Title
Latent Order Logistic Models for Social Network Analysis & Causal Inference for Stochastic Networks

Author
Clark, Duncan Andrew

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Latent Order Logistic Models for Social Network Analysis

&

Causal Inference for Stochastic Networks

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Duncan Andrew Clark

2022

ABSTRACT OF THE DISSERTATION

Latent Order Logistic Models for Social Network Analysis

&

Causal Inference for Stochastic Networks

by

Duncan Andrew Clark

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2022

Professor Mark Stephen Handcock, Chair

This dissertation considers credible models for social network data. In the classical social network analysis setting, nodes are connected to each other with the connections between the nodes stochastic and interdependent. As a result social network modeling has a sample size of one, where each network observed must be regarded as a single realization of a stochastic process. I consider three distinct but related topics in this field in each of the chapters.

I first consider the Latent Order Logistic (LOLOG) model, a recently proposed model class for social network modeling. I take the data centric viewpoint, that is, how does the LOLOG model perform in terms of fit, qualitative interpretation and scientific conclusions on data that practitioners fit with commonly used exponential-family random graph models (ERGM).

I also propose a method for Bayesian inference for the LOLOG model. This method also yields insight into the posterior distribution of the latent edge ordering, which is fundamental to the LOLOG model. This approach allows for deeper insight into the LOLOG process, as well as insight into the reasons for LOLOG's desirable properties for social network analysis.

The final chapter of the thesis considers a separate topic. I consider an approach for causal inference with observational data in a network setting. Current approaches assume the

network is observed and exogenous. The main contribution of this chapter is to regard the edges and nodal covariates as jointly distributed in a causal model. We consider the causal structure of this problem, and provide an extension to current approaches to account for this.

Throughout this dissertation, I use real-data networks to demonstrate the ideas and methodology. Examples include the National Longitudinal Study of Adolescent to Adult Health and a well known network of New England Lawyers.

The dissertation of Duncan Andrew Clark is approved.

Yingnian Wu

Mason Alexander Porter

Qing Zhou

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Mark S. Handcock for advising me throughout the course of my studies. His exceptionally broad knowledge of statistiscal concepts always held an answer or new direction of approach, whilst his deep knowledge of the theory of social network modeling, as well practical implementation issues have guided me throughout this work. On account for his reading of this dissertation I will not add more for the sake of his modesty. Though I wish to thank him greatly for his patience with me and unending optimism, which always left me believing whichever problems I or the world in general was experiencing in that moment, were unlikely to be insurmountable.

I would also like to acknowledge that both the models studied in this disseration were innovations of Ian Fellows and Mark S. Handcock. I particularly wish to thank Ian for writing code for the ERNM and LOLOG models without which the work in this dissertation would not have been possible.

Lastly, I would like to thank Elizabeth for loving me, always being by my side, reminding me to smell the roses whilst balancing out my many shortcomings.

VITA

2011–2015	MSci Mathematics 1st Class Honours, University College London
2015–2017	Actuarial Analyst, Xafinity Consulting
2017–2018	Grader, Department of Statistics, UCLA
2018–2020	Teaching Fellow, Department of Statistics, UCLA
2020	Research Data Science Intern, Facebook Inc.
2021–2022	Graduate Student Researcher, Departments of Health Policy, UCLA
2021	Research Data Science Intern, Facebook Inc.

PUBLICATIONS

Comparing the real-world performance of exponential-family random graph models and latent order logistic models for social network analysis. *Journal of the Royal Statistical Society: Series A*, with Mark S. Handcock (2021)

Introduction

This dissertation consists of three self contained articles, on separate but related topics. The general theme binding them is the need for credible models, for inference from social network data. In these cases the primary feature of the process, is the lack of an often assumed statistical property: independently distributed data. For network data where ties between nodes are regarded as random, the "sociality" of such settings dictates that assuming independence between dyads is implausible. For example, friendships between people are universally understood to be related to other friendships of both prospective friends. As a result social network analysis must be regarded as an extreme case of small data, where the whole network is one observation, not many observations of independent dyads. As a result credible generative models for social network data have proven challenging to develop. It is to this literature that this dissertation primarily contributes.

The first chapter of the dissertation considers the recently proposed Latent Order Logistic (LOLOG) model for modelling social network data. We assess the real-world performance of these models when applied to typical networks modelled by researchers. Specifically, we model data from an ensemble of articles in the journal *Social Networks* with published Exponential Random Graph Model (ERGM) fits, and compare the ERGM fit to a comparable LOLOG fit. ERGMs are widely used in social network analysis and are typically interpreted as a snapshot of a network at a given point in time or in a final state whereas LOLOG directly allows for a latent network formation process. This chapter demonstrates that the LOLOG models are, in general, in qualitative agreement with the ERGM models, and provide at least as good a model fit. In addition, they are typically faster and easier to fit to data, without the tendency for degeneracy that plagues ERGMs. The results support the general use of LOLOG models in circumstances where ERGMs are considered.

In the second chapter, we derive a Bayesian approach for inference with LOLOG models, which requires consideration of the posterior distribution of a random edge ordering. Bayesian methods have been derived and shown to be useful for ERGM, however due to the latent edge ordering formulation of LOLOG, different methods are required to facilitate Bayesian inference for LOLOG models. The chapter gives example analyses on the well known Kapferer's Tailors network, Lazega's New England lawyers network and one of the schools in the Add-Health survey of adolescent health behaviours in schools. Sampling from the posterior distribution of edge orderings also yields the Maximum Likelihood Estimator (MLE) operationalized through the variational estimate of the MLE.

The third chapter of this dissertation connects the tools developed for social network analysis, to a rigorous framework for causal inference. This chapter extends an existing chain graph approximation to a directed acyclic graph (DAG) to express a possible causal structure for the evolution of a social network over time. The connection to social network modeling is made by noting that the chain graph may be modeled as an Exponential Random Network Model (ERNM). Bayesian machinery is employed to infer missing potential outcomes to estimate causal quantities of interest.

CHAPTER 1

Comparing the real-world performance of exponential-family random graph models and latent order logistic models for social network analysis

Exponential-family Random Graph models (ERGM) are widely used in social network analysis when modelling data on the relations between actors. ERGMs are typically interpreted as a snapshot of a network at a given point in time or in a final state. The recently proposed Latent Order Logistic model (LOLOG) directly allows for a latent network formation process. We assess the real-world performance of these models when applied to typical networks modelled by researchers. Specifically, we model data from an ensemble of articles in the journal *Social Networks* with published ERGM fits, and compare the ERGM fit to a comparable LOLOG fit. We demonstrate that the LOLOG models are, in general, in qualitative agreement with the ERGM models, and provide at least as good a model fit. In addition, they are typically faster and easier to fit to data, without the tendency for degeneracy that plagues ERGMs. Our results support the general use of LOLOG models in circumstances where ERGMs are considered.

1.1 Introduction

Social network analysis has become increasingly important in recent decades, with particular need in the social sciences to elucidate relational structure (Goldenberg et al., 2010). However, developing generative models for social networks has proven challenging (Chatterjee and Diaconis, 2013). Here we consider a social network a collection of fixed nodes, each with

fixed covariates and with edges stochastically present or absent between every pair of nodes. The chief problems for modelling such data are the vast space of possible networks and likely highly complex dependence structures of the network edges.

The Exponential-family Random Graph Models (ERGM) framework is widely used to represent the stochastic process underlying social networks (Frank and Strauss, 1986; Hunter and Handcock, 2006). ERGMs allow researchers to quantitatively evaluate the impact of local social processes and nodal attributes on the probability of edges between nodes forming. However, these models are prone to near-degeneracy (Handcock, 2003) and can not naively be applied to large networks (Schweinberger, 2011; Chatterjee and Diaconis, 2013). Model degeneracy is the application specific tendency of the model to concentrate probability mass on a small subset of graphs, especially those which are not similar to realistic networks for that application.

Much progress has been made on managing model degeneracy by introducing local neighbourhood structures (Schweinberger and Handcock, 2015) or tapering (Fellows and Handcock, 2017). The presence of degeneracy in many fitted ERGMs motivates the search for alternative model classes with similar or complementary modelling capacity that are less susceptible to these challenges.

While ERGMs are descriptive, they are often embedded as the equilibrium distribution of a social process. The Latent Order Logistic model (LOLOG) (Fellows, 2018b) is a related model that uses an edge formation process to develop a general probability model over the space of graphs. It is motivated by using the so-called change statistics, the change in the specified graph statistics resulting from toggling an edge on or off, as predictors in a sequential logistic regression for each possible edge. Noting that an ERGM specified with independent edge variables, reduces to a sequential logistic regression on its change statistics, ERGM and LOLOG are equivalent in the independent dyad case (Fellows, 2018b). LOLOG models also allow non-independent dyads, and graph statistics that depend on the order of edge formation, which result in models different than ERGM.

LOLOG models have the advantage that they are straightforward to sample from, and

can be used with simpler model terms, that would for an ERGM almost certainly result in near-degeneracy. This allows for a fast and user-friendly fitting procedure, with easily interpretable model terms.

How can we assess and compare differing model classes? Both ERGM and LOLOG are fully general and able to represent arbitrary distributions over the set of graphs (Fellows, 2018b, Theorem 1). As ERGMs are the equilibrium distribution of a relatively general Markov chain Monte Carlo (MCMC) process, there are many mechanisms that can lead to them, as there are for LOLOG. Hence both model classes have strong theoretical and modelling motivations, although the ERGM class to this point has been much more extensively explored (Schweinberger and Stewart, 2020; Schweinberger et al., 2020). In this paper, we provide a separate and novel contribution to the assessment on the model classes. Our objective is to compare the models by a pairwise assessment on the population of networks that the research community would choose to fit them on. The idea here is to move the perspective from that of the model viewpoint (i.e., given we have a model, what can we fit with it?) to a data-centric viewpoint (i.e., given that this is the data we have, what are the best modelling approaches?). The latter is the question facing the real-world users of these models, while the"inverse problem'' addressed by the former is commonly taken as it does not require the population of networks to be specified.

However, to take the data-centric viewpoint, we need to specify the population. We operationalised this in this paper by taking a population of networks that ERGM models have been applied to in the premier journal for publishing social network analyses, *Social Networks* (Everett and Valente, 2020). *Social Networks* is an interdisciplinary journal for those with "interest in the study of the empirical structure of social relations and associations that may be expressed in network form''. While the sub-population of networks in *Social Networks* for which ERGMs have been fit is a sample of the population of interest, we believe that it is a salient and (non-statistical) representative sub-population of the broader population.

Our selection of ERGM papers was at first a census of papers in the journal *Social Networks* using the ERGM framework, published from the journal's founding in 1979 up to and in-

cluding the January 2016 issue. Note that we have chosen a population of networks that are biased toward ERGM. These networks have successfully completed the peer-review process of *Social Networks*. In particular, the ERGM fit and analyses have passed peer-review and are deemed of sufficient scientific interest to appear in this premier journal. Clearly this is not sufficient to ensure the fit and models are appropriate for the data, although they represent a strong selectiveness relative to the population of networks that researchers would consider for analysis (without regard to a model class choice). Hence a comparable or competitive fit for LOLOG models to this sub-population presents stronger evidence for the value of LOLOG models than a comparison to the broader population. In particular it seems likely that in published papers that fit an ERGM model, ERGM performs well on this data set. Thus we expect a publication bias towards networks that suit ERGM well, which may not necessarily suit LOLOG well. We therefore suggest that good performance on data published with ERGM fits is a conservative indicator that LOLOG is a useful model for analysing social networks.

Identifying, assembling and fitting ERGMs and LOLOGs to an ensemble of networks, analysing their goodness of fit (GOF) and interpreting the results, is a significant undertaking. For brevity we give the fit of networks from a case-study (Sailer and McCulloch, 2012) in detail and provide summaries for the remaining networks in the supplement.

The structure of this paper is as follows. In Section 1.2 we briefly introduce ERGMs and LOLOG models, reviewing work in Fellows (2018b) as well as discussing the theoretical similarities and differences. Section 1.3 gives a description of the ensemble of networks and discusses the motivation for selecting such an ensemble. Section 1.4 shows both the LOLOG and ERGM fit of office layout networks with the data from Sailer and McCulloch (2012). Section 1.5 presents a summary of all the LOLOG and ERGM fits to each of the networks in the ensemble. Section 1.6 discusses the results of the fitting, as well as its implications regarding the utility of the LOLOG model.

1.2 ERGM and LOLOG Model Classes

Let Y be a random graph whose realisation is $y \in \mathscr{Y} = \{a \in \mathbb{R}^{n \times n} \quad | \quad \forall i, j \quad y_{i,i} = 0 \quad y_{i,j} \in \{0, 1\}\}$. We regard the number of nodes and any nodal covariates as fixed and known. For undirected networks, the additional restriction that $y_{i,j} = y_{j,i} \quad \forall i, j$ can be added. Let $|y| = n(n-1)$ denote the number of possible edges in y ($|y| = n(n-1)/2$ for undirected graphs). A dyad in a graph is a sub-graph of two nodes and any edges between them.

1.2.1 Model Specification

LOLOG and ERGM are alternative specifications of the distribution of Y. An ERGM for the network can be expressed as

$$p_E(y|\theta) = \frac{\exp(\theta \cdot g(y))}{c(\theta)} \qquad y \in \mathscr{Y} \tag{1.1}$$

where $g(y)$ is a $d-$vector valued function defining a set of sufficient statistics; $\theta \in \mathbb{R}^d$ is a vector of parameters; and $c(\theta)$ the normalising constant. Each ERGM family is defined by the choice of sufficient statistics.

These are chosen by the researcher, depending on domain knowledge, to specify the generating social processes. They can be any statistical summary of network properties and are typically motivated by social theory (Goodreau et al., 2009) or symmetry arguments (Strauss, 1986). In this way, ERGMs constitute a family of models across different choices of the sufficient statistics. Typically graph statistics are the density and degree counts, as well as nodal or edge covariate terms such as sociability and homophily (Morris et al., 2008). Geometrically weighed edgewise shared partner (GWESP) and geometrically weighted degree (GWDEG) terms are often included (Snijders et al., 2006) as they capture complex structure while reducing the effects of near degeneracy (Handcock, 2003). A very large number of terms are used by researchers in applications. Explicit definitions of almost all terms used in this paper can be found in Morris et al. (2008) or the documentation of Handcock

et al. (2018). Regardless of which sufficient statistics are used, the ERGM will have the maximal entropy of any distribution satisfying the d-dimensional mean constraints placed on $g(y), E[g(Y)] = \mu$. LOLOG models posit the existence of a latent discrete temporal dimension, $t = 1, \ldots, |y|$ so that the edges form in a sequence. Fellows (2018b) defines the latent random variables $Y_t,\ t = 1, \ldots, |y|$ representing the sequential formation of Y. Y_t has exactly t edges and is formed from Y_{t-1} by the addition of an edge. A LOLOG model is specified by two components, the first is the probability of observing a graph given a specified order of edge formation, s:

$$p(y|s, \theta) = \prod_{t=1}^{|y|} \frac{1}{Z_t(s)} \exp\left(\theta \cdot C_{s,t}\right) \tag{1.2}$$

where $s = \{s_1, s_2, \ldots, s_{|y|}\} \in \mathscr{S}_{|y|}$ is the set of possible edge formation orders with $|y|$ dyads, and

$$C_{s,t} = g(y_t, s_{\leq t}) - g(y_{t-1}, s_{\leq t-1}) \tag{1.3}$$

where $s_{\leq t}$ denotes the first t elements of $s \in \mathscr{S}_{|y|}$. The $C_{s,t}$ are the difference in the graph statistics from the y_{t-1} network to the y_t network and are informally called the "change statistics" of the formation process. The $Z_t(s)$ sequentially specify the normalising constants. Let y_t^+ be the graph y_{t-1} with the edge s_t added, then

$$Z_t(s) = \exp\left(g(y_t^+, s_{\leq t}) - g(y_{t-1}, s_{\leq t-1})\right) + 1 \tag{1.4}$$

The second component is the model for the edge order permutations, $p(s)$. The LOLOG distribution for Y is:

$$\begin{aligned} p_L(y|\theta) &= \sum_s p(y|s, \theta) p(s) \\ &= \sum_s \left(p(s) \prod_{t=1}^{|y|} \frac{1}{Z_t(s)} \exp\left(\theta \cdot C_{s,t}\right) \right) \end{aligned} \tag{1.5}$$

1.2.2 Model Interpretation

For the LOLOG model, conditioning on an edge permutation s, at each step t, we have $\text{logit}\left(p(y_t^+|s_{\leq t}, y_{t-1}, \theta)\right) = \theta \cdot C_{s,t}$. Thus, at each time t, conditional on the network already formed by that point, each dyad is a logistic regression on the change statistics associated with the edge. For ERGMs, equation (1.1) yields the auto-logistic interpretation of the θ parameter $\log\left(\frac{p(y_{i,j}^+|y_{i,j}^c,\theta)}{p(y_{i,j}^-|y_{i,j}^c,\theta)}\right) = \theta\cdot(g(y_{i,j}^+)-g(y_{i,j}^-))$, where $y_{i,j}^c$ is $y\backslash y_{i,j}$, $y_{i,j}^+ = y_{i,j}^c \cup \{y_{i,j}=1\}$ and $y_{i,j}^- = y_{i,j}^c \cup \{y_{i,j}=0\}$. Thus, conditional on the rest of the graph, each dyad can be thought of as an (auto)-logistic regression on change statistics. This gives a helpful interpretation for the parameters but does not help interpret the probability distribution of each edge unconditional of the rest of the graph.

1.2.3 Model Identifiability

Conditional on an edge ordering, $p(y|s,\theta)$ is an exponential family distribution, with θ identifiable subject to our chosen statistics $g(y)$ not containing any statistics that are linear combinations of each other (Barndorff-Nielsen, 1978, Corollary 8.1).

The edge ordering itself, whilst not the focus of our inference, is also identifiable. Proposition 1 in Ye et al. (2022), establishes identifiability of the causal structure of a DAG, under mild conditions, if the nodes are parametrized as generalized linear models (GLMs). Each edge in the LOLOG model is a Bernoulli family GLM with a logit link function, and predictors the change statistics (see equations 1.2 and 1.3). These change statistics only depend on the previous edges in the ordering (see equation 1.3). The LOLOG model is therefore a special case of the setting of Ye et al. (2022) where nodes in the DAG are the edge variables (y_t) and the ancestors are the previous edges in the edge ordering ($s_{\leq t}$).

However we note that for practical applications in the social network domain, we almost never observe enough data to make use of the methods of Ye et al. (2022). Each edge is a realization from a GLM and we typically observe a single realization of a network and thus observe each edge once. There are also a large number of edges, and hence nodes in the DAG

to be identified; for a network with n nodes $\mathcal{O}(n^2)$ edges. These properties of the typical social network analysis setting limit the applicability of DAG identification methods.

1.2.4 Model Estimation

Due to the intractability of summing over all possible edge permutations in the LOLOG model, the likelihood or likelihood ratio, cannot be evaluated and the maximum likelihood estimate (MLE) is intractable. Fellows (2018b) proposed a method of moments (MOM) approach to estimate model parameters. The idea is to seek θ_{MOM} such that $g(y) - \mathbb{E}_{\theta_{\text{MOM}}}[g(Y)] = 0$. Fellows (2018b) developed a Newton-Raphson approach as it is possible to differentiate the $\mathbb{E}_{\theta_{\text{MOM}}}[g(Y)]$ with respect to θ and approximate its value by sampling from the LOLOG model. Along with introducing LOLOG models in Fellows (2018b), the `lolog` R package (Fellows, 2018a) provides a sophisticated, fast and user-friendly method to fit LOLOG models to data.

ERGM parameters are typically estimated using an MCMC procedure to compute the MLE (Snijders, 2002; Hunter and Handcock, 2006). This is computationally demanding and there are sophisticated R packages available to perform this estimation (Handcock et al., 2018).

For both LOLOG and ERGM models we approximate standard errors derived from MCMC estimated inverse Fisher information matrices.

1.2.5 Model Discussion

A key advantage of the LOLOG model is the ease of simulation from the model. To simulate a network, we simply draw s from $p(s)$ and perform a sequential logistic regression simulation on the change statistics (Fellows, 2018a). The ERGM by comparison requires a full MCMC procedure to simulate networks (Handcock et al., 2018).

For LOLOG models we are required to model $p(s)$, the probability mass function (PMF) on the space of possible edge permutations. In the absence of strong substantive reason for a particular partial ordering of the edges, a uniform PMF can be used. However, many natural

reasons exist to constrain the edge ordering. For example, in schools that welcome a new cohort each year, edges in upper years could reasonably be constrained to have been formed before edges in lower years.

For ERGMs the interpretation is conditional on the entire rest of the network, whilst for LOLOG models with a specified edge ordering, the interpretation is conditional only on the network formed up until that point. We emphasise that the network formed up until that point will depend on the particular edge permutation. We note that in the case where the tie variables are independent, the edge ordering s does not matter (as the dyads do not depend on each another) and LOLOG reduces to logistic regression on change statistics, as does ERGM and to the same model.

We may also compare dyad-dependent ERGM and LOLOG models through their simulation algorithms A network from an ERGM is simulated through an MCMC procedure where dyads are considered conditional on the rest of the network. Often many thousands of steps are required to converge to the stationary distribution. As noted in Section 1.2.2, the log odds is equal to the inner product of the parameter and the change statistics of that dyad. The LOLOG model is formed by first sampling a dyad ordering, then starting with an empty network, adding an edge based on the log odds (being the inner product of the parameter and the change statistic). Each dyad is considered for edge formation and then the process is terminated leaving the simulated graph.

LOLOG considers each dyad exactly once, whereas the ERGM process can consider dyads multiple times for both edge formation and dissolution. We suggest a reason that LOLOG models do not suffer from the same degeneracy is that, in the simulation, each dyad is considered exactly once. This limits the scope for the explosive edge formation or deletion that often occurs when simulating from ERGM models.

More broadly, we argue that the LOLOG, motivated as a model with an easy simulation method with parameters that remain interpretable, is more desirable than ERGM. Whilst ERGMs are straightforward to write down, the MCMC procedures required to sample from them are more problematic.

Whist the LOLOG model has a latent temporal mechanism, through the edge ordering, it does not directly model temporal network dynamics. We primarily consider LOLOG as a method to draw inferences from a single observation of a social process, where the intermediate steps are unobserved. There has been much work on models for the temporal dynamics of networks. Utilizing exponential family models there are temporal exponential-family random graph models (TERGM) (Hanneke et al., 2010), and their widely used, special case separable exponential-family random graph models (STERGM) (Krivitsky and Handcock, 2014). There are also agent based models, particularly the widely used stochastic actor oriented models (SOAM) (Snijders, 1996). Outside of the statistical network modeling literature, there has also been significant interest in temporal dynamics of so called kinetic network models from statistical physics (Krapivsky et al., 2010), though this area focuses on large networks and their properties, rather than our focus on statistical inference of social processes in this dissertation. In addition, in recent years neural networks have been adapted for graph inputs (Scarselli et al., 2009), which removes the need for specification of statistics. Such latent representations are not dissimilar to the latent space modeling of Hoff et al. (2002), though the neural network framing allows for additional flexibility and computational advantages. These topics are not the focus of this dissertation, so we do not consider them further.

1.2.6 Assessing Goodness of Fit

For the interpretation of model parameters to be valid, we must show that the model is a plausible generating process for the observed network. In Section 1.4 we follow the goodness of fit procedure in Hunter et al. (2008). That is, we graphically compare the simulated distribution of chosen graph statistics to the observed values of those graph statistics. Whilst our models are highly parsimonious representations of complex social processes, the goodness of fit method highlights, that at a minimum, we should expect the observed statistics to be plausible realisations from a well-fitting model.

1.3 Description of the Ensemble

We considered papers in the journal *Social networks* where ERGMs were fit to data. We included papers up to and including the January 2016 issue. There were 45 such papers, of which we selected 18 papers as follows. First, we excluded bipartite ERGMs (5), we then included all networks with publicly available data (7) and selected a further 11 papers out of the remaining 33 based on their, subjectively assessed, novelty as well as the likely availability and ability to share data. We contacted the authors of the 11 papers and received the data for 7 of the papers. This gave an ensemble of 137 networks in 14 peer reviewed published papers, as many papers contained multiple networks. We note that 102 of these networks were from a single paper (Lubbers and Snijders, 2007), which were omitted from our analyses, leaving 35 networks. Table 1.1 shows a brief summary for each of the networks.

Our selection of ERGM papers was at first a census of papers in the journal *Social Networks* using the ERGM framework. The conclusions drawn from this study should be considered stronger than if the networks selected were sampled at random or through convenience. We do note that we did take a selective sample as described above as a first wave of networks to request data for, though this was also chosen based on our thoughts on which networks the authors would be able and willing to share.

We considered papers that used ERGMs for their statistical analyses as the ERGM class of models is arguable the most widely used descriptive statistical model for network analyses (Amati et al., 2018). Both LOLOG models and ERGMs are typically used to model global network structure using local network structure, thus comparing the two models is appealing. While both LOLOG and ERGM can represent any given PMF over the space of networks (Fellows, 2018b), specifying interpretable models that fit the data is often the practical challenge. There is no obvious reason to suspect similar performance in terms of fit and interpretability, when fit with similar network statistics, on the same network. In particular it seems likely that in papers published that fit an ERGM, ERGM should perform well thus we expect a publication bias towards networks that suit ERGM well compared to

Table 1.1: Properties of each network contained in the ensemble. The ensemble includes directed and undirected networks from various applications ranging in size from 16 nodes to 1681 nodes

Description	Network	Nodes	Edges	Directed	Citation
Add Health		1681	1236	Undirected	Harris et al. (2007b)
School Friends		Various	Varies	Directed	Lubbers and Snijders (2007)
Kapferer's Tailors		39	267	Undirected	Robins et al. (2007)
Florentine Families		16	15	Undirected	Robins et al. (2007)
German Schoolboys		53	53	Directed	Heidler et al. (2014)
Employee Voice	1	27	104	Directed	Pauksztat et al. (2011)
Employee Voice	2	24	53	Directed	Pauksztat et al. (2011)
Employee Voice	3	30	126	Directed	Pauksztat et al. (2011)
Employee Voice	4	31	139	Directed	Pauksztat et al. (2011)
Employee Voice	5	37	149	Directed	Pauksztat et al. (2011)
Employee Voice	6	39	155	Directed	Pauksztat et al. (2011)
Office Layout	University	67	211	Directed	Sailer and McCulloch (2012)
Office Layout	University	69	203	Directed	Sailer and McCulloch (2012)
Office Layout	Research	109	458	Directed	Sailer and McCulloch (2012)
Office Layout	Publisher	119	872	Directed	Sailer and McCulloch (2012)
Disaster Response		20	148	Directed	Doreian and Conti (2012)
Company Boards	2007	808	1997	Undirected	Wong et al. (2015)
Company Boards	2008	808	1740	Undirected	Wong et al. (2015)
Company Boards	2009	808	1682	Undirected	Wong et al. (2015)
Company Boards	2010	808	1622	Undirected	Wong et al. (2015)
Swiss Decisions	Nuclear	24	282	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Pensions	23	294	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Foreigners	20	169	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Budget	25	224	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Equality	24	248	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Education	20	227	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Telecoms	22	256	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Savings	19	138	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Persons	26	280	Directed	Fischer and Sciarini (2015)
Swiss Decisions	Schengen	26	316	Directed	Fischer and Sciarini (2015)
University Emails		1133	10903	Undirected	Toivonen et al. (2009)
School Friends	grade 3	22	177	Directed	Anderson et al. (1999)
School Friends	grade 4	24	161	Directed	Anderson et al. (1999)
School Friends	grade 5	22	103	Directed	Anderson et al. (1999)
Online Links	Hyperlinks	158	1444	Directed	Ackland and O'Neil (2011)
Online Links	Framing	150	1382	Undirected	Ackland and O'Neil (2011)

LOLOG. We therefore suggest that good performance on data published with ERGM fits, is a conservative indicator that LOLOG is a useful model for analysing social networks.

We also note that the LOLOG model allows for the consideration of information on the order of the edge formation within a network the researcher may have. This is currently implemented by allowing edge orderings to be constrained to those orderings compatible with the sequential adding of nodes to the network, followed by the consideration of all possible new edges. This is not possible in ERGM and few of the available networks had plausible ordering mechanisms. However, this may not be entirely due to the lack thereof: without the ability to model such an ordering process with ERGM, it seems likely that even if there is a compelling sequential node adding process the data would not be reported or even collected.

1.4 Case Study of LOLOG and ERGM fits: Complex networks where ERGM is insufficient

In this section we consider a case study from a single published paper where the networks in question are sufficiently complex to demonstrate that ERGM can be insufficient and LOLOG can help in modelling social network data.

We consider four networks of daily social interactions between workers within four different office spaces. An ERGM based analysis was originally carried out in Sailer and McCulloch (2012). Ties are present between person i and person j if person i reported daily social interaction with person j. Two of the networks are of a British university faculty before and after an office refurbishment, the remaining two are a German research institute and a corporate publishing company. The networks are directed and have 69, 63, 109 and 120 people/nodes, respectively.

The research question of interest in Sailer and McCulloch (2012) is the effect of spatial distance in the formation of social interactions within an office environment. The authors specified an ERGM with terms to represent the potential complex structure. These are listed

in the first column of Table 2 and detailed here.

The edges term models the overall propensity for social interactions, it has a similar role to an intercept term in regression. The reciprocity term measures the propensity for both people in a dyad to report social interaction with the other. The GWESP term, with decay parameter 0.5, is an integrated measure of the transitivity of social interactions (See Snijders et al., 2006, for a detailed explanation). The usefulness term is an edge-covariate term, with value equal to the sum over edges of the usefulness measure: for dyad (i, j) being person i's self-reported perception of the usefulness of person j. It measures the direct dependence of the propensity to have a social interaction on the usefulness of the person nominated. The team match term is the number of ties between people from the same team. It measures the propensity of teams to influence the density of social interaction. The floor match term is similar to team match, except it measures the importance of being on the same floor for social interaction. The metric distance term is the sum of the shortest walking distance in meters between the socially interacting peoples normal place of work. Similarly, the topo distance is the sum of measures of how far the desks could be perceived to be apart given the topography of the office (See Sailer and McCulloch, 2012, for precise definitions). The coefficients of the metric and topo distances measure the increase in log-odds of a social interaction given the distance they are apart. These coefficients are generally negative, indicating that social interactions become less common as the distance increases.

The best fitting model was then selected using the Akaike Information Criterion (AIC) and then a variety of different distance metrics were added individually as edge-covariates. The best model in terms of AIC was once again selected and analysed. Notably, no analysis of the goodness of fit for the models was provided.

1.4.1 Model Fits

We were able to recreate the selected ERGM fit for all four networks, shown in Table 1.2. The reciprocity coefficient is positive in two of the networks, indicating that the conditional

Table 1.2: Office layout ERGM fits as per the published results. In all cases the selected measure of distance is negative and significant suggesting that close office workers, are more likely to interact, even after allowing for team, floor, usefulness as well as social structure in the form of reciprocity and transitivity.

	University 2005	University 2008	Research Institute	Publisher
Edges	-3.4 (0.37)***	-4.41 (0.2)***	-4.1 (0.12)***	-5.07 (0.15)***
Reciprocity	0.38 (0.45)	0.62 (0.31)***	2.39 (0.2)***	-1.26 (0.19)***
GWESP(0.5)	1.36 (0.14)***	1.24 (0.11)***	0.92 (0.07)***	2.09 (0.09)***
Usefulness	0.7 (0.15)***	0.54 (0.11)***	0.81 (0.04)***	1.31 (0.05)***
Team Match	0.78 (0.18)***	0.56 (0.1)***	NA	NA
Floor Match	0.15 (0.26)	0.58 (0.14)***	NA	NA
Metric Distance	-0.04 (0.01)***	-0.01 (0)***	-0.01 (0)***	NA
Topo Distance	NA	NA	NA	-0.06 (0)***

*** p-value < 0.001 , ** p-value < 0.01, * p-value < 0.05

log-odds of a social interaction is positive if the social interaction is mutual. The GWESP coefficient is positive for all four networks, indicating that the log-odds of a social interaction existing is positive if the social interaction increases this measure of transitivity. The usefulness coefficient is positive for all four networks, indicating that the log-odds of a social interaction existing is positively related to the usefulness of the nominated person. The team match coefficient is positive for all four networks, indicating that the log-odds of a social interaction existing is positive if the social interaction is within the same team (as distinct from between people in different teams). Floor match coefficients are also positive, indicating that the log-odds of a social interaction existing is positive if the social interaction is within the same floor (as distinct from between people in different floors). The metric and topo coefficients are generally negative, indicating that social interactions become less common as the distance increases.

Overall, the Sailer and McCulloch (2012) concluded that daily social interactions of people in offices exhibit a tendency for mutuality and social closure. Interactions are also more likely to occur where there is a high level of usefulness of the receiver to the sender as well as within teams. While being on the same floor plays a role in some cases, the distance apart plays a role in all cases, with social interactions more likely for people closer together.

Table 1.3: Office layout LOLOG fit with the same terms as the published ERGM. Model fits show broad quantitative agreement with the published results using the ERGM in Table 1.2

	University 2005	University 2008	Research Institute	Publisher
Edges	-1.69 (0.38)***	-3.67 (0.36)***	-3.18 (0.13)***	-1.63 (0.09)***
Reciprocity	1.99 (0.34)***	1.96 (0.31)***	3.9 (0.25)***	0.64 (0.2)***
GWESP(0.5)	0.55 (0.12)***	0.87 (0.13)***	0.73 (0.09)***	-0.22 (0.06)***
Usefulness	1.02 (0.15)***	0.81 (0.14)***	1.21 (0.05)***	1.89 (0.06)***
Team Match	1.29 (0.19)***	0.72 (0.19)***	NA	NA
Floor Match	-0.28 (0.3)	1.08 (0.29)***	NA	NA
Metric Distance	-0.07 (0.01)***	-0.02 (0.01)***	-0.02 (0)***	NA
Topo Distance	NA	NA	NA	-0.1 (0)***

*** p-value < 0.001 , ** p-value < 0.01, * p-value < 0.05

We were able to obtain LOLOG fits with the same covariates, as the ERGM fits for all networks, we summarise the fits in Table 1.3. In addition, we show the LOLOG fit using GWESP, 2- and 3- in- and out-stars, together with all covariate matches and metric distance in Table 1.4. For $k = 1, 2, \ldots,$ a k-out-star centred on a node i and a set of k different nodes $\{i_1, \ldots, i_k\}$ such that the tie from i to i_j exists for $j = 1, \ldots, k$. The k-out-star statistic is the number of distinct k-out-stars in the network (i.e., summing over the centring nodes). The k-in-star statistic is the same except the ties from i_j to i must exist for $j = 1, \ldots, k$ (rather than the out-ties from i). As noted in Section 2.2, the qualitative interpretation of the LOLOG coefficients is similar to ERGM with the primary difference being the log-odds is conditional on the network at the point the edge is added. We directly compare the qualitative fits in Section 4.3.

We were also able to fit LOLOG models to each of the networks when the GWESP term is replaced with a triangle term. This is not possible with ERGM due to near-degeneracy. We summarise this in Table 1.5. The estimated standard errors for the Publisher network are very high, suggesting there is great uncertainty in the data generating process. The estimated standard errors for the mutual and triangle terms for the University in 2005 and 2008 are also high though not as severe and they fall out of significance for these model fits.

Table 1.4: Office layout LOLOG fit with GWESP and 2- and 3- in- and out-stars. Significant out-star terms may suggest there is social structure unaccounted for with just the published ERGM terms. Despite additional significant structural terms, the LOLOG models still show broad quantitative agreement with the published results using ERGM

	University 2005	University 2008	Research Institute	Publisher
Edges	-3.2 (0.67)***	-5.04 (0.59)***	-4.04 (0.22)***	-4.87 (1.19)***
Reciprocity	2.03 (0.77)***	1.11 (0.45)***	4.7 (0.52)***	3.16 (1.27)***
GWESP(0.5)	0.33 (0.2)	0.49 (0.16)***	0.77 (0.11)***	0.01 (0.26)
Out-2-Star	1.39 (0.26)***	0.65 (0.16)***	0.41 (0.07)***	0.69 (0.15)***
Out-3-Star	-0.28 (0.07)***	-0.07 (0.03)***	-0.04 (0.01)***	-0.02 (0)***
In-2-Star	0.26 (0.22)	0.25 (0.15)	0.21 (0.12)	0.73 (0.54)
In-3-Star	-0.04 (0.05)	-0.03 (0.02)	-0.09 (0.03)***	-0.18 (0.1)
Usefulness	1.07 (0.2)***	0.75 (0.16)***	1.28 (0.07)***	2.98 (0.61)***
Team Match	1.93 (0.31)***	1.14 (0.25)***	NA	NA
Floor Match	-0.24 (0.47)	1.35 (0.43)***	NA	NA
Metric Distance	-0.09 (0.01)***	-0.02 (0.01)***	-0.02 (0)***	NA
Topo Distance	NA	NA	NA	-0.24 (0.06)***

*** p-value < 0.001 , ** p-value < 0.01, * p-value < 0.05

As the triangle term increases the estimated standard errors and does not improve the GOF (see next section), we suggest using the GWESP term.

We also fitted the LOLOG model where the people are added in the order of their average usefulness, as reported by the other people. As we suspect more useful people may have been in the office longer or should be the first point of contact for new employees, we suggest this as a plausible ordering mechanism. The fit was comparable to the fit without the ordering, and the GOF was not improved, so we do not consider it further.

We tried to fit an ERGM model with the in- and out- geometrically weighted degree (GWDEG) terms but this was degenerate for the University 2005 and 2008 networks. The in-GWDEG term adds one network statistic to the model equal to a weighted sum of the in-degree counts with weights decreasing geometrically. The out-GWDEG is similar with the out-degree counts (See Hunter, 2007, for a detailed explanation). For the Research institute and Publisher out-GWDEG was negative and significant, in line with the LOLOG model positive 2-star and negative 3-star parameters. However, the fit was still poor and inferior

Table 1.5: Office Layout LOLOG fit with triangles instead of gwesp term, shows broad quantitative agreement with the published results on nodal covariates, however, suggests little tendency for reciprocity and transitivity in the university networks.

	University 2005	University 2008	Research Institute	Publisher
Edges	-2.05 (0.82)***	-3.9 (0.67)***	-3.36 (0.15)***	-5.4 (49.3)
Reciprocity	-2.96 (5.83)	-0.08 (1.6)	3.34 (0.36)***	-24.8 (367.3)
Triangles	2.69 (2.95)	1.2 (0.73)	0.61 (0.13)***	3.71 (50.63)
Usefulness	1.35 (0.53)***	0.83 (0.16)***	1.21 (0.06)***	6.57 (88.39)
Team Match	1.8 (0.85)***	0.9 (0.35)***	NA	NA
Floor Match	-0.29 (0.88)	1.07 (0.46)***	NA	NA
Metric Distance	-0.1 (0.05)***	-0.02 (0.01)***	-0.02 (0)***	NA
Topo Distance	NA	NA	NA	-0.39 (5.82)

*** p-value < 0.001 , ** p-value < 0.01, * p-value < 0.05

to the LOLOG model. We do not comment further on this, though it is reassuring that the ERGM with GWDEG gives similar interpretations to LOLOG with star terms. The GWDEG terms were not discussed in Sailer and McCulloch (2012).

