Bayesian Analysis of Network Data
–
Model Selection and Evaluation of the Exponential Random Graph Model

Dissertation

Presented to the Faculty for
Social Sciences, Economics, and Business Administration
at the
Otto-Friedrich-University of Bamberg
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR RERUM POLITICARUM

by
Julian Körber, Master of Science in Survey Statistics
born July 3, 1984 in Forchheim, Germany

Bamberg
2018

Examination Committee:

Principal advisor:	Professor Dr. Susanne Rässler *University of Bamberg, Germany*
Reviewers:	Professor Dr. Kai Fischbach *University of Bamberg, Germany*
	Professor Dr. Lasse Gerrits *University of Bamberg, Germany*
Date of submission:	November 27, 2017
Date of defence:	March 29, 2018

URN: urn:nbn:de:bvb:473-opus4-516206

DOI: https://doi.org/10.20378/irbo-51620

Abstract

The exponential random graph model (ERGM) is a class of stochastic models for network data widely applied in statistical social network analysis. The ERGM can be used to model a wide range of social processes. However, it is generally difficult to estimate due to its intractable normalizing constant. Markov chain Monte Carlo maximum likelihood (MCMC-ML) ERGM estimation is available but tends to be numerically unstable due to model degeneracy of particular specifications. Bayesian ERGM estimation is robust to model degeneracy and is a practical alternative to the MCMC-ML approach.

Bayesian model selection is based on the Bayes factor which is the ratio of marginal likelihoods of concurring models. The research aim of this thesis is to estimate the marginal likelihood of the ERGM class using path sampling which is also called thermodynamic integration. Power posterior sampling is a discretized version of thermodynamic integration using a fixed path of tempering steps to transition from the prior distribution to the posterior distribution of interest. In this thesis, power posterior sampling is used both to integrate over the parameter space of the ERGM posterior distribution of interest and to yield an estimate of the respective intractable ERGM normalizing constant. Existing approaches of estimating the ERGM marginal likelihood rely on a non-parametric density approximation or a Laplace approximation. The proposed power posterior exchange algorithm with explicit evaluation of the likelihood (PPEA-EEL) does not require such approximations and yields a valid estimate of the ERGM marginal likelihood. As the PPEA-EEL is a computationally expensive approach involving many MCMC samples, new graphical methods to evaluate power posterior samples are developed.

In this thesis a brief introduction to random graphs and network dependencies is given. The ERGM class is discussed and various dependency assumptions are illustrated. MCMC-ML ERGM estimation is applied to policy networks in Ghana, Senegal and Uganda. Bayesian ERGM estimation and Bayesian model selection are discussed. An overview is given on methods of estimating the marginal likelihood originating from importance sampling, namely bridge sampling, path sampling and power posterior sampling. The PPEA-EEL is applied to social network data and the numerical stability of the approach is evaluated.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Networks represent relations between entities. These entities are called nodes and the relations are called edges. Depending on the type of nodes and the type of edges considered, a variety of networks can be defined. Social networks indicate the relations between social actors and represent social processes. Such social processes are e.g. the communication of political actors, relations within an organization or friendship within a class room. This thesis is restricted to human social networks where social actors are human individuals or collectives of individuals like organizations. Social processes typically generate patterns like reciprocity, transitivity and hierarchy which are not easy to model statistically. Stochastic models for network data have to capture such patterns of tie variable formation while recognizing the variability that cannot be modeled explicitly. Most stochastic models rely on the assumption of independent and identically distributed observations which would risk incorrect inference in network analysis, while stochastic models for network data explicitly try to formulate the dependence between social actors and network ties. Wasserman and Faust (1994) and Jackson (2008) give a general introduction to statistical social network analysis. Kolaczyk (2009) and Snijders (2011) give an overview on stochastic models for network data.

The exponential random graph model (ERGM) is a class of stochastic models for binary network ties as dependent variables and is by far the most widely used model for statistical social network analysis. It can model a wide range of hypothesized patterns of network tie formation originating from many strands of theories. Network ties can be modeled conditional on endogenous patterns of network self organization and exogenous covariates based on actor attributes which allows for a

high flexibility in social network analysis. The ERGM class can also be applied to networks where the nodes are non-human entities, e.g. technological or biological networks.

1.1 Exponential random graph models

An observed binary network is considered as the realization of a random graph. A random graph is a collection of random binary tie variables on a fixed set of nodes. If all tie variables are assumed to be independent, network tie formation can be modeled with a Bernoulli process. However, this would be a very unrealistic data generating mechanism in most cases, therefore stochastic models are required which can capture realistic patterns of network tie formation. The ERGM class can be used to explicitly model patterns of network dependence such as endogenous network self organization and the influence of exogenous nodal attributes. Patterns of tie formation are represented by fundamental subgraph configurations like triangles or stars and configurations depending on nodal attributes. Counts of such configurations are used as sufficient statistics for the ERGM probability distribution. While the ERGM is not a new model class at all, it did not gain popularity in statistical social network analysis for a long time until severe problems of model estimation were solved.

The most simplistic Bernoulli random graph model dates back to Erdös and Renyi (1959). Besag (1974) show how a random graph model can be formulated in an exponential form using counts of subgraph configurations as sufficient statistics. Holland and Leinhardt (1981) develop the first ERGM specification that is able to represent patterns of reciprocity. The Markov model by Frank and Strauss (1986) is able to capture patterns of transitivity, hierarchy and centrality. The Markov model was a break trough in statistical social network analysis. However, it is very hard to estimate, which prevented its application for two decades. Wasserman and Pattison (1996) introduce a log-linear formulation of the ERGM which facilitates the interpretation and allows for the inclusion of exogenous covariates. A more general dependence assumption formulated by Pattison and Robins (2002) and implemented by Snijders et al. (2006) generated a break through for the ERGM class. Hunter and Handcock (2006) develop Markov Chain Monte Carlo maximum likelihood (MCMC-ML) ERGM estimation based on simulated networks. Bayesian

ERGM estimation is introduced by Koskinen (2008) and Caimo and Friel (2011).

The major drawback of the ERGM class is the difficulty of model estimation which is caused by model degeneracy and the unavailability of the likelihood normalizing constant. The latter requires auxiliary network simulations and renders methods of model estimation computationally expensive. The first can cause convergence issues as it may be impossible to find suitable starting values for the MCMC-ML approach. Bayesian ERGM estimation using the exchange algorithm (EA) by Murray et al. (2006) is robust to model degeneracy. Today, the ERGM class is widely applied and many extensions are available, see the Lusher et al., eds (2012) for an overview.

1.2 Bayesian model selection

In Bayesian statistics the prior distribution $p(\theta)$ represents the assumptions about the population characteristics before data are observed. The likelihood function $p(y|\theta)$ represents the data generating process and the posterior distribution

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{1.1}$$

represents the updated assumptions if data are observed. The normalizing constant $p(y)$ is called the marginal likelihood. In most cases the analytical evaluation of the posterior distribution is not possible and MCMC techniques are required. The Bayesian approach allows for model comparison using the posterior odds of concurring models m_1 and m_2

$$\frac{\Pr(m_1|y)}{\Pr(m_2|y)} = \frac{\Pr(m_1)}{\Pr(m_2)} \times \frac{p(y|m_1)}{p(y|m_2)}. \tag{1.2}$$

The prior odds $\Pr(m_1)/\Pr(m_2)$ are updated with the Bayes factor $p(y|m_1)/p(y|m_2)$ which is a ratio of marginal likelihoods. The marginal likelihood is the normalizing constant of the respective posterior distribution. Typically, the prior odds are assumed to be one so the Bayes factor is the relevant quantity for model comparison. In this role the marginal likelihood is also called the model evidence. Computation of the model evidence is a complicated task and requires techniques of numerical integration for most cases.

Importance sampling can be used to estimate the model evidence, see Geweke (1989), although it may be very difficult to specify an efficient importance function. Bridge sampling introduced by Meng and Wong (1996) estimates the ratio of normalizing constants using importance samples from two non-normalized distributions. If a bridging distribution with known normalizing constant is used, this approach can be applied to estimate the evidence of the target distribution. If the Kullback-Leibler divergence (Kullback, 1968) between the target distribution and the bridging distribution is too large, bridge sampling will be inefficient. Gelman and Meng (1998) introduce path sampling which uses an infinite number of bridging distributions in order to estimate the ratio of normalizing constants. In statistical physics this approach is also called thermodynamic integration, see Ogata (1989), where integration over a parameter space is achieved by transiting over a temperature range. In Bayesian statistics this corresponds to a transition from the prior distribution to the posterior distribution of interest. Power posterior sampling, which is a discretized version of path sampling, uses a fixed number of temperature steps transiting from the prior to the posterior, see Friel and Pettitt (2008). A class of tempered posterior distributions is constructed which can be used to integrate over the parameter space of the posterior of interest. MCMC simulations from the tempered posteriors yield an estimate of the evidence using a trapezoidal approximation. If the discretized temperature path is well chosen and a sufficiently large number of temperature steps is used, the Kullback-Leibler divergence between subsequent tempered posteriors will be small. This results in a low discretizational error in the trapezoidal approximation. The estimation of the ERGM evidence is especially challenging as the normalizing constant of the likelihood is not available. Path sampling can be used to yield an estimate of this normalizing constant, see Hunter and Handcock (2006) and Friel (2013).

1.3 Problems and research aims

The MCMC ML approach introduced by Hunter and Handcock (2006) is a popular method for ERGM estimation. However, it can be numerically very unstable and finding suitable starting values can be difficult due to model degeneracy. A Bayesian approach introduced by Caimo and Friel (2011) is robust to model degeneracy and is a practical alternative for ERGM estimation. The Bayes factor can be used

for ERGM selection which requires estimation of the model evidence.

The main research aim for this thesis is to yield a valid estimate of the ERGM evidence using power posterior sampling. This approach requires MCMC simulation from tempered posterior distributions and evaluation of the intractable ERGM likelihood. A combination of the EA and power posterior sampling is proposed which will be referred to as power posterior exchange algorithm (PPEA). The normalizing constant of the ERGM likelihood is estimated using the identity introduced by Meng and Wong (1996) which allows for the explicit evaluation the likelihood (EEL) of the intractable ERGM class. The EEL can be achieved as a byproduct of power posterior sampling as it uses the same discretized temperature path to estimate the normalizing constant of the likelihood. This approach yields a valid estimate of the ERGM evidence and shall be referred to as power posterior exchange algorithm with explicit evaluation of the likelihood (PPEA-EEL).

Two other approaches of estimating the ERGM evidence are known. Similar to the PPEA-EEL, Caimo and Friel (2013) and Friel (2013) use path sampling to estimate the normalizing constant of the ERGM likelihood. The ERGM evidence is estimated using the identity introduced by Chib (1995) while the posterior is evaluated using a non-parametric density estimate of a MCMC sample simulated with the EA. This approach is restricted to ERGM specifications with at most five parameters. Thiemichen et al. (2016) apply a Laplace approximation to estimate the evidence of more complex ERGM specifications. They also use path sampling to estimate the ERGM normalizing constant and apply a Laplace approximation to a MCMC sample. This approach requires the strong assumption that the posterior can be approximated by a normal distribution which in many cases will not hold. The PPEA-EEL requires no approximation assumptions of the posterior and can be used if a non-parametric kernel density estimate is not available. However, the PPEA-EEL is computationally very expensive and not easy to implement. It requires complicated specifications and the evaluation of numerous MCMC simulations. At the end of this thesis, best practice recommendations for the PPEA-EEL specification are given.

The second research aim is to develop tools for the evaluation of power posterior sampling algorithms. Graphical methods are developed that allow for the inspection of the tempered posteriors transiting from the prior to the posterior. As numerous MCMC chains have to be inspected, aggregating methods are proposed which help to evaluate the convergence of the power posterior sampler. Further-

more, a method to evaluate the trapezoidal approximation is proposed which helps to judge the reliability of the PPEA-EEL. This methods suggests to adapt the importance sample size in the EEL-step to the step size of the discretized path which is a new insight in this field of research.

1.4 Overview

In chapter 2, the notation for network data and random graphs used throughout this thesis is introduced. It is shown how network statistics can be used to describe observed networks. The ERGM class is introduced and various assumptions of network dependency are illustrated. Methods of network simulation are introduced and ERGM estimation using the MCMC-ML approach is discussed. Finally, model degeneracy is illustrated and an overview on ERGM extensions and alternative stochastic models for network data is given.

In chapter 3, MCMC-ML ERGM estimation is applied to policy networks in Ghana, Senegal and Uganda. The communication ties between relevant actors of policy formulation are modeled using network statistics derived from various theories on policy formation and political communication. While the ERGM class has been used to model policy networks in developed countries before, this approach is completely new to developing countries. The MCMC-ML method used in chapter 3 is plagued by model degeneracy which causes numerical instability.

Bayesian ERGM estimation using the EA is robust to model degeneracy. In chapter 4, the fundamentals of Bayesian statistics are discussed. The EA uses auxiliary data simulated from the non-normalized ERGM likelihood in order to circumvent the evaluation of its intractable normalizing constant. The EA is applied to estimate ERGM specifications for Krackhardt's friendship network, see Krackhardt (1987) and the expert network in Ghana.

In chapter 5, Bayesian model selection using the marginal likelihood is discussed. The focus is on methods descending from importance sampling which are bridge sampling, path sampling and power posterior sampling. The PPEA is introduced as a new method of power posterior sampling for the ERGM class using the EA. It is shown how the temperature path constructed for the PPEA can be used to estimate the normalizing constant of the ERGM likelihood which allows for the EEL-step. The proposed new PPEA-EEL approach uses MCMC simulations from

the tempered posterior distributions and the corresponding sequence of explicitly evaluated tempered likelihood functions in order to yield an estimate of the ERGM evidence. The PPEA-EEL is applied to the Krackhardt's managers network and the expert network in Ghana. New graphical methods to evaluate the behaviour of the PPEA transiting over the temperature path are introduced.

Chapter 6 summarizes the thesis and discusses limitations of the PPEA-EEL and alternatives to this approach. Finally, an outlook is given on future research and how the efficiency of the PPEA-EEL can be increased.

Chapter 2

Exponential random graph models

Statistical models for binary outcomes like the logistic regression approach rely on the assumption of independence of observations. When analyzing network data this assumption is unrealistic as it is well known that individual behaviour depends on the interaction with others. In the simplest case of tie variable interdependence the probability of a tie from *alter* to *ego* might depend on the existence of a tie from *ego* to *alter*. Further, it is well known that social interaction is often structured after the principle *'a friend of a friend is a friend'*. The process of network tie formation has to be modeled in such a way that observations are independent *conditional* on explicitly formulated features of dependency. The exponential random graph model (ERGM) popularized by Wasserman and Pattison (1996) solves this problem for binary network data.[1] The ERGM can be used to test for a lot of hypothesized patterns of network interdependence including the influence of exogenous covariates. However, as will be discussed in section 2.5, this potential is limited by the numerical effort required for parameter estimation.

In section 2.1 a notation for random graphs and network data is introduced and the description of networks using appropriate summary statistics is discussed. It is shown how binary network data can be understood statistically as realizations of a random graph. In section 2.3 the ERGM model class is defined. It is illustrated how a dependence graph is used to deduce affiliated sufficient network statistics.

[1] Wasserman and Pattison (1996) refer to the p^* model. This term is no longer in use in the literature.

Various dependence assumptions are discussed in section 2.4 which help to formulate realistic models of network tie formation. Maximum likelihood (ML) ERGM estimation, which is introduced in section 2.5, requires Markov chain Monte Carlo (MCMC) methods of network simulation. MCMC-ML ERGM estimation which helps to circumvent the explicit evaluation of the intractable ERGM likelihood function is illustrated. The common problem of model degeneracy is discussed. Finally, a short overview of ERGM extensions and modeling alternatives is given.

2.1 Networks as random graphs

In social network analysis the relations between a fixed set of units (nodes) are considered. Typically nodes are actors such as friends within a class room, employees within an organization, organizations within a political system or national states in the global context. The relational links between two nodes are called edges or ties. Throughout this work we consider only binary ties which may be present or absent but have no relational weight such as the relative importance of a friendship. Consider a fixed set of n nodes that may or may not be connected by a tie. The relation between the pair of nodes i and j may be indicated by a binary random tie variable

$$Y_{ij} = 1, \quad i > j$$

if the two nodes are connected and

$$Y_{ij} = 0, \quad i > j$$

else. If $Y_{ij} = 1$, an edge y_{ij} exists between the two nodes whereas self ties y_{ii} are not allowed for any i. If directed relations are considered, the random tie variables $Y_{ij} \neq Y_{ji}$ i.e. it is of concern which of the nodes is sending a tie towards the other node. For undirected relations $Y_{ij} = Y_{ji}$. The collection of all N random tie variables on a fixed set of n nodes is called a random graph Y which may be represented by a $n \times n$ random adjacency matrix. Throughout this work capital letters denote random variables and small letters denote realizations: Y denotes a random graph on n nodes, Y_{ij} denotes a random tie variable. y denotes an observed network represented by an $n \times n$ adjacency matrix and y_{ij} denotes an observed tie. The diagonal of the adjacency matrix y is always empty as self ties

are not allowed. An entry in row i and column j represents the observed tie variable y_{ij}. For undirected graphs the adjacency matrix y is symmetric. If directed graphs (digraphs) are considered the upper and the lower triangular matrices of y are distinct. For undirected graphs the number of tie variables is

$$N = \binom{n}{2} = \frac{n(n-1)}{2}. \tag{2.1}$$

Considering directed relations the number of tie variables is

$$N = n(n-1). \tag{2.2}$$

The set of all possible random graphs $\mathcal{Y}$ on n nodes has a size of

$$\mathcal{G} = 2^{\binom{n}{2}} = 2^{n(n-1)/2} \tag{2.3}$$

for undirected graphs and a size of

$$\mathcal{G} = 2^{n(n-1)} \tag{2.4}$$

for digraphs. In social network analysis it is useful to define subgraphs of k nodes with all the corresponding tie variables. These are called k-subgraphs where the most common subgraphs are on two and three nodes. A 2-subgraph containing the nodes i and j is called a dyad $d_{i,j}$. For directed networks the dyad

$$d_{i,j} = (Y_{ij}, Y_{ji}), \quad i \neq j \tag{2.5}$$

consists of the two random tie variables Y_{ij} and Y_{ji} and the possibly connected nodes i and j. For undirected networks $d_{i,j}$ contains only one tie variable Y_{ij} possibly connecting i and j. If Y is undirected, the 3-subgraph on the triple of actors h, i, j is the triad

$$t_{h,i,j} = (Y_{hi}, Y_{ij}, Y_{jh}), \quad h \neq i \neq j \tag{2.6}$$

where the number of tie variables doubles if Y is directed. We make the assumption that there are isomorph states that can be observed on k-subgraphs if the labeling of the contained nodes is ignored. There are four isomorph possible realizations of edge counts within the undirected triad $t_{h,i,j}$: zero, one, two and three edges. For

example, if there is one single edge in the triad $t_{h,i,j}$, the tie $Y_{hi} = 1$ is isomorph to $Y_{ij} = 1$ and $Y_{jh} = 1$ as the nodes could be simply relabeled. If Y is a digraph, there are six isomorph single edge states in the triad $t_{h,i,j}$:

$$\{(Y_{hi} = 1); (Y_{ih} = 1); (Y_{ij} = 1); (Y_{ji} = 1); (Y_{hj} = 1); (Y_{jh} = 1)\}.$$

Isomorphism is especially important for states of a triad where the edges form closed triangular configurations. The directed triangle (y_{hi}, y_{ij}, y_{jh}) is isomorph to the directed triangle (y_{hj}, y_{ji}, y_{ih}) if the labeling of the nodes is ignored: both states of the triad $t_{h,i,j}$ represent a triangle constructed from non-reciprocal edges with every node sending to only one other node. This state is called a cyclic triad, see figure 2.1, middle panel. As there are 2^2 states per directed dyad (empty dyad, two states with a non-reciprocal edge, reciprocal edge) and three dyads nested within a triad there are $2^6 = 64$ possible states of a directed triad. If we ignore the labeling of the nodes there are 16 isomorphic states left. The state of a k-subgraph is called a configuration, so there are 16 possible isomorph configurations on a directed triad. Wasserman and Faust (1994) give details on what they call the triad census and on configurations of k-subgraphs with $k > 2$. Isomorphism is important for the homogeneity assumptions needed to identify an ERGM, see section 2.3.

2.2 Network statistics

A model for social networks needs to capture interdependencies like reciprocity, homophily due to actor attributes, transitivity and differences in nodal degrees due to activity and popularity of actors, see Snijders (2011). There are various statistics that may be used to describe such patterns in an observed network. Further, some of these statistics may be used as sufficient statistics of a model for network tie formation. An observed network y is a realization of the random graph Y and may be described using a vector of network statistics $s(y)$ containing a collection of functions of y. We mainly consider counts of elementary k-subgraphs a random graph may be constructed with where usually $k << n$. The simplest network statistic

$$L(y) = \sum_{i<j} y_{ij} \tag{2.7}$$

is the number of edges within an observed network. It is nested within all other network count statistics. The density of a network

$$D(y) = \frac{L(y)}{N} \tag{2.8}$$

is the share of the realized edges on all possible edges. (2.8) describes the propensity of the nodes in y to form ties. For social networks typically $0 < D(y) < 0.5$. If both edges within a directed dyad exist, this is called a reciprocal (or mutual) tie with the corresponding network statistic

$$M(y) = \sum_{i<j} y_{ij} y_{ji} \tag{2.9}$$

called reciprocity (or mutuality). This statistic describes the propensity of nodes in y to answer ties once received. As (2.7) is nested within the mutuality statistic, $M(y) = 1$ corresponds to $L(y) \geq 2$: if there is one mutual tie observed in the network, there must be at least two edges observed.

Network count statistics defined on 3-subgraphs are especially important to social networks analysis. Triads where three nodes are part of a non-empty dyad are called connected triads. Configurations where the edges y_{hi} and y_{ij} share the connected node i but where $Y_{hj} = 0$ are called 2-stars with the corresponding network statistic

$$S_2(y) = \sum_{h<i<j} y_{hi} y_{ij}. \tag{2.10}$$

For directed 2-stars there are three cases of interest: if both h and j are sending to i, the configuration is called a 2-in-star which has the interpretation of popularity of node i. If i is sending to h and j, the configuration is called a 2-out-star and has the interpretation of activity of node i. If h is sending to i and i is sending to j, this special case of $S_2(y)$ is called a 2-path $TP(y)$, see figure 2.1, right panel. This configuration is very important for the concept of transitivity. As not all nodes in the triad are connected in a $S_2(y)$ configuration, the undirected 2-star and the directed 2-path may have the interpretation of skipping others instead of being friends with everyone. Imagine a situation where people prefer to be connected to only one person instead of a group of people which is typical for romantic relations. This tendency is contrasted by a behaviour of nodes similar to *'a friend*

of a friend is a friend'. Human beings tend to have multiple friends resulting in clustered structures of social networks. It is commonly observed that every node in a triad is connected to the other two nodes, so nodes have the tendency to form closed triangles. A network statistic counting such triangular configurations for undirected networks is

$$T(y) = \sum_{h<i<j} y_{hi} y_{ij} y_{jh} \tag{2.11}$$

where there is only one way to form a closed triangle given an ordered permutation of a triple $h < i < j$. If y is directed there are seven isomorph configurations of a triad that form a closed triangle resulting in $T(y) = 1$: Two configurations resulting in

$$[L(y) = 3, M(y) = 0],$$

three configurations with

$$[L(y) = 4, M(y) = 1],$$

one configuration with

$$[L(y) = 5, M(y) = 2]$$

and one configuration with

$$[L(y) = 6, M(y) = 3].$$

Note that a closed triangle always consists of at least three nested $S_2(y)$ configurations which highlights how network statistics are nested within each other. Of particular interest is the directed transitive triangle

$$TT(y) = \sum_{h<i<j} y_{hi} y_{ij} y_{hj}, \tag{2.12}$$

see figure 2.1, left panel. The two ties $y_{hi} = 1$ and $y_{ij} = 1$ form a 2-path which is closed by $y_{hj} = 1$. This is the reason why 2-paths have an interpretation of potential transitive triangles. The occurrence of triangular structures is also referred to as network closure which is important for the concept of clustering and transitivity, see Wasserman and Faust (1994) and Lusher et al., eds (2012).

Configurations of the 3-subgraph are separated into transitive and intransitive

triads. Triads containing an empty dyad cannot be transitive, so 2-paths and 2-stars are intransitive configurations. For undirected networks the closed triangle is a transitive configuration as every node is connected to each other node. A directed triad is transitive if it contains a 2-path, a 2-out-star and a 2-in-star. A directed triangle containing only 2-paths is intransitive as every node is only sending a tie a single other node, resulting in a cyclic triangle, see figure 2.1, middle panel. The concept of transitivity is important to social network analysis as clustered regions in an observed network may often be constructed from transitive triad configurations. This is typical for friendship networks. If predominantly patterns of intransitivity are at work, this has an interpretation of brokerage which is not expected among friends. Davis (1970) conceptualize transitivity and hierarchy in social networks and examines the role of transitive configurations contributing to clustered regions in an observed graph.

Global transitivity of an undirected network may be measured using the clustering coefficient

$$C(y) = \frac{3 \cdot T(y)}{S_2(y)}. \tag{2.13}$$

Note that every closed triad consists of three nested 2-stars, so the numerator $3 \cdot T(y)$ leads to a a range of $C(y)$ from 0 to 1. If y is directed, $C(y)$ may be computed using directed transitive triangles $TT(y)$ and directed 2-paths $TP(y)$ resulting in

$$C(y) = \frac{3 \cdot TT(y)}{TP(y)}. \tag{2.14}$$

In both cases $C(y)$ may be interpreted as a measure of transitivity as it is a ratio of closed transitive triads and *potential* transitive triads that are not closed: $S_2(y)$ could be turned into $T(y)$ and $TP(y)$ could be turned into $TT(y)$ if a closing edge was added. If y contains a lot of connected triads without observing a single transitive triad, this would result in $C(y) = 0$. No clustered regions would be observed but rather loosely connected strands of ties forming chain like structures. In the extreme case of $C(y) = 1$ a graph would be completely constructed from transitive triangles resulting in one single dense cluster of nodes.

More complex network statistics may be defined on higher order k-subgraphs. The transitive k-triangle configuration $T_k(y)$ consists k of transitive triangles that

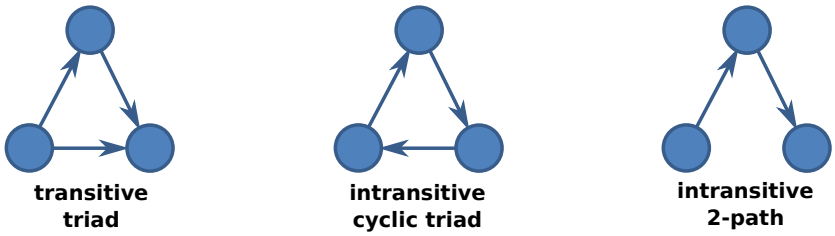

Figure 2.1: Transitive and intransitive triads:
Left panel: Transitive triad $TT(y)$: a closed triangle where one node is sending to two others.
Middle panel: Intransitive cyclic triad, every node is sending to only one other node.
Right panel: Intransitive triad $TP(y)$: the triangle is not closed.

are stacked on the shared edge $y_{hj} = 1$, so the dyad $d_{h,j}$ must not be empty. $T_k(y)$ is also called the edge-wise shared partner statistic $EP_k(y)$ which is an important statistic in modelling observed networks realistically. In addition to the tie $y_{hj} = 1$, the nodes h and j are also connected indirectly via k shared partners. If y is directed, $EP_k(y)$ corresponds to stacking k 2-paths on the directed edge $y_{hj} = 1$.

$$EP_k(y) = \sum_{i=1}^{k} \sum_{h<i<j} y_{hi} y_{ij} y_{hj} \tag{2.15}$$

Similarly, the directed k-2-path statistic $TP_k(y)$ counts the number of 2-paths that can be stacked on the shared dyad $d_{i,j}$ while the nodes i and j *do not have to* be connected, so the dyad $d_{i,j}$ may be empty. Therefore $TP_k(y)$ is also called the dyad-wise shared partner statistic $DP_k(y)$: it is not relevant whether the dyad $d_{i,j}$ contains any edge. The nodes i and j are connected *at least* via k shared partners.

$$DP_k(y) = \sum_{i=1}^{k} \sum_{h<i<j} y_{hi} y_{ij} \tag{2.16}$$

It is easy to see that $DP_k(y)$ is nested within $EP_k(y)$. Both statistics can also be computed for undirected networks. The directed k-triangle and k-2-path statistics are illustrated in the center of figure 2.2. The other most common class of network statistics defined on k-subgraphs are k-star (or k-degree) statistics. For undirected

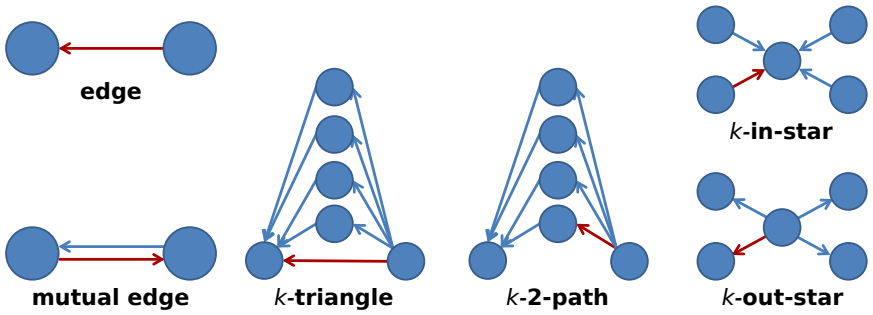

Figure 2.2: Examples of common directed network statistics, $k = 4$

networks the k-star statistic is defined as

$$S_k(y) = \sum_{i=1}^{k} \sum_{i<j} y_{ij}. \tag{2.17}$$

$S_1(y)$ is equivalent to $L(y)$ and $S_2(y)$ is equivalent to an undirected 2-path. In-stars and out-stars are also called in-degrees and out-degrees for directed networks. They have the interpretation of popularity (or attractivity) and activity (or outgoingness) of actors. k-star statistics are illustrated on the right hand side of figure 2.2.

There are many more ways to describe network data. We cover only the most fundamental statistics which are needed to understand ERGM model estimation and are commonly used as sufficient statistics, see section 2.3. We do not cover network paths, connectivity and centrality of networks. Wasserman and Faust (1994), Jackson (2008) and Morris et al. (2008) are rich sources on these topics. Also, we do not deal with the visualization of any but very simple graphs using basic plot algorithms. This work will be focused on the distribution of network statistics in order to describe relational data.

2.3 Model definition and interpretation

Instead of modeling independent binary tie variables of a network a model is needed for the joint random tie variables in the random graph Y. If one wishes to model realistic network interdependencies, the probability distribution of a random tie

variable needs to be modeled conditional on the rest of the graph as

$$\Pr(Y_{ij} = y_{ij}|Y_{-ij}).$$

The ERGM popularized by Wasserman and Pattison (1996) offers the potential to model such tie variable interdependencies. It has the basic form of an exponential family distribution

$$\Pr(Y = y|\theta) = \frac{\exp\{\theta' \cdot s(y)\}}{z(\theta)}. \tag{2.18}$$

$s(y) = (s_1(y), \ldots, s_P(y))$ is a collection of P network statistics of the observed network y discussed in section 2.2. The choice of sufficient network statistics $s(y)$ corresponds to a particular pattern of network dependency. $\theta' = (\theta_1, \ldots, \theta_P)$ is a vector of P corresponding model parameters and $z(\theta)$ is a normalizing constant insuring that (2.18) is a proper probability distribution. The normalizing constant

$$z(\theta) = \sum_{\tilde{y} \in \mathcal{Y}} \exp\left\{\theta' \cdot s(\tilde{y})\right\} \tag{2.19}$$

requires summation over all $\mathcal{G}$ elements in the space of possible graphs

$$\mathcal{Y} = \{\tilde{y}_1, \ldots, \tilde{y}_\mathcal{G}\}$$

on n nodes. It is easy to see that even for small networks (2.19) is analytically unavailable as there are too many elements in $\mathcal{Y}$. An undirected graph on $n = 10$ nodes has

$$\mathcal{G} = 2^{n(n-1)/2} = 2^{45}$$

possible realizations, whereas a directed random graph on $n = 46$ nodes has

$$\mathcal{G} = 2^{n(n-1)} = 2^{2070}$$

possible realizations, a digit which cannot be represented on a standard Windows computer.[2] This is the major problem with ERGM estimation: while evaluation

[2] In section 2.5 it will be shown how network simulation can be used to circumvent the evaluation of $z(\theta)$ for ML ERGM estimation.

of the non-normalized kernel of the probability distribution

$$\exp\left\{\theta' \cdot s(y)\right\}$$

is straightforward, the normalized probability distribution

$$\frac{\exp\left\{\theta' \cdot s(y)\right\}}{z(\theta)}$$

is unavailable due to the intractability of (2.19). A log-linear representation of (2.18) can be used to avoid the evaluation of (2.19), see section 2.3.2. Methods based on path sampling are available which allow for the estimation of (2.19) and will be discussed in chapter 5.

2.3.1 Dependence graphs and sufficient statistics

The a *priori choice* of statistics in $s(y)$ corresponds to a hypothesized pattern of self organization of network tie variables, see Snijders et al. (2006). (2.18) depends on the linear combination $\theta' \cdot s(y)$ where the goal of ERGM estimation is to infer on the unknown parameter vector θ. The statistics in $s(y)$ determine the assumption of network dependence and, vice versa, the assumption of conditional *independence* of network tie variables. Models of the form (2.18) represent a distribution of random graphs which are constructed from the smaller subgraph configurations in $s(y)$.

Frank and Strauss (1986) impose the assumption of homogeneity of isomorphic network configurations which holds for all models of the form (2.18). E.g. all parameters for reciprocal ties are equated assuming that all nodes have the same propensity to answer received ties. This might not be a very realistic assumption as people do tend to differ in such propensities, but it greatly reduces the number of parameters to be estimated.

Early predecessors of the ERGM like the Bernoulli graph model and the Markov model discussed in section 2.4 are limited in the choice of functions in $s(y)$. The general form of an ERGM popularized by Wasserman and Pattison (1996) which is further developed by Snijders et al. (2006) may contain arbitrary statistics in $s(y)$. These statistics may go beyond basic network configurations like the number of edges $L(y)$ or the number of triangles $T(y)$ described in section 2.1. $s(y)$ may also include statistics $s(y, x)$ which are functions of exogenous covariates x, see section 2.4.5. Further, geometrically weighted network statistics can model complex

patterns of transitivity and clustering, see section 2.4.4.

Models of the form (2.18) have their origin in spacial statistics and statistical mechanics. They are developed from models for spatial interactions in Markov random fields like the Ising model of ferromagnetism in lattice structures, see Ripley (1991) for an overview. Besag (1974) show how an exponential family conditional probability distribution for tie variables in random graphs of size n of the form

$$\Pr(Y_{ij}|Y_{-ij}) \tag{2.20}$$

can be constructed using subgraph configurations discussed in section 2.2 as sufficient network statistics. Y_{-ij} is a set of tie variables neighbouring Y_{ij}, so Y_{ij} depends on its neighbours but is conditionally independent from all other tie variables in the graph. Any singular tie Y_{ij} is assumed to have non-zero probability

$$\Pr(Y_{ij}) > 0$$

and it is further assumed that ties can occur together so that

$$\Pr(Y_{ij_1}, \ldots, Y_{ij_{n-1}}) > 0.$$

This is what Besag (1974) call the positive condition which shall be assumed throughout this work. The neighbourhood Y_{-ij} is defined by the functional form of (2.20).

Frank and Strauss (1986) extend (2.20) to what they call Markov random graphs where Y_{ij} is a random tie variable in a graph on n nodes. Y_{-ij} are the other tie variables in the graph so that

$$\Pr(Y_{ij} = 1|Y_{-ij}, \mathfrak{G}).$$

The dependence of Y_{ij} on other tie variables Y_{-ij} is defined by a so called dependence graph $\mathfrak{G}$ which implies certain network configurations like triangles and k-stars as sufficient network statistics of (2.20), see section 2.4. $\mathfrak{G}$ is an undirected graph where the N tie variables of the observed network y are nodes. $\mathfrak{G}$ connects the tie variables that are assumed to be dependent. If all tie variables are assumed to be independent, $\mathfrak{G}$ is an empty graph as illustrated on the left panel of figure

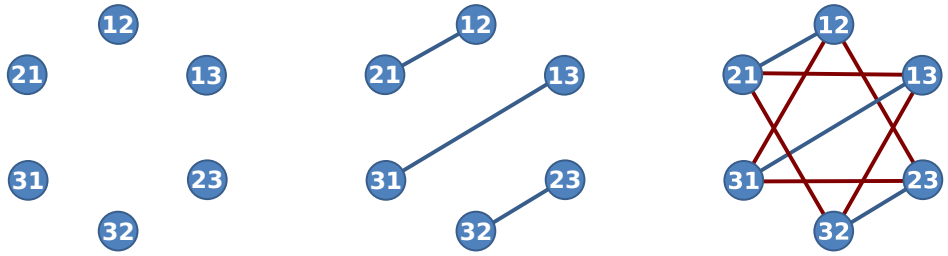

Figure 2.3: Three dependence graphs on a small directed network on $n = 3$ nodes with $N = 6$ tie variables resulting from different assumptions of conditional independence. Unconnected tie variables are assumed to be conditionally independent.
Left panel: Tie variables are assumed to be independent, empty graph.
Middle panel: Reciprocal ties are assumed to depend on each other, blue lines indicate the dependence assumption of reciprocity.
Right panel: In addition ties within cyclic triads like (y_{12}, y_{23}, y_{31}) are assumed to depend on each other, red lines indicate the dependence assumption of cyclic triangulation.

2.3: no dependence relations exist in $\mathfrak{G}$, where

$$\Pr(Y_{ij} = 1 | Y_{-ij}, \mathfrak{G}) = \Pr(Y_{ij} = 1).$$

If a pattern of reciprocity is assumed for a directed graph, mutual tie variables Y_{ij} and Y_{ji} are connected in $\mathfrak{G}$, see the middle panel of figure 2.3. The probability of a directed tie depends only on the other tie variable in the dyad so that

$$\Pr(Y_{ij} = 1 | Y_{-ij}, \mathfrak{G}) = \Pr(Y_{ij} = 1 | Y_{ji}).$$

If a particular pattern of triadic interdependence like cyclic triangulation is assumed, $\mathfrak{G}$ gets denser, see the right panel of figure 2.3. The two possible cyclic triangles on $n = 3$ nodes are (y_{12}, y_{23}, y_{31}) and (y_{13}, y_{32}, y_{21}). If all triangles of the triad census were allowed, $\mathfrak{G}$ on $n = 3$ nodes would be a full graph with every tie variable connected to each other (not shown). Robins and Pattison (2005) give details on the dependence graph ranging from basic lattice models to the ERGM (2.18). More detailed illustrations of various dependence graphs are given in Koskinen and Daraganova (2012).

Frank and Strauss (1986) use $\mathfrak{G}$ to translate assumptions of conditional independence into counts of network statistics. They use the Hammersley-Clifford

theorem of Besag (1974) to proof that a random graph model can be formulated in an exponential form using counts of network subgraphs as sufficient statistics. $\mathfrak{G}$ directly defines the sufficient statistics of an ERGM, see also Wasserman and Pattison (1996). These statistics can be interpreted as elementary subgraphs a random graph Y may be constructed with, which was a major breakthrough in modeling random graphs in those days.

An important assumption is that all subgraph configurations representing a sufficient network statistic are homogenous, e.g. all edges and triangles have the same probability to occur in the graph. The homogeneity assumption allows for a log-linear interpretation of (2.18) where simple counts of sufficient network statistics can be used. Throughout the rest of this work we refrain from explicitly conditioning on $\mathfrak{G}$. We will define the set of sufficient network statistics $s(y)$ *a priori* and implicitly assume the corresponding dependence graph $\mathfrak{G}$.

2.3.2 Log-linear formulation

Strauss and Ikeda (1990) give a log-linear formulation of (2.18) where the probability of an existing tie from i to j is modeled conditional on all other tie variables of the observed network y:

$$\Pr(Y_{ij} = 1|Y_{-ij}) = \frac{\Pr(Y = y_{ij}^+)}{\Pr(Y = y_{ij}^+) + \Pr(Y = y_{ij}^-)} \tag{2.21}$$

where y_{ij}^+ is the observed network y with the tie variable $Y_{ij} \stackrel{!}{=} 1$ being forced to be one and y_{ij}^- is the observed network y with $Y_{ij} \stackrel{!}{=} 0$ being forced to be zero, see Wasserman and Pattison (1996). y_{ij}^+ and y_{ij}^- differ only in the value of the tie variable Y_{ij}. Using (2.18), equation (2.21) can be reformulated as

$$\Pr(Y_{ij} = 1|Y_{-ij}, \theta) = \frac{\exp\{\theta' \cdot s(y_{ij}^+)\}}{\exp\{\theta' \cdot s(y_{ij}^+)\} + \exp\{\theta' \cdot s(y_{ij}^-)\}.} \tag{2.22}$$

A single tie can be modeled conditional on the rest of the graph while the intractable normalizing constant $z(\theta)$ cancels in (2.22). Logistic regression allows to analyze the odds of a binary outcome which in this case is

$$\frac{\Pr(Y_{ij} = 1|Y_{-ij}, \theta)}{\Pr(Y_{ij} = 0|Y_{-ij}, \theta)} = \frac{\exp\{\theta' \cdot s(y_{ij}^+)\}}{\exp\{\theta' \cdot s(y_{ij}^-)\}} = \exp\{\theta'[s(y_{ij}^+) - s(y_{ij}^-)]\}. \tag{2.23}$$

Figure 2.4: Change in network statistic counts on edges, 2-stars and triangles of a small undirected network on $n = 3$ nodes if the random tie variable Y_{ij} is changed from zero (left panel) to one (right panel).

The change statistics

$$\delta_{ij} = \left[s(y_{ij}^{+}) - s(y_{ij}^{-}) \right] \tag{2.24}$$

represent the $1 \times P$ vector of change in the sufficient network statistics $s(y)$ if the tie variable Y_{ij} is toggled from 0 to 1. A simple example of change statistics for an undirected graph is given in figure 2.4. Three statistics are considered in $s(y)$: the number of edges $L(y)$, the number of 2-stars $S_2(y)$ and the number of triangles $T(y)$. If the tie y_{ij} is added, the number of edges increases by one to $L(y) = 3$. As this tie now closes a triangle, $T(y)$ is increased from zero to one. But an undirected triangle also consists of three 2-stars increasing the corresponding count from one to three. So the resulting change statistics are

$$\delta_{ij} = \left[L(y)_{\delta} = +1;\ S_2(y_{\delta}) = +2;\ T(y)_{\delta} + 1 \right].$$

Using (2.24) in (2.23) yields the odds

$$\frac{\Pr(Y_{ij} = 1 | Y_{-ij}, \theta)}{\Pr(Y_{ij} = 0 | Y_{-ij}, \theta)} = \exp\{\theta' \cdot \delta_{ij}\}. \tag{2.25}$$

Taking the log from (2.25) yields what Wasserman and Pattison (1996) call the logit p^* model

$$\omega_{ij} = \ln \left\{ \frac{\Pr(Y_{ij} = 1 | Y_{-ij}, \theta)}{\Pr(Y_{ij} = 0 | Y_{-ij}, \theta)} \right\} = \theta' \cdot \delta_{ij}. \tag{2.26}$$

ω_{ij} are the log odds (logit) of a tie being present using the change statistics this tie induces. In the logit formulation of the ERGM the change statistics δ_{ij} have the role of explanatory variables in modeling $\Pr(Y_{ij} = 1 | Y_{-ij}, \theta)$. The random tie

variable depends on the change in network statistic counts δ_{ij} induced by toggling Y_{ij} from zero to one. This allows for a log linear interpretation of the ERGM similar to logistic regression. Examples will be given in section 2.4.

2.4 Dependence assumptions

Before encountering the general form of the ERGM simpler exponential family models for random graphs are considered. All models presented in this section are of the form (2.18) but differ in the choice of sufficient networks statistics in $s(y)$ which define different dependence assumptions.

2.4.1 The Bernoulli random graph model

The Bernoulli random graph model of Erdös and Renyi (1959) represents the simplest form of an ERGM with the number of observed edges $L(y)$ as the only sufficient network statistic in $s(y)$. It assumes the tie variable Y_{ij} to be completely independent from the rest of the graph Y_{-ij}. This might be a rather unrealistic assumption but it will facilitate an understanding of how the ERGM class works. All independent $N = n(n-1)$ tie variables in the directed random graph Y on n nodes follow a Bernoulli distribution with

$$\Pr(Y_{ij} = 1|Y_{-ij}) = \Pr(Y_{ij} = 1) = \vartheta$$

as the only parameter. So the probability of observing the network y can simply be expressed using the constant probabilities of the independent ties resulting in

$$\Pr(Y = y|\vartheta) = \prod_{i>j} \vartheta^{y_{ij}} \cdot (1-\vartheta)^{1-y_{ij}} = \frac{\exp\left\{L(y)\ln(\frac{\vartheta}{1-\vartheta})\right\}}{\exp\left\{-N\ln(1-\vartheta)\right\}}, \tag{2.27}$$

see appendix A.1 for details. If we consider

$$\theta = \ln\left(\frac{\vartheta}{1-\vartheta}\right)$$

it is easy to see that the enumerator of (2.27)

$$\exp\left\{L(y)\ln\left(\frac{\vartheta}{1-\vartheta}\right)\right\} = \exp\{\theta \cdot L(y)\}$$

is the kernel of an ERGM in the sense of (2.18) with the only parameter θ and the only sufficient network statistic $L(y)$. The denominator of (2.27)

$$\exp\{-N\ln(1-\vartheta)\} = z(\theta)$$

is the corresponding normalizing constant, see the proof in appendix A.1. $z(\theta)$ does not depend on y: it does not contain $L(y)$ and is a constant if the number of nodes n and the probability of tie formation ϑ are known. If $\vartheta = 0$ no ties are possible and the empty graph is the only element in $\mathcal{Y}$. This leads to the simplest ERGM possible

$$\Pr(Y = y|\theta) = \frac{\exp\{\theta \cdot L(y)\}}{z(\theta)} \tag{2.28}$$

where $\theta = \ln(\vartheta/(1-\vartheta))$ directly corresponds to the Bernoulli distribution of the independent tie variables. Note the assumption of tie variable homogeneity for the ERGM, so in the model (2.28) all ties have the same probability

$$\Pr(Y_{ij} = 1|\theta) = \vartheta = \frac{e^\theta}{1+e^\theta}$$

to occur.

The log-linear interpretation of the ERGM model class is based on the change statistics (2.24), see section 2.3.2. For any ERGM of the form (2.28) it holds that $\delta_s(y_{ij}) = 1$ as $s(y)$ contains only the edge count $L(y)$. This results in the log odds of a present tie

$$\omega_{ij} = \theta \cdot \delta_s(y_{ij}) = \theta$$

as changing the tie variable Y_{ij} from zero to one will increase $L(y)$ by $+1$. A negative value of θ has the interpretation that the mechanism of tie formation at work prefers to have less ties present in the network. In this case the odds

$$\frac{\Pr(Y_{ij} = 1|\theta)}{\Pr(Y_{ij} = 0|\theta)} = \exp\{\omega_{ij}\} = e^\theta$$

of a tie present are smaller than one. This is typical for social networks which tend to be sparse graphs with $L(y) < N$ and $D(y) < 0.5$.

Due to the simplicity of (2.28) the Bernoulli random graph model is well suited to get a deeper understanding of the ERGM class. Consider a simple directed network on $n = 2$ nodes consisting of $N = n(n-1) = 2$ tie variables (Y_{ij}, Y_{ji}). The number of possible graphs is $\mathcal{G} = 2^{n(n-1)} = 4$ which can easily be listed:

$$\begin{aligned}\mathcal{Y} = \{&\tilde{y}_1 = (Y_{ij} = 0, Y_{ji} = 0),\\ &\tilde{y}_2 = (Y_{ij} = 1, Y_{ji} = 0),\\ &\tilde{y}_3 = (Y_{ij} = 0, Y_{ji} = 1),\\ &\tilde{y}_4 = (Y_{ij} = 1, Y_{ji} = 1)\}.\end{aligned}$$

The possible realizations of sufficient network statistics are

$$s(\mathcal{Y}) = \{L(\tilde{y}_1) = 0, L(\tilde{y}_2) = 1, L(\tilde{y}_3) = 1, L(\tilde{y}_4) = 2\}$$

using $s(\tilde{y}) = L(\tilde{y}) = \sum_{i>j} \tilde{y}_{ij}$. If we observe the network

$$y = (y_{ij} = 1, y_{ji} = 0),$$

we yield the ML estimate

$$\hat{\vartheta} = \frac{\sum y_{ij}}{n(n-1)} = \frac{L(y)}{N} = \frac{1}{2},$$

thus

$$\hat{\theta} = \ln\left(\frac{\hat{\vartheta}}{1-\hat{\vartheta}}\right) = 0.$$

This value has the interpretation that the mechanism of tie formation at work is indifferent to adding any tie as the log odds of observing y

$$\omega_{ij} = \hat{\theta} = 0.$$

The odds of a tie present are exactly one. Adding a tie does not render y more unlikely as it would be the case with $\hat{\theta} < 0$. The kernel of the ERGM is

$$\exp\{\theta \cdot s(y)\} = \exp\{0 \cdot 1\} = 1,$$

and the normalizing constant[3] may easily be calculated as

$$\sum_{\tilde{y} \in \mathcal{Y}} \exp\{\theta \cdot s(\tilde{y})\} = \mathcal{G} \cdot 1 = 4.$$

So the ERGM can be evaluated explicitly[4] as

$$\Pr(Y = y) = \Pr(Y_{ij} = 1, Y_{ji} = 0) = \frac{1}{4}.$$

The Bernoulli random graph model is also suitable to illustrate the relation between the ERGM parameter vector θ and the expected value of sufficient networks statistics $\mathrm{E}_\theta[s(y)]$ for the random graph Y. If $\theta = 0$, we expect that 50% of all possible network ties are observed resulting in a rather dense networks, so

$$\mathrm{E}_\theta[L(Y)] = N/2.$$

Typically, social networks are sparse graphs with $D(y) < 0.5$, resulting in a negative parameter values of the edge statistic $L(y)$ for almost any ERGM and $\mathrm{E}_\theta[L(Y)] < N/2$.

