMs from the fact that $q_\theta(X^*)/q_\theta(X)$ may be written in terms of the change statistics.

```
LISAforMissingDyads(X)

Initialize(X)

OldState:=X

for t:=1 to NumberOfSamples {
            output MetropolisHastingsUpdatingStep()

            for j:=1 to NumberOfElementsOfV {
                        output GibbsUpdatingStep()
                        OldState:=NewState(X,V$_{t,j}$)
            }
}
```

```
Initialize(X)
V:=<arbitrary state>
$\theta_0$:=MPLE(U,V)
H:=-Hessian(U,V,$\theta_0$)
ProposalVariance:=$\frac{c}{\sqrt{2+p}}$H$^{-1}$
N:=NumberOfActors(X)
/*It is usually efficient to set $\theta_0$ to the MLE and H to the Information Matrix
using importance sampling or Robins-Monro with the MPLE as starting value*/
/*How to set NumberOfIterations, K, m, burnin, and c depends on among other things
size of graph*/
return NumberOfIterations,K,m,burnin,$\theta_0$,V,N,ProposalVariance
```

```
MetropolisHastingsUpdatingStep()
X:=OldState
$\theta$*:=MultivariateNormal($\theta_{t-1}$,ProposalVariance)/*What proposal distribution is suitable
is case specific but for e.g. ERGMs this multivariate normal proposal is a convenient
choice*/
$\lambda(\theta^*,\theta_0)$:=LogLIS($\theta$*,$\theta_0$,K,m,OldState,burnin)
u:=Uniform
if log(u)< $\log q_{\theta^*}(\mathbf{X})/q_{\theta_{t-1}}(\mathbf{X})+\lambda(\theta^*,\theta_0)-\lambda(\theta_{t-1},\theta_0)$ {
            $\theta_t$:=$\theta$*
}
else {
            $\theta_t$:=$\theta_{t-1}$
}
return $\theta_t$
```

```
Uniform(A)
/*Generate a variate from the uniform distribution on the set A; If no argument
supplied generate a variate from the rectangular distribution on the unit interval*/
```

```
GibbsUpdatingStep()
X:=OldState
X*:=NewState(X,1-$V_{t-1,j}$)

Probability:= $\{1+q_{\theta_t}(\mathbf{X})/q_{\theta_t}(\mathbf{X}^*)\}^{-1}$
u:=Bernoulli(Probability)
$V_{t,j}$:= $V_{t-1,j}$ +u(1-2 $V_{t-1,j}$)
return $V_{t,j}$
```

```
NewState(X,$X_e$)
X*=X
/*Arguments symbolical shorthand for: set element e in X* equal to $X_e$ */
/*When network undirected e represents a dyad*/
return X*
```

```
LogLIS(θ*,$θ_0$,K,m,OldState,burnin)
for i:=1 to burnin {
            OldState:=MetropolisHastingsUpdatingStepGraph(θ*,OldState)
}
for j:=1 to m {
            θ(j):= θ*[1-(j-1)/(m-1)]+$θ_0$(j-1)/(m-1)/*θ is a smoth mapping bridging θ*
            and $θ_0$*/
}
for j:=1 to m {
            $\nu_j$:= Uniform([K])/*[K]={1,...,K}*/
            $\mathbf{Y}_j^{(\nu_j)}$:=OldState
            for k:= $\nu_j+1$ to K {
                        $\mathbf{Y}_j^{(k)}$:=MetropolisHastingsUpdatingStepGraph(θ(j), $\mathbf{Y}_j^{(k-1)}$)
                        $w_{j,j+1,k}=\sqrt{q_{\theta(j)}(\mathbf{Y}_j^{(k)})q_{\theta(j+1)}(\mathbf{Y}_j^{(k)})}\Big/q_{\theta(j)}(\mathbf{Y}_j^{(k)})$ /*NB when log[$q_θ$(X)]=
                        θ'z(X); $w_{j,j+1,k}$ = exp[1/2(θ(j+1)- θ(j))'z(Y)]*/
                        if j>1
                                    $w_{j,j-1,k}=\sqrt{q_{\theta(j)}(\mathbf{Y}_j^{(k)})q_{\theta(j-1)}(\mathbf{Y}_j^{(k)})}\Big/q_{\theta(j)}(\mathbf{Y}_j^{(k)})$
            }
            for k:= $\nu_j-1$ to 1 {
                        $\mathbf{Y}_j^{(k)}$:=MetropolisHastingsUpdatingStepGraph(θ(j), $\mathbf{Y}_j^{(k+1)}$)
                        $w_{j,j+1,k}=\sqrt{q_{\theta(j)}(\mathbf{Y}_j^{(k)})q_{\theta(j+1)}(\mathbf{Y}_j^{(k)})}\Big/q_{\theta(j)}(\mathbf{Y}_j^{(k)})$
                        if j>1
                                    $w_{j,j-1,k}=\sqrt{q_{\theta(j)}(\mathbf{Y}_j^{(k)})q_{\theta(j-1)}(\mathbf{Y}_j^{(k)})}\Big/q_{\theta(j)}(\mathbf{Y}_j^{(k)})$
            }
            if j < m {
                        $\mu_j$:=RandomWithProbabilities($(w_{j,j+1,k})_{k=1}^{K}/\Sigma\, w_{j,j+1,k}$)
                        OldState:= $\mathbf{Y}_j^{(\mu_j)}$
            }
}
λ(θ*,$θ_0$):= $\log\sum_{j=1}^{m-1}\sum_{k=1}^{K}w_{j,j+1,k}-\log\sum_{j=1}^{m-1}\sum_{k=1}^{K}w_{j+1,j,k}$
return λ(θ*,$θ_0$)
```

```
MetropolisHastingsUpdatingStepGraph(θ,OldState)
X:=OldState
Element:=Uniform($_NC_k$(N,2))/*When network directed the set is $N^{(2)}$ instead*/
X*:=NewState(OldState,1-$X_{Element}$)
u:=Uniform
```

```
if log(u)< $\log q_\theta(\mathbf{X}^*)/q_\theta(\mathbf{X})$ {
            OldState:=X*
}
return OldState
```

H. List of Figures