We note the computation time difference in the LOLOG and ERGM parameter estimation. We ran each with a single core with Intel(R) Xeon(R) Platinum 8160 CPU @ 2.10GHz processor. The recreated ERGM took around 35 seconds, and the LOLOG took around 8 seconds. For larger networks, we found parallelisation in the network simulation step of the fit to be extremely helpful for both the LOLOG and ERGM models. From our experience for larger networks the performance differential between LOLOG and ERGM can be much greater, in particular when the ERGM MCMC simulation is computationally expensive.

1.4.2 Goodness of Fit

Firstly, we consider the goodness of fit for the published ERGM model, and the LOLOG model with the same terms. Figures 1.1 show the comparison of simulated distribution of the in-degree with the observed network statistics. Figures 1.3 and 1.4 contained in Appendix 1.8.1 show the same comparison for edgewise shared partners (ESP) and out-degree.

Table 1.6 shows comments on the goodness of fit for each network, using the recreated

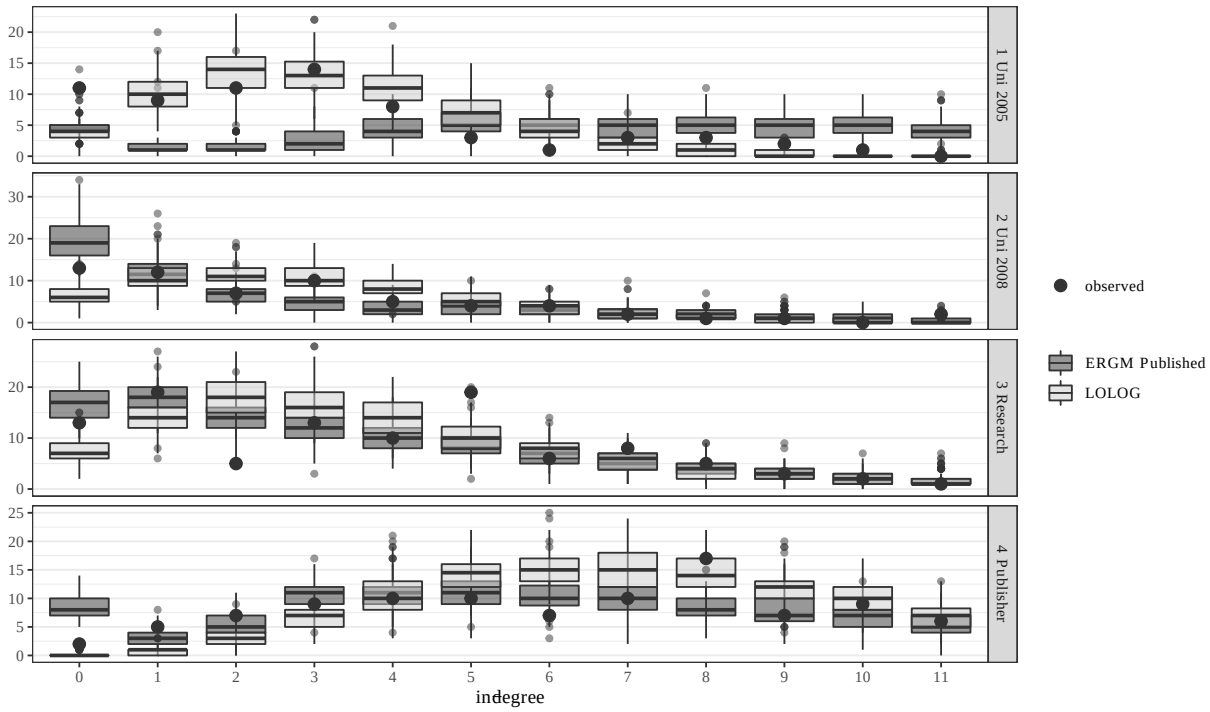


Figure 1.1: In-degree goodness of fit comparison plot for Office layout networks. The comparison is between ERGM and LOLOG fits both using the published ERGM terms

Table 1.6: Summary of GOF for ERGM and LOLOG with published terms for Office layout networks. For all networks neither the LOLOG model nor ERGM provide satisfactory fit.

Network	ERGM	LOLOG
2005 University	Fits poorly on out-degree Fits poorly on ESP	Fits poorly on in-degree Fits poorly on ESP but much better than ERGM
2008 University	Fits poorly on out-degree Fits poorly on ESP	ERGM convex, LOLOG concave on in-degree Fits poorly on out-degree Fits poorly on ESP
Research Institute	Fits poorly on out-degree Fits poorly on ESP	Fits poorly on in-degree Fits poorly on out-degree Fits poorly on ESP
Publisher	Fits poorly on in-degree Fits poorly on out-degree Fits poorly on ESP	Fits poorly on in-degree Fits poorly on out-degree Fits poorly on ESP

published ERGM and the LOLOG model with published ERGM terms. Where no comment is made for any of the goodness of fit terms or any model, the model fits well on that statistic.

All models for all networks have a least one of the in-degree, out-degree or ESP statistic of the observed network not being a typical value for the fitted models. As a result, the models do poorly on recreating networks similar to the observed, and thus inference based on the parameter estimates and standard errors should be treated with caution. In particular we note that the LOLOG model with identical terms to the published ERGM does not seem to help improve the fit for any of the networks in question here.

We also show the GOF for in-degree the LOLOG model with GWESP and 2- and 3- in- and out-stars for each model in Figure 1.2. Figures 1.5 and 1.6 contained in Appendix 1.8.1 show the plots for out-degree and ESP. We note here that all models fit the in-degree distribution well, all models except for the Publisher network fit the out-degree distribution well and the University 2005 and 2008 models fit well on the ESP distribution. This is an improvement in all cases versus the ERGM models published in Sailer and McCulloch (2012).

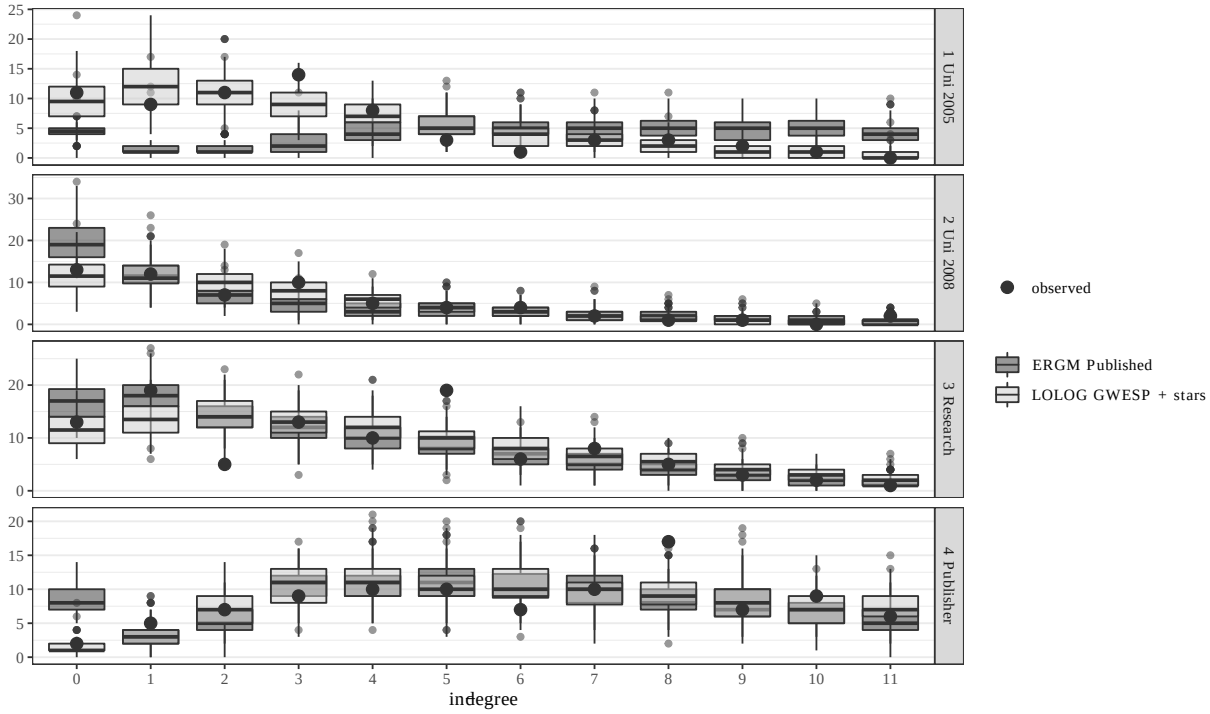


Figure 1.2: In-degree goodness of fit comparison plot for Office layout networks. The comparison is between the published ERGM and LOLOG fits with GWESP and stars terms included.

1.4.3 Model Comparison

These networks are of particular interest as they represents a real world-cases of applied researchers seeking a statistical tool to represent and test their social hypotheses and analyse their collected data. Good performance in such settings for the LOLOG model suggests the model could be of real use to the applied social network research community. Using these four complex networks as an example, helps us to present the the utility of the LOLOG model. The ERGM and LOLOG models with the terms as in Sailer and McCulloch (2012) produced the same qualitative interpretation.

However, it is important to note that neither the specific LOLOG or the ERGM fitted the data well in terms of all in-degree, out-degree and ESP distributions. Therefore, the models are not capturing basic aspects of the observed network data and the above interpretation should be treated with caution. In particular, the Publisher network proved especially hard to fit.

Using the triangle term in the LOLOG model in place of the GWESP term did not improve the fit. Including 2- and 3- in- and out-star terms yields models that fit much better on the in- and out-degree distribution as well and the ESP distribution. We therefore have a stronger belief that inferences from these models are valid. They show similar conclusions to the published ERGM, though in addition we observe a significant positive out-2-star coefficient and a significant negative out-3-star coefficient, suggesting that there is a tendency for some people to have social interactions with many more people than others. This tendency for super-daily interactors was not captured in the published ERGM fit. We also note that the lack of a significant in-2-star parameter suggests that there is not a corresponding tendency for some people to attract more interactions, when their usefulness had already been accounted for. We can infer that perhaps there is a surplus of unwanted daily interaction due to people with a tendency for high out-degree. Thus, the LOLOG model allowed for a better fit, as well as a deeper interpretation of the social interaction process.

1.5 Summary of Results for the Ensemble

The comparison of the value of models rarely will come down to a quantitative measure on a single dimension. The social processes that produce network data are typically complex and our choice of which data to analyse tends to favour complex structures. The models typically only approximate that structure and some features of the data are not represented in the models. Scientists that model social network data typically have multiple objectives, with some models more suited to some of those objectives rather than others. Having said this, we constructed a rubric of criteria to assess the models, both relatively and absolutely. We follow each criterion with a brief justification for why it was included.

1. Are we able to recreate the published ERGM qualitatively?
   We asked this to screen out network data where our usage differs qualitatively from the original, for whatever reason. This is to help ensure we were using the data correctly, so that our comparison is valid.

2. Do the recreations of the published ERGM fit the network well?
   This is to assess the validity of the published ERGM results, and to assess if ERGM is a good model for the published case study.

3. Are we able to fit the LOLOG with the published ERGM terms?
   This is to assess the LOLOG on terms likely favourable to the ERGMs. Typically, published ERGMs will have undergone model selection criteria to choose terms that had better fit compared to other possible ERGMs. This criterion assesses the flexibility of the LOLOG model class.

4. Does the LOLOG model with the published ERGM terms fit well?
   This is an absolute measure of the LOLOG goodness-of-fit with the ERGM terms.

5. Are we able to fit the LOLOG model with ERGM Markov terms (that are often degenerate in ERGM)?
   Markov terms, such as k-stars and triangles, often lead to near-degenerate models

despite their conceptual appeal (Frank and Strauss, 1986). This criterion assesses if the LOLOG can aid in interpretability by using simpler terms that are not possible in ERGM.

6. Is a better fit achieved with LOLOG than the published ERGM?

   This is a direct comparison to judge if the LOLOG is a better model for the observed data than the published ERGM.

7. Do the published ERGM and best-fitting LOLOG models have consistent interpretations?

   This assesses if qualitative substantive conclusions drawn from each model are consistent with the other. If affirmative, this gives some confidence that qualitative conclusions are not simply an artifact of the chosen modelling approach.

8. Which model do we believe to be more useful?

   This is a subjective judgement criterion. A major component is the goodness-of-fit criteria (Section 2.5). These criteria measure the degree to which important statistical characteristics of the network data are reproduced by the model. These focus on characteristics not explicitly in the model. A second component is the substantive interpretability of the terms (i.e., are they socially salient?). A third is the complexity of the model terms (i.e., the value of simplicity).

Table 1.7 provides a summary of the ERGM and LOLOG model fits for the networks in our ensemble, the columns are binary answers (1=Yes, 0 = No), to the above criteria. The fits were carried out in R using the `ergm` package (Handcock et al., 2018), and the `lolog` package (Fellows, 2018a). For the GWESP, GWDSP and GWDEG terms decay parameters were used as stated. If they were not available, $\alpha = 0.5$ was used.

Finally, we make some general comments regarding the significant amount of information on the hundreds of models fitted to the data that we gathered, more detailed summaries for each individual network are contained the supplement. More detailed overall comments on the study are in the discussion in Section 1.6.

Table 1.7: Summary table for LOLOG and ERGM Fits. Criteria a-g are defined and explained in the text and broadly address general modelling concerns.

Description	Network	Nodes	a	b	c	d	e	f	g	h
Add Health		1618	1	0	1	0	1	1	1	LOLOG
School Friends		Various								
Kapferer's Tailors		39	1	0	1	0	1	1	0	LOLOG
Florentine Families		16	1	1	1	1	1	1	0	ERGM
German Schoolboys		53	1	1	0	NA	1	1	1	Both
Employee Voice	1	27	0	NA	1	1	1	1	NA	LOLOG
Employee Voice	2	24	1	1	0	NA	0	0	NA	ERGM
Employee Voice	3	30	0	NA	1	1	1	1	NA	LOLOG
Employee Voice	4	31	0	NA	1	1	1	1	NA	LOLOG
Employee Voice	5	37	0	NA	1	1	1	1	NA	LOLOG
Employee Voice	6	39	0	NA	1	1	1	1	NA	LOLOG
Office Layout	University	67	1	0	1	0	1	1	1	LOLOG
Office Layout	University	69	1	1	1	0	1	1	1	LOLOG
Office Layout	Research	109	1	1	1	0	1	1	1	LOLOG
Office Layout	Publisher	119	1	0	1	0	1	1	1	LOLOG
Disaster Response		20	0	0	0	0	1	1	0	LOLOG
Company Boards	2007	808	0	0	0	0	1	1	NA	LOLOG
Company Boards	2008	808	0	0	0	0	1	1	NA	LOLOG
Company Boards	2009	808	0	0	0	0	1	1	NA	LOLOG
Company Boards	2010	808	0	0	0	0	1	1	NA	LOLOG
Swiss Decisions	Nuclear	24	0	1	0	NA	1	1	1	ERGM
Swiss Decisions	Pensions	23	0	1	1	0	1	0	0	ERGM
Swiss Decisions	Foreigners	20	0	1	0	NA	1	0	0	ERGM
Swiss Decisions	Budget	25	0	1	0	NA	1	1	0	ERGM
Swiss Decisions	Equality	24	0	0	0	NA	1	1	0	LOLOG
Swiss Decisions	Education	20	0	0	1	0	1	1	NA	LOLOG
Swiss Decisions	Telecoms	22	0	0	0	NA	1	1	NA	LOLOG
Swiss Decisions	Savings	19	1	1	0	NA	1	1	0	ERGM
Swiss Decisions	Persons	26	0	1	0	NA	1	1	0	ERGM
Swiss Decisions	Schengen	26	0	0	0	0	1	1	NA	LOLOG
University Emails		1133	0	0	0	0	0	0	NA	Neither
School Friends	grade 3	22	1	0	0	0	1	1	NA	LOLOG
School Friends	grade 4	24	1	0	0	0	1	1	NA	ERGM
School Friends	grade 5	22	1	0	0	0	1	1	NA	ERGM
Online Links	Hyperlinks	158	1	0	1	0	1	1	1	LOLOG
Online Links	Framing	150	1	0	1	0	1	0	1	LOLOG
Column Proportion	NA	NA	0.43	0.37	0.46	0.23	0.94	0.86	0.5	NA

Overall, we see that in many cases, we were not able to recreate the published ERGM (Table 1.7 column a), and often when we could, the model did not fit the data well using the GOF methodology of Hunter et al. (2008) (Table 1.7 column b). We were sometimes able to use the same terms as the published ERGM to fit a LOLOG model, however there were also some networks where we could not fit the LOLOG model with ERGM terms.

Where it was possible to fit a LOLOG model with the published ERGM terms,the model usually did not fit the data well (Table 1.7 column c). However, in almost all cases, we were able to fit the LOLOG model, with terms that usually result in degenerate ERGMs e.g., triangles and stars (Table 1.7 column e), and usually could achieve at least as good a fit as the published ERGM (Table 1.7 column f). Where it was possible to fit both a LOLOG and ERGM model the qualitative interpretations were equivalent on all parameters for half of the networks (Table 1.7 column g).

In general, our experience in fitting the LOLOG model was that it was easier and faster to fit than ERGM (Table 1.7 column h). The MOM estimation typically required little to no tuning, in contrast to MLE computation for ERGM models. In addition, the triangles and star terms that can be readily fit with LOLOG models provide a simple and intuitive interpretation for users of the model.

1.6 Discussion

We have shown that the LOLOG model can be fit to most members of an ensemble of network data sets that have published ERGM fits in the journal *Social Networks*. We report a case-study of a complex data set and show that the LOLOG model is at least the equal of the ERGM, in terms of goodness of fit and interpretability. We carried out fits to 35 networks in total and provided a summary of each of the networks' fits. We regard this as strong evidence that the LOLOG model is a useful model for modelling real social network data, as journal articles with published ERGM fits likely have a selection bias towards data sets that are well suited to ERGMs.

In carrying out this study, we have gained a great deal of practical experience in the types of tasks for which ERGMs are used, as well as practical problems in fitting them, in particular code run time and degeneracy issues. We have found the LOLOG model to be in general more user-friendly and faster to fit, leading to easier identification of poor models, and a much faster data analysis procedure. These benefits should not be overlooked, in particular, when social network analyses are often of interest to applied researchers whose expertise is not statistical modelling. As a result, LOLOG models seem particularly better suited to feasibly analysing larger networks, which whilst possible to fit with ERGMs (Stivala et al., 2020), often require significant tuning and computational resources.

LOLOG models can usually be fit with terms that are almost always degenerate for ERGMs on even small networks. Using this greater flexibility of specification, we were often able to achieve a better fit. In addition, the need to use complex geometrically weighted statistics is reduced, aiding interpretability of the LOLOG model. In practice we also believe LOLOG models could facilitate more robust model selection procedures. The degeneracy issues of ERGM as well as the time taken to fit the model, can result in researchers omitting terms based on their degeneracy, as well as considering fewer models than they would want. The fast fit and robust to degeneracy properties of the LOLOG model should help alleviate these practical issues. This should increase the scope of terms that researchers use, as they can focus on their representation of the underlying social processes rather than being restricted by computational and class specific representation issues.

We have also seen that qualitative interpretations of analyses carried out with both ERGMs and LOLOG models are generally in agreement. We do note, however, from our experiences that the LOLOG model applied to small networks can result in parameter estimates with high variance, where the ERGM model parameters have lower variances, more amenable to interpretation.

The goodness of fit of the LOLOG models also compare favourably with the ERGMs, with little drop in quality, for the same terms. In particular with the ability to use simpler terms for the LOLOG model we were often able to achieve improved fit over the published ERGMs

in the ensemble of networks that we fit.

The LOLOG model has the advantage of being able to account for edge orderings. We believe that this may be helpful for analysing network data, although we have not seen clear benefits in the ensemble of network data in this study. It is worth noting that there are many settings where the ability to model the edge ordering process is a great advantage of the LOLOG model. A clear case is citation networks where the temporal directional is fundamental (McLevey et al., 2018). Another case is where preferential attachment type processes are thought to be strong. A third is where the edge ordering is known exactly or thought to be strongly influenced by a covariate or contingency. The further consideration of edge ordering processes is beyond the scope of this paper. However, we hope the availability of a latent ordering network model like LOLOG, will spur the development of edge ordering processes models. We also note that the LOLOG model is a fully general model in the sense that it can represent any distribution over the space of networks. Therefore, even if it is hard to justify such an edge formation procedure, the LOLOG model may still be a useful approach to understanding the social processes producing network data.

All analysis was done in the R environment (R Core Team, 2020) primarily with the `lolog` (Fellows, 2018a) and `ergm` packages (Handcock et al., 2018). The code and available data to reconstruct the analyses of this paper are available at https://github.com/duncan-clark/lolog_catalog_paper/tree/main/example_fit.

1.7 Acknowledgements

The project described was supported by grant number 1R21HD075714-02 from NICHD, and grant numbers SES-1230081 and IIS-1546300 from the NSF.

We would like to acknowledge and thank all of the authors that provided data that made this study possible. We would like to thank the following, for taking the time to correspond with us and for providing their data: Greetje Van Der Werf, Lotte Vermeij, Miranda Lubbers, Mikko Kivelä, Riitta Toivonen, Jari Sarimäki, Jukka-Pekka Onnela, Robert Ackland, Birgit

Pauksztat, Kerstin Sailer, Dean Lusher, André Gygax, Roger Guimera, and Manuel Fischer. We note that there is uncertain personal benefit as well as some risk in doing so. We greatly appreciate their time and effort in preserving their data and providing it when we requested. They have made significant contributions to reproducibility of research in its many forms.

1.8 Appendix

1.8.1 Additional Goodness of Fit Figures

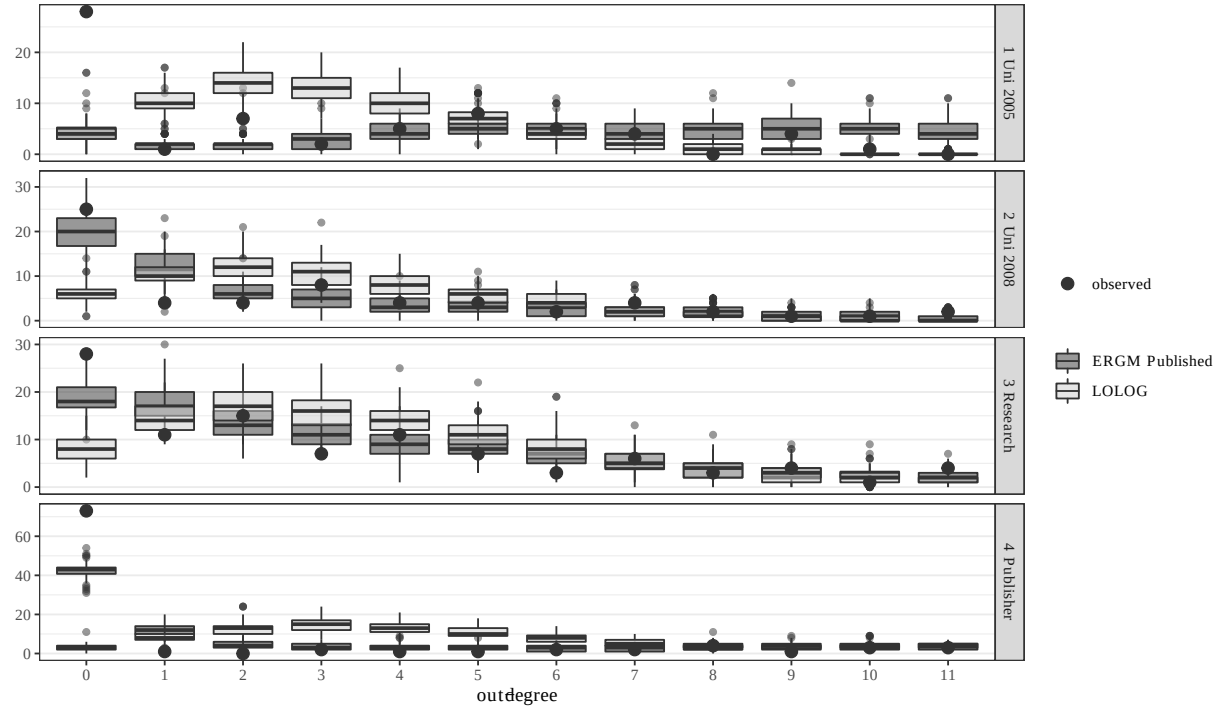


Figure 1.3: Out-degree goodness of fit comparison plot for Office layout networks. The comparison is between ERGM and LOLOG fits both using the published ERGM terms

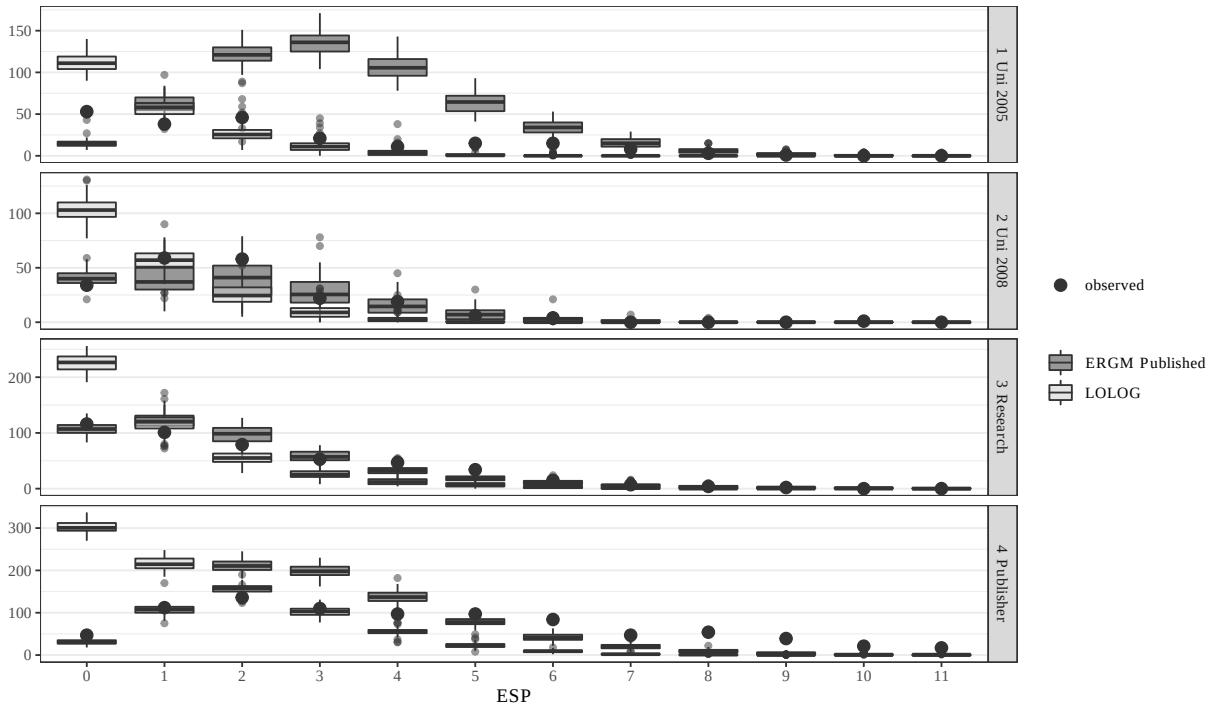


Figure 1.4: ESP goodness of fit comparison plot for Office layout networks. The comparison is between ERGM and LOLOG fits both using the published ERGM terms

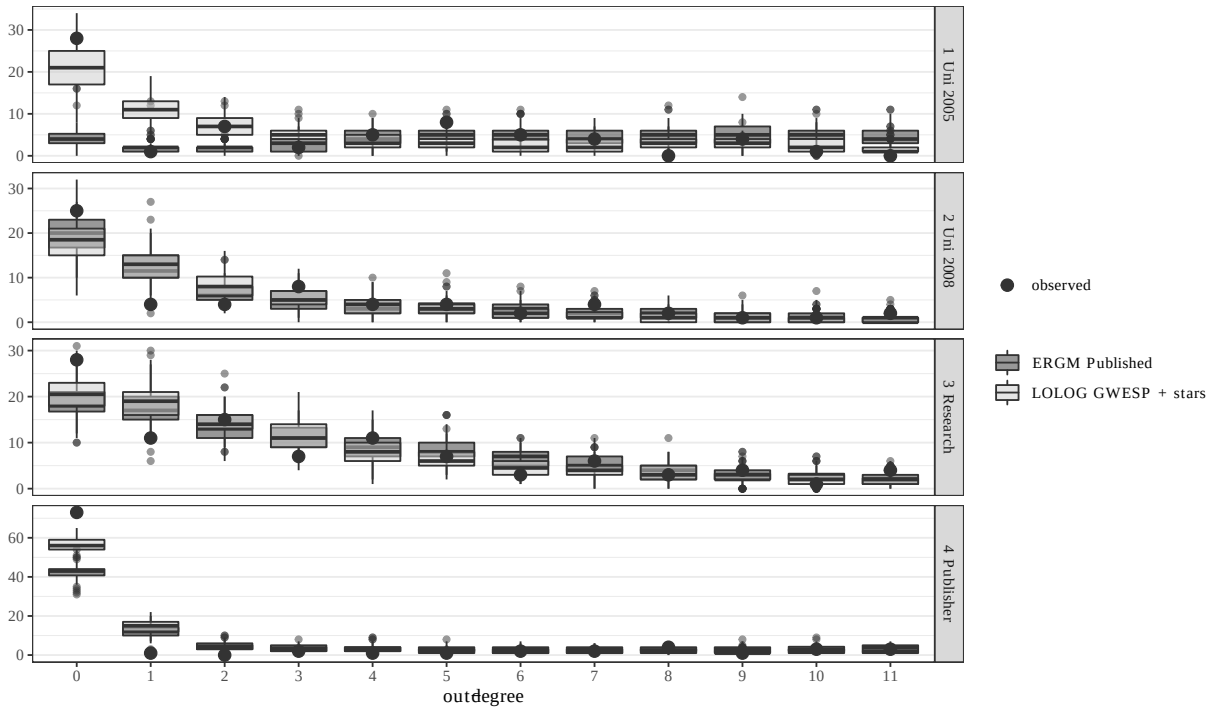


Figure 1.5: Out-degree goodness of fit comparison plot for Office layout networks. The comparison is between the published ERGM and LOLOG fits with GWESP and stars terms included.

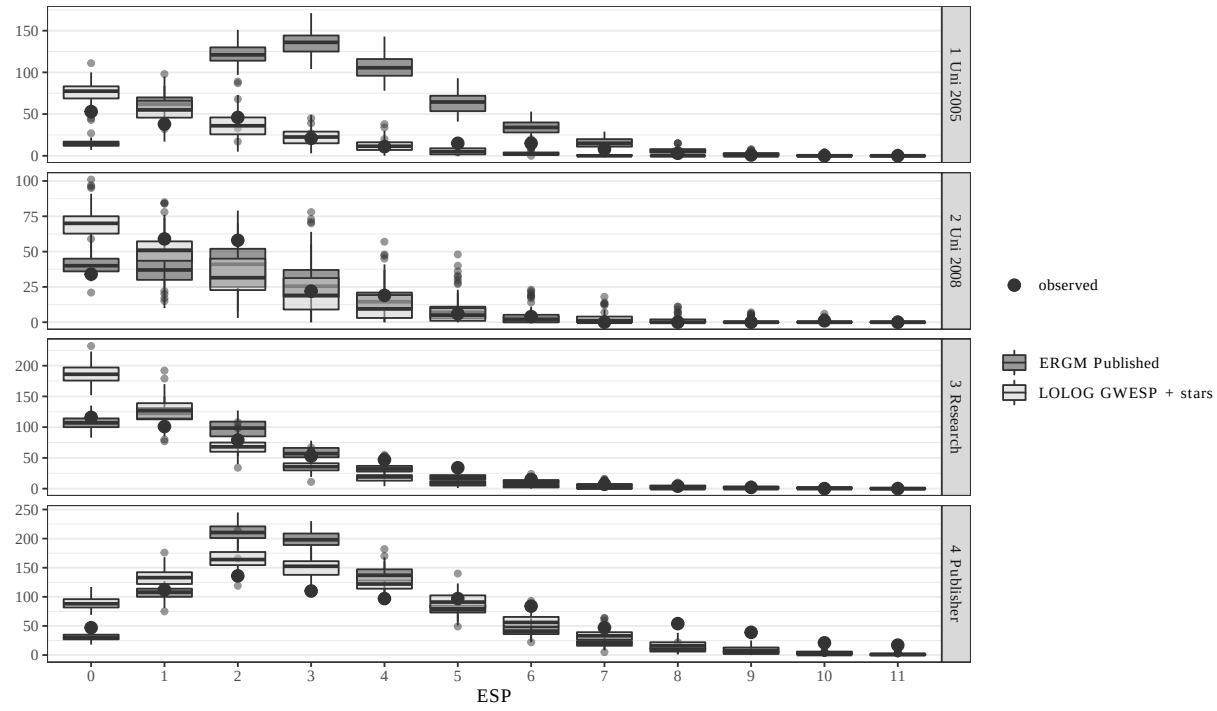


Figure 1.6: ESP goodness of fit comparison plot for Office layout networks. The comparison is between the published ERGM and LOLOG fits with GWESP and stars terms included.

1.8.2 Links to publically available data

Table 1.8 provides hyperlinks to the publicly available datasets used in our ensemble.

Table 1.8: Links to publicly available datasets

Network	Links
Add Health	addhealth.cpc.unc.edu/
Elementary School	moreno.ss.uci.edu/data.html#children
Florentine Families	sites.google.com/site/ucinetsoftware/datasets/padgettflorentinefamilies
Kapferer's Tailors	sites.google.com/site/ucinetsoftware/datasets/kapferertailorshop
Natural Disasters	vlado.fmf.uni-lj.si/pub/networks/data/GBM/kansas.htm
German Schoolboys	github.com/gephi/gephi/wiki/Datasets

Company Boards	corp.boardex.com

CHAPTER 2

Bayesian Inference for Latent Order Logistic Network Models

The Latent Order Logistic (LOLOG) model is a flexible model class for modelling social network data. Likelihood based parameter estimation is challenging due to the complex likelihood function. Bayesian methods have been shown to be useful for the related Exponential Random Graph Model (ERGM). We develop a Bayesian approach for inference using the LOLOG model. We give detailed example analysis on the well known Lazega's New England lawyers network. We also demonstrate the method's feasibility for networks of $1000s$ of nodes using the Add-Health survey data of adolescent health behaviours in schools.

2.1 Introduction

Fully Bayesian approaches to modelling social network data are not yet widely used in real data social network analyses. Perhaps not due to a lack of appeal of such methods, but to practical implementation difficulties when the likelihood for most realistic social network models are mathematically intractable. Bayesian methods provide a axiomatically consistent framework for inference on the parameters of interest, that do not violate the likelihood principle O'Hagan and Kendall (1993). In contrast, the classical argument for inference on the quantities of interest, often derived from model parameters, is based on the assertion that the true data generating process and its parameters are fixed. Thus following the classical approach it is impossible to attach uncertainty to one's inference in the way that practitioners intuitively desire, that is explicit uncertainty in the parameters of the data generating

process, not randomness in the observations derived from a fixed generating process. Consequently classical inference requires somewhat shoe-horned arguments for confidence interval interpretation, that are often ignored by researchers wanting to arrive practical conclusions for their particular problem. In particular in the field of social network analysis, where the primary questions of interest are posed by social scientists, the shortcomings of classical inference are likely under appreciated.

We also suspect that in the majority of social network problems researchers have strong priors on the types of relationships that are likely to form. For example homophilly on nodal attributes or social closure in friendship networks, are almost always observed. Therefore any analysis of social network data should build in the researchers' prior knowledge derived from many years developing expertise in their field. The difficulty in specifying prior knowledge should not obscure its importance, or excuse a researcher from using this knowledge. Indeed ignoring prior information in the Bayesian paradigm is as egregious as, say ignoring difficult to deal with features of the data.

The Exponential-family Random Graph Models (ERGM) framework is widely used to represent the stochastic process underlying social networks [Frank and Strauss (1986); Hunter and Handcock (2006)]. Bayesian approaches to ERGMs have been developed recently, [Caimo and Friel (2011), Caimo and Friel (2014), Wang (2011)], which use a Markov Chain Monte Carlo (MCMC) method, for doubly intractable distributions [Murray et al. (2006), Møller et al. (2006)], so called due to the dual difficulties of sampling as well as the evaluation of the likelihood.

In this paper, we present a Bayesian inferential framework for the LOLOG model. We discuss choices of priors for social network models and present a class of pseudo conjugate priors for the LOLOG model Fellows (2018b), with a practical empirical Bayes flavoured suggested prior. We note that conditional on a specified edge ordering, the LOLOG model is a logistic regression on change statistics of the partial network, to sample for the posterior, we first sample for the posterior edge ordering distribution, followed by sampling from the posterior parameter distribution conditional on that edge ordering. This allows us to sample from

the full posterior parameter distribution of LOLOG model parameters conditional on the observed data.

The computational burden is significantly less than a comparable Bayesian fit to an ERGM model as conditional on the edge ordering the LOLOG model is a logistic regression. In addition the method is "embarrassingly parallel" and our examples show that the method can be applied to networks with $1000s$ nodes with reasonably many model parameters.

The structure of this paper is as follows. In Section 2.2 we briefly introduce the LOLOG model, as per Fellows (2018b). In Section 2.3 we introduce the MCMC algorithm for the two stages of sampling from the LOLOG posterior. In Section 2.4 we discuss the possible prior distributions for the parameters in the LOLOG model, and introduce the pseudo conjugate prior for the LOLOG model as well as its practical empirical Bayes implementation. Section 2.5 shows an example with the well known New England lawyers data set (S. Kraatz et al., 2003), with full discussion of tuning parameters and diagnostics for the fit, we also present the fit to a larger network of friendships in adolescents using the Add-Health data set (Harris et al., 2007b). We conclude with discussion in Section 2.6.

2.2 LOLOG Model Formulation

We introduce the LOLOG model and define the standard notation that will be used henceforth. We define y to be a graph, in particular a graph realised from the random variable Y. Since we regard any nodal covariates as fixed, for a network of size n, Y takes values in the space $\mathscr{Y} = \{a \in \mathbb{R}^{n\times n} \quad | \quad \forall i,j \quad a_{i,i} = 0 \quad a_{i,j} \in \{0,1\}\}$. For undirected networks the additional restriction that $a_{i,j} = a_{j,i} \quad \forall i,j$ is added. Note that the sample space, even in the restricted undirected case is finite and size $2^{\frac{n(n-1)}{2}}$, this become astronomically large even for small networks.