Note that the log-linear interpretation of (2.28) is equivalent to a logistic regression model with the intercept as only parameter. As long as no subgraph configurations except of the edge statistic $L(y)$ are used, an ERGM could also be estimated using logistic regression as the tie variables are assumed to be independent. However, it is the strength of the ERGM class to model such interdependencies which will be discussed in the next subsections.

2.4.2 The dyad independence model

The dyad independence model is the first model to capture at least some interdependence within a directed graph, see Holland and Leinhardt (1981) who call it

[3]Note that the normalizing constant of any ERGM with all elements in the parameter vector θ equal to zero is the number of possible graphs $\mathcal{G}$. This will be useful in estimating the normalizing constant explicitly, see section 5.3.2.

[4]While the evaluation of a Bernoulli random graph model is trivial, the calculation of $z(\theta)$ is impossible for networks of even moderate size of say $n = 6$ nodes if the ERGM specification is more complex. The `R` (R Development Core Team, 2011) package `ergm` (Handcock et al., 2008) offers a function to explicitly list $\mathcal{Y}$ for directed networks of up to 5 nodes and evaluating the ERGM with exact calculation of $z(\theta)$.

the p-1 model, and Fienberg and Wasserman (1981). While two dyads $d_{i,j}$ and $d_{k,l}$ are independent, the two tie variables (Y_{ij}, Y_{ji}) within the dyad $d_{i,j}$ do depend on each other. This represents a typical reciprocal pattern of tie variable self organizations in social networks as actors tend to answer ties they once received. The dyad independence ERGM is

$$\Pr(Y = y|\theta) = \frac{\exp\{\theta_L \cdot L(y) + \theta_M \cdot M(y)\}}{z(\theta)}. \tag{2.29}$$

In contrast to (2.28), now the number of reciprocal edges $M(y)$ is added as sufficient network statistic. The log-linear interpretation of (2.29) requires the evaluation of the change statistics using $\theta' \cdot \delta_{ij}$ for the added tie variable $Y_{ij} = 1$. In contrast to the Bernoulli random graph model, now there are two possible values. If $Y_{ji} = 0$, the additional tie $Y_{ij} = 1$ will only increase the edge count $L(y)$ by one and will not create a mutual edge. An isolated tie within an empty dyad will be generated. Assume the parameter values $\theta' = (\theta_L = -1;\ \theta_M = 2)$. This results in the change statistics

$$\delta_{ij} = \{L(y)_\delta = +1;\ M(y)_\delta = 0\}$$

and the log odds

$$\omega_{ij} = -1 \cdot 1 + 2 \cdot 0 = -1.$$

Adding an isolated tie renders the network more unlikely.

If the directed tie $Y_{ji} = 1$ already exists, $Y_{ij} = 1$ will create a mutual edge within a non-empty dyad, resulting in

$$\delta_{ij} = \{L(y)_\delta = +1;\ M(y)_\delta = +1\}$$

as both a new tie and a mutual tie are created. In this case the log odds of observing $Y_{ij} = 1$ is

$$\omega_{ij} = \theta_L \cdot L(y)_\delta + \theta_M \cdot M(y)_\delta = -1 \cdot 1 + 2 \cdot 1 = 1.$$

If $Y_{ij} = 1$ creates a mutual edge, the log odds of that tie ω_{ij} will be increased from -1 to 1 by 2. Equivalently, the odds of a tie is increased by the factor

$$e^2 = 7.389$$

if this tie creates a mutual edge. The mechanism of tie formation at work prefers

to have edges within non-empty dyads rather than isolated edges. Non-reciprocal edges are much more unlikely than mutual edges within a dyad. This is a typical pattern of human behaviour, as for example friendship, cooperation and romantic relations are reciprocal by nature or are at least unlikely to persist if they are not reciprocated.

2.4.3 The Markov model

In order to capture patterns of transitivity common in social networks, Frank and Strauss (1986) introduced the Markov assumption of network dependence: all random tie variables are independent *unless* they share a node. If the random variables Y_{ij} and Y_{kl} are independent conditional on all other random variables in the graph $Y_{-(i,j,k,l)}$, but Y_{ij} and Y_{jk} do depend given $Y_{-(i,j,k)}$, Y is a Markov graph. Y_{ij} and Y_{jk} depend on each other as they share the node j. This requires sufficient network statistics defined at least on the 3-subgraph such as triangles and 2-stars. Frank and Strauss (1986) discuss the basic triad model defined on the 3-subgraph using triangle and 2-star counts

$$\Pr(Y = y|\theta) = \frac{\exp\{\theta_L \cdot L(y) + \theta_S \cdot S_2(y) + \theta_T \cdot T(y)\}}{z(\theta)}. \tag{2.30}$$

2.30 is the simplest ERGM that fulfills the assumption of Markov dependence and can be extended by including higher order star configurations $S_{k=3}(y), \ldots, S_{k=n-2}(y)$ defined on k-subgraphs up to $k = n - 2$. Historically, the Markov model is the first ERGM that can model complex patterns of social behaviour like transitivity by dropping the strong and unrealistic assumption of dyadic independence.

Assume the parameter vector of (2.30) to be

$$\theta' = (\theta_L = -1,\ \theta_S = 0.5,\ \theta_M = 1.5)\,.$$

If $Y_{ij} = 1$ closes a triangle, the resulting change statistics are

$$\delta_{ij} = \{L(y)_\delta = +1;\ S_2(y)_\delta = +2;\ T(y)_\delta = +1\}\,,$$

see the example in figure 2.4. The resulting logit of $Y_{ij} = 1$ is positive as

$$\omega_{ij} = -1 \cdot 1 + 0.5 \cdot 2 + 2 \cdot 1 = 1.5.$$

If no triangle is closed but a two star is created the logit is

$$\omega_{ij} = -1 \cdot 1 + 0.5 \cdot 1 + 2 \cdot 0 = -0.5$$

which has the interpretation that ties are more likely to occur within triangles than within 2-stars. If neither a triangle is closed nor a 2-tar is created, the logit is

$$\omega_{ij} = -1 \cdot 1 + 0.5 \cdot 0 + 2 \cdot 0 = -1.$$

This has the interpretation that isolated ties are the most unlikely edges to occur whereas ties are most likely to occur within triangles.

A logistic regression interpretation of (2.30) requires an assumed *ceteris paribus* unit increase in the change statistic of the corresponding parameter. If all change statistics except $L(y)_\delta$ are zero, $\omega_{ij} = -1$ represents the baseline propensity to form isolated ties. θ_L has an interpretation equivalent to the intercept in logistic regression models. If $S_2(y)_\delta$ is increased by one, *ceteris paribus*, the logit of a tie creating a 2-star is increased by the factor $e^{\theta_S} = 1.649$. If $T(y)$ is increased by one, *ceteris paribus*, the logit of a tie closing a triangle is increased by the factor $e^{\theta_T} = 4.482$. Note that $T(y)$ cannot be increased without increasing the nested $S_2(y)$ and $L(y)$, thus the *ceteris paribus* interpretation.

2.4.4 The social circuit model

While Wasserman and Pattison (1996) popularized the ERGM for social network modeling using the Markov dependence assumption, the resulting ERGM model specifications often lead to unreasonable parameter estimates of θ. The Markov model is plagued by a phenomenon called model degeneracy as discussed in Handcock (2003a) and Handcock (2003b). It causes non-convergence of the MCMC-ML scheme used to estimate ERGM parameters, see section 2.5.4. A degenerate model will provide useless parameter estimates $\hat{\theta}$ and nonsensical network simulations given those parameter estimates. Simulating data from $p(y|\hat{\theta})$ the resulting networks tend to be mixtures of empty graphs with not a single tie present and full graphs with all possible ties present. Using network simulations model degeneracy will be illustrated in section 2.5.4.

What first was considered an algorithmic problem in optimizing the likelihood function was found to be a problem of the network statistics used in the Markov mo-

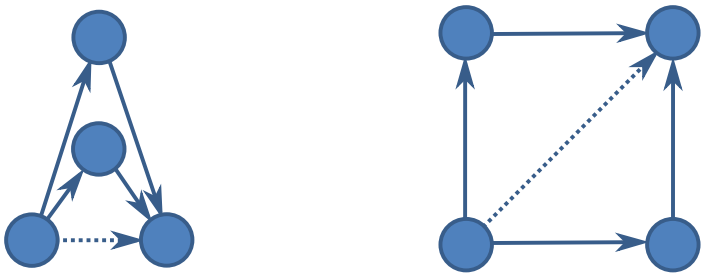

Figure 2.5: A directed 4-cycle *(right panel)* may be represented as $EP_2(y)$ or as $DP_2(y)$ (*left panel)*, depending on whether the dashed tie exists.

del, see Snijders et al. (2006) and Schweinberger (2011). The Markov assumption is still too restrictive and unrealistic to capture tie variable dependence of observed social networks. To ameliorate this problem Snijders et al. (2006) develop a new ERGM specification to capture complex patterns of transitivity. This development stage of the ERGM is called the social circuit model. New parameters capturing network transitivity are introduced which require a less restrictive assumption of network dependency. The assumption of the Markov graph by Frank and Strauss (1986) is replaced with the partial conditional dependence assumption introduced by Pattison and Robins (2002): two dyads are dependent if they share a node *or* if they are part of a 4-cycle. Imagine the situation of married couples: if the two husbands are friends, the wives are also more likely to know each other. Clustering and transitivity may not only occur in the form of triangular subgraphs like $T(y)$ but also in 4-cycles with potential diagonal tie variables. Note that a 4-cycle may also be represented by a shared partner statistic like the 2-triangle $EP_2(y)$ or the 2-2-path $DP_2(y)$, see figure 2.5. The partial conditional dependence assumption is more realistic than the Markov assumption but also requires more complex network statistics based on k-triangles and k-2-paths. Snijders et al. (2006) develop statistics such as the alternating k-triangle and the alternating k-2-path:[5] a single statistic represents the whole distribution of possible k-triangles and k-2-paths with k ranging from 1 to $n-2$ shared partners. The sign of subsequent k-statistics is toggled, so if $k-1$ has a positive sign, k will be negative, $k+1$ again positive

[5]Snijders et al. (2006) also propose a new specification for k-stars capturing network hierarchy which do not require the partial conditional dependence assumption. In this work the discussion is restricted to network statistics capturing transitivity.

and so on. This helps to prevent the avalanche effect causing model degeneracy, see section 2.5.4.

Hunter (2007) propose a slight modification of the statistics introduced by Snijders et al. (2006) using geometrically weighted $EP_k(y)$ and $DP_k(y)$ statistics. A dampening parameter regulates the influence of higher degrees of shared partners to the respective change statistics. The most important configuration for modeling transitivity in the social circuit model is the geometrically weighted edge-wise shared partner statistic (GWESP): the distribution of EP_k with $(k = 1, \ldots, n-2)$ is represented using a geometric series resulting in the $GWESP(y)$ statistic

$$GWESP(y) = e^{\alpha_E} \sum_{k=1}^{n-2} \Big(1 - \big(1 - e^{-\alpha_E}\big)^k\Big) EP_k(y). \tag{2.31}$$

$EP_k(y)$ is the number of transitive k-triangles in the network, see section 2.2. The geometric series

$$1 - \big(1 - e^{-\alpha_E}\big)^k$$

helps to prevent the avalanche effect of model degeneracy described in Snijders et al. (2006), see also section 2.5.4. The tuning parameter α_E controls the weight of the contribution of higher degrees of shared partners to the change statistics used in the log linear ERGM formulation. Goodreau et al. (2009) give a detailed illustration of how α_E affects the change statistics. Their work is also a rich source on patterns of transitivity and how they can be modeled using the $GWESP(y)$ statistic. Tuning of α_E may be used to focus on smaller or larger clusters of nodes, e.g. $\alpha_E = 0.1$ puts most weight on small clusters with $k \leq 2$, whereas $\alpha_E = 1.5$ puts substantial weight on larger cluster with $k \geq 10$. This is important for two reasons: First, it is possible to tune (2.31) for a good representation of the observed network as $\alpha_E = 10$ would not make any sense in a tiny network of only five nodes. Second, tuning of α_E can be used to prevent model degeneracy. It may happen that a particular value of α_E is plausible for the observed network which might show substantial transitivity, but the parameter estimate $\hat{\theta}$ lead to a degenerate distribution $S(y|\hat{\theta}, \alpha_E)$ of the relevant network statistics. While reducing α_E will reduce the potential of the ERGM to completely explain the observed transitivity, this reduction might prevent model degeneracy. A more limited model might be better than a model which cannot be estimated.

The second important statistic introduced by Hunter (2007) is the geometrically

weighted dyad-wise shared parter (GWDSP) statistic

$$GWDSP(y) = e^{\alpha_D} \sum_{k=1}^{n-2} \Big(1 - \big(1 - e^{-\alpha_D}\big)^k\Big) DP_k(y). \tag{2.32}$$

Using $DP_k(y)$ this results in a different interpretation than (2.31): (2.32) represents structural holes and the behavior to skip other nodes in the network. This behaviour is the opposite of transitive triad closure so $GWDSP(y)$ is some kind of counterpart to $GWDSP(y)$. Again, the tuning parameter α_D is used to focus on a certain range of shared partners. The same geometric series as in (2.31) is used to prevent model degeneracy. (2.31) and (2.32) may be used together in a social circuit ERGM where α_E and α_D do not have to be equal. Typically, the parameter estimate for $GWESP(y)$ is positive representing the transitive pattern of *'a friend of a friend is a friend'*. The $GWDSP(y)$ parameter is typically negative if a single dense cluster is observed and may be positive if separate clustered cliques of nodes exist. A positive parameter for $GWDSP(y)$ and a negative parameter for $GWESP(y)$ is untypical for friendship networks.

Morris et al. (2008) give an overview of possible network statistics in $s(y)$ which may also include geometrically weighted degree distributions (or geometrically weighted stars) and a variety of networks statistics involving exogenous covariates x. Hunter et al. (2013) show how almost any sufficient ERGM statistic may be constructed and implemented in the `ergm` suite (Hunter et al., 2008a).

2.4.5 Including exogenous covariates

The Markov model is restricted to endogenous sufficient network statistics $s(y)$ which are counts of network subgraph configurations. Wasserman and Pattison (1996) extend the ERGM family to exogenous network statistics $s(y, x)$ which may also be functions of exogenous covariates x. Such network statistics may be used to model dependencies within the graph on nodal and dyadic attributes. Network statistics depending on a function of nodal attributes $b(x_i)$ of actor i are of the form

$$\sum_{i>j} y_{ij} b(x_i) \tag{2.33}$$

which indicates a higher propensity to form ties if the value of $b(x_i)$ is higher and the corresponding parameter is positive. If y is directed, (2.33) may represent a higher popularity or activity with $b(x_i)$ increasing. If the attributes of two actors within a dyad are considered, effects like attribute homophily may be modeled. Dyadic attribute network statistics are of the form

$$\sum_{i>j} y_{ij} b(x_i, x_j). \tag{2.34}$$

The main effect of both attributes is simply

$$b(x_i, x_j) = x_i + x_j.$$

A second-order effect indicating homophily or similarity on the attribute is

$$b(x_i, x_j) = I\{x_i = x_j\}$$

where $I\{x_i = x_j\}$ may by a binary indicator of i and j belonging to the same group. Continuous measures of similarity are also possible. (2.34) may represent a higher propensity to share ties with nodes that are similar on a particular measure or that belong to the same group.

The general ERGM including endogenous and exogenous network statistics takes the form

$$\Pr(Y = y|x, \theta) = \frac{\exp\left\{\sum_{l=1}^{P} \theta_l \cdot s_l(y, x)\right\}}{z(\theta)} \tag{2.35}$$

where $\theta = (\theta_1, \ldots, \theta_P)$ is the vector of parameters assigned to the network statistics $s_1(y, x), \ldots, s_P(y, x)$. The total number of parameters is

$$P = \mathcal{Q} + \mathcal{R}$$

where $\mathcal{Q}$ is the number of sufficient statistics based on exogenous covariates and $\mathcal{R}$ is the number of endogenous subgraph configurations. Throughout this work we will assume the partial conditional dependence assumption of the social circuit model discussed in section 2.4.4. We will use the corresponding set of sufficient network statistics in $s(y, x)$, so (2.35) is a generalization of (2.18).

Models of the form (2.35) are appealing as they are able to incorporate three sources of transitivity in social networks: Hierarchy in the degree distribution due to nodal attributes, social selection processes due to homophily effects and social processes of endogenous network self organization, see Snijders et al. (2006). After including effects based on exogenous covariates, endogenous network statistics like transitive triads can be added evaluating whether there is transitivity beyond social selection and hierarchy at work in the network. In chapter 3 transitivity caused by exogenous covariates and endogenous network self organization will be discussed.

2.5 Maximum likelihood estimation and network simulation

In frequentistic statistics parameter estimation requires the maximization of the likelihood function. The maximum likelihood (ML) estimate $\hat{\theta}$ is the value maximizing the likelihood function. Typically, networks are unique and repeated sampling from a population of graphs is not possible. Therefore the ERGM likelihood is equivalent to (2.18). This is the reason why some authors, e.g. Lusher et al., eds (2012), do not distinguish between the ERGM probability distribution and the ERGM likelihood. Simulating networks is crucial for ML ERGM parameter estimation. As the normalizing constant (2.19) of the ERGM (2.18) is analytically not tractable, MCMC methods are needed for ML parameter estimation which will be discussed in section 2.5.2. This requires efficient MCMC simulation of networks, see section 2.5.1. Given a vector of parameter estimates $\hat{\theta}$ network simulation can further be used for goodness-of-fit evaluation, see section 2.5.3 and chapter 3. Network simulation may further be used to circumvent the evaluation of (2.19) in Bayesian ERGM estimation, see chapter 4, and to explicitly estimate the ERGM normalizing constant (2.19) as discussed in chapter 5.

2.5.1 Simulating Networks

Before discussing ML estimation of the ERGM class a MCMC approach for network simulation has to be introduced. A Metropolis-Hastings (MH) sampler may be used

to simulate a sequence of R networks on n nodes

$$\mathbf{y}_\theta = y^{(1)}, \ldots, y^{(R)}$$

from $p(y|\theta_h, m_h)$ given an ERGM specification m_h represented by a fixed vector of parameters θ_h. This vector may be obtained from ML or Bayesian parameter estimation, see section 2.5.2 and chapter 4. It may also be chosen arbitrarily to compare the influence of M different model specifications $m_1, \ldots, m_M$ as will be illustrated in the simulation example in this section. The MH sampler is initialized with an empty graph. For each iteration r a new network y^* is proposed by randomly selecting a single pair of nodes i and j in the previous state of the chain $y^{(r-1)}$ and toggling it to

$$y^*_{ij} = 1 - y^{(r-1)}_{ij}.$$

The proposed network y^* and the previous draw $y^{(r-1)}$ differ only in the value of the toggled tie y^*_{ij}. The probability of accepting the proposed network y^* is

$$a_{TNT} = \min \left\{ 1, \frac{\Pr(Y = y^*|\theta_h, m_h)}{\Pr(Y = y^{(r-1)}|\theta_h, m_h)} = \frac{\exp\{\theta_h' \cdot s(y^*)\}}{\exp\left\{\theta_h' \cdot s(y^{(r-1)})\right\}} \right\} \tag{2.36}$$

where the normalizing constant $z(\theta_h)$ cancels. It is important to note that simulating from an ERGM is possible without knowing its normalizing constant. This allows for MCMC parameter estimation, see section 2.5.2. Such a MH sampler may require a substantial number of burn-in draws to be discarded before it converges to the distribution of interest $p(y|\theta_h, m_h)$. Everitt (2012) recommends that every tie variable should have the chance to get toggled which requires a burn-in period of at least N iterations. After convergence a set of networks may be sampled using a long enough thinning interval between subsequent draws. So there is no need to restart the algorithm if multiple samples are required.

Selecting a pair of nodes at random may lead to bad mixing of the MH sampler as social networks are typically sparse with a relatively low share of present ties and a high share of empty dyads. Way more toggles from an empty dyad to a new tie would be proposed than vice versa. In order to improve mixing, Morris et al. (2008) propose to select an empty dyad or an existing tie with a probability of 50% which dramatically accelerates the mixing and thus the convergence of the MH sampler. Their approach is called the *"tie-no-tie"* sampler (TNT) and is the default method

of simulating networks from a likelihood in the `ergm` package, see Hunter et al. (2008a). Throughout this work the TNT sampler is used for network simulation. See also Snijders (2002) on MCMC network simulation. There, Gibbs sampling is discussed in order to simulate random graphs by subsequent full conditional draws of single tie variables given the rest of the graph. However, Gibbs sampling for network simulation is computational less efficient than the TNT sampler, see Hunter et al. (2008a), and is usually not applied.

In a small simulation study it shall be illustrated how the TNT sampler works. We mimic the data structure of the well known friendship network of Krackhardt (1987) on managers in a high tech company. A toy network with an reduced set of only $n = 10$ actors with two nodal covariates is generated which are also found in the original study.[6] The nodal attributes are the hierarchical position (three ordered levels) and the department (four unordered categories) a node is affiliated with, see table 2.1.

Table 2.1: Simulated toy networks: Exogenous covariates of simulated networks

node	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
level	1	2	2	2	3	3	3	3	3	3
department	0	1	2	3	1	2	3	1	2	3

Obviously, node 1 is the chief executive officer (CEO) of the company being the only actor with a level value of 1 and working alone in department 0, just like in the original data set. Using these two nodal attributes three undirected dyadic covariates are computed:

1. *Level difference:*
   Absolute value of the difference in hierarchical levels of two nodes ranging from 0 to 2.

2. *Same level:*
   Binary covariate indicating whether two actors have the same hierarchical level.

[6] While the real network on $n = 21$ nodes will be analyzed in chapter 4, a smaller toy network is suited for the purpose of illustration. Also, the TNT sampler is much easier to control for small graphs. Convergence issues and computational time for larger networks in repeated MCMC simulations will be discussed in chapter 5.

Table 2.2: Simulated toy networks: Parameter specifications used for network simulation

$s(y)$	m_1	m_2
edges	-3.0	-3.0
reciprocity	0.5	1.0
GWESP, $\alpha_E = 0.1$	1.0	
same department	2.0	2.0
same level	1.0	1.0
level difference	-1.0	

3. *Same department:*
   Binary covariate indicating whether two actors belong to the same department.

Two specifications are compared with respected to simulated networks they produce, see table 2.2: m_1 uses the three exogenous covariates where the same department and the same hierarchical level contribute positively to friendship tie formation, whereas an increase in level distance shall have a negative effect on friendship. In addition, patterns of endogenous network self organization shall be at work with a tendency to share partners on existing edges. This should result in transitive structures captured by a positive parameter for $GWESP(y)$. With $\alpha_E = 0.1$, connected actors should prefer to have one or two shared partners, but not too many. Further, there should be a tendency to answer once received ties captured by a positive parameter for reciprocity. m_2 shall lack the tendency for transitivity and the effect of level difference. The parameter value for reciprocity is set to 1.0, otherwise the two specifications are identical. $R = 10,000$ networks are simulated from $p(y|\theta_1, m_1)$ and $p(y|\theta_2, m_2)$. The TNT sampler is allowed to burn in for $10,000$ iterations and a thinning interval of $1,000$ draws is chosen between subsequent draws.

The simulated networks are described using three commonly used network statistics: the distribution of shares of in-degrees of all n nodes, the distribution of

shares of out-degrees of all n nodes and the distribution of proportions of edge-wise shared partners $EP_k(y)/L(y)$ on all $L(y)$ edges. These statistics are also the default choice of the `ergm` package by Hunter et al. (2008a) used for goodness-of-fit evaluation (GOF),[7] see section 2.5.3. Figure 2.6 shows the distribution of those three statistics on the simulations from $p(y|\theta_1, m_1)$: the boxplots show the distributions of shares of nodes and edges having degree k of the respective statistic over the $10,000$ simulated networks. The thin grey lines indicate the 2.5% and 97.5% quantiles of the simulations. The distributions of in- and out-degrees look identical in the results for m_1 as neither activity nor popularity are modeled. There is an almost even tendency to have between 0 and 4 friends while having exact 1 friend is less likely. More than 7 friends are never simulated. Specification 1 causes some clustering in the simulated networks, as the median of $EP_1(y)/L(y)$ is 0.5. A share of $EP_1(y)/L(y) < 0.1$ is never simulated so there is always at least some transitivity at work while shares above 0.8 are possible. Edges outside a triangle are relatively rare as the median of $EP_0(y)/L(y)$ is below 0.2. The share of $EP_2(y)$ and $EP_3(y)$ rapidly decreases and $EP_{k\geq 5}(y)$ is never simulated.

The results for m_2 look very different on the distribution of $EP_k(y)/L(y)$. More than two shared partners are not simulated, the median of $EP_1(y)/L(y)$ is below 0.2 and the median of $EP_0(y)/L(y)$ is above 0.8 while networks without a single edge-wise shared partner are possible. The in- and out-degrees look not too different from specification 1 but there is a clear tendency to have only a single friend in the network. The red lines in the figures 2.6 and 2.7 represent the distribution of a single network $y_{typ}|\theta_h$ which is selected to be typical for the particular specification. The network minimizing the sum of mean squared deviations from the three GOF-statistics is chosen. The share of nodes and edges over the degrees of the two typical networks is almost always equal to the respective median values of the simulations. In figure 2.8 the two typical networks are plotted: the network simulated from m_1 shows substantial clustering, most edges are parts of triangles, forming a dense region. The isolated node has to be the CEO as level difference and within-department homophily lead to a very low propensity of tie formation for that particular actor. The network simulated from m_2 looks totally different

[7] Using the default settings, the `gof()` function of the `ergm` package (Hunter et al., 2008a) further computes the distribution of minimum geodesic distances. Throughout this work we refrain from using this statistic as it was of no use to distinguish between model specifications: the simulated distributions always look indistinguishable.

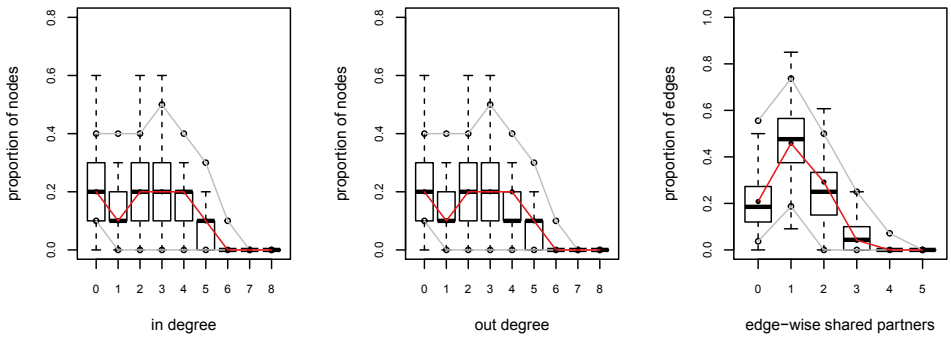

Figure 2.6: Simulated toy networks: Goodness-of-fit plots, m_1

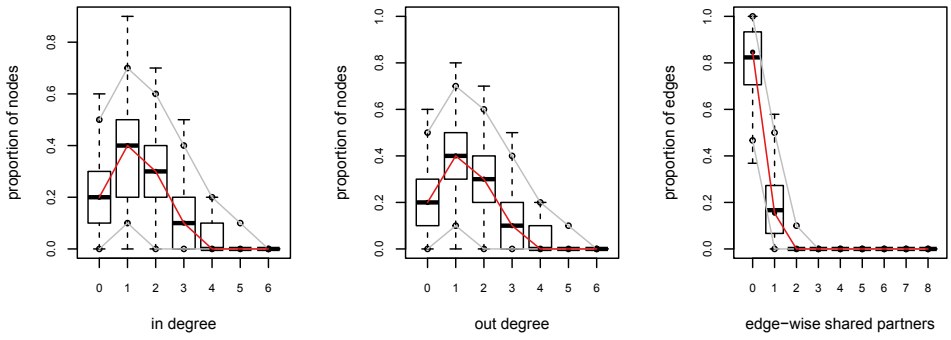

Figure 2.7: Simulated toy networks: Goodness-of-fit plots, m_2

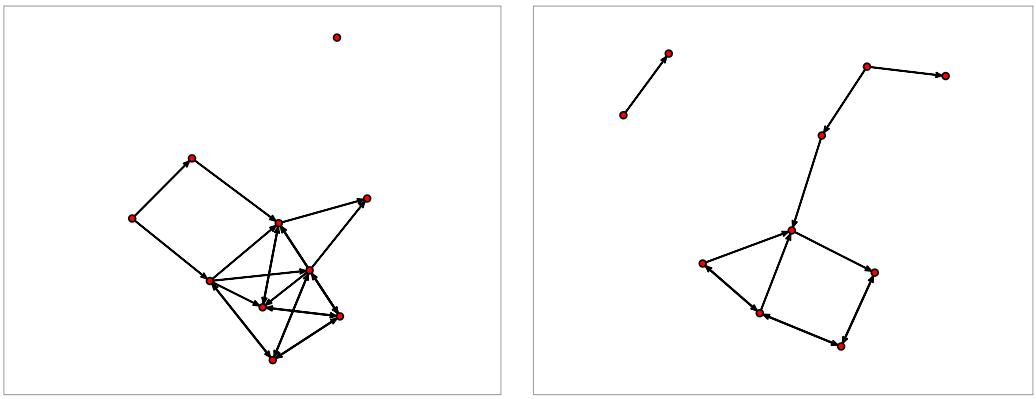

Figure 2.8: Typical simulated toy network for m_1 (left panel) and m_2 (right panel)

containing only a single triangle and many nodes being connected to only one or two others. The comparison of the two plots and the distribution in GOF-statistics highlights the impact of adding $GWESP(y)$ to an ERGM specification. This particular statistic is very powerful in modeling tie variable formation of observed social networks.

Figures A.1 and A.2 show the trace plots, histograms and the autocorrelation (ACF) plots of the simulated network statistics of m_1 and m_2. For both simulations, the ACF is negligible. Considering the recommendations of Everitt (2012) convergence of the TNT sampler may be assumed as the first $10,000 >> N$ iterations are discarded where $N = n^2 - n = 90$. The histograms of the respective sufficient network statistics may also be used to compare the respective typical networks to the mean $\hat{E}\left[s(y|\theta)\right]$. The blue lines indicate the value of the typical network $s(y_{typ}|\theta)$ and the red lines indicate the mean of the simulated networks $\hat{E}\left[s(y|\theta)\right]$. The typical networks are not extreme in their values of the sufficient network statistics. Tables A.1 and A.2 summarize the simulated sufficient network statistics.

2.5.2 Markov chain Monte Carlo maximum likelihood estimation

Frank and Strauss (1986) discover severe difficulties in parameter estimation of models of the form (2.30) which they introduce. They realize that the standard ML approach is not applicable to the ERGM family unless for trivially simple specifications which are of no practical value. The major problem with this model class

is the normalizing constant $z(\theta)$ which cannot be computed analytically. Strauss and Ikeda (1990) introduce pseudo likelihood estimation for the dyad independence model (2.29), while Geyer and Thompson (1992) state that the pseudo ML approach for the ERGM family overestimates the dependence in social networks, see also Handcock (2003b). Snijders et al. (2006) render pseudo ML methods suspect for social network analysis, so generally this approach should be avoided.

As a solution to this problem Geyer and Thompson (1992) introduce a ML approach for intractable likelihoods using MCMC methods to simulate data from the likelihood $p(y|\theta)$ which will be referred to as MCMC-ML. The goal of MCMC-ML is to solve the moment equation

$$\mathrm{E}_\theta\left[s(Y)\right] = s(y) \tag{2.37}$$

where Y is a random graph defined on the space of possible graphs $\mathcal{Y}$ on n nodes, θ is the vector of model parameters to be estimated and $s(y)$ is a vector of sufficient network statistic computed on the observed network y. The ML estimate is the solution to the moment equation for the exponential family, see Lehmann and Casella (1998). The maximum likelihood criterion should lead to parameter estimates $\hat{\theta}$ for which the observed network statistics $s(y)$ have the highes probability so the expected value of sufficient network statistics is equal to the observed value. For theoretical details on ERGM inference and solving the moment equation we refer to Handcock (2003b).

Due to the intractability of the ERGM likelihood the expected value of the sufficient statistics $\mathrm{E}_\theta\left[s(y)\right]$ and the covariance matrix $\Sigma(\theta)$ of the parameter vector θ are analytically not available. No standard Newton-Raphson algorithm can be applied to find an ML estimate $\hat{\theta}$ solving (2.37). Geyer and Thompson (1992) construct a stochastic approximation to the likelihood which is based on importance sampling. It allows for maximization of the likelihood function using an approximative Fisher scoring algorithm. This approach is implemented for ERGM estimation by Hunter and Handcock (2006). Instead of maximizing the likelihood directly the ratio of likelihoods $p(y|\theta)/p(y|\theta_0)$ is maximized where θ_0 is a fixed and known vector of reference parameters. Handcock (2003b) call this the maximization of the relative likelihood. Consider the log ERGM likelihood

$$\mathcal{L}(\theta) = \ln p(y|\theta) = \theta' \cdot s(y) - \ln\left(z(\theta)\right) \tag{2.38}$$

and the log of the likelihood ratio $p(y|\theta)/p(y|\theta_0)$

$$\mathcal{L}(\theta) - \mathcal{L}(\theta_0) = (\theta - \theta_0)' \cdot s(y) - \ln\left(\frac{z(\theta)}{z(\theta_0)}\right). \tag{2.39}$$

The value $\hat{\theta}$ maximizing (2.38) is also the maximizer of (2.39) but the problem of the intractable normalizing constants $z(\theta)$ and $z(\theta_0)$ remains. The ratio $z(\theta)/z(\theta_0)$ may be estimated using

$$\begin{aligned}
\frac{z(\theta)}{z(\theta_0)} &= \frac{\sum_{\tilde{y}\in\mathcal{Y}} q(\tilde{y}|\theta)}{z(\theta_0)} = \sum_{\tilde{y}\in\mathcal{Y}} \frac{q(\tilde{y}|\theta)}{q(\tilde{y}|\theta_0)} \frac{q(\tilde{y}|\theta_0)}{z(\theta_0)} \\
&= \mathrm{E}_{y|\theta_0}\left[\frac{q(y|\theta)}{q(y|\theta_0)}\right] = \mathrm{E}_{y|\theta_0}\left[\frac{\exp\{\theta' \cdot s(y)\}}{\exp\{\theta_0' \cdot s(y)\}}\right] \\
&= \mathrm{E}_{y|\theta_0}\left[\exp\left\{(\theta - \theta_0)' \cdot s(y)\right\}\right]
\end{aligned}$$

where $q(y|\theta) = \exp\{\theta' \cdot s(y)\}$ is the non-normalized likelihood of interest and $\mathrm{E}_{y|\theta_0}$ is the expected value with respect to $p(y|\theta_0)$. The summation over all elements $\tilde{y} \in \mathcal{Y}$ is not possible but a large set of networks may be sampled using MCMC draws from $p(y|\theta_0)$. This yields an approximation to the log likelihood ratio (2.39)

$$\mathcal{L}(\theta) - \mathcal{L}(\theta_0) \approx (\theta - \theta_0)' \cdot s(y) - \ln\left[\frac{1}{R}\sum_{r=1}^{R} \exp\left\{(\theta - \theta_0)' \cdot s(y^{(r)})\right\}\right] \tag{2.40}$$

where random networks $\mathbf{y}_{\theta_0} = y_1, \ldots, y_R$ are simulated from $p(y|\theta_0)$. This can be interpreted as an importance sampling approach estimating $z(\theta)/z(\theta_0)$ using $p(y|\theta_0)$ as importance distribution, see also Everitt (2012). If the reference vector θ_0 was badly chosen, a restart of the process might be necessary using an updated reference parameter vector. Methods for the estimation of normalizing constants based on importance sampling, bridge sampling and path sampling are discussed in chapter 5.

Hunter and Handcock (2006) propose a method which iteratively maximizes the relative log likelihood (2.39). Given the initial reference vector θ_0 an importance sample $\mathbf{y}_{\theta_0}$ is drawn using MCMC methods. An approximate Fisher scoring

algorithm is applied which moves from θ_0 to a value θ^* until

$$\mathrm{E}_{\theta^*}\left[s(y)\right] \approx s(y).$$

The mean and the covariance of the simulated data $\mathbf{y}_{\theta_0}$ are used to approximate the Fisher information matrix, see algorithm 3.4 in Hunter and Handcock (2006), which is based on the algorithm of Geyer and Thompson (1992). If the scoring algorithm converges, θ^* is accepted as the ML estimate $\hat{\theta}$, otherwise the process is restarted with θ^* replacing the reference vector θ_0. Convergence may be assumed at iteration c of the scoring algorithm if

$$\kappa(\theta^{(c)}, \theta^{(c-1)}) = \left[\mathcal{L}(\theta^{(c)}) - \mathcal{L}(\theta_0)\right] - \left[\mathcal{L}(\theta^{(g-1)}) - \mathcal{L}(\theta_0)\right] < \kappa_{min}$$

where κ_{min} is a specified minimal improvement in the log likelihood ratio between subsequent iterations $(c-1)$ and (c).

Algorithm 1: MCMC-ML ERGM estimation

```
Initialize θ_0
Simulate y_{θ_0} = y^(1), ..., y^(R) from p(y|θ_0)
Solve E_{θ*}[s(y)] = s(y) using approximate Fisher scoring: θ_0 → θ*
if κ(θ^(c), θ^(c−1)) < κ_min then
    accept θ* = θ̂.
else
    restart with θ_0 = θ*.
end
```

Algorithm (1) is the default implementation of ERGM estimation in the package `ergm`, see Hunter et al. (2008a). The convergence criterion κ_{min} of the approximate Fisher scoring algorithm and a maximum number of required restarts in the case of non-convergence have to be specified. This is an important feature as a large number of restarts indicates problems with the model specification or badly chosen initial value θ_0.

Geyer and Thompson (1992) warn that the required sample size R might be enormous if θ_0 is far from $\hat{\theta}$ which is a major disadvantage of algorithm (1). Badly chosen values for θ_0 lying in the degenerate area of the specified model may cause non-convergence of both the approximate Fisher scoring algorithm and the TNT

sampler used for network estimation, see 2.5.4. In this case it is not possible to calculate the MCMC-ML estimate even thought the ML estimate may exist. While Hunter et al. (2008a) propose to use a pseudo ML estimate to initialize the MCMC-ML algorithm, Handcock (2003b) warn that there is no guarantee that this approach will result in θ_0 being close enough to $\hat{\theta}$. Despite the popularity of algorithm (1), the serious problem of finding a suitable value for θ_0 has not been solved yet. Snijders (2002) apply MCMC-ML for ERGM estimation using the Robins-Monroe algorithm instead of approximate Fisher scoring. This approach is less efficient but is also less prone to convergence failure due to badly chosen starting values. Handcock (2003a) propose Bayesian model estimation as an alternative which was not implemented until the MCMC methods of Koskinen (2008) and Caimo and Friel (2011), see chapter 4. Indeed, the Bayesian approach is more robust than MCMC-ML as it may converge even with badly chosen initial values and is less sensitive to the problem of model degeneracy discussed in section 2.5.4. The ERGM likelihood cannot be evaluated without an estimate of the intractable normalizing constant $z(\theta)$. Bridge sampling, which is an extension of importance sampling, may be applied to get an estimate $\hat{z}(\theta)$.[8] This approach will be discussed in detail in section 5.3.2. Given a parameter estimate $\hat{\theta}$ the value of the log-likelihood $\mathcal{L}(\hat{\theta})$ can be used to calculate information criteria in order to compare concurring models, see Hunter and Handcock (2006). In section 3 nested models will be compared using such criteria. Section 5.3 will discuss Bayesian non-nested ERGM comparison using the marginal likelihood.

2.5.3 Goodness of fit evaluation

Posterior predictive methods to evaluate the model goodness-of-fit (GOF) are popular for ERGM selection as likelihood based criteria which are not available without estimating the normalizing constant of the likelihood. After obtaining a ML estimate $\hat{\theta}_h$ for a given model specification m_h, the TNT sampler may be used to simulate networks from the likelihood $p(y|\hat{\theta}_h)$. The simulated networks should capture important features of the observed network y. GOF-statistics $s_{GOF}(y)$ have to be chosen according to the features which shall be modeled correctly. The

[8] The `R` (R Development Core Team, 2011) package `ergm` (Hunter et al., 2008a) uses a mapping of multiple bridges to estimate $z(\theta)$, see Hunter and Handcock (2006). In fact this can be interpreted as a form of discretized path sampling, see section 5.2.3, and is similar to the approach introduced by Friel (2013).

choice of $s_{GOF}(y)$ may be different from the sufficient ERGM statistics $s(y)$. If the simulated networks are in line with y, the observed values $s_{GOF}(y)$ should be close to the mean of the simulated data $\hat{E}_{\hat{\theta}_h}[s_{GOF}(y)]$. This may be evaluated using the GOF plots introduced in section 2.5.1. In practical terms the line indicating the distribution of $s_{GOF}(y)$ should be as centered as possible to the boxplots of the distribution of GOF-statistics of the simulated networks. If the respective typical network $y_{typ}|\hat{\theta}_{m_h}$ was indeed an observed network, figures 2.6 and 2.7 would represent almost perfect fit. Typically, the GOF plots of models fit to real data show much larger deviations from the observed network, see chapters 3 and 4 for applications. This approach is suitable to check whether the model is reasonable at all but also requires a decision on which network features should be captured. Any subgraph configuration could be used for model evaluation but there are best-practice recommendations available, see Hunter et al. (2008b). If a model specifications shall be compared, the same set of GOF-statistics has to be used. Throughout this work the in-degree, the out-degree and the EP_k-statistic will be used for model evaluation. These statistics are part of the default GOF-statistics used in the `ergm` package by Hunter et al. (2008a). The EP_k-statistic is of particular importance as it represents the common feature of transitivity in social networks.

Posterior predictive checks are the most prominent method of non-nested ERGM model selection. In chapter 5 a procedure for computing the model evidence in Bayesian ERGM estimation will be discussed which can be used for model selection. However, this method is computationally extremely expensive while GOF simulations are easy to generate.

2.5.4 Model degeneracy

While parameter estimation of the ERGM class is known to be difficult since Frank and Strauss (1986) it was not until the advent of efficient network simulation methods that problems with particular model specifications became obvious. A common problem is that for particular choices of sufficient network statistics algorithms like (1) are not able to find a maximum to the log likelihood ratio (2.39). Snijders (2002) speculate that the convergence issues are an algorithmic shortcoming but Snijders et al. (2006) realize that the problems are caused by some specifications which are intended to model transitivity. If triangles are used as sufficient network

statistics in the ERGM specification, the networks simulated from $p(y|\hat{\theta})$ may not look like the observed network at all even though $\hat{\theta}$ appears to be a reasonable parameter value. Rather the simulated networks are either completely empty with no tie existing or are full graphs with all possible ties present. This is the result of what Handcock (2003b) defines as degeneracy of an ERGM model specification. The expected value $\mathrm{E}_{\hat{\theta}}[s(y)]$ is close to the relative boundary of the convex hull of the set

$$\{s(\tilde{y}) : \tilde{y} \in \mathcal{Y}\},$$

see Rinaldo et al. (2009). As a result, most of the probability of $p(y|\hat{\theta})$ is concentrated on the full graph and the empty graph but little or no probability is placed on realistic configurations similar to the observed network y. MCMC-ML estimation might fail or the estimates might have huge variance. This is the reason why finding a suitable starting value for algorithm (1) is so difficult as a value θ_0 which falls into the degenerate region will cause non-convergence of the approximate Fisher scoring algorithm. Also, the TNT sampler used for network simulation will not converge.

Model degeneracy is related to phase transitions in the Ising model as discussed by Besag (1974) and Frank and Strauss (1986) and becomes apparent in network simulation as a sudden jump in the expected value $\mathrm{E}_{\theta}[s(y)]$ as a function of θ. This function may be continuous but shows a sudden increase in its gradient resulting in a jump right over the value $\hat{\theta}$ which satisfies $\mathrm{E}_{\hat{\theta}}[s(y)] = s(y)$, rendering $\mathrm{E}_{\theta}[s(y)]$ a near discontinuous function. This phase transition occurs at the value $\hat{\theta}$ which otherwise would be accepted as ML estimate. Unfortunately, this estimate produces only nonsensical simulated networks $\mathbf{y}_{\hat{\theta}}$ which are a mixture of full and empty graphs. The result will be a bimodal probability distribution of the network statistic $s(y|\theta)$ for parameters near $\hat{\theta}$. A parameter estimate from that degenerate region on average reproduces the observed value $s(y)$ so that indeed

$$\hat{E}_{\hat{\theta}}[s(y)] = s(y)$$

but both modes of $s(y|\hat{\theta})$ are far from the mean $\hat{E}_{\hat{\theta}}[s(y)]$. The fitted model is not able to reproduce meaningful network data. The TNT sampler used for network simulation will not converge but rather jump between the two modes of $s(y|\hat{\theta})$ while almost no probability mass will fall in between. Schweinberger (2011) give details on ERGM degeneracy and under which model specifications this problem is likely

to occur.

Snijders et al. (2006) hint that the reason for model degeneracy is not an algorithmic problem but is inherent to the ERGM likelihood under the assumption of Markov dependence. They describe an avalanche effect that can occur under model degeneracy if a graph has moderate to low density but shows substantial transitivity which is typical for social networks. If the ERGM shall include a parameter of transitive triangles or $k-$stars, there is a tendency towards huge parameter values for these statistics. The alternating sings of the shared parter statistics discussed in section 2.4.4 are meant to prevent this avalanche effect. Hunter (2007) introduce a model specification similar to Snijders et al. (2006) based on the $GWESP(y)$ and $GWDSP(y)$ statistics discussed in section 2.4.4 which can prevent model degeneracy. These sufficient network statistics require the partial conditional dependence assumption of Pattison and Robins (2002). Further, the assumption of nodal homogeneity may contribute to ERGM instability, see Thiemichen et al. (2016). Handcock (2003b) proposes to ameliorate the algorithmic difficulties caused by near degeneracy with a Bayesian approach using suitable prior distributions on the sample space of θ. Indeed, Bayesian ERGM estimation is more robust to model degeneracy than MCMC-ML and will be discussed in chapter 4.

The phenomenon of model degeneracy is illustrated using simulations of small networks on $n = 10$ nodes. Two sets of networks are simulated from different ERGM specifications m_1 and m_2 both containing $L(y)$ and $GWESP(y)$ as sufficient network statistics. $\theta_L = -2.4$ is fixed in m_1 and θ_{GWESP} is ranging from 0.5 to 1.5 in steps of 0.01. $GWESP(y)$ is specified with $\alpha_E = 1.2$ which is a way too high value for such a small network. In m_2 $\alpha_E = 0.2$ which is a much better value as only few shared partners contribute to the change statistics of $GWESP(y)$. For each step of θ_{GWESP}, $R = 1,000$ networks are simulated and the mean of $GWESP(y)$ is calculated. For $\alpha_E = 1.2$ in m_1 there is strong evidence of model degeneracy as there is a massive jump in the expected value of $\mathrm{E}_\theta\left[GWESP(y)\right]$ around $\theta = 0.68$, see figure 2.9. The resulting probability distribution of $GWESP(y)$ is bimodal where substantial probability mass is placed on the empty graph at $GWESP(y) = 0$ and on a high density graph whereas the density at the mean of $GWESP(y)$ is very low. Obviously $\alpha_E = 1.2$ is too large for this particular network and a social circuit model should be specified that puts more weight on a lower degree of shared partners. If this weight is reduced to $\alpha_E = 0.2$ as in m_2, see figure 2.10, the phase transition disappears and only at the

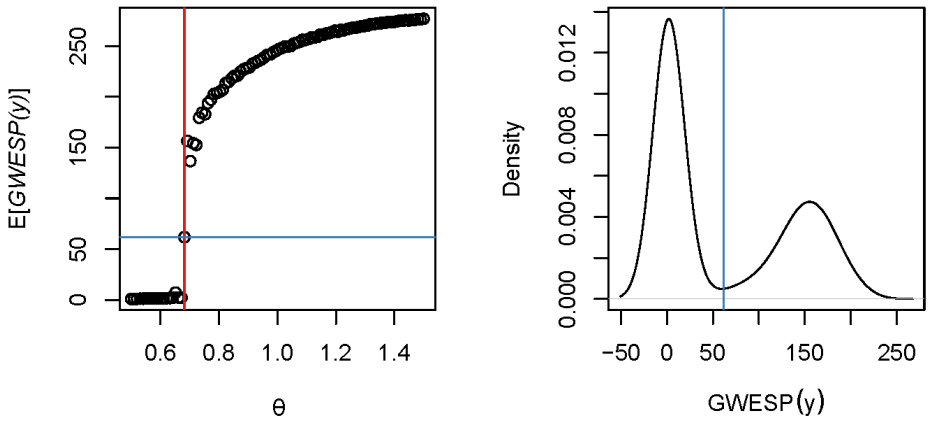

Figure 2.9: Model degeneracy resulting from m_1:
Left panel: The phase transition is obvious around the value of $\theta = 0.68$ indicated by the red line.
Right panel: The resulting distribution of $GWESP(y)$ given that parameter value is bimodal where both modes are far from the mean of the network statistic (blue line). $\alpha_E = 1.2$ causes model degeneracy.

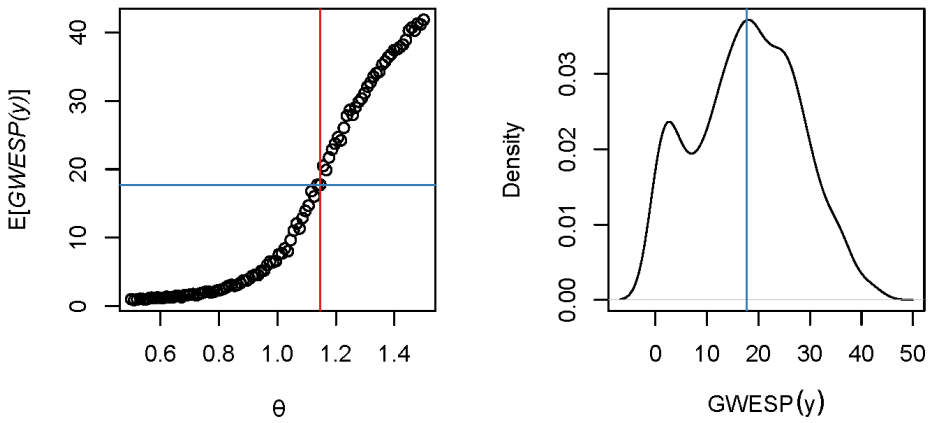

Figure 2.10: No model degeneracy with m_2: *Left panel:* No phase transition observable. *Right panel:* The resulting distribution of $GWESP(y)$ at $\theta = 1.14$ is weakly bimodal with the major mode being identical to the mean. The low value of $\alpha_E = 0.2$ greatly ameliorates model degeneracy compared to $\alpha_E = 1.2$

steepest point of $\mathrm{E}\left[GWESP(y)|\theta_{GWESP}\right]$ the density of $GWESP(y)$ shows some weak bimodality. If the weight is further reduced to $\alpha_E = 0.1$ (not shown), the bimodality completely disappears.

Hunter and Handcock (2006) show how to estimate the tuning parameter α_E in geometrically weighted network statistics. This approach has the drawback of drastically decreasing the numerical stability and speed of MCMC parameter estimation. Throughout this work we specify α_E a priori and keep the value relatively low, see chapter 3 and 4.

2.6 Extensions and modeling alternatives

The ERGM for binary tie variables is not only the most popular model for social network data, it also has by far the most extensions. A major limitation of the binary ERGM is the restriction to dichotomous relations. Krivitsky (2012) introduce the generalized ERGM which allows for the estimation of relational count data. Estimation of network dynamics is possible as well, Hanneke et al. (2010) introduce the discrete temporal ERGM and Snijders and van Duijn (1997) follow an actor oriented approach which is similar to the ERGM but computes sufficient network statistics on the nodal level. Nodes are assumed to decide at discrete time points whether ties should be kept, created or withdrawn. Pattison and Wasserman (1999) extend the ERGM class to multivariate networks on n nodes with u layers. This approach must not be confused with the ERGM for bipartite networks with two sets of nodes where there are links between but not within the sets, see Wang et al. (2009). Think of people in the first set that are members to organizations in the second set. Thiemichen et al. (2016) offer Bayesian ERGM estimation with nodal random effects which does not require the strong assumption of nodal homogeneity. Missing values are not an issue in this work. However, it shall be noted that Wang et al. (2016) introduce multiple imputation of missing tie variables to the ERGM framework. Koskinen et al. (2013) introduce data augmentation for networks with missing tie information and partially observed covariates.

The ERGM is by far the most developed and popular model class for social networks. However, there are alternative approaches available which might have their limitations as well as advantages. Snijders (2011) give an overview of statistical models for network data, including the ERGM. A rather heterogenous class of

models are latent variable approaches. Similar to logistic regression using a probit link function the probability of a binary relation is modeled using a latent continuous variable representing the propensity to form a tie. In order to obtain realistic models assumptions of network interdependency have to be made. Stochastic block models popularized by Nowicki and Snijders (2001) have the interpretation of mixture models for random graphs, see also Daudin et al. (2008). They allow for the modeling of communities and clusters which are important features of social networks. Compared to the ERGM class they are more restricted in modeling complex network interdependencies but are much easier to estimate and interpret. The latent factor model (LFM) introduced by Hoff (2005) and Hoff (2009) is less limited in its capacity to model higher order network dependencies. This model class can be seen as a real alternative to the ERGM. The idea of latent classes and distances between nodes in a latent social space is combined in order to model patterns of reciprocity and network transitivity. The LFM is more generally applicable than the binary ERGM (2.18) discussed in this work as it allows for modeling non-binary relations in y. It can easily be applied to continuous and count data, see Hoff (2005). The nodal homogeneity assumption of the standard ERGM is not needed as nodal random effects are used to model actor heterogeneity. Furthermore, the LFM is extended by Hoff (2011) and Hoff (2015) to model high dimensional networks using a latent tensor normal distribution. E.g. y may be an $u \times n \times n$ array of tie variables representing a u-layered network. It is possible to analyze arrays of even higher dimensionality such as a $t \times u \times n \times n$ dynamic multiple network consisting of u layers sampled at t time points. Even though the LFM class has huge potential to model various types of relational data it is not yet well-known in the field of social network analysis. To our knowledge no systematic comparison of the ERGM and the LFM has been conducted yet.

Chapter 3

Determinants of communication in policy networks in Ghana, Senegal and Uganda

In this chapter exponential random graph model (ERGM) specifications are fit to several agricultural policy networks in Africa. While this method previously was applied to policy networks in industrial countries, see Leifeld and Schneider (2012), this approach is completely new to developing countries. Personal interviews have been conducted with stake holders in Ghana, Senegal and Uganda resulting in social network data representing the communication structure of agricultural policy formulation in these countries. This chapter is the result of a cooperation between the Institute of Agricultural Economics, Christian-Albrechts-Universität Kiel and the Chair of Statistics and Econometrics, Otto-Friedrich-Universität Bamberg.

3.1 Introduction

Communication and participation in policy networks is determined by structural settings and policy preferences of political actors as discussed among others by Carpenter et al. (2004), Adam and Kriesi (2007) or Weible et al. (2010). Using quantitative methods of policy network analysis Henry et al. (2011), Leifeld and Schneider (2012), and Lee et al. (2012) examine the main determinants of political communication and participation in industrialized countries. As countries differ

structurally in the distribution of power and political communication, Adam and Kriesi (2007) point at the importance of the national context when analyzing determinants of political communication. Using the classification of Lijphart (1999), determinants of communication in consensual-federal democracies like the Federal Republic of Germany studied by Leifeld and Schneider (2012) may differ substantially from majoritarian-unitarian democracies like Ghana, Senegal and Uganda studied in this paper. In addition, the higher informal concentration of power in African democracies around the president, see Bratton (2007) and van der Walle (2003), may alter the results on determinants of network tie formation as obtained so far for western democracies. The important role of political actors like non-governmental organizations (NGO) and different framing conditions as implied by dependence on external funding increase the difference in network tie formation between African democracies and developed countries.

Given this, we analyze the determinants of communication in the policy network related to the implementation of the *Comprehensive Africa Agriculture Development Programme* (CAADP) in Ghana, Senegal and Uganda. Initiated by the African Union in 2003, the main goals of the CAADP program are to achieve agricultural growth and poverty reduction through investments in the agricultural sector and harmonization of policy programs. Each government of the three countries implements the approach of CAADP by inviting local stakeholder organizations to design, monitor and evaluate policies. Beyond political actors and donor organizations, the umbrella organizations of the civil society organizations, farmer organizations and private sector organizations signed the national CAADP compacts.[1] Ghana signed the *Medium Term Agriculture Sector Investment Plan* in 2009, Senegal signed the *Programme National d'Investissement Agricole* in 2010 and Uganda signed the *Agricultural Sector Development Strategy and Investment*

[1] The main medium term goals of the CAADP are to achieve agricultural gross domestic product growth and to half the poverty by 2015 compared to 1990 in accordance with the first United Nations Millennium Development Goal. The national CAADP compacts shall offer an efficient communication platform enabling a participatory policy process involving development partners, financial institutions, ministries, governmental agencies and the private sector including poor smallholders. However, a CAADP working group on non state actor participation critically assesses the ability of stakeholders to use the newly created opportunities of participation, see Randall (2011). Using information gathered by a qualitative stakeholder survey and desk research, Randall (2011) point out that CAADP has not consistently achieved high quality inclusion of non-state actors at national, regional and local levels.

Plan in 2010. Several studies document with regard to determinants of political communication, that next to preferences of political actors, see among others Carpenter et al. (2004), Henry et al. (2011) and Sabatier and Weible (2007), structural factors influence the formation of policy network ties, see Lubell et al. (2010). As suggested by Leifeld and Schneider (2012), we assess political communication related to policy formulation via expert knowledge exchange and bargaining for political support among actors. Network data on political communication are collected via face-to-face interviews with the political elite of Ghana, Senegal and Uganda in 2012. We use the ERGM framework to estimate and test the impact of structural variables and actor preferences on political communication in terms of expert knowledge exchange and bargaining for political support. Estimation is performed via a maximum likelihood (ML) approach based on Markov chain Monte Carlo (MCMC) methods, see Hunter and Handcock (2006). In all three countries the perceived power of an actor and existing communication structures are important determinants of political communication. If actors cooperate by exchanging agricultural expert knowledge, they are also more likely to cooperate in policy formulation and vice versa. Political communication occurs in a situation of social control with actors seeking for multiple partners in a trusted environment. Our empirical findings suggest that only in the case of Ghana the government is a particularly important partner of political communication. Furthermore, it is the only country where international donor organizations actively try to influence the search for political support and where commonly attended committees play an important role for expert information exchange. In Ghana a pattern of communication driven by ideological similarities is apparent. In the case of Senegal ideology is important only for communication related to political support. Uganda is the only country where interest groups form coalitions. This article proceeds as follows: Determinants of political communication are discussed in section 3.2. Section 3.3 describes the survey design and empirical data, as well as the network model and the applied estimation approach. Section 3.4 provides the empirical results. Section 3.5 concludes.

3.2 Determinants of communication in policy networks

Various definitions of policy networks are discussed within the literature, see Janning et al. (2009) for an overview. Following Leifeld and Schneider (2012), we distinguish between expert information exchange and the seeking of political support in policy formulation. While the first is the communication in terms of expert knowledge exchange[2] on agricultural processes among actors, the second is lobbying by interest groups and bargaining for political support among actors in order to influence the legislative process. We thus assume two related but distinguishable forms of political communication influencing policy formulation, where each form is captured within a corresponding binary network, and assess the influence of a set of determinants as suggested by theoretical considerations.

Following Carpenter et al. (2004), theoretical explanations with regard to the formation of ties within policy networks can either be characterized as preference-driven or structure-driven. Theories emphasizing the role of preferences focus on similarities and discrepancies of policy beliefs among actors, whereas structural theories emphasize the role of communication choices and social contexts of actors. To accommodate both theoretical approaches, we follow Wasserman and Faust (1994) and Adam and Kriesi (2007) and consider two main categories of determinants of political communication. Firstly, actor attributes capturing similarity in political preferences and individual contexts, and secondly structural network attributes. With regard to similarity of political preferences, several studies argue for the informational role of lobbying, see e.g. Austen-Smith (1993), Ball (1995), and Lohmann (1993). These studies emphasize that lobbying in terms of knowledge exchange may help to choose efficient policies. However, expert information may be costly and not always publicly available. Receiving information from sources with similar interests to oneself lowers the likelihood of receiving information that does not match one's own interests, see Festinger (1954) and Austen-Smith (1993). Accordingly, approaching organizations with similar political interests is rationale as it reduces the financial, emotional and processing costs of political communication.[3] Therefore, the influence of preference similarity is assessed in form of the

[2] An example of expert information is, for instance, the knowledge about the effects of farm input subsidies on the welfare of different social groups.

[3] For experimental evidence on political preference similarity as a driver of tie choice,

following hypothesis.

> **Hypothesis 1a (Process efficiency):** Political communication is more likely between actors with similar political preferences.

Further, the advocacy coalition framework, see Sabatier (1987), Sabatier and Weible (2007), and Weible et al. (2010), assumes that actors have a continuum of political preferences. This continuum of preferences ranges from fundamental beliefs valid for all policy fields to instrumental attitudes necessary to implement specific policy outcomes in a particular policy field, see Janning et al. (2009). While the fundamental beliefs are ideology driven and typically time invariant, the instrumental attitudes are issue specific and open to adaptation. This differentiation of political preferences should thus be included into a model explaining policy network tie formation. Even though an actor perceives only a limited set of other organizations as influential in a policy field, it has to gather information about their political preferences. The more detailed such information needs to be, the more costly is the process of gathering it. Therefore actors seek to minimize information costs by relying on the fundamental policy beliefs of others instead of informing themselves about specific instrumental attitudes, which is captured in the following hypothesis.

> **Hypothesis 1b (Ideology):** Actors seek to minimize the costs of informing themselves about political preferences of others by relying on broad ideological attitudes.

Next to preference similarity Leifeld and Schneider (2012) stress the importance of structural context factors, e.g., commonly attended political committees, see Lubell et al. (2010). Membership in umbrella organizations or common membership in political committees indicates meeting opportunities, increasing the probability that a pair of actors forms a communication tie.

> **Hypothesis 2a (Membership):** Actors seek to minimize transaction costs of political communication by using the meeting opportunities of umbrella organizations.

Leifeld and Schneider (2012) argue further that it is more cost efficient for actors to use existing network ties than to create new ones. Hence, received ties should be preserved and also get answered.

---

see Ahn et al. (2013).

Hypothesis 2b (Reciprocity): Actors tend to answer received communication ties.

As tie formation between organizations is costly, actors seek to form ties with actors they can trust.[4] If *ego* has many shared neighbors with *alter*, then *alter* is more likely to trust *ego* as a sender of high quality information, see Berardo and Scholz (2010) and Carpenter et al. (2004). Following Leifeld and Schneider (2012), policy network tie formation is hence scrutinized by common neighbors resulting in a situation of social control. It is also easier for *alter* to get in contact with other actors if a lot of common neighbours exist, resulting in self strengthening clustered regions in the network as suggested by Henry et al. (2011) and Sabatier and Weible (2007). We expect patterns of transitivity to be at work in policy networks. That is, an organization will seek information from another organization if a third party links them both, see also Holland and Leinhardt (1971) and Berardo and Scholz (2010).

Hypothesis 2c (Clustering and social control): Actors seek for cooperation in a trusted environment.

Even though we assume the process of political communication to be twofold, we are aware that expert information exchange and political bargaining are not conducted by a different set of actors. Similar to Leifeld and Schneider (2012), we assume both fields of communication to be overlapping but distinguishable. If two actors cooperate in policy formulation they are likely to exchange expert information as well since using an already existing tie creates no additional costs.

Hypothesis 2d (Twofold communication): Actors tend to communicate both in expert information exchange and the field of political bargaining.

As stressed by Huckfeldt and Sprague (1995) and Knoke (1996), another important determinant is an agent's power to influence legislation. Given the purpose of lobbying as an interest-mediation mechanism, lobbying organizations contact highly influential actors within the political elite in order to ensure that their members benefit from final policy decisions. In line with Weible and Sabatier (2005), we

[4]Note that common preferences might also determine membership in an umbrella organization and thereby increase the trust an organization has in the information of other organizations with the same memberships.

therefore expect that the higher the perceived influence of an actor, the more likely it is that organizations will send information to this actor. We choose perceived influence for two main reasons. First, we argue in line with Shepsle and Weingast (1987) that consideration of formal political power only would dismiss the informal influence of international organizations in developing countries. Second, we argue that formal political power is usually highly correlated with the perceived influence of actors endowed with formal power. Moreover, employing the concept of perceived influence has the advantage of reflecting both informal and formal political power distributions with one measure. Political power tends to be intensely concentrated around the president, and as a result the cabinet is more powerful in policy-making, see also van der Walle (2003). In particular, we formulate the following two hypotheses.

> **Hypothesis 3a (Influence attribution):** Actors seek to minimize transaction costs of political communication by sending ties to actors they perceive as influential.

> **Hypothesis 3b (Executive):** Members of the government are the most attractive partners in political communication.

International donor organizations play an important role in the policy reform context of developing countries. They aim to influence the process of policy formulation in order to establish effective agricultural policy programs. The more local governmental and non-governmental organizations will feel responsible for the implementation of policy reform the more efficient the policy programs will be. Therefore international donor organizations should try to promote participatory policy-making in order to increase ownership and commitment of local stakeholders.

> **Hypothesis 4 (International organizations):** International donor organizations seek to influence the processes policy formulation.

As tie formation is costly, actors seek to receive high quality information that can be trusted. Organizations with the reputation of being well informed are especially attractive to others. Policy proposals promoted by well informed actors are more likely to find political support. As discussed by Sabatier (1987), scientific organizations are perceived as sources of high quality information which can be

used to strengthen own policy proposals. It can be expected that other actors try to receive expert information from scientific organizations in the process of policy formulation.

> **Hypothesis 5 (Research):** Scientific research organizations are more likely to send expert information.

Non-governmental interest groups should be more likely to form coalitions among themselves than with other types of organizations for two reasons: first, as suggested by Leifeld and Schneider (2012), they are more interested in changing or maintaining the status quo than others and second they should have a special interest in persuading other interest groups of their own preferred policy outcomes, see also Sabatier and Weible (2007). Thus, interest group homophily should be at work in policy networks.

> **Hypothesis 6 (Coalitions):** Non-governmental interest groups seek to form coalitions among themselves.

Given this set of hypotheses on the processes governing political participation in policy networks of developing countries, we will assess them for the case of agricultural policy reform in the three countries of Ghana, Senegal and Uganda.

3.3 Survey design and statistical framework

The units of observation in an elite network study are organizations understood as corporative actors. Respondents are considered as experts of this corporative actor for the specific policy field. The boundaries of an elite communication network must be specified in order to minimize the probability that important players are missing non-randomly. The edges of the network under scrutiny should constitute of all ties relevant to the implementation of CAADP related agricultural investment programs. We use an elite network survey design applied among others by Pappi et al. (1995) for examining policy networks and processes in the USA and Germany. Interviewees were asked to check those organizations on a list compiled in advance with which they maintain a specific relation. Interviewees always have the option to add organizations. This approach tackles the problem of underreporting in a free recall and failures in setting the theoretical network boundaries. We distinguish between two types of political participation as dependent variables,

i.e. expert information exchange and the seeking of political support. To collect data on expert information exchange, interviewees were asked to check those organizations on a list of organizations with which they share information about the consequences of agricultural policies. In particular, expert information transfers have been collected from the suppliers' perspective, e.g. interest groups, and from the consumers' side, e.g. governmental institutions. Therefore, we are able to construct a confirmed and complete expert knowledge network. A knowledge transfer is considered confirmed if both the supplier and demander of knowledge independently report this transfer. To collect the political support network actors were asked which organizations are important for them to formulate policies supported by a majority of voters, while representatives of non-governmental organizations were asked to which political institutions they intermediate their clientele's interests. The corresponding questions from the survey interview are given in appendix B.1.

We identify organizations with formal political power and organizations that have access to formal powerful actors due to their institutional position by desk research. In Ghana, Senegal and Uganda members of the executive, the legislative, local government institutions, and public sector agencies will have formal political power or at least access to members endowed with formal political power. We further include two groups of organizations which may be sources of knowledge for actors involved in or affected by agricultural policy-making. As a first group we consider policy analysts, i.e. donor organizations and research organizations, which provide actors with information that will enable them to choose or to lobby for political strategies compatible with their goals, see also Sabatier and Weible (2007). Non-farmer, farmer and civil society organizations constitute another set of actors with potentially valuable expertise. Relevant actors of these two groups are identified through information from participant lists of policy workshops, official policy documents and web-based member directories of umbrella organizations. We classify organizations according to the categories in table B.1. Our final analysis is based on 46 realized interviews in Ghana and Senegal and 43 realized interviews in Uganda.

We calculate various network statistics corresponding to the hypothesized mechanisms of network tie formation. Table B.2 gives an overview of the network statistics used as model terms and the hypotheses of Section 3.2 they are affiliated with. Following Leifeld and Schneider (2012) the preference similarity between

two actors is operationalized by two different distance measures approximating similarity in policy interests. The first distance is a measure of general preference similarity (prefsim), representing beliefs about broad policy goals. The interviewees could distribute 100 points of relative importance on several ideologically determined policy goals like poverty reduction and gender equity not necessarily related to agricultural policies. These items model an actors fundamental policy beliefs according to the advocacy coalition framework of Sabatier and Weible (2007). The second distance focuses on specific programs of agricultural policy reform and requires knowledge of this concrete policy field. It is thus less ideology driven than the preference similarity measure and models instrumental attitudes. We label it political similarity (polsim). Political similarities of organizations have been calculated on individual relevance attribution to concrete agricultural policy programmes related to the national implementation of CAADP. The interviewees were asked to distribute budget shares to those six programmes proportional to their relative importance.

Further, the perception of an organization's influence in policy-making will influence its probability of receiving ties. Therefore, we use a reputation network for identifying an organization's perceived political influence. Respondents were asked to mark organizations on the list that, according to their opinion, stand out as especially influential. The perceived influence of *ego* is measured by the share of all other actors nominating *ego* as influential.

In order to model opportunity structures for political cooperation, interviewees were asked which political committees they are members of. As suggested by Leifeld and Schneider (2012), we calculate a dyad-specific count variable that indicates how often two organizations are members of the same umbrella organization (membership).

Throughout this chapter the notation for network data and network statistics introduced in chapter 2 are used. Let y denote a $n \times n$ directed adjacency matrix on a set of n nodes. Y is a random graph consisting of $N = n(n-1)$ directed random tie variables. A random tie variable $Y_{ij} = 1$ if actor i sends a directed tie to actor j, $Y_{ij} = 0$ else. y_{ij} is an observed tie (edge). As y is a digraph, $y_{ij} \neq y_{ji}$, resulting in an asymmetric adjacency matrix. Self ties are not permitted, so the diagonal of y is always empty. $\mathcal{Y}$ is the set of all possible graphs on a fixed set of n nodes. Further, let x be an $n \times n \times \mathcal{Q}$ array of exogenous covariates like the preference similarity of two nodes (a dyadic attribute) or the type of an organization (a nodal

attribute).

The endogenous network configurations are micro sub graphs a network graph could be constructed with and can be used as a control for the effect of exogenous factors. They explain the internal self-organizing structure of the dependent network variable. Counts of edges $L(y)$ are used to model the general propensity of tie formation. The number of reciprocal edges $M(y)$ can be used to measure the tendency to answer received ties. The geometrically weighted edge-wise shared partner statistic $GWESP(y)$ and the geometrically weighted dyad wise shared partner statistic $GWDSP(y)$ are used to capture effects of transitivity, see section 2.2. These statistics can be formulated as

$$GWESP(y) = e^{\alpha_E} \sum_{k=1}^{n-2} \Big(1 - \big(1 - e^{-\alpha_E}\big)^k\Big) EP_k(y), \tag{3.1}$$

$$GWDSP(y) = e^{\alpha_D} \sum_{k=1}^{n-2} \Big(1 - \big(1 - e^{-\alpha_D}\big)^k\Big) DP_k(y), \tag{3.2}$$

where the k edge-wise shared partners statistic $EP_k(y)$ is the number of directed edges that are the base for k transitive triads. Therefore the tie y_{ij} must exist and the connected nodes i, j must have k shared partners. It can be imagined as stacking k transitive triads having the base edge y_{ij} in common. The dyad-wise shared partner statistic $DP_k(y)$ is the count of pairs of nodes i, j that share k partners but unlike to $EP_k(y)$ the dyad $d(i, j)$ may be empty. $GWESP(y)$ represents multiple triangulation and the propensity to form closed clustered structures, contrasted by $GWDSP(y)$ representing multiple independent 2-paths. A positive $GWESP(y)$ and a negative $GWDSP(y)$ parameter can be interpreted as a propensity to avoid structural holes like 4-cycles with no diagonal ties, see Lee et al. (2012). The two shared partner statistics use a geometric series $(1 - e^{\alpha_E})^k$ and $(1 - e^{\alpha_D})^k$ in order to model the distribution of shared partners relevant for tie formation, see section 2.5.4 for an interpretation. α_D and α_E are tuning parameters weighting the number of shared partners, see Hunter and Handcock (2006), Hunter (2007) and section 2.5.4. For our analysis the tuning parameters are both chosen to be fixed at relatively low baseline values[5] of $\alpha_D = \alpha_E = 0.1$. These low values facilitate model estimation but risks the underestimation of the relevance

[5]The parametrization for $GWESP(y)$ and $GWDSP(y)$ is taken from Leifeld and Schneider (2012).

of configurations with many shared partners.[6] We illustrate the $GWESP(y)$ and $GWDSP(y)$ statistics in section 2.4.4. For more details on network statistics for the analysis of policy networks, see Robins et al. (2012).

As social networks typically show patterns of tie variable interdependence like reciprocity or triangulation, these features have to be considered during model formulation. Sending or receiving network ties happens within a social context which renders Bernoulli graphs assuming independence of tie variables as introduced by Erdös and Renyi (1959) a rather unrealistic choice for modeling social behavior. A well established model class for social networks is the ERGM framework developed by Wasserman and Pattison (1996) and modified by Snijders et al. (2006). Lusher et al., eds (2012) illustrate a wide range of ERGM applications, giving also a detailed introduction to ERGM theory. This model class can represent the structure and the driving factors of a network by using an *a priori* defined set of sufficient network statistics. These network statistics are subgraphs representing particular patterns of social behavior and thus allow for the modeling of the endogenous self organization of a network. The ERGM class can also represent the influence of exogenous covariates on network tie formation, see section 2.4.5.

The ERGM probability distribution can be formulated as

$$\Pr(Y = y|x, \theta) = \frac{\exp\left\{\sum_{l=1}^{P} \theta_l \cdot s_l(x, y)\right\}}{z(\theta)}. \tag{3.3}$$

$s(x, y)' = (s_1(x, y), \ldots, s_P(x, y))'$ is a vector of $P = \mathcal{R} + \mathcal{Q}$ observed sufficient network statistics which may contain $\mathcal{R}$ endogenous configurations of network self organization and $\mathcal{Q}$ exogenous covariates. The $\mathcal{R}$ endogenous sufficient statistics are network counts for directed subgraph configurations, e.g. $GWESP(y)$, $GWDSP(y)$ or k-star configurations. $\theta' = (\theta_1, \ldots, \theta_P)$ is a vector of P model parameters. Each θ_l corresponds to a network statistic $s_l(x, y)$. The normalizing constant $z(\theta) = \sum_{\tilde{y} \in \mathcal{Y}} \exp\{\theta' \cdot s(x, \tilde{y})\}$ ensures that Equation 3.3 is a probability distribution and requires summation over all possible network realizations. The most appropriate *a priori* set of sufficient statistics has to be chosen before an ERGM can be estimated. Such a particular choice depends on the research que-

[6] The tuning parameters can be fixed or may be a free parameter to be estimated. Such a parametrization can be analyzed using a curved ERGM, see Hunter and Handcock (2006), complicating parameter estimation by increasing the risk of non-convergence of the MCMC-ML algorithm.

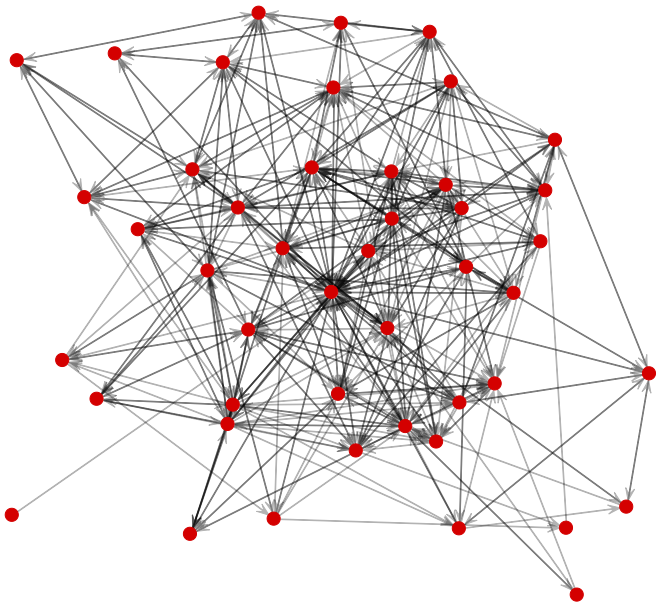

Figure 3.1: Plot of the expert network in Ghana on $n = 46$ nodes. Due to the relatively high density of the network, it is hard to detect any striking pattern of network tie formation with the naked eye.

stion and the underlying hypotheses on network tie formation. Details on the ERGM class are given in in section 2.3. The interpretation of parameter values in θ_l are given in section 2.4.

Due to the enormous number of possible realizations in $\mathcal{Y}$ the normalizing constant is intractable even for networks of moderate size. This makes parameter estimation difficult within the ERGM framework. The analytical evaluation of the normalizing constant can be circumvented by using a simulation based Markov chain Monte Carlo maximum likelihood (MCMC-ML) approach, see Snijders (2002) and Hunter and Handcock (2006). Random graphs are sampled in order to approximate the likelihood function, obtaining an ML estimate of the model parameters $\hat{\theta}$ by maximizing the simulated likelihood. MCMC-ML estimation of the ERGM family is a computational intensive task frequently complicated by non-convergence. Details on the MCMC-ML approach used in this chapter are given in section 2.5.2.

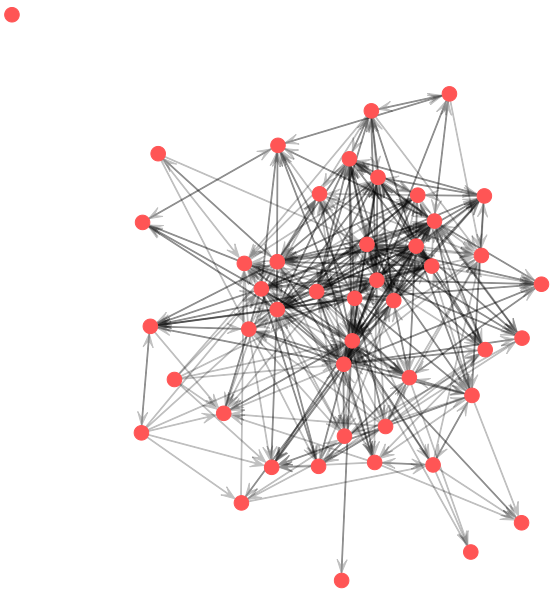

Figure 3.2: Plot of the support network in Ghana on $n = 46$ nodes.

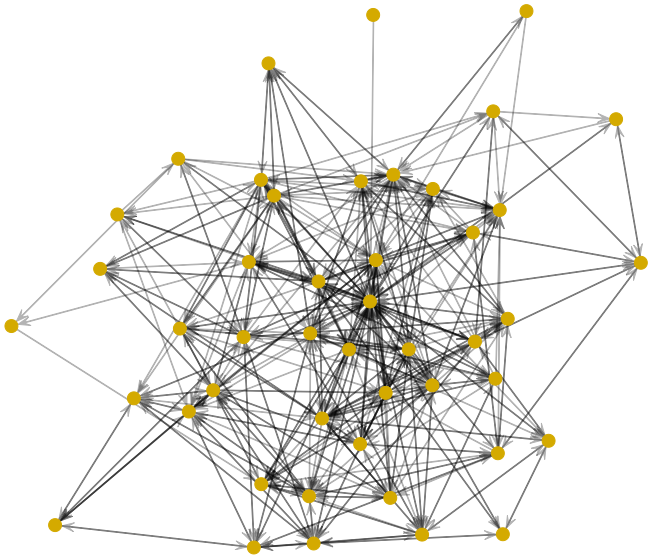

Figure 3.3: Plot of the expert network in Senegal on $n = 46$ nodes.

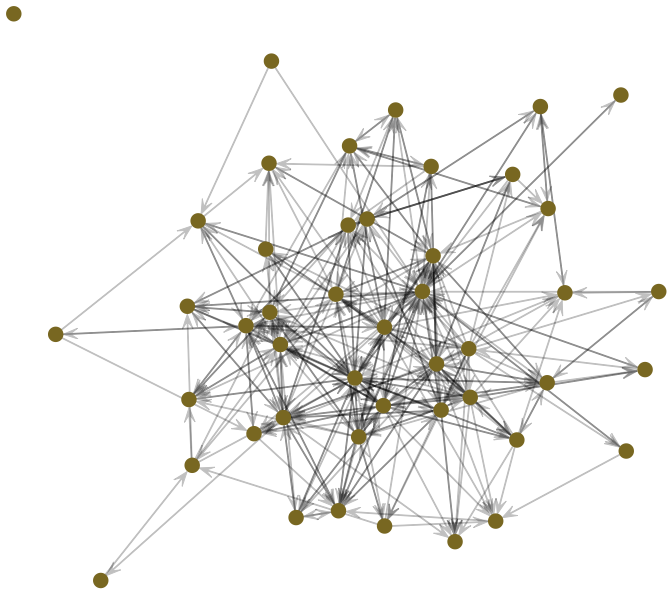

Figure 3.4: Plot of the support network in Senegal on $n = 46$ nodes.

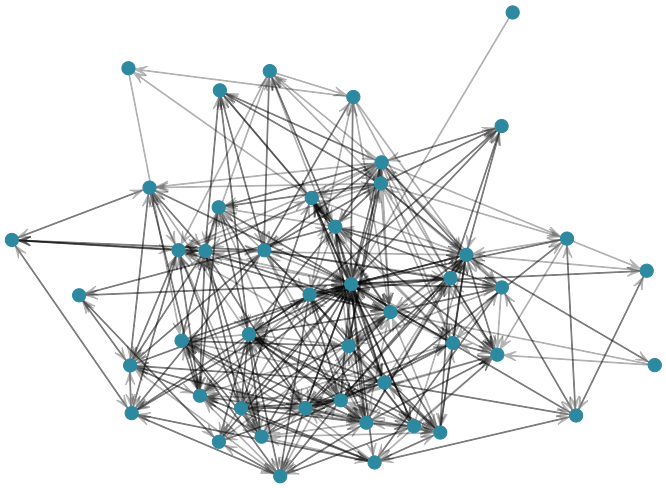

Figure 3.5: Plot of the expert network in Uganda on $n = 43$ nodes.

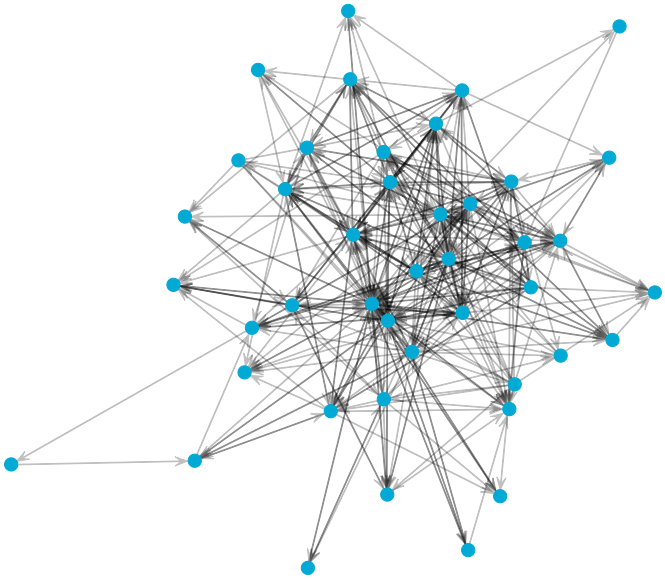

Figure 3.6: Plot of the support network in Uganda on $n = 43$ nodes.

3.4 Empirical analysis

Figures 3.1 to 3.6 show plots of the six realized networks.[7] All networks are very dense: the only striking pattern that can be seen with the naked eye is that each of them forms a single dense cluster and that there are few central nodes having a much higher degree than other nodes. The ERGM framework shall be used to test the hypotheses formulated in section 3.3 on the determinants of political communication in Ghana, Senegal and Uganda. For each dependent variable, the expert and the support network respectively, two models are estimated: an endogenous model, containing only endogenous network statistics, and a structural model, containing the same endogenous statistics plus exogenous covariates as a control.

The exogenous covariates used for our analysis contain nodal and dyadic attributes. The expert structural model contains the support network as explanatory dyadic attribute and vice versa. The membership variable is a dyadic attribute counting how many attended committees two actors have in common. An organization's global reputation is calculated as an actor attribute using a network of perceived importance. It is counted how many of the $n - 1$ other actors re-

[7] The force-directed placement algorithm of Fruchterman and Reingold (1991) is used to create the layouts, default specifications of the `R` (R Development Core Team, 2011) package `statnet` (Handcock et al., 2008) package are used.

port node *ego* as an influential actor. The governmental popularity configuration measures whether organizations belonging to the executive receive more ties than other organizations, whereas the two activity parameters measure whether donor and research organizations send more ties than others. The political similarity and preference similarity variables are based on Euclidean distances between two nodes. Political similarity represents similar instrumental attitudes on a set of concrete policy programmes, preference similarity represents fundamental policy beliefs like protecting the environment and women's rights.[8] The interest group homophily configuration measures whether interest groups are more likely to form ties with each other than with other types of organizations. Summary statistics of the six communication networks and the corresponding covariates can be found in table B.2.

MCMC-ML ERGM estimation is done using the `R` (R Development Core Team, 2011) package `ergm` (Hunter et al., 2008b) included in the `statnet` environment, see Handcock et al. (2008) which is an implementation of algorithm (1), see section 2.5.2. The TNT sampler generates $R = 10,000$ effective MCMC network simulations using a thinning interval of $R_{thin} = 100$ simulations after discarding the first $R_{burn} = 1024 \cdot 16$ networks (`ergm` package default). The algorithm is allowed to restart up to 1,000 times if non-convergence of the approximate Fisher scoring algorithm is detected. Otherwise default settings of the `ergm` package are used.

It must be noted that the application of the MCMC-ML approach of Hunter and Handcock (2006) to our data is plagued by numerical instability. The algorithm is very sensitive to the initial parameter value θ_0, see section 2.5.2. The same model specification is used for all three countries which requires a substantial number of restarts for some networks. Also, the algorithm does not converge for all specified seed values. The weights of $GWESP(y)$ and $GEDSP(y)$ have to be lowered in the support network of Ghana in order to achieve convergence. Finding model specifications for which the MCMC-ML algorithm converges is not easy an requires a lot of tweaking which is a well known problem with this approach. An alternative to MCMC-ML is Bayesian ERGM estimation introduced by Caimo and Friel (2011) which will be discussed in chapter 4.

The baseline parameter specification is based on Leifeld and Schneider (2012). Similar results (not shown) are achieved with a Bayesian approach of ERGM esti-

[8] On the calculation of those nodal similarities, see Leifeld and Schneider (2012).

mation using the `R` (R Development Core Team, 2011) package `bergm`, see Caimo and Friel (2014). The coefficients of Model 1 to Model 4 in table B.3 may be interpreted as change in the conditional log odds of a tie present if, *ceteris paribus*, the change statistic of the respective configuration is increased by one. The pattern of endogenous self organization is the same across all countries and networks. The positive parameters of the mutuality statistics $M(y)$ indicate a certain willingness to answer ties once received, supporting hypothesis 2b. The networks are governed by clustered structures as the $GWESP(y)$ parameters are always positive. The $GWDSP(y)$ parameters are always negative, so tie formation shows a clear pattern of transitivity without creating separate, bridged regions. There is no tendency to skip other actors or to address only a few hub actors. Political communication takes place within closely connected groups involving many organizations, integrating many shared partners. The overall pattern of the data seems to speak in favor of Hypothesis 2c. This pattern is persistent even if controlled for exogenous covariates as the level of significance and the parameter signs of the endogenous statistics generally stay the same. Model 4 in Ghana shows an insignificant parameter for the $GWESP(y)$ statistic indicating that clustering is completely explained by exogenous covariates.

Adding exogenous covariates always improves the model fit indicated by larger values of the log-likelihood and smaller values for the Akaike information criterion and the Bayesian information criterion, see table B.3. An actor's reputation is an important exogenous factor of network tie formation across all countries, increasing the probability of a sent tie if *ego* perceives *alter* as influential, so Hypothesis 3a can be supported. The positive parameters for the support dyadic covariate (Model 2) and for the expert dyadic covariate (Model 4) are in line with hypothesis 2d: actors tend to cooperate in both fields of political participation using existing communication ties. The activity parameter for research organizations is never significant and positive so there is no evidence that research organizations are consulted more frequently than others during expert information exchange or actively seek to influence the bargaining for political support. Hypothesis 5 must be disproved.

In Ghana common membership in committees has a stake only in the exchange of expert information. Attendance to an additional committee increases the odds

of expert information exchange by the factor

$$\exp\{1.065\} = 2.900.$$

As Hypothesis 2a postulates, actors seek to minimize transaction cost of political communication by using umbrella organizations, but only in the course of policy information exchange. Hypothesis 2a cannot be supported in the other countries. Ghana's executive is more popular than other types of organizations: the probability of receiving a tie is increased significantly if an actor belongs to the government. This supports hypothesis 3b in Ghana but again not in the other two countries. Donor organizations in Ghana are less likely to exchange expert information than other organizations, indicated by a negative activity parameter. The odds of expert information exchange are decreased by the factor

$$\exp\{-0.316\} = 0.729$$

if *ego* is a donor organization. For the support network the image is reversed: the odds of sending a tie is increased by the factor

$$\exp\{0.928\} = 2.529$$

if *ego* is a donor organization. In Ghana, donor organizations are less dedicated to expert information exchange but try to influence the process of policy implementation. Hypothesis 4 cannot be supported generally, in the other two countries the activity parameter for donor organizations is never significant and positive. In Ghana similarity in instrumental attitudes is never significant. However, actors look for cooperation with partners showing the same fundamental policy beliefs, indicated by positive and significant parameters for preference similarity in Model 2 and Model 4. E.g. the odds of an expert information tie in Ghana is increased by the factor 3.208 if the preference similarity between two actors is increased by 10%. Instead of reading up on detailed policy concepts of other actors, organizations are satisfied knowing the ideological preferences of other actors which is more cost efficient. In Senegal ideological preference similarity is important for the support network but not for the export network. Beyond that, only the significant pattern all countries have in common is observed. The data underpin Hypotheses 1a and 1b for Ghana and partially for Senegal. In Uganda again only the mechanisms

which all countries have in common are apparent, with the exception that interest groups seem to form coalitions to find political support. Hypothesis 6 is supported only in Uganda.

The TNT sampler used in the MCMC-ML algorithm can also be used for model evaluation, see section 2.5.1. Repeated simulations of random graphs based on previously obtained parameter estimates are compared to the observed network by evaluating goodness-of-fit statistics. Figures B.1 to B.6 in appendix B.3 show the distribution of goodness-of-fit statistics for the estimated endogenous (upper row) and exogenous (lower row) models. The thick black lines describe the observed distributions of in-degrees, out-degrees and edge-wise shared partners EP_k, each plotted over a simulation of 100 random graphs. Details on these plots can be found in section 2.5.3. It can be seen that the model fit improves if exogenous network statistics are added. The plots in the lower panel always show better goodness-of-fit than those in the upper panel. Consider e.g. figure B.2, upper row: the endogenous model slightly underestimates the occurrence $EP_k(y), \quad k > 8$ but clearly overestimates $EP_k(y), \quad k = 3$. The model containing exogenous network statistics predicts the distributions of edge-wise shared partners correctly. The exogenous covariates generally help to explain higher degrees of triangulation and the high popularity and activity of certain nodes, improving the GOF compared to the endogenous models for all countries and networks.

3.5 Conclusion

In summary, the proposed framework reflects the policy process as a country-specific mechanism, embedded in a particular structural setting, aggregating policy preferences of divergent actors to a policy decision. It is capable of considering the influence of actors with vested interests in the specific policy domain that are not endowed with formal political power by constitution. Social network analysis using the ERGM enables us to include and test many influence factors from several strands of theories. Based on this framework, we take some first steps towards describing the agricultural policy landscape in Ghana, Senegal and Uganda. An actor's power perceived by others is a crucial driving force of network tie formation. Preference similarity in fundamental policy beliefs can be important but similarity in specific policy programs is not. Opportunity structures matter as poli-

tical information exchange and political support do coincide. Common attendance to political committees is only relevant for expert information exchange in Ghana. There is no clear evidence that members of the executive and international donor organizations are more important than other actors. Research organizations seem to behave passively and do not seek to deliver high quality information to other actors. There is no clear evidence that interest groups build coalitions. The network shows significant endogenous self organization even if the models are controlled for exogenous covariates. Actors tend to cooperate under social control including many common neighbours, not trying to skip others. Scrutinizing each other makes tie formation more reliable and generates social trust. Even though not all of the hypothesized mechanisms seem to be at work, the models provide reasonable fit to the observed data and thus can explain the processes of agricultural policy formulation and implementation in the three countries.

CAADP has been criticized for the inefficient communication between state and non-state actors at the national level. Quite often non-state actors are not even aware of the opportunity structures of political communication the CAADP shall offer, see Randall (2011). In our data there is no evidence that actors are well informed as they do not consult scientific organizations and do not seem to know about the issue-specific preferences of each other. Non-state interest groups are generally not well organized and umbrella organizations are not used as a platform for efficient communication. International donor organizations are not able to influence the process of political communication and non-state actors do not seem do be able to easily address the government.

MCMC-ML estimation for the ERGM is heavily plagued by convergence issues due to starting values of algorithm (1) which are hard to choose. The default method of the `ergm` package by Hunter et al. (2008b) is to initialize θ_0 with the pseudo ML estimate of the parameter vector. If this estimate lies within the degenerate region of the model, it may happen that the approximate Fisher scoring algorithm will never converge, no matter how many restarts of algorithm (1) are allowed. Finally, MCMC-ML estimation is possible for our data but required tedious fine tuning like trying several seed values. Furthermore, not all plausible specifications are estimable. E.g. m_1 for the expert network in Ghana converges but if the $GWDSP(y)$ statistic is removed from $s(y)$ MCMC-ML estimation is not possible anymore: even though the number of parameters is reduced, it is not possible to find starting values that let the MCMC-ML algorithm converge. A solution to this

problem is Bayesian ERGM estimation discussed in the next chapter which is much more robust to model degeneracy.

Chapter 4

Bayesian exponential random graph model estimation

In this chapter Bayesian model estimation of the exponential random graph model (ERGM) using Markov chain Monte Carlo methods (MCMC) based on the Metropolis-Hastings (MH) sampler is discussed. No conjugate prior distribution is known to the intractable ERGM likelihood and sampling from full conditional distributions using a Gibbs sampler is not an option. In addition, as the ERGM normalizing constant is analytically not available, a standard MH algorithm cannot be applied. A solution to this problem is the exchange algorithm (EA) introduced by Murray et al. (2006). This approach relies on the simulation of auxiliary network data in order to evaluate the analytically intractable ERGM likelihood. In section 4.1 a short introduction to Bayesian inference and the MH algorithm is given. The EA is discussed in section 4.2 and in section 4.3 it is shown how adaptive direction sampling introduced by Gilks et al. (1994) can help to make the EA more efficient. In section 4.4 the approach is applied by fitting ERGM specifications to a well studied network of friendship relations among managers in a high-tech company, see Krackhardt (1987), and the expert network in Ghana, see section 3. Section 4.5 gives a summary.

4.1 Bayesian inference

Bayesian inference is a process of inductively learning from data using Bayes' rule. Before obtaining data y, *a priori* beliefs about a population characteristic θ are represented by the prior distribution $p(\theta)$. The model (or data generating process) represents the beliefs of observing y if θ was the population characteristic captured in the likelihood function $p(y|\theta)$. After observing the data the *a priori* beliefs are updated obtaining the *a posteriori* beliefs about θ which are represented by the posterior distribution $p(\theta|y)$. This updating step from prior beliefs under a particular model to posterior beliefs uses Bayes' rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4.1}$$

where

$$p(y) = \int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta} \tag{4.2}$$

is the normalizing constant of (4.1) which requires integration over all possible values $\tilde{\theta} \in \Theta$. In most cases (4.2) is not available as it requires integration over a high dimensional parameter space Θ. MCMC methods are needed to summarize (4.1) and circumvent the explicit evaluation of (4.2). This allows for Bayesian inference which typically results in an estimate of θ and a respective posterior highest density region (HDR) of $p(\theta|y)$. $p(y)$ is called the marginal likelihood. It has the interpretation of a normalizing constant insuring that the posterior is a proper probability distribution if data are available. But it is also the marginal distribution of the data and has the role of an *a priori* predictive distribution of y before actual data are available. Furthermore, (4.2) can be used for Bayesian model selection discussed in chapter 5.

MCMC algorithms for Bayesian inference have to sample parameter values from

$$p(\theta|y) \propto p(\theta)\ q(y|\theta)/z(\theta) \tag{4.3}$$

where $q(y|\theta)$ is the non-normalized kernel of the likelihood function and $z(\theta)$ is the corresponding normalizing constant of the likelihood. (4.3) is known only up to its normalizing constant $p(y)$. A Markov chain has to be constructed which, after initialization, converges towards a stationary distribution that is equal to

$p(\theta|y)$ if run long enough. Stochastic simulation using MCMC techniques generates autocorrelated samples from the posterior. The simulations needed before the chain has converged have to be discarded and a sufficient number of samples has to be drawn to give a good summary of the posterior (4.1). Thus evaluation of convergence and autocorrelations is crucial in Bayesian inference based on MCMC techniques.