To develop models which account for partially formed networks Fellows (2018b) introduced new notation. The random variable Y_t takes values in $\mathscr{Y}$, realisations y_t denote the partial graph formed at time t. That is, the first t edges have been considered in the LOLOG

generating process. Let $\mathscr{S}_{n_d}$ be the set of possible edge orderings on a network with n_d dyads. For a time t, $s_{\leq t}$ is used to denote the first t elements of s.

The first step in specifying the LOLOG model is to state the probability of observing a graph given a specified order of edge formation s i.e.:

$$p(y|s,\theta) = \prod_{t=1}^{n_d} \frac{1}{Z_t} \exp\left(\theta \cdot C_{s,t}\right) \tag{2.1}$$

Where, for g a function giving specified graph statistics we define the change statistics

$$C_{s,t} = g(y_t, s_{\leq t}) - g(y_{t-1}, s_{\leq t-1}) \tag{2.2}$$

The change statistics are the difference in the graph statistics, between the y_t network and the y_{t-1} network. The Z_t are the usual normalising constants so the sum of the probabilities of the edge forming and not forming are 1. That is letting y_t^+ be the graph y_{t-1} with the edge s_t added and y_t^- the graph without the edge added i.e. unchanged. Then we have Z_t as follows:

$$Z_t = \exp\left(C_{s,t}^+\right) + \exp\left(C_{s,t}^-\right) = \exp\left(g(y_t^+, s_{\leq t}) - g(y_{t-1}, s_{\leq t-1})\right) + 1 \tag{2.3}$$

That is each edge is considered in turn as a formal logistic regression with the so-called change statistics used as predictors. To obtain the full unconditional distribution we sum over the space of possible edge permutations as follows.

$$\begin{aligned} p(y|\theta) &= \sum_s p(y|s,\theta)p(s) \\ &= \sum_s \left(p(s) \prod_{t=1}^{n_d} \frac{1}{Z_t} \exp\left(\theta \cdot C_{s,t}\right) \right) \end{aligned} \tag{2.4}$$

A key advantage of the LOLOG model is the ease of simulation from the model, to simulate

a network we simply draw s from $p(s)$ and then do a sequential logistic regression simulation on the change statistics.

The intractability of ERGMs, is due to the normalizing constant being a sum over the space of all possible graphs, the intractability in LOLOG is summing over all possible edge permutations. As the likelihood or a likelihood ratio (as is used for ERGMs), cannot be evaluated, the maximum likelihood estimate (MLE) for LOLOG is intractable, See Fellows (2018b) for details of a method of moments (MOM) approach to estimate model parameters in a frequentist sense. We note that the general idea is leverage the ease of simulation to estimate quantities required in the fitting procedure, something which we will use to efficiently explore the posterior parameter distribution also.

We also note that the MOM estimate approximates the MLE only asymptotically as the size of the network become large, and its finite sample distribution is unknown. As the size of a network is considered a fixed property of the network, the notion of the MOM estimate's asymptotic approximation to the MLE is not appealing. In addition since every LOLOG model (and ERGM) places non zero mass on the empty and full graph, for which both the MLE and MOM do not exist, the expectations and standard errors of the MLE and MOM parameter distributions do not exist either. The Bayesian approach we present has no such theoretical problems, as inference is based on a sample from the posterior parameter distribution.

Along with developing LOLOG models in Fellows (2018b), the `lolog` R package Fellows (2018a) has been developed. It provides a sophisticated, fast and user friendly method to fit LOLOG models to data utilising the `Rcpp` package to leverage the speed of pure C++ code.

2.3 Bayesian Inference for LOLOG Models

Typically the researcher fitting LOLOG to a network likely has an understanding of the possible values of the parameters, if a LOLOG model generated the data, that is they have a prior as in the standard Bayesian approach. As is usually the case, the analytic posterior

distribution is intractable, and we resort to sampling from the posterior using an MCMC approach.

As the application of Bayes' theorem on the full sum in Equation 2.4 is intractable, we proceed conditional on a selected edge ordering. We have:

$$p(\theta|y) = \sum_s p(\theta, s|y) \tag{2.5}$$

$$= \sum_s p(\theta|s, y)p(s|y) \tag{2.6}$$

To draw a sample of θ unconditional on the edge ordering i.e. from $p(\theta|y)$, we first draw θ_i from the prior $p(\theta)$ then we draw samples $\{s_1, \dots s_n\}$ from the edge ordering distribution $p(s_i|y, \theta_i)$ and sample from each of $p(\theta|y, s_i)$.

Fellows (2018b) noted that the distribution of $p(S|y, \theta)$ is intractable, sampling from $p(S|y, \theta)$ is non trivial, and of interest. The distribution represents the types of orderings that are likely to have formed the network, which may be of interest to the researcher.

2.3.1 Sampling from Posterior Edge Permutation Distribution

We can sample from $p(S|y, \theta)$ using the Metropolis Hastings (MH) algorithm (Hastings, 1970), the acceptance ratio between two permutations s and s', using a symmetric proposal step is as follows:

$$A(s', s) = min\left(1, \frac{p(s'|y, \theta)h(s'|s)}{p(s|y, \theta)h(s|s')}\right) \tag{2.7}$$

$$= min\left(1, \frac{p(y|s', \theta)p(s')h(s'|s)}{p(y|s, \theta)p(s)h(s|s')}\right) \tag{2.8}$$

assuming all permutations are equally likely a priori, and a symmetric step

$$= min\left(1, \frac{p(y|s',\theta)}{p(y|s,\theta)}\right) \tag{2.9}$$

$$= min\left(1, \frac{\prod_{t=1}^{n_d} \frac{1}{Z^{s'_t}} exp\left(\theta \cdot c(y_{s'_t}|y^{t-1}, s'_{\leq t})\right)}{\prod_{t=1}^{n_d} \frac{1}{Z^{s_t}} exp\left(\theta \cdot c(y_{s_t}|y^{t-1}, s_{\leq t})\right)}\right) \tag{2.10}$$

$$= min\left(1, \prod_{t=1}^{n_d} \frac{Z^{s_t}}{Z^{s'_t}} exp\left(c(y_{s_t}|y^{t-1}, s_{\leq t}) - c(y_{s'_t}|y^{t-1}, s'_{\leq t})\right)\right) \tag{2.11}$$

This acceptance probability is easy and fast to calculate. The main computational burden of the acceptance probability calculation and hence sampling from the distribution $P(S|y,\theta)$ is the calculation of the change statistics for each of the permutations s' and s.

As with most applications of the MH algorithm, the choice of proposal step is important. With a simple random sample of the space of edge permutations, we experienced very poor exploration of the space. The sampler would become stuck on a edge permutation of high probability and take a very long time to move away, i.e. the chain took a very long time to converge to its stationary distribution and samples where highly correlated with each other. As a result a local sampler was considered, the approach was to select a small subset of edges in s and permute their position at random to get to s'.

The remaining choice is that of what proportion of the edges to switch, there are two competing problems. Switch too many, and the sampler will not explore the space, since random samples are likely to be rejected once the sampler reaches a permutation with high probability. Switch too few and the sampler will not explore the space of permutations very well, and the permutations will be heavily auto correlated. Thus we expect there to be an optimal proportion for a given network. We examine the required burnin length; the number of chains required before we claim that the chain is at it's stationary distribution, using the method due to Gelman and Rubin (1992). This considers the variance within and between several Monte Carlo Markov chains, little difference in between chain and within chain variation suggests that the chain is at its stationary distribution. We show a full example of the tuning procedure in Section 2.5.1.

As a distribution of permutations on the edges it is a discrete distribution on a sample space of order $\mathcal{O}((n^2)!)$. Edge permutation distributions are hard to visualise and understand, in our examples we present different edge permutations distributions in terms of their change statistic distribution since this is the only thing that impacts the LOLOG model. We also show the distributions of formation locations for structures of interest. That is, if permutation s did generate y the time t at which each observed structure would have formed. The distribution of formation locations gives the researcher insight into whether complex structures in the network where more or less likely to have formed earlier or later in the network formation process.

2.3.2 Sampling from Posterior Parameter Distribution Conditional on Edge Permutation

Applying Bayes' theorem and assuming the prior distribution of θ is independent of s we have:

$$p(\theta|y,s) \propto p(y|\theta,s)p(\theta|s) \tag{2.12}$$

$$= p(y|\theta,s)p(\theta) \tag{2.13}$$

So the Metropolis Hastings ratio for a proposal function $h(\theta_2|\theta_1)$ stepping from θ_1 to θ_2 is as follows:

$$A(\theta_2,\theta_1) = \min\left(1, \frac{p(y|\theta_2,s)p(\theta_2)h(\theta_2|\theta_1)}{p(y|\theta_1,s)p(\theta_1)h(\theta_1|\theta_2)}\right) \tag{2.14}$$

Assuming the symmetry of the proposal distribution

$$A(\theta_2,\theta_1) = \min\left(1, \frac{p(y|\theta_2,s)p(\theta_2)}{p(y|\theta_1,s)p(\theta_1)}\right) \tag{2.15}$$

$$= \min\left(1, \prod_{t=1}^{n_d} \frac{Z(t,y,\theta_1,s)}{Z(t,y,\theta_2,s)} exp\left(\theta_2 \cdot C_{y,s,t} - \theta_1 \cdot C_{y,s,t}\right)\right) \tag{2.16}$$

This can be computed and utilised to sample from the posterior, it could even be estimated with a sampled version for speeding up computations in large networks. Next we note that the set of change statistics $\{C_{y,s,t}\}_{t=1}^{n_d}$ does not depend on θ. The calculation of the change statistics for large networks can be a computational burden, however for each permutation s it needs to be calculated once only, and can be reused as the sampler steps around the θ parameter space.

Adding parameters to the model only increases computation time in the time taken to explore the parameter space. We have found in our examples it does not greatly increase the computation time required to calculate the acceptance probabilities, thus we expect this method to be suitable for LOLOG models with a reasonable number of parameters of interest.

Typically LOLOG parameters are on different scales and cannot be directly compared, intuitively since there are many more possible complex structures than simple ones. As a result the parameter space can be hard to sample from with a random walk. We consider 3 MCMC methods for sampling from the posteriors

1. Random Metropolis Hastings walk with a Gaussian step with a diagonal covariance matrix scaled ad hoc with the total possible number of each parameter's structure

2. Hamiltonian Monte Carlo (HMC) (Duane et al., 1987) with leapfrog integrator, with momentum as estimated Fisher information estimated with simulated networks and their change statistics and tuned step number and step size.

3. Riemannian Manifold Hamiltonian Monte Carlo (RMHMC) (Girolami and Calderhead, 2011). Hessian as estimated Fisher information which is estimated with simulated networks and their change statistics. One Step is used as we wish to minimise the number of required steps since every subsequent step requires the estimation of the Hessian with the burden of simulating networks.

For larger networks the scaling issues in the parameter space are more acute and we expect the random walk to be ineffective and HMC and RMHMC methods to be necessary to

achieve feasible sampling. In particular some second order knowledge as used in the RMHMC method is required to step around parameter spaces in any kind of efficient manner. For large networks we suggest an initial estimation of the Fisher information matrix to be used in the HMC algorithm as the momentum matrix, which can be estimated through simulation as in Equation 2.23. While an initial calculation to be used for all HMC steps will not capture the curvature of the space as the parameter values change, it will at least capture some notion of the different scales in different dimensions for which we have no a priori knowledge. This is particularly a problem fitting models with more than a few terms to larger networks.

To implement HMC we require the calculation of the gradient of the negative log likelihood of the LOLOG posterior, this is referred to as the gradient of the energy function in the jargon of HMC. Letting the prior on θ be $\pi(\theta)$ we can express the negative energy function as:

$$-U(\theta|y,s) = -\log\left(p(\theta|y,s)\right) = -\log(p(y|\theta,s)\pi(\theta)) + C \tag{2.17}$$

$$= -\log(\pi(\theta)) - \sum_{t=1}^{n_d} \log\left(\frac{1}{Z_t} e^{\theta \cdot C_{s,t}}\right) \tag{2.18}$$

we can simply take the gradient

$$\nabla \log(p(\theta|y,s)) = \frac{\nabla\pi(\theta)}{\pi(\theta)} + \sum_{t=1}^{n_d} C_{s,t} - \sum_{t=1}^{n_d} \frac{\nabla Z_t}{Z_t} \tag{2.19}$$

$$= \frac{\nabla\pi(\theta)}{\pi(\theta)} + g(y) - \sum_{t=1}^{n_d} \frac{C^+_{s,t} e^{\theta \cdot C^+_{s,t}}}{1 + e^{\theta \cdot C^+_{s,t}}} \tag{2.20}$$

This is relatively simple to calculate, although it does require a n_d dimensional sum for the last term. We note that n_d grows with $(n^2)!$ where n is the number of nodes. For large networks, sampling this sum to calculate the gradient may be desirable, that is employing the Stochastic Gradient RMHMC Ma et al. (2015).

For RMHMC, we attempt to use second order information of the parameter space in order

to improve the sampling properties. In the Bayesian setting the natural matrix is the Fisher information matrix, we estimate this as the expected Hessian of the likelihood.

$$I(\theta|s) = \mathbb{E}_{X|\theta,s}\left[H(\theta|x,s)\right] \tag{2.21}$$

$$= \mathbb{E}_{X|\theta,s}\left[(\nabla U(\theta|x,s))^\top \nabla U(\theta|x,s)\right] \tag{2.22}$$

Method 3 simulates k new networks from the distribution $X|\theta, s$ to obtain

$$I(\theta|s) \approx \frac{1}{k}\sum_{i=1}^{k}(\nabla U(\theta|x_i,s))^\top \nabla U(\theta|x_i,s) \tag{2.23}$$

This requires the simulation of k networks from the LOLOG model, which requires the calculation of the change statistics for those networks. These change statistics are also used in computing the required gradients.

2.3.3 Assessing Convergence of MCMC for posterior parameter distribution

As with any MCMC approach checking the convergence of the MCMC chain to the stationary distribution is crucial to the validity of the analysis. This is an inherently difficult question and to some degree cannot be completely answered. Attempting to ensure the chains are close to their stationary distributions we follow the approach of Gelman and Rubin (1992). This approach requires a real number to represent each state of the Markov chain. As we have two MCMC procedures, one for the edge distribution and one for the parameter distribution we require a numerical value for each state of both of these Markov chains.

For edge permutation samples, we propose using a weighted sum, that is if a permutation of e edges s is represented as a vector i.e $s \in \mathbb{R}^{n_d}$. With $s_i = j$ implying that the i^{th} edges to

be considered is that indexed with a j we have:

$$f(s) = \sum_{k=1}^{e} k \times s_i \tag{2.24}$$

Note that the space of permutations is $\mathscr{O}(e!)$ but the maximum value for the function $f(s)$ is the sum of squares which is $\frac{e(e+1)(2e+1)}{6}$, which is much smaller than $e!$ for even small graphs. For illustration a 10 node undirected network has 45 possible edges, $45! \approx 10^{56}$ whereas $45^3 \approx 10^5$. Thus there will be multitudes of permutations that take the same value of $f(s)$. In our examples we noted that the simple weighted sums did occasionally produce duplicate values for different permutations in examples we sampled but not at a high rate.

For the posterior parameter distribution, we considered each parameter value individually in the convergence checking proceedure.

For the edge permutations we initialised 8 chains at random permutations, for the posterior parameters we initialised at the variational LOLOG fit (Fellows, 2018b). We have experimented with adding additional dispersion, these required longer burn-in times but not excessively. We computed the ratio of within to between variance as per Gelman and Rubin (1992), values close to 1 indicate convergence of the Markov chain. We also visually inspect the trace plots for each of the MCMC chains.

2.3.4 Simulating from Full Posterior Distribution

Simulating from the posterior requires the drawing of a parameter θ and an edge permutation s from the joint posterior given the observed graph y. Once these are drawn the usual sequential logistic regression simulation can be carried out.

Once we have samples from the posterior parameter distribution, we have several choices when it comes to simulating new networks. Noting that since $p(\theta, s|y) = p(\theta|y, s)p(s|y)$, so we could simulate from $p(s|y)$ and then from $p(\theta|y, s)$. Or we have $p(\theta, s|y) = p(s|y, \theta)p(\theta|y)$ so could simulate from $p(\theta|y)$ and then from $p(s|y, \theta)$.

1. Sample from $p(\theta|y)$, then run additional MCMC proceedure to sample from $p(s|y,\theta)$

2. Use the samples from the fitting proceedure. Sample from the $\{s_1, \dots s_n\}$ and then from the corresponding sample from $p(\theta|s_i, y)$.

3. Sample independently. Sample from $p(\theta|y)$ given from fitting proceedure. Sample from $p(s|y)$ from fitting proceedure.

Method (a) is valid assuming our sample from $p(\theta|y)$ has low dependence on the permutations used to generate it, the downside of this method is it requires a further MCMC proceedure to sample from $p(s|y,\theta)$. Method (b) samples preserving the dependence of θ and s, it assumes that we sampled enough permutations to have a plausible sample from the distribution of $s|y$. Method (c) disregards that the samples from $p(\theta|y)$ and $p(s|y)$ were formed by combining samples, throwing away the knowledge of coupling of s and θ essentially assuming they are independent.

We use Method (b) for simulating from the fitted model, to minimise the computational burden while accounting for the joint distribution of θ and s conditional on y.

2.3.5 Assessing Model Fit

To assess goodness of fit of our model, with only one observed network, we use Bayesian posterior prediction goodness of fit graphics in the style of Hunter et al. (2008). The difference in the Bayesian case being that the simulations now are derived from the posterior, that is new networks are simulated by first drawing θ and s from the posterior distribution sample derived via MCMC and then simulating a network with that θ and s. The distribution of the simulated networks' statistics is then compared to the observed network statistics. Here we consider the degree and edgewise shared partner distribution, though others are possible.

2.4 Pseudo Conjugate Priors

Prior probability distributions are subjective probabilities distributions that require specifying the probability distribution of the parameters conditional on the total of all the knowledge the researcher has, excepting the data itself (O'Hagan and Kendall, 1993). This is, at best, challenging, but likely impossible. In particular, often uninformative priors are used (Jeffreys, 1946), as a means to specify little prior knowledge, though also likely due to the difficulty in formalising one's prior knowledge. In the context of ERGMs Caimo and Friel (2011) utilised a flat Gaussian prior in their analysis and Wang (2011) discussed the advantages of using the conjugate prior distribution. A flat prior is dissatisfying as it places significant mass on areas of the natural parameter space that are near degenerate (Handcock, 2003), the researcher likely has strong prior knowledge that the model producing the observed network was not near degenerate. Indeed a non degeneracy prior may be a useful, i.e. specifying the prior to place mass on parameter values that do not produce degenerate models, though this is at present an open area of research. The conjugate prior is a convenient method of specifying useful prior information, but does still leave the researcher highly constrained in their input of the sum total of their knowledge.

Specifying the conjugate prior in the case of an exponential family is well understood (Diaconis and Ylvisaker, 1979), however due to the summing over all possible edge permutations in the LOLOG model, developing a conjugate prior for the full unconditional likelihood is intractable. We suggest considering the distribution of graphs conditional on a particular edge permutation. Since the LOLOG model reduces to a logistic regression on the change statistics (which depend on the edge permutation), we can use exponential family priors to specify a conjugate prior distribution, conditional on the edge permutation. We can then sample permutations and take an average of the conjugate priors to form a pseudo-conjugate prior.

In particular given the density of the graph, one can place the prior that every edge is equally likely to form. Clearly this prior depends on the observed data, so cannot be interpreted as

the researcher's knowledge prior to seeing any of the data. However it only depends on the density statistic of the network, which is at best weakly informative of the structure of the network, which is usually of interest. Indeed this is a form of the general empirical Bayes methodology,which can be framed in terms of hierarchical model prior[Casella (1985), Efron (2010)].

Chen and Ibrahim (2003) specify conjugate priors for generalised linear models (GLMs) in terms of a prior guess of each data point y_i for each of the observed covariate vector x_i. Proceeding as in Chen and Ibrahim (2003) for an exponential family with dispersion parameter τ:

$$P(y_i|\theta_i, \tau) = \exp\left(\tau\left(y_i\eta_i - b(\eta_i)\right) + c(y_i, \tau)\right) \tag{2.25}$$

Letting $\eta_i = \eta(X_i\beta)$ for some link function η

$$= \exp\left(\tau\left(y_i\eta(X_i\beta) - b(\eta(X_i\beta))\right) + c(y_i, \tau)\right) \tag{2.26}$$

As per Diaconis and Ylvisaker(1979) the conjugate prior is

$$\Rightarrow P(\eta|\tau, y_0, a_0) \propto \exp(a_0\tau\left(y_0^\top \cdot \eta - b(\eta)\right)) \tag{2.27}$$

Noting that for exponential families we have $\mathbb{E}(y|\eta) = \nabla b(\eta)$. We have

$$\mathbb{E}\left[y\right] = \mathbb{E}_\eta\left[\mathbb{E}\left[y|\theta\right]\right] \tag{2.28}$$

$$= \mathbb{E}_\theta\left[\nabla b(\eta)\right] \tag{2.29}$$

Using (Diaconis and Ylvisaker(1979)) Theorem 2

$$\Rightarrow \mathbb{E}\left[y\right] = y_0 \tag{2.30}$$

Thus the specification of the prior parameter y_0 can be thought of as specifying a prior guess for $\mathbb{E}[y]$. The a_0 parameter corresponds to the effective sample size of hypothetical data that was used to formulate this prior guess, in particular $a_0 = \frac{n_0}{n}$ where n_0 is the prior sample size and n is the number of observations in the data. So an a_0 parameter equal to 1 would correspond to the prior being derived from data the same size as the observed data. In the context of the LOLOG model with change statistics as the predictors, n is the number of edges in the network.

Rewriting η as a function of the regression parameters β results in a proper conjugate prior for β Chen and Ibrahim (2003)

$$P(\beta|\tau, y_0, a_0) \propto \exp\left(a_0\tau\left(y_0^\top \cdot \eta(X\beta) - b\left(\eta(X\beta)\right)\right)\right) \tag{2.31}$$

For the LOLOG model, since only the distribution conditional on the edge permutation is an exponential family distribution, we cannot form such a conjugate prior. However we form a pseudo conjugate prior by simulating k matrices of change statistics of the model to be fit $\{X_i\}_{i=1}^k$ with permutations $\{s_i\}_{i=1}^k$. For each row of a matrix X_i we need corresponding guess vector y_i. We then combine these matrices and guesses to form the X^* and y_0^* for the GLM conjugate prior. For rows of X that are repeated, we retain a single row, and take the mean of the guesses for that row as the value for y_0. This forms the LOLOG pseudo conjugate prior setting b and η to be the standard functions for logistic regression:

$$p(\theta) = \exp\left(\alpha_0\left(\left(y_0^*\right)^\top \cdot X^*\beta - \sum_{i=1}^{n}\log\left(1+\exp\left(X^*\beta\right)\right)\right)\right) \tag{2.32}$$

Specifying the pseudo conjugate prior essentially means choosing a guess y_0 for different change statistics. As a default prior we recommend specifying the y_0^* as equal to the density of the observed network, and tuning the a_0 to obtain a prior distribution that aligns with the researcher's uncertainty in the model parameters. Following the empirical Bayes paradigm,

we also consider using a MOM fit on the network with just the edge and triangle parameters to generate predictions y_0^*.

We also emphasise that as in equation 2.13 we are assuming that $p(\theta) = p(\theta|s)$ that is that the prior on the parameter does not depend on the permutation. This necessarily results in a poor prior since we expect the parameters and the permutations to be correlated in a complex way, however specifying that relationship a priori is essentially impossible so is not attempted.

2.5 Examples

We present two examples of Bayesian fitting of the LOLOG model. We fit the well known New England lawyers data set (S. Kraatz et al., 2003) as well as a larger network of friendships in adolescents using the Add-Health data set (Harris et al., 2007b). We show the Lawyer fit with a fuller discussion of tuning and diagnostics and show the Add-health fit to demonstrate the method's ability to fit larger networks.

2.5.1 New England Lawyers

The network is undirected with 36 nodes. Ties in the networks represent collaboration between lawyers in a law firm in New England spread across 3 offices in different geographic locations. The data has nodal covariates for seniority, type of law practice, gender, and office location.

We show the Bayesian LOLOG fit using the empirical Bayes style pseudo conjugate edge-triangle prior, that is we carry out a MOM fit just using the edge and triangle term and use this to generate predictions y_0 for simulated change statistics. Further fits with a variety of priors can be found in Appendix 2.8.

Table 2.1: MOM fit for LOLOG model to New England Lawyers Network

	Parameter Estimate	Std. Error	P-value
Edges	-7.6589	1.0777	0.0000
Triangles	1.0585	0.4400	0.0161
Node Seniority	0.0280	0.0108	0.0099
Node Practice	0.7118	0.1838	0.0001
Dyad Gender Match	1.1633	0.4313	0.0070
Dyad Practice Match	0.9434	0.2761	0.0006
Dyad Office Match	1.6142	0.3028	0.0000

2.5.1.1 MOM fit

We fit the same LOLOG model as in Fellows (2018b) using the Bayesian approach and compare the results. We fit using an edges term as usual, together with the triangle term to account for social transitivity. We also include main covariate effects for seniority and practice, since different practices and seniorities could conceivably be associated with different levels of collaboration, as well as matching effects on gender, practice and office location. For discussion details see Fellows (2018b). In addition we also specify a constraint to the edge ordering, we add the vertices sequentially in order of their seniority, that is, as each step the next most senior node is added, then all possible new edges are considered in a random ordering, repeated until all nodes have been added. Table 2.1 shows the MOM fit.

2.5.1.2 Prior Distribution

Table 2.2 shows a summary of a sample of 1000 draws from the pseudo conjugate prior based on predictions using the MOM fit to the data using only the edges and triangles term.

2.5.1.3 Assessing Convergence of MCMC

We consider a convergence checking proceedure for the RMHMC sampler, for the edge-triangle prior. As there are separate MCMC proceedures for the distributions of $S|y$ and

Table 2.2: Edge Triangle prior distribution summary for New England Lawyers

	Prior Mean	Prior Std. Error	Prior P-value
Edges	-2.0421	1.9310	0.1466
Triangles	1.6272	0.3101	0.0000
Node Seniority	-0.0023	0.0201	0.4483
Node Practice	-0.0418	0.3829	0.4584
Dyad Gender Match	-0.0608	0.7469	0.4644
Dyad Practice Match	-0.0135	0.5601	0.4863
Dyad Office Match	-0.0233	0.5661	0.4868

$\theta|y, s$ we show diagnostics for both.

As we constrained the edge permutations to be permutations achieved by adding nodes in order of seniority, there are two parameters for the local sampler, the number of nodes in which edges were swapped and the proportion of these edges that were swapped. We tuned the edge permutation sampler, to swap a number of edge positions to achieve a median step probability of around 50% this resulted in a sampler swapping 30% of the edges within 10% of the nodes.

For the RMHMC proceedure we used 10 network simulations to estimate the fisher information matrix and, with a step size of 0.3. We note that a higher number of network simulations can be used to facilitate a higher acceptance rate with a step size of close to 1, however simulating networks is the computationally burdensome portion of the proceedure, so we often preferred a smaller step size with a higher variance estimate of the Fisher information matrix in practice.

We examine the trace plots of each MCMC proceedure and following the approach in Gelman and Rubin (1992) we also consider the ratio of the between chain variance to the within chain variance for 19 MCMCs. A ratio of close to 1 suggests that the chain has converged to its stationary distribution.

Figure 2.1 shows the trace plot and MCMC density for the posterior edge distribution sampler. Figure 2.2 shows the progression of the Gelman shrinkage ratio for the same distribution. They suggest the MCMC procedure for the $S|y$ distribution is likely converged after

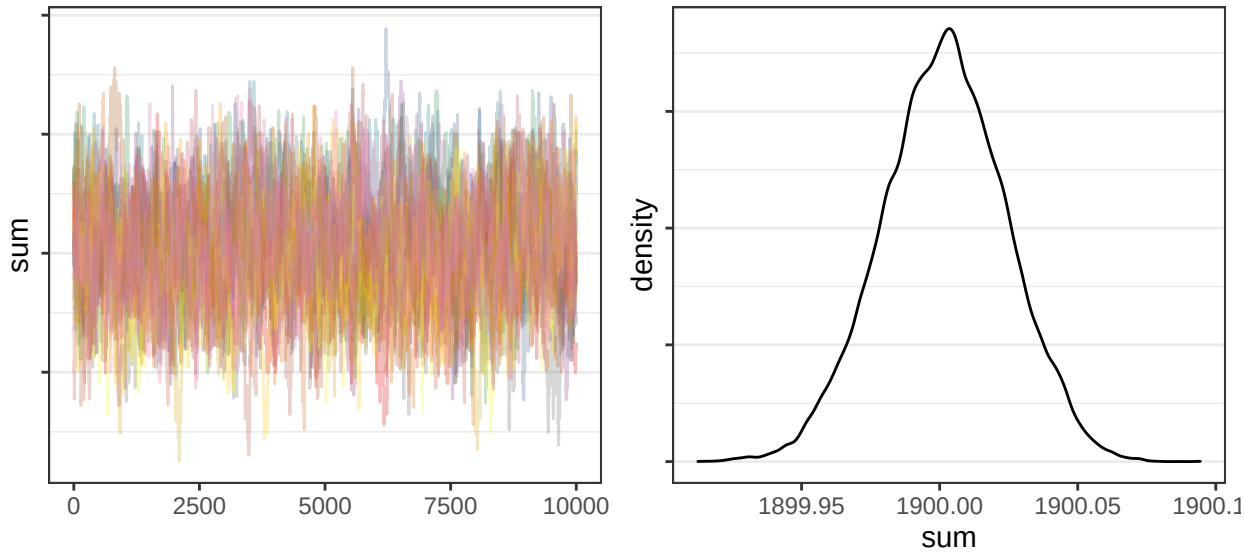


Figure 2.1: Plot of MCMC trace and density for edge permutation sampler for LOLOG model for New England Lawyers network

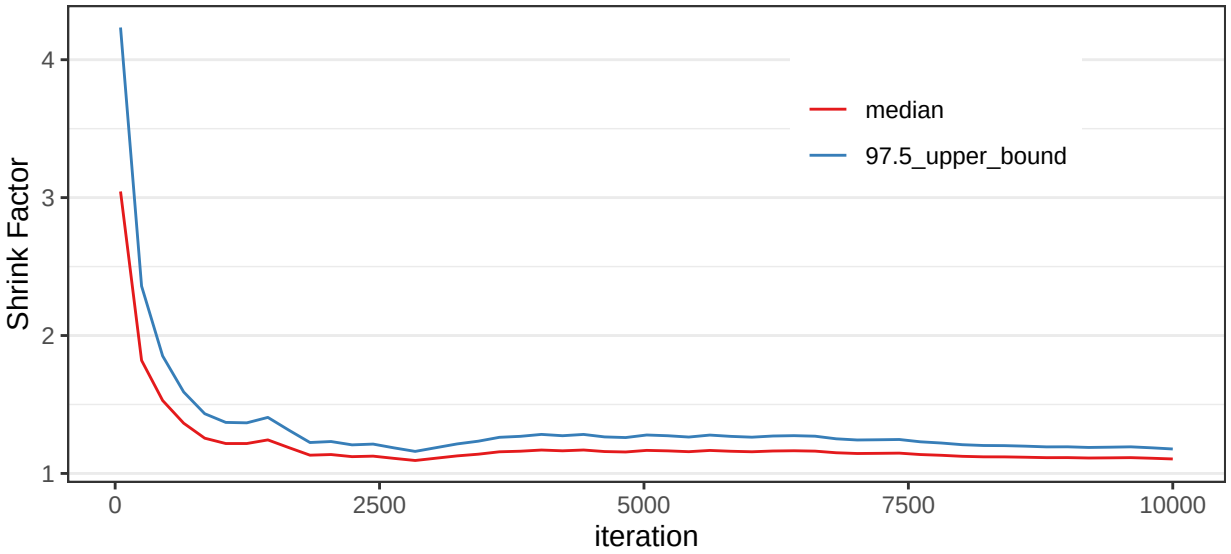


Figure 2.2: Plot of Gelman diagnostics for edge permutation sampler for LOLOG model for New England Lawyers network

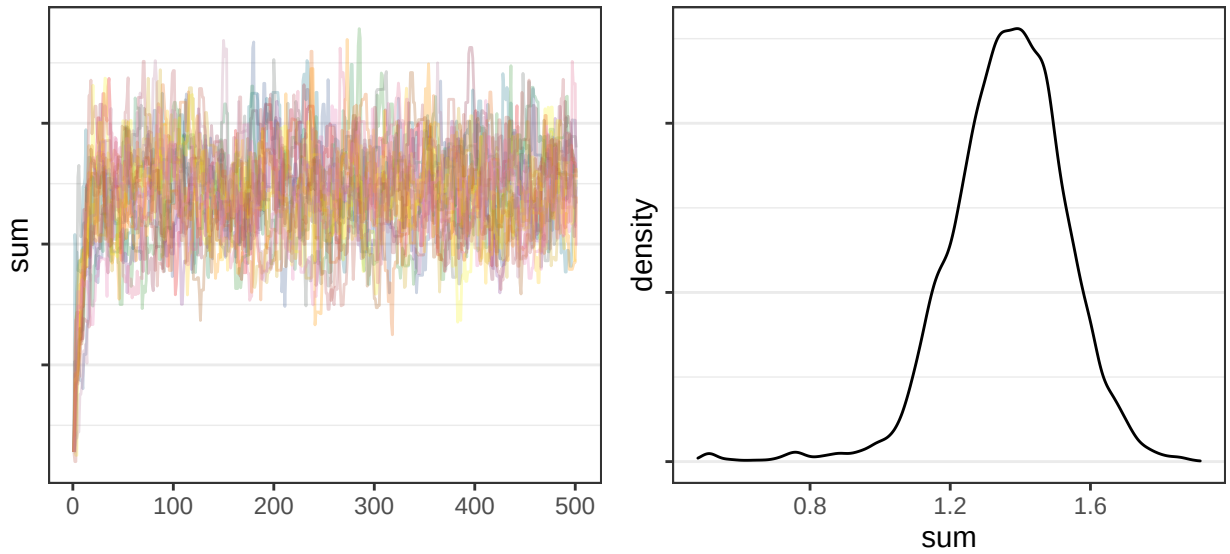


Figure 2.3: Plot of MCMC traces for triangle parameter using RMHMC sampler for LOLOG model for New England Lawyers network

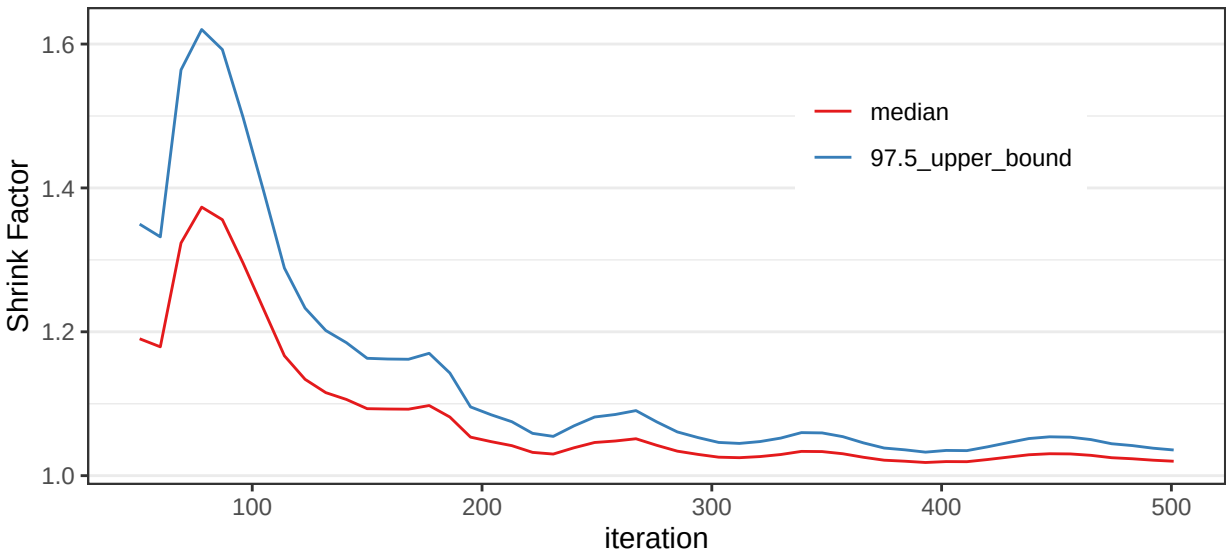


Figure 2.4: Gelman variance ratio plot for triangle parameter for LOLOG model for New England Lawyers network

around 1000 steps.

For the posterior parameter distribution, we only show the trace and Gelman diagnostics plot for the triangle parameter, when examining the convergence we also looked at the other model parameters but omit them for brevity here. Figure 2.3 shows the trace plot and Figure 2.4 shows the Gelman shrinkage ratio. The figures suggest convergence after around 500 iterations.

2.5.1.4 Sampling Permutations

Visualising distributions of edge permutations is challenging, we obtain the change statistic distributions derived from samples from $p(s|y, \theta)$. Figure 2.5 compares the distribution of triangle change statistics for the edge-triangle prior sample and an unconditional sample from $p(s)$. The difference in distribution for other complex structures such as 2- and 3- stars is similar. We show each distribution in comparison to the change statistics of the unconditional permutation distribution. We can see that the triangle change statistic distribution conditional on the observed graph and the using the edge triangle prior is different from the unconditional sample distribution. This suggests the distribution of edge permutations is important.

2.5.1.5 Goodness of Fit

Figures 2.6 show the observed values of the network edgewise shared partners (ESP) and degree statistics, with the boxplot derived from simulations of the fitted posterior model.