If such techniques shall be applied to ERGM estimation, additional difficulties arise. Full conditional distributions of the ERGM model parameters are not known so Gibbs sampling is not an option, see Everitt (2012). In such a case the MH algorithm, introduced by Hastings (1970), could be used to sample from the posterior by proposing a new vector θ^* drawn from a proposal distribution $\mathcal{H}(\theta^*|\theta)$. This distribution requires considerate specification by the user in order to avoid low acceptance rates and excessive autocorrelation of the MCMC samples. While the MH algorithm is widely applicable, finding a suitable proposal distribution can be very challenging. The proposals generated by $\mathcal{H}(\theta^*|\theta)$ are accepted with the probability

$$a_{MH} = \min\left\{1, \frac{\mathcal{H}(\theta|\theta^*)}{\mathcal{H}(\theta^*|\theta)} \frac{p(\theta^*)}{p(\theta)} \frac{q(y|\theta^*)}{q(y|\theta)} \frac{z(\theta)}{z(\theta^*)}\right\}. \tag{4.4}$$

Details on the MH sampler can be found in Chib and Greenberg (1995) and Chib and Jeliazkov (2001). For the ERGM class the ratio of normalizing likelihood constants $z(\theta)/z(\theta^*)$ is not available, so the standard MH algorithm cannot be applied.

4.2 The exchange algorithm

Møller et al. (2006) avoid the evaluation of the intractable normalizing constants $z(\theta)$ and $z(\theta^*)$ in (4.4) by using auxiliary data y^* simulated from $p(y^*|\theta^*)$ with $\{y, y^*\} \in \mathcal{Y}$. Their approach is called the single auxiliary variable MH algorithm (SAV) and is the first MCMC method to be known to sample from the correct posterior if the likelihood has an intractable normalizing constant, see also Everitt (2012). Koskinen (2008) apply the SAV to ERGM estimation which was the first fully Bayesian approach for this model class.

Murray et al. (2006) develop the so called exchange algorithm (EA) which is similar to the SAV but more efficient. The intractable ratio $z(\theta)/z(\theta^*)$ in (4.4) is

estimated using importance sampling similar to the MCMC-ML approach discussed in section 2.5.2. The main idea behind the EA is to propose a parameter vector bringing its own simulated data generated from the same intractable likelihood as the observed data. Given a parameter proposal θ^* drawn from $\mathcal{H}(\theta^*|\theta)$, auxiliary data y^* are simulated from $p(y^*|\theta^*)$ with $\{y, y^*\} \in \mathcal{Y}$ where the functional form of $p(y^*|\theta^*)$ is identical to the likelihood $p(y|\theta)$. The EA samples from the augmented posterior

$$p(\theta^*, y^*, \theta|y) \propto p(y|\theta)p(\theta)\mathcal{H}(\theta^*|\theta)p(y^*|\theta^*). \tag{4.5}$$

$\mathcal{H}(\theta^*|\theta)$ may be an arbitrary proposal distribution, e.g. a symmetric random walk distribution centered at θ. As with the standard MH algorithm tuning of $\mathcal{H}(\theta^*|\theta)$ is required in order to efficiently sample θ from the posterior. Suitable choices for $\mathcal{H}(\theta^*|\theta)$ will be discussed in section 4.3.

The probability of moving from a current value θ to the new proposed value θ^* in the EA can be formulated as

$$a_{EA} = \min\left\{1, \frac{\mathcal{H}(\theta|\theta^*)}{\mathcal{H}(\theta^*|\theta)}\frac{p(\theta^*)}{p(\theta)}\frac{q(y|\theta^*)}{q(y|\theta)}\frac{q(y^*|\theta)}{q(y^*|\theta^*)} \times \frac{z(\theta)z(\theta^*)}{z(\theta)z(\theta^*)}\right\} \tag{4.6}$$

where the ratio of unavailable normalizing constants in the standard MH ratio (4.4) is replaced with $q(y^*|\theta)/q(y^*|\theta^*)$. Note that the normalizing constants $z(\theta)$ and $z(\theta^*)$ cancel so (4.6) uses ratios of non-normalized likelihoods $q(y|\cdot)$ and $q(y^*|\cdot)$. Murray et al. (2006) refer to an exchange move if θ^* is accepted as $q(y^*|\theta)/q(y|\theta)$ measures the "affinity" between θ and the auxiliary data y^* and $q(y|\theta^*)/q(y^*|\theta)$ measures the "affinity" between the proposed θ^* and the observed data y, see also Caimo and Friel (2011). Murray (2007) proof that the EA has in fact the posterior of interest as invariant distribution. Everitt (2012) show that

$$\mathrm{E}_{y^*|\theta^*}\left[\frac{q(y^*|\theta)}{q(y^*|\theta^*)}\right] = \frac{z(\theta)}{z(\theta^*)}, \tag{4.7}$$

thus the ratio of non-normalized likelihoods is an unbiased estimator of the ratio of normalizing constants. The ratio $q(y^*|\theta)/q(y^*|\theta^*)$ can be interpreted as single point importance sampling estimate of $z(\theta)/z(\theta^*)$ which may be improved using a large number of R importance draws, see Alquier et al. (2016). If $R \to \infty$ the EA is equivalent to the MH.

As the EA uses (4.7) to directly estimate the ratio of normalizing constants in contrast to the SAV which estimates this ratio indirectly the EA is more efficient. However, it has to be noted that due to the requirement of simulating auxiliary data y^* the EA is less efficient than the standard MH algorithm if the normalizing constant $z(\theta)$ is available. The algorithmic scheme of the EA is described in algorithm (2). `R` code for the EA implemented in this thesis can be found in the attached DVD. More details on the theoretical justification of the EA can be found in Murray (2007), Everitt (2012) and Liang et al. (2016). It is crucial for the EA

Algorithm 2: The exchange algorithm

1 Initialize $\theta^{(0)}$
2 **for** $i = 1, \ldots, I$ **do**
  Gibbs update of θ^* and y^*:
3 Draw $\theta^* \sim \mathcal{H}(\theta^*|\theta^{(i-1)})$ where $\mathcal{H}(\cdot|\cdot)$ is assumed to be symmetric.
4 Draw $y^* \sim p(y^*|\theta^*)$.
  Exchange move:
5 Move from $\theta^{(i-1)}$ to θ^* with probability

$$a_{EA} = \min\left\{1, \frac{p(\theta^*)}{p(\theta^{(i-1)})} \frac{q(y|\theta^*)}{q(y|\theta^{(i-1)})} \frac{q(y^*|\theta^{(i-1)})}{q(y^*|\theta^*)}\right\}.$$

6 Repeat step 5 until $a_{EA} = 1$; count the number of retries.
7 **end**

to work that the auxiliary data y^* are exactly simulated from the likelihood and not just a crude approximation. As perfect sampling for the ERGM class is not available, MCMC methods like the TNT sampler simulating from a long Markov chain are required, see section 2.5.1. Everitt (2012) recommend to use at least $N = n^2 - n$ samples of directed networks before the TNT sampler may be assumed to draw from the correct distribution of random graphs.

Applied to ERGM estimation the EA is less prone to model degeneracy than the MCMC-ML algorithm discussed in section 2.5.2. Initial values of θ_0 may simply be sampled from the prior $p(\theta)$. Caimo and Friel (2011) show that if $p(\theta)$ is rather

flat and the proposal distribution $\mathcal{H}(\cdot|\cdot)$ is symmetric

$$\ln a_{EA} \approx \min\left\{0, (\theta - \theta^*)' \left[s(y^*) - s(y)\right]\right\}. \tag{4.8}$$

The probability of the move from θ to θ^* is higher if $||s(y^*) - s(y)||$ is close to zero. If θ^* lies in the degenerate region, a move from θ to θ^* causes a disproportionate increase in $||s(y^*) - s(y)||$ rendering this move unlikely. Vice versa, the EA will quickly move away from initial values $\theta^{(0)}$ in the degenerate region. The MCMC-ML approach discussed in section 2.5.2 would fail in such a situation. The EA may be initialized by simply sampling $\theta^{(0)}$ from the prior $p(\theta)$. Once converged to a high posterior density region the EA is unlikely to leave this area.

Throughout this work the EA is specified in such a way that only accepted draws are stored and that step 5 in algorithm (2) is repeated until a proposal θ^* is accepted. A mean number of four retries is equivalent to a acceptance rate of 25% in the standard MH algorithm.

4.3 Adaptive direction sampling

A major problem in Bayesian ERGM estimation is the thin and correlated support of the posterior not located in the degenerate region of a model, see Rinaldo et al. (2009), whereas the largest part of the parameter space Θ yields degenerate parameter values. The EA may help to avoid the acceptance of degenerate parameter proposals but this might also cause low acceptance rates resulting in inefficient sampling from the posterior and slow convergence of the EA. As a solution, Caimo and Friel (2011) propose population MCMC approaches using parallel chains of the EA resulting in a more efficient exploration of the posterior. These chains should communicate in such a way that chains closer to the non-degenerate region are able to update other chains and "pull" them away from the degenerate region. Such an approach can dramatically improve the mixing and speed up the convergence of all chains together.

The adaptive direction sampler (ADS) introduced by Gilks et al. (1994) is such a technique. It automatically helps to orientate the move of a chain towards the target area of the posterior using parallel MCMC chains. The chains update each other and an adaptive mechanism automatically steers the direction of proposed moves in the parameter space. Only the scaling of those moves needs to be specified.

The ADS may be used to generate proposals $\theta^* \sim \mathcal{H}(\theta^*|\theta^{(i-1)})$ in step 3 of the EA, see algorithm (2). The algorithmic scheme of this step is given in algorithm (3): V parallel ADS chains are used where typically $V \geq 4 \cdot P$. P is the number of parameters of the posterior to be evaluated. In order to generate a move from θ^{i-1} to θ^* for a particular chain, two other parallel chains are randomly selected without replacement and the difference of their states at iteration $(i-1)$ is scaled with the factor γ. An error term ϵ is added which is centered at the proposed value and has variance which is small compared to the size of the proposed moves. `R` code for the implementation of ADS in this thesis can be found in the attached DVD. A theoretical justification of the ADS is given by Gilks and Roberts (1994). Ter

Algorithm 3: Adaptive direction sampling step at iteration i of the EA

1 **for** $v = 1, \ldots, V$ **do**
2 Select two chains without replacement v_1 and v_2 from $\{1, \ldots, V\} \setminus v$
3 Draw $\epsilon \sim N(0, \sigma_{ADS} \cdot \mathbb{I}_P)$
4 Propose $\theta_v^* = \theta_v^{i-1} + \gamma \cdot \left(\theta_{v_1}^{(i-1)} - \theta_{v_2}^{(i-1)}\right) + \epsilon$
5 **end**

Braak (2006) and Ter Braak and Vrugt (2008) propose more advanced methods of automatically selecting a suitable scaling of the proposed move. However, we follow best practice recommendations by Caimo and Friel (2014) for the specification of the ADS used in combination with the EA which work well for ERGM estimation. We refer to the EA combined with ADS as EA-ADS.

Throughout this work we will use $V = 20$ parallel chains for all variants of the EA-ADS. We follow the recommendation by Ter Braak and Vrugt (2008) to use a scaling factor

$$\gamma = 2.387/\sqrt{2 \cdot P}.$$

Like Caimo and Friel (2011) we use a P-variate normal distribution for ϵ with

$$\epsilon \sim N(0, \sigma_{ADS} \cdot \mathbb{I}_P)$$

where $\mathbb{I}_P$ is a $P \times P$ unit matrix and $\sigma_{ADS} \leq 0.01$. The EA-ADS can drastically improve mixing of the chains and reduce autocorrelations compared to a single-site update EA, see Caimo and Friel (2011). In the next section we will see that this

approach leads to surprisingly fast convergence of the parallel EA chains with short burn-in periods required and little autocorrelation of the chains.

4.4 Application

4.4.1 Krackhardt's Managers

We apply Bayesian ERGM estimation using EA-ADS to a famous and well studied friendship network of managers in a high-tech company, see Krackhardt (1987). This is a directed network with $n = 21$ nodes and covariates representing the hierarchical level and the department of nodes. Details can be found in Wasserman and Faust (1994). Snijders (2002) analyze this network and apply MCMC-ML parameter estimation. Note that their model specifications does not contain parameters like $GWESP(y)$ as in those days the social circuit dependence assumption by Snijders et al. (2006) was not developed yet. The network is plotted in figure 4.1. The network shows some substantial clustering in edge-wise shared partners with $EP_2(y) = 29$, $EP_3(y) = 14$ and $EP_4(y) = 15$. There are even edges with six and seven shared partners, see the right panel in figure 4.6.

The distribution of $EP_k(y)$ statistics suggests a social circuit model parametrization including $GWESP(y)$ with a geometrical weight α_E near 0.75. Before specifying a model, we check for model degeneracy using simulations from an ERGM

$$\Pr(Y = y|\theta_1, \theta_2) \propto \exp\{\theta_1 \cdot L(y) + \theta_2 \cdot GWESP(y)\}$$

where $\theta_1 = -2$ fixed and θ_2 ranging from 0 to 1 in steps of 0.01. For each value of θ_2 a set of $R = 10,000$ networks is simulated using the TNT sampler discarding the first $10,000$ iterations and using a thinning interval of $1,000$ draws. This approach is also used to illustrate model degeneracy due to miss-specification of the $GEWSP(y)$ statistic in section 2.5.4. For $\alpha_E = 0.75$ no model degeneracy is detected, see figure 4.2. But it shall be noted that $\mathrm{E}\left[GWESP(y)|\theta_2\right]$ as a function of θ_2 becomes rather steep around $\theta_2 = 0.43$. This suggest a specification with a lower value of α_E to be on the safer side and make MCMC simulation more efficient.

We estimate two model specifications for the Krackhardt friendship network, see table 4.1. m_1 contains the edge parameter, a dyadic covariate indicating whet-

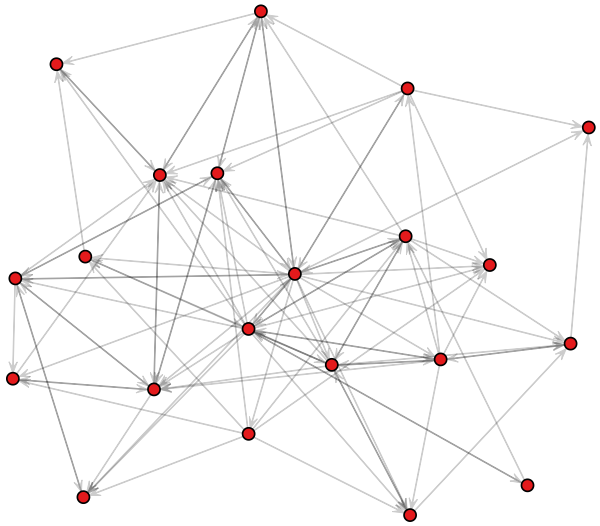

Figure 4.1: Plot of Krackhardt's friendship network on $n = 21$ nodes. The layout is generated using the algorithm of Fruchterman and Reingold (1991).

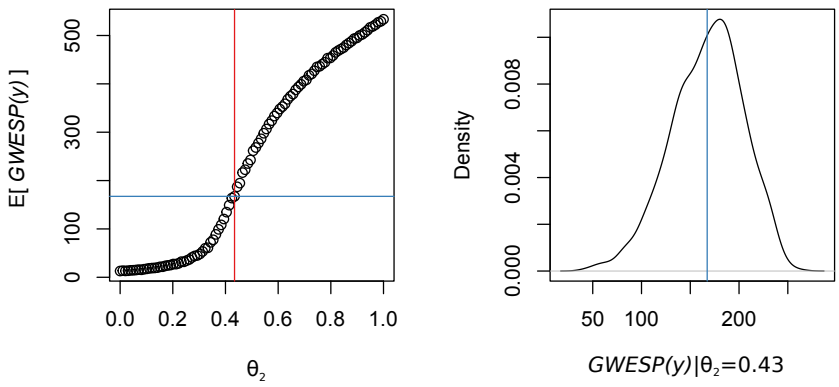

Figure 4.2: *Left panel:* No phase transition observable for $\alpha_E = 0.75$, but the slope gets rather steep around $\theta_2 = 0.43$ so α_E should not be increased further. *Right panel:* The resulting probability distribution for $GWESP(y)$ given $\theta_2 = 0.43$ is unimodal, this mode is close to the mean of the network statistic (blue line).

her two actors belong to the same department (same department) and a dyadic covariate indicating whether two actors have the same hierarchical level (same level). m_1 represents a homophilic network dependence structure. m_2 contains only endogenous network statistics which are the count of edges $L(y)$, the count of mutual ties $M(y)$ and the $GWESP(y)$ statistic with $\alpha_E = 0.2$. m_2 represents endogenous network self organization. We set $\alpha_E = 0.2$ instead of $\alpha_E = 0.75$ which practically ignores the influence of $EP_3(y)$ and $EP_4(y)$ to the change statistics of $GWESP(y)$ and puts only little weight on $EP_2(y)$. Keeping α_E low reduces the potential of the model to fit the observed clustering within the network. But it also facilitates parameter estimation as it will be easier for the EA-ADS to generate acceptable parameter proposals. Note that m_1 could be estimated using logistic regression but m_2 requires an ERGM specification. The *a priori* information on θ is kept low: only weakly informative multivariate normal priors $N(0, 100 \cdot \mathbb{I}_P)$ are specified for both models where $\mathbb{I}_P$ is a $P \times P$ unit matrix. Both models are estimated using the EA-ADS with $V = 20$ parallel chains each of length $I = 1,500$ iterations. The chains are initialized by random draws from a P-variate normal distribution $N(0, \mathbb{I}_P)$. The ADS used for $\mathcal{H}(\theta^*|\theta^{(i-1)})$ in step 3 of algorithm (2) is specified with $\sigma_{ADS} = 0.001$ and $\gamma = 2.387/\sqrt{2P}$. The TNT sampler for network simulation in step 4 of algorithm (2) is allowed to burn in for 420 iterations which allows every tie variable to be toggled, see Everitt (2012).

The run times are 1.09 hours for m_1 and 1.19 hours for m_2 on a 3 GHz Windows machine. On average 5.86 retries are required for the EA-ADS to generate accepted proposals for m_1 while for m_2 on average 5.81 retries are needed. By inspecting the traceplots in figure 4.3 and C.1 it can be seen that all chains converge rapidly and mix well. A relative short burn-in period of only $I_{burn} = 50$ iterations is sufficient which highlights the efficiency of the ADS method. As can be seen in figure 4.4 and figure C.2 the ACF of all chains is negligible after lag 20, for most chains the ACF declines even faster. Using parallel chains the Gelman-Rubin scale reduction factor $\hat{R}$ can be computed, see Gelman et al. (2013). After discarding the first $I_{burn} = 50$ iterations each of the $V = 20$ chains is split into the first and the second half resulting in $2 \cdot V = 40$ chains each of length 725 iterations. $\hat{R}$ is smaller than 1.01 for all parameter estimates from both models. Convergence may be assumed for all chains. The parallel chains are merged to one sample of size

$$I_{merge} = m \cdot (I - I_{burn}) = 20 \cdot (1,500 - 50) = 29,000.$$

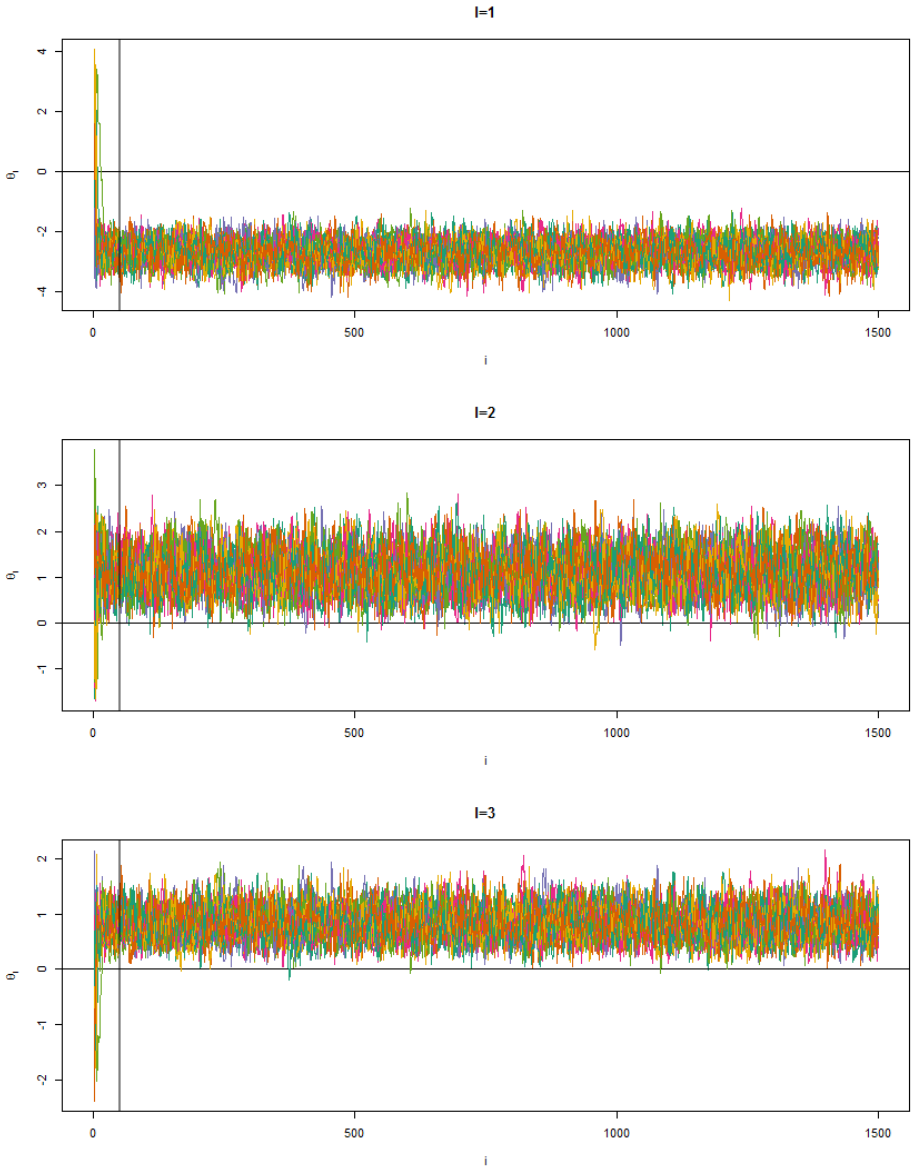

Figure 4.3: Krackhardt's network, m_2:
Traceplots of the $V = 20$ parallel ADS chains of the parameters θ_l. The transparent gray line indicates the discarded $I_{burn} = 50$ samples used as burn-in period.

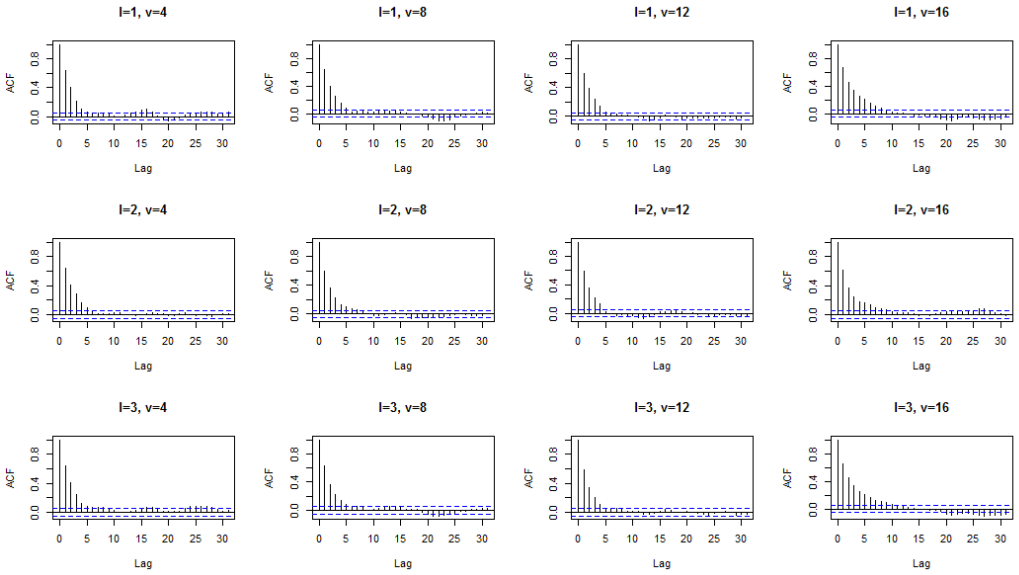

Figure 4.4: Krackhardt's network, m_2:
ACF of some of the parallel ADS chains used in the EA. Four chains displayed for each parameter.

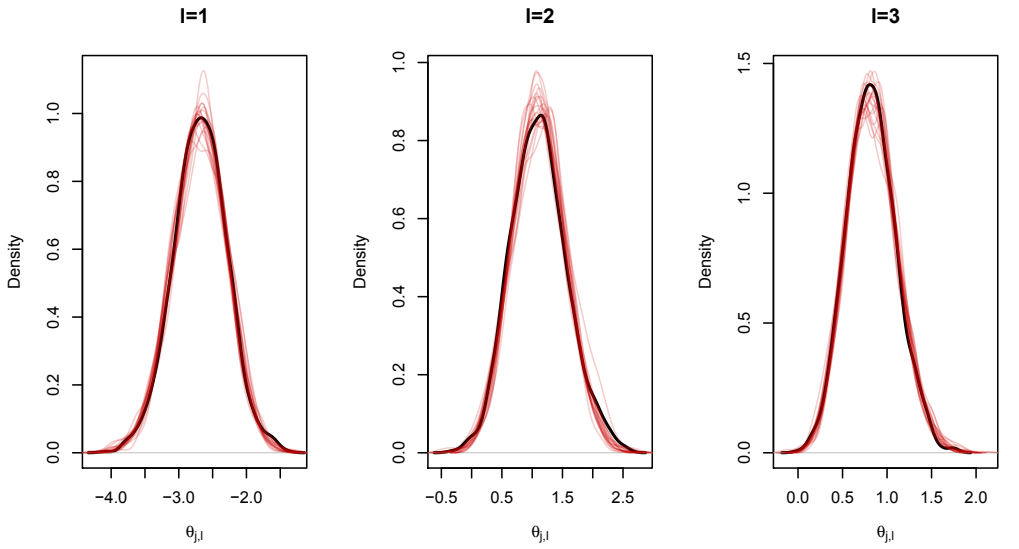

Figure 4.5: Krackhardt's network, m_2:
Densities of merged EA sample:
The thick black line represents the merged sample, the transparent red lines represent the $V = 20$ parallel chains used for ADS sampling.

Table 4.1: Krackhardt's managers: EA parameter estimates

	m_1	m_2
edges	-2.045^{***}	-2.692^{***}
	(0.279)	(0.400)
same department	1.139^{***}	
	(0.345)	
same level	0.944^{***}	
	(0.313)	
reciprocity		1.113^{**}
		(0.440)
GWESP, $\alpha_E = 0.2$		0.834^{***}
		(0.282)
Notes:	MCMC mean value	
	(MCMC standard deviation)	
	*** $0 \notin 99\%$ posterior HDR	
	** $0 \notin 95\%$ posterior HDR	

This merged sample is used to compute the parameter estimates. In figure 4.5 and C.3 the density of the merged sample is compared to the densities of the parallel chains. The results of MCMC parameter estimation are summarized in table 4.1. In both models the edge parameter is negative indicating a low propensity to form ties if the change statistics of all other configurations are zero. Note that the edge parameter has an interpretation similar to the intercept in linear regression, see section 2.4. In m_1 the parameter value of 1.139 for the same level statistic has the interpretation that the odds of forming a tie is increased, *ceteris paribus* (c.p.), by the factor

$$\exp\{1.139\} = 3.124$$

if both actors work in the same department. The odds of tie formation is increased, c.p., by the factor

$$\exp\{0.944\} = 2.570$$

if both actors have the same hierarchical level. The parameter value 1.113 of reciprocity in m_2 indicates that the odds of a tie are increased, c.p., by the factor

$$\exp\{1.113\} = 3.043$$

if it answers an incoming tie. The positive parameter value for $GWESP(y)$ indi-

cates a tendency to form ties within a transitive clustered structure sharing one or two partners.

Posterior predictive checks are applied to compare the fit of the two models. The EA-ADS parameter estimates are used to simulate $1,000$ networks from $p(y|\hat{\theta}_{m_1})$ and $p(y|\hat{\theta}_{m_2})$ using the TNT sampler. A burn-in period of $10,000$ simulations and a thinning interval of $1,000$ simulations are specified. We evaluate the model fit using the distributions of in-degrees, out-degrees and edge-wise shared partners, see figure 4.6.

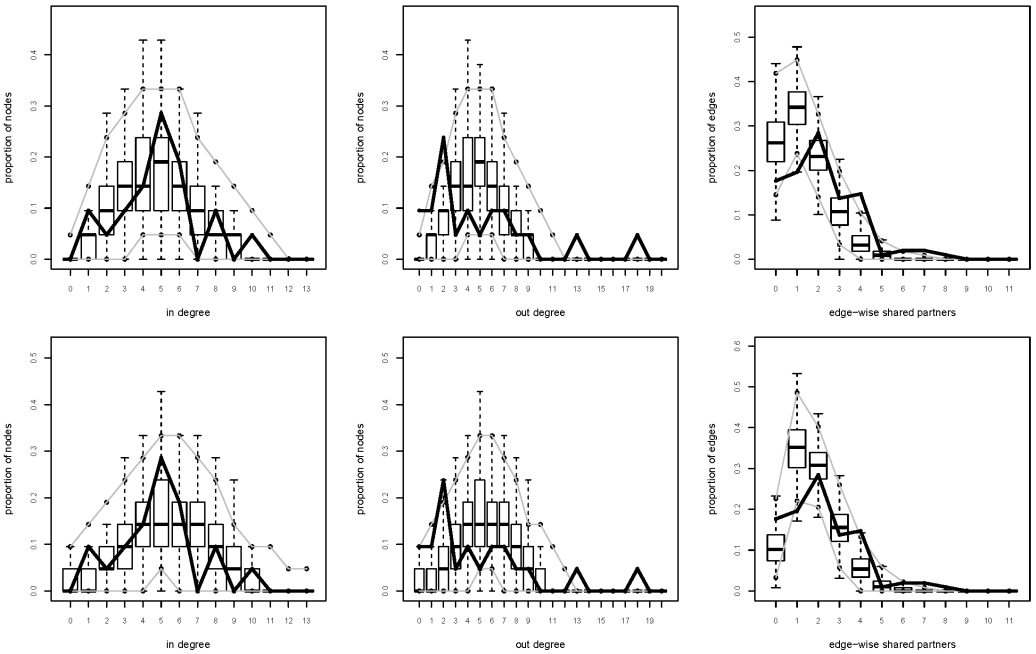

Figure 4.6: Krackhardt's network: GOF EA-ADS m_1 *(top panel)* and m_2 *(bottom panel)*

Using only the default summary statistics it is not possible to select one of the two models over the other. Both models fit well for in-degrees but clearly overestimate out-degrees of 4 and 10. They underestimate out-degrees below 4 and do not represent the two actors with an high out-degree of 13 and 18. Also, both models clearly overestimate the proportion of $EP_1(y)$. The fit for higher degrees of shared partners is acceptable. Both models are comparable in the amount of transitivity they do explain. Nevertheless, it is not clear whether the observed clustering in the network is due to homophily or due to endogenous network self organization. Following the recommendations of Hunter et al. (2008b) no superior

model can be selected: m_2 seems to fit better on the $EP_k(y)$ statistics but fits worse on the distribution of in-degrees. This result highlights why model comparison based on the marginal likelihood is useful: if posterior predictive checks do not help to decide between non-nested ERGM specifications, the model with the largest value of $p(y|m_h)$ should be selected. In chapter 5 it will be apparent that indeed m_2 has to be preferred.

4.4.2 Expert network in Ghana

MCMC-ML estimation for the ERGM specifications in chapter 3 are heavily plagued by convergence issues due to starting values for algorithm (1) which are hard to choose. The EA-ADS is applied to estimate two nested model specifications for the expert network in Ghana: m_1 contains the edge parameter $L(y)$, reciprocity $M(y)$ and $GWESP(y)$ with $\alpha_E = 0.1$ as sufficient network statistics. Note that MCMC-ML will fail in estimating this particular specification as it is not possible to find suitable starting values for algorithm (1). m_2 in addition contains a binary dyadic covariate indicating whether two actors also cooperate in the support network (support). Flat $N(0, 100 \cdot \mathbb{I}_P)$ multivariate normal priors are used for both models where $\mathbb{I}_P$ is a $P \times P$ unit matrix. The EA-ADS is using $V = 20$ parallel chains to sample from the posterior. Each chain has a length of $I = 1,500$ iterations. ADS is specified with $\sigma_{ADS} = 0.001$ and $\gamma = 2.387/\sqrt{2P}$. Following Everitt (2012), a burn-in period of $N = 2,070$ iterations is set for the TNT sampler simulating networks. On average 5.87 retries are required to accept a proposed parameter vector θ^* for m_1 and 7.06 tries are required on average for m_2. The run times are 3.58 hours (m_1) and 4.54 hours (m_2) on a 3 GHz Windows machine using a single core. Traceplots of the chains are depicted in figure C.4 and figure C.7. All chains rapidly converge after less than $I_{burn} = 50$ iterations, which is used as burn-in period, and mix well. The Rubin-Gelman diagnostic of convergence $\hat{R}$, see Gelman et al. (2013), is smaller than 1.01 for all chains. For most chains the ACF is negligible after lag 10, see figure C.5 and figure C.8. For both EA-ADS runs the parallel chains are merged after discarding the burn-in draws creating a single MCMC sample of $I_{merge} = 29,000$ draws each. Density plots of the parallel chains are compared to the density of the merged sample in figure C.6 and figure C.9 where it is obvious that the samples are consistent.

The obtained ERGM parameter estimates and the respective HDR intervals

Table 4.2: Expert network Ghana: EA parameter estimates

	m_1	m_2
edges	-4.210^{***}	-4.361^{***}
	(0.463)	(0.464)
reciprocity	3.786^{***}	3.629^{***}
	(0.271)	(0.276)
GWESP, $\alpha_E = 0.1$	1.339^{***}	1.247^{***}
	(0.394)	(0.393)
support		1.162^{***}
		(0.202)
Notes:	MCMC mean value	
	(MCMC sd)	
	*** $0 \notin$ 99% posterior HDR	

are in line with the results of chapter 3 where in addition the $GWDSP(y)$ statistic is used. This statistic has a parameter estimate for all four models in chapter 3 which are very small compared to the $GWDSP(y)$ statistic. In m_1 the parameter estimates of the other three statistics do not change much compared to the endogenous model in chapter 3, compare table 4.2 and table B.3. Adding the dyadic support covariate in m_2 renders isolated ties more unlikely compared to m_1 as the edge parameter is slightly decreased. The parameter estimates of reciprocity and $GWESP(y)$ are smaller in m_2 so the added support covariate explains some of the observed reciprocity and transitivity. For m_2 the odds of expert information exchange is increased, c.p., by the factor

$$\exp\{1.162\} = 3.196$$

if two actors give political support to each other.

We check for model degeneracy like we did in section 2.5.4. Figure C.10 shows that the MCMC estimates for the $GWESP(y)$ parameter in m_1 and m_2 are far from any degeneracy. Interestingly, $\mathrm{E}[GWESP(y)]$ for m_1 might get close to degeneracy around values of $\theta_3 = 0.9$ which could be the reason why it is not possible to estimate this model specification using MCM-ML methods in chapter 3.

GOF simulations are conducted using the TNT sampler with the same specifications as in section 4.4.1. It can be noticed that the fit of m_2 is better than m_1,

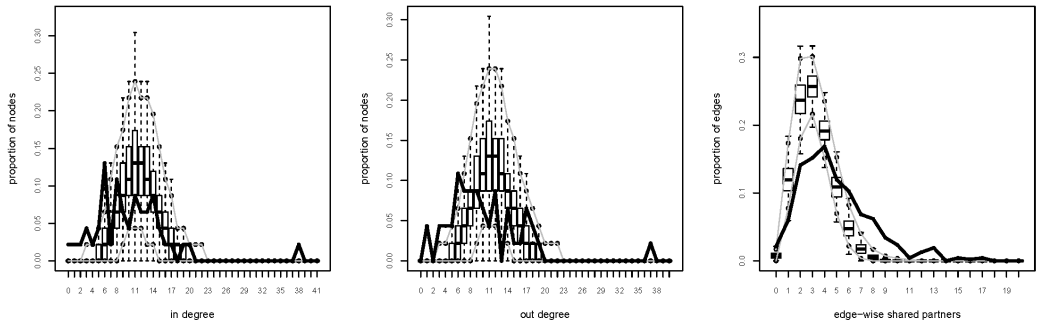

Figure 4.7: Expert network Ghana: GOF EA m_1

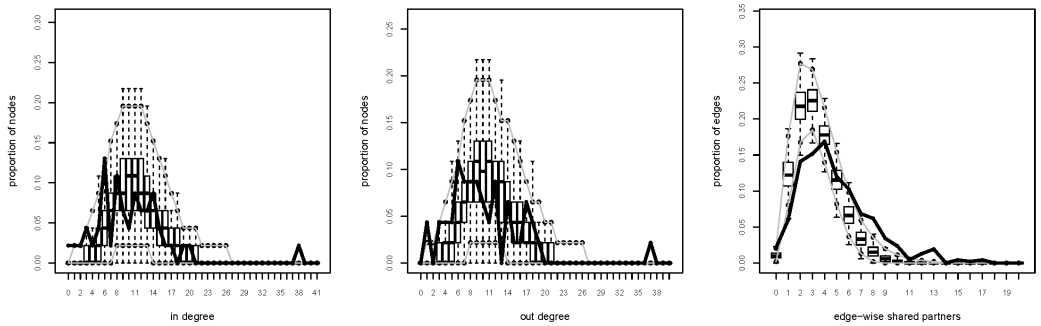

Figure 4.8: Expert network Ghana: GOF EA m_2

see figure 4.7 and 4.8, and is almost as good as the full model in chapter 3, compare figure B.1. m_2 cannot capture the very high degree of some nodes and does not fit the $EP_k(y)$ distribution as well as the full model. Nevertheless, the GOF plots highlight how important the support covariate is to explain tie variable formation as adding this single variable yields goodness-of-fit almost equal to the full model. However, it may be difficult to choose the best model by GOF plots only. Model selection based on the marginal likelihood is discussed in chapter 5.

4.5 Summary

Bayesian ERGM estimation using the EA is robust to starting values near or inside the degenerate region which allows for the estimation of models that would not converge using MCMC-ML methods. The relative advantage in computational speed of the latter approach is useless if the starting values are badly chosen and algorithm (1) is not able to converge, irrespective of the number of specified retries. The bottleneck of the EA-ADS is the effort to generate auxiliary data as in every

iteration the TNT sampler has to converge to sample from the correct distribution of random graphs. Liang et al. (2016) introduce the adaptive exchange algorithm using parallel MCMC chains combined with importance sampling which does not require perfect sampling. This approach might speed up the process of Bayesian ERGM estimation.

Throughout this work we will use rather flat multivariate normal priors centered at zero. Handcock (2003a) propose informative prior distributions on θ to avoid ERGM model degeneracy discussed in section 2.5.4. Unfortunately, the location of the degenerate region of a model specification is unknown *a priori*. Snijders et al. (2006) give some heuristics on these locations for the social circuit ERGM. We consider model degeneracy not a major issue for Bayesian ERGM estimation using the EA as long as a reasonable model is specified. Caimo and Friel (2011) use slightly informative priors as it generally may be assumed that social networks have a negative density parameter for $L(y)$, a positive parameter for reciprocity $M(y)$ and a positive parameter for configurations capturing transitivity like $T(y)$, $EP_k(y)$ or $GWESP(y)$. They propose multivariate normal priors with $\mathrm{E}[\theta_k] \in \{-1, 0, 1\}$ and a diagonal matrix for $\Sigma(\theta) = \sigma_0 \mathbb{I}$ with diagonal elements $10 \leq \sigma_0 \leq 100$.

The Bayesian approach is especially powerful if partially observed networks with missing data shall be analyzed. Koskinen et al. (2010) use data augmentation for Bayesian ERGM estimation with missing tie variable information. Koskinen et al. (2013) extend this work to partially observed networks with missing attribute values. As an alternative to the ERGM class, Fosdick and Hoff (2015) extend the latent factor model, see Hoff (2005) and Hoff (2009), to partially observed networks with missing covariate information. Missing values are not an issue in this work, yet we want to highlight that snowball sampling applied in the survey design in chapter 3 is prone to missing values as it is not always clear which actors belong to a particular network. Wang et al. (2016) introduce multiple imputation of missing information, see Rubin (1976), to the ERGM framework.

One mean of assessing the model fit are posterior predictive checks but in this chapter it became obvious that this approach is not always helpful in selecting a superior model. In the next chapter Bayesian model selection using the marginal likelihood $p(y)$ will be discussed. The EA-ADSä will be combined with thermodynamic integration in order to estimate $p(y)$ and non-nested ERGM specifications for social network data will be compared.

Chapter 5

Bayesian model selection for network data

In this chapter Bayesian model selection for the ERGM class is discussed. Concurring model specifications $m_1, \ldots, m_M$ are compared using the marginal likelihood $p(y|m_h)$ as selection criterion. A new approach for computing the ERGM marginal likelihood is proposed which combines the power posterior sampling approach by Friel and Pettitt (2008), a version of path sampling introduced by Gelman and Meng (1998), with explicit evaluation of the ERGM likelihood (EEL) by estimating its intractable normalizing constant. A path sampling version of the exchange algorithm (EA) discussed in chapter 4 is applied which shall be referred to as the power posterior exchange algorithm (PPEA).

In Bayesian statistics the posterior distribution given a particular model specification m_h is

$$p(\theta_h|y, m_h) = \frac{p(y|\theta_h, m_h)p(\theta_j|m_h)}{p(y|m_h)} \tag{5.1}$$

where $p(y|\theta_h, m_h)$ is the likelihood of model m_h and $p(\theta_h|m_h)$ is the prior on θ_h. The normalizing constant

$$p(y|m_h) = \mathrm{E}_{\theta_h}\left[p(y|\theta_{m_h})\right] = \int_{\theta_h} p(y|\theta_h, m_h)p(\theta_h|m_h)d\theta_h \tag{5.2}$$

insures that (5.1) is a proper probability distribution if data y are available and requires integration over the whole parameter space of $\theta_h \in \Theta_h$. In this role 5.2

is called the marginal likelihood or integrated likelihood. MCMC methods like the Gibbs sampler and the MH sampler circumvent the evaluation of the marginal likelihood as it is usually not of interest for Bayesian inference using a single model. $p(y|m_h)$ can also be used for Bayesian model selection of concurring specifications. In this role it is referred to as the evidence of model m_h. Usually the evidence is analytically not available and Monte Carlo (MC) methods are needed to yield an estimate $\hat{p}(y|m_h)$ unless a prior $p(\theta_h|m_h)$ conjugate to the likelihood $p(y|m_h)$ is used. In most cases, this is not possible so an MC estimate of (5.1) is required which can be computational demanding. The estimation of the ERGM evidence is further complicated by the intractability of the likelihood normalizing constant. Until recently, posterior predictive checks were the only option for ERGM model selection. The GOF plots introduced in chapter 2 are well suited to evaluate whether a model makes sense at all. But it may not be possible to select concurring models because the predictions look indistinguishable considering default summary statistics.

We use power posterior sampling introduced by Friel and Pettitt (2008), which is a version of thermodynamic integration, in order to integrate over the parameter space and yield an estimate of the ERGM evidence. Power posterior sampling is combined with the exchange algorithm of Murray et al. (2006) resulting into the power posterior exchange algorithm (PPEA). Like Friel (2013) we use thermodynamic integration to estimate the normalizing constant of the ERGM likelihood. The PPEA is combined with an explicit evaluation of the likelihood (EEL) which shall be referred to as PPEA-EEL. This fully Bayesian approach offers an estimate of the ERGM evidence while explicitly evaluating the likelihood function. PPEA-EEL is not restricted to small networks and may be used with model specifications that are able to capture realistic patterns of tie variable formation. Previous implicit approaches of estimating the ERGM evidence rely on Laplace approximations or non-parametric density estimates based on MCMC posterior draws. Also, existing approaches are restricted to networks with a more simplistic data structure like indirected ties or small networks. Most approaches use limited model specifications. The PPEA-EEL is applied to directed networks of considerable size using model specifications that are able to capture realistic patterns of social behaviour. Using the evidence for ERGM selection is superior to the established posterior predictive checks as no summary statistics have to be chosen *a priori*. It can help to select model specifications that are indistinguishable by posterior predictive approaches

discussed in section 2.5.3. The evidence may be used to compare non-nested model specifications and offers the potential to compare different model classes which are fit to the same data. The major drawback of the PPEA-EEL is its computational intensity. In order to explicitly evaluate ERGM likelihoods of the demanded complexity an enormous computational effort is required. The PPEA-EEL is not easy to implement and a large number of MCMC samples has to be inspected. Heuristics are offered on how to implement and control the MCMC sampling techniques. Graphical methods are proposed to evaluate the numerical stability of evidence estimation.

In section 5.1 the Bayes factor is discussed and how to select concurring models in the Bayesian framework. In section 5.2 sampling techniques are introduced that can be used to estimate the model evidence. The focus of this work is on approaches related to importance sampling, namely bridge sampling and path sampling. Both parts of the PPEA-EEL, power posterior sampling the the explicit evaluation of the likelihood, are based on path sampling techniques. Other approaches of Bayesian model selection such as the across-model approach introduced by Green (1995) and the method by Chib (1995) are discussed in section 5.2.5. The PPEA-EEL is introduced in section 5.3.1. In section 5.4 we apply this approach to the Krackhardt's managers network known from section 4.4 and the expert network in Ghana known from chapter 3. Geweke (1999) and Ando (2010) give an introduction to Bayesian model selection and review analytical and approximative methods of computing the evidence. A more recent review including more advanced computation methods is given by Friel and Wyse (2012).

5.1 The Bayes factor

If the evidence (5.2) was available for all competing models $m_1, \ldots, m_M$ we could compute the posterior probability of m_h as

$$\Pr(m_h|y) = \frac{p(y|m_h)\Pr(m_h)}{\sum_{h=1}^{M} p(y|m_h)\Pr(m_h)}. \tag{5.3}$$

The model specification maximizing (5.3) should be selected. Typically *a priori* indifference between models is assumed with

$$\Pr(m_1) = \Pr(m_2) =, \ldots, = \Pr(m_J)$$

Table 5.1: Interpretation of the Bayes factor according to Kass and Raftery (1995)

$2 \ln B_{12}$	Evidence against m_2
0 to 2	Not worth more than a bare mention
2 to 6	Substantial
6 to 10	Strong
> 10	Decisive

so (5.3) simplifies to

$$\Pr(m_h|y) = \frac{p(y|m_h)}{\sum_{h=1}^{M} p(y|m_h)}. \tag{5.4}$$

If the observed data are unlikely under a particular model m_h, the value of $p(y|m_h)$ will be smaller than under a model where the observed data are very likely. The posterior odds of two concurring models m_1 and m_2 are

$$\frac{\Pr(m_1|y)}{\Pr(m_2|y)} = \frac{\Pr(m_1)}{\Pr(m_2)} \times \frac{p(y|m_1)}{p(y|m_2)}. \tag{5.5}$$

The Bayes factor of the two concurring models m_1 and m_2

$$B_{12} = \frac{p(y|m_1)}{p(y|m_2)} \tag{5.6}$$

is the ratio of the respective model evidences. (5.5) is obtained by updating the prior odds $\Pr(m_1)/\Pr(m_2)$ with B_{12}. In most cases *a priori* indifference between two models is assumed with $\Pr(m_1) = \Pr(m_2)$. In this case (5.5) is identical to B_{12}. Calculation of the evidence $p(y|m_h)$ is crucial for Bayesian model comparison using Bayes factors, see Geweke (1999) for a review. The obvious application of (5.6) is to pairwise compare models. Kass and Raftery (1995) give an interpretation of the scale of $2 \ln BF_{12}$ as evidence against m_2, see table 5.1. Note that the evidence $p(y|m_h)$ and the Bayes factor B_{12} can be used to compare non-nested models and different model classes fit to the same data. E.g. this would allow for the comparison of an ERGM and a latent factor model (Hoff, 2005, 2009) fit to the same network.

After computing the posterior probabilities $\Pr(m_1|y), \ldots, \Pr(m_M|y)$ of a set of model specifications is computed, Bayesian model averaging can be used to

capture the uncertainty in model selection. Instead of discarding all but the single selected model a predictive distribution averaging over all estimated models is constructed. The predictive distribution of m_h is weighted with its respective posterior probability $Pr(m_h|y)$. See Madigan and Raftery (1994) and Raftery et al. (1997) and Hoeting et al. (1999) for a tutorial.

Many methods of estimating $p(y|m_h)$, see section 5.2, require an explicit evaluation of the likelihood function

$$p(y|\theta_h, m_h) = \frac{q(y|\theta_h, m_h)}{z(\theta_h)}$$

with its non-normalized kernel $q(y|\theta_h)$ and the normalizing constant $z(\theta_h)$. Unfortunately, for the ERGM model class the likelihood is only known up to the normalizing constant $z(\theta_h)$. Methods based on importance sampling to estimate $z(\theta_h)$ of the ERGM are available where it is sufficient to know the non-normalized likelihood $q(y|\theta_h, m_h)$. The power posterior approach by Friel and Pettitt (2008) can be used to estimate the ERGM evidence using path sampling to explicitly estimate $z(\theta_h)$, see Friel (2013). Everitt et al. (2016) review methods of Bayesian model comparison based on the evidence of models using likelihoods with unknown normalizing constants. The information criteria introduced by Akaike (1998)

$$AIC_h = 2P - 2\ln\hat{p}(y|\theta, m_h) \tag{5.7}$$

and Schwarz (1978)

$$BIC_h = \ln(N)P - 2\ln\hat{p}(y|\theta, m_h) \tag{5.8}$$

require an explicit evaluation of the likelihood.[1] These criteria are available for the ERGM model class only after estimation of the normalizing constant $z(\theta_h)$. Hunter and Handcock (2006) offer the estimation of $z(\theta_h)$ using a simplified version of path sampling considered in section 5.2.3.[2] Their approach is implemented in the `R` (R Development Core Team, 2011) package `ergm` (Hunter et al., 2008a) and is used to estimate the log-likelihood, the AIC and the BIC of the model specifications

[1] Note that for network data N is the number of tie variables.

[2] Hunter and Handcock (2006) refer to their approach as *bridge sampling*. In fact they use a discretized version of path sampling with J fixed steps which is similar to the approach proposed by Friel (2013).

estimated in chapter 3. Posterior predictive checks discussed in section 2.5.1 can be used to evaluate the capability of a model to predict realistic data, e.g. data that are close to the observed data. Appropriate measures have to be chosen and it can be difficult to decide which model offers a better fit.

5.2 Computing the model evidence

Bayesian model estimation requires Monte Carlo (MC) methods for most model classes as the normalizing constant of the posterior distribution is analytically unavailable. In this section numerical methods of computing the evidence are discussed. Importance sampling is an option but in most cases this methods will be very inefficient. Meng and Wong (1996) introduce bridge sampling which uses samples from the posterior of interest and a bridging distribution with known normalizing constant. The marginal likelihood of interest can be estimated using a ration of normalizing constant. If the Kullback–Leibler divergence (Kullback, 1968) between the bridging distribution and the target distribution is large, it may be difficult to find a suitable bridging distribution. Path sampling (or thermodynamic integration) introduced by Gelman and Meng (1998) solves this problem by using a path of multiple bridging distributions where the Kullback–Leibler divergence of subsequent bridging distributions is small. Power posterior sampling introduced by Friel and Pettitt (2008) is a version of path sampling where a discretized path from the prior to the posterior of interest is used the estimate the evidence of the model of interest.

For the ease of notation we will suppress the conditioning on the model index m_h and drop the subscript index on the model parameters θ_h. We assume the evidence to be estimated within a particular model specification, referring simply to $p(y)$ instead of $p(y|m_h)$.

5.2.1 Importance sampling

Importance sampling is a rather simple MC method to compute the normalizing constant $p(y)$ of a posterior distribution $p(\theta|y)$, see Kass and Raftery (1995) for a general introduction. $q_0(\theta)$ is the (non-normalized) importance density. The

evidence may be expressed as

$$p(y) = \frac{\mathrm{E}_\theta \left[p(y|\theta) \frac{p(\theta)}{q_0(\theta)} \right]}{\mathrm{E}_\theta \left[\frac{p(\theta)}{q_0(\theta)} \right]} \tag{5.9}$$

where the expectation $E_\theta [\ldots]$ is taken with respect to $q_0(\theta)$. $p(\theta)/q_0(\theta)$ is the importance weight, $p(\theta)$ is the prior distribution and $p(y|\theta)$ is the likelihood. The evidence may be estimated using an importance sample $\theta^{(1)}, \ldots, \theta^{(R)}$ generated from $q_0(\theta)$ with

$$\hat{p}(y) = \frac{\sum_{r=1}^{R} p(y|\theta^{(r)}) \frac{p(\theta^{(r)})}{q_0(\theta^{(r)})}}{\sum_{r=1}^{R} \frac{p(\theta^{(r)})}{q_0(\theta^{(r)})}}. \tag{5.10}$$

See Geweke (1989) for a discussion. The main difficulty using (5.10) is to find a suitable importance function $q_0(\theta)$.

McCulloch and Rossi (1991) discuss the prior $p(\theta)$ as importance function, which simplifies the evidence to

$$p(y) = \frac{\mathrm{E}_\theta \left[p(y|\theta) \frac{p(\theta)}{p(\theta)} \right]}{\mathrm{E}_\theta \left[\frac{p(\theta)}{p(\theta)} \right]} = \mathrm{E}_\theta \left[p(y|\theta) \right]. \tag{5.11}$$

(5.11) is the expectation of the likelihood $p(y|\theta)$ taken with respect to the prior $p(\theta)$. Using a sample $\theta^1, \ldots, \theta^{(R)}$ simulated from $p(\theta)$ yields the prior arithmetic mean estimator of the evidence

$$\hat{p}(y) = \frac{1}{R} \sum_{r=1}^{R} p(y|\theta^{(r)}). \tag{5.12}$$

The problem using (5.12) is that in most cases the area of high likelihood is rather small whereas the prior is chosen to be rather diffuse. Thus most sample points will have very small likelihood values resulting in an inefficient simulation process, see McCulloch and Rossi (1991).

Newton and Raftery (1994) propose to use the non-normalized posterior $q(\theta|y) = p(y|\theta)p(\theta)$ as importance function. The evidence may be expressed as

$$p(y) = \frac{\mathrm{E}_\theta\left[p(y|\theta)\frac{p(\theta)}{q(\theta|y)}\right]}{\mathrm{E}_\theta\left[\frac{p(\theta)}{q(\theta|y)}\right]} = \frac{\mathrm{E}_\theta\left[p(y|\theta)\frac{p(\theta)}{p(y|\theta)p(\theta)}\right]}{\mathrm{E}_\theta\left[\frac{p(\theta)}{p(y|\theta)p(\theta)}\right]} = \frac{1}{\mathrm{E}_\theta\left[\frac{1}{p(y|\theta)}\right]} \tag{5.13}$$

Using a sample $\theta^{(1)}, \ldots, \theta^{(R)}$ simulated from the posterior yields the harmonic mean estimator

$$\hat{p}(y) = \left[\frac{1}{R}\sum_{r=1}^{R}\frac{1}{p(y|\theta^{(r)})}\right]^{-1}. \tag{5.14}$$

(5.14) is appealing as in most cases the MCMC sample used for Bayesian parameter estimation could be used. However, the harmonic mean estimator of the evidence has serious drawbacks. If the prior is more diffuse than the posterior, which is the case for most model specifications, high likelihood values will be overrepresented in the sample so $\hat{p}(y)$ will be insensitive to the prior. The approach also requires an unpractical high number of MCMC simulations. Even worse, estimates tend to have infinite variance, see Friel and Wyse (2012) for a discussion. Xie et al. (2011) show that the harmonic mean estimator generally overestimates the evidence.

Finding a suitable importance distribution to estimate the evidence is complicated. If the distance in the sense of the Kullback-Leibler (KL) divergence between the prior and posterior is large, importance sampling will always be inefficient.

5.2.2 Bridge sampling

Newton and Raftery (1994) try to ameliorate the problem of finding a suitable importance density for (5.9) by using a mixture of prior and posterior as importance density. This is a first step towards bridge sampling introduced by Meng and Wong (1996): instead of using a single importance density to estimate the evidence, multiple densities are used. The densities are constructed in such a way that they can "bridge" the large KL divergence between the prior and the posterior. In its fundamental form bridge sampling uses two non-normalized sampling densities $q_0(\theta)$ and $q_1(\theta)$ generating two independent importance samples. Typically, $q_1(\theta)$ has an unknown normalizing constant z_1. $q_0(\theta)$ is a bridging distribution with known normalizing constant z_0. In most cases $q_1(\theta)$ is the posterior of interest and

the normalizing constant z_1 is the evidence. $q_0(\theta)$ is the non-normalized kernel of the proper prior $p_0(\theta)$ with known z_0. Note that bridge sampling can also be used to estimate the normalizing constant of an intractable likelihood function, see e.g. Hunter and Handcock (2006).

Meng and Wong (1996) define the bridge sampling identity for a ratio of two normalizing constants z_1 and z_0 as

$$\eta \equiv \frac{z_1}{z_0} = \frac{\mathrm{E}_0[q_1(\theta)\varphi(\theta)]}{\mathrm{E}_1[q_0(\theta)\varphi(\theta)]}. \tag{5.15}$$

$q_j(\theta), \quad j \in (0,1)$ is a non-normalized sampling distribution. The normalizing constant z_0 of $q_0(\theta)$ is known while z_1 of $q_1(\theta)$ is unknown. $\mathrm{E}_j[\ldots]$ is the expectation taken with respect to the normalized sampling distribution $p_j(\theta) = q_j(\theta)/z_j$. $\varphi(\theta)$ is an arbitrary function satisfying

$$0 < \left| \int_{\Omega_0 \cap \Omega_1} \varphi(\theta) p_0(\theta) p_1(\theta) d\theta \right| < \infty.$$

Ω_j is the support of $p_j(\theta)$ where in most cases $\Omega_0 = \Omega_1$. The general bridge sampling estimate of η using draws from both $q_0(\theta)$ and $q_1(\theta)$ is

$$\hat{\eta} = \frac{1/R_0 \sum_{r_0=1}^{R_0} q_1(\theta_0^{(r_0)})\varphi(\theta_0^{(r_0)})}{1/R_1 \sum_{r_1=1}^{R_1} q_0(\theta_1^{(r_1)})\varphi(\theta_1^{(r_1)})}. \tag{5.16}$$

Various choices of $\varphi(\theta)$ are discussed by Meng and Wong (1996), Frühwirth-Schnatter (2004) and Chen et al. (2002). Note that importance sampling is a special case of bridge sampling as

$$\varphi(\theta) = q_0(\theta)^{-1}, \quad \Omega_1 \subset \Omega_0$$

yields

$$\eta = \mathrm{E}_0 \left[\frac{q_1(\theta)}{q_0(\theta)} \right]. \tag{5.17}$$

where $\mathrm{E}_0[\ldots]$ is taken with respect to the normalized importance density $p_0(\theta)$.

The resulting importance sampling estimate of η is

$$\hat{\eta} = \frac{1}{R_0} \sum_{r_0=1}^{R_0} \frac{q_1(\theta^{(r_0)})}{q_0(\theta^{(r_0)})} \tag{5.18}$$

with $\theta^{(1)}, \ldots, \theta^{(R_0)}$ sampled from $q_0(\theta)$. This is the importance sampling estimate used by Geyer and Thompson (1992) which allows for the explicit estimation of the ERGM normalizing constant, see section 2.5.2 for a discussion.[3]

In most cases bridge sampling is implemented with $q_1(\theta)$ as the model posterior of interest with unknown normalizing constant z_1 which is the model evidence. MCMC draws are simulated from $q_1(\theta)$ and are used to construct the bridging distribution $q_0(\theta)$ in such a way that $z_0(\theta)$ is known. The model evidence may be estimated using

$$z_1 = \hat{\eta} \cdot z_0.$$

Lopes and West (2004) use draws from the non-normalized posterior and a multivariate normal approximation to the posterior after obtaining MCMC draws to estimate the evidence of factor models. Note that this approach requires the posterior to be well behaved and unimodal, see Diciccio et al. (1997). Frühwirth-Schnatter (2004) use a similar approach to estimate the evidence of mixture and Markov switching models where $q_1(\theta)$ is the non-normalized posterior of interest and $q_0(\theta)$ is an unsupervised bridging distribution constructed from an MCMC sample simulated from $q_1(\theta)$.

5.2.3 Path sampling

It may be very difficult to find a suitable bridging distribution in order to estimate (5.15). If the KL divergence between $q_0(\theta)$ and the $q_1(\theta)$ is large, (5.16) might yield a very inefficient estimate of the ratio of normalizing constants η, see Calderhead and Girolami (2009) for a discussion. As a solution to this problem Gelman and Meng (1998) develop path sampling as an extension to bridge sampling. Instead of using a single bridging distribution and the target distribution $J \to \infty$ distributions

[3] Finding a suitable importance distribution is very difficult. Hunter and Handcock (2006) propose a simplified path sampling approach to estimate the intractable ERGM normalizing constant which they refer to as multiple bridge sampling.

are used. A class of non-normalized densities

$$q(\theta|t), \quad t \in [0,1]$$

has to be defined joining $q_0(\theta)$ and $q_1(\theta)$. This class has to be constructed in such a way that

$$q(\theta|t=0) = q_0(\theta)$$

and

$$q(\theta|t=1) = q_1(\theta)$$

defining a continuous and differentiable path in the distributional space from $q_0(\theta)$ to $q_1(\theta)$. This path helps to "bridge" die KL divergence between $q_0(\theta)$ and $q_1(\theta)$ if both densities have the same support. The scalar parameter t is a continuous random variable in $[0,1]$ with prior distribution $p(t)$. t may be an uniform random variable in $[0,1]$ or follow a beta distribution, see Xie et al. (2011). z_t is a normalizing constant so that

$$p(\theta|t) = \frac{1}{z_t} q(\theta|t).$$

Gelman and Meng (1998) define the ratio of normalizing constants known from bridge sampling, see equation (5.15), on the log-scale. This is called the path sampling identity:

$$\lambda = \ln\left(\frac{z_1}{z_0}\right) = \ln(z_1) - \ln(z_0) = \int_0^1 \mathrm{E}_{\theta|t}[U(\theta,t)]dt. \tag{5.19}$$

$\mathrm{E}_{\theta|t}$ is the expectation taken with respect to $p(\theta|t)$.

$$U(\theta,t) = \frac{d}{dt} \ln q(\theta|t) \tag{5.20}$$

is called the potential and t is called the temperature.[4] This terminology is chosen due to the analogy to thermodynamic integration used in the field of statistical

[4]Considering the analogy to thermodynamic integration, t is precisely an *inverse* temperature: if $t = 0$, the free energy may be imagined to be maximal with atoms moving freely. $t = 1$ corresponds to the minimal temperature possible corresponding to zero free energy, the movement of atoms is minimal. Somehow simplifying the path from $t = 0$ to $t = 1$ corresponds to cooling a system from its gaseous phase to its solid phase. However, as this is just an analogy, we refrain from calling t an inverse temperature and call it simply the temperature.

physics, see Ogata (1989). Conceptually, the methods are identical, yet the fields of application are very different, see Gelman and Meng (1998) for a discussion. The log ratio of normalizing constants λ can be found by solving an integral over the unit interval, see appendix D.1 for details. As

$$t \sim p(t),$$

(5.19) can be expressed as

$$\lambda = \mathrm{E}_{\theta,t}\left[\frac{U(\theta,t)}{p(t)}\right] \tag{5.21}$$

where $\mathrm{E}_{\theta,t}[\ldots]$ is taken with respect to the joint sampling distribution

$$p(\theta,t) = p(\theta|t)p(t). \tag{5.22}$$

This suggests the estimate of (5.19)

$$\hat{\lambda} = \frac{1}{R}\sum_{r=1}^{R}\frac{U(\theta^{(r)},t^{(r)})}{p(t^{(r)})} \tag{5.23}$$

where $\left((\theta^{(1)},t^{(1)}),\ldots,(\theta^{(R)},t^{(R)})\right)$ is sampled from $p(\theta,t)$.

It may not always be practical to sample t from a continuous distribution. In some cases $U(\theta,t)$ is difficult to evaluate and may require additional computational effort, see section 5.2.4. In such a situation it is more practical to evaluate $U(\theta,t)$ for a fixed number of temperatures instead of using $t^{(1)},\ldots,t^{(R)}$ sampled points. The integral over the unit interval required to solve (5.19) may be approximated by discretizing the path from $p_0(\theta)$ to $p_1(\theta)$ on J finite points applying a trapezoidal rule or Simpson's rule, see among others Lartillot and Philippe (2006) and Friel and Pettitt (2008). Using a discretized approximation of that path, t is not a random variable anymore but is chosen from a fixed grid. For each point $t_j,\ j = 1,\ldots,J$ the expected value of the potential given the temperature step t_j may be estimated as

$$\hat{E}_{\theta|t_j}\left[U(\theta,t_j)\right] = \frac{1}{R}\sum_{r=1}^{R}\frac{d}{dt}\ln q(\theta^{(r)}|t_j) \tag{5.24}$$

using a MC sample $\theta^{(1)},\ldots,\theta^{(R)}$ drawn from $p(\theta|t_j)$. The standard trapezoidal

rule of integrating a function $f(x)$ between points a and b is

$$\int_a^b f(x)dx = (b-a)\left[\frac{f(b)-f(a)}{2}\right]. \tag{5.25}$$

Applying (5.25) yields an estimate of (5.19)

$$\hat{\lambda} = \sum_{j=2}^{J} \xi_j = \sum_{j=2}^{J} (t_j - t_{j-1}) \left(\frac{\hat{E}_{\theta|t_{j-1}}[U(\theta, t_{j-1})] + \hat{E}_{\theta|t_j}[U(\theta, t_j)]}{2} \right). \tag{5.26}$$

ξ_j is the mean of the estimated expected potentials at subsequent temperature steps t_{j-1} and t_j weighted by the temperature step size $(t_j - t_{j-1})$.

Path sampling may be used with Bayesian statistics in order to estimate the model evidence $p(y)$. In such a case $q_1(\theta)$ is the posterior of interest with unknown normalizing constant

$$z_1 = z_{t=1} = p(y).$$

$q_0(\theta)$ is a proper prior distribution with known normalizing constant

$$z_0 = z_{t=0} = 1.$$

Using (5.19) the model evidence may be estimated as

$$\ln \hat{p}(y) = \hat{\lambda}$$

as it is known that

$$\ln z_0 = \ln z_{t=0} = 0.$$

The path of t over the unit interval then corresponds to a transition from the prior at $t = 0$ to the posterior at $t = 1$. Lartillot and Philippe (2006) and Friel and Pettitt (2008) construct the class of tempered distributions in such a way that t has the role of a data weight where $t = 0$ corresponds to zero information obtained from data. Applying path sampling the problem of integrating over the whole parameter space Θ in order to compute the evidence may be replaced by the integral over the unit interval.

In this work path sampling is used in order to obtain the model evidence during Bayesian ERGM estimation by combining path sampling with the exchange algorithm, see section 5.3.1. This requires an explicitly estimate of the normalizing

constant of the intractable ERGM likelihood, see section 5.3.2.

5.2.4 Power posterior sampling

Power posterior sampling introduced Friel and Pettitt (2008) is a version of path sampling where $q_1(\theta)$ is the non-normalized posterior and $q_0(\theta)$ is the non-normalized prior. The continuous path from $q_0(\theta)$ to $q_1(\theta)$ is constructed by raising the likelihood to a power t

$$p(\theta|y,t) \propto p(y|\theta)^t p(\theta)$$

which is called the power posterior. The non-normalized target posterior of interest at the temperature $t = 1$ is

$$p(\theta|y,t=1) \propto (y|\theta)^{t=1} p(\theta) = q_1(\theta). \tag{5.27}$$

The prior is

$$p(\theta|y,t=0) \propto p(y|\theta)^{t=0} p(\theta) = p(\theta) = q_0(\theta) \tag{5.28}$$

and is assumed to be proper. In analogy to thermodynamic integration we refer to t as the temperature, $p(y|\theta)^t$ as the tempered likelihood and $p(\theta|y,t)$ as the tempered posterior. The normalizing constant of the tempered posterior $p(\theta|y,t)$ is

$$p(y|t) = \int_\theta p(y|\theta)^t p(\theta) d\theta, \quad t \in [0,1] \tag{5.29}$$

where

$$p(y|t=1) = p(y)$$

is the normalizing constant of the target posterior and thus the model evidence. If a proper prior is used,

$$p(y|t=0) = 1$$

as it is known that the prior integrates to 1. These results can be applied to the path sampling identity (5.19). The resulting power posterior sampling identity

expresses the log of the model evidence as

$$\begin{aligned} \ln\left(p(y)\right) &= \ln \frac{p(y|t=1)}{p(y|t=0)} \\ &= \ln p(y|t=1) - \ln p(y|t=0) \\ &= \int_0^1 \mathrm{E}_{\theta|y,t}\left[\ln p(y|\theta)\right] dt. \end{aligned} \tag{5.30}$$

$\mathrm{E}_{\theta|y,t}\left[\ldots\right]$ is the expectation taken with respect to the tempered posterior $p(\theta|y,t)$. Details on (5.30) are given in appendix D.2. Friel and Pettitt (2008) express the model evidence $p(y)$ as an integral over the unit interval from 0 to 1 with respect to the power t. Just like with path sampling the log evidence can be calculated by evaluation of the one dimensional integral over the unit interval. At low temperatures with t close to 0 the tempered posterior will strongly resemble the prior which typically has a much higher variance than the target posterior. This allows the sampler to move freely and explore the parameter space for t close to zero. The closer the temperature gets to 1 the more restricted the sampler will be and the less free are its movements. Calderhead and Girolami (2009) give a detailed illustration on sampling from tempered posteriors where the target posterior has a complex multimodal shape. They highlight that a strength of tempering algorithms like power posterior sampling is their ability to facilitate sampling from complex distributions without getting caught inside local maxima. On the advantages of tempering algorithms see also Neal (2001).

The integral

$$\int_0^1 \mathrm{E}_{\theta|y,t}\left[\ln p(y|\theta)\right] dt$$

can be approximated using J discretized steps $j = 1, \ldots, J$ transiting from $t_1 = 0$ to $t_J = 1$ over the unit interval instead of sampling t from a continuous distribution. If the normalizing constant $z(\theta|t)$ of the tempered likelihood $p(y|\theta,t)$ is not available analytically, as it is the case for the ERGM, this approach has the advantage that $z(\theta|t)$ needs to be estimated only for a fixed number of J tempered likelihoods. Sampling t from the continuous distribution $p(t)$ is not an option for intractable likelihoods as the computational burden of estimating $z(\theta|t)$ for every sampled t would be too high. For each discrete temperature step t_j the expected value of the tempered likelihood $\mathrm{E}_{\theta|y,t_j}\left[\ln p(y|\theta)\right]$ with respect to the tempered posterior is estimated using MCMC simulations from $p(\theta|y,t_j)$. This corresponds to

a discretized transition on J steps from the prior

$$p(y|\theta)^{t_j=0}p(\theta)$$

to the non-normalized posterior

$$p(y|\theta)^{t_j=1}p(\theta).$$

Subdividing the whole integrating range from $t = 0$ to $t = 1$ into $J - 1$ intervals

$$[t_{j-1}, t_j],\ j = 2, \ldots, J$$

and applying the trapezoidal rule (5.25) yields a discretized approximation of the log evidence

$$\ln p(y) \approx \sum_{j=2}^{J} \xi_j \tag{5.31}$$

where

$$\xi_j = (t_j - t_{j-1}) \times \left(\frac{\mathrm{E}_{\theta|y,t_{j-1}} [\ln p(y|\theta)] + \mathrm{E}_{\theta|y,t_{t_j}} [\ln p(y|\theta)]}{2} \right). \tag{5.32}$$

The expectation $\mathrm{E}_{\theta|y,t}[\ldots]$ in (5.32) is taken with respect to the tempered posterior $p(\theta|y,t)$. ξ_j is the mean of the expected values of the log likelihood taken with respect to neighbouring tempered posteriors $p(\theta|y,t_{j-1})$ and $p(\theta|y,t_j)$ weighted by the temperature step length $(t_j - t_{j-1})$. The estimate (5.31) is obtained using J MCMC samples each of size I

$$\theta_j^{(i)}, \ldots, \theta_j^{(I)}, \quad (j = 1, \ldots, J)$$

taken from the respective tempered posterior $p(\theta|y,t_j)$ and estimating

$$\hat{E}_{\theta|y,t_j} [\ln p(y|\theta)] = \frac{1}{I} \sum_{i=1}^{I} \ln p \left(y|\theta_j^{(i)} \right). \tag{5.33}$$

Note that (5.33) requires an explicit evaluation of the likelihood $p(y|\theta)$ at the sampled points $\theta_j^{(i)}$. If the normalizing constant of the likelihood is not available,

this requires additional computational effort, see section 5.3.2.

The major drawback of using a fixed schedule of temperature steps is the induced discretizational error. Calderhead and Girolami (2009) find that equidistant spacing of steps as used by Lartillot and Philippe (2006) is biased and generally the spacing should be skewed towards the prior, see also Xie et al. (2011). Calderhead and Girolami (2009) give theoretical details on how the optimal path minimizes the KL divergence between subsequent tempered posteriors $p(\theta|y, t_{j-1})$ and $p(\theta|y, t_j)$. They also give empirical evidence that an adequate temperature spacing is more important than the number of discretizing steps. Friel and Pettitt (2008) use the discretized path

$$t_j = \left(\frac{j}{J}\right)^w \tag{5.34}$$

and set $w = 5$ focusing on low values of t_j corresponding to tempered posterior distributions which are very close to the prior. This temperature spacing shall both minimize the discretizational error and the bias in estimating $p(y)$. Friel et al. (2014) suggest to reduce the discretizational error either by correcting the trapezoidal rule given a particular temperature path or by adaptively optimizing the placement of the steps in the path itself. On other approaches to find an optimal temperature spacing for power posterior sampling, see Lefebvre et al. (2009), Behrens et al. (2012), Hug et al. (2016) and Oates et al. (2016). Throughout this work the spacing (5.34) with $w = 5$ is used.

5.2.5 Other approaches

There are many more MC techniques available for the computation of the evidence. A short overview is given on those techniques that have been applied for ERGM selection. An early method is the Laplace approximation by Tierney and Kadane (1986) which makes the strong assumption that the posterior may be approximated by a normal distribution. The posterior should at least be highly peaked around its mode which has to be close to the ML estimate. Thiemichen et al. (2016) us a Laplace approximation for the estimation of the ERGM evidence after obtaining MCMC samples simulated with the EA. This method has the advantage that it is fast and rather easy to implement, see Friel and Wyse (2012) for a discussion. However, it shall be noted that the strong approximation assumption of an elliptical

posterior distribution will not always hold.

A very popular method of computing the model evidence is the approach introduced by Chib (1995) for the output of the Gibbs sampler. The formula of the posterior (4.1) can be rearranged to

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}. \tag{5.35}$$

The evidence can be estimated by plugging a value θ^* of high posterior density into (5.35). Chib and Jeliazkov (2001) apply (5.35) to the output of the MH sampler. Friel (2013) use (5.35) to estimate the evidence the ERGM. After obtaining a MCMC sample from the posterior its density is evaluated using a non-parametric kernel density estimation approach which is restricted to models with not more than than five parameters. The evidence may be estimated as

$$\hat{p}(y) = \frac{q(y|\theta^*)p(\theta^*)}{\hat{z}(\theta^*)\hat{p}(\theta^*|y)} \tag{5.36}$$

where $\hat{p}(\theta^*|y)$ is a kernel density estimate of the posterior at the point θ^*. Both the Laplace approximation by Thiemichen et al. (2016) and the usage of (5.36) require $\hat{z}(\theta^*)$ which is an explicit estimate of the normalizing constant of the ERGM likelihood, see section 5.3.2. Friel (2013) estimate the ERGM evidence for non-elliptic posterior distributions where a Laplace approximation could not be applied. The major drawback of (5.36) is that the non-parametric kernel density estimation proposed by Friel (2013) is limited to models with a low number of parameters.

Green (1995) propose reversible jump MCMC (RJ-MCMC) which exploits the joint distribution of all models of interest where the joint parameter space may contain models of different dimensionality. A single MCMC chain is constructed crossing the joint model and parameter space of concurring models. Hastie and Green (2012) give a review of RJ-MCMC methods for Bayesian model selection. Caimo and Friel (2013) apply RJ-MCMC to Bayesian ERGM comparison for nested models. It can be very difficult to achieve reasonable mixing across the model space and practical implementation of RJ-MCMC may be hindered by the sensitivity to the specification of the jump proposal.

There are many more MC and MCMC methods to compute the evidence which have not been applied to the ERGM. A popular MC technique is annealed importance sampling introduced by Neal (2001) which uses a tempering transition similar

to path sampling. The MC methods of annealed important sampling and more recently sequential MC methods, see Del Moral et al. (2006), are powerful tools for computing the evidence as they are computationally more efficient than MCMC methods, see Friel and Wyse (2012) and Everitt et al. (2016) for a review. In contrast to MCMC samplers which generate autocorrelated samples, MC methods are easy to parallelize and to run on multi core computers. Tavaré et al. (1997) propose approximate Bayesian computation (ABC) to estimate the evidence where the likelihood $p(y|\theta)$ is approximated by simulation methods using summary statistics $s(y)$, Marin et al. (2012) for a review. To our knowledge ABC has not been applied to the ERGM. An advantage could be that suitable summary statistics $s(y)$ to evaluate the simulations are already defined for the ERGM class.

5.3 The power posterior exchange algorithm

The PPEA-EEL consists of two main steps. First, MCMC samples have to be simulated from the tempered posteriors along the path of the temperature steps (power posterior step), see section 5.3.1. After obtaining these simulations, in a second step importance sampling is used to estimate the normalizing constants of the tempered likelihoods

$$z(\theta_j|t_j),\ j = 1, \ldots, J$$

which which allows for an explicit evaluation of the likelihoods (EEL) and the estimation of $\mathrm{E}_{\theta|y,t_j}\left[\ln p(y|\theta_j)^{t_j}\right]$. The evidence of an ERGM can be estimated using (5.31). The scheme of the PPEA-EEL is described in algorithm (4).

In contrast to Friel (2013) and Thiemichen et al. (2016) the PPEA-EA does not require a posterior density approximation as it solve the integral over the parameter space using (5.30). It can be applied to non-elliptic posteriors where the assumptions for a Laplace approximations are not met.

5.3.1 Power posterior step

In this section power posterior sampling is combined with the EA in order to estimate the ERGM evidence. The trapezoidal approximation to the evidence (5.32) requires MCMC samples from the tempered posteriors

$$p(\theta_j|y, t_j),\ j = 1, \ldots, J.$$

Algorithm 4: The PPEA-EEL approach

1 Specify a discrete transition on J steps from $t_{j=1} = 0$ to $t_{j=J} = 1$.

2 **PPEA step:** Use the PPEA to sample

$$\theta_j, \ (j = 1, \ldots, J)$$

from the tempered posterior $p(\theta_j|y, t_j)$, see algorithm (5).

3 **EEL step:** Use importance sampling to estimate the normalizing constants

$$z(\theta_j|t_j), \ (j = 1, \ldots, J)$$

of the tempered likelihoods, see equation (5.42) .

4 Estimate $\mathrm{E}_{\theta|y,t_j}\left[\ln p(y|\theta_j)^{t_j}\right], \ (j = 1, \ldots, J)$.

5 Estimate the log evidence $\ln p(y)$ using a trapezoidal approximation, see equation (5.31).

As the normalizing constant of the ERGM likelihood is not available the EA, see algorithm (2), is applied at every temperature step t_j. The algorithmic scheme of the PPEA is described in algorithm (5). `R` code for the PPEA implemented in this thesis can be found in the attached DVD.[5] For each temperature t_j the PPEA samples from an augmented posterior distribution

$$p(\theta_j, \theta_j^*, y^*|y, t_j) \propto p(y|\theta_j)^{t_j} p(\theta_j) p(y^*|\theta_j^*)^{t_j} \mathcal{H}(\theta_j^*|\theta_j, t_j). \tag{5.37}$$

θ_j^* is a proposed vector of model parameters sampled from a proposal distribution $\mathcal{H}(\theta_j^*|\theta_j, t_j)$. V parallel adaptive direction sampling (ADS) chains may be used in order to improve convergence, see section 4, which helps to explore the support of the tempered posterior. y^* is an auxiliary network simulated conditional on θ_j^* from $p(y|\theta_j^*)^{t_j}$. As perfect sampling from the ERGM is not possible, the tie-not-tie sampler (Hunter et al., 2008b) can be applied to sample auxiliary network data, see section 2.5.1. y^* is used in the exchange move, see step **7** of algorithm (5), to circumvent the evaluation of the intractable ERGM likelihood. The temperature spacing is constructed in such a way that the KL-divergence between subsequent temperature steps (t_{j-1}, t_j) is minimized, see section 5.2.4. Thus the chains in step

[5] Note that this implementation requires parallel multi core computation and will not run on a single core.

Algorithm 5: The power posterior exchange algorithm

Sample $\theta_{j=1}^{(i=1)}, \ldots, \theta_{j=1}^{(i=I)}$ from the prior $p(\theta)$.
for $j = 2, \ldots, J$ **do**
  Initialize $\theta_j^{(i=1)}$
  for $i = 2, \ldots, I$ **do**
    Gibbs update of θ_j^* and $y^*|\theta_j^*, t_j$:
    Draw $\theta_j^* \sim \mathcal{H}\left(\theta_j^* | \theta_j^{(i-1)}\right)$ where $\mathcal{H}(\cdot|\cdot)$ is assumed to be symmetric.
    Draw $y^* \sim p(y^*|\theta_j^*)^{t_j}$.
    Exchange move:
    Move from $\theta_j^{(i-1)}$ to θ_j^* with probability

$$a_{PPEA} = \min\left\{1, \frac{p\left(\theta_j^*\right)}{p\left(\theta_j^{(i-1)}\right)} \frac{q\left(y|\theta_j^*\right)^{t_j}}{q\left(y|\theta_j^{(i-1)}\right)^{t_j}} \frac{q\left(y^*|\theta_j^{(i-1)}\right)^{t_j}}{q\left(y^*|\theta_j^*\right)^{t_j}}\right\}.$$

    Repeat step 7 until $a_{PPEA} = 1$, count the number of retries.
  end
end

3 of algorithm (5) of the temperature (j) may be initialized using the mean of the chains at temperature $(j-1)$. After convergence the state j is independent of the state $j-1$. At temperature $j=1$ the tempered posterior is equal to the prior. Usually a normal prior centered to zero will be used which does not require the EA to be simulated from, see step **1** of algorithm (5).

5.3.2 Explicit evaluation of the intractable likelihood

The normalizing constant $z(\theta)$ of the ERGM likelihood is impossible to evaluate analytically in most cases as it requires summation over all possible network graphs in $\mathcal{Y}$, see section 2.3. In the EEL step of the PPEA-EEL importance sampling is used to explicitly estimate the normalizing constants of the tempered ERGM likelihoods along the temperature path after obtaining MCMC simulations from the tempered posteriors. In contrast to implicit approaches of estimating the ERGM evidence, see Friel (2013) and Thiemichen et al. (2016), the proposed PPEA-EEL does not approximate the density of the posterior and does not require particular approximation assumptions. While MCMC methods like the EA circumvent the necessity to evaluate $z(\theta)$, explicit evaluation of the normalized likelihood is a prerequisite if one is interested in the estimation of the evidence $p(y)$ without relying on posterior approximations.

Importance sampling, see section 5.2.2, can be used to estimate the ratio of normalizing constants $z_1(\theta)/z_0(\theta)$ where $z_1(\theta)$ is the unknown normalizing constant of the likelihood which shall be evaluated explicitly. $z_0(\theta)$ is the known normalizing constant of the importance function. If all parameter values of an ERGM specification are equated to zero, it is known that the resulting normalizing constant is the number of possible graphs $\mathcal{G}$, see section 2.4.1. The non-normalized kernel of such an ERGM assuming independence of all tie variables may be used as importance function $q_0(y|\theta_0)$ with known normalizing constant $z_0(\theta) = \mathcal{G}$.[6] Using equation (5.18) the normalizing constant of interest can be estimated as

$$\hat{z}_1(\theta) = \frac{1}{R}\sum_{r=1}^{R}\frac{q_1(y^{*(r)}|\theta_1)}{q_0(y^{*(r)}|\theta_0)} \times z_0(\theta) \tag{5.38}$$

where $\left(y^{*(1)},\ldots,y^{*(R)}\right)$ is an importance sample simulated from $q_0(y|\theta_0)$. For al-

[6] As $\mathcal{G}$ can be a very large figure the `R` (R Development Core Team, 2011) package `Brobdingnag` (Hankin, 2007) is required for networks with $n \geq 32$ nodes.

most any ERGM the importance function $q_0(y|\theta_0)$ assuming tie variable independence will be a very bad choice to be used in (5.38) as a simulated network $y^{*(r)}$ will have little in common with the observed network y. This problem can be solved by using a discretized path of tempered importance functions which help to transition from the tractable importance function $q_0(y|\theta_0)$ to $q_1(y|\theta_1)$. The path of temperatures

$$(t_{j=1} = 0, \ldots, t_{j=J} = 1)$$

known from power posterior sampling can be used for that transition on J steps. A class of tempered importance functions $q(y|\theta_j)^{t_j}$ can be constructed with its normalizing constant $z(\theta_j|t_j)$. The ratio of normalizing constants

$$\frac{z_1(\theta)}{z_0(\theta)} = \frac{z(\theta_{j=J}|t_{j=J} = 1)}{z(\theta_{j=1}|t_{j=1} = 0)}$$

can be expressed as

$$\frac{z_1(\theta)}{z_0(\theta)} = \frac{z(\theta_2|t_2)}{z_0(\theta)} \times \frac{z(\theta_3|t_3)}{z(\theta_2|t_2)} \times \frac{z(\theta_4|t_4)}{z(\theta_3|t_3)} \times \cdots \times \frac{z(\theta_J|t_J)}{z(\theta_{J-1}|t_{J-1})}. \tag{5.39}$$

The direct estimation of $z_1(\theta)/z_0(\theta)$ is replaced by a product of ratios of neighbouring tempered normalizing constants. While the KL-divergence between the importance function and the target might be too large to make (5.38) an efficient estimate of $z_1(\theta)$, the discretized path on J steps used in (5.39) yields $J-1$ importance estimates of ratios where the importance function $q(y|\theta_j)^{t_j}$ and the subsequent target $q(y|\theta_{j-1})^{t_{j-}}$ have very small KL-divergence. The ratio of two neighbouring tempered normalizing constants can be approximated using the ratio of the corresponding tempered kernels

$$\frac{z(\theta_{j+1}|t_{j+1})}{z(\theta_j|t_j)} \approx \frac{q(y|\theta_{j+1})^{t_{j+1}}}{q(y|\theta_j)^{t_j}}. \tag{5.40}$$

For each temperature step

$$t_j,\ (j = 1, \ldots, J)$$

sample R importance simulations from the corresponding tempered importance function

$$y_j^{*(1)}, \ldots, y_j^{*(R)} \sim q(y|\theta_j^{(i)})^{t_j}$$

where the sample from the tempered posterior

$$\theta_j^{(i)},\ i = 1, \ldots, I$$

is obtained in step **2** of algorithm (4). This yields an unbiased importance estimate of the ratio of normalizing constants of neighbouring tempered likelihoods

$$\left(\frac{z\left(\widehat{\theta_{j+1}|t_j}_{+1}\right)}{z\left(\theta_j|t_j\right)}\right) = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{R}\sum_{r=1}^{R}\frac{q\left(y_j^{*(r)}|\theta_{j+1}^{(i)}\right)^{t_{j+1}}}{q\left(y_j^{*(r)}|\theta_j^{(i)}\right)^{t_j}}\right) \tag{5.41}$$

by averaging over all I simulations from the tempered posteriors $p(\theta_j|y, t_j)$. The normalizing constant of the likelihood of interest $z(\theta_{j=J}|t_{j=J})$ can be estimated as

$$\hat{z}(\theta_{j=J}|t_{j=J}) = \prod_{j=1}^{J-1}\left(\frac{z\left(\widehat{\theta_{j+1}|t_j}_{+1}\right)}{z\left(\theta_j|t_j\right)}\right) \times z_0(\theta) \tag{5.42}$$

where $z_0(\theta) = \mathcal{G}$. Equation (5.42) can be used to estimate any normalizing constant $z(\theta_j|t_j)$ on the temperature path. The tempered likelihood at t_j for a given value $\theta_j^{(i)}$ can be evaluated explicitly as

$$\hat{p}(y|\theta_j^{(i)})^{t_j} = \frac{q(y|\theta_j^{(i)})^{t_j}}{\hat{z}(\theta_j|t_j)}. \tag{5.43}$$

The expected value $\mathrm{E}_{\theta_j|y,t_j}\left[\ln p(y|\theta_j)\right]$ can be estimated by averaging (5.43) over all I MCMC simulations obtained with the PPEA, see equation (5.33). This result is used in the trapezoidal approximation (5.31) of the ERGM evidence, see section 5.2.4. `R` code for the EEL-step implemented in this thesis can be found in the attached DVD.[7]

5.4 Application

The PPEA-EEL is applied to estimate the ERGM evidence for Krackhardt's friendship network introduced in chapter 4 and the expert network in Ghana known from chapter 3. Estimates of the evidence are used to select various model specifications

[7] Note that the EEL-step requires draws from the tempered posteriors simulated in the PPEA-step. One such a sample can be found on the DVD.

over each other. This task is computationally very intensive: especially for the expert network in Ghana it becomes obvious that the PPEA-EEL is limited by the available computational resources. The runtimes of the PPEA-step and the EEL-step can both easily exceed several days on multi core machines. Graphical methods are introduced which allow for the evaluation of the estimated evidence. This is not a trivial task as multiple MCMC chains simulated on parallel computer cores have to be evaluated over many temperature steps. Our method helps to examine the behaviour of the PPEA and evaluate the reliability of the estimated evidence obtained in the EEL-step.

The PPEA-EEL can be applied to estimate the ERGM evidence where the assumptions required for a Laplace approximation are not met. In section 5.4.1 the result of the PPEA-EEL is compared to the results of Bayesian logit and probit models which use a Laplace approximation to yield an estimate of the evidence. Inspecting the respective PPEA density plots suggests that the assumption of elliptic posterior densities is met and that a Laplace approximation is in deed applicable. However, we refrain from Laplace approximations for the expert network in Ghana as the assumption of elliptical posterior distributions is not met. Inspecting the density plots of the PPEA samples in section D.4 makes clear that the densities tend to be skewed, especially for the edges parameters. In fact, ERGM posterior densities can be strictly non-elliptical, see the application in Friel (2013).

5.4.1 Krackhardt's managers

The PPEA-EEL is applied to the same two non-nested model specifications m_1 and m_2 which are estimated with the EA-ADS in chapter 4, see table 5.2. m_1 contains an intercept and two exogenous covariates and can also be estimated with a standard logistic regression approach. m_2 contains an intercept, the reciprocity statistic $M(y)$ and the $GWESP(y)$ statistic with $\alpha_E = 0.2$ which requires an ERGM approach to be estimated. The prior specifications used are identical to chapter 4. The schedule of the temperature path of the PPEA-EEL is specified with

$$\left(t_j = \frac{j}{J}\right)^5$$

where $J = 100$ for all specifications in this chapter, see figure 5.1.

In this section the graphical results are shown for m_2 which is of more inte-

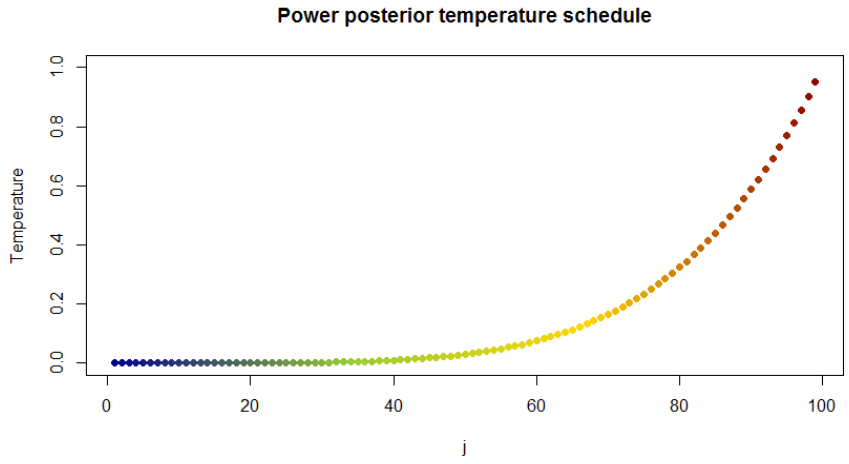

Figure 5.1: The temperature schedule of the tempering steps used for the PPEA with $t_j = (j/J)^w, w = 5$. Every temperature t_j has its own color assigned which will be used throughout this work. The blue colors represent temperatures very close to zero with $t_{j<20} < 0.0003$. The corresponding tempered posteriors are practically identical to the prior. Note that the first 50 temperature steps represent temperatures with $t_{j\leq 50} < 0.032$ so most tempered posteriors have very little data weight. Only during the last 25 temperature steps (orange to dark red colors) the data weight rapidly increases.

rest as it requires an ERGM specification and is more complicated to estimate. The plots for m_1 are shown in the appendix D.3. The MCMC simulations of the PPEA-step are computed on the CoolMUC-1 cluster of the Leibnitz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities. Each model specification is run on $c = 16$ cores in parallel. As the PPEA is a MCMC algorithm generating autocorrelated samples it cannot be parallelized. Our approach is to run independent PPEA simulations on each core which are merged after obtaining a certain number of simulations. Each core is simulating $V = 20$ parallel ADS chains of the PPEA over the $J = 100$ temperatures. Due to run time limitations every chain is run for only $I = 200$ iterations. This might appear a rather short chain length but it has to be pointed out that the EA-ADS approach used for the PPEA dramatically reduces the burn-in period, see section 4.3. All chains mix well and converge after less than $I_{burn} = 50$ iterations which is used as a default burn-in period. Traceplots of the edge parameter $\theta_{j,l=1}$ can be inspected in figure 5.2 for some temperatures. The Gelman-Rubin scale reduction factor $\hat{R}$, see Gelman et al. (2013), is computed for each parameter at each core at $j = J$ using the $V = 20$ parallel ADS chains. For all cores and parameter chains of m_1 and m_2 convergence

can be assumed as $\hat{R} < 1.04$. Figure 5.4 indicates the mean required retries to accept a draw over all $c \cdot V = 320$ chains of the PPEA at each temperature step j for m_2. At $j < 20$ the data weight is so low that the tempered target distribution is practically identical to the prior resulting in an average of about 2.5 retries required. This is equivalent to an acceptance rate of about 0.4. Note that for $j = 1$ the PPEA is not required as the parameter values $\theta_{j=1,l}$ are directly simulated from the prior. As the data weight increases it is getting harder to draw from $p(\theta|y, t_j)$. At $j > 70$ the number of required retries only slowly grows resulting in about 5 retries required on average. This pattern is typical for the PPEA. It is more difficult to sample from the tempered posteriors at higher temperatures, yet the increase is not linear. There are two regimes apparent, one closer to the prior and one closer to the posterior of interest.

The ACF for a sample of chains at some temperatures can be inspected in figure 5.5. Inspection of the ACF of all $P \cdot J \cdot c \cdot V = 96.000$ PPEA-ADS-chains over all parameters, temperatures and cores is impossible so aggregation is required. In figure 5.6 it can be seen how the mean ACF at lag=1, lag=2 and lag=3 changes with the temperature step j for all $c \cdot V = 320$ chains. The ACF increases with j where a sudden increase can be observed around $t_{j=60} = 0.078$. Note that at $t_{j=1} = 0$ the tempered posterior is identical to the prior and no MCMC techniques are required resulting in ACF values of zero. Figure 5.6 also gives evidence for the two regimes of difficulty in sampling from the tempered posteriors. We consider this novel visualization an efficient way to aggregate the ACF for path sampling approaches.

The independent simulations on the $c = 16$ cores are merged to a total sample of $c \cdot V \cdot (I - I_{burn}) = 48,000$ MCMC simulations for each $\theta_{j,l}$. For parameter estimation and the EEL-step the merged sample is thinned out to a size of $I_{thin} = 10,000$ simulations. Compare the ACF of the merged sample (5.7) and the thinned sample (5.8) at some temperatures for the edge parameter $\theta_{j,l=1}$.