Noting that as a majority of the observed values fall within the models' simulated networks' statistics, we consider the model to fit the data well. Further goodness of fit plots for the MOM fit and fits using other priors are included in Appendix 2.8

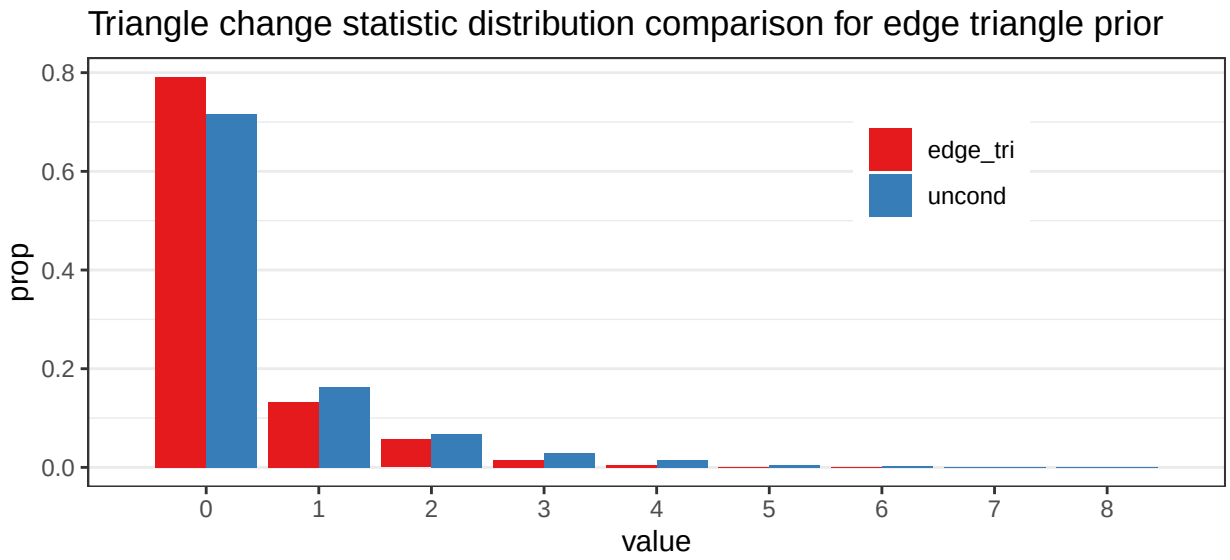


Figure 2.5: Triangle change statistic for posterior edge permutation distribution using edge triangle prior for LOLOG model for New England Lawyers network

2.5.1.6 Posterior Interpretation

Table 2.3 shows the summary of our sample from the posterior distribution.

Our interpretation of the MOM fit summarised in Table 2.1, is as follows:

- The triangle parameter is positive and significant suggesting social transitivity in the collaboration network.
- All the included covariate main and matching parameter estimates are positive and significant suggesting that each covariate, and matching covariate increases the chance of an edge forming.
- The matching effect for office is likely larger than that of gender and practice, suggesting that geographic location is more important in determining likely collaborations.

Using the edge triangle prior, which is not a true prior distribution our interpretation of the Bayesian fit is in agreement with the MOM interpretation, and the same qualitative conclusions can be made, without the need to invoke asymptotic arguments regarding standard

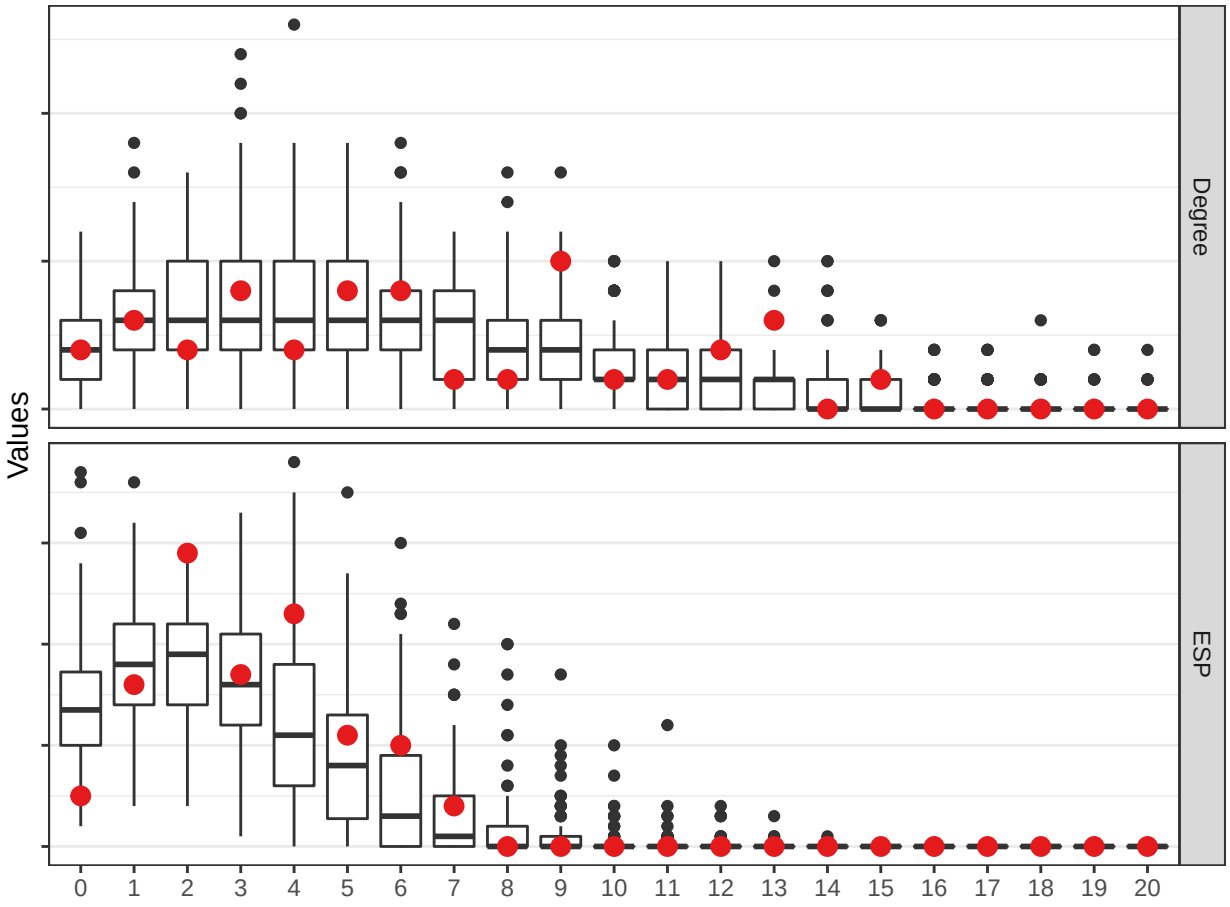


Figure 2.6: Degree and ESP goodness of fit plot for posterior simulations using edge triangle prior for LOLOG model for New England Lawyers network

Table 2.3: Edge Triangle posterior distribution summary for Lazega's Lawyers

	Posterior Mean	Posterior Std. Error	Posterior P-value
Edges	-7.9979	0.6904	0.0000
Triangles	1.4612	0.2519	0.0000
Node Seniority	0.0418	0.0067	0.0000
Node Practice	0.8076	0.1211	0.0000
Dyad Gender Match	0.9116	0.3503	0.0044
Dyad Practice Match	0.6311	0.2502	0.0088
Dyad Office Match	1.1939	0.2679	0.0000

error estimation which do not hold. In addition in the Bayesian framework we can comment directly on the uncertainty of parameter estimates. We stress that the MOM interpretation is theoretically suspect, as the parameter uncertainty estimates rely on asymptotic results, with a network size of only 36 nodes.

2.5.1.7 Computational Performance

We tuned the RMHMC sampling procedure to obtain samples from the posterior parameter distribution. We utilised 20 cores with Intel(R) Xeon(R) Platinum 8160 CPU @ 2.10GHz processor on our departmental server, sampling θ from 19 different s_k samples from $p(s|y)$. For the edge triangle prior run we used an s swap rate of 30% of the edges within 10% of the nodes, with 10000 burn-in s samples. We used 10 RMHMC network simulations per step of size 0.3, with 500 burn-in theta samples. The total run time was approximately 5 minutes.

2.5.2 Adolescent Health

We next consider fitting models a network of high school students, obtained from the well studied National Longitudinal Study of Adolescent to Adult Health (Harris et al., 2007b). Networks from the survey have been fit using ERGMs (Goodreau (2007), Hunter et al. (2008)). There are multiple networks available but the particular network in this case obtained from Harris et al. (2007a) has 1681 nodes with covariates for grade, gender and race

Table 2.4: Add Health LOLOG MOM fit

	Parameter	Std. Error	P-value
Edges	-9.6693	0.0945	0
Triangles	4.2779	0.1715	0
2-Star	0.3828	0.0413	0
3-Star	-0.1865	0.0266	0
Dyad Gender Match	0.6755	0.0634	0
Dyad Grade Match	2.7549	0.0716	0
Dyad Race Match	1.0902	0.0608	0

provided, this matches the data used in the detailed ERGM fit in Goodreau (2007).

2.5.2.1 MOM fit

It is natural to consider students to be added to the network in the order of their grade, for fitting the LOLOG model we constrain edge orderings to orderings, such that all students from each grade are added and then all the new possible edges considered before adding the next grade and considering the new possible edges.

Table 2.4 shows the MOM fit of the model with the edges, triangle, 2-star and 3-star structural terms, as well as matching covariates for race, sex and grade.

Figure 2.7 shows the goodness of fit plot of the degree, ESP and geodesic distance distribution. We note that the MOM estimated LOLOG fits poorly on the ESP distribution and does not recreate the exact shape of the geodesic distance distribution. This is similar to the ERGM fitting issues observed on the network in Goodreau (2007).

2.5.2.2 Bayesian Fit

For smaller networks it is possible to sample from the distribution $S|y$, for example in the Lawyers network there were 36 nodes so 630 edges so the space was discrete of order 630!. Note that using Stirling's approximation $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ we have $630! \geq \sqrt{2\pi \cdot 630}\,(232)^{630} \approx 63\,(10^{2.4})^{630} > 10^{1500}$. The sample space was enormously large, yet

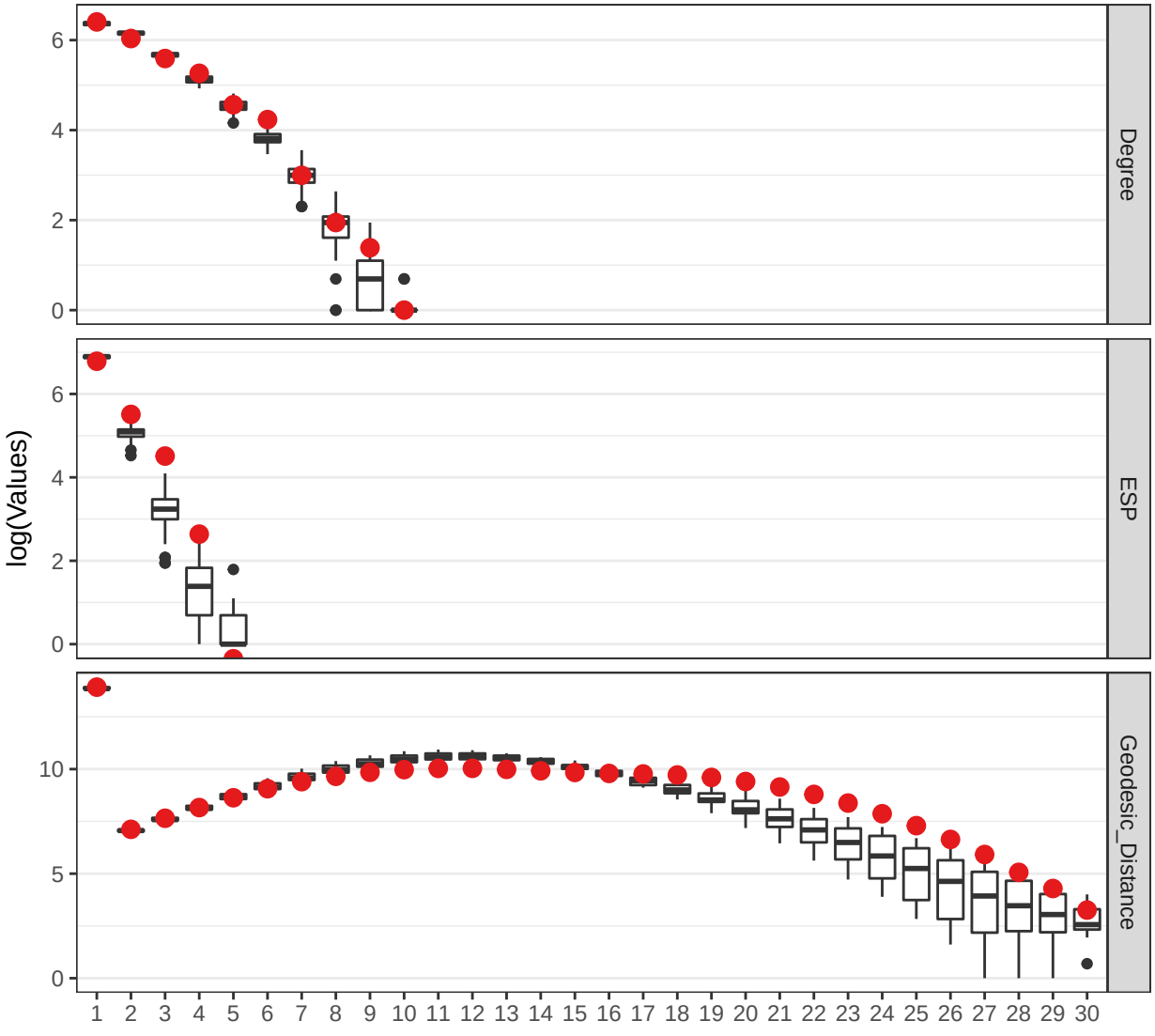


Figure 2.7: Goodness of fit plots for Add Health LOLOG on logarithmic scale

Table 2.5: Edge Triangle posterior distribution summary for Add Health network

	Posterior Mean	Posterior Std. Error	Posterior P-value
Edges	-9.5266	0.0830	0
Triangles	3.3955	0.0602	0
2-Star	0.2832	0.0306	0
3-Star	-0.1238	0.0196	0
Dyad Gender Match	0.6548	0.0594	0
Dyad Grade Match	2.7244	0.0663	0
Dyad Race Match	1.0627	0.0607	0

as each edge permutation is of length 630 we are able to assess the burnin and step until convergence as in section 2.5.1.3. However with a network of 1681 nodes we have 1412040 edges. Stepping around this edge permutation space becomes impossible. In addition assessing the burnin of a chain of objects each of size $\sim 10^6$ also becomes difficult, not to mention storing such a chain of any length.

As a result, in order to utilise the Bayesian methodology for LOLOG,we make the assumption that $p(s|y) = p(s)$. We suspect that as the network becomes large the impact of the network on possible edge permutations is reduced as the dependence between edges is likely reduced, but we have no theoretical results regarding this.

We still sample from $\theta|y, s$ using the RMHMC method outlined above. Table 2.5 shows a summary of a sample from posterior LOLOG model taken using the edge triangle prior. It is qualitatively consistent with the MOM estimated LOLOG model.

Figure 2.8 shows the goodness of fit plot of the degree, and ESP distribution. We note that the Bayesian LOLOG fits similarly to the MOM LOLOG, struggling with the ESP and geodesic distance distribution.

2.5.3 Interpretation

Our interpretation of the MOM fit and the Bayesian fit are consistent and as follows:

- There is a tendency for social closure due to the positive significant triangle term.

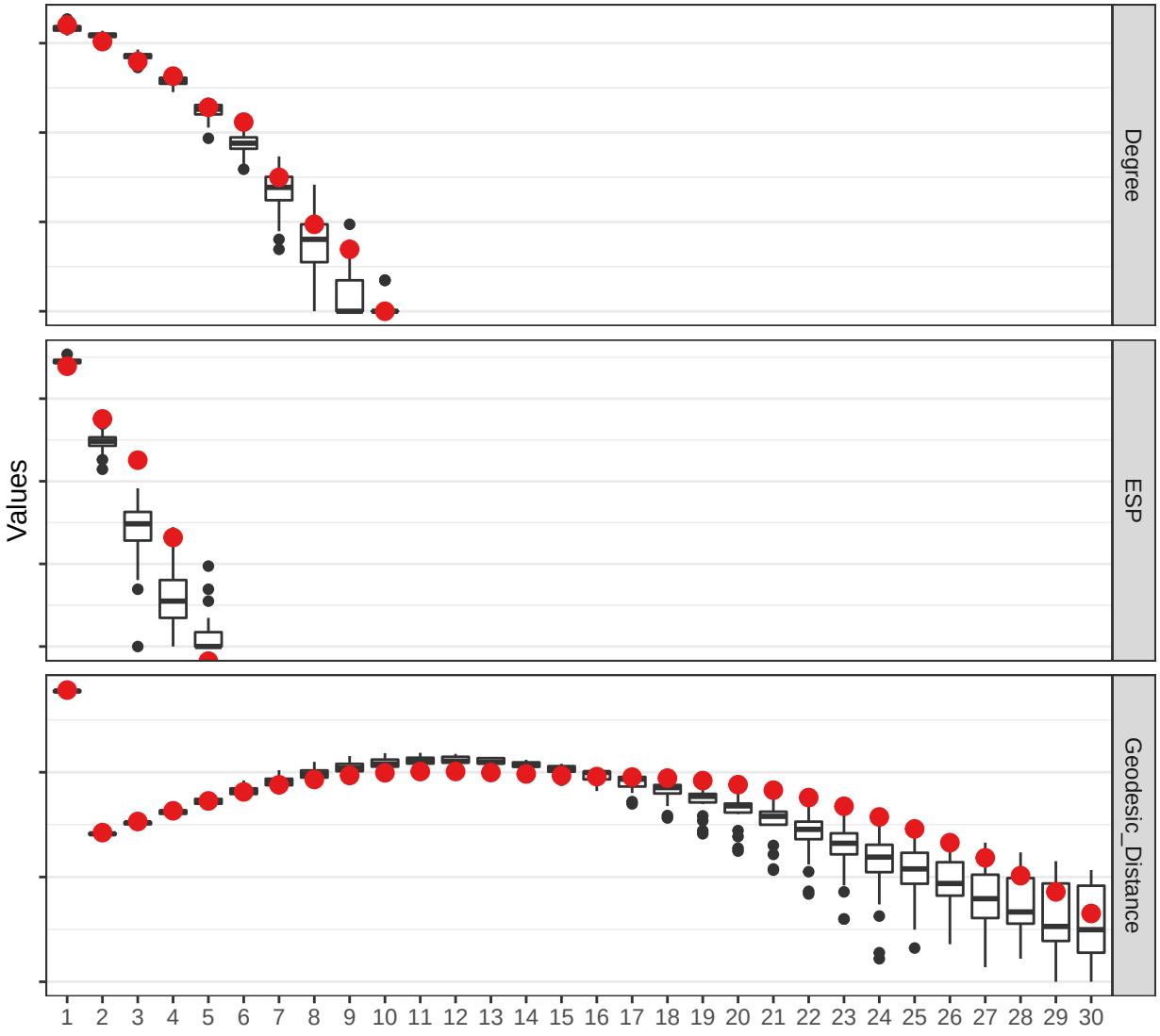


Figure 2.8: Goodness of fit plots for Add Health bayesian LOLOG on logarithmic scale

- The positive significant 2 star and negative 3 star term suggest there is a tendency for popular nodes to form, but with some limiting process on the number of ties a node can maintain.
- There is a tendency for edges to form in dyads that match on grade, race and sex.
- The matching tendency from strongest to weakest is as follows: grade, race, sex.

2.6 Discussion

We have developed a fully Bayesian approach to fitting the LOLOG model to social network data sets. We have justified the use of the pseudo conjugate prior for such analysis both theoretically and practically. We have demonstrated that it is possible to fit the fully Bayesian LOLOG model to a network of 36 nodes and with a plausible relaxation to a network of $\sim 1000s$ of nodes. Through the two examples we able to show consistent interpretation of the posterior distribution with the MOM estimation proceedure.

For smaller networks, it is possible to sample from $p(s|y,\theta)$ the distribution of edge orderings conditional on the observed graph and the model parameter. We showed examples where samples of edge permutations from this distribution, differ in the distribution of change statistics. This suggests that at least for small networks generated by the LOLOG model, the edge ordering has some effect on the nature of the networks generated. Indeed sampling from $p(s|y,\theta)$ may itself be of interest to the applied researcher interested in the order of formation of complex structures in a social network.

The Bayesian approach is theoretically appealing, as classical estimates of parameter estimate uncertainty are only asymptotically valid. In the context of network data, this means as the number of nodes increases the distribution of the estimates converge in distribution to the normal distribution with covariance equal to the fisher information matrix. However modelling with the LOLOG model, as with other generative network models, we assume that the number of nodes is fixed, therefore asymptotic notions are not appealing. The Bayesian

approach allows for clear, theoretically consistent interpretation of parameter uncertainty, subject to the specified prior.

Specification of the prior is challenging but most researchers using a network model to study their data likely have some idea, based on previous literature and empirical experience, of what they believe. However specifying this in terms of a probability distribution over the parameters of a complex model is likely not possible, we proposed using an Empirical Bayes approach, using some features of the data to form our prior, whilst not strictly Bayesian, it does help include prior information which could be conditional on the observed graph. We proposed using the so called edge and edge triangle prior to represent the priors that the model should produce the same edge density and edge and triangle density of the observed network respectively. We showed examples where such a prior produces a well fitting model as well as providing consistent interpretation with the MOM parameters for nodal attribute effects. We note that as researchers are often interested in nodal attribute effects, a prior on the structure which provides interpretation on the attribute effect is a valuable tool.

For larger networks e.g. $\sim 1000s$ nodes sampling from $p(s|y,\theta)$ proved impossible as the size of each edge permutation balloons rapidly with $\mathscr{O}(n^2)$, resulting in an inability to assess MCMC convergence as well as practical storage issues. However we suspect that $p(s|y,\theta) \to p(s)$ as $n \to \infty$, since as the network grows the observed networks likely gives less information about the ordering that formed it. We justify this by thinking conversely, consider a small network with a LOLOG model with a strong triangulation effect but also a strong ceiling effect on degree nodes, triangles in this kind of network are likely to have formed early in the edge formation process as the ceiling effect likely prevents late forming triangles, therefore the observed network and model parameter have a strong bearing on possible edge formation orders. For a large network assuming $p(s|y,\theta) = p(s)$, we were able to fit to a network of size 1681 nodes. We obtained the approximate posterior distribution using the edge triangle prior and demonstrated a good fit as well as consistent interpretation with the MOM approach.

The Bayesian approach allows for a natural extension for Bayesian model averaging which could be explored, as has been developed for the Bayesian ERGM Caimo and Friel (2013).

This could be particularly useful for smaller networks where asymptotic arguments for MOM parameter estimates are not valid, as it is often easy to under or over fit due to the lack of data. We note that as the RMHMC proceedure for exploring the parameter space often works well with a step size of close to 1 when simulating sufficiently many networks for an accurate Fisher information matrix estimation, the posterior parameter space may be in many cases approximately Gaussian, this may lend itself to a variational Bayes approach which could speed up the model fitting procedure, or even be used as an informative prior for the fully Bayesian approach.

It is also possible to pursue an approximate Bayesian computation (Sisson et al., 2018; Marjoram et al., 2004) method. This method is useful when the likelihood of a model is intractable as in LOLOG, but the model can be simulated from with relative ease. Ease of simulation is one of the great strengths of the LOLOG model, we believe this to be a promising method for the LOLOG model, and in particular note that it has been implemented successfully in the related ERGM models Yin and Butts (2020), despite the difficulty in simulating networks from ERGMs.

We note that the computational bottleneck for large networks is the cost of storing the edge permutation of size n^2 which is not sparse. One possible relaxation is to require all edge permutations to be obtained from a sequential node added procedure, thus we could consider edge permutations equivalent if they have the same node ordering, this only require storing node permutations, which would be of size n which would be much more feasible.

Acknowledgements

The project described was supported by grant number 1R21HD075714-02 from NICHD, and grant numbers SES-1230081 and IIS-1546300 from the NSF.

2.7 Appendix

2.8 New England Lawyers Additional LOLOG Model Fits

In this section include details on Bayesian LOLOG model fits with a variety of priors that we considered for the for New England lawyers network.

2.8.1 Comparing Priors

We consider the following priors:

- Flat : Uniform distribution over a wide range for each parameter
- Edge : Empirical Pseudo conjugate prior, predicting with equal probabilty for each edge equal to the density of the observed network.
- Edge Triangle : Empirical Pseudo conjugate prior, using MOM fit with just edges and triangles to predict edges
- MOM : Empirical Pseudo conjugate prior, using MOM fit with all terms to predict edges

Summaries of the priors are shown in tables 2.6, 2.2 and 2.8. We note that the flat prior is not preferred, since it does not capture any of the scale differences between the parameters unlike the edge prior. The MOM prior uses the data pervasively and is therefore not considered suitable as a true prior but is shown for comparison.

2.8.2 Assessing Convergence of MCMC

We omit our assessment of the convergence of the MCMC procedures for each prior, as the prior did not strongly effect the rate of convergence.

Table 2.6: Edge prior distribution summary for LOLOG model for New England Lawyers

	Prior Mean	Prior Std. Error	Prior P-value
Edges	-1.6975	1.9183	0.1819
Triangles	-0.0115	0.1768	0.4804
Node Seniority	0.0000	0.0190	0.5022
Node Practice	0.0185	0.3847	0.4840
Dyad Gender Match	0.0284	0.7316	0.4962
Dyad Practice Match	-0.0280	0.5400	0.4807
Dyad Office Match	0.0145	0.5460	0.4895

Table 2.7: Edge Triangle prior distribution summary for LOLOG model for New England Lawyers

	Prior Mean	Prior Std. Error	Prior P-value
Edges	-2.0421	1.9310	0.1466
Triangles	1.6272	0.3101	0.0000
Node Seniority	-0.0023	0.0201	0.4483
Node Practice	-0.0418	0.3829	0.4584
Dyad Gender Match	-0.0608	0.7469	0.4644
Dyad Practice Match	-0.0135	0.5601	0.4863
Dyad Office Match	-0.0233	0.5661	0.4868

Table 2.8: MOM prior distribution summary for LOLOG model for New England Lawyers

	Prior Mean	Prior Std. Error	Prior P-value
Edges	-7.4394	1.3385	0.0001
Triangles	1.0432	0.4493	0.0113
Node Seniority	0.0269	0.0114	0.0122
Node Practice	0.6903	0.2111	0.0080
Dyad Gender Match	1.0926	0.4741	0.0213
Dyad Practice Match	0.9178	0.2835	0.0016
Dyad Office Match	1.5780	0.3285	0.0004

2.8.3 Sampling Permutations

From these priors we obtain the change statistic distributions derived from samples from $p(s|y,\theta)$. The histograms in Figures 2.9, 2.10 and 2.11 show the different distributions in terms of the triangle, two-star and three-star change statistics. We show each distribution in comparison to the change statistics of the unconditional permutation distribution. We can see that the triangle change statistic distributions become progressively more different to the unconditional sample as the priors use more of the data, suggesting the distribution of edge permutations may be important.

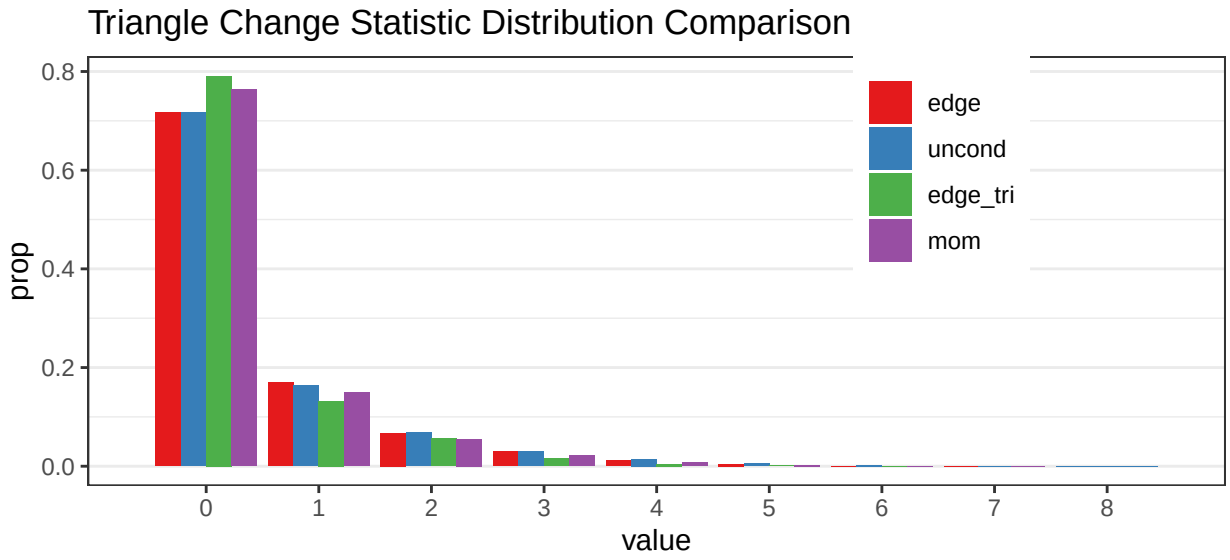


Figure 2.9: Triangle change statistic distribution comparison for posterior edge permutation samples for LOLOG model for New England lawyers network

2.8.4 Sampling from the posterior

For each prior we tuned the sampling procedure to obtain samples from each posterior. Table 2.9 summarises the values of the tuning parameters used. We note that different priors require different tuning parameters to fit in a reasonable time. We suggest this shows that the shape of the posterior parameter space for both θ and s can be different for different priors. For each run we utilised 20 cores on our departmental compute node, sampling θ

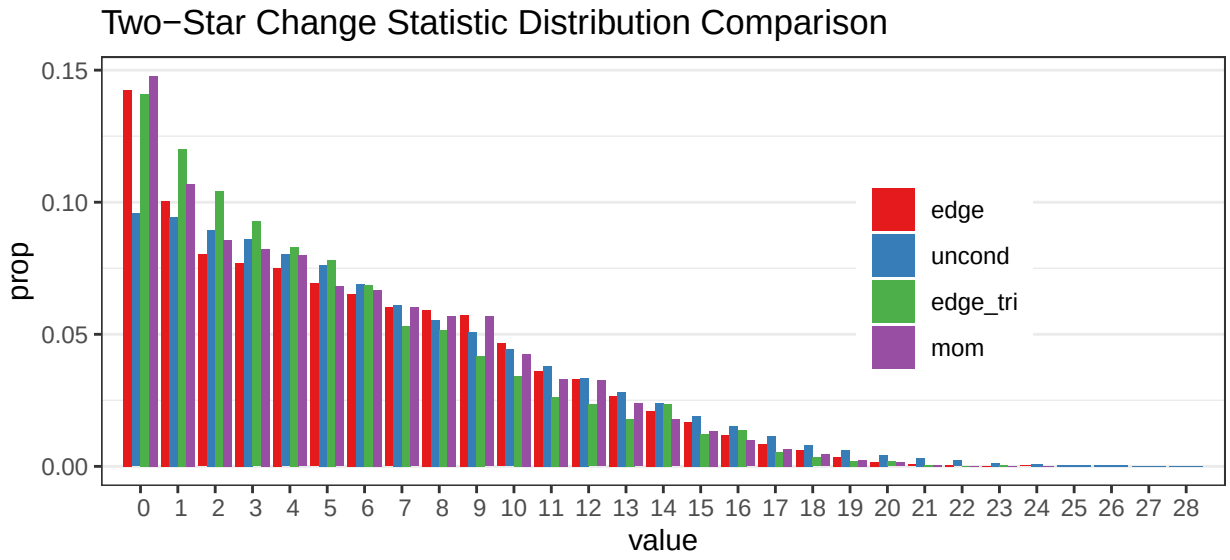


Figure 2.10: Two-star change statistic distribution comparison for posterior edge permutation samples for LOLOG model for New England lawyers network

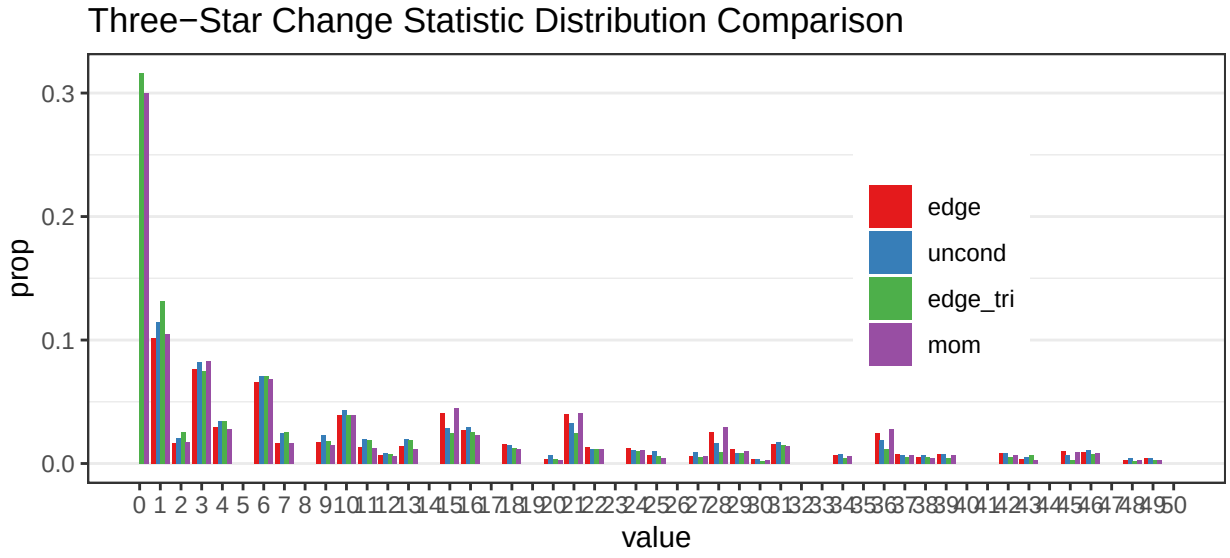


Figure 2.11: Three-Star change statistic distribution comparison for Lazega's Lawyers

Table 2.9: Tuning summary of New England Lawyers

Prior	s_swap	s_burn	RMHMC_sims	RMHMC_step	Theta_burn
flat	NA	NA	200	0.9	100
edge	25%	5000	100	0.5	1000
edge_tri	4%	2000	50	0.3	500
MOM	4%	2000	200	0.9	200

Table 2.10: Flat posterior distribution summary for New England Lawyers

	Posterior Mean	Std. Error	P-value
Edges	-7.6902	0.9535	0.0000
Triangles	0.4255	0.1382	0.0000
Node Seniority	0.0391	0.0095	0.0000
Node Practice	0.7560	0.1638	0.0000
Dyad Gender Match	0.9956	0.3493	0.0005
Dyad Practice Match	0.8117	0.2434	0.0000
Dyad Office Match	1.5110	0.2595	0.0000

from 19 different s_k samples from $p(s|y)$.

We show summaries for the posteriors sampled using each of the priors, in tables 2.10, 2.11, 2.12 , and 2.13.

Table 2.11: Edge posterior distribution summary for New England Lawyers

	Posterior Mean	Std. Error	P-value
Edges	-6.4434	0.4873	0.0000
Triangles	0.2355	0.0823	0.0023
Node Seniority	0.0292	0.0053	0.0000
Node Practice	0.6274	0.0922	0.0000
Dyad Gender Match	0.7974	0.2916	0.0000
Dyad Practice Match	0.6913	0.2033	0.0000
Dyad Office Match	1.2185	0.2183	0.0000

Table 2.12: Edge Triangle posterior distribution summary for New England Lawyers

	Posterior Mean	Std. Error	P-value
Edges	-7.9979	0.6904	0.0000
Triangles	1.4612	0.2519	0.0000
Node Seniority	0.0418	0.0067	0.0000
Node Practice	0.8076	0.1211	0.0000
Dyad Gender Match	0.9116	0.3503	0.0044
Dyad Practice Match	0.6311	0.2502	0.0088
Dyad Office Match	1.1939	0.2679	0.0000

Table 2.13: MOM posterior distribution summary for New England Lawyers

	Posterior Mean	Std. Error	P-value
Edges	-7.6296	1.0406	0.0000
Triangles	1.1066	0.3431	0.0000
Node Seniority	0.0261	0.0089	0.0001
Node Practice	0.7348	0.1632	0.0000
Dyad Gender Match	1.1459	0.3468	0.0000
Dyad Practice Match	0.9367	0.1873	0.0000
Dyad Office Match	1.5675	0.2290	0.0000

2.8.5 Goodness of Fit

We next show the goodness of fit plots for each of the priors using the ESP terms. All the models fitted well on degree terms. Figures 2.12, 2.13, 2.14 and 2.15 show the goodness of fit on the ESP terms. We show the goodness of fit utilising the sample from $p(s|y)$.

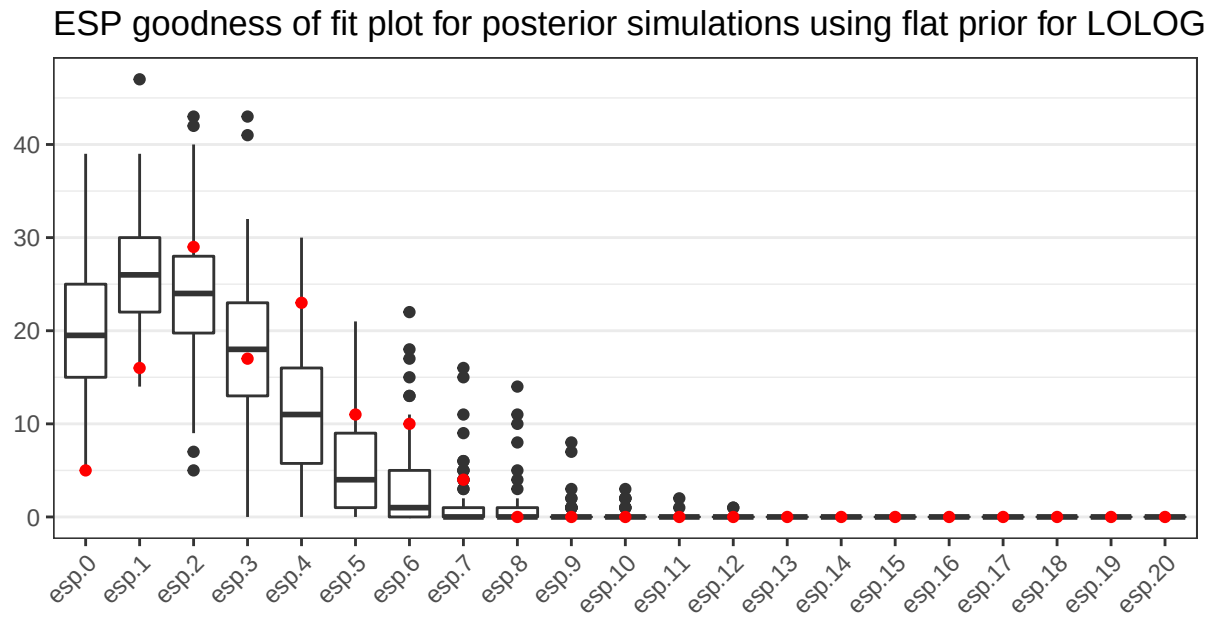


Figure 2.12: ESP goodness of fit plot for flat prior for New England Lawyers

Clearly the flat and edge priors produce models that do not produce the observed graph's esp as a typical observation. The edge triangle and MOM prior produce better fitting models on the esp distribution. Comparing the box plots we can see that in edge triangle prior fit, the observed graph is a less extreme observation than it would have been for the MOM fit.

2.8.6 Interpretation

In this case all priors produced fits that had the same interpretation, this need not be the case. The edge triangle prior LOLOG model fits the ESP distribution best out of the priors considered. This suggests that careful specification of the prior can be important in terms of the fitted model being a likely model to have generated the observed data. We note that

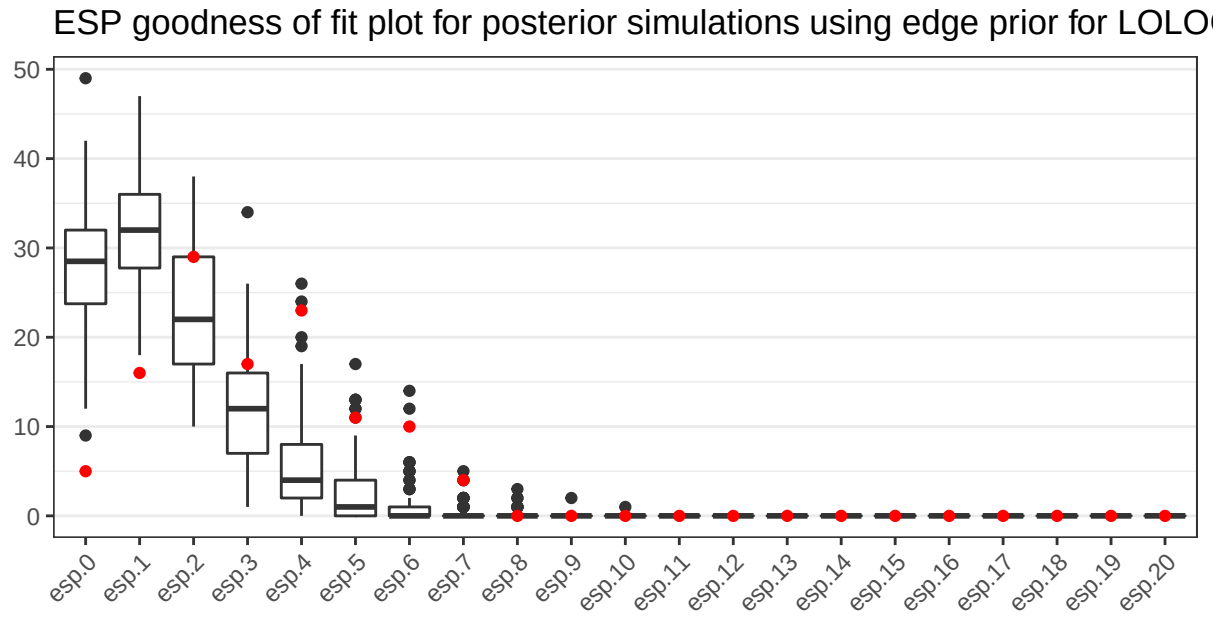


Figure 2.13: ESP goodness of fit plot for edge prior New England Lawyers

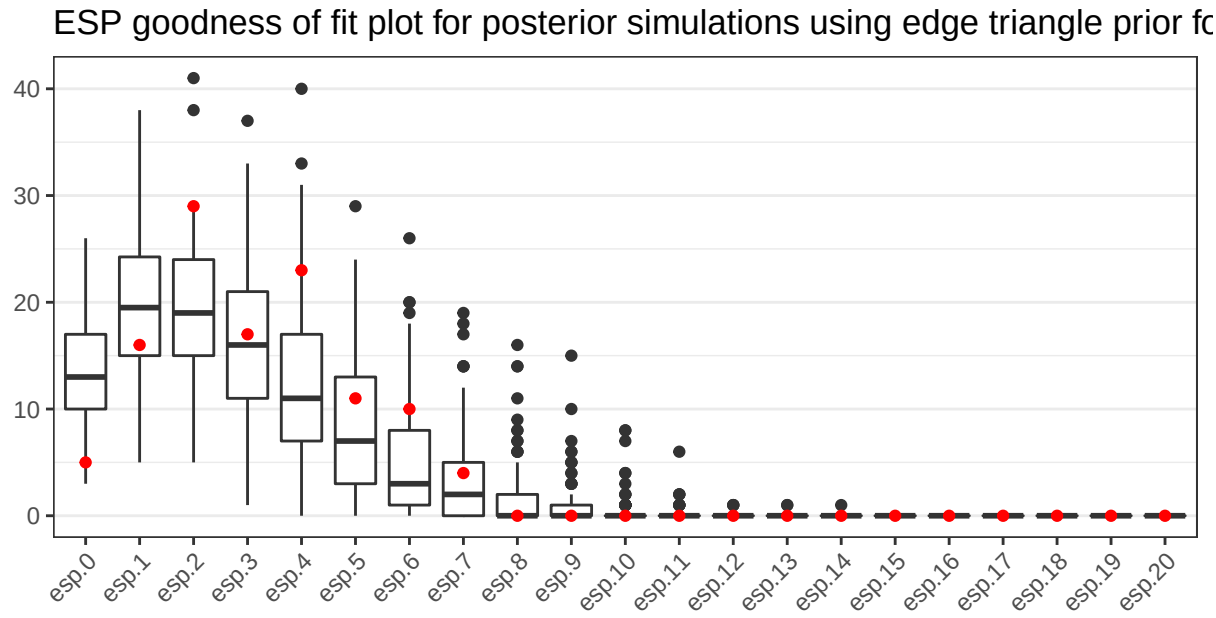


Figure 2.14: ESP goodness of fit plot for edge tri prior for New England Lawyers

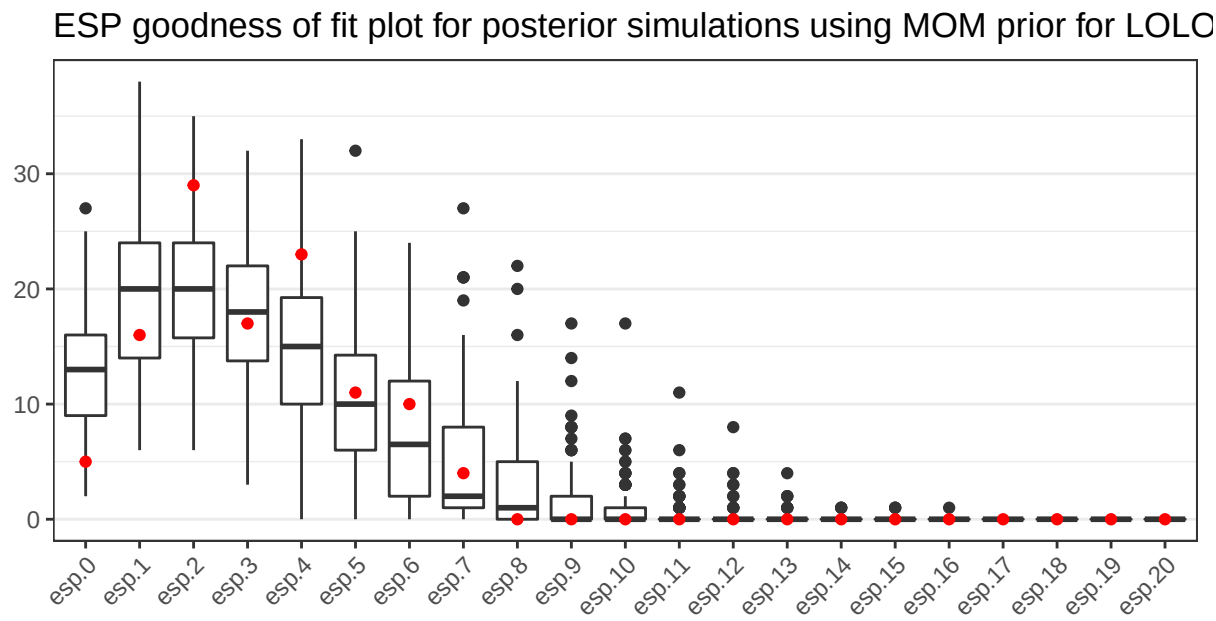


Figure 2.15: ESP goodness of fit plot for mom prior for New England Lawyers

we needed to use some information contained in the data to develop a prior that allowed for tractable inference.

CHAPTER 3

An Approach to Causal Inference over Stochastic Networks

Claiming causal inferences in network settings necessitates careful consideration of the often complex dependency between outcomes for actors. Of particular importance are treatment spillover or outcome interference effects. We consider causal inference when the actors are connected via an underlying network structure. Our key contribution is a model for causality when the underlying network is unobserved and co-evolves with the actor covariates stochastically over time. We develop a joint model for the relational and covariate generating process that avoids restrictive separability and deterministic network assumptions as these rarely hold realistic social network settings. Our framework utilizes the highly general class of Exponential-family Random Network models (ERNM) of which Markov Random Fields (MRF) and Exponential-family Random Graph models (ERGM) are special cases. We present potential outcome based inference within a Bayesian framework, and propose a modification to the exchange algorithm to allow for sampling from ERNM posteriors. We present results of a simulation study demonstrating the validity of the approach. Finally, we demonstrate the value of the framework in a case-study of smoking over time in the context of adolescent friendship networks.

3.1 Introduction

Causal inference is difficult, especially in systems with partially known and likely complex structure. There is an extensive literature on causal inference methods for so called "network

settings" from a variety of perspectives (Hudgens and Halloran, 2008; Shalizi and Thomas, 2011; Ogburn and VanderWeele, 2014; van der Laan, 2014; Sofrygin and van der Laan, 2017; DeAmour, 2016; Aronow and Samii, 2017; Shpitser et al., 2021; Tchetgen et al., 2021; Ogburn et al., 2020).

There are recent empirical studies claiming strong causal results in network settings that have been controversial. For example, claims about social contagion (Christakis and Fowler, 2007, 2008, 2010) with corresponding methodological criticism from others (e.g. Ogburn et al., 2020). Such results are controversial due not only to their surprising and perhaps provocative substantive nature, for example, statements such as "obesity is socially contagious", but also due to the strong assumptions required to justify the methodology.

Much of the problem stems from the unknown underlying social processes. For example, as explicitly noted in Shalizi and Thomas (2011), contagion is often confounded with homophily. They consider Directed Acyclic Graphs (DAGs) (Pearl, 1995; Spirtes et al., 2000) and demonstrate that contagion can be confounded when latent homophily is present.

In this paper we consider social structures represented by networks with stochastic links and covariates that stochastically co-evolve over time. We present a generalized chain graph approximation to a credible social process DAG, which allows for a dependence structure that we believe to be compatible with such problems. We then frame causal inference in terms of network equilibrium potential outcomes, that is, potential outcomes derived from the chain graph structure, and use an augmented variable Markov Chain Monte Carlo (MCMC) algorithm to sample from their posterior distributions. The key contribution of our approach is to allow for uncertainty in the network structure and for codependence of edges and nodal covariates in the underlying social process over time.

Chain graphs have been posited as a possible representation of this causal structure (Tchetgen et al., 2021; Shpitser et al., 2021). By considering a DAG of a network generating process over time, Ogburn et al. (2020) suggest that estimating causal effects in this setting is not viable unless fine grained temporal data is available and the causal structure is simple.

We consider the situation where we observe a single network, considered a realization of a

random social process over both nodes and edges. Causal inference in this setting cannot be reduced to an independent and identically distributed (IID) problem in the same fashion as Sofrygin and van der Laan (2017). The chain graph approximation employed in Tchetgen et al. (2021) and Ogburn et al. (2020) cannot be used as it does not allow for stochastic connections. There have been efforts to allow for network uncertainty (Toulis and Kao, 2013; Kao, 2017). However these methods require unconfoundedness of the edges and the random nodal covariates, given enough fixed information about the nodes. In not making such an assumption we require a joint probability model for the random edges and nodal covariates in the graph.

We note that the literature of network causal inference is at present sharply disconnected from the literature concerning generative models for social networks. As most approaches have considered the networks as fixed, there has been little interest in placing a probability distribution over the space of possible networks. Typically generative models for social networks for example the commonly used Exponential-family Random Graph Models (ERGM) (Frank and Strauss, 1986) consider the edges as the random variables to be modelled and nodal covariates as fixed. There has been much sophisticated work on understanding such models (Handcock, 2003; Robins et al., 2007; Schweinberger and Handcock, 2015), and well developed MCMC based fitting procedures (Snijders, 2002; Handcock, 2002) with associated complex software (Handcock et al., 2018). However due to the assumed fixed nature of the nodal covariates, these are of limited use for causal inference on nodal outcomes. Markov random field (MRF) models treat the nodal covariates as stochastic but the connections between nodes fixed. Encompassing both model classes are the novel Exponential-family Random Network Model (ERNM) (Fellows and Handcock, 2012). ERNM are a class of exponential family models that encompasses ERGMs and MRF as special cases. ERNM allows for the edges and nodal covariates to be stochastic, thus the nodal covariates and the edges can be co-dependent. The focus of this paper is causal inference based on the plausible representation of complex social structure via ERNM.

We utilize a Bayesian framework for the causal quantities and the network model. This

allows for the incorporation of prior information, as well the automatic accounting of uncertainty in a theoretically consistent manner. The Bayesian approach does not require appeals to asymptotic arguments for its validity. Indeed, asymptotics for causal quantities in our setting are conceptually difficult as there is no single asymptotic framework that is compelling. In particular, the number of nodes, N, is a fundamental characteristic of the social process and not a sampling design characteristic, as it is in most of Statistics. For example, the interactions of a class with 5 students will be quite different from a class of size 75. Hence asymptotic approximations must identify credible invariant parametrisations (Krivitsky et al., 2011). Different values of N change the fundamental structure of the social network, as dependent edge behaviour is strongly related to the number of nodes in a network. We develop a modification to the exchange algorithm (Murray et al., 2006) to allow for sampling from ERNM posteriors, which we use to infer the posterior distributions of potential outcomes and, hence, estimate causal estimands of interest.

This paper is structured as follows. Section 3.2 introduces our general network setting and our notation. Section 3.4 considers the DAG of a network process over time allowing for network uncertainty with a chain graph approximation. Section 3.3 defines causal quantities of interest in terms of equilibrium potential outcomes. Section 3.6 briefly describes a simulation study, the details of which are contained in the supplement. Section 3.7 considers a case-study of a network from the National Longitudinal Study of Adolescent Health (Harris et al., 2007b) and gives estimates of unknown causal quantities. Section 3.8 provides general discussion of the method, and its ability to generate credible causal inference.

3.2 Notation and Setting

We consider a known fixed set of N nodes. Each node has a random nodal outcome Y_i, thus the whole network outcome is $Y = (Y_1, \dots Y_N)$. Realizations of the random nodal covariates Y_i and Y are denoted with lower case y_i and y. For this paper we only consider binary outcomes (although the ideas are easily extended to non-binary outcomes). Each node is also

permitted to have further multivariate nodal covariates similarly denoted $X = (X_1, \dots X_N)$ with $X_i \in \mathbb{R}^p$ for some p.

We denote the random edges between nodes as the random variable $A = \{A_{i,j}\}_{i=1,j=1}^{N}$, with realizations a. A can be considered a random adjacency matrix. We also restrict $A_{i,j}$ to be binary with 1 indicating a connection and 0 representing the absence of a connection. For this paper, we make the restriction that our networks are undirected i.e. $A_{i,j} = A_{j,i} \quad \forall i, j$.

A network realization is defined to be a set $\{y, a, x\}$. When considering the dynamic network of the process over time we indicate the outcomes at time t with superscript, for example, the outcome random variable for node i at time t is Y_i^t and the whole network random variable at time t is $\{Y^t, A^t, X^t\}$. If the temporal dynamics result in an equilibrium distribution, we denote it by $\lim_{t\to\infty} (Y^t, A^t, X^t) = (Y, A, X)$.

As the node set is fixed, the nodal covariates X are often in practice fixed throughout the evolution of the social process. Going forward we omit X from our notation, that is for clarity, we consider our networks as realizations of $(Y, A)\,|X = x_{observed}$ but write (Y, A).

We represent the treatment of nodes via the treatment vector $Z = (z_1, \dots z_N)$ with realizations z. For the purposes of this paper we consider the treatment to be applied prior to the evolution of the network process though perhaps conditional on the fixed nodal covariates. We leave allowing for time varying treatments assignments and outcome evolution to future research, though we believe it to be compatible with our approach.

In the following section we will introduce a DAG to represent the dependence structure of the social network. We emphasize that the nodes in the DAGs represent random variables in the stochastic social process underlying the above described social network setting (rather than actors in the social network). That is, nodes in the DAG represent random variables in the social process generating the network, they may either be treatments, random nodal covariates, or random edges.

3.3 Network Potential Outcomes and Causal Estimands of Interest

Hudgens and Halloran (2008), Toulis and Kao (2013) and Aronow and Samii (2017) all considered potential outcome-based frameworks of assumptions and definitions as a basis for causal inference for nodal covariates. We consider network potential outcomes as realizations of an equilibrium distribution of a social process that evolved over time.

Our causal estimands should be interpreted as the effect of an intervention on the equilibrium distribution, this is implicit in Toulis and Kao (2013) and Aronow and Samii (2017). Estimands in Tchetgen et al. (2021) and Ogburn et al. (2020) are based interventions on nodal statuses prior to the evolution of the social process, and estimate network effects, rather than nodal effects. We note that these whole network direct or spillover treatment outcomes for pre social processes interventions, the ERNM (Fellows and Handcock, 2012) presented in Section 3.5.2 is still fully compatible with such regimes, and indeed a generalization of the MRF models used in Tchetgen et al. (2021) and Ogburn et al. (2020).

We set $Y_i(Z, A, Y_{-i})$ to be the equilibrium potential outcome of node i, given the treatment vector of all nodes Z, the network A, and the outcomes of all the other nodes Y_{-i}. In this definition, each potential outcome i depends fully on the entire rest of the edges and nodal potential outcomes in the network. However as our estimands of interest relate to one-step neighbourhoods of nodes, we make the following assumption.

The one-step neighbourhood, assumptions also allows us to dramatically reduce the required number of simulations to estimate missing potential outcomes in section 3.5.

Assumption 1. *One Step Neighbourhoods*

$$\forall i \qquad Y_i(Z, A, Y_{-i}) = Y_i(Z_i, Z_{\mathcal{N}_i(A)}, \mathcal{N}_i(A), Y_{\mathcal{N}_i(A)}) \tag{3.1}$$

where $\mathcal{N}_i(A)$ is the neighbourhood of node i in the network A. That is, the individual outcomes only depend on treatments, edges and outcomes in their own one-step neighbourhood.

We next state the causal estimands that we will pursue in their full generality. In section 3.5 we will consider models, and simplifying assumptions that allow for inference in practice.

Definition 1. *Primary Effects*

$$\xi_i = \frac{1}{|(Z^{-i}, A, Y^{-i})|} \sum_{(z^{-i},a,y^{-i})\in(Z^{-i},A,Y^{-i})} Y_i(z_i = 1, z, a, y) - Y_i(z_i = 0, z, a, y) \tag{3.2}$$

$$\xi^{ave} = \frac{1}{N}\sum_{i=1}^{N} \xi_i \tag{3.3}$$

where (Z^{-i}, A, Y) *are all possible combinations of edges, outcomes and treatments excluding the outcome and treatment of node* i*.*

Definition 2. k*-peer Treatment Effect*

$$\delta_i^k = \frac{1}{|(Z, A)_{\mathscr{N}_i}|} \sum_{(z,a,y)\in(Z,A)^k_{\mathscr{N}_i}} Y_i(z_i = 0, z, a, y) - Y_i(z_i = 0, (z, a) \in (Z, A)^0_{\mathscr{N}_i}, y) \tag{3.4}$$

$$\delta_k = \frac{1}{N}\sum_{i=1}^{N} \delta_i^k \tag{3.5}$$

where $(Z, A)^k_{\mathscr{N}_i}$ *is the set of combinations of node-i-local edge realizations and treatment assignments such that node* i *is exposed to* k *other treated nodes.*

Definition 3. k*-peer Outcome Effect For Binary Outcomes*

$$\delta_i^k = \frac{1}{|(A, Y)_{\mathscr{N}_i}|} \sum_{(z,a,y)\in(A,Y)^k_{\mathscr{N}_i}} Y_i(z_i = 0, z, a, y) - Y_i(z_i = 0, z, (a, y) \in (A, Y)^0_{\mathscr{N}_i}) \tag{3.6}$$

$$\delta_k = \frac{1}{N}\sum_{i=1}^{N} \delta_i^k \tag{3.7}$$

where $(A, Y)^k_{\mathscr{N}_i}$ *is the set of combinations of node-i-local edge and outcome realizations such that node* i *is exposed to* k *other nodes with outcome* 1*.*

We note that the k-peer outcome effect could be considered as a special case of a k-peer treatment effect where the treatment results in outcome 1, almost surely. However we find it

convenient to express the estimand separately as it is often of substantive interest in networks where some other treatment is administered, but the k-peer outcome effect is important.

3.4 Causal Framework

DAGs provide a means for precisely specifying the structure of relationships between random variables (see Pearl (2009) for an introduction). Pearl (1995) developed strict criteria for identification of causal effects.Formal equivalence with the potential outcome framework in Richardson and Robins (2013)

Nodes in the DAG represent random variables, with edges drawn between nodes being strictly uni-directional, and cycles of edges prohibited, so that so called "feedback" loops are not permitted. The interpretation of a directed line from node i to node j is that i causally effects j. For a node indexing set V, denote the corresponding random variables $\{x_v\}_{v\in V}$. This graphical structure encodes the following factorization of their joint probability distribution $p(x) = \prod_{v\in V} p(x_v|x_{pa(v)})$, where $pa(v)$ are the parents of v, that is the nodes in the DAG without edges into v.

The causal effect of X on Y is represented by $p(y|do(X = x))$, the distribution of the random variable Y when X has been externally set to x.Pearl (2009) provides transformations for expressing this new distribution $p(y|do(X = x))$ in terms distribution on observed random variables e.g. $p(y|x)$ or $p(y|x, z)$ for z some other variable in the system.

Chain graphs permit undirected as well as directed edges, which allow for different Markov properties from DAGs. In fact chain graphs can represent dependence structures that are not possible under a DAG. We include some discussion in the supplement however we omit subtleties of their Markov properties discussed in Frydenberg (1990) and utilized for causal analysis in Lauritzen and Richardson (2002). In their full generality, chain graphs can express complex dependence structures. However, in our case, our example chain graph only has one chain component which results in none of the outcomes being rendered conditionally independent, thus we do not require an in-depth review for the purposes of this paper.

Practically, one possible interpretation of the undirected edges in a chain graph is that the two variables interact in a causal feedback sense over time.

DAGs representing the causal structure of outcomes of nodes in networks under both interference and contagion are given a clear and detailed treatment in Ogburn and VanderWeele (2014). We generalize the related conjecture in Ogburn et al. (2020) for a chain graph approximation of causal structure of a social network, which slowly evolves over time. This is based on an interpretation of causality with feedback relationships over time which chain graphs can be used to explain (Lauritzen and Richardson, 2002). Our generalization is to consider a social network where the connections are not fixed and are motivated by the empirical observation that social connections are often strongly dependent on other nodes' connections as well as other nodes' covariates, treatments and outcomes.

As an illustration for our chain graph approximation, Figure 3.1 represents the structure of a network evolving over time for a three node network similar to that of Ogburn et al. (2020) and Tchetgen et al. (2021). Note the analogous connections between node 3 and 1 are omitted for clarity. Block arrows denote multiple arrows into the variables described therein. The full DAG with all arrows becomes quickly unwieldy. In the notation of Section 3.2, the nodal treatments are $\{Z_1, Z_2, Z_3\}$, nodal outcome variables $\{Y_1, Y_2, Y_3\}$, and we denote the undirected edges between nodes i and j as $A_{i,j}$. Directed network edges as well as fixed nodal covariates can also be built-in but are omitted here for clarity. Superscripts denote the status of a variable at that time step e.g. Y_2^t is the outcome for node 2 at time t.

Under this DAG, each of the treatment variables Z_j are permitted to causally effect all of the other outcome variables, as well as edges involving j. Outcomes Y_i^t and Y_j^t are permitted to causally effect edges $A_{i,j}^t$, as well as outcomes Y_i^{t+1} and Y_j^{t+1}. Edges $A_{i,j}^t$ are permitted to causally effect edges $A_{k,l}^{t+1}$ with no restrictions on k and l.

We note that there are key edges that are not present in this DAG. Edges and outcomes, from time step t can only influence edges and outcomes at time step $t+1$, not others. We argue that this is plausible under slow evolution of a network process over time. In general, it is exceedingly rare that enough data are available to identify all the relationships posited

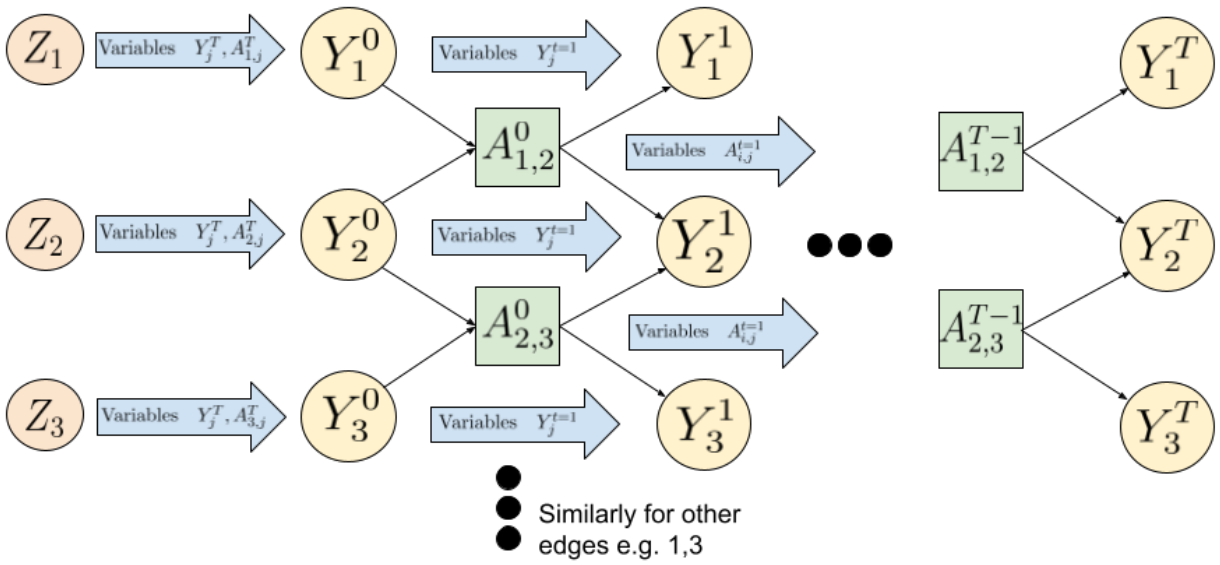


Figure 3.1: Full DAG of temporal network formation feedback process, in the three node network case. Treatment variables Z_j are permitted to causally effect all of the other variables. Outcomes Y_i^t and Y_j^t are permitted to causally effect edges $A_{i,j}^t$, as well as outcomes Y_i^{t+1} and Y_j^{t+1}. Edges $A_{i,j}^t$ are permitted to causally effect edges $A_{k,l}^{t+1}$ with no restrictions on k and l. In practice the functional form of models will usually restrict the causal impact of node i's outcome and treatment to only the set of neighbouring nodes $\{j\}$, that is, where there is an edge present between node i and j.

in Figure 3.1.

The DAG for our system is highly complex, and we usually observe little incremental time data on the social process. Thus operations as introduced in Pearl (1995) to reduce $p(y|do(X = x))$ to expressions involving only distributions, from which we have realizations, e.g., $p(y|x)$, are not possible. That is the true causal effect is unidentifiable. We pursue "approximate causal inference" through approximating the DAG for our social process over time with a chain graph.

As noted in Lauritzen and Richardson (2002), the equilibrium distribution of a so called infinite DAG can be represented as a chain graph. Similarly to Ogburn et al. (2020), but modelling edges as random, the DAG in Figure 3.1 can be approximated by the chain graph shown in Figure 3.2. The chain component (undirected component) of this chain graph is close to being complete since we allow every nodal outcome to influence every other nodal

outcome, however we only allow outcomes Y_i and Y_j to influence edge $A_{i,j}$. In the full DAG, backdoor paths through previous time steps result in this not being the case.

This suggests that the complex structure of such a DAG can be approximated with the chain graph factorization:

$$p(Y, A, Z) = \left(\prod_{i=1}^{n} p(z_i) \right) p(Y, A|Z) \tag{3.8}$$

As an approximation to the true temporal DAG a chain graph model of causality serves to render the causal effects tractable in practice. See Ogburn et al. (2020) and Lauritzen and Richardson (2002) for a fuller explanation.

As an illustration of the generality of this chain graph's dependence structure we define a conditional independence property, that we refer to as *local conditional independence* as follows:

$$Y_i^T \perp\!\!\!\perp Y_j^T, Z_j \quad \Big| \quad A_{i,j}^T = 0, \quad Z_i, \quad \{Y_l^T : A_{i,l}^T = 1\} \tag{3.9}$$

That is, nodal outcomes are conditionally independent given all neighbours, and that they are not connected. Equation (3.9) does not hold *a priori* due to the dependence induced by the random edge $A_{i,j}$ unlike in the fixed network case where it does. That is, the chain graph retains dependencies between outcomes of non-connected nodes, even when conditioning on neighbours. In fact, due to the close-to-complete nature of the chain component, there are few conditional independence assumptions that can be concretely made. Indeed the complexity of such systems is the reason why social network modelling has proven to be difficult.

Conditional independence properties have been considered for ERGMs (Snijders et al., 2006). Common properties induced by model choice in ERGMs are the so-called *Markov* assumption

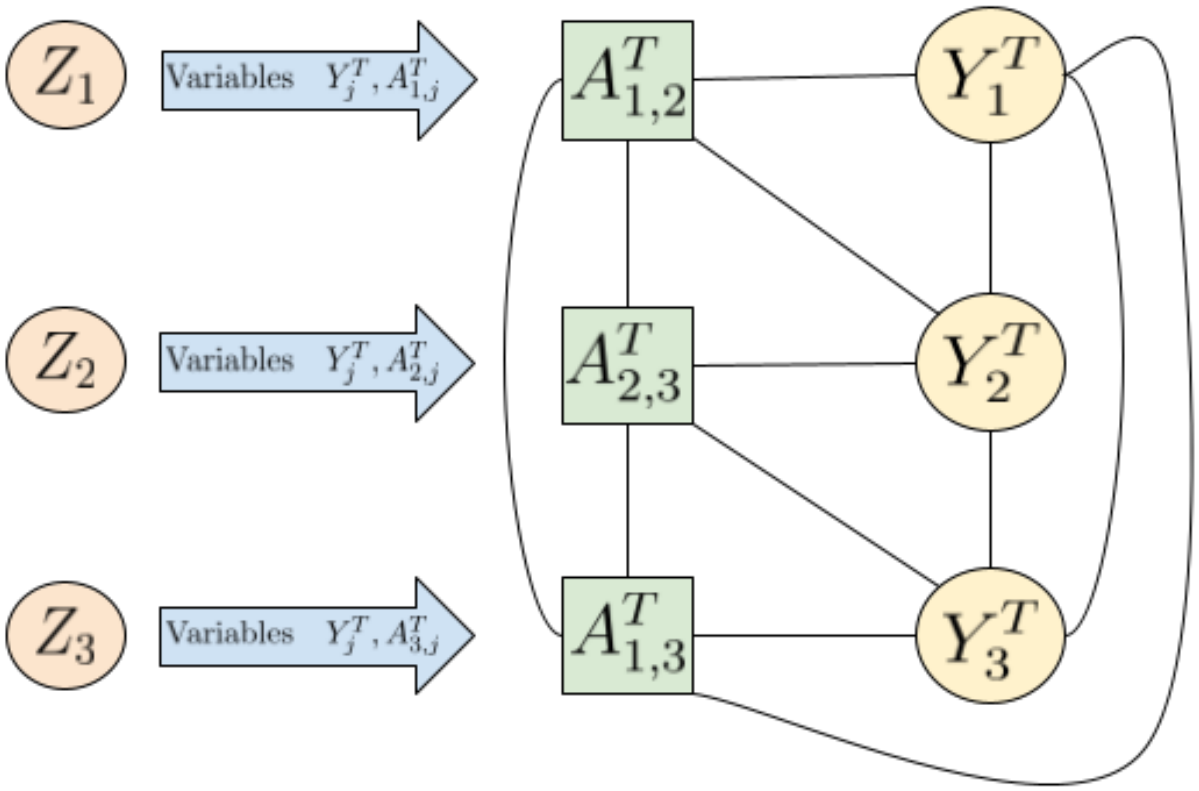


Figure 3.2: Chain graph approximation of a three node temporal social network process with treatment

(Frank and Strauss, 1986) and the so called *social circuit* assumption that requires the Markov assumption condition in addition to a cycle condition (Snijders et al., 2006). Both of these commonly imposed assumptions severely limit possible dependence structures. In practice the social network modelling community has found the social circuit assumption to yield well fitting models in a wide variety of situations (Goldenberg et al., 2010).

We do not comment on the validity of the approximation of the DAG by the chain graph. Ogburn et al. (2020) gave simulations supporting their version of this approximation. We will only consider recovering causal estimands based on the assumption that the chain graph approximation holds.

The remainder of this paper concerns modelling $P(Y, A|Z)$ in order to estimate causal quantities. We specify our causal estimands using equilibrium distribution potential outcomes.

Implicitly the assumption we make in the DAG formulation, under some intervention on the equilibrium denoted with a superscript $*$.

$$P_{equil}(Y_i = y|(A, Z, Y_i^c) = (A, Z, Y_i^c)^*) = P_{equil}(Y_i = y|\text{do}((A, Z, Y_i^c) = (A, Z, Y_i^c)^*) \quad (3.10)$$

We note that intervening on an equilibrium distribution is not possible in practice. However in order to claim the estimation of causal quantities, which are intertwined with social processes, we believe it to be necessary as well as implicit in current potential outcome based approaches (Aronow and Samii, 2017; Toulis and Kao, 2013). We interpret our causal estimands as an average, of the possible treatments, and social processes that led to a given treatment in equilibrium.

The lack of conditional independence assumptions that we are able to make in Section 3.5, is the direct consequence of the nearly complete chain graph model specified in this section.

3.5 Estimation and Identifying Assumptions

Section 3.3 introduced network potential outcomes and our estimands of interest. Section 3.4 justified that they have a casual interpretation. In this section we will introduce models that can be practically used to infer the missing potential outcomes, to produce concrete estimates of causal effects.

Our strategy is to pursue model-based Bayesian imputation of potential outcomes (Imbens and Rubin, 2015).

3.5.1 Structural Assumptions

We next state possible assumptions restricting the dependence structure of the process that will enable us to feasibly impute missing potential outcomes from network simulations. In

our assumptions we explicitly include fixed nodal covariates X.

Assumption 2. *Unconfounded Treatment Assignment Assumption Under Network Assignment*

$$P(Z|X, A, \mathbb{Y}) = P(Z|X, A, \mathbb{Y}') \quad \forall Z, X, A, \mathbb{Y} \quad and \quad \mathbb{Y}'$$

where A is a network, X is a set of nodal covariates on the nodes, and $\mathbb{Y}$ are the potential outcomes.

To account for uncertainty in the network it is possible to make the assumption, as in Kao (2017), that the causal link between the network and outcomes can be broken by the inclusion of nodal covariates. Specifically:

Assumption 3. *Unconfounded Network Assumption under Network Interference*

$$P(A|X, \mathbb{Y}) = P(A|X, \mathbb{Y}') \quad \forall A, X, \mathbb{Y} \quad and \quad \mathbb{Y}'$$

We note that Assumption 2 also follows as a consequence of the chain graph approximation, as in the three node example in Figure 3.2 (See Lauritzen and Richardson, 2002, for discussion on this).

Assumption 3 essentially assumes away the main problem of network causal inference, that the network structure and the potential outcomes are related. The idea is that after including enough nodal covariates, the association between the network and the nodal potential outcomes breaks down. Note that this assumption trivially holds if the network is considered fixed.

Denoting the missing potential outcomes as $\mathbb{Y}_{miss}$ and the observed as $\mathbb{Y}_{obs}$. Using Assumptions 1 and 2, Kao (2017) show that :

$$P(\mathbb{Y}_{miss}|X, Z, A, \mathbb{Y}_{obs}) = P(\mathbb{Y}_{miss}|X, \mathbb{Y}_{obs}) \tag{3.11}$$

Kao (2017) propose choosing covariates X to include so that Assumption 3 is met, and suggest that the addition of covariates derived from modelling the network A may be included to achieve this.

We require only Assumptions 1 and 2, and modelling the potential outcomes jointly with the network. We argue that this is more realistic in most situations, and forces the researches into more realistic modelling choices to arrive at causal inference for nodal outcomes.

The relaxation of Assumption 3 breaks down the proof of Equation (3.11), with the analogous result under the relaxed assumption being:

$$P(\mathbb{Y}_{miss}|X, Z, A, \mathbb{Y}_{obs}) = P(\mathbb{Y}_{miss}|X, A, Y_{obs}) \tag{3.12}$$

The retention of the network A in the conditioning variables, requires that we model its posterior distribution in the imputation of the missing potential outcomes.

3.5.2 Modelling

To facilitate modelling when the network is not fixed, we use the ERNM (Fellows and Handcock, 2012). The ERNM model can be viewed as a generalization of ERGM to allow for random nodal covariates, alternatively and equivalently it can be viewed as generalization of Markov Random Fields (MRF) that allows for random edges. The basic formulation is an exponential family for network a with nodal covariates y:

$$P(A = a, Y = y|\theta) = \frac{1}{c(\theta, \mathscr{N})} \exp\left(\theta \cdot g(a, y)\right) \tag{3.13}$$

The sample space $\mathscr{N}$ is the space of all possible binary edge realizations together with all of the possible random nodes, e.g., $\mathscr{N} \subset 2^{\mathbb{A}} \times \mathscr{Y}^n$, where $2^{\mathbb{Y}}$ is the power set of all the dyads and $\mathscr{X}^n$ is the joint sample space of the nodal covariates. For formal details see Fellows and Handcock (2012).