The behaviour of the PPEA for ERGM estimation is not easy to understand and requires some illustration. We propose graphical analysis for the path of $\theta_{j,l}$ over the temperatures. To our knowledge this is the first kind of graphical inspection of the behaviour of power posterior algorithms like the PPEA. For the purpose of illustration the densities of the edge parameter $\theta_{j,l=1}$ of m_2 are plotted over the temperatures, see figure 5.9. The dark blue density corresponds to the prior distribution at $t_{j=1} = 0$ and the dark red density corresponds to the poste-

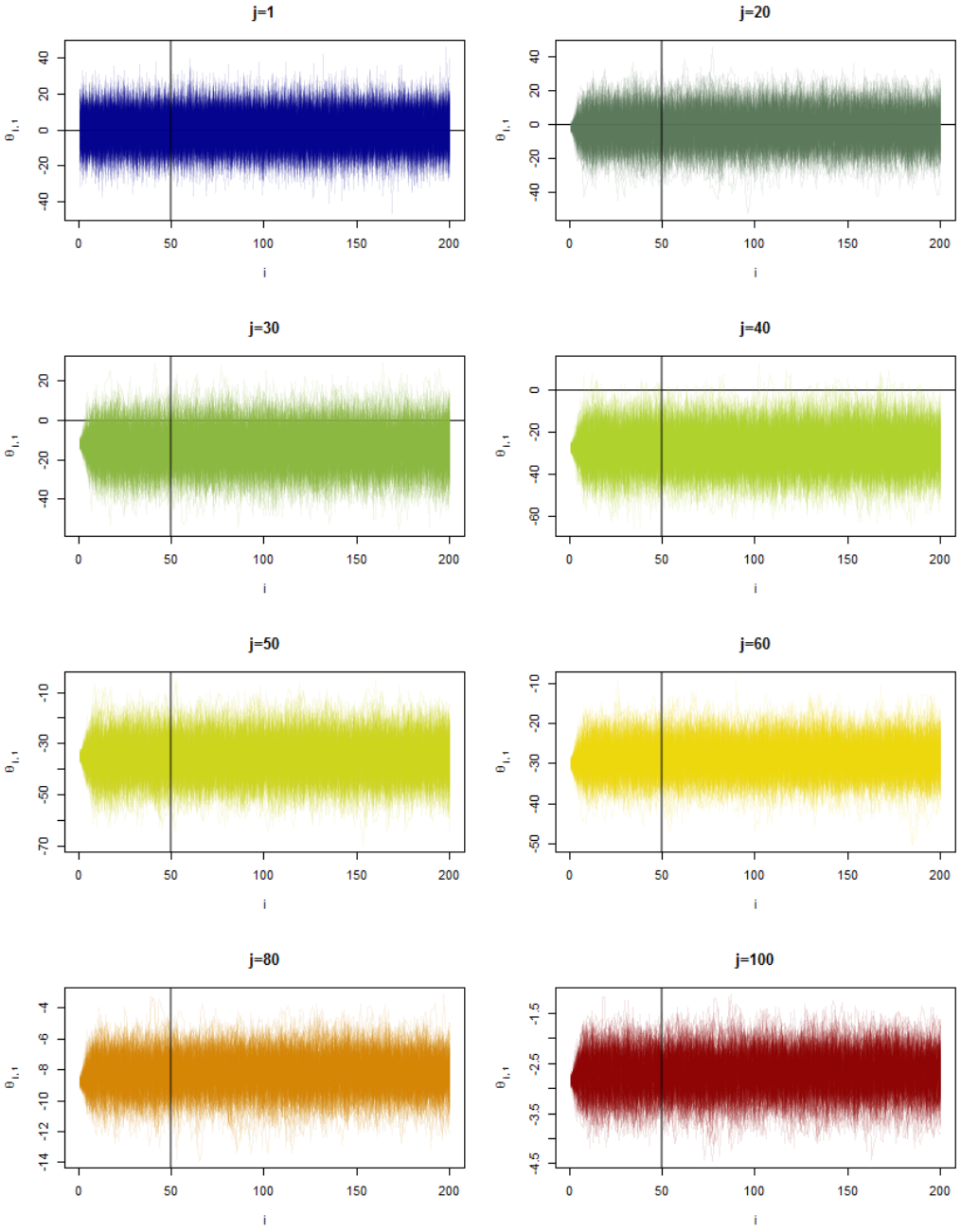

Figure 5.2: PPEA Krackhardt's managers, m_2:
Traceplots of the $c \cdot V = 320$ chains of the edge parameter $\theta_{j,l=1}$ at some temperatures j of the PPEA. The colors of the traceplots represent the respective temperature illustrated in figure 5.1. The horizontal grey line indicates the burn-in period of $I_{burn} = 50$ iterations used. Note that the draws at $j = 1$ are directly simulated from the prior and do not need the PPEA, thus the immediate convergence. At $j = 20$ the tempered posterior is still very close to the prior, both in location and scale.

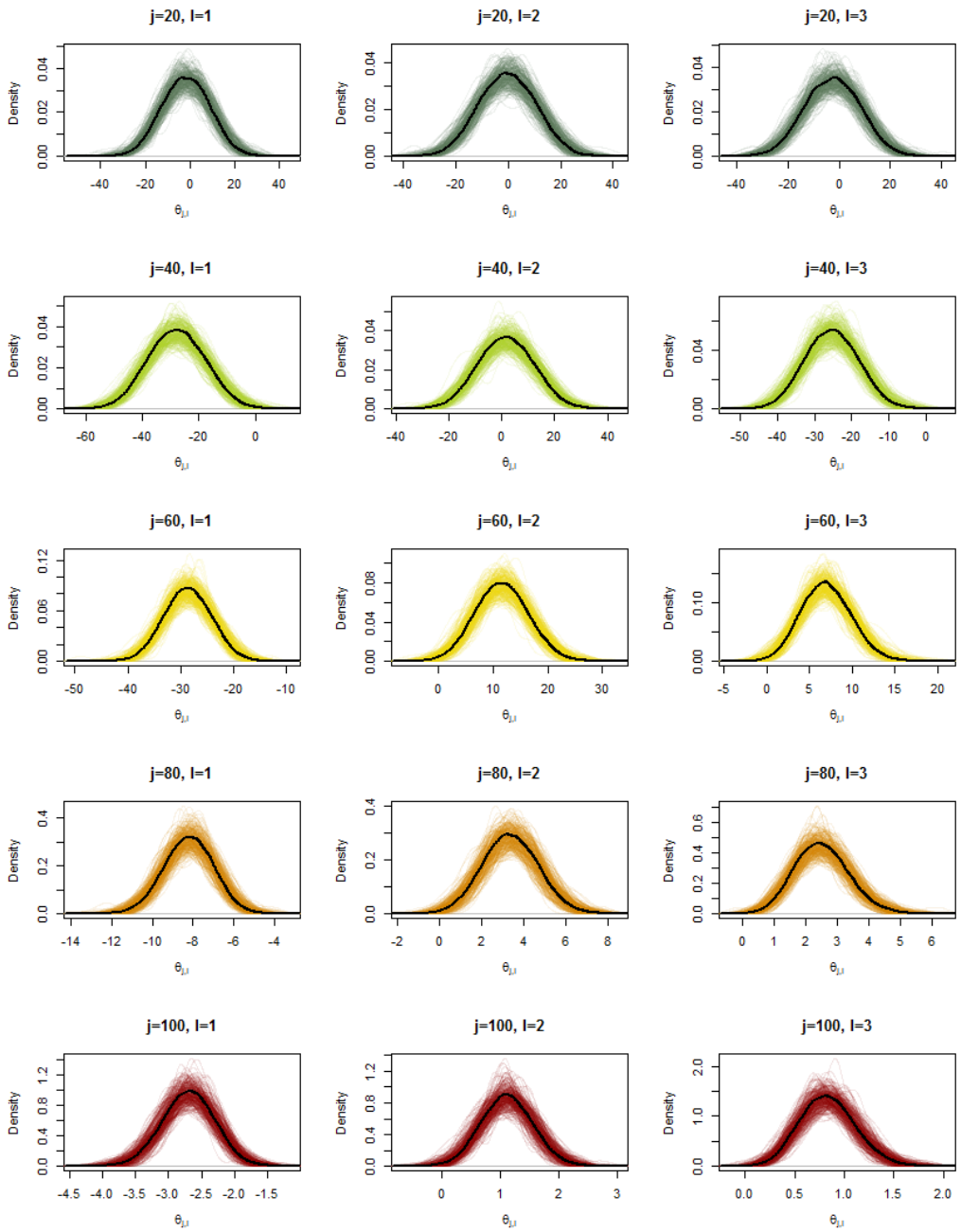

Figure 5.3: PPEA Krackhardt's managers, m_2:
Density plots of the $c \cdot V = 320$ chains of the parameters $\theta_{j,l}$ at some temperatures j of the PPEA after discarding the first $I_{burn} = 50$ iterations. The black line indicates the density of the merged sample.

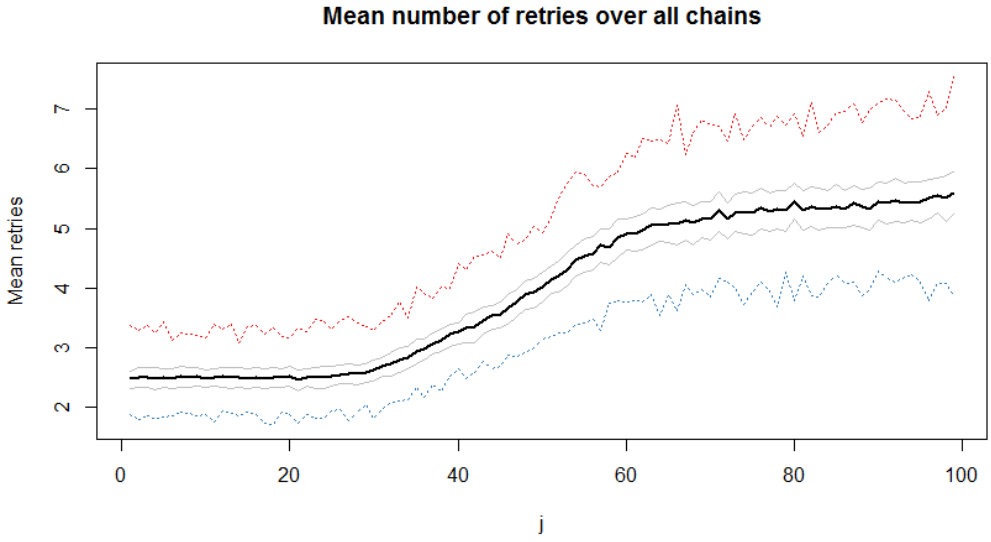

Figure 5.4: PPEA Krackhardt's managers, m_2:
Mean required retries to accept a draw over all $c \cdot V = 320$ chains of the PPEA at each temperature step j. The thick black line indicates the mean value, the thin grey lines indicate the 25% and the 75% quantiles, the dashed blue line indicates the minimum and the dashed red line the maximum of required retries at temperature step j.

rior of interest at $t_{j=J} = 1$. Note the shift in location and scale of the densities $p(\theta_{j,l=1}|y, t_j)$. The parameter estimates of $\theta_{j,l}$ reach huge values around tempering step $j = 60$, see figure 5.10. The dramatic changes in the value of $\theta_{j,l}$ occurs at rather low temperatures $t_j < 0.2$ with $j < 73$, see figure 5.12. The rapid change of tempered posterior parameter values has to be captured properly by the temperature spacing. This is the reason why so many temperature steps are concentrated at very low values of t_j which are much closer to the prior than to the target posterior, see Calderhead and Girolami (2009) on this issue. The impact of the large parameter values $\theta_{j,l},\ j < 73$ on the tempered posteriors are compensated by very low data weights $t_j < 0.2,\ j < 73$. Note that the non-normalized kernel of the tempered ERGM log likelihood is

$$\ln q(y|\theta_j)^{t_j} = \ln \left\{ \exp \left\{ (\theta_j \cdot s(y))^{t_j} \right\} \right\} = t_j \cdot \theta_j \cdot s(y) \tag{5.44}$$

with the constant vector of sufficient statistics $s(y)$. The path of the product $t_j \cdot \theta_j$ over the temperature steps is rather smooth for the edge parameter $\theta_{j,l=1}$ and the reciprocity parameter $\theta_{j,l=2}$, see figure 5.11. The dramatic behavior of $\theta_{j,l}$ almost

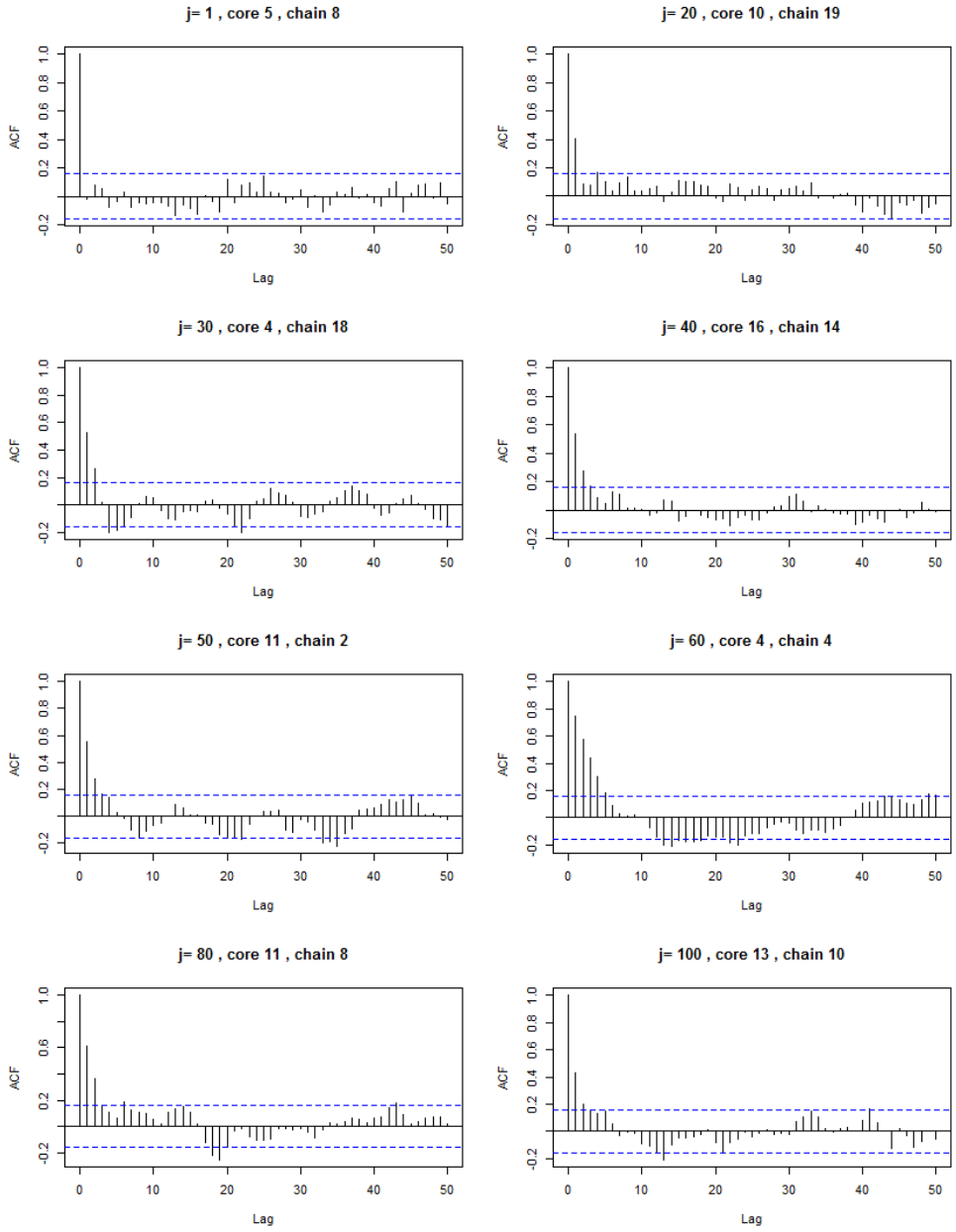

Figure 5.5: PPEA Krackhardt's managers, m_2:
ACF of a sample of the $c = 16$ cores and the respective $V = 20$ ADS chains.

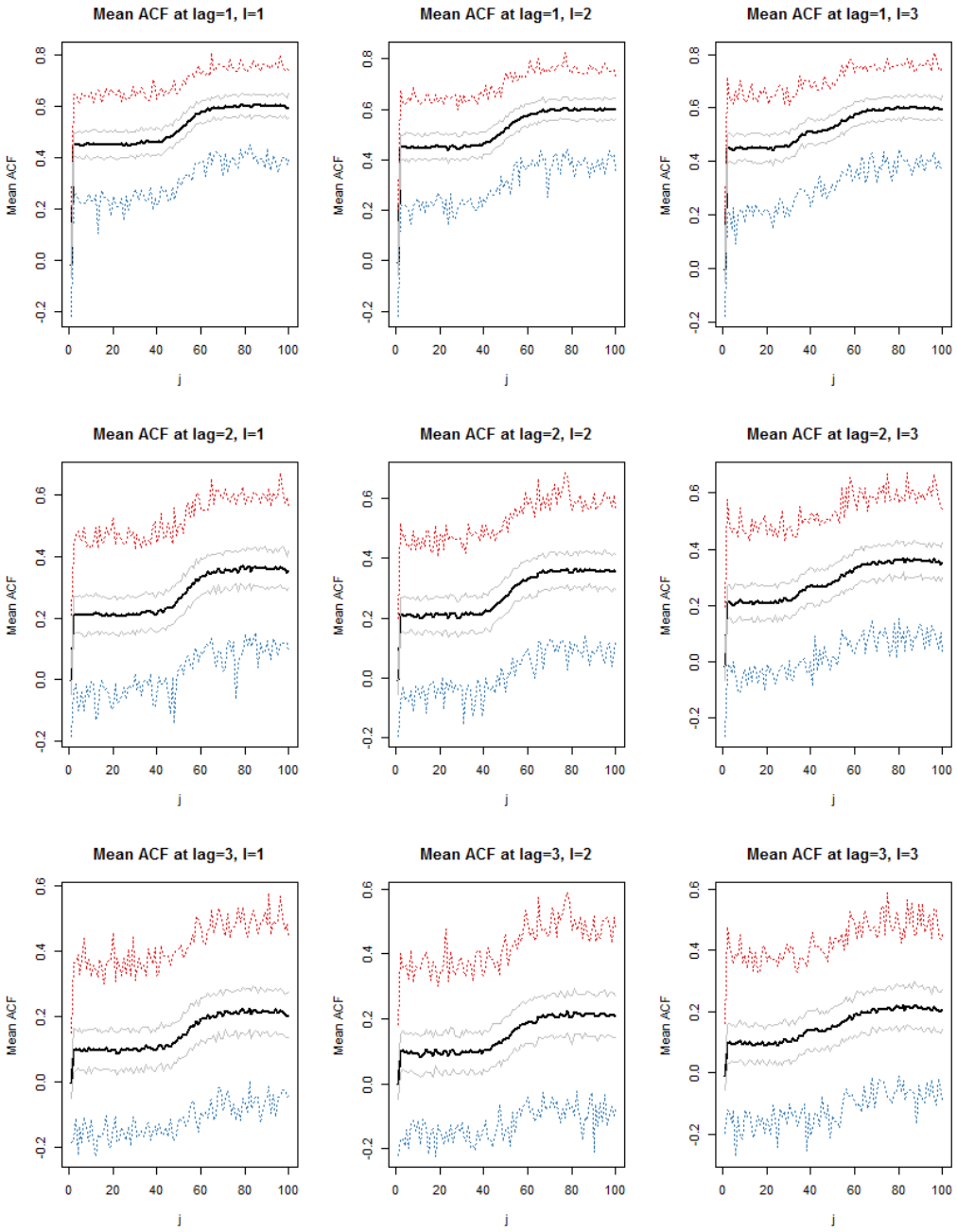

Figure 5.6: PPEA Krackhardt's managers, m_2:
Black line: mean value of the ACF of all $c \cdot V = 320$ PPEA chains of $\theta_{j,l}$. *Grey lines*: upper and lower quartile. *Red line*: maximum ACF. *Blue line*: minimum ACF.
Upper panel: lag=1. *Middle panel:* lag=2. *Lower panel:* lag=3.

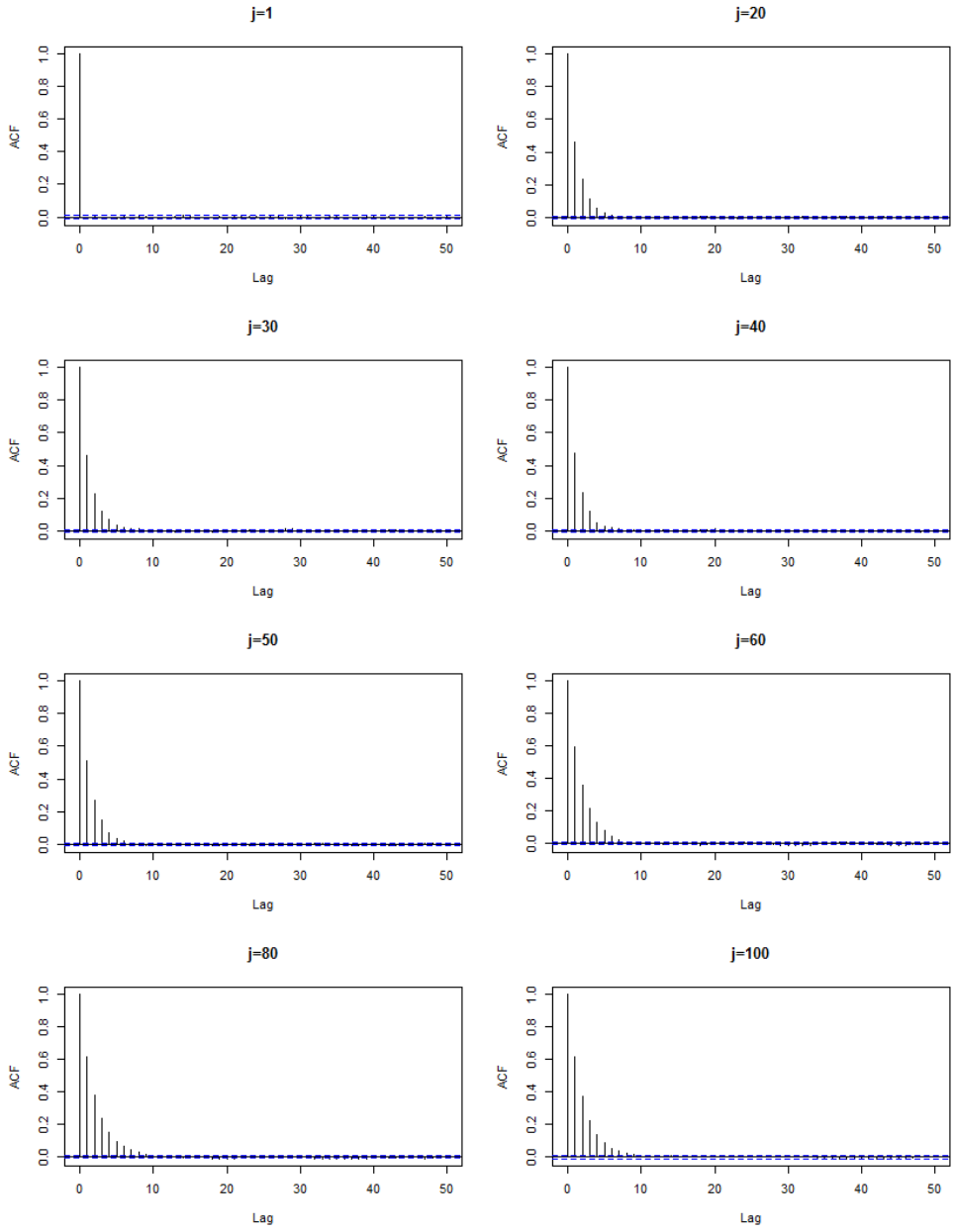

Figure 5.7: PPEA Krackhardt's managers, m_2:
ACF of the merged chains at some temperatures j. For each temperature the $c \cdot V = 320$ chains are merged to a single sample of size $I_{merge} = 48,000$.

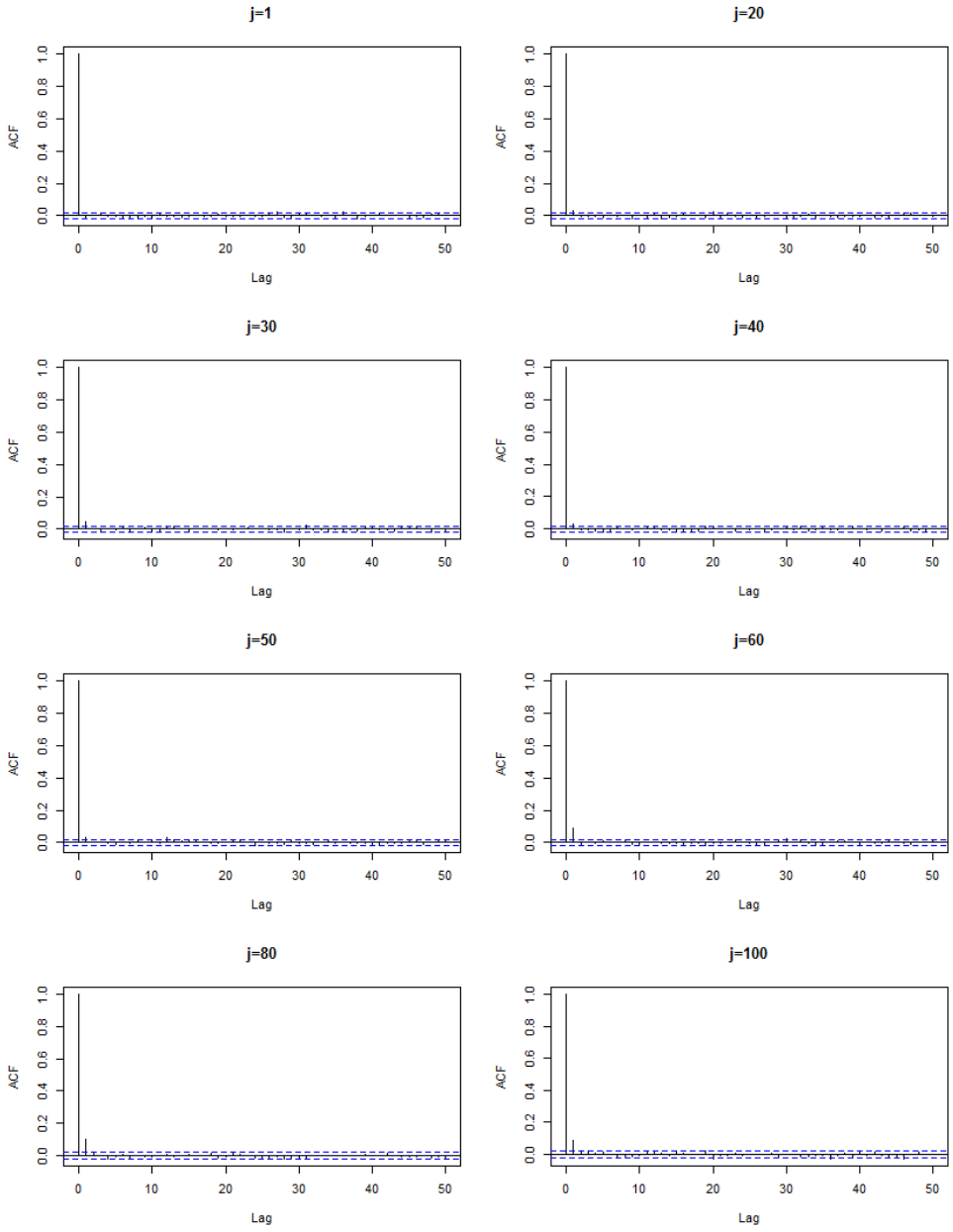

Figure 5.8: PPEA Krackhardt's managers, m_2:
ACF of the thinned merged chains at some temperatures j. For each temperature the merged chain is thinned out to have sample size $I_{thin} = 10,000$. The thinned out sample is used for MCMC parameter estimation, see table 5.2.

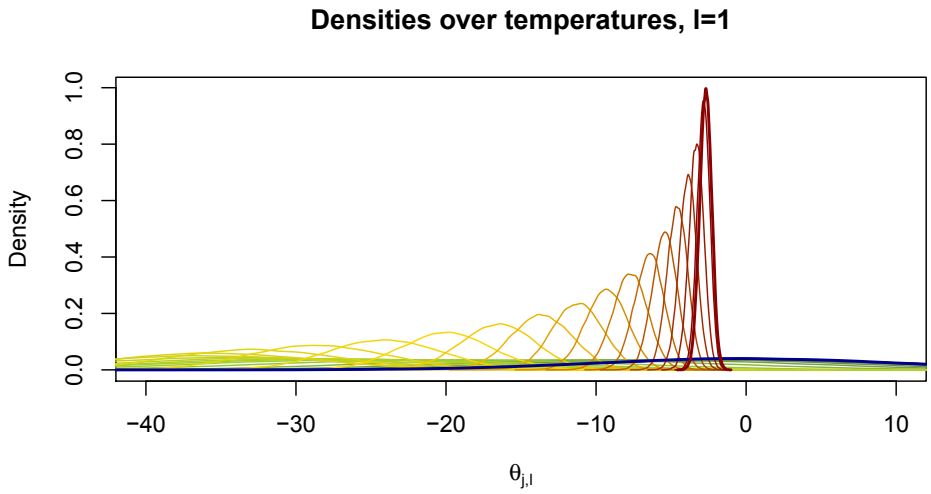

Figure 5.9: PPEA Krackhardt's managers, m_2:
Densities of the merged PPEA draws of the edge parameter $\theta_{j,1}$ transiting over the temperature path. The dark blue line represents draws from the prior at $j = 1$ and the dark red line represents draws from the target posterior at $j = 100$. This visualization highlights the path from the prior to the target posterior both in location and scale of the tempered posteriors.

completely disappears if the path of $t_j \cdot \theta_{j,l}$ is plotted over the temperature t_j, see figure 5.13. Note that the path of $t_j \cdot \theta_{j,l}$ at $t_j > 0.2$ is not perfectly smooth at $t_j > 0.4$. This might be caused by the restricted sample size of the PPEA due to run time limitations.

The EEL-step uses importance sampling to estimate the normalizing constants of the tempered likelihoods. Due to strict run time limitations of 48 hours the trajectories $\theta_{j,l}^{(1)}, \ldots, \theta_{j,l}^{(I)}$ are thinned out using only $I_{thin} = 10,000$ samples

$$\theta_j^{(i)}, \ (i = 1, \ldots, I_{thin} = 10,000), \ (j = 1, \ldots, J = 100)$$

of the $I = 48,000$ samples generated with the PPEA-step. The TNT sampler is used to simulate the R required networks using a burn in period of $N = 420$ draws and a thinning interval of 100 draws. This has to be done $J \cdot I_{thin} = 10^6$ times obtaining $J \cdot I_{thin} \cdot R = 10^8$ simulated networks which is a computational intensive task. As the 10^6 TNT samples are independent from each other this task can easily be parallelized on a multi core computer. The EEL-step is run on the CoolMUC-2 cluster on $c = 28$ cores. For m_1 and m_2 the simulations finish within the run time

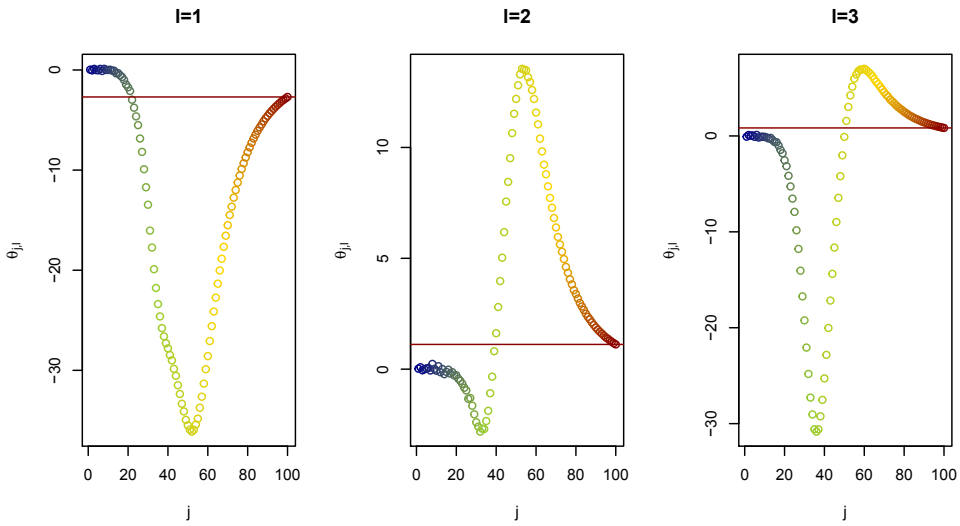

Figure 5.10: PPEA Krackhardt's managers, m_2:
MCMC estimates of $\theta_{j,l}$. The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $j = 100$.

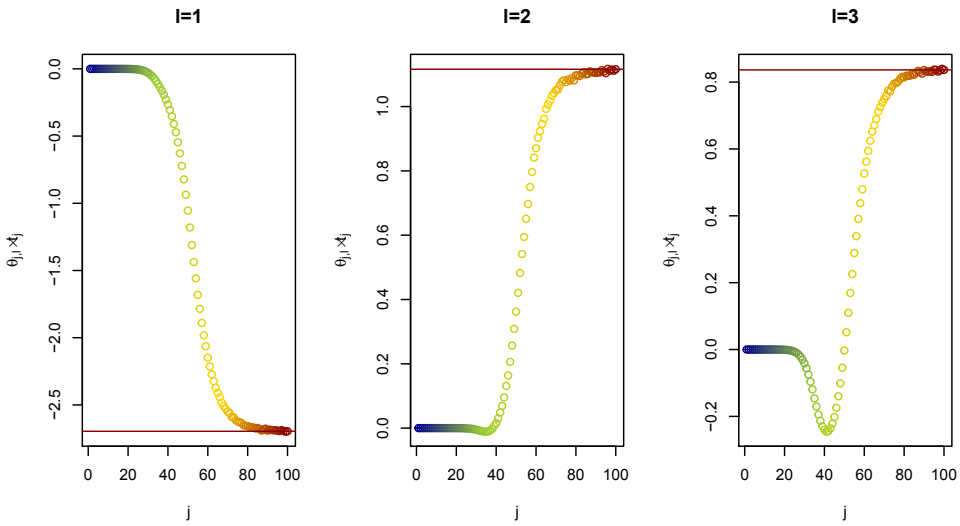

Figure 5.11: PPEA Krackhardt's managers, m_2:
MCMC estimates of $\theta_{j,l}$ multiplied by t_j. This visualization highlights how the parameter estimates of the tempered posterior influence the tempered likelihood through $t_j \cdot \theta_{j,l}$. The horizontal dark red line indicates $t_j \cdot \theta_{j,l}, \quad j = J$.

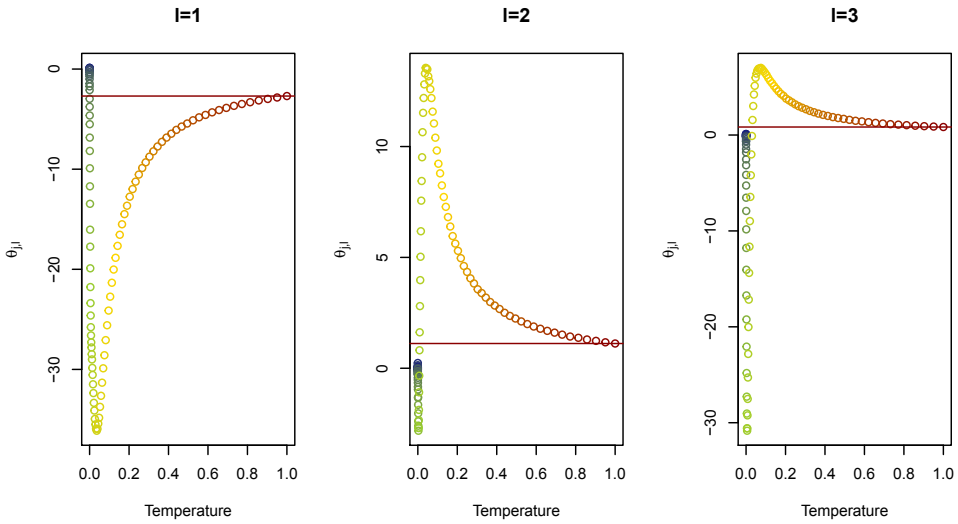

Figure 5.12: PPEA Krackhardt's managers, m_2:
Path of the MCMC estimates of $\theta_{j,l}$ in over t_j. In contrast to figure 5.11 it is obvious that the dramatic changes in $\theta_{j,l}$ occur at very low temperatures where the data have little influence on the posterior. The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t_{j=J} = 1$.

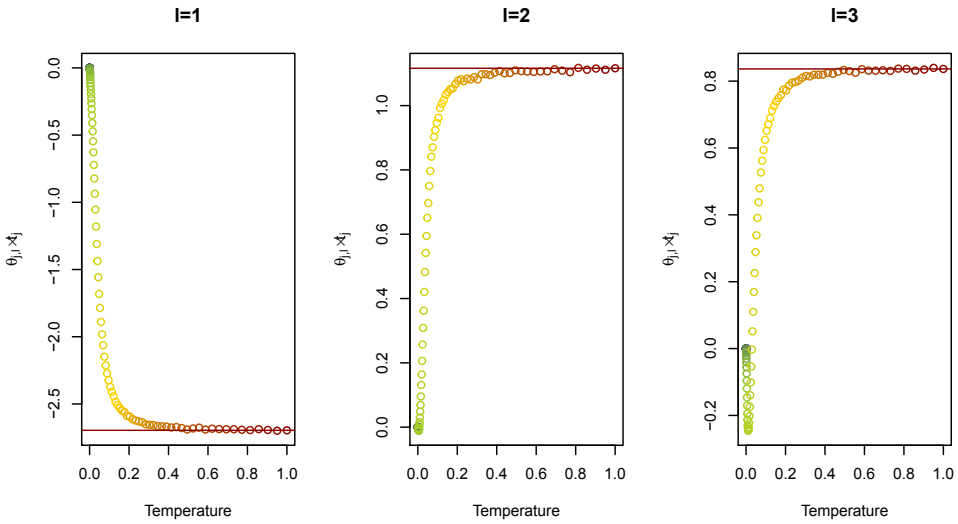

Figure 5.13: PPEA Krackhardt's managers, m_2:
MCMC estimates of $\theta_{j,l}$ multiplied by t_j over the temperature. At temperatures $t_j > 0.5$ the tempered likelihoods are practically identical to the target likelihood at $t_{j=J} = 1$. The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t_{j=J} = 1$. The horizontal dark red line indicates $t_j \cdot \theta_{j,l}, \quad j = J$.

limit of 48 hours.

We also offer graphical analysis of the EEL-step. The estimates

$$\hat{E}_{\theta_j|y,t_j}\left[\ln p(y|\theta_j)\right]$$

given the PPEA samples, see the upper panel in figure 5.14, reflect the unsmooth path of $t_j \cdot \theta_{j,l}$ at the highest temperature steps. The path of the log normalizing constants $\ln \hat{z}(\theta_j|t_j)$ estimated in the EEL-step is illustrated in the middle panel of figure 5.14. $\hat{E}_{\theta_j|y,t_j}\left[\ln q(y|\theta_j)\right]$ and $\ln \hat{z}(\theta_j|t_j)$ yield the normalized estimate

$$\hat{E}_{\theta_j|y,t_j}\left[\ln p(y|\theta_j)\right] = \hat{E}_{\theta_j|y,t_j}\left[\ln q(y|\theta_j)\right] - \ln \hat{z}(\theta_j|t_j).$$

The respective path is shown in the lower panel of figure 5.14. These expected values are used for the trapezoidal approximation of the evidence by summing up the elements ξ_j, see equation (5.32) and (5.33). ξ_j is the average of neighbouring expected tempered log likelihood values weighted by the respective stepsizes $(t_j - t_{j-1})$. In figure 5.15 (upper panel) the path of $\hat{E}_{\theta_j|y,t_j}\left[\ln p(y|\theta_j)\right]$ over the temperature is illustrated where the vertical grey lines indicate the respective stepsize. This highlights how little values at low temperatures contribute to the estimation of the evidence and how large the weight of the last temperature steps is. The middle panel of figure 5.15 shows the path of ξ_j over the temperature steps and the lower panel shows the same path but in relation to the temperature t_j. Both paths look rather smooth, only the very last value of x_j seems to be a bit too large. The results for m_1 look very similar, see the figures in appendix D.3.

The proposed graphical methods illustrate the behaviour of the PPEA and can be used heuristically to evaluate the quality of the PPEA-EEL estimate of the evidence. If the paths of the tempered parameter estimates, the path of the normalizing constants of the tempered likelihoods and the path of the elements ξ_j look very noisy, the result $\hat{p}(y)$ should be questioned.

Table 5.2 summarizes the results of the PPEA-EEL for m_1 and m_2. As m_1 can be estimated with a standard logistic regression approach the results of Bayesian logit and probit regression model estimation[8] are compared to the results of the PPEA-EEL and the EA-ADS. For both m_1 and m_2 the results of the PPEA-EEL

[8]Bayesian logit and probit model estimation is done using the `R` package `MCMCpack`. Default settings are used, the results in table 5.2 are based on 10^6 MCMC draws and a burn in period of $2 \cdot 10^5$ draws. The log evidence is estimated using a Laplace approximation.

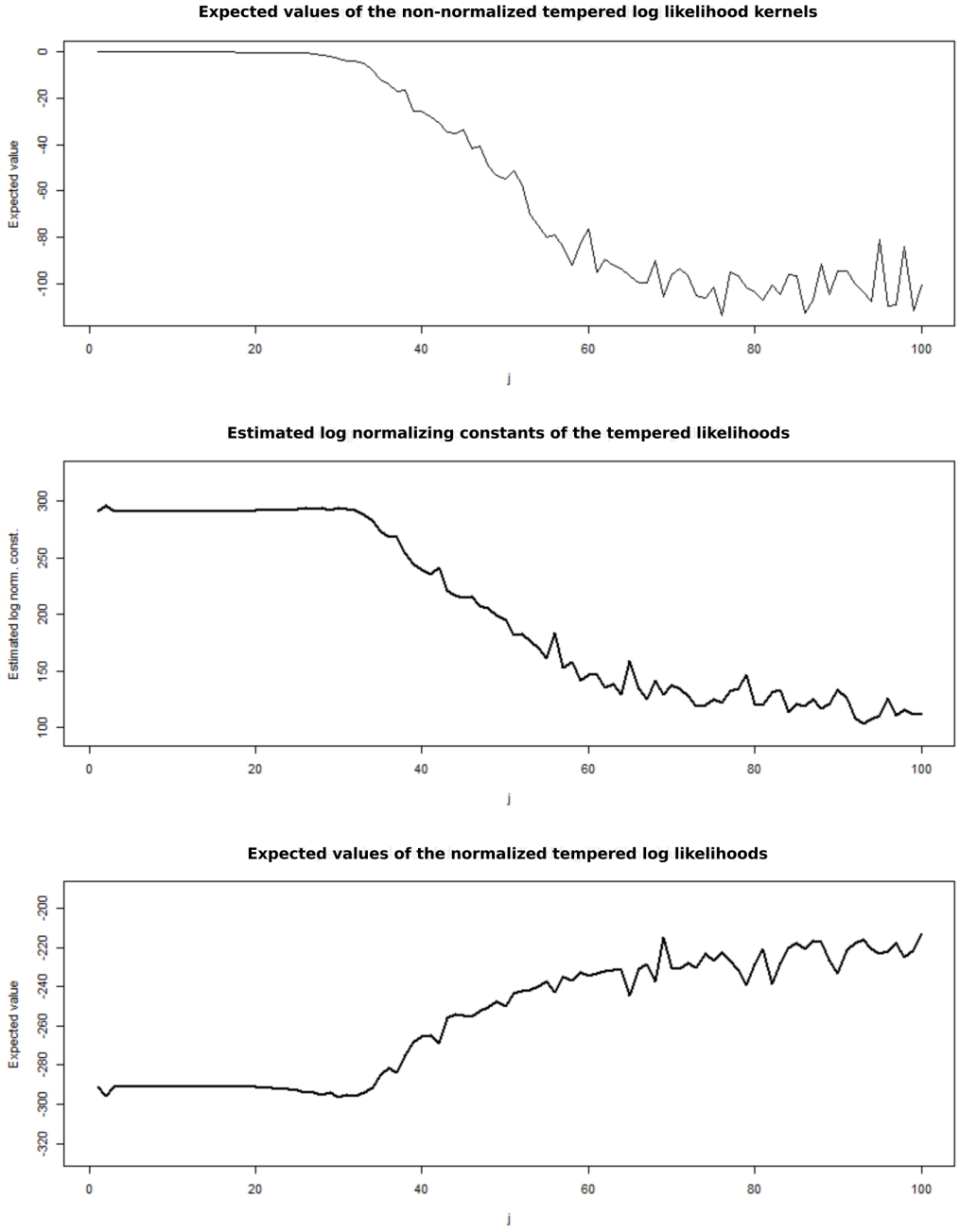

Figure 5.14: PPEA-EEL Krackhardt's managers, m_2:
Top panel: Expected values of the non-normalized log likelihood kernels.
Middle panel: Importance sampling estimates of the normalizing constants of the tempered log likelihoods.
Lower panel: Expected values of the normalized tempered log likelihoods.

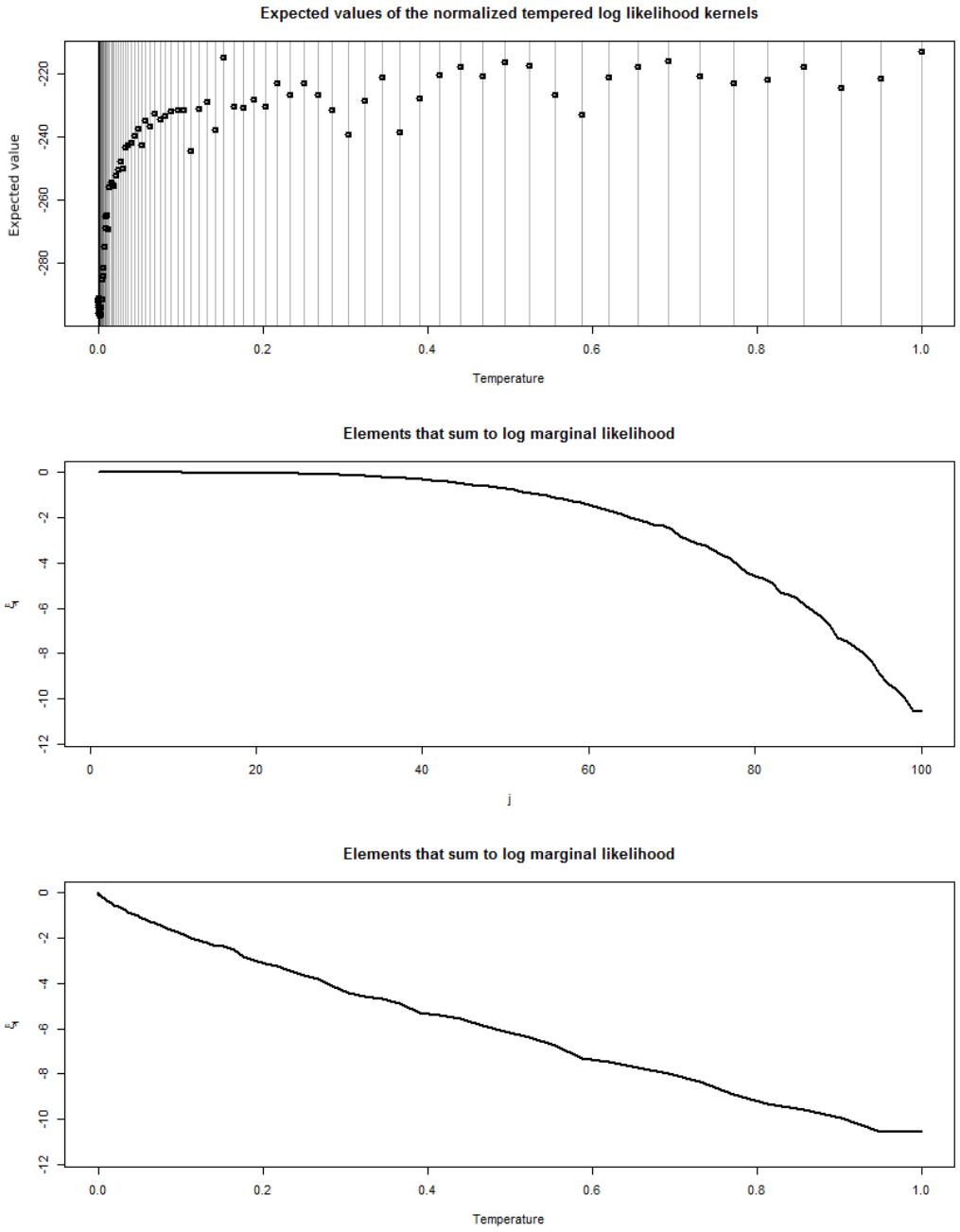

Figure 5.15: PPEA-EEL Krackhardt's managers, m_2:
Top panel: Expected values of the normalized log likelihood kernels in relation to the temperature steps $t_j - t_{j-1}$.
Middle panel: Path of ξ_j, the summands of the trapezoidal approximation.
Lower panel: Path of ξ_j, the summands of the trapezoidal approximation in relation to the temperature.

Table 5.2: Krackhardt's managers: PPEA-EEL results

	Model 1				**Model 2**	
	PPEA-EEL ERGM	EA-ADS ERGM	Logit	Probit	PPEA-EEL ERGM	EA-ADS ERGM
edges / intercept	−2.035*** (0.278)	−2.045*** (0.279)	−1.997*** (0.225)	−1.185*** (0.129)	−2.696*** (0.400)	−2.692*** (0.400)
same department	1.137*** (0.351)	1.139*** (0.345)	1.095*** (0.256)	0.612*** (0.155)		
same level	0.938*** (0.312)	0.944*** (0.313)	0.926*** (0.253)	0.502*** (0.148)		
reciprocity					1.116** (0.442)	1.113** (0.440)
GWESP					0.836*** (0.280)	0.834*** (0.282)
log evidence: $2\ln BF_{21}$	-228.4	-	-229.2	-230.3	-223.77 9.26	-
Notes:	MCMC estimates (MCMC sd) **$0 \notin$ 95% HDR; *** $0 \notin$ 99% HDR					

and the EA-ADS are very similar. Note that due to run time limitations large MCMC sample sizes are difficult to obtain for those approaches. This might explain the slight deviation of m_1 from the logit model which is based on 10^6 MCMC draws. For the logit and the probit model a Laplace approximation is used to estimate the log evidence. Theoretically all models should yield the same value of $\ln \hat{p}(y|m_1)$. The log evidence of the PPEA-EEL is close to the logit and the probit value. This result and the graphical analysis support the conclusion that the PPEA-EEL is able to yield a valid estimate of the model evidence. Using the scale of Kass and Raftery (1995) the Bayes factor gives strong evidence against m_1 as $2 \ln BF_{21} > 6$. Note that the two models are indistinguishable using standard posterior predictive checks, see chapter 4. The endogenous specification m_2 is superior in explaining tie variable formation to the specification m_1 using exogenous covariates.