MRF models can be seen as the ERNM conditional on the network in Equation (3.14), and ERGMs the ERNM conditional on the nodal attributes. The normalising for the conditional distribution constant for the MRF is written $c(\theta, \mathscr{N}(a), a)$ to reflect summing over the restricted space (Fellows and Handcock, 2012).

$$P(Y = y|A = a, \theta) = \frac{1}{c(\theta, \mathscr{N}(a), a)} \exp\left(\theta \cdot g(a, y)\right) \tag{3.14}$$

For example a typical MRF model on a network with treatment effect, outcome homophily as well as outcome and treatment neighbour effects can be written as:

$$p(Y = y|A = a, Z = z) = \frac{1}{c(\theta, \mathscr{N}(a), a)} \exp\Big(\theta_1 \cdot \sum_{i=1}^{N} y_i + \theta_2 \cdot \sum_{i=1}^{N} y_i z_i + \tag{3.15}$$

$$\theta_3 \cdot \sum_{i,j=1,a_{i,j}=1}^{N} \mathbb{I}(y_i = y_j) + \tag{3.16}$$

$$\theta_4 \cdot \sum_{i,j=1,a_{i,j}=1}^{N} y_i Y_j + \theta_5 \cdot \sum_{i,j=1,a_{i,j}=1}^{N} y_i z_j\Big) \tag{3.17}$$

The full ERNM, however, is permitted to include terms typically included to account for common social phenomenon, e.g. social transitivity - the tendency for edges to complete triangles of edges, or social popularity - the tendency for some nodes to have many more connection than others. For example the number of edges, triangles, two-stars and three-stars may used. A typical ERNM model which in addition the the MRF terms in equation 3.15 accounts for transitivity, and centralisation with triangles and two-stars can be written as:

$$p(Y = y|A = a, Z = z) = \frac{1}{c(\theta, \mathcal{N}(a), a)} \exp\Big(\theta_1 \cdot \sum_{i=1,j=1}^{N} a_{i,j} + \tag{3.18}$$

$$\theta_2 \cdot \sum_{i,j,k=1}^{N} \mathbb{I}(a_{i,j} = a_{i,k} = a_{j,k} = 1) + \tag{3.19}$$

$$\theta_3 \cdot \sum_{i,j,k=1}^{N} \mathbb{I}(a_{i,j} = a_{i,k} = 1) + \theta_4 \cdot \sum_{i=1}^{N} y_i + \tag{3.20}$$

$$\theta_5 \cdot \sum_{i=1}^{N} y_i z_i + \theta_6 \cdot \sum_{i,j=1,a_{i,j}=1}^{N} \mathbb{I}(y_i = y_j) + \tag{3.21}$$

$$\theta_7 \cdot \sum_{i,j=1,a_{i,j}=1}^{N} y_i y_j + \theta_8 \cdot \sum_{i,j=1,a_{i,j}=1}^{N} Y_i z_j \Big) \tag{3.22}$$

Tchetgen et al. (2021) utilized Markov random field (MRF) models with coding or pseudo likelihood estimators to power a Gibbs sampling procedure, from which they estimated causal effects of treatment in situations where a single network was observed and the observed data treatment was permitted to be affected by covariates. It is not clear why modern MCMC methods were not used, as pseudo likelihood methods are known to have undesirable properties (Duijn et al., 2009). Ogburn et al. (2020) gave an example with parameters estimated from multiple observations from a MRF using the to "maximize penalized node-conditional likelihoods."

Fellows and Handcock (2012) extended extensive work on MCMC MLE estimation for ERGM (Snijders, 2002; Hunter and Handcock, 2006) to ERNM. However following the Bayesian paradigm and we simulate from the posterior distribution of the missing potential outcomes conditional on the observed data. This accounts for uncertainty in a theoretically consistent manner O'Hagan and Kendall (1993), and removes the need for asymptotic assumptions on the node set which are unrealistic, or bootstrapping Efron (1979) as in Ogburn et al. (2020).

Toulis and Kao (2013) also followed this paradigm, allowing for edge uncertainty with a Poisson edge model, and a linear outcome model. We note that their model does not account

for the dyad dependent nature of real social processes, which is perhaps the most important feature of social network data.

We note that sampling from the posterior distribution of an ERNM is non-trivial, details are contained in the supplement, which require the use of a the exchange algorithm for so called doubly intractable distributions (Murray et al., 2006).

With a suitably simple MRF model with a fixed network or a separable network and outcome model, the causal effects can usually be computed directly from the realized parameter values. For an ERNM this is not the case. Noting that, in full generality, each of the nodal potential outcomes depends on the whole network and all other nodal potential outcomes. The equilibrium distribution of a missing binary potential outcome for node i can be written:

$$P(\mathbb{Y}_{miss,i}(a, y^{-i})|X, A_{obs}, Y_{obs}) = \int_{\Theta} P(\mathbb{Y}_{miss,i}(a, y^{-i})|\theta, X, A_{obs}, Y_{obs}) p(\theta|X, A_{obs}, Y_{obs}) d\theta \tag{3.23}$$

We can then approximate this by simulating a large number of networks $\{(A, Y)_j\}_{j=1}^{M}$, letting $M(a, y^{-i}) = \{(a', y') \in \{(A, Y)_j\}_{j=1}^{M} : a' = a, y'^{,-i} = y^{-i}\}$ be the number of these networks with the required network a and other nodal covariates y^{-i}. This yields

$$P(\mathbb{Y}_{miss,i}(a, y^{-i}) = 1|X, A_{obs}, Y_{obs}) \approx \frac{1}{|M(a, y^{-i})|} \sum_{(a,y) \in M(a,y^{-i})} Y_{miss,i}(a, y) \tag{3.24}$$

Simulating enough networks that have the required network and nodal covariates is infeasible for networks of realistic size. We consider one-step neighbourhoods, that is we allow only nodes connected to the ego to effect the nodal outcome. Thus we dramatically reduce the number of unique potential outcomes for any given node, by requiring that only the treatment assignment, edges involving a node and outcomes of the neighbours of the nodes matter, not the whole network.We note that this is, in fact, a highly restrictive assumption, though is

required to feasibly simulate the missing potential outcomes.

Concretely to estimate the potential outcome for the k peer effect estimand, for each i, instead of restricting to $M(a, y^{-i})$ we restrict to $M^k_{\mathbb{N}_i}(a, y^{-i}) = \{(a', y') \in \{(A, Y)_j\}_{j=1}^M : (a', y') \in (A, Y)^k_{\mathbb{N}_i}\}$. Where $(A, Y)^k_{\mathbb{N}_i}$ is defined in Definition 3. That is we only restrict to simulations where the correct neighbourhood is achieved to estimate the expected value of the missing potential outcomes.

$$P(\mathbb{Y}_{miss,i}(a, y^{-i}) = 1 | X, A_{obs}, Y_{obs}) \approx \frac{1}{|M^k_{\mathbb{N}_i}(a, y^{-i})|} \sum_{(a,y) \in M^k_{\mathbb{N}_i}(a, y^{-i})} Y_{miss,i}(a, y) \tag{3.25}$$

We can then use these expected potential outcomes, to estimate (Bayesian) expected versions of the causal estimands conditional on the observed data.

3.6 Example : Simulation Study

3.6.1 A DAG compatible data generating process

In this section we consider simulating from a 100 node network, with a procedure that follows the true DAG for a social process. Letting A now be the edge random variables, and Y the outcome random variable. We propose the following simulation procedure.

Algorithm 1: Figure 3.2 DAG simulation procedure

Result: (A, Y) sampled from $2^{\mathbb{A}} \times \mathscr{Y}^n$

Assign treatments Z

Let $Y_i = 0 \quad \forall \quad i$

Let $A_{i,j} = 0 \quad \forall \quad i, j$

for $t = 1, 2, \ldots T$ **do**

  Simulate Y^t from $p(Y^t | Y^{t-1}, A^{t-1}, Z)$

  Simulate A^t from $p(A^t | Y^t, A^{t-1}, Z)$

end

The algorithm specified in Algorithm 1 is deliberately abstract, we do not specify the probability functions $p(Y^t | Y^{t-1}, A^{t-1}, Z)$ or $p(A^t | Y^t, A^{t-1}, Z)$ yet.

For our simulations we suggest that choosing $p(Y^t | Y^{t-1}, A^{t-1}, Z)$ and $p(A^t | Y^t, A^{t-1}, Z)$ as a logistic regression, using change statistics as predictors. That is we allow for a proposed tie or node change to be more or less probable based the corresponding change to some specified network statistics. We suggest choosing change statistics in line with our intuition on social processes, for example edges that complete triangles of edges are, all else equal, more likely to form than other edges.

We make a slight simplification: we only allow a single edge or node to toggle at each time step. This results in the probability of a step being exactly the acceptance probability that would be used if we were using a Markov chain to sample from an ERNM, with a simple 1 edge or 1 vertex toggle proposal step. Thus sampling from the DAG with this kind of model for a large enough T is equivalent to sampling from a Markov chain for the corresponding ERNM.

The purpose of implementing simulations is to provide empirical evidence to convince the reader that our method can recover causal effects in the real world case where the true data generating process (DGP) is complex and unknown. The credibility generated by the exercise depends on the credibility of the chosen DGP. If a simplistic DGP is chosen that deliberately fits the method well, the simulations generate less credibility. The fact that

an ERNM sampling procedure can well represent the suggested DAG, should not dissuade the reader of the value of the simulations, in fact that the DAG is compatible with ERNM suggests that our model may in fact be less mis-specified than feared.

In particular with the ERNM DGP it is possible to generate networks with transitivity, homophily and contagion, yet still know the posterior distribution of the causal estimands, for a given network simulation. Other generating processes could have been selected, but would limit the scope of understanding the performance of our method as the true posterior would not be available.

3.6.2 Model Specification

We consider 4 possible DGPs for 100 node networks where 50% for the nodes are treated before the social process evolves. We simulate networks from the ERNM DGP then fit the posterior distributions under that DGP which generates the ground truth posterior distribution. We then fit the posteriors of the remaining three mis-specified DGPs to the simulated networks, and compare the resulting posteriors to the ground truth posteriors.

Table 3.1 shows the proposed parameter values and key properties of the model classes for the 4 DGPs. These ERNM parameters were chosen for simplicity and to achieve a mean degree of close to 3, which might be reasonable in for example a friendship network. We only present the results of simulation from the ERNM model, as it is the only model that represents the DAG.

The ERNM includes a mild spill over through a peer treatment effect, and contagion through a peer outcome effect, where in addition to homophily, peer outcomes and treatments also increase the chance of a positive nodal outcome. We also include a homophilous GWESP term on outcome, that is a GWESP term where edgewise shared partners are only creditted if they match on outcome. This term represents outcome transitivity. We suggest that in many cases a researcher would often believe that these effects are present in a social network formation process, and would fit such an ERNM to observed data.

Table 3.1: Data Generating Process Summary. First block is parameter values for the edge model, second block is the parameter values for the node model, which in the case of the ERNMs are not separable. The third block gives a basic summary of the model classes. We consider networks generated by the ERNM and fit with the other DGPs, the parameter values for the other DGPs are shown to demonstrate the terms included in those models, not for model fitting.

	ERNM	MRF	ERGM+Logistic	Logistic
Edges	4.5	NA	4.5	NA
GWESP	1	1	1	NA
GWDEG	1	NA	1	NA
Outcome Homophily	1	NA	1	NA
Intercept	1	1	1	1
Treatment	1	1	1	1
Neighbors Treated	0.1	0.1	0.1	0.1
Neighbors Outcomes	0.1	0.1	0.1	0.1
Stochastic Edges	Yes	No	Yes	No
Separable Likelihood	No	NA	Yes	NA

The MRF formulation assumes a fixed network with parameters for outcome GWESP, outcome homophily, number of treated neighbours, positive outcome neighbours, main effect and intercept. This may represent a model that a researcher assuming a fixed network, with the simplistic chain graph approximation may adopt. Note that the MRF model can include terms that are functions of both edges and nodes, e.g., outcome GWESP and outcome homophily, but the calculation of these statistics only changes due to the nodes changing, not the edges changing.

The ERGM augmented with logistic regression accounts for network uncertainty with the ERGM, and the nodal outcomes with the logistic regression.

We also consider a pure logistic regression model where the network is only allowed for through the neighbour covariates in the logistic regression.

3.6.3 Results

The causal estimands we consider are the treatment main effect, 1 to 5 peer outcome effects, 1 to 5 peer treatment effect. We simulate 100 network realizations from the ERNM. For each of these realizations we generated samples from the posterior parameter distribution for each of the 4 models. For each of the 400 posterior parameter distributions we sampled 100 parameter realizations and estimated the causal estimands for those realizations. Thus the output of the simulation was 100 simulated networks with 4 posterior distributions for each network, for each of the causal estimands.

We note that this was a computationally demanding simulation. For each of 4 DGPs, we fit 100 posterior distributions to simulated networks from the ERNM. For each of these 100 distributions we then drew 100 samples from the posterior, for each of which we simulated 100 networks to infer the missing potential outcomes. The fitting of each of the posterior distributions typically required of the order of 10^4 burn-in simulations with each step requiring a new ERNM MCMC, which required a toggle burn-in of order 10^4. So each posterior fitting procedure required, the 10^8 ERNM toggles with associated change statistic calculation. As there were 4×10^2 posterior distribution required to be fit, the posterior fitting step required 4×10^{10} ERNM network toggles. Simulating and inferring the missing potential outcomes also requires MCMC burn-ins, though as multiple steps were not required it is a lower order component of the computation time.

Ordinarily the researcher would observe one network, fit one posterior and simulate networks to infer the causal effect, which is feasible for networks of the order of hundreds of nodes.

Table 3.2 shows the mean posterior-mean and the Frequentist coverage rates of the 95% Bayesian credible intervals, together with the true causal estimands of the DGP. The coverage rates are included to enable calibration of the credible intervals (Little, 2011). We also show the mean mean-a-posteriori to justify that, on average, the posteriors are centred around the true value.

We note that the ERGM logistic and pure logistic models recover some outcome effects on

Table 3.2: Mean mean a-posteriori causal estimands fitted to 100 network simulations from ERNM. The coverage of the true mean by the 100 estimated 95% credible intervals is shown in brackets.

	True	ERNM	MRF	ERGM+Logistic	Logistic
main	0.28	0.27 (65%)	0.28 (67%)	-0.03 (0%)	-0.17 (0%)
1-peer-out	0.28	0.27 (69%)	0.36 (31%)	0.18 (59%)	0.16 (8%)
2-peer-out	0.50	0.5 (68%)	0.64 (21%)	0.45 (98%)	0.39 (53%)
3-peer-out	0.66	0.65 (63%)	0.77 (34%)	0.66 (97%)	0.58 (72%)
4-peer-out	0.77	0.74 (57%)	0.82 (52%)	0.76 (95%)	0.7 (77%)
5-peer-out	0.82	0.8 (58%)	0.83 (62%)	0.8 (95%)	0.76 (81%)
1-peer-treat	0.13	0.14 (80%)	0.16 (69%)	-0.06 (0%)	0 (11%)
2-peer-treat	0.27	0.27 (77%)	0.28 (70%)	-0.12 (0%)	0 (13%)
3-peer-treat	0.39	0.39 (71%)	0.4 (70%)	-0.16 (0%)	0 (14%)
4-peer-treat	0.50	0.48 (72%)	NA	-0.2 (0%)	0.01 (15%)
5-peer-treat	0.56	0.54 (70%)	NA	-0.23 (1%)	0.01 (14%)

average, but perform very poorly on treatment effects. The ERNM and MRF posteriors seem to broadly be centred close to the true effects, though the MRF posteriors have much lower Frequentist coverage than the ERNM model, perhaps suggesting optimistically low variance in the posterior distribution. This is as expected as the MRF model does not account for randomness in the edges of the network. In addition the MRF model was unable to identify higher order peer treatment effects, denoted as NA in table 3.2. This is because 4 and 5 peer treatments were not observed in any of the simulated networks.

However if we work consistently in the Bayesian framework, for the networks were generated from an ERNM, the posterior causal estimand distribution derived from the ERNM posterior, is the "ground truth" in the Bayesian sense. Thus the correct assessment of the performance of any given method should be comparing its posterior causal estimand distribution to the ground truth distribution. The comparison for a given method is to compare its posterior fit based on each of the 100 simulated networks to the corresponding posterior derived from the true DGP. Therefore understanding the performance of each method reduces to comparing distributions. We use the relative distribution (Handcock and Morris, 1999) to this end. We consider the relative rank distribution of each of the pairs of models, using boundary

Table 3.3: Mean KL divergence of relative rank distributions of posteriors for causal estimands across 100 network simulations from the ERNM

	ERNM	MRF	ERGM+Logistic	Logistic
main	0	1.09	3.77	4.44
1-peer-out	0	2.45	2.18	3.33
2-peer-out	0	2.59	1.18	1.88
3-peer-out	0	2.2	1.08	1.26
4-peer-out	0	1.56	1.03	1.13
5-peer-out	0	1.22	0.93	1.08
1-peer-treat	0	1.24	4.23	3.52
2-peer-treat	0	1.25	4.24	3.52
3-peer-treat	0	1.34	4.23	3.48
4-peer-treat	0	NA	4.21	3.44
5-peer-treat	0	NA	4.19	3.41

adjusted kernel density estimation using the `reldist` package (Handcock, 2015).

Table 3.3 shows the estimated Kullback-Leibler (KL) divergences between the relative rank distribution and the uniform distribution. To calibrate the size of the divergences, the KL divergence between two unit variance Gaussian distributions is equal to one-half the squared difference between their means. On this scale, a KL divergence of d corresponds to a $\sqrt{2d}$ mean difference. As the ERNM model is being compared against itself, the expect the divergence to be 0. The others have large KL divergences from the ERNM posterior, suggesting that they are not able to recreate the true posterior distribution of important causal estimands when misspecified.

3.7 Case-Study of Smoking Behavior within a High School

In this section we give a real data case-study utilizing ERNM to relax the fixed network assumption as well as conditional unconfoundedness of the edges and nodal potential outcomes. In this case we do not know the ground truth, so the purpose of this section is to demonstrate that our method produces plausible posterior estimand distributions and to

highlight the differences between methods for these data. We also performed a simulation study with known data generating processes which is contained in the supplement.

We consider one of the school social networks from the National Longitudinal Study of Adolescent Health (Harris et al., 2007b). The network we used for this example has 869 nodes of which 462 were male and 407 were female, with 344 having reported trying a cigarette at least once. For consistency with Fellows and Handcock (2012), gender was coded as 1 for male and 0 for females, hence the gender coefficients reported correspond to males.

We consider the following estimands and estimate them under different frameworks:

1. k-peer effect of the gender of peers on smoker status of the ego
2. k-peer outcome effect of peer smoker status on smoker status of the ego.

Within our Bayesian framework we consider the following models for imputing the required potential outcomes to claim causal inference:

1. Full ERNM model with potential outcome as a random nodal covariates.
2. Markov random field model with fixed network
3. Logistic regression.

In our framing the outputs are posterior distributions of causal estimands. For information we also show the results of the MRF model, with parameters estimated through maximum pseudo-likelihood estimation and potential outcomes derived from these as in Tchetgen et al. (2021). In line with the known bias of pseudo likelihood estimates for ERGM (Duijn et al., 2009) we believe this method will perform poorly.

We used a version of the exchange algorithm (Murray et al., 2006), with an extension which allows for efficient sampling from the ERNM posterior. The development is given in the supplement. While informative priors are compatible with the computational framework,

Table 3.4: Summary of posterior distributions of network models, posterior means are shown with posterior standard errors in parentheses

	ERNM	MRF	Logistic Regression
Edges	-4.93 (0.03)	NA	NA
Grade GWESP	0.11 (0)	NA	NA
GWDEG	-1.67 (0.36)	NA	NA
Grade Homophily	4.13 (0.05)	NA	NA
Sex Homophily	0.67 (0.08)	NA	NA
Smoke Homophily	0.48 (0.06)	0.43 (0.07)	NA
Intercept	0.95 (0.11)	0.75 (0.12)	-1.07 (0.2)
Gender	-0.07 (0.03)	0.02 (0.04)	0.47 (0.16)
Female neighbors	0.11 (0.02)	0.08 (0.02)	-0.17 (0.05)
Smoker neighbours	-0.64 (0.12)	-0.46 (0.12)	0.54 (0.06)
Stochastic Edges	Yes	No	No
Stochastic Covariates	Smoker Status	Smoker Status	Smoker Status
Separable Likelihood	No	NA	NA

here we report based on a uniform prior over all parameters. The results do not appear to be sensitive to the choice of prior.

Table 3.4 gives a summary of the posterior distributions each of the models, showing the posterior means with the posterior standard errors in parentheses. We note that parameters should not be compared across models, as the functional forms are different, we show this table to summarize the posteriors, but to also highlight the differences between the models.

For GWESP and GWDEG terms decay parameters were fixed at 0.5. The use of these geometrically weighted terms is in part necessary to avoid degeneracy issues (Handcock, 2003; Snijders et al., 2006), but also implicitly induces the social circuit dependence assumption for the ERGM, rather than the more restrictive Markov assumption. However we used the ERNM style homophily terms (Fellows and Handcock, 2012) for both the ERNM and ERGM model, which in fact induce non local dependencies. Thus the dependence structures of the edges in these models are unknown and best described as "complex." In addition we enforce homogeneity in school grade for the edgewise shared partners in the GWESP term as there

is a very strong grade structure to transitivity in this network.

We do not interpret the posterior parameter distributions directly, rather we make comparison through the smoker peer effect on the smoker status of a node.

Table 3.5 and Figure 3.3 show the the k-smoker-peer effect estimates. These are is estimated as the additional chance of smoking that having k smoker friends has over having no smoker friends. The ERNM and MRF model are in agreement for peers one to three, with some divergence after this. The logistic regression model is markedly different from the ERNM for one and two peer effects, while for higher effects the estimates are closer to the ERNM estimates. The pseudo likelihood estimated MRF model, as expected, are quite different. This helps confirms our prior belief that this estimation method is likely biased in the Frequentist sense, consequently the fitted model does not fit the data well, and does accurately estimate causal estimands.

We believe the ERNM to be most plausible from a theoretical perspective. Whilst in this example the effect size difference from the MRF model was not large, we believe it to be a more robust approach when estimating network causal effects. In particular where there is strong transitivity interacting with nodal outcomes as well as for smaller networks, we expect the effect would be larger.

In addition we show Bayesian posterior prediction goodness-of-fit graphics in the style of Hunter et al. (2008) in the supplement. These demonstrate that networks simulated from the ERNM model posterior, correspond closely to the observed data.

3.8 Discussion

In this paper we model causality when the underlying population is networked and that network is unobserved. Our approach jointly stochastically models the links and nodal covariates in the network, better representing our state of knowledge and their codependency.

Considering a DAG, we suggest that the estimating true causal effects in this setting, is

Table 3.5: Posterior means of the k peer smoker outcome effect ATEs, for various methods. The pseudo MRF value is the mean simulated from the parameter estimate.

ERNM	MRF	pseudo_MRF	Logistic Regression
0.12 (0.01)	0.13 (0.01)	0.17 (0.02)	0.08 (0)
0.22 (0.02)	0.23 (0.02)	0.28 (0.02)	0.18 (0.01)
0.35 (0.02)	0.33 (0.02)	0.35 (0.02)	0.3 (0.02)
0.48 (0.04)	0.42 (0.03)	0.4 (0.02)	0.42 (0.03)
0.59 (0.04)	0.5 (0.03)	0.45 (0.02)	0.53 (0.04)
0.68 (0.05)	0.57 (0.04)	0.49 (0.02)	0.63 (0.04)
0.75 (0.04)	0.63 (0.04)	0.52 (0.02)	0.7 (0.04)

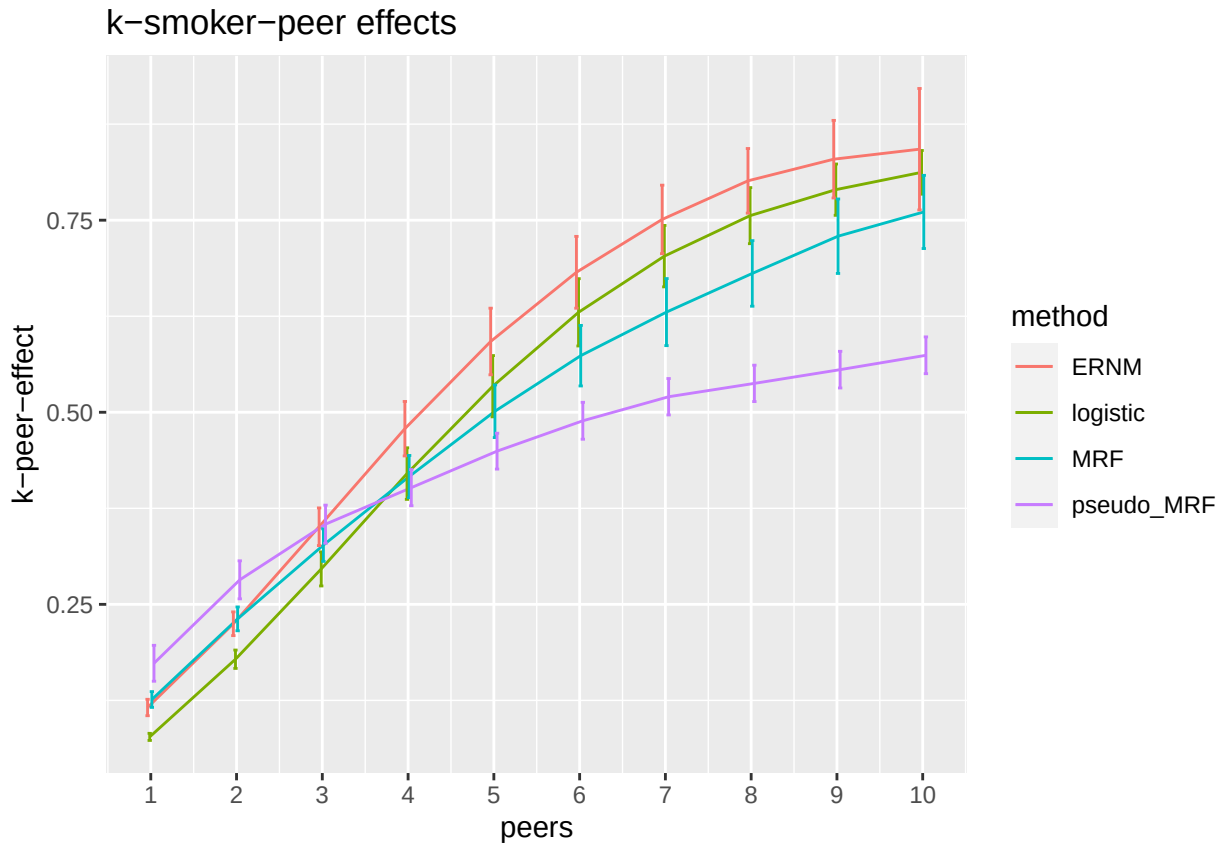


Figure 3.3: Plot comparing the posteriors distribution of the ERNM, MRF and logistic regression estimated k-peer outcome effect. The MRF with parameters equal to the maximum pseduo likelihood esimtate is also shown

almost always intractable due to the usual lack of fine grained temporal data, as well as the highly complex causal structure. We present a chain graph approximation to the DAG, which allows for a dependence structure that we believe to be compatible with such problems. We then frame the approximate causal inference in terms of network equilibrium potential outcomes, that is, potential outcomes that are free to depend on nodes in the neighbourhood of the node in question. We propose the use of ERNMs to jointly model both random edges and nodal covariates. We also develop a simple modification to the exchange algorithm allowing for a feasible sampling from the posterior distribution. We use the posterior distribution, through simulation, to impute the distribution of the missing potential outcomes, allowing the consideration of the distribution of the causal estimands. We showed, using a school network from the National Longitudinal Study of Adolescent Health, that failing to account for the network structure of the problem could lead to misleading qualitative conclusions, in particular when considering the one- and two- peer outcome effect.

Our primary contributions lie in considering the consequences of relaxing of the commonly made fixed network assumption and proposing the use of suitable social network models to estimate causal effects. This relaxation complicates the causal structure considerably and necessitates the use of complex models to derive causal estimands. Clearly the relaxation of this assumption also allows for greater generalizability. Our inferences hold for the given node set and social process, whereas assuming the fixed network narrows the scope to that observed network only. Our method is only applicable to the given node set. However we suggest that the posteriors derived from the given network can serve as strong priors for "similar" networks. In general, qualitative features of the posteriors from the given network can useful inform other analyses. It is not possible to make further statements than this for networks of different sizes.

We believe that in the context of social network analysis, where individual attributes are heavily influenced by social context, such covariation of edges and nodal covariates is overwhelmingly more representative of many social processes. That is, in many social networks we believe it highly likely that individual characteristics and the connections that form be-

tween individuals are strongly dependent. This is especially true if the network evolves over time. We also note that avoiding such issues by letting the size of the network become large and invoking asymptotic arguments, fundamentally mis-interprets the problem. The phenomenon of interest; the inter dependencies of actors, is a result of the small size of the network, for example dependence assumptions implicitly made in modelling a 30 node network usually do not make sense to apply to say a network of 3000 nodes. Thus arguments that rely on the number of nodes being large, are incoherent, as they rely on changing the structure of the problem itself, to understand uncertainty. As our approach is fully Bayesian on the fixed node set, we do not require asymptotic arguments on the number of nodes in a network.

Notable by its absence is discussion of suitable prior distributions for ERNMs. We note there has been some work developing conjugate prior distributions for ERGMs (Wang, 2011) and strongly suspect that such an approach may also be applicable in our setting. In practice flat Gaussian prior are often used for ERGMs (Caimo and Friel, 2011). For the purpose of demonstrating our approach, we used uniform priors, which make no account for the geometry of exponential families, but allow us defer careful consideration of possible priors to future work, while demonstrating the utility of our method. We note that we performed similar posterior fits with flat Gaussian priors, which did not effect the posteriors substantially.

The cost of our approach is the strong assumption that such a complex process, can be adequately modelled by an ERNM. This is in general the main criticism of model based causal inference approaches, that models are mis-specified with unknown consequences. In a network setting this mis-specification is often acute e.g. constant marginal effect of additional smoker friends in the linear potential outcomes model. Our central argument is that a complex model is much less mis-specified than current approaches. We have sought to justify this with real data and simulations, though propose this as a future area of research. For example, how dependent do outcomes in networks need to be to invalidate conclusions made with unrealistic models?

The mis-specification may seem to be cause for pessimism, however we emphasize that

network settings are indeed the extreme case of small data, as we usually only have one observations of a network on a fixed set of nodes. Thus intuitively we should expect strict functional form assumptions to be required to generate any meaningful statistical, and especially causal inference. In fact we argue that approaching network problems with simpler assumptions is problematic, whilst potentially less prone to mis-specification in the sense that simple models can be used, this easily glosses over the inherent difficulty of dealing with network data where nodes and edges are strongly dependent on other nodes and edges.

we also note that specifying a model for the full network data generation process also allows inference in cases where only a subset of the network is sampled. Accounting for such sampling structure is likely analogous to the method for ERGMs in Gile and Handcock (2016). Accounting for this is not possible with the other methods considered in this work.

We believe that meaningful steps can be made towards causal inference on networks, through careful consideration of the complex causal structure of such problems. Whilst we make strict assumption on the function form of this, if the researcher is unwilling to make such assumptions, we opine casual inference is out of reach. We suggest it is better to acknowledge the complexity of the situation, and therefore claim that causal inference is not possible, than employing highly restrictive assumptions on the dependence structure of the data generating process, to allow simpler models to be employed.

3.9 Acknowledgements

This article is based upon work supported by the National Science Foundation(NSF, MMS-0851555, SES-1357619, IIS-1546259) and National Institute of Child Health and Human Development (NICHD, R21HD063000, R21HD075714 and R24-HD041022). The content is solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

3.10 Supplementary Materials

Supplement The supplement contains an additional chain graph approximation diagram, a review of ERNM and Bayesian computation for them and a goodness of fit analysis for the modelling of the adolescent health network, including a MCMC convergence analysis. (pdf)

3.11 Chain Graph Introduction

This section gives a very brief introduction to chain graphs, to help understand why they are appropriate for approximating the structure of a social process that evolves over time. This section exclusively repeats material from Frydenberg (1990), Lauritzen and Richardson (2002), and is included for completeness as the entire article relies on chain graph concepts.

3.11.1 Graph Definitions

We first define a graph $\mathscr{G}$ as a set $\{V, E(V)\}$ where V is a finite set of vertices, and $E(V)$ is a set of edges between the vertices contained in V, both directed $\rightarrow$ and undirected $-$ edges are permitted. We denote the subgraph induced by a vertex a as $\mathscr{G}_a$

A partially directed cycle is defined as subset of $\{v_i, \ldots, v_k\} \subseteq V$, re-indexing as necessary, such that for each i, $v_i \rightarrow v_{i+1}$ or $v_i - v_{i+1}$ with at least one of the edges directed. A chain graph is defined as a graph with at no partially directed cycles.

There exists a path between u and v if $u = v$ or if u and v are connected by a sequence of edges through vertices $\{v_1, \ldots, v_k\}$. A path is directed if any of edges between v_i and v_{i+1} are directed and in the same direction i.e. v_i to v_{i+1} or v_i to v_{i-1}.

For vertices $v, u \in V$ we define u as a parent of v if $u \rightarrow v$ and write $u \in pa(v)$ for undirected edges we define u and v neighbours if $u - v$ and write $u \in ne(v)$ and $v \in ne(u)$. The boundary of v is defined as the union of the neighbours and the parents, the closure $cl(v)$ is the $v \cup bd(v)$

A subgraph induced by a vertex set A is denoted $\mathscr{G}_A$ and contains all vertices in A and all edges involving vertices in A. For vertices a and b minimal complex is an induced subgraph of the form $a \to v_1 - \cdots - vr \leftarrow b$.

The chain components $\mathscr{T}$ of a chain graph G are the connected components of G with all directed edges deleted.

We define the future $\phi(U)$ and past $\pi(U)$ for a vertex u as $\phi(u) = \{v \in V | \exists \text{ a path from } u \text{ to } v\}$ and $\pi(u) = \{v \in V | \exists \text{ a path from } v \text{ to } u\}$. The future and past of a vertex set U is defined as the union of the futures and past of all the nodes in the set. A set is terminal if $\phi(U) = \emptyset$ and initial if $\pi(U) = \emptyset$.

A subset of vertices U is anterior if it can be generated by removing terminal chain components. This allows us to finally define the moral graph, which will be used to define the Markov property for chain graphs. The moral graph $\mathscr{G}^m$ is the graph formed by considering the chain components of the graph, and completing their boundaries, i.e. replacing directed edges with undirected edges between chain components and their parents, and connecting all parents with undirected edges.

3.11.2 Chain Graph Markov Property

Graphical models serve to encode the structure of conditional independence relationships between variables. The Markov property of a graphical model is this encoding. The chain graph Markov property was defined and investigated in Frydenberg (1990).

We restate Theorem 3.3 of Frydenberg (1990) which establishes the equivalence of 4 proposed Markov properties.

Theorem 1. *Chain Graph Markov Property Equivalences Theorem 3.3 of Frydenberg (1990)*

For a chain graph $\mathscr{G}$ and probability measure P such that:

$$(A \perp B|D \cup C \quad and \quad A \perp C|D \cup B) \quad implies \quad A \perp B \cup C|D$$

The following are equivalent:

1. *Global $\mathscr{G}$-Markovian : $A \perp B|C$ whenever C separates A and B in $\left(\mathscr{G}_{an(A \cup B \cup C)}\right)$ - the moralization of the anterior set of the union.*
2. *Local $\mathscr{G}$-Markovian : $u \perp [V \setminus \phi(u)] \setminus cl(u)|bd(u) \quad \forall \quad u$*
3. *Pairwise $\mathscr{G}$-Markovian $u \perp v|[V \setminus \phi(u)] \setminus \{u, v\}$ for $v \notin \phi(u)$, and u and v not adjacent*
4. *$\mathscr{G}_A^m$-Markovian for every anterior set A of $\mathscr{G}$. Where $\mathscr{G}_A^m$-Markvoian refers to the undirected Markov property in Frydenberg (1990)*

These Markov properties lead to the factorization of all chain graph models in Theorem 4.1 of Frydenberg (1990).

Theorem 2. *Chain Graph Factorizations (Partial Version of Theorem 4.1 of Frydenberg (1990))*

For a distribution P with positive density p w.r.t. some product measure $\mu = \bigtimes_{v \in V} \mu_v$ on space $\mathscr{H}$. If P is $\mathscr{G}$-Markovian then p can be factorized as :

$$p(x) = \prod_{\tau \in \mathscr{T}} \prod_{C \in \mathscr{C}_\tau} \psi_\tau^C (X_C) \tag{3.26}$$

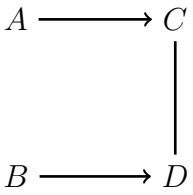
Where $\mathscr{T}$ denotes the set of chain components in $\mathscr{G}$ and $\mathscr{C}_\tau$ denotes the set of cliques in $\left(\mathscr{G}_{cl(\tau)}\right)^m$

This theorem provides that factorisation of chain graphs as described in Lauritzen and Richardson (2002) as a "DAG of boxes" where the boxes are the set of chain components.

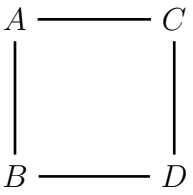
3.11.3 Chain Graph Example Comparison with DAG

In this section we follow the example given in Section 5 of Lauritzen and Richardson (2002). We give a simple example of conditional dependence structures that cannot be represented with a DAG, but can be with a chain graph.

Figure 3.4a shows a simple chain graph. The density can be factorised under the chain graph factorisation theorem as $f(a,b,c,d) = f(c,d|a,b)f(a)f(b)$. However the chain graph markov property places restrictions on the form of $f(c,d|a,b)$.



(a) Full chain graph



(b) Moralization of smallest anterior set containing $\{a,b,c,d\}$. Moralization, connects the largest chain component to its parents with undirected edges, and connects these parents with undirected edges also

Figure 3.4: Full chain graph example and moral graph

Figure 3.4b shows the transformation to the moralization of the anterior set of the union, so we can use the global Markov property. As the sets $\{b,c\}$ and $\{a,d\}$ separate a and d and b and c respectively we have:

$$a \perp d|\{b,c\} \qquad b \perp c|\{a,d\} \tag{3.27}$$

In addition using the pairwise Markov property and noting that $\phi(a) = \{c,d\}$ we have that:

$$a \perp b | [V \setminus \phi(a)] \setminus \{a, b\} \tag{3.28}$$

$$\rightarrow a \perp b | \emptyset \tag{3.29}$$

$$\rightarrow a \perp b \tag{3.30}$$

Figure 3.11.3 shows 2 DAGs that one might think encode the Markov property as Figure 3.4a. Figure 3.5a shows of which is an common parent for vertices c and d the second of which is a common child of vertices c and d. We note that following Pearl (1995) since vertices c and d are colliders, conditioning on them leads to dependence so that $a \not\perp d | \{b, c\}$ and $b \not\perp a | \{a, d\}$. Figure 3.5a shows a DAG where we have $a \not\perp b$ there is conditioning on a common child.

(a) DAG with common parent

(b) DAG with conditioning on common child

Figure 3.5: Example DAGs representing different Markov properties to the Chain graph in Figure 3.4a

3.12 Four Node Chain Graph Example

Figure 3.6 shows a chain graph approximation to the true DAG for a four node social process. Although there are six edge variables $A_{i,j}$, each edge variable is only connected to four other edges. In general the full deletion of connections in the chain graph is regarded as a very restrictive assumption on the network dependence structure in the social networks literature. We note that for networks of larger size the proportion of connections missing compared to the complete chain graph grows. Thus for networks of size in the hundreds of nodes, as in our

examples, the chain graph approximation strongly restricts the full dependence structure. It is this restriction that allows for the modelling of the complete network.

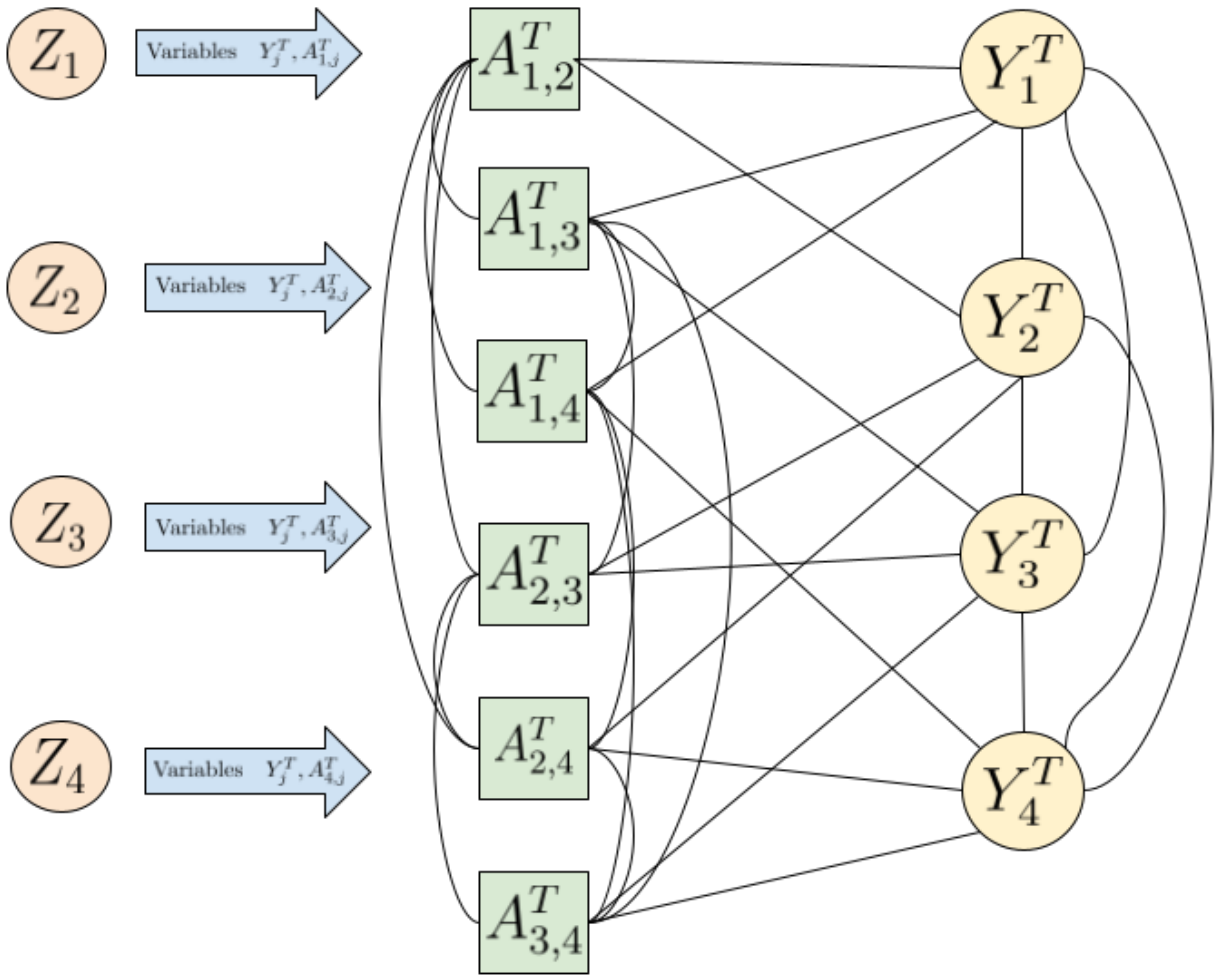


Figure 3.6: Chain graph approximation of a 4 node temporal social network process with pre process treatment and the Markov assumption for edge dependence

3.13 Exponential-family Random Network Models

In this section we briefly introduce the Exponential-family Random Network model (Fellows and Handcock, 2012) and explain the exchange algorithm (Murray et al., 2006). This section gives a brief introduction to a complex method, this paper uses the ERNM and exchange

algorithm machinery, though it is not the primary purpose, so the unfamiliar reader should consult the above citations. The exchange algorithm is used for sampling from ERNM posterior distributions, sampling from ERNMs is completely analogous to the exchange algorithm for ERGM (Wang, 2011; Caimo and Friel, 2011).

The ERNM model can be viewed as a generalization of ERGM to allow for random nodal covariates, alternatively and equivalently it can be viewed as generalization of Markov Random Fields (MRF) that allows for random edges. The basic formulation is an exponential family as follows, for network y with nodal covariates X.

$$P(Y = y, X = x|\eta) = \frac{1}{c(\eta, \mathscr{N})} \exp\left(\eta \cdot g(y, x)\right) \tag{3.31}$$

The sample space $\mathscr{N}$ is the space of all possible binary edge realizations together with all of the possible random nodes, i.e $\mathscr{N} \subset 2^{\mathbb{Y}} \times \mathscr{X}^n$, where $2^{\mathbb{Y}}$ is the power set of all the dyads and $\mathscr{X}^n$ is the joint sample space of the n nodal covariates. For formal details see Fellows and Handcock (2012).

In the ERGM framework it is usual to include statistics that are counts of specific sub graph realizations, for example the number of triangles and stars in a network. This accounts for the social structure of many real networks, which often exhibit more social closure or transitivity than can be explained by covariates alone. In addition one can include terms typically included in a MRF models e.g. the number of neighbours a node has with a given covariate.

Similar to ERGMs, ERNMs have problems with degeneracy (Handcock, 2003). Fellows and Handcock (2012) comment on this, in particular they propose the use of a somewhat unusual term for homophily which is usually required to be able to fit ERNMs. We use this term in our examples, as well as geometrically weighted edgewise shared partner and degree terms (Snijders et al., 2006) to allow us to fit models that are not degenerate.

ERGMs have well defined dependence structures due to the Hammersley-Clifford Theorem

(Besag, 1974). The dependence structure of an ERGM is directly related to the choices of sufficient statistics. A version of the Hammersley-Clifford Theorem is applicable to ERNM models, though we do not prove it here. Thus, though we consider a close to complete chain graph to represent the true dependence structure of the network, we implicitly impose strict dependence structure on the edges of the network, through our choice of statistics in the ERNM. However in practice the required choice of homophily statistic for ERNM renders the dependence structure complex. See Snijders et al. (2006) for a discussion of the dependence structures associated with various sufficient statistics in the related ERGMs.

Sampling from ERNMs directly is intractable due to the normalizing constant $c(\eta, \mathscr{N})$ which is in fact a astronomically high dimensional sum over the space of all possible graph realizations $\sum_{(y,x)\in 2^{\mathbb{Y}}\times\mathscr{X}^n} \exp(\eta \cdot g(y,x))$. As with ERGMs an MCMC routine is required to sample from the distribution.

ERNMs also have the so called "doubly intractable" property that both the likelihood and the posterior contain intractable normalizing constants. This renders a fully Bayesian approach challenging. As a result we use MCMC sampling with the exchange algorithm analogously to the procedure developed for ERGM. That is instead of sampling from the distribution $p(y, x|\eta)$, we sample from an augmented distribution as following. Going forward we abuse notation and denote (y, x) and simply y, here y are now realizations of random edges and nodal covariates.

$$p(\eta', y', \eta|y_{obs}) \propto p(y_{obs}|\eta)p(\eta'|\eta)p(y'|\nu') \tag{3.32}$$

The idea is to at each step propose a ν' from some proposal distribution q, simulate y' from $p(y'|\eta')$ and then accept η' with probability:

$$\alpha = min(1, \frac{p(\eta')p(y_{obs}|\eta')q(\eta|\eta')p(y'|\eta)}{p(\eta)p(y_{obs}|\eta)q(\eta'|\eta)p(y'|\eta')}) \tag{3.33}$$

We note that simulating y' from $p(y'|\eta')$ is not simple, as sampling from an ERNM itself requires an MCMC procedure, thus the exchange algorithm for ERNM is computationally demanding.

We note that often the challenge with achieving plausible convergence to the stationary distribution with this method has been framed as a difficulty in choosing a "good" the proposal distribution $q(\eta'|\eta)$ (Caimo and Friel, 2011). Here "good" in this case relates to the number of steps before the claim of reaching the stationary distribution is made, though of course in infinite time generic proposal step will theoretically lead to convergence. Indeed naive approaches to adapting the exchange algorithm to ERGMs result in inability to achieve so called "burn-in." Poor proposals can be viewed as proposals that produce extreme values of $\frac{p(y'|\eta)}{p(y'|\eta')}$ either extremely large, or close to 0.

We believe this is often observed as a consequence of the non obvious geometry of the parameter space, coupled with the degenerate nature of ERNMs outside areas of high probability. Outside plausible areas the models produce pathological graphs, for example full or empty graphs, and therefore such proposals are rarely accepted. However if the sampler does ever end up or even start in these locations, it often rarely escapes in a practical time. Intuitively we imagine this as the shape of the posterior parameter space being like a knife edge ridge, to explore we need to move back and forth along the ridge. The procedure will rarely accept a proposal falling far of the ridge, but if we get close to the edge, the sampler can easily fall off and then rarely be able to get back on to the ridge. Sampling naively, is like trying to explore the ridge without any knowledge of how wide it is, inevitably the intrepid explorer falls off and is lost for ever in low posterior probability land.

The theoretically natural approach to accounting for this geometry is to enforce a strict prior to this effect. The problem is specifying such a prior is itself intractable. Some efforts have been made to specify better priors for this problem (Wang, 2011), using conjugate priors, or so called non-degeneracy priors. However these approaches by themselves have not yet yielded a practical sampling methodology.

With the absence of the ability to specify better priors, there have been efforts to choose

the proposal distribution, so that practical burn in can be achieved. The current state of the art as proposed in Wang (2011) and Caimo and Friel (2011), use adaptive techniques to learn the geometry. However we have not achieved efficient sampling with these methods in networks in the with one-hundred or more nodes with complex models and 10 or more parameters.

We found it to be very helpful to let $q(\eta'|\eta)$ to be the multivariate Gaussian centred at η with covariance matrix Σ set to be a scaled inverse Fisher information matrix with tuning parameter α as follows:

$$\Sigma = \alpha \cdot I(\eta)^{-1} \tag{3.34}$$

$$= \alpha \cdot \left(-\mathbb{E}\left[\left(\frac{\partial}{\partial \eta} log(p(y|\eta)) \right) |\eta \right] \right)^{-1} \tag{3.35}$$

$$= \alpha \cdot \left(-\mathbb{E}\left[(g(y) - \mathbb{E}\left[g(y)\right])^{\top} (g(y) - \mathbb{E}\left[g(y)\right]) \right] \right)^{-1} \tag{3.36}$$

$$\tag{3.37}$$

We can estimate this as the sample version, utilizing the simulations $\{y_i\}_{i=1}^{m}$ generated from the MCMC, which we already had to run to the stationary distribution to generate y', so there is little additional computational cost.

We do not explore principled methods of tuning α (Hummel et al., 2012). Typically we found an α between 0.1 and 0.5 to facilitate fast exploration of the space. We found that this proposal distribution allowed us to sample from the posterior ERNM and ERGM distributions for our 869 node example in the order of minutes.

3.14 Goodness of Fit for National Longitudinal Study of Adolescent Health High School Smokers

In this section we present a goodness of fit analysis for the real data example of smoking behaviour in adolescents in a high school.

In social network analyses usually only one network is observed, we use Bayesian posterior prediction goodness of fit graphics in the style of Hunter et al. (2008), where simulated distributions of network summary statistics are compared to the observed values. The simulations are derived from the posterior, that is new networks are simulated by first drawing η from the posterior distribution sample and then simulating a network with that η. The choice of the statistics of interest is subjective, in the social networks literature degree, edgewise shared partner, and geodesic distance distributions are often considered.

Figures 3.7, 3.8 and 3.9 compare the fits of the ERNM and ERGM models, for the degree, ESP and geodesic distance distributions respectively. In the case of the logistic regression and MRF models, we do not assess the goodness of fit of summary statistics involving edges only, as the edges are not considered random. We note that both models do not fit very well on ESP, but do well on the degree distribution. In our experience the fit on geodesic distance distribution is comparable to fits often observed when fitting these models to social network data.

We note that all these models are in fact highly parsimonious for a complex social process and suffer from degeneracy issues when additional terms are added, thus we expect achieving good fit to be very challenging. It may be the case that tapering (Fellows and Handcock, 2017), which would allow for more realistic terms without degeneracy may be required to obtain a well fitting model.

In the context of our problem, we suspect the particular advantage of ERNM is that for simulated networks the smoker nodes are observed within network sub-structures consistent with the observed network. Figure 3.10 compares the degree distribution of smoker nodes for the ERNM and ERGM models, with the ERNM model capturing this distribution better

than the fixed node ERGM. Figures 3.12 and 3.11 compare the distributions of the proportion of smoker edges and triads in the networks simulated from each model. We consider the proportion of smoker triads, as all models' simulations underestimate the absolute number of triads, as expected when considering the ESP goodness of fit in Figure 3.8. We note that the ERNM model and the full MRF fits considerably better that the pseudo MRF and the logistic regression model.

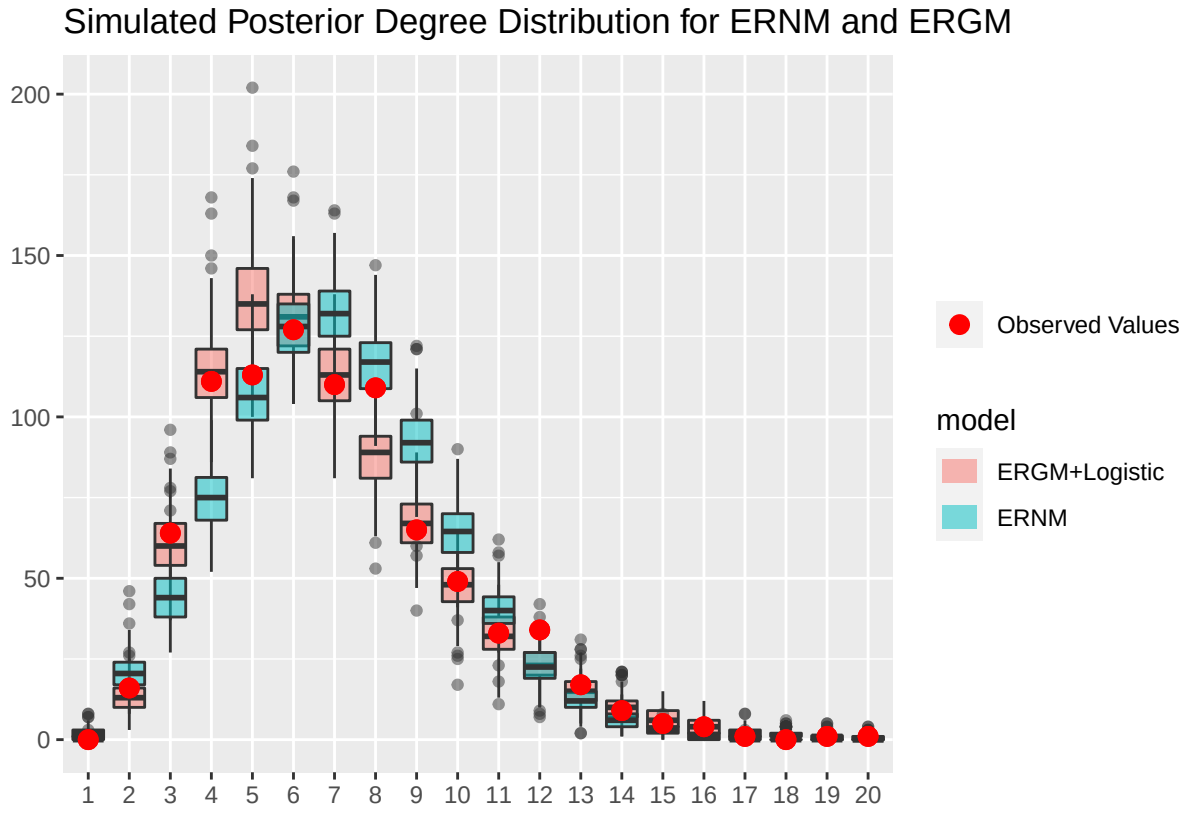


Figure 3.7: ERNM and ERGM simulated posterior degree distribution

3.15 MCMC Convergence Checks for National Longitudinal Study of Adolescent Health High School Smokers Network

We examine the trace plots of the MCMC procedure. Figures 3.13, 3.14 and 3.15 show the trace plots for the edge, GWESP and smoker neighbours terms respectively. We can see that

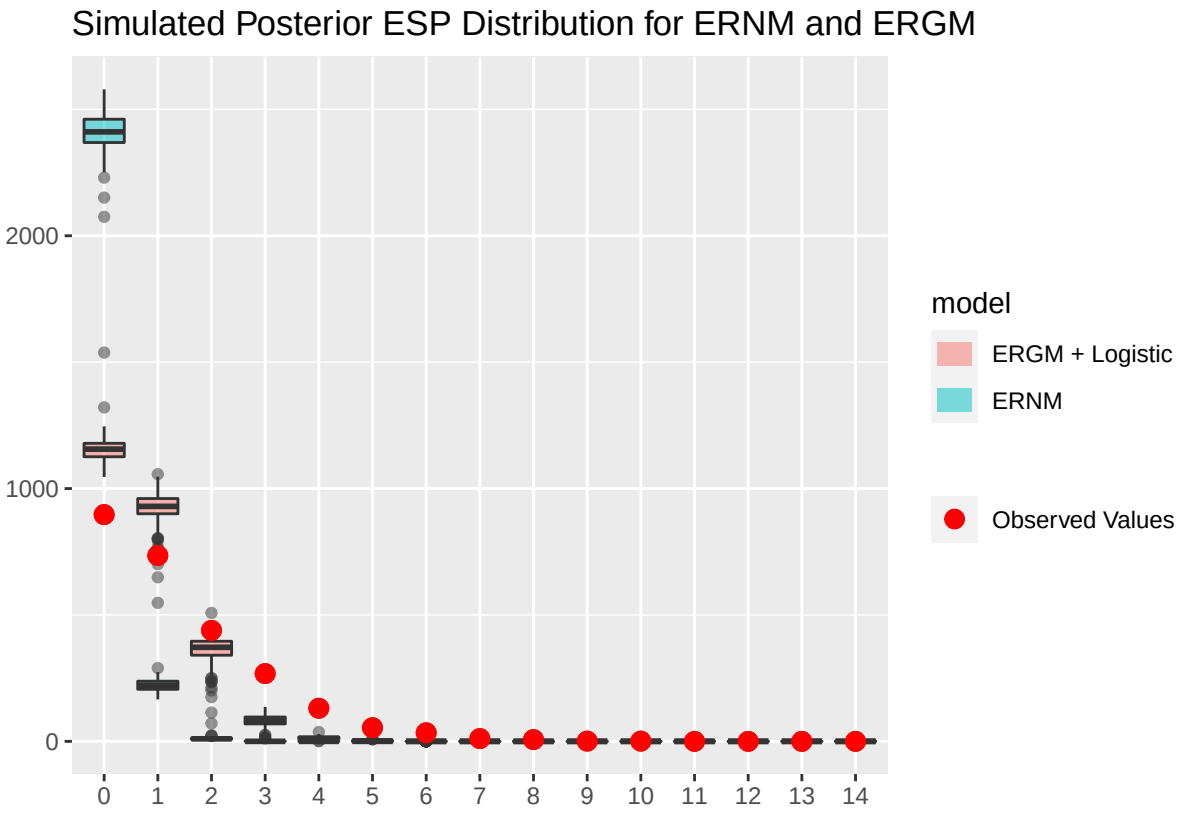


Figure 3.8: ERNM and ERGM simulated posterior ESP distribution

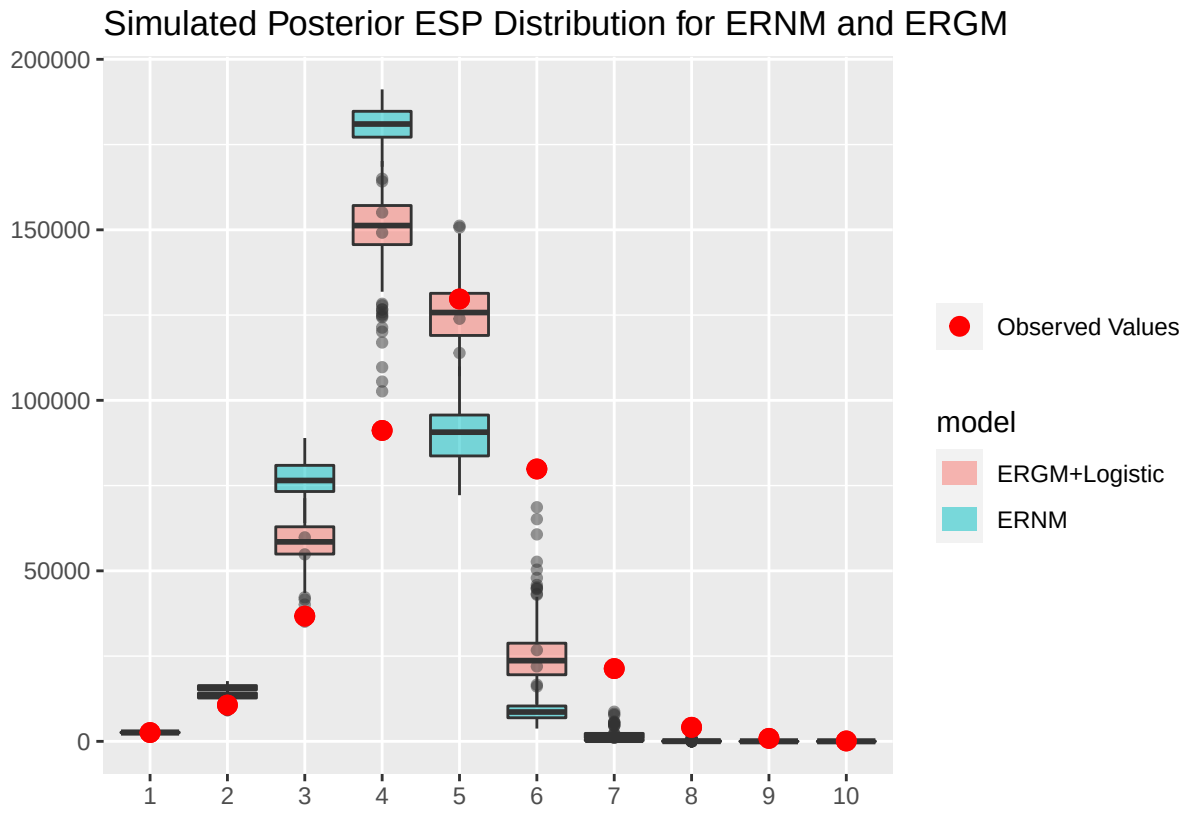


Figure 3.9: ERNM and ERGM simulated posterior geodesic distance distribution

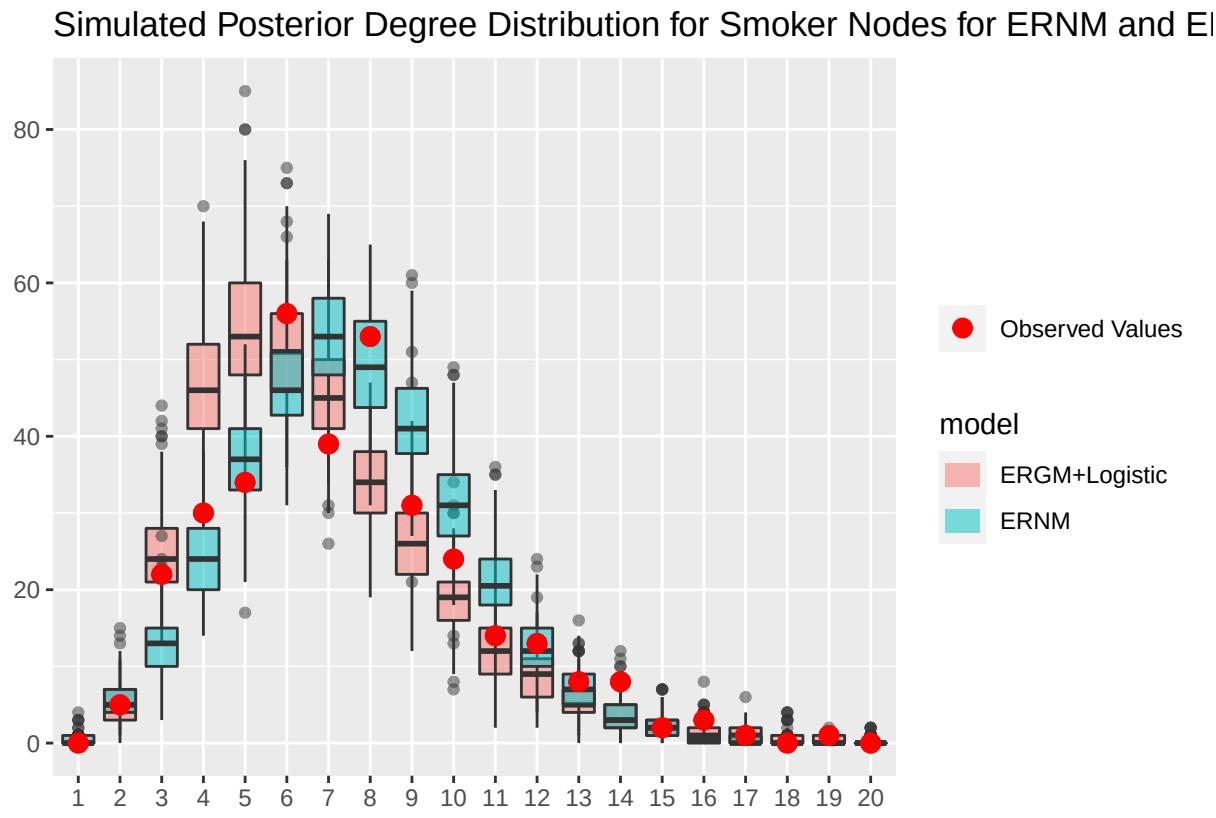


Figure 3.10: ERNM and ERGM simulated posterior degree distribution for smoker nodes only

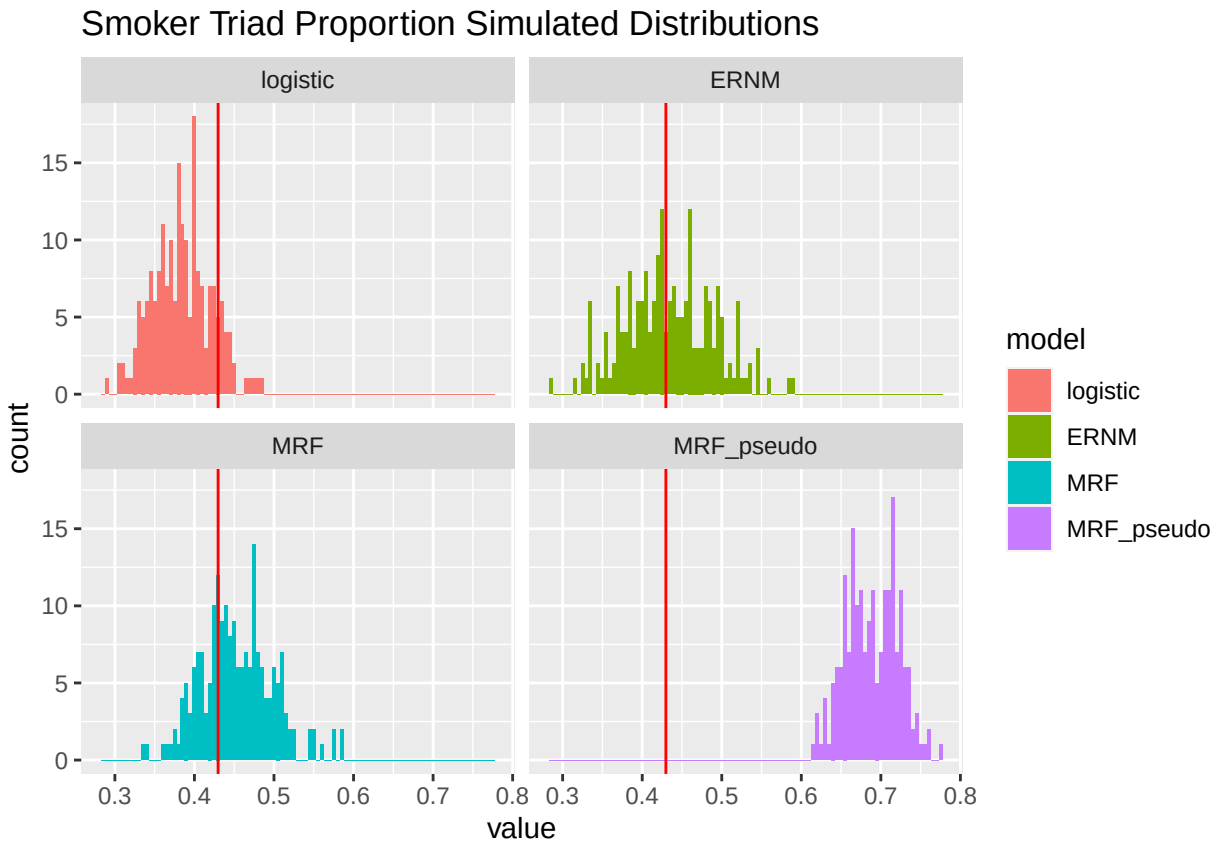


Figure 3.11: Distribution of proportion of smoker triads

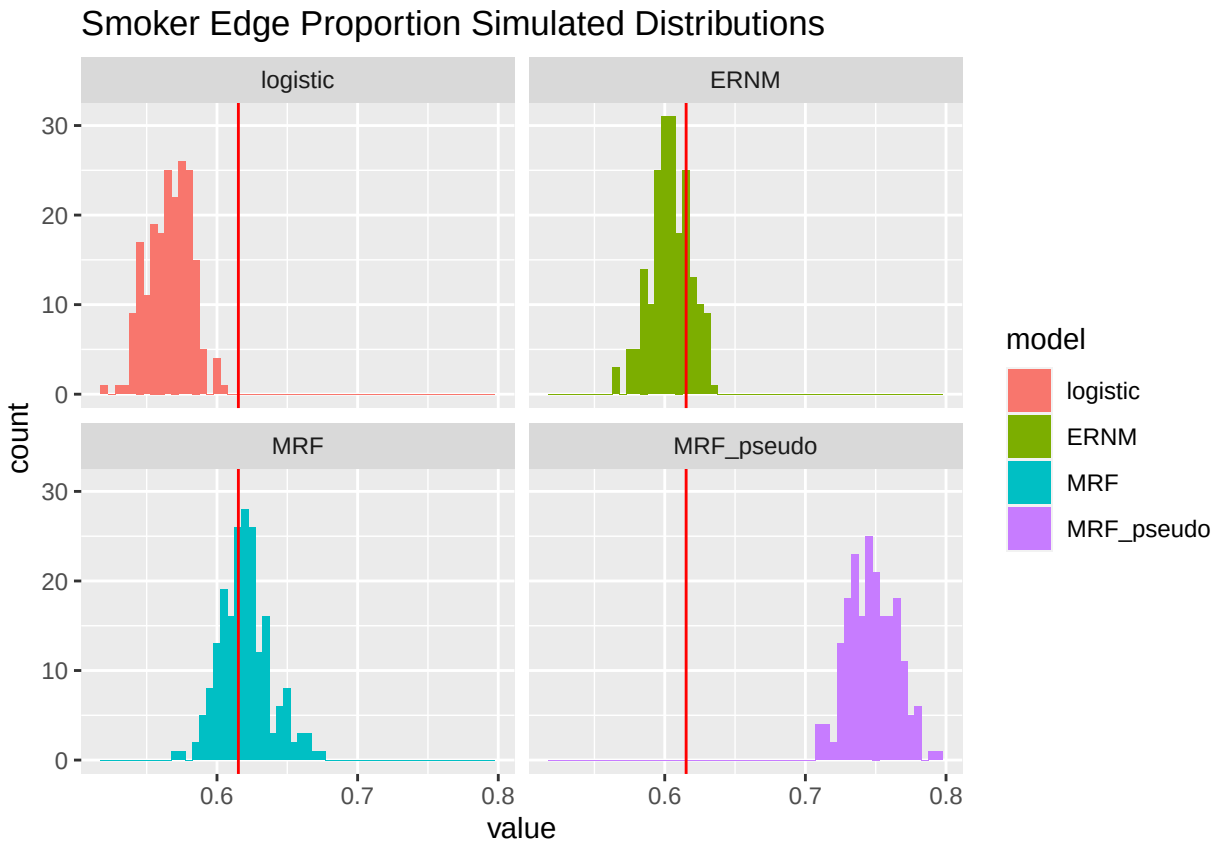


Figure 3.12: Distribution of proportion smoker edges

after around 2500 proposals for these parameters the chains appear to vary independently of their starting point, suggesting the convergence of the MCMC procedure. We show only three parameters for brevity.

We also follow the approach in Gelman and Rubin (1992) and consider the ratio of the between chain variance to the within chain variance for 8 MCMCs. A ratio of close to 1 suggests that the chain has converged to its stationary distribution. Figures 3.16, 3.17 and 3.18 show the ratio plots for the edge, GWESP and smoker neighbours terms respectively.