5.4.2 The PEBAP expert network in Ghana

In this section the evidence for several ERGM specifications of the PEBAP expert network in Ghana is estimated. As this task is very intensive in computational resources we refrain from estimation the full model specification with 13 variables

from chapter 3. Instead, an endogenous baseline model specification m_1 is estimated including only three variables which is nested in all other model specifications $m_2, \ldots, m_6$, see table 5.3. For the ease of interpretation the preference similarity in m_6 is a dummy variable indicating high similarity for those pairs of nodes where the preference similarity used in chapter 3 is larger than its median value. We use the same specifications for the PPEA-EEL as in section 5.4.1 on the CoolMUC-1: each of the $c = 16$ cores is computing $V = 20$ EA-ADS chains for each $J = 100$ temperature. The respective chain length is $I = 200$ and the burn in period is $I_{burn} = 50$ resulting in $c \cdot V \cdot (I - I_{burn}) = 48,000$ MCMC simulations for each $\theta_l|t_j$. All MCMC simulations finish narrowly within six days which is the maximum run time allowed on the CoolMUC-1. Models with $P > 4$ parameters are not considered in this chapter: this would require less iterations or a reduction of the number of temperature steps J to let the the PPEA-step finish within six days. In section D.4 the traceplots and the density plots of the PPEA-step are shown. Note e.g. figures D.8 and D.14 (lowest panel): the densities of $\theta_{j,l}, \quad j = 100$ are skewed, the density parameters tend to have heavy tails to negative values whereas the $GWESP(y)$ parameters have heavy tails towards positive values. As the assumption of elliptical posterior distributions is violated we refrain from Laplace approximations in this section.

The importance sampling specifications from section 5.4.1 of the EEL-step appear to be insufficient for the expert network in Ghana. Initially $R_j = 1,000$ importance samples are specified for each temperature step j using $N = n^2 - n = 2,070$ simulations as burn-in period and a thinning interval of 100 draws. These simulations finish within 13 hours for all M specifications. However, with $R_j = 1,000$ the estimates of $z(\theta_j|t_j)$ and the resulting estimates of $p(y|m_h)$ are unstable for all M specifications. The EEL specification working well for the smaller network in section 5.4.1 seems not to suffice for the larger expert network with $n = 46$ nodes. The instability of $\hat{z}(\theta|t_j, m_2)$ and of $\hat{p}(y|m_2)$ is illustrated in figure 5.16 and figure 5.17. Additional importance simulations should yield a smoother path of ξ_j. As the total computation time is limited to 48 hours R_j is incrementally increased for $j \geq 50$ resulting in $R_{j=J} = 10,000$. The increase is proportional to the step length $t_j - t_{j-1}$ in equation (5.32): the larger the weight of a sample in the trapezoidal approximation of the evidence the larger the sample size R_j. For $j < 50$ the sample size is constant with $R_j = 1,000$. Using only $R_j = 1,000$ results in lower values for $\hat{z}(\theta|t_j, m_2)$, see the blue line in figure 5.16, and thus in higher values

Table 5.3: Expert network Ghana: PPEA-EEL results

	m_1	m_2	m_3	m_4	m_5	m_6
edges	-4.212^{***}	-4.357^{***}	-4.222^{***}	-5.219^{***}	-4.198^{***}	-6.259
	(0.468)	(0.463)	(0.465)	(0.545)	(0.463)	(6.734)
mutual	3.785^{***}	3.624^{***}	3.647^{***}	3.802^{***}	3.775^{***}	3.782^{***}
	(0.273)	(0.278)	(0.278)	(0.277)	(0.274)	(0.282)
GWESP	1.343^{***}	1.247^{***}	1.286^{***}	1.197^{***}	1.269^{***}	1.347^{***}
	(0.398)	(0.393)	(0.398)	(0.388)	(0.397)	(0.417)
support		1.164^{***}				
		(0.205)				
membership			1.295^{***}			
			(0.318)			
nodal reput.				3.355^{***}		
				(0.934)		
popul. exec.					0.504^{*}	
					(0.263)	
pref. simil.						2.041
						(6.741)
log likelihood:	-860.7	-683.4	-747.4	-762.8	-750.0	-834.8
AIC:	1727	1375	1503	1534	1508	1678
BIC:	1767	1428	1556	1587	1561	1731
log evidence:	**-859.8**	**-741.0**	**-757.8**	**-803.5**	**-802.9**	**-848.1**
$2\ln BF_{h1}$	-	237.6	204	112.6	113.8	23.4
$2\ln BF_{2h}$	237.6	-	33.6	125	123.8	214.2

Notes: MCMC estimates
(MCMC standard deviation)
* $0 \notin$ 90% HDR; ** $0 \notin$ 95% HDR; *** $0 \notin$ 99% HDR

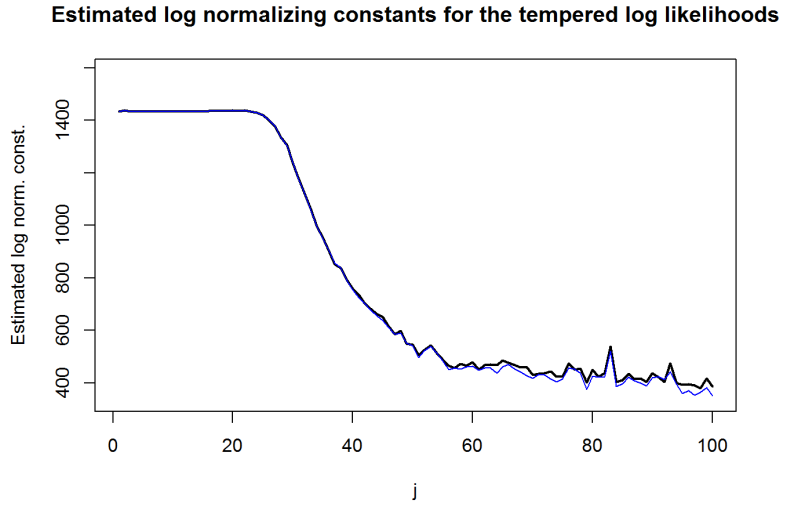

Figure 5.16: PPEA-EEL Ghana, model 2:
Estimates of the normalizing constants $z(\theta|t_j, m_2)$ of the tempered likelihoods $p(y|\theta_j, m_2)^{t_j}$. The blue line indicates the estimates with $R_j = 1,000$ for all temperatures. The black line indicates the results with an incremental increase in R_j for $j \geq 50$ with the final value $R_J = 10,000$.

of the respective tempered likelihoods. With $R_j = 1,000$ the path of ξ_j is not as smooth as with additional draws. The higher values of the tempered likelihoods also yield a higher estimate of the log evidence, see figure 5.17. Using the increased sample sizes the path looks smooth and the estimate of the evidence should be realistic, compare the black line and the blue line in figure 5.17. Note the substantial difference in the estimated values of the log evidence of 15.4. All simulations finish narrowly within the 48 hours time limit. Given the values of $\ln \hat{p}(y|m_h)$ it is obvious that m_2 should be selected over all other specifications. Compared to the endogenous baseline specification m_1 cooperation for political support is the most important exogenous covariate to explain network tie formation. Note that using the goodness-of-fit, compare figures 4.7 and 4.8 m_1 looks very similar to m_2. These plots are of little help if a clearly superior model has to be selected whereas $2 \ln BF_{21} = 237.6$ is decisive. As $2 \ln BF_{2h} \geq 33.6$ the evidence against all other models is decisive, too. Preference similarity is the least important exogenous covariate considered with an insignificant dummy variable of hight preference similarity. As the PPEA-EEL allows for an explicit evaluation of the ERGM likelihood $\hat{p}(y|\theta, m_h)$ the AIC 5.7 and the BIC 5.8 can be computed. The values

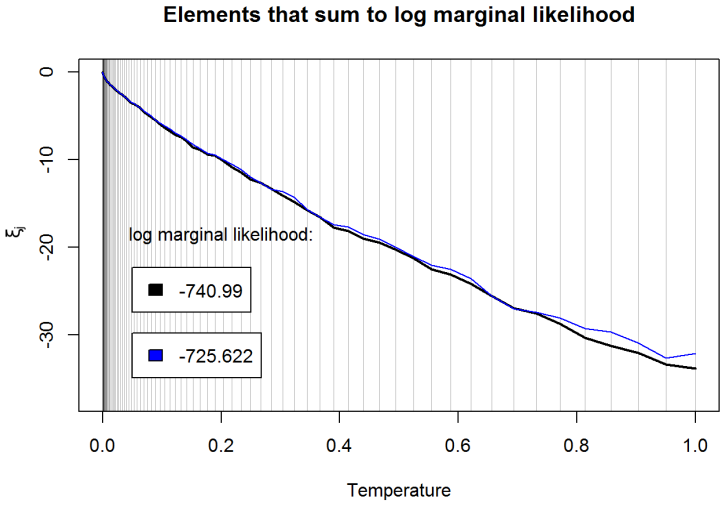

Figure 5.17: PPEA-EEL Ghana, m_2:
Path of ξ_j and the estimates of the evidence $p(y|m_2)$ resulting from different importance sample sizes used for the EEL. The blue color indicates results with $R_j = 1,000$ for all temperatures. The black color indicates results with an incremental increase in R_j for $j \geq 50$ with the final value $R_{j=J} = 10,000$. Note that $t_{j\leq 50} \leq 0.031$.

of the log likelihood, the AIC and the BIC are consistent with most estimates of the model evidence except for m_4 and m_5. With $2 \ln BF_{54} = 1.2$ the difference between those two models is not worth more than a bare mention. Considering the AIC and the BIC the endogenous model m_1 is almost as good as the endogenous model from chapter 3 (AIC=1694, BIC=1717), compare table 5.17 and table B.3. The additional $GWDSP(y)$ statistic does not give much improvement compared to m_1. Note that the estimates of the AIC and BIC in chapter 3 are based on the simplified path sampling algorithm by Hunter and Handcock (2006). m_2 is clearly superior to the full model from chapter 3 (AIC=1499, BIC=1573) which has 13 parameters.

5.5 Summary

In this chapter it is shown how to yield the marginal likelihood of an ERGM using power posterior sampling. As the ERGM likelihood is analytically not tractable the EA has to be applied in order to circumvent the evaluation of the unavailable ERGM normalizing constant. Path sampling is one possible technique to estimate

the marginal likelihood in Bayesian inference. Power posterior sampling is a version of path sampling which uses a discretized path on J steps to transition from the prior to the posterior. We introduce the PPEA in order to sample from the J tempered posteriors which allows for the estimation of the ERGM evidence. The temperature path used in the PPEA-step of the PPEA-EEL yields a sequence of tempered importance distributions. This allows for the estimation of the ERGM normalizing constant in the EEL-step. The PPEA-EEL can be applied to ERGM posterior distributions for which neither a Laplace approximation is possible nor a non-parametric density approximation would work.

New graphical methods are introduce for both the analysis of the PPEA-step and the EEL-step. These new methods help to illustrate the behavior of the PPEA while transiting over the temperature path. Heuristics are proposed indicating the reliability of the PPEA-EEL. The results of the PPEA-EEL are compared to Bayesian logit and probit regression. It can be concluded that the approach is able to yield a reliable estimate of the ERGM evidence if specified correctly. For the expert network in Ghana the most important exogenous covariate is identified by comparing partially nested models using the ERGM evidence.

If a superior ERGM specification has to be selected, the marginal likelihood has higher sensitivity than the posterior predictive GOF plots commonly used. In section 4.4.1 it is not possible to select a superior model for the Krackardt's managers network using such plots whereas in section 5.4.1 the PPEA-EEL yields strong evidence against m_1. In section 5.4.2 the evidence of m_2 against m_1 is decisive for the expert network in Ghana while using GOF plots the two models look very similar. The PPEA-EEL yields an explicit estimate of the ERGM likelihood which allows for a comparison to other approaches using the AIC and the BIC where no estimate of the marginal likelihood is available, e.g. maximum likelihood results.

The major drawback of the PPEA-EEL is the enormous computational effort required to estimate the ERGM evidence. Path sampling methods are per se computationally expensive. The PPEA-EEL requires estimation of the normalizing constants of the tempered likelihoods which roughly doubles the computational effort compared to power posterior sampling. Given the computational resources available for this work it is possible to estimate the evidence of ERGM specifications with $P \leq 4$ parameters and directed networks with $n < 50$ nodes. For larger networks and more complex ERGM specifications the PPEA-EEL is too impractical

as it would reach run times of several weeks. In this work it is not possible to compute the evidence of complex models analyzed in chapter 3.

Furthermore, the PPEA-EEL is not easy to implement as it consists of several steps which all require attentive specification and inspection. First of all, a discretized temperature path has to be constructed. The more temperature steps are chosen, the smaller the KL-divergence between subsequent tempered posteriors and thus the discretizational error of the trapezoidal approximation. However, increasing the number of steps also increases the computational effort. In this work heuristics introduced by Friel and Pettitt (2008) are applied. The PPEA-step requires tuning of the scale of the proposal distribution. We apply best practices proposed by Ter Braak and Vrugt (2008) which work well resulting in fast convergence of the PPEA chains. Inspecting the MCMC chains of the PPEA is not a trivial task as in this work it comprises the inspection of $32,000$ for each parameter used. We introduce graphical methods of aggregating the inspection of the ACF, the traceplots and densities of the tempered posteriors and the path of the MCMC parameter estimates. These new methods help to keep track of the behaviour of all PPEA chains. The EEL-step requires adaption of the importance sample sizes over the temperature path. The larger the temperature spacing between the tempered importance function and the subsequent target, the more samples should be used. This is a new insight to the EEL approach introduced by Friel (2013).

Chapter 6

Summary and discussion

The research aim of this thesis is to estimate the model evidence of the exponential random graph model (ERGM) and apply Bayesian model selection. The ERGM represents the distribution of a random graph on a fixed set of nodes using subgraph configurations as sufficient statistics. These statistics may contain counts of endogenous network configurations and counts of exogenous covariates. The ERGM can model various patterns of network tie formation under different assumptions of tie variable dependence. In most cases, the normalizing constant of the ERGM is analytically not available. This renders parameter estimation difficult as auxiliary network simulations have to be used in order to circumvent the evaluation of the intractable normalizing constant. Efficient simulation of network data and its usage for posterior predictive model evaluation is discussed. A commonly used Markov chain Monte Carlo maximum likelihood (MCMC-ML) approach for ERGM estimation, see Hunter and Handcock (2006), is introduced. This approach is plagued by model degeneracy which can cause non-convergence of the approximate Fisher scoring algorithm.

MCMC-ML ERGM estimation is applied to policy networks in Ghana, Senegal and Uganda. Hypothesized patterns of political communication are modeled using network statistics. This is the first kind of such an analysis in developing countries. There is evidence that the communication between actors is not as efficient as requested by political stake holders in the three countries. The study also shows the limitations of MCMC-ML ERGM estimation as model degeneracy impedes the estimation of some specifications.

Bayesian ERGM estimation using the exchange algorithm (EA) to circumvent

the intractable normalizing constant, see Murray et al. (2006), is robust to model degeneracy. The approach introduced by Caimo and Friel (2011) is illustrated and applied to the famous Krackardts's managers network, see Krackhardt (1987), and the expert network of political communication in Ghana. Posterior predictive goodness-of-fit plots are applied which are not always helpful in selecting a superior ERGM specification.

In this thesis, a new approach for Bayesian ERGM estimation is proposed which yields an estimate of the marginal likelihood. This quantity has the interpretation of the evidence of a particular model and may be used for Bayesian model selection. Power posterior sampling introduced by Friel and Pettitt (2008) is a version of path sampling, see Gelman and Meng (1998) and can be used to estimate the model evidence. A discretized path is defined transiting from the specified prior distribution to the posterior of interest. A trapezoidal approximation can be used to integrate over the parameter space of the posterior and yield an estimate of the evidence. We propose a combination of power posterior sampling and the EA to simulate from tempered ERGM distributions defined by the discretized path. The proposed method is referred to as power posterior exchange algorithm (PPEA). After obtaining a collection of MCMC simulations in the PPEA-step, a method proposed by Friel (2013) is applied which yields an estimate of the normalizing constant of the ERGM likelihood. In order to do so, a sequence of importance distributions defined by the power posterior temperature path is constructed. This estimate finally allows for the explicit evaluation of the likelihood (EEL) and is used in the trapezoidal approximation to estimate the ERGM evidence. The whole approach is referred to as PPEA-EEL.

The PPEA-EEL is applied for Bayesian ERGM selection and the results are compared to Bayesian logit and probit models that do not require an ERGM specification. The results are consistent and it can be concluded that the PPEA-EEL yields valid estimates of the ERGM evidence. New graphical methods are proposed which help to inspect the behavior of the PPEA transiting over the temperature path. Further, methods are proposed indicating the reliability of the PPEA-EEL results. These methods suggest the new insight that the importance sample size used for the EEL-step should be adapted to the step size of the discretized temperature path. This adaption yields a smooth path of tempered likelihood normalizing constants and improves the estimate of the ERGM evidence. If specified correctly, the PPEA-EEL yields a valid estimate of the ERGM evidence which does not rely

on strong approximation assumptions and is not limited by non-parametric kernel density estimates.

6.1 Limitations and alternatives

The PPEA-EEL is limited only by the computational resources available. However, the approach is computationally extremely expensive and in this work the analysis is restricted to directed networks with less than 50 nodes and ERGM specifications with no more than four parameters. The methods of ERGM estimation discussed in this work e.g. MCMC-ML, the EA and the PPEA, all rely on the simulation of auxiliary network data to circumvent the evaluation the analytically intractable ERGM normalizing constant. The PPEA-EEL uses a sequence of importance samples of simulated networks in order to estimate the ERGM normalizing constant. All these simulations are generated with the TNT sampler introduced by Morris et al. (2008) which is the computational bottleneck of the methods proposed in this work. If faster simulation of network data was available, the computation time of the PPEA-EEL could be reduced without increasing the discretizational error of estimating the ERGM evidence.

As the computational effort required for the PPEA-EEL is high, it can be more practical to use simpler approaches where applicable. Friel (2013) and Caimo and Friel (2013) estimate the evidence of the ERGM using an identity introduced by Chib (1995) and non-parametric density estimation of the posterior. Thiemichen et al. (2016) use a Laplace approximation which is very fast and easy to implement if the approximation assumptions hold. We recommend the following best practice for the estimation of the ERGM evidence. Use the EA-ADS to yield a MCMC sample from the ERGM posterior. If the sample suggests the posterior to be elliptical, use a Laplace approximation to estimate the ERGM evidence, see Thiemichen et al. (2016). If the posterior cannot be approximated by a normal distribution, use the identity by Chib (1995) and a non-parametric density estimate of the posterior, see Friel (2013). If the non-parametric density estimation should fail, use the PPEA-EEL and start with $J = 25$ temperature steps. Construct the temperature path like Friel and Pettitt (2008). Apply the methods of graphical inspection proposed in section 5.4.1. If the paths of the relevant quantities look noisy while transiting over the temperature space, increase the number of temperature steps. Doubling

J roughly corresponds to doubling the computational effort required.

There are alternatives to the ERGM evidence available if the computational costs of estimating this quantity are too high. The likelihood based information criteria like the AIC and the BIC can be used for model selection. To be applicable for the ERGM class, they require an estimate of the normalizing constant of the likelihood. The simplified path sampling approach introduced by Hunter and Handcock (2006) is an option as it is much faster than the approach by Friel (2013) implemented in the PPEA-EEL. However, little is known about the accuracy of the simplified path sampling approach compared to the EEL approach used in this work. The posterior predictive plots discussed in section 2.5.3 are much easier to implement than estimation of the ERGM evidence. Such plots can be a good guideline to evaluate the compatibility of the model with the observed data. However, in this work it becomes apparent that the sensitivity of these plots is inferior to the model evidence. Cross validation approaches can be used for model selection, too. A model is trained on a random subset and the rest of the data is predicted in order to evaluate the goodness-of-fit of the model. This can be repeated by defining mutually exclusive training subsets. We consider this a dangerous approach for ERGM selection as the observations within a network are not independent. A defined holdout sub sample cannot be expected to show the same patterns of tie variable formation as the rest of the network. The removal of a single central node from the training data could dramatically alter the network structure e.g. resulting in a totally different degree distribution or could disconnect regions of the graph. The ERGM models the global topological structure of a network which might be changed substantially by excluding subsets of nodes.

6.2 Outlook

In this thesis, the range of estimable ERGM specifications is limited due to the high computational costs of the PPEA-EEL. With more computational resources available in the future, more complex models for larger networks will be comparable using the ERGM evidence. The PPEA-EEL consists of several steps which all have potential to improve the computational efficiency.

Throughout this work a high number of $J = 100$ temperature steps is used for the PPEA-EEL which keeps the discretizational error of the trapezoidal approxi-

mation low. If the number of temperature steps used for power posterior sampling was lowered, the computational costs would be reduced. Calderhead and Girolami (2009) conclude that the optimal placement of the steps is more important than the number of steps. Little is known about the optimal placement of those steps for the ERGM class. In this work, a non-adaptive approach proposed by Friel and Pettitt (2008) is implemented concentrating the largest number of steps at very low temperatures. Adaptive approaches for optimizing the placement of temperature steps are available, see among other Lefebvre et al. (2009) and Hug et al. (2016). More research needs to be done on this issue for the ERGM class.

The EA-ADS used for simulation from the tempered posteriors yields fast convergence but requires many parallel ADS chains which makes it slow compared to a standard MH sampler. Recent developments lead to more efficient versions of the EA which can reduce the computing time and could be implemented to the PPEA-EEL. Caimo and Friel (2014) propose the approximate exchange algorithm with improved direction sampling which can drastically reduce the sampling variance compared to the EA-ADS applied and in this work. Friel et al. (2016) explore how multi-core computation can be used to reduce the sampling variance of MCMC techniques dealing with intractable likelihoods. Everitt et al. (2016) discuss sequential Monte Carlot methods and modified importance sampling approaches for models with intractable likelihood and computation of the marginal likelihood. Alquier et al. (2016) discuss population MCMC methods similar to the PPEA for the same problems.

The TNT sampler is the bottleneck of the PPEA-EEL as it is used in all computational expensive steps of the approach. In this work, established routines for network simulation are used. These routines are fast enough for most applications like MCMC-ML ERGM estimation and the EA-ADS, but result in a long computation time for the PPEA-EEL. Caimo and Friel (2013) use a implementation of the TNT sampler written in `C++` which could speed up the PPEA-EEL compared to an implementation in `R`.

The marginal likelihood allows for the comparison of non-nested ERGM specifications. Furthermore, it allows for the comparison of different model classes fit to the same data. This would allow for the comparison of the ERGM and the latent factor model introduced by Hoff (2005) and its extension by Hoff (2009). In a simplified version, the latent factor model is identical to a probit model assuming independence of tie variables. In section 5.4.1 an ERGM assuming tie variable

independence and an equivalent probit specification yield the same estimate of the marginal likelihood. However, the latent factor model uses a different approach to capture patterns of transitivity and reciprocity. It is an open research question to compare these two model classes.

Appendix A

Exponential random graph models

A.1 The Bernoulli random graph model is an ERGM

It shall be proved that the Bernoulli random graph model discussed in section 2.4.1 is an ERGM. y is an observed directed network on n nodes with

$$N = n(n-1)$$

possible edges. Y is the corresponding random graph Y consisting of N random tie variables. Consider a simple Bernoulli ERGM with the sum of observed edges

$$L(y) = \sum_{i>j} y_{ij}$$

as the only sufficient statistic. All N random tie variables are assumed to be independent. They follow a Bernoulli distribution with the constant tie probability

$$\Pr(Y_{ij} = 1) = p.$$

The probability distribution of the random graph Y is

$$\begin{aligned}
\Pr(Y = y|p) &= \prod_{i>j} p^{y_{ij}} \cdot (1-p)^{1-y_{ij}} \\
&= p^{L(y)} \cdot (1-p)^{N-L(y)} \\
&= \exp\{L(y)\ln(p) + (N - L(y))\ln(1-p)\} \\
&= \exp\{L(y)\left[\ln(p) - \ln(1-p)\right] + N\ln(1-p)\} \\
&= \frac{\exp\left\{L(y)\ln(\frac{p}{1-p})\right\}}{\exp\{-N\ln(1-p)\}}.
\end{aligned}$$

An ERGM using only $L(y)$ as sufficient statistic takes the form

$$\Pr(Y = y|\theta) = \frac{\exp\{L(y)\theta\}}{z(\theta)}.$$

Using the the relation

$$\theta = \ln\left(\frac{p}{1-p}\right)$$

it is clear that

$$\exp\left\{L(y)\ \ln(\frac{p}{1-p})\right\} = \exp\{L(y)\theta\}$$

is the non-normalized kernel of the Bernoulli ERGM.
Further it needs to be shown that the normalizing constant of the respective ERGM

$$z(\theta) = \exp\{-N\ln(1-p)\}.$$

$z(\theta)$ requires summation over all elements $\tilde{y} \in \mathcal{Y}$ in the space of possible graphs $\mathcal{Y}$ on n nodes. Consider $L(\tilde{y})$ as a possible number of realized edges in a graph,

$L(\tilde{y}) = (0, \ldots, N)$:

$$\begin{aligned} z(\theta) &= \sum_{\tilde{y} \in \mathcal{Y}} \exp\{L(\tilde{y})\theta\} \\ &= \sum_{L(\tilde{y})=0}^{N} \binom{N}{L(\tilde{y})} \exp\{L(\tilde{y})\ \theta\} \\ &= \sum_{L(\tilde{y})=0}^{N} \binom{N}{L(\tilde{y})} \exp\{\theta\}^{L(\tilde{y})} \end{aligned}$$

Using the binomial theorem yields

$$\begin{aligned} \sum_{L(\tilde{y})=0}^{N} \binom{N}{L(\tilde{y})} \exp\{\theta\}^{L(\tilde{y})} &= (1 + \exp\{\theta\})^N \\ &= \exp\{N \ln(1 + \exp\{\theta\})\} \\ &= \exp\{-N \ln(1-p\}\,. \end{aligned}$$

Note that

$$\begin{aligned} \ln(1 + \exp\{\theta\}) &= \ln\left(1 + \exp\left\{\ln \frac{p}{1-p}\right\}\right) \\ &= \ln\left(1 + \frac{p}{1-p}\right) \\ &= \ln\left(\frac{1-p+p}{1-p}\right) \\ &= \ln\left(\frac{1}{1-p}\right) \\ &= -\ln(1-p). \end{aligned}$$

□

A.2 Tables

Table A.1: Simulated toy networks: Summary of sufficient network statistics, specification 1

	edges	mutual	GWESP	samedept	samelev	levdiff
min	5.00	0.00	0.00	1.00	4.00	0.00
1st Qu.	20.00	5.00	15.29	7.00	16.00	3.00
median	23.00	7.00	18.98	8.00	18.00	5.00
mean	22.64	6.67	18.87	8.55	17.94	4.79
3rd Qu.	25.00	8.00	22.60	10.00	20.00	6.00
max	39.00	15.00	40.50	18.00	30.00	17.00

Table A.2: Simulated toy networks: Summary of sufficient network statistics, specification 2

	edges	mutual	samedept	samelev
min	3.00	0.00	1.00	0.00
1st Qu.	12.00	2.00	7.00	6.00
median	14.00	3.00	8.00	8.00
mean	14.44	3.16	8.09	8.10
3rd Qu.	17.00	4.00	10.00	10.00
max	30.00	9.00	15.00	17.00

A.3 Figures

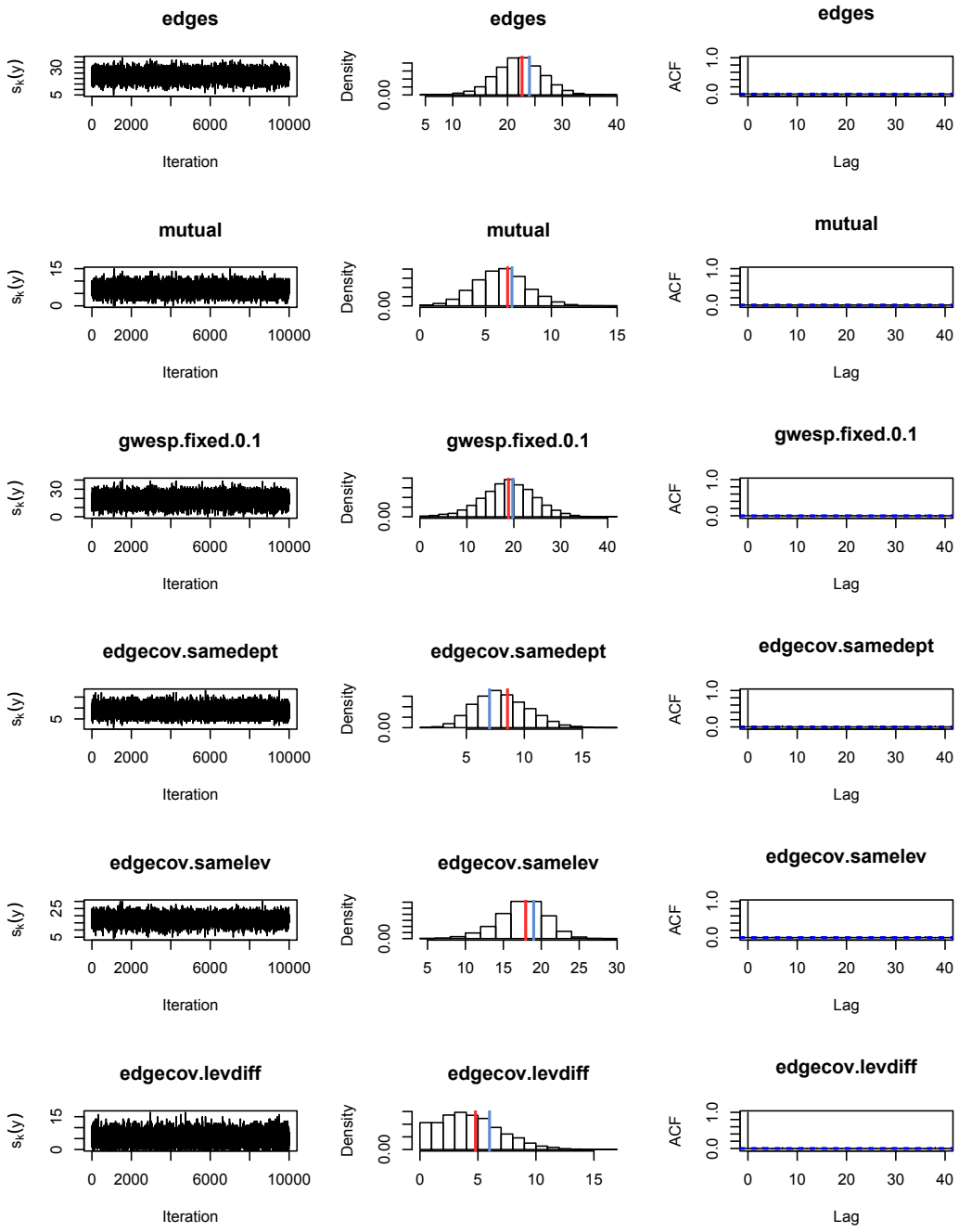

Figure A.1: Simulated toy networks: Summary of sufficient statistics, m_1: Traceplots, histograms and ACF.

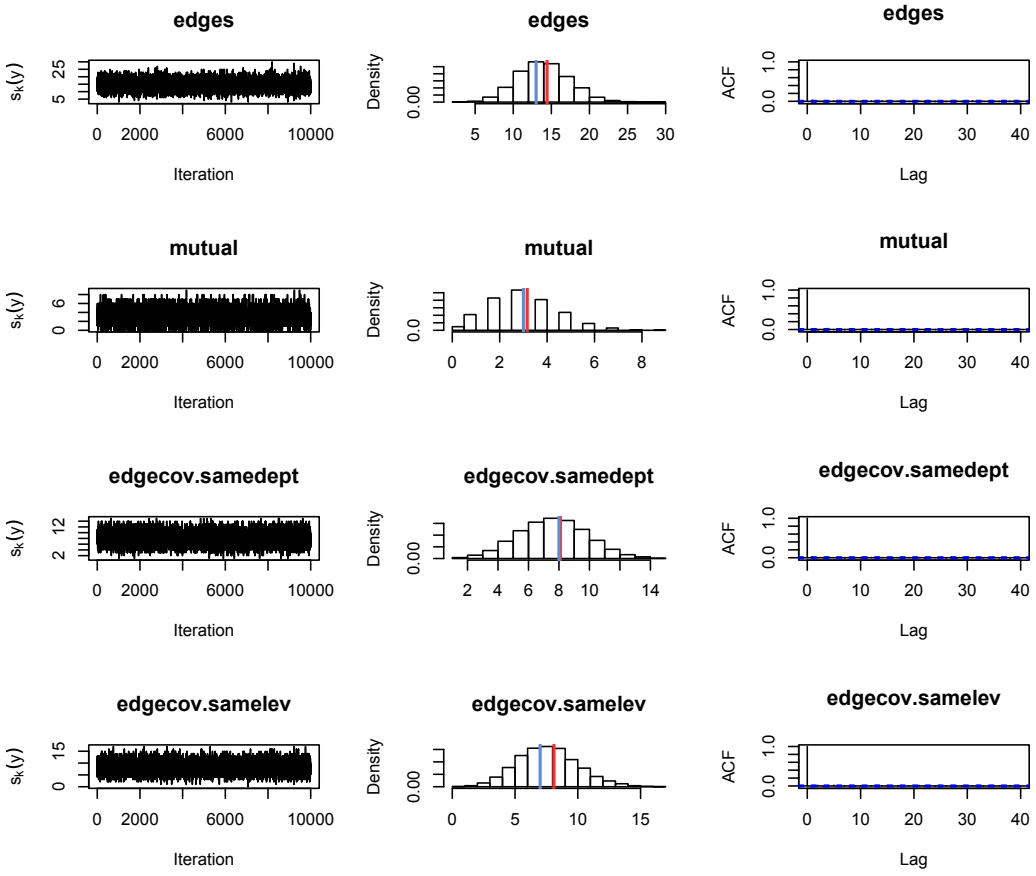

Figure A.2: Simulated toy networks: Summary of sufficient statistics, m_2: Traceplots, histograms and ACF.

Appendix B

Determinants of communication in policy networks

B.1 Survey questions

Expert information:

"Stakeholder organizations, research institutes or political actors can frequently provide expert information to other organizations, especially when consequences of complex policies have to be evaluated. Such kind of expert information comprises the knowledge of the effects of different policy instruments on the welfare of different social groups. Therefore expert information is very interesting for political organizations as well as for other interest groups when designing and influencing agricultural policy programmes."

Sending information:
"Using the list of organizations again, please check all organizations to which your organization provides expert information on agricultural policies."

Receiving information:
"Using the list of organizations again, please check all organizations from which your organization receives expert information on agricultural policies."

Political support:

"In democracies stakeholder organizations are representatives of their members and their interests. Therefore the policy position of such a group is highly connected with the resulting welfare for their members. Thus, a major role of stakeholder organizations in democracies is intermediating their clientele's interest to politicians, i.e. trying to influence policy or politicians to generate as much welfare as possible for their members. Obviously, politicians won't support a stakeholder organization's position without any reward. On their part they expect in return the political support of members of the stakeholder organization. However, political agents also represent their electorate in parliament. Therefore, political agents are interested to find political solutions supported by a majority of their electorate."

Governmental actor:

"Please check those organizations which are important for you regarding the intermediation of political positions supported by voters."

Non-governmental actor:

"Taking now the above described kind of support relation between organizations and political agents into account, please check those political institutions on the list with which your organization has such a relationship."

B.2 Tables

Table B.1: Classification of actors with absolute and relative frequency

Category	Group	Ghana	Senegal	Uganda
State actors	Executive (EXE)	6 *(0.130)*	7 *(0.152)*	7 *(0.163)*
	Public Sector Agencies (PUB)	5 *(0.109)*	5 *(0.109)*	6 *(0.140)*
	Legislative (LEG)	2 *(0.044)*	1 *(0.022)*	2 *(0.047)*
Int. organizations	Donor (DON)	7 *(0.152)*	7 *(0.152)*	6 *(0.140)*
	International NGOs (INGO)	5 *(0.109)*	4 *(0.087)*	3 *(0.070)*
Research organizations	(RES)	7 *(0.152)*	10 *(0.217)*	5 *(0.116)*
Interest groups	(IG)	14 *(0.304)*	12 *(0.261)*	14 *(0.326)*
n		**46**	**46**	**43**
Note:		absolute frequency *(relative frequency)*		

Table B.2: Model terms and affiliated hypotheses

	Ghana		Senegal		Uganda		Hypothesis
	Expert	Support	Expert	Support	Expert	Support	
edges							
network density[a]	0.226	0.215	0.281	0.159	0.254	0.226	
political similarity							1a
mean simil.[e]	0.433		0.348		0.613		
	(0.116)		*(0.111)*		*(0.142)*		
preference similarity							1a,1b
mean simil.[e]	0.236		0.250		0.259		
	(0.115)		*(0.111)*		*(0.141)*		
membership							2a
network density[a]	0.082		0.162		0.059		
mean memberhsip[e]	0.93		2.54		1.05		
	(0.68)		*(2.54)*		*(1.45)*		
mutual							2b
reciprocity[b]	0.387	0.360	0.359	0.317	0.340	0.314	
GWESP & GWDSP							2c
transitivity[d]	0.439	0.3847	0.506	0.380	0.480	0.458	
expert & support							2d
mean degree[c]	20.35	19.30	25.30	14.26	21.35	18.98	
	(11.79)	*(12.02)*	*(13.40)*	*(9.95)*	*(11.74)*	*(12.24)*	
nodal reputation							3a
mean reputation[e]	0.339		0.690		0.4670		
	(0.116)		*(0.144)*		*(0.129)*		
popularity EXE							3b
mean i.degree[g]	17.33	16.33	13	7.711	13.57	11	
	(10.33)	*(6.74)*	*(7.42)*	*(5.77)*	*(8.83)*	*(11.82)*	
activity DON							4
mean o.degree[f]	13.29	17.71	11.43	9.14	12.50	7.33	
	(4.39)	*(3.68)*	*(6.27)*	*(6.62)*	*(5.36)*	*(5.96)*	
activity RES							5
mean o.degree[f]	8.14	8.86	13.60	6.60	2.80	6.60	
	(4.10)	*(7.01)*	*(7.78)*	*(5.89)*	*(1.64)*	*(6.35)*	
IG homophily							6
share IG:IG [h]	0.253	0.203	0.175	0.175	0.271	0.328	

Note:
[a] share of directed ties among all possible $n^2 - n$ ties
[b] share of reciprocal ties
[c] mean degree *(standard deviation)*
[d] clustering coefficient, see Wasserman and Faust (1994)
[e] mean value *(standard deviation)*
[f] mean out degree *(standard deviation)*
[g] mean in degree *(standard deviation)*
[h] share of homophilic ties among all IG ties

Table B.3: ERGM parameter estimates all countries

	Ghana				**Senegal**				**Uganda**			
	Expert network		Support network		Expert network		Support network		Expert network		Support network	
	m_1	m_2	$m_3{}^+$	m_4	m_1	m_2	m_3	m_4	m_1	m_2	m_3	m_4
edges	-4.539^{***}	-16.496^{***}	-2.683^{***}	-15.306^{***}	-3.375^{***}	-3.402^{***}	-3.208^{***}	-16.517^{***}	-3.368^{***}	-4.586^{***}	-2.801^{***}	-3.195^{**}
	(0.452)	(0.224)	(0.174)	(0.153)	(0.494)	(0.984)	(0.247)	(0.673)	(0.365)	(0.848)	(0.238)	(1.350)
mutual	3.669^{***}	3.420^{***}	3.291^{***}	3.094^{***}	2.888^{***}	2.757^{***}	2.870^{***}	2.665^{***}	2.702^{***}	2.554^{***}	2.451^{***}	2.260^{***}
	(0.210)	(0.216)	(0.200)	(0.202)	(0.175)	(0.178)	(0.212)	(0.218)	(0.187)	(0.193)	(0.189)	(0.194)
support		0.955^{***}				1.388^{***}				1.231^{***}		
		(0.111)				(0.121)				(0.123)		
expert				1.003^{***}				1.336^{***}				1.215^{***}
				(0.117)				(0.120)				(0.120)
GWESP	1.714^{***}	1.330^{***}	0.492^{***}	0.105	1.283^{***}	1.281^{***}	0.886^{***}	0.649^{***}	1.277^{***}	0.787^{***}	0.888^{***}	0.578^{***}
	(0.366)	(0.325)	(0.133)	(0.143)	(0.411)	(0.425)	(0.147)	(0.146)	(0.297)	(0.297)	(0.188)	(0.195)
GWDSP	-0.078^{***}	-0.028	-0.142^{***}	-0.122^{***}	-0.183^{***}	-0.138^{***}	-0.102^{***}	-0.067^{**}	-0.154^{***}	-0.024	-0.155^{***}	-0.119^{***}
	(0.027)	(0.022)	(0.028)	(0.012)	(0.019)	(0.021)	(0.029)	(0.030)	(0.018)	(0.030)	(0.015)	(0.026)
membership		1.065^{***}		0.039		0.087		-0.112		0.220		0.047
		(0.154)		(0.156)		(0.113)		(0.153)		(0.199)		(0.213)
nodal reput.		1.443^{**}		2.112^{***}		1.015^{***}		1.407^{***}		3.374^{***}		2.019^{***}
		(0.586)		(0.643)		(0.390)		(0.506)		(0.547)		(0.515)
popul. exec.		0.448^{***}		0.564^{***}		-0.188		-0.109		0.188		0.092
		(0.164)		(0.182)		(0.149)		(0.177)		(0.167)		(0.159)
activ. donor		-0.316^{*}		0.928^{***}		-0.409^{***}		0.225		0.057		-0.429^{**}
		(0.178)		(0.179)		(0.147)		(0.172)		(0.174)		(0.172)
activ. research		-0.206		0.134		0.083		-0.328^{**}		-1.474^{***}		-0.031
		(0.167)		(0.158)		(0.124)		(0.158)		(0.315)		(0.179)
pol. simil.		-0.083		-0.001		-0.001		-2.219^{*}		-0.001		-0.912
		(0.186)		(0.004)		(0.004)		(1.304)		(0.004)		(1.068)
pref. simil.		11.657^{***}		11.849^{***}		-0.862		14.268^{***}		-0.246		0.298
		(0.224)		(0.153)		(0.802)		(0.673)		(0.735)		(0.740)
IG homophily		0.092		-1.465		0.034		0.041		0.018		0.613^{***}
		(0.161)		(1.018)		(0.185)		(0.221)		(0.167)		(0.155)
log Likelihood	-843.211	-736.632	-858.21	-749.335	-1019.76	-924.536	-744.799	-655.487	-860.369	-743.863	-826.363	-731.055
Akaike Inf. Crit.	1694.422	1499.265	1724.420	1524.671	2047.520	1875.072	1497.598	1336.974	1728.739	1513.727	1660.726	1488.111
Bayes. Inf. Crit.	1716.964	1572.524	1746.961	1597.930	2070.061	1948.331	1520.139	1410.233	1750.734	1585.212	1682.721	1559.596

Note: MCMC mean value
(MCMC sd)
*p<0.1; **p<0.05; ***p<0.01
$^{+}$: $\alpha_E = 0.05, \alpha_D = 0$

B.3 Figures

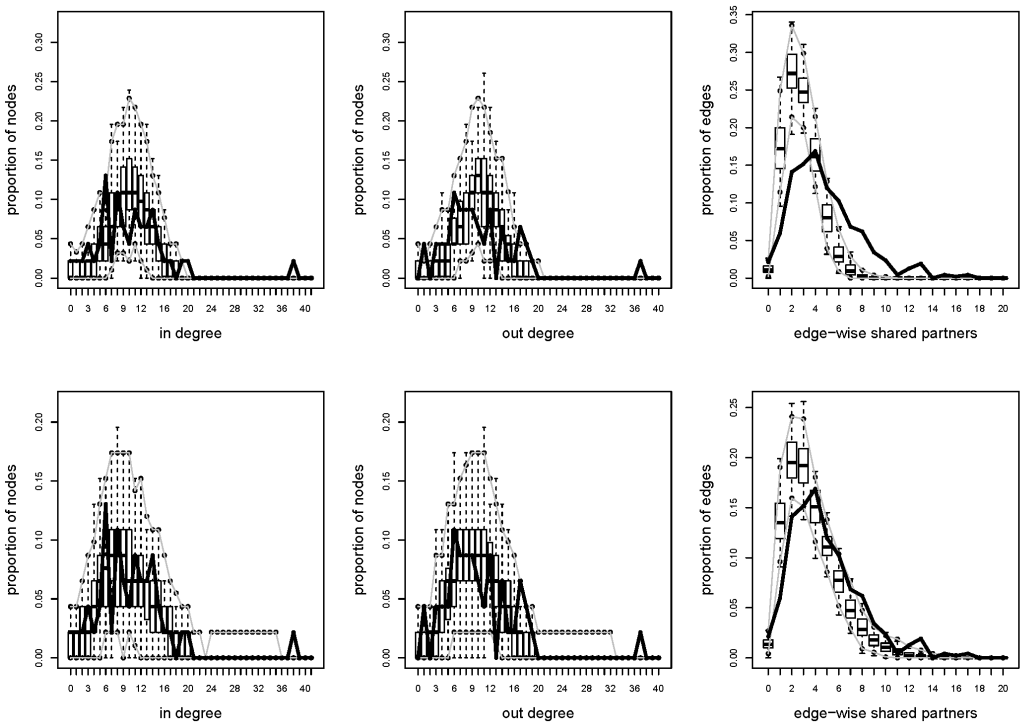

Figure B.1: GOF: Ghana Expert network m_1 (upper row) and m_2 (lower row)

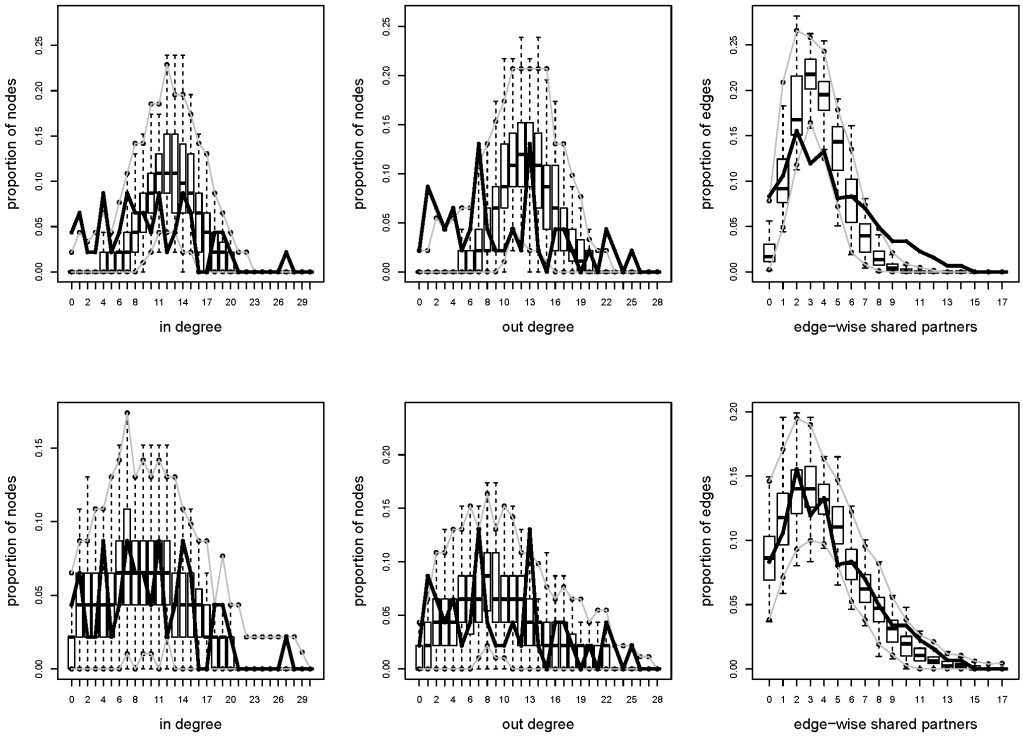

Figure B.2: GOF: Ghana Support network m_3(upper row) and m_4 (lower row)

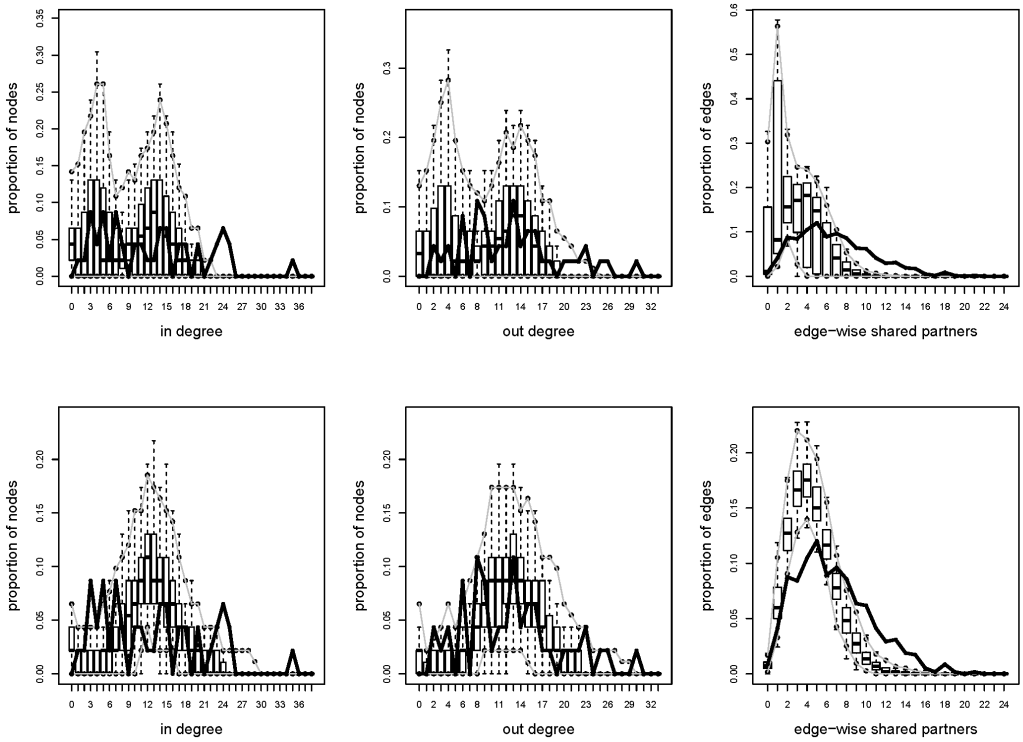

Figure B.3: GOF: Senegal Expert network m_1 (upper row) and m_2 (lower row)