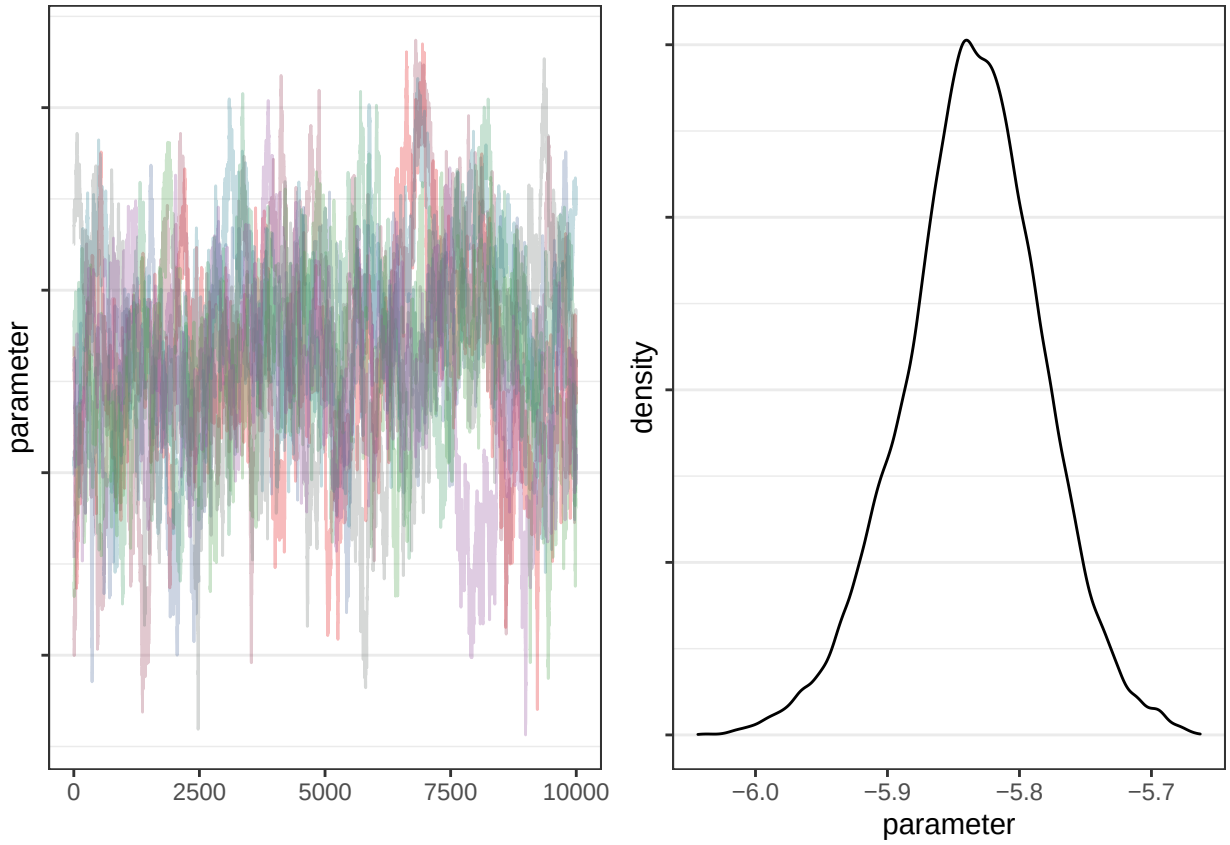


Figure 3.13: Trace plot for edge parameter

3.16 Gender Peer Effects for Addolescent Health Network

We consider the k-peer-gender effect on the ego smoking of having k additional friends has over having no friends of specified gender. We note that the causal effect of gender

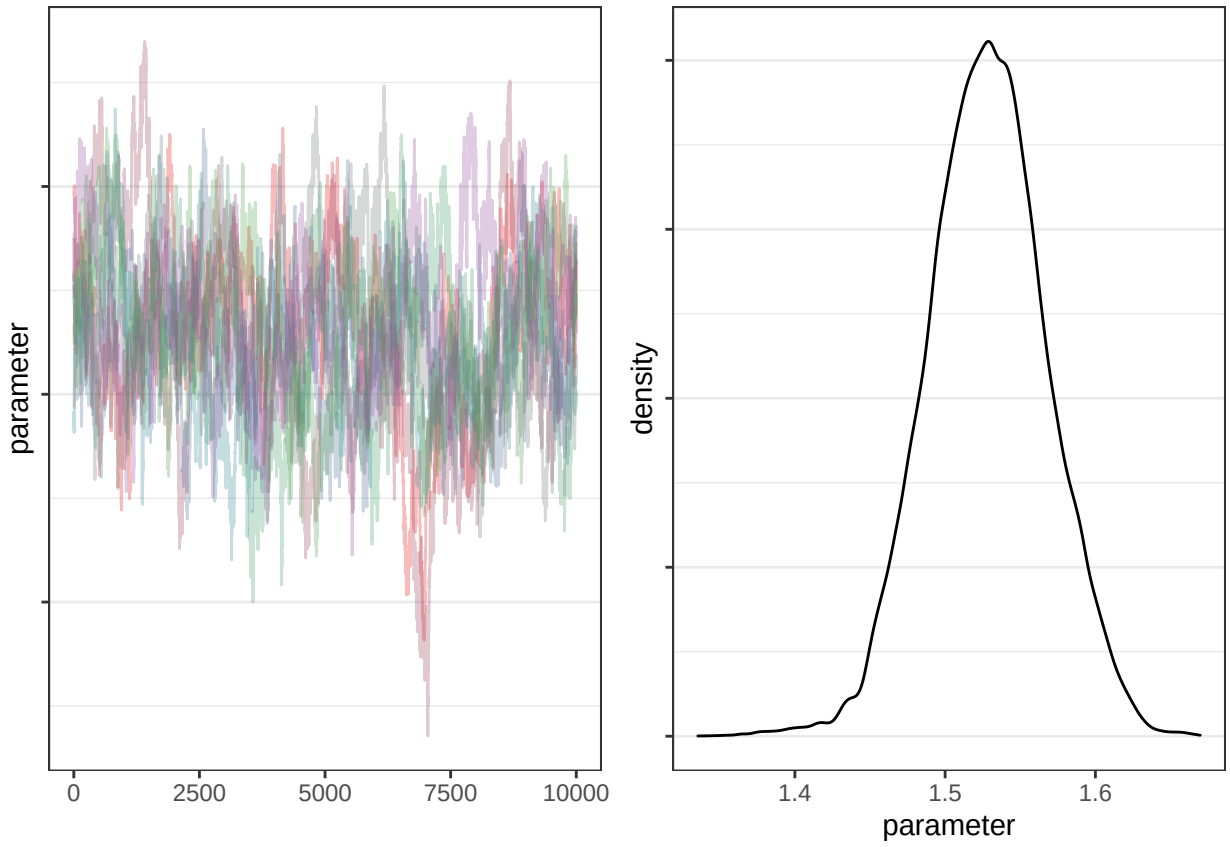


Figure 3.14: Trace plot for GWESP parameter

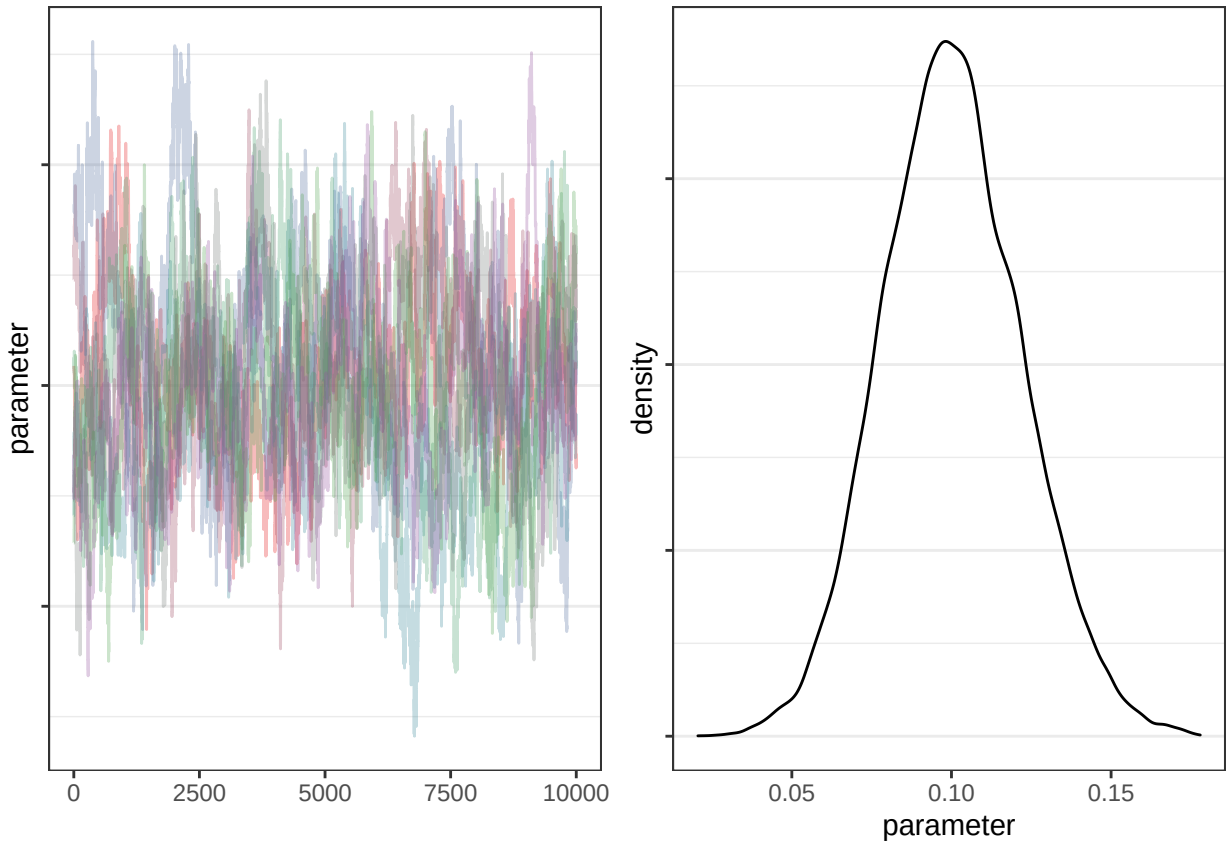


Figure 3.15: Trace plot for smoker neighbor parameter

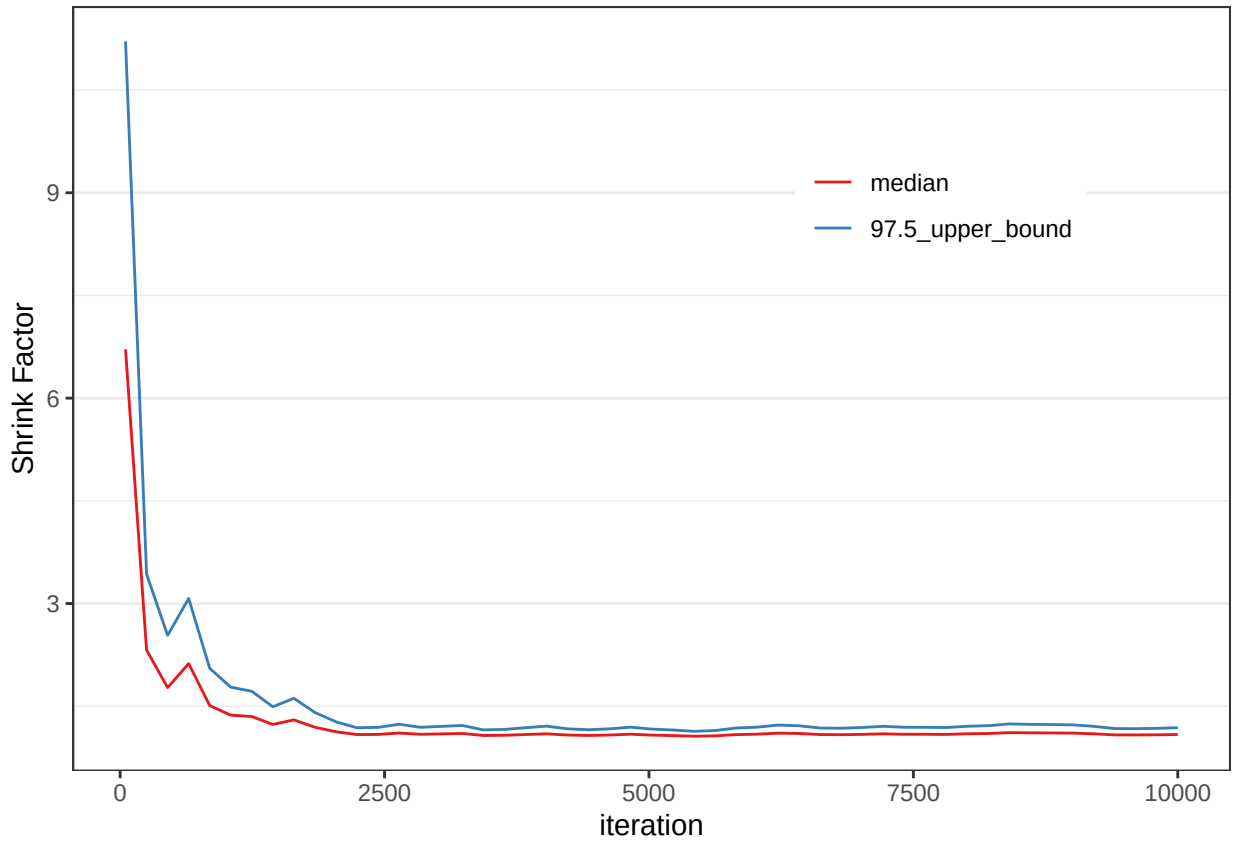


Figure 3.16: Plot of Gelman diagnostics for edge parameter

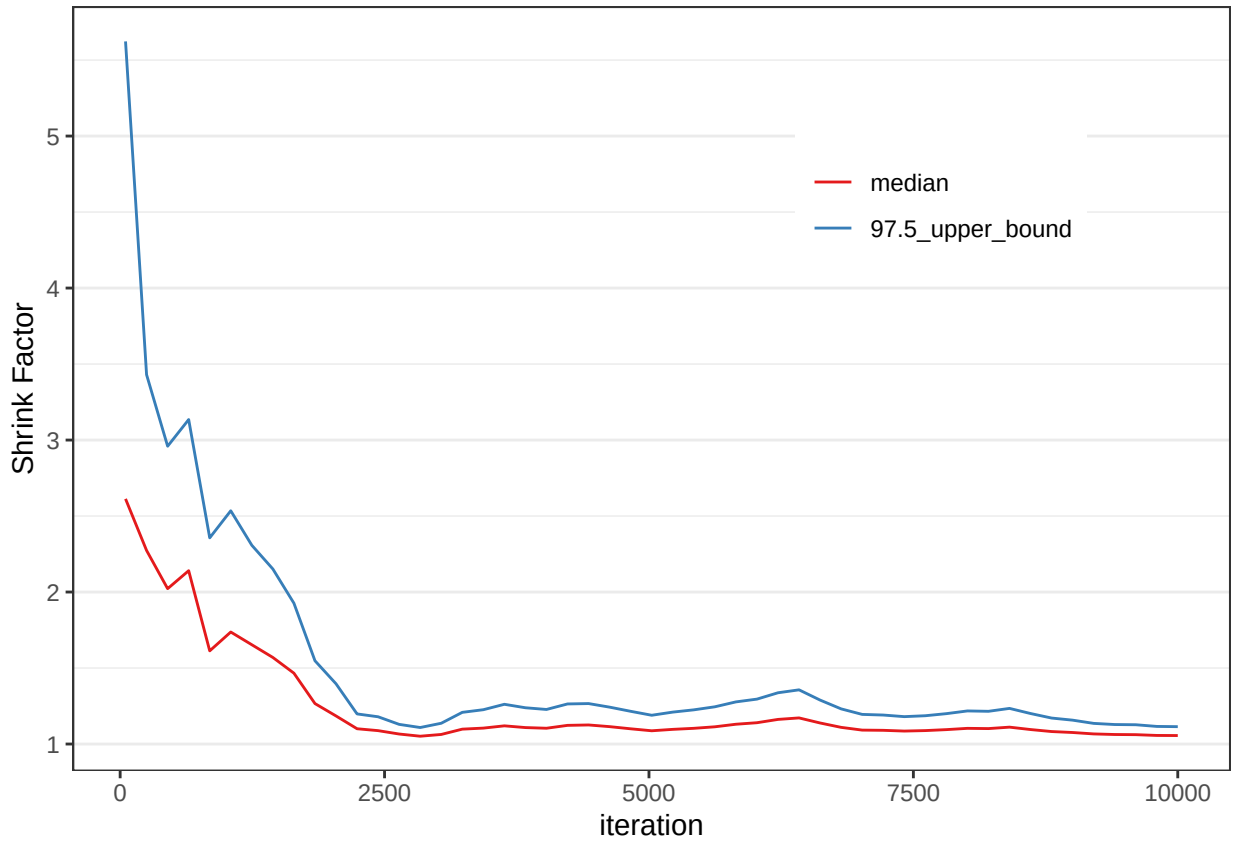


Figure 3.17: Plot of Gelman diagnostics for GWESP parameter

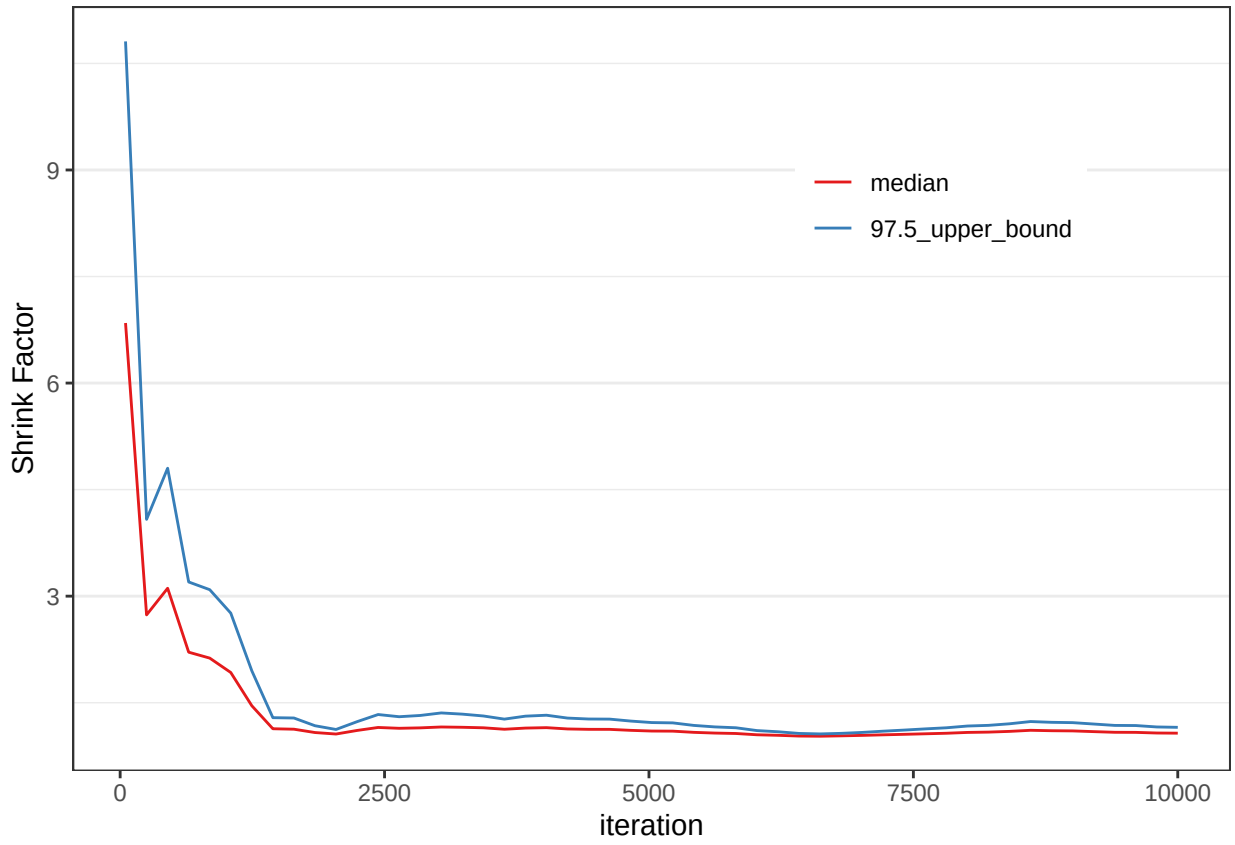


Figure 3.18: Plot of Gelman diagnostics for smoker neighbors parameter

is controversial to interpret, however considering say the k-female-peer effect, may prove insightful, as one of the consequences of being female is a significantly higher liklihood of having many female friends.

Figure 3.19 shows the estimated effects, where the gender of the nodes and gender of the neighbours considered is varied. We concentrate on the female neighbours for female nodes, and male neighbours for male nodes, since there is strong homophily on gender, these outcomes are most likely to occur in the social process that generated the network. Interventions that set a node to have more friends of the opposite gender are likely to need to be extreme to counteract the overall social forces, inherent in the network formation process.

We note that little difference is shown between the effect of male neighbours on male nodes and the effect of female neighbours on female nodes. We suggest that this shows the effect of more neighbours in general is the driving force behind increased likelihood of smoking adoption among nodes with many friendships. Figure 3.20 shows the is little difference in the marginal increase in effect from each additional friend under all models.

Clearly the logistic regression model performs poorly here, suggesting that more neighbours has a negative effect on smoking. This is due to the confounding of the number of female neighbours, with the number of smoker neighbours, which results in a negative estimate for this coefficient. We suggest that this could be mitigated with a different specification, allowing for interactions, however we include this model as precautionary tale against using naive models for network data. The ERNM and MRF are statistically indistinguishable, whilst the pseudo MRF also seems similar.

However as the raw number of friendships a female node has is highly correlated with the number of female friendships a node has, we caution interpreting the effect, as that of gendered friends, rather that of friends in general. In particular Figure 3.21 shows the effect of the proportion of genders an egos friends are for the ERNM model. The proportion of an egos friends that are homogeneous on gender does not seem to greatly effect the likelihood of smoking. We note the top right panes on this have lower overall smoking levels as observed for male nodes in the data. This shows that there is limited evidence that the gender of a

nodes friends has an effect on the likelihood of smoking.

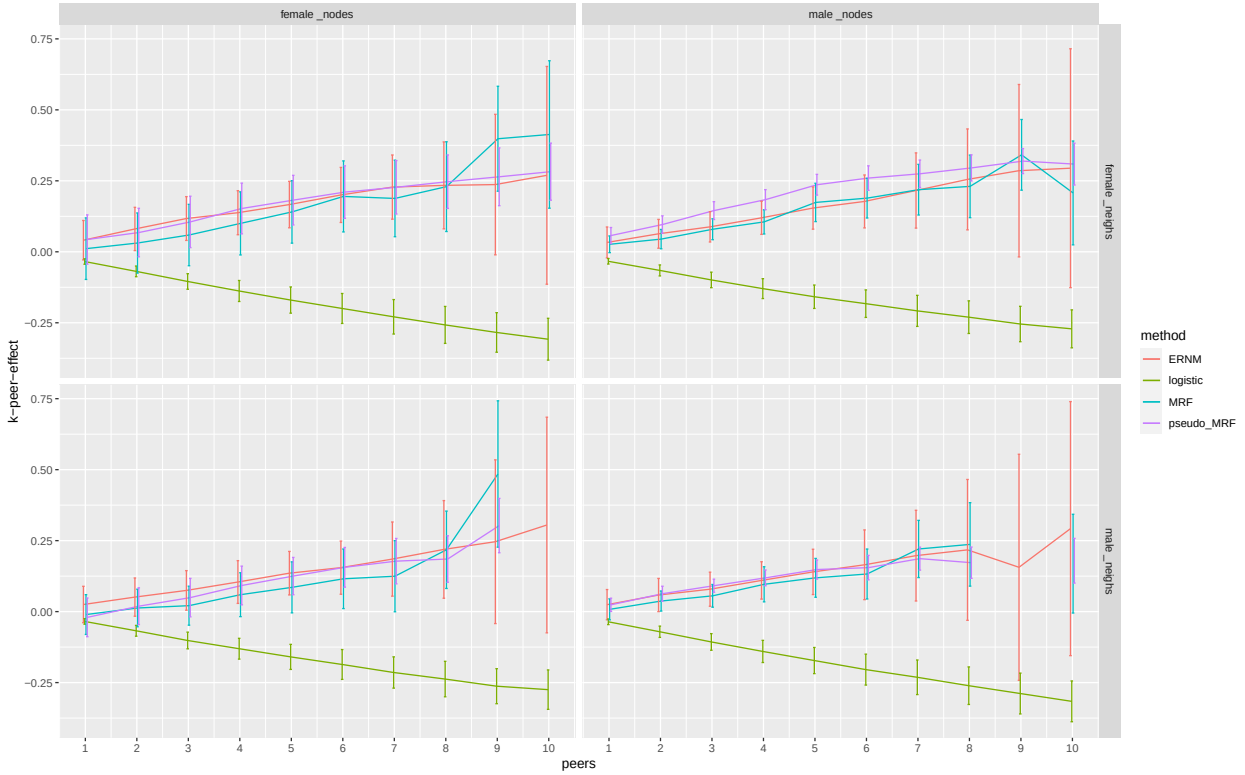


Figure 3.19: Plot of k-gender effects for Addolescent Health network, split by node gender and neighbour gender

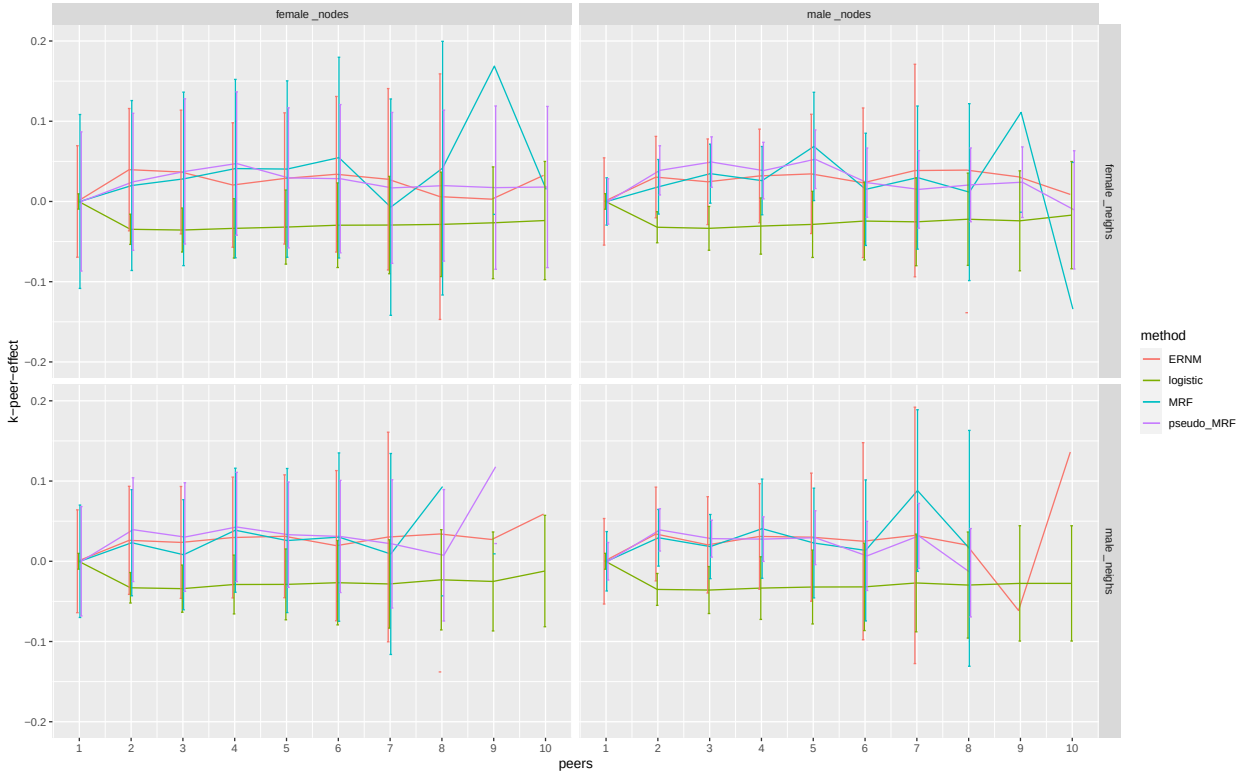


Figure 3.20: Plot of k-gender marginal effects for Addolescent Health network, split by node gender and neighbour gender

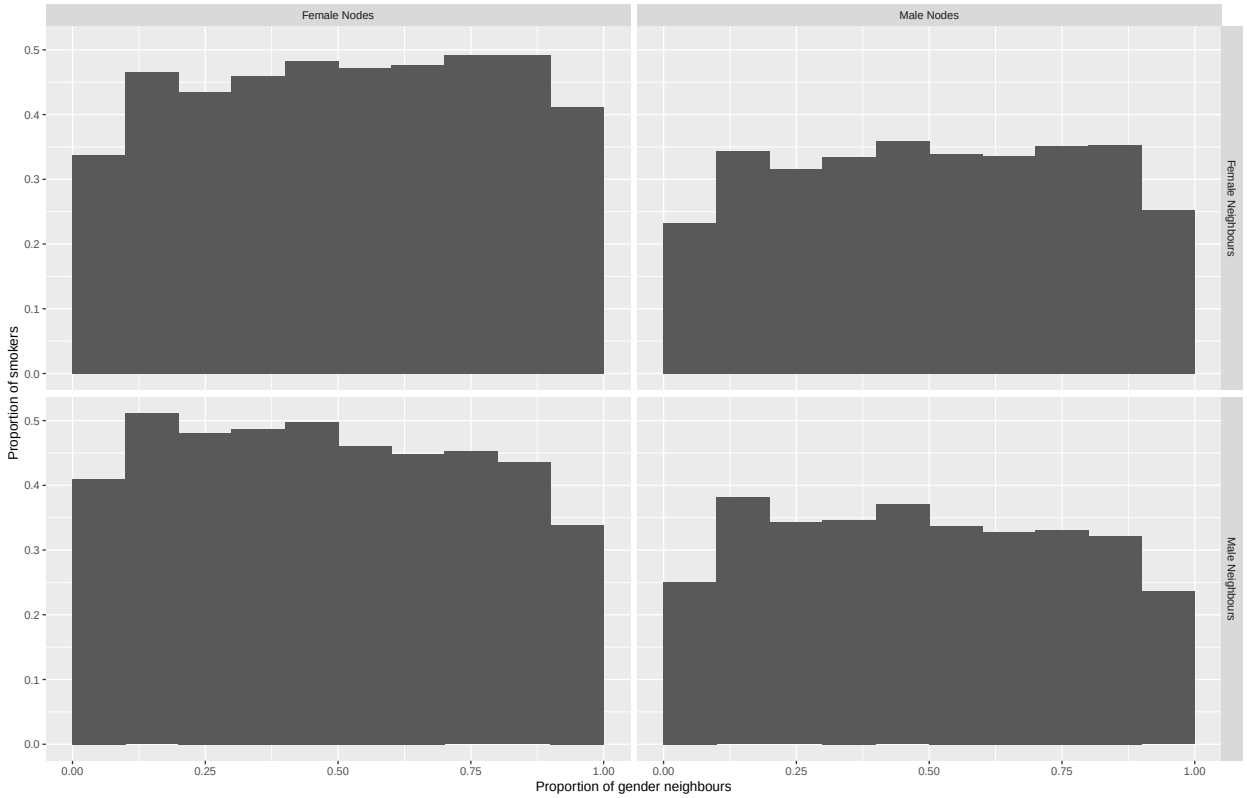


Figure 3.21: ERNM model barplot of mean smoker status, for each proportion of neighsbours of each gender, for each ego gender. We see that the proportion of neighbors of either particular gender does notimpact the likelihood of smoking.

BIBLIOGRAPHY

Ackland, R. and M. O'Neil (2011). Online collective identity: The case of the environmental movement. *Social Networks 33*(3), 177 – 190.

Amati, V., A. Lomi, and A. Mira (2018). Social network modeling. *Annual Review of Statistics and Its Application 5*(1), 343–369.

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

Aronow, P. M. and C. Samii (2017, December). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat. 11*(4), 1912–1947.

Barndorff-Nielsen, O. (1978). *Information and exponential families: in statistical theory*. Wiley series in probability and mathematical statistics: Tracts on probability and statistics. Wiley.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological) 36*(2), 192–225.

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

Caimo, A. and N. Friel (2013, 01). Bayesian model selection for exponential random graph models. *Social Networks 35*, 11–24.

Caimo, A. and N. Friel (2014). Bergm: Bayesian exponential random graphs in r. *Journal of Statistical Software 61*(2), 1–25.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician 39*(2), 83–87.

Chatterjee, S. and P. Diaconis (2013, 10). Estimating and understanding exponential random graph models. *Ann. Statist. 41*(5), 2428–2461.

Chen, M.-H. and J. G. Ibrahim (2003). Conjugate priors for generalized linear models. *Statistica Sinica 13*(2), 461–476.

Christakis, N. A. and J. H. Fowler (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine 357*(4), 370–379. PMID: 17652652.

Christakis, N. A. and J. H. Fowler (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine 358*(21), 2249–2258. PMID: 18499567.

Christakis, N. A. and J. H. Fowler (2010, September). Social network sensors for early detection of contagious outbreaks. *PLOS ONE 5*(9), 1–8.

DeAmour, A. (2016). *Misspecification, Sparsity, and Superpopulation Inference for Sparse Social Networks*. Ph. D. thesis, Harvard University.

Diaconis, P. and D. Ylvisaker (1979). Conjugate priors for exponential families. *The Annals of Statistics 7*(2), 269–281.

Doreian, P. and N. Conti (2012). Social context, spatial structure and social network structure. *Social Networks 34*, 32–46.

Duane, S., A. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics Letters B 195*(2), 216 – 222.

Duijn, M., K. Gile, and M. Handcock (2009, 01). A framework for the comparison of maximum pseudo likelihood and maximum likelihood estimation of exponential family random graph models. *Social networks 31*, 52–62.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 1–26.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Everett, M. and T. Valente (2020). *Social Networks: An International Journal of Structural Analysis*. Elsevier.

Fellows, I. (2018a). Latent order logistic (lolog) graph models. https://github.com/statnet/lolog.

Fellows, I. and M. Handcock (2017, 20–22 Apr). Removing Phase Transitions from Gibbs Measures. In A. Singh and J. Zhu (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Volume 54 of *Proceedings of Machine Learning Research*, Fort Lauderdale, FL, USA, pp. 289–297. PMLR.

Fellows, I. and M. S. Handcock (2012). Exponential-family random network models.

Fellows, I. E. (2018b, April). A new generative statistical model for graphs: The latent order logistic (lolog) model.

Fischer, M. and P. Sciarini (2015). Unpacking reputational power: Intended and unintended determinants of theassessment of actors' power. *Social Networks 42*, 60–71.

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

Frydenberg, M. (1990). The chain graph markov property. *Scandinavian Journal of Statistics 17*(4), 333–353.

Gelman, A. and D. B. Rubin (1992, 11). Inference from iterative simulation using multiple sequences. *Statist. Sci. 7*(4), 457–472.

Gile, K. and M. Handcock (2016, September). Analysis of networks with missing data with application to the national longitudinal study of adolescent health. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 66*.

Girolami, M. and B. Calderhead (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(2), 123–214.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning 2*(2), 129–233.

Goodreau, S. (2007, 06). Advances in exponential random graph (p*) models applied to large social networks. *Social networks 29*, 231–248.

Goodreau, S. M., J. Kitts, and M. Morris (2009). Birds of a feather, or friend of a friend? Using statistical network analysis to investigate adolescent social networks. *Demography 46*, 103–125.

Handcock, M. S. (2002). Degeneracy and inference for social network models. In *Paper presented at the Sunbelt XXII International Social Network Conference in New Orleans, LA*.

Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. Working paper #39, University of Washington.

Handcock, M. S. (2015). *Relative Distribution Methods*. Los Angeles, CA. Version 1.6-4. Project home page at urlhttp://www.stat.ucla.edu/ handcock/RelDist.

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, P. N. Krivitsky, and M. Morris (2018). *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (http://www.statnet.org). R package version 3.9.4.

Handcock, M. S. and M. Morris (1999). *Relative Distribution Methods in the Social Sciences*. New York: Springer. ISBN 0-387-98778-9.

Hanneke, S., W. Fu, and E. P. Xing (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics 4*(none), 585 – 605.

Harris, K., C. Halpern, A. Smolen, and B. Haberstick (2007a, January). *Addolescent Health Survey*. Available online at: http://moreno.ss.uci.edu/data.html#adhealth, last accessed: 2019-12-06.

Harris, K., C. Halpern, A. Smolen, and B. Haberstick (2007b, 01). The national longitudinal study of adolescent health (add health) twin data. *Twin research and human genetics : the official journal of the International Society for Twin Studies 9*, 988–97.

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

Heidler, R., M. Gampner, A. Herz, and F. Esser (2014). Relationship patterns in the 19th century: The friendship network in a german boys' school class from 1880 to 1881 revisited. *Social Networks 37*, 1–13.

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

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842. PMID: 19081744.

Hummel, R. M., D. R. Hunter, and M. S. Handcock (2012). Improving simulation-based algorithms for fitting ergms. *Journal of Computational and Graphical Statistics 21*(4), 920–939.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks 29*, 216–230.

Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2008). Goodness of fit of social network models. *Journal of the American Statistical Association 103*(481), 248–258.

Hunter, D. R. and M. S. Handcock (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics 15*(3), 565–583.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 186*(1007), 453–461.

Kao, E. (2017). *Causal Inference Under Network Interference: A Framework for Experiments on Social Networks*. Ph. D. thesis, Harvard University.

Krapivsky, P. L., S. Redner, and E. Ben-Naim (2010). *Complex networks*, pp. 441–470. Cambridge University Press.

Krivitsky, P. and M. Handcock (2014, 01). A separable model for dynamic networks. *Journal of the Royal Statistical Society. Series B, Statistical methodology 76*, 29–46.

Krivitsky, P. N., M. S. Handcock, and M. Morris (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology 8*(4), 319–339.

Lauritzen, S. L. and T. S. Richardson (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 321–348.

Little, R. (2011). Calibrated Bayes, for Statistics in General, and Missing Data in Particular. *Statistical Science 26*(2), 162 – 174.

Lubbers, M. J. and T. A. Snijders (2007). A comparison of various approaches to the exponential random graph model: A reanalysis of 102 student networks in school classes. *Social Networks 29*(4), 489 – 507.

Ma, Y., Y.-A. Ma, T. Chen, and E. B. Fox (2015). A complete recipe for stochastic gradient mcmc. In *NIPS*.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2004, 01). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America 100*, 15324–8.

McLevey, J., A. V. Graham, R. McIlroy-Young, P. Browne, and K. S. Plaisance (2018). Interdisciplinarity and insularity in the diffusion of knowledge: an analysis of disciplinary boundaries between philosophy of science and the sciences. *Scientometrics 117*(1), 331–349.

Morris, M., M. S. Handcock, and D. R. Hunter (2008). Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software 24*(4).

Murray, I., Z. Ghahramani, and D. J. C. MacKay (2006). Mcmc for doubly-intractable distributions. In *UAI*.

Møller, J., A. N. Pettitt, R. Reeves, and K. K. Berthelsen (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika 93*(2), 451–458.

Ogburn, E. L., I. Shpitser, and Y. Lee (2020, 07). Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 183*(4), 1659–1676.

Ogburn, E. L., O. Sofrygin, I. Diaz, and M. J. van der Laan (2020, February). Causal inference for social network data.

Ogburn, E. L. and T. J. VanderWeele (2014, November). Causal diagrams for interference. *Statist. Sci. 29*(4), 559–578.

O'Hagan, A. and M. Kendall (1993). *Bayesian Inference*. Kendall's advanced theory of statistics. Arnold.

Pauksztat, B., C. Steglich, and R. Wittek (2011). Who speaks up to whom? a relational approach to employee voice. *Social Networks 33*(4), 303–316.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*(4), 669–688.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). USA: Cambridge University Press.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richardson, T. S. and J. M. Robins (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128*(30), 2013.

Robins, G., T. Snijders, P. Wang, M. Handcock, and P. Pattison (2007). Recent developments in exponential random graph (p) models for social networks. *Social Networks 29*(2), 192–215.

S. Kraatz, M., N. Shah, and E. Lazega (2003, 09). The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership. *Administrative Science Quarterly 48*, 525.

Sailer, K. and I. McCulloch (2012). Social networks and spatial configuration—how office layouts drive social interaction. *Social Networks 34*, 47–58.

Scarselli, F., M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini (2009). The graph neural network model. *IEEE transactions on neural networks 20*(1).

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association 106*(496), 1361–1370. PMID: 22844170.

Schweinberger, M. and M. S. Handcock (2015). Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(3), 647–676.

Schweinberger, M., P. N. Krivitsky, C. Butts, and J. Stewart (2020). Exponential-family models of random graphs: Inference in finite-, super-, and infinite population scenarios. *Statistical Science*.

Schweinberger, M. and J. Stewart (2020, 02). Concentration and consistency results for canonical and curved exponential-family models of random graphs. *Ann. Statist. 48*(1), 374–396.

Shalizi, C. and A. Thomas (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological methods and research 40*, 211–239.

Shpitser, I., E. T. Tchetgen, and R. Andrews (2021). Modeling interference via symmetric treatment decomposition.

Sisson, S., Y. Fan, and M. Beaumont (2018). *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Taylor & Francis Group.

Snijders, T. (2002, 06). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure 3*.

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

Snijders, T. A. B., P. E. Pattison, G. L. Robins, and M. S. Handcock (2006). New specifications for exponential random graph models. *Sociological Methodology 36*(1), 99–153.

Sofrygin, O. and M. J. van der Laan (2017). Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population. *Journal of Causal Inference 5*(1), 20160003.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search* (2nd ed.). MIT press.

Stivala, A., G. Robins, and A. Lomi (2020, 01). Exponential random graph model parameter estimation for very large directed networks. *PLOS ONE 15*(1), 1–21.

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

Tchetgen, E. J. T., I. R. Fulcher, and I. Shpitser (2021). Auto-g-computation of causal effects on a network. *Journal of the American Statistical Association 116*(534), 833–844.

Toivonen, R., L. Kovanen, M. Kivelä, J.-P. Onnela, J. Saramäki, and K. Kaski (2009). A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks 31*(4), 240 – 254.

Toulis, P. and E. Kao (2013). Estimation of causal peer influence effects. In S. Dasgupta and D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, Volume 28 of *Proceedings of Machine Learning Research*, Atlanta, Georgia, USA, pp. 1489–1497. PMLR.

van der Laan, M. J. (2014). Causal inference for a population of causally connected units. *Journal of Causal Inference 2*(1), 13–74.

Wang, R. (2011). Likelihood-based inference of exponential-family random graph models for social networks. Dissertation.

Wong, L. H. H., A. Gygax, and P. Wang (2015). Board interlocking network and the design of executive compensation packages. *Social Networks 41*, 85–100.

Ye, Q., A. A. Amini, and Q. Zhou (2022). Distributed learning of generalized linear causal networks. *arXiv Preprint* (2201.09194).

Yin, F. and C. T. Butts (2020). Kernel-based approximate bayesian inference for exponential family random graph models.