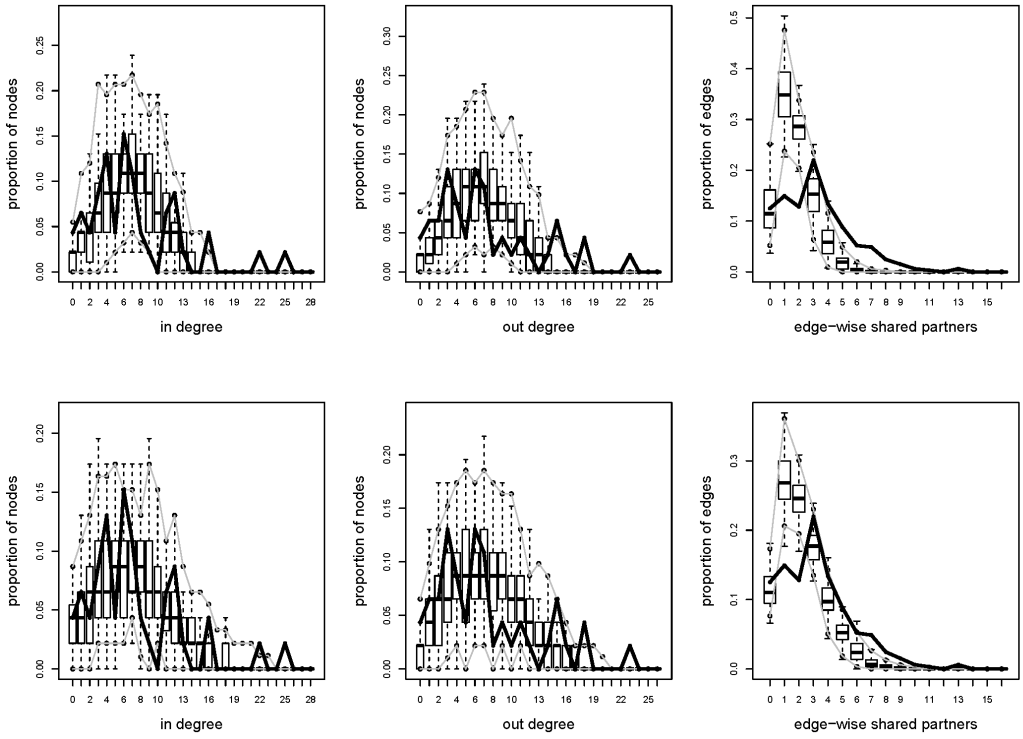

Figure B.4: GOF: Senegal Support network m_3 (upper row) and m_4 (lower row)

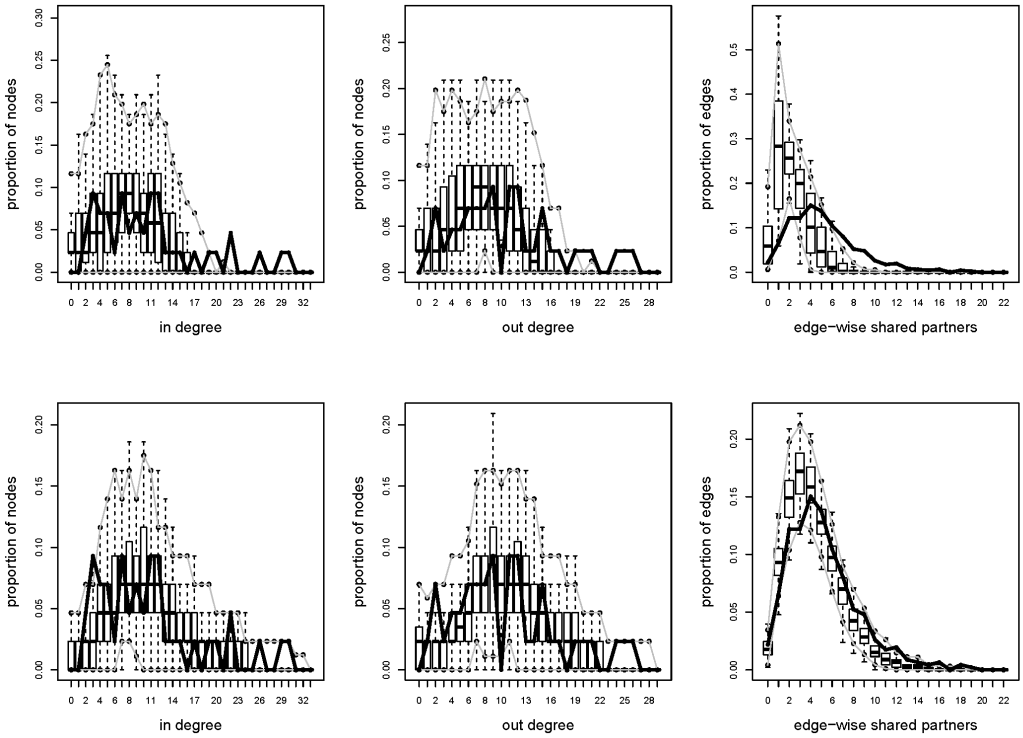

Figure B.5: GOF: Uganda Expert network m_1 (upper row) and m_2(lower row)

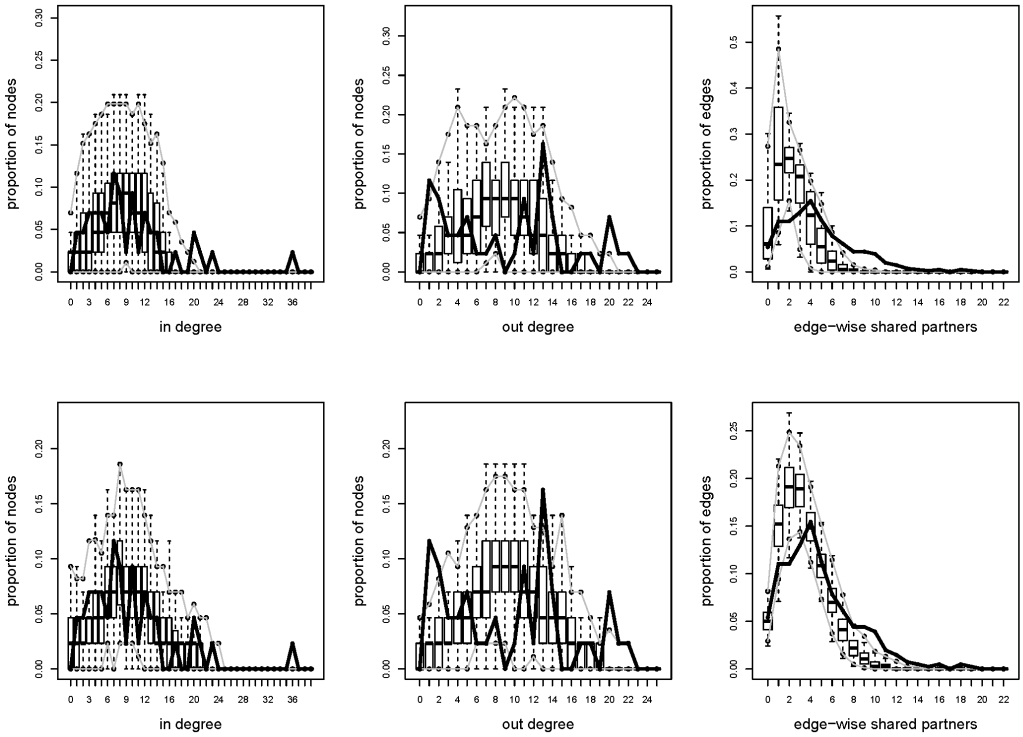

Figure B.6: GOF: Uganda Support network m_3 (upper row) and m_4 (lower row)

Appendix C

Bayesian exponential random graph model estimation

C.1 Krackhardt's managers

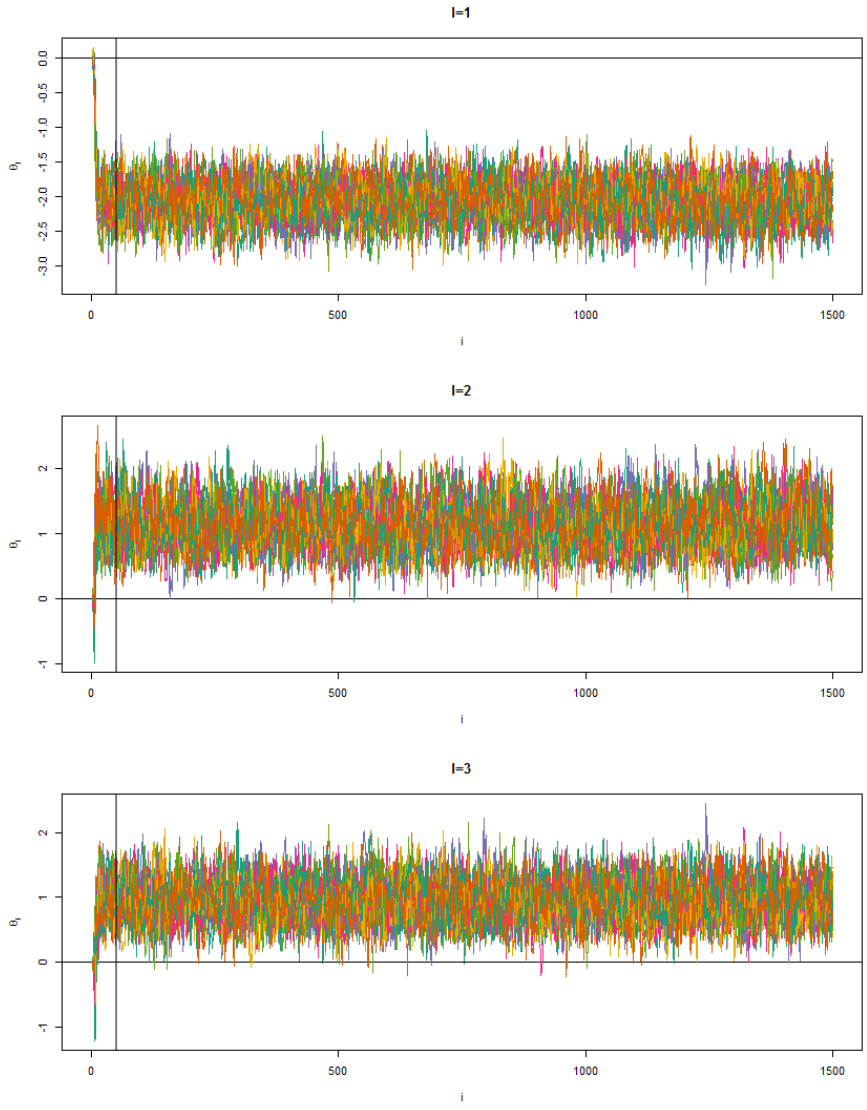

Figure C.1: Krackhardt's network, m_1:
Traceplots of the $V = 20$ parallel ADS chains of the parameters θ_l. The transparent gray line indicates the discarded $I_{burn} = 50$ samples used as burn-in period.

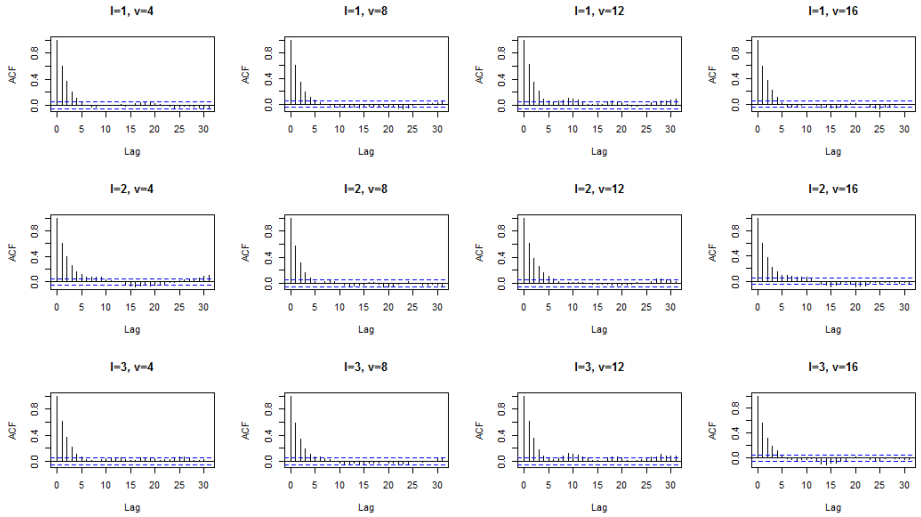

Figure C.2: Krackhardt's network, m_1:
ACF of some of the parallel ADS chains used in the EA. Four chains displayed for each parameter.

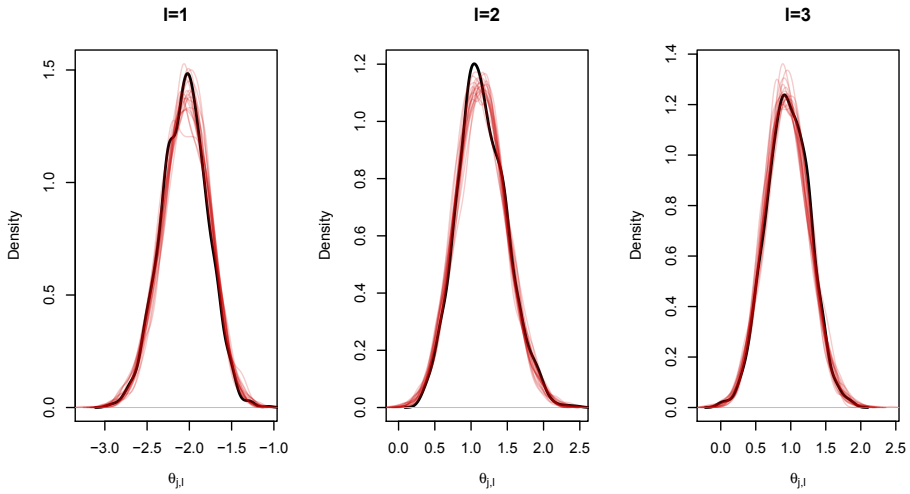

Figure C.3: Krackhardt's network, m_1:
Densities of merged EA sample: The thick black line represents the merged sample, the transparent red lines represent the $V = 20$ parallel chains used for ADS sampling.

C.2 Ghana

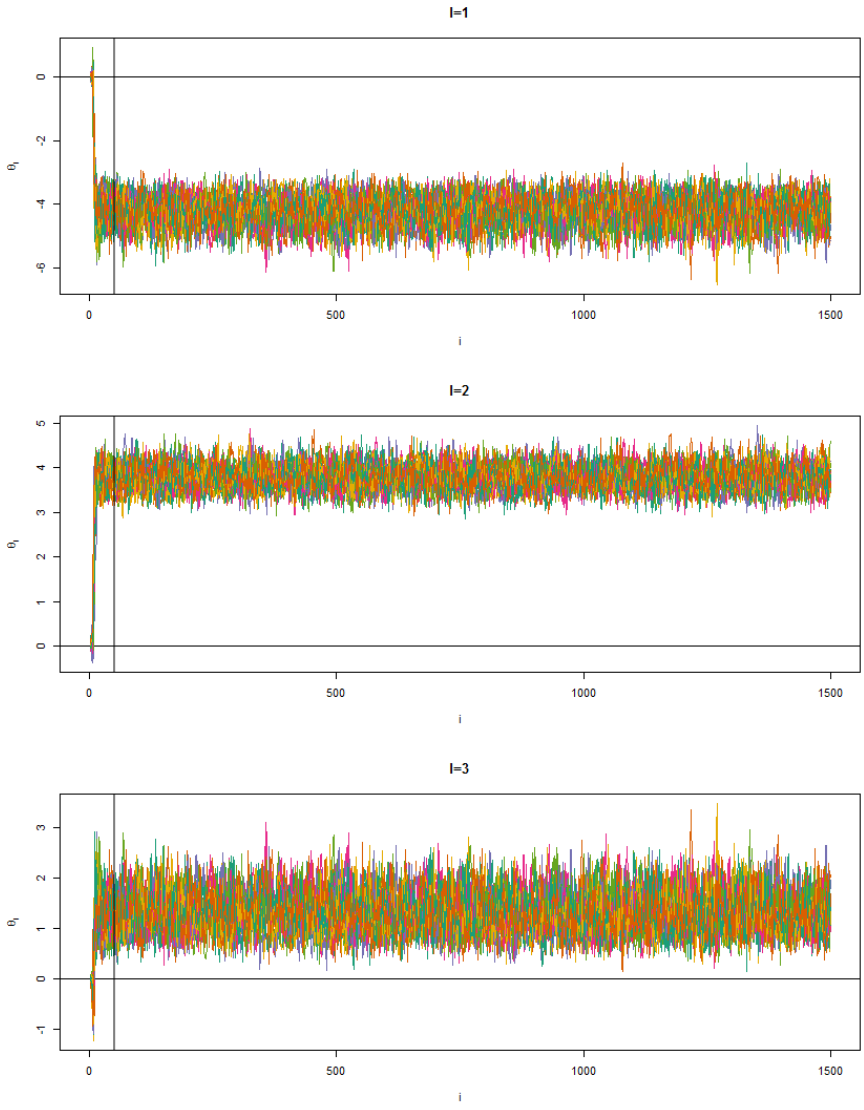

Figure C.4: Ghana expert network, m_1:
Traceplots of the $V = 20$ parallel ADS chains of the parameters θ_l. The transparent gray line indicates the discarded $I_{burn} = 50$ samples used as burn-in period.

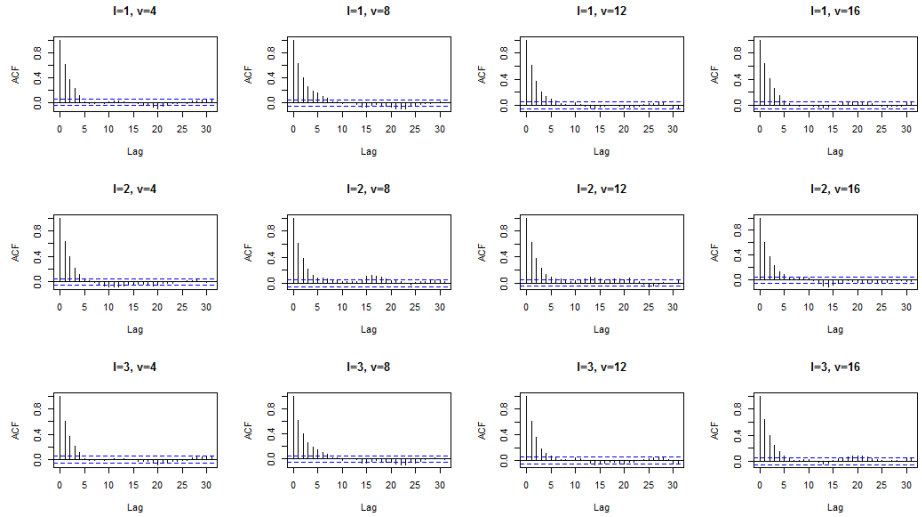

Figure C.5: Ghana expert network, m_1:
ACF of some of the parallel ADS chains used in the EA. Four chains displayed for each parameter.

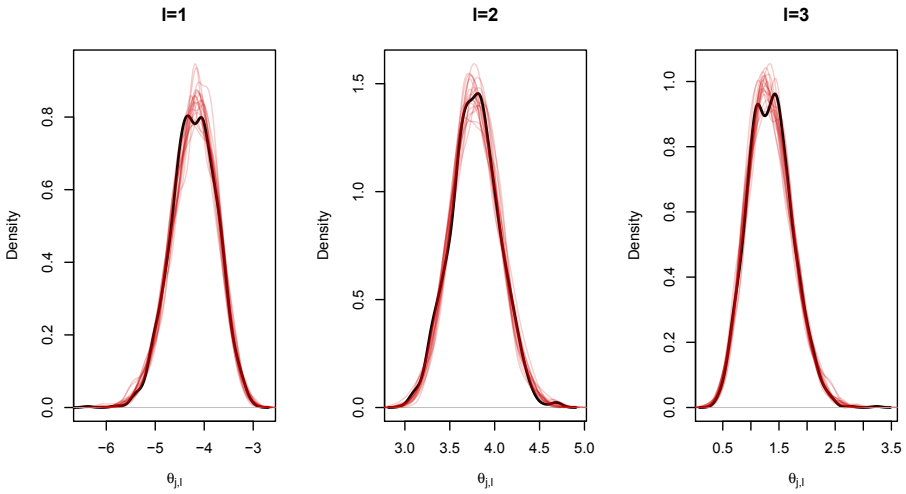

Figure C.6: Ghana expert network, m_1:
Densities of merged EA sample: The thick black line represents the merged sample, the transparent red lines represent the $V = 20$ parallel chains used for ADS sampling.

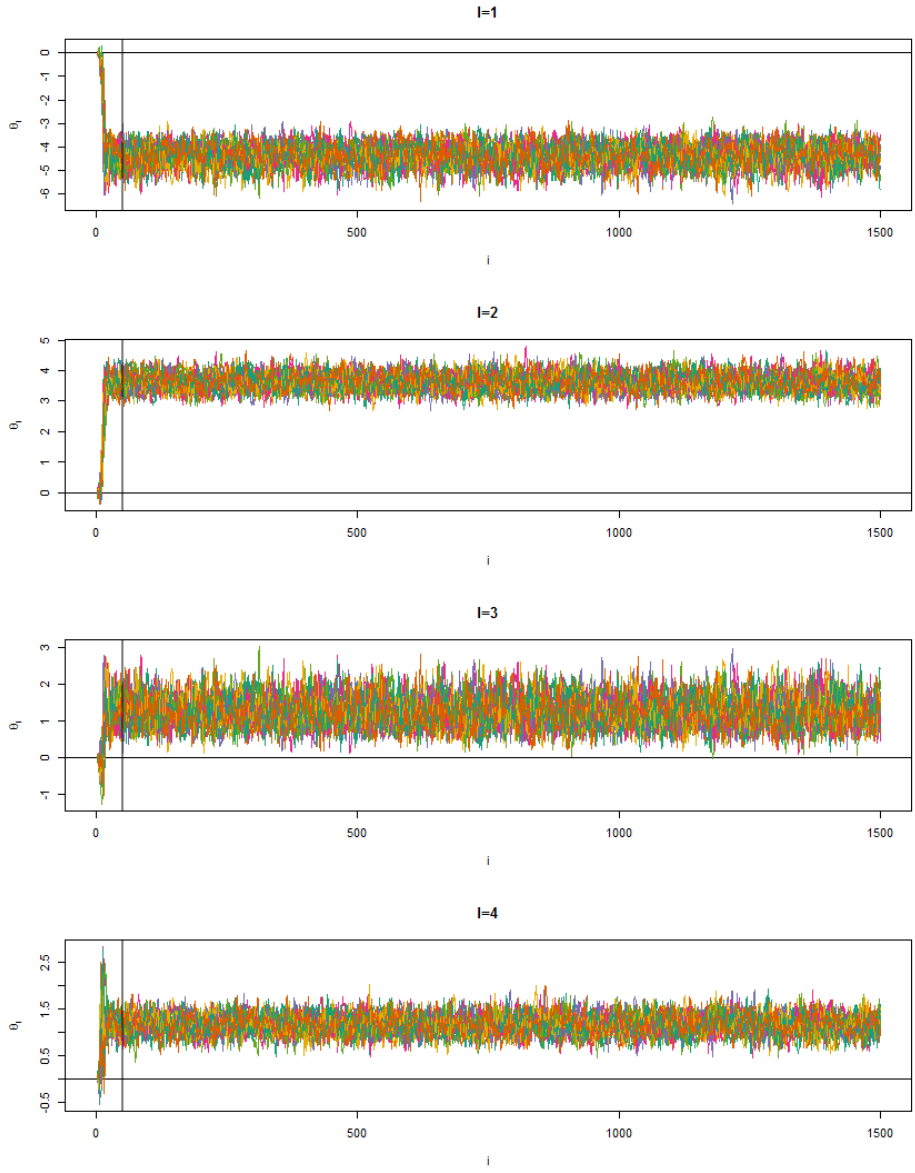

Figure C.7: Ghana expert network, m_2:
Traceplots of the $V = 20$ parallel ADS chains of the parameters θ_l. The transparent gray line indicates the discarded $I_{burn} = 50$ samples used as burn-in period.

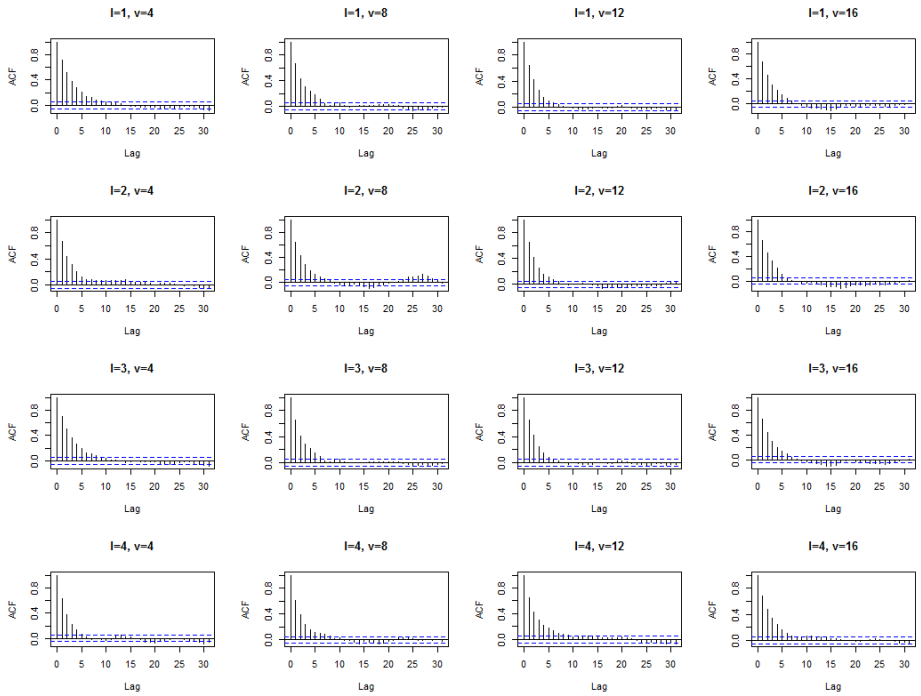

Figure C.8: Ghana expert network, m_2:
ACF of some of the parallel ADS chains used in the EA. Four chains displayed for each parameter.

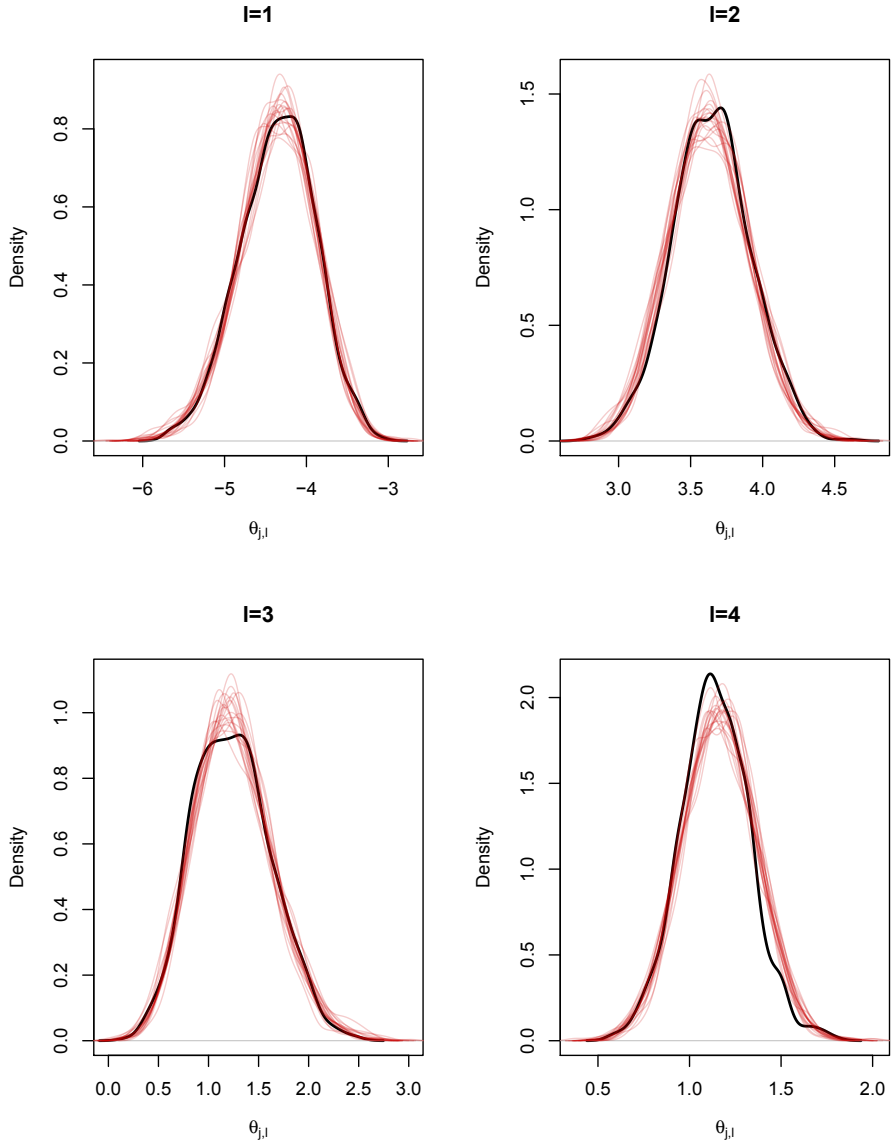

Figure C.9: Ghana expert network, m_2:
Densities of merged EA sample: The thick black line represents the merged sample, the transparent red lines represent the $V = 20$ parallel chains used for ADS sampling.

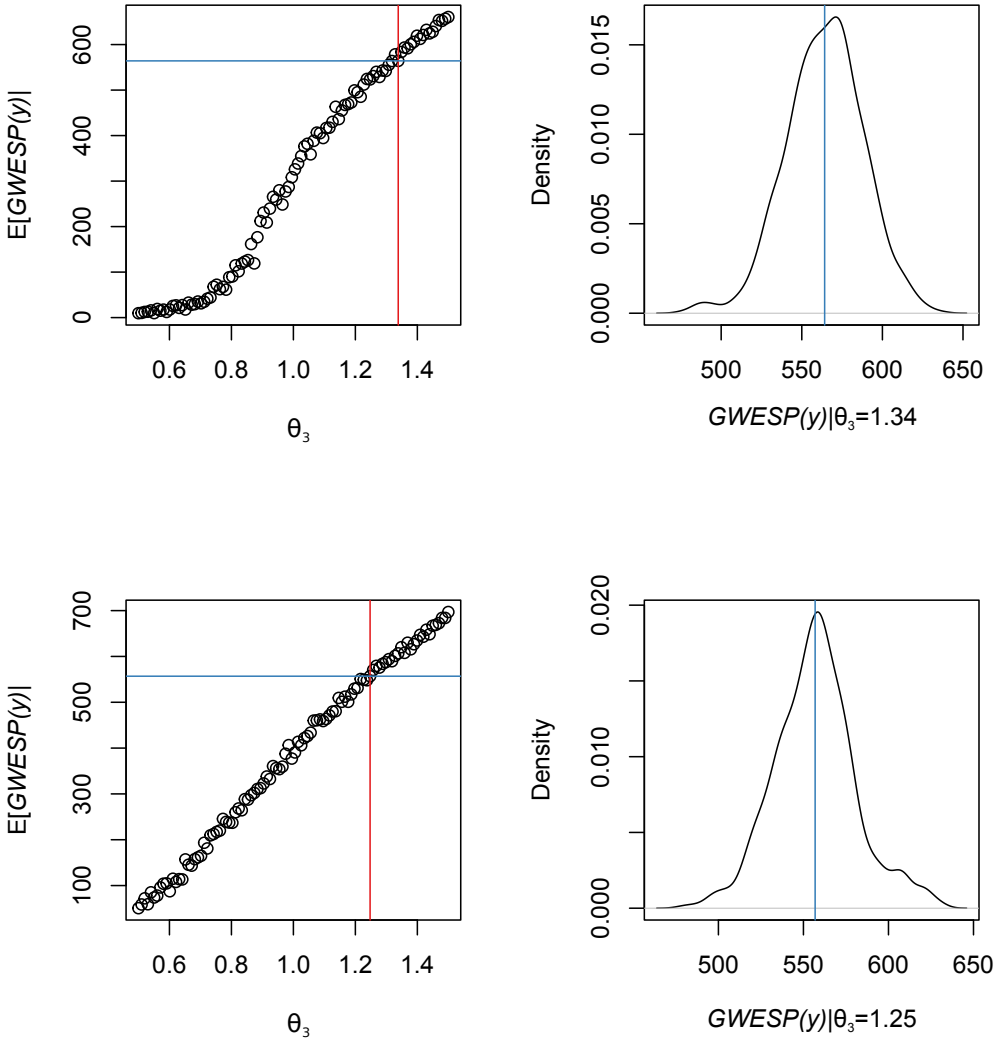

Figure C.10: Ghana: Check for degeneracy, m_1 and m_2:
The red line indicates the MCMC parameter estimate of θ_3, the blue line indicates $\hat{E}\left[GWESP(y)|\hat{\theta}_3\right]$.
Top panel: No degeneracy observable for m_1, but $\mathrm{E}\left[GWESP(y)|\theta_3\right]$ is rather steep around $\theta_3 = 0.9$. This might explain why it is not possible to estimate m_1 using MCMC-ML.
Lower panel: No degeneracy observable for m_2.

Appendix D

Bayesian model selection for network data

D.1 The path sampling identity

Gelman and Meng (1998) introduce the path sampling identity (5.19) where the log ratio of normalizing constants can be found by solving an integral over the unit interval. $t \in [0, 1]$ is a random variable defined on the unit interval with the prior distribution $t \sim p(t)$. Let $p(\theta|t)$ denote a normalized sampling density with the non-normalized kernel $q(\theta|t)$ and the normalizing constant

$$z_t = \int q(\theta|t)d\theta.$$

The potential

$$U(\theta, t) = \frac{d}{dt} \ln q(\theta|t)$$

is the derivative of the log non-normalized kernel with respect to t. The derivative

of the log normalizing constant can be expressed as

$$\begin{aligned}
\frac{d}{dt}\ln z_t &= \frac{1}{z_t}\frac{d}{dt}z_t \\
&= \frac{1}{z_t}\frac{d}{dt}\int q(\theta|t)d\theta \\
&= \frac{1}{z_t}\int \frac{d}{dt}q(\theta|t)d\theta \\
&= \int \frac{1}{q(\theta|t)}\frac{d}{dt}q(\theta|t)\frac{q(\theta|t)}{z_t}d\theta \\
&= \int \frac{d}{dt}\ln q(\theta|t)p(\theta|t)d\theta \\
&= \mathrm{E}_{\theta|t}\left[\frac{d}{dt}\ln q(\theta|t)\right] \\
&= \mathrm{E}_{\theta|t}[U(\theta,t)]
\end{aligned}$$

The log ratio of normalizing constants (5.19) can be found by integration over the unit interval:

$$\begin{aligned}
\lambda &= \ln z_1 - \ln z_0 \\
&= \int_0^1 \frac{d}{dt}\ln z_t dt \\
&= \int_0^1 \mathrm{E}_{\theta|t}[U(\theta,t)]dt
\end{aligned}$$

Note that the definite integral

$$\int_0^1 \frac{d}{dt}\ln z_t dt$$

of the differentiable function $\frac{d}{dt}\ln z_t$ over the interval $[0,1]$ is equal to the difference of antiderivatives

$$\ln z_1 - \ln z_0$$

where $\ln z_t$ is the antiderivative of $\frac{d}{dt}\ln z_t$.

D.2 Power posterior sampling

Friel and Pettitt (2008) apply the path sampling identity to power posterior sampling. The tempered posterior distribution is

$$p(\theta|y,t) = \frac{p(y|\theta)^t p(\theta)}{p(y|t)} \tag{D.1}$$

where $p(y|\theta)^t$ is the tempered likelihood, $p(\theta)$ is the prior and

$$p(y|t) = \int p(y|\theta)^t p(\theta) d\theta \tag{D.2}$$

is the normalizing constant of the tempered posterior. Similar to appendix D.1 the derivative of the log normalizing constant can be expressed as

$$\begin{aligned}
\frac{d}{dt}\ln p(y|t) &= \frac{1}{p(y|t)}\frac{d}{dt}p(y|t) \\
&= \frac{1}{p(y|t)}\frac{d}{dt}\int_\theta p(y|\theta)^t \quad p(\theta)d\theta \\
&= \frac{1}{p(y|t)}\int \frac{d}{dt}p(y|\theta)^t \quad p(\theta)d\theta \\
&= \frac{1}{p(y|t)}\int p(y|\theta)^t \ln p(y|\theta) \quad p(\theta)d\theta \\
&= \int_\theta \frac{p(y|\theta)^t p(\theta)}{p(y|t)} \ln p(y|\theta) \\
&= E_{\theta|y,t}\left[\ln p(y|\theta)\right]
\end{aligned} \tag{D.3}$$

Note that

$$\frac{d}{dx}a^x = \ln(a)a^x.$$

$\mathrm{E}_{\theta|y,t}[\ldots]$ is the expectation taken with respect to the tempered posterior $p(\theta|y,t)$. (D.3) is used in (5.30).

D.3 PPEA Krackhardt's managers

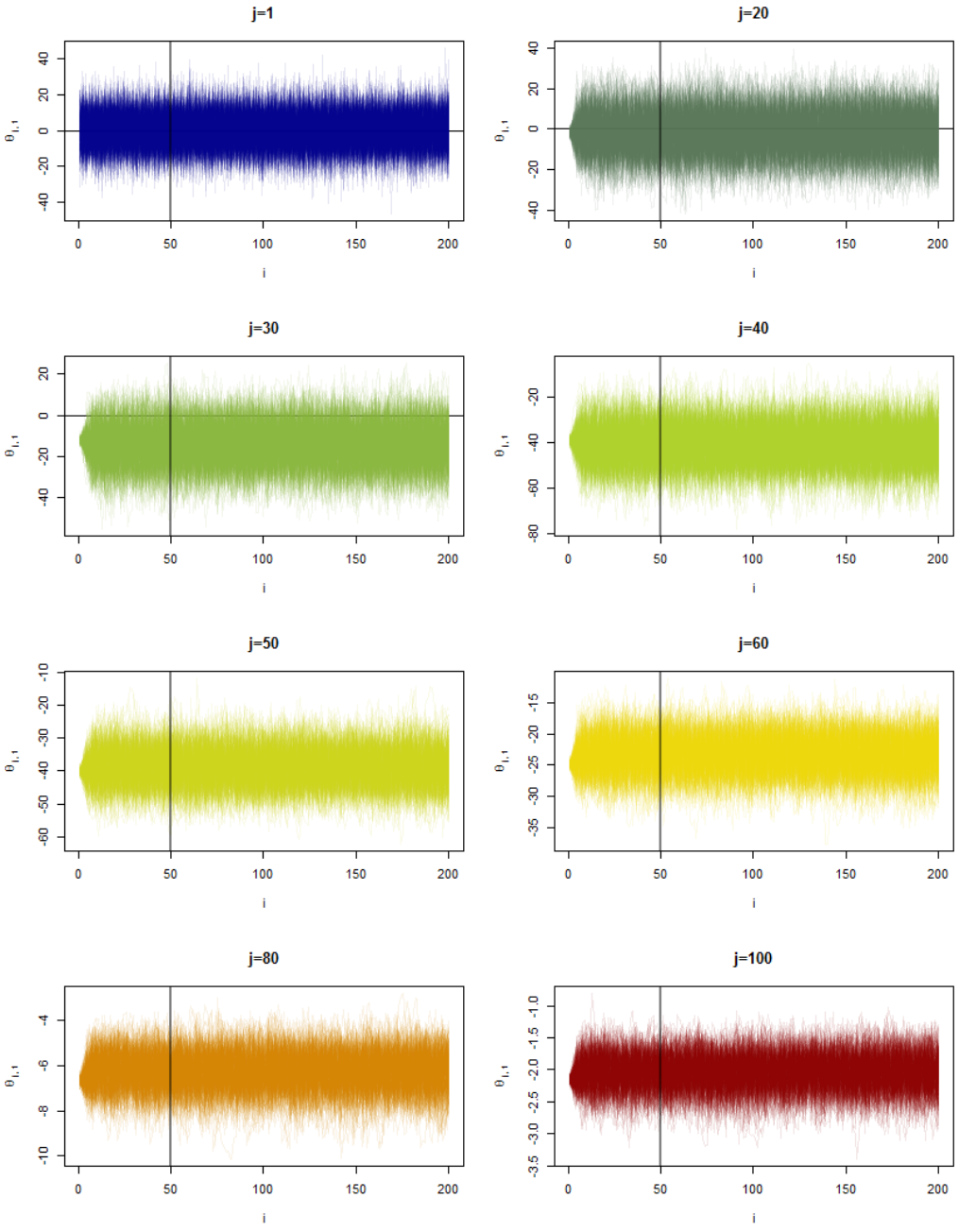

Figure D.1: PPEA Krackhardt's managers, m_1:
Traceplots of the $c \cdot V = 320$ chains of the edge parameter $\theta_{j,l=1}$ at some temperatures j of the PPEA. The horizontal grey line indicates the burn-in period of $I_{burn} = 50$ iterations used.

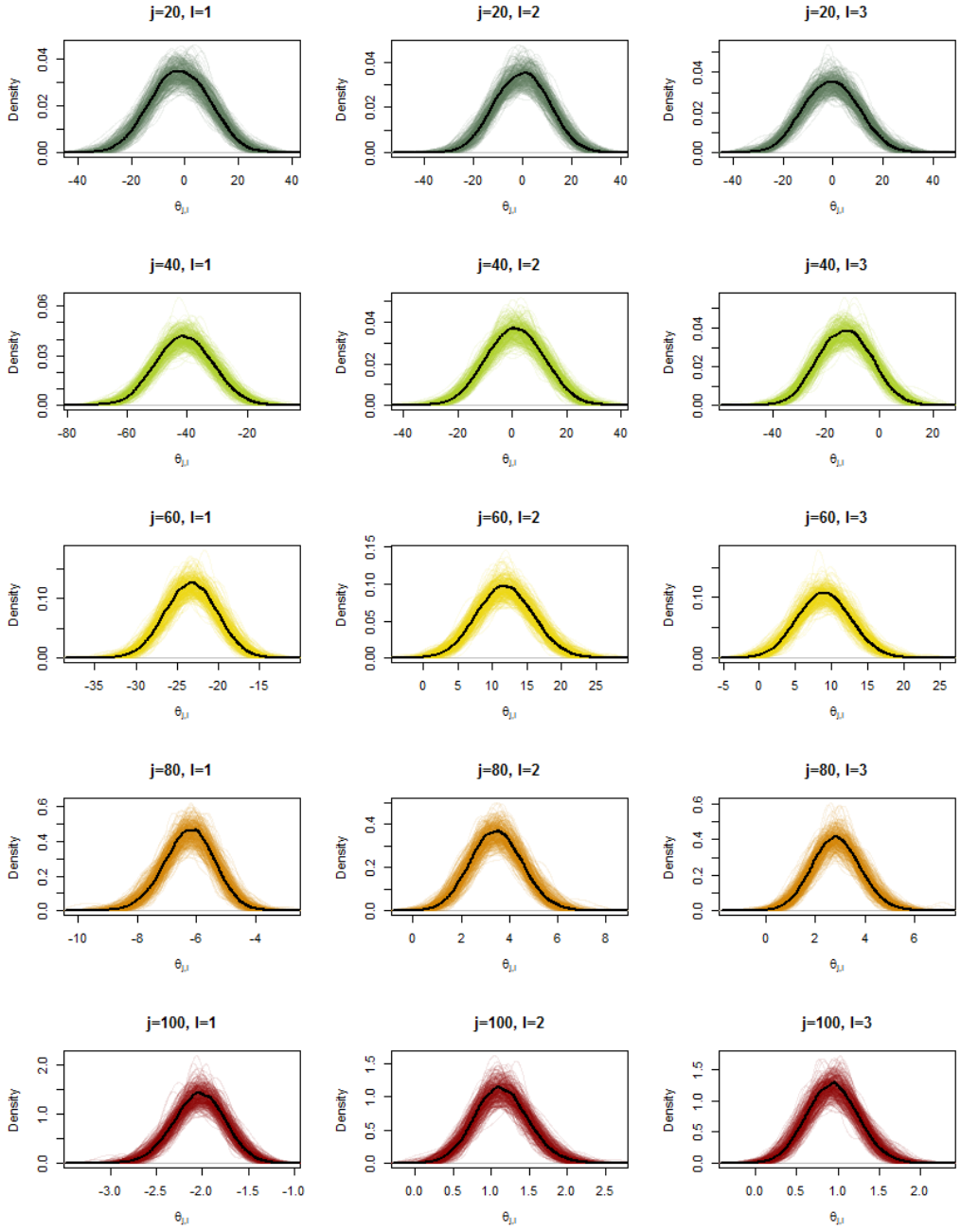

Figure D.2: PPEA Krackhardt's managers, m_1:
Density plots of the $c \cdot V = 320$ chains of the parameters $\theta_{j,l}$ at some temperatures j of the PPEA. The black line indicates the density of the merged sample.

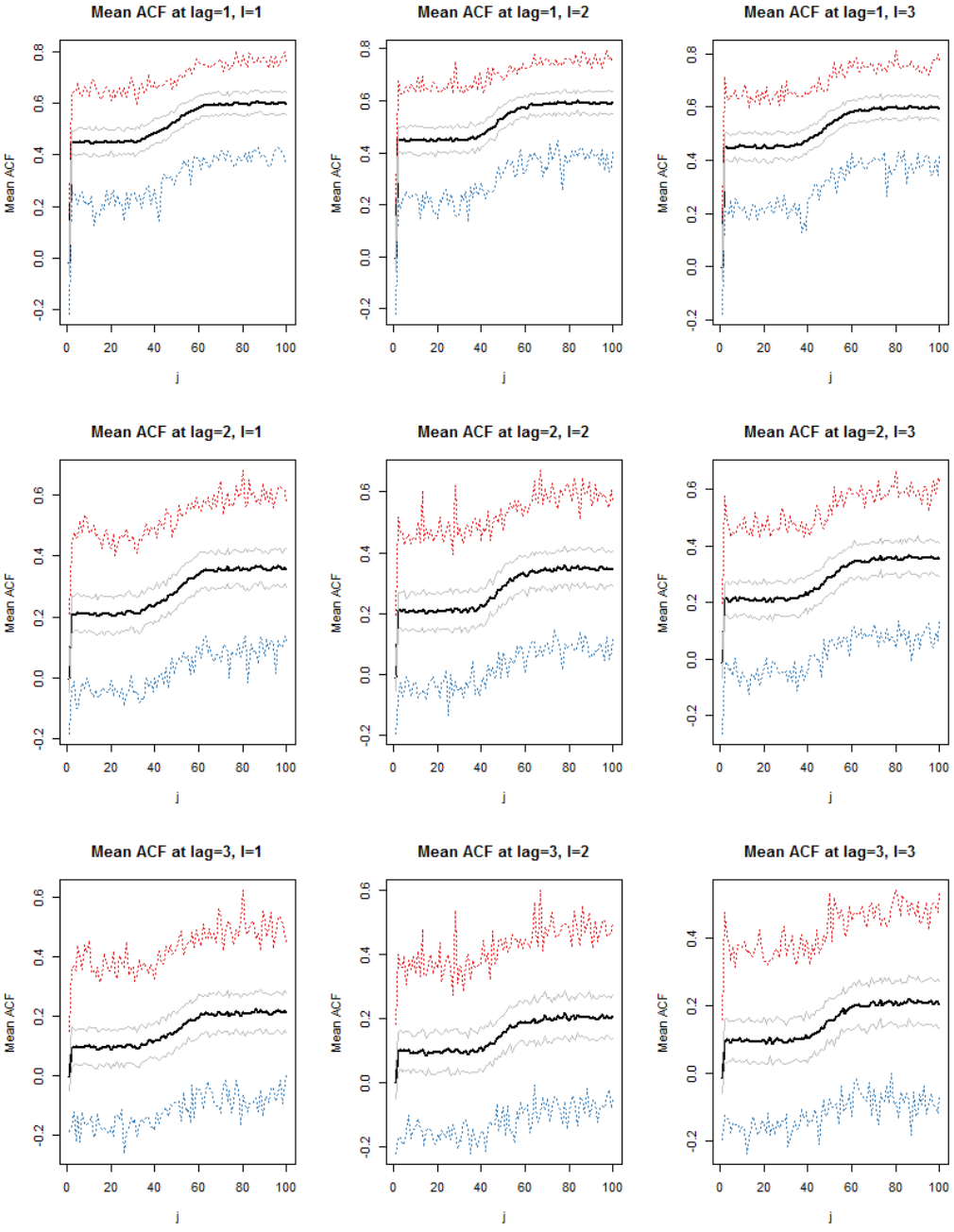

Figure D.3: PPEA Krackhardt's managers, m_1:
Black line: mean value of the ACF of all $c \cdot V = 320$ PPEA chains of $\theta_{j,l}$. *Grey lines*: upper and lower quartile. *Red line*: maximum ACF. *Blue line*: minimum ACF.
Upper panel: lag=1. *Middle panel:* lag=2. *Lower panel:* lag=3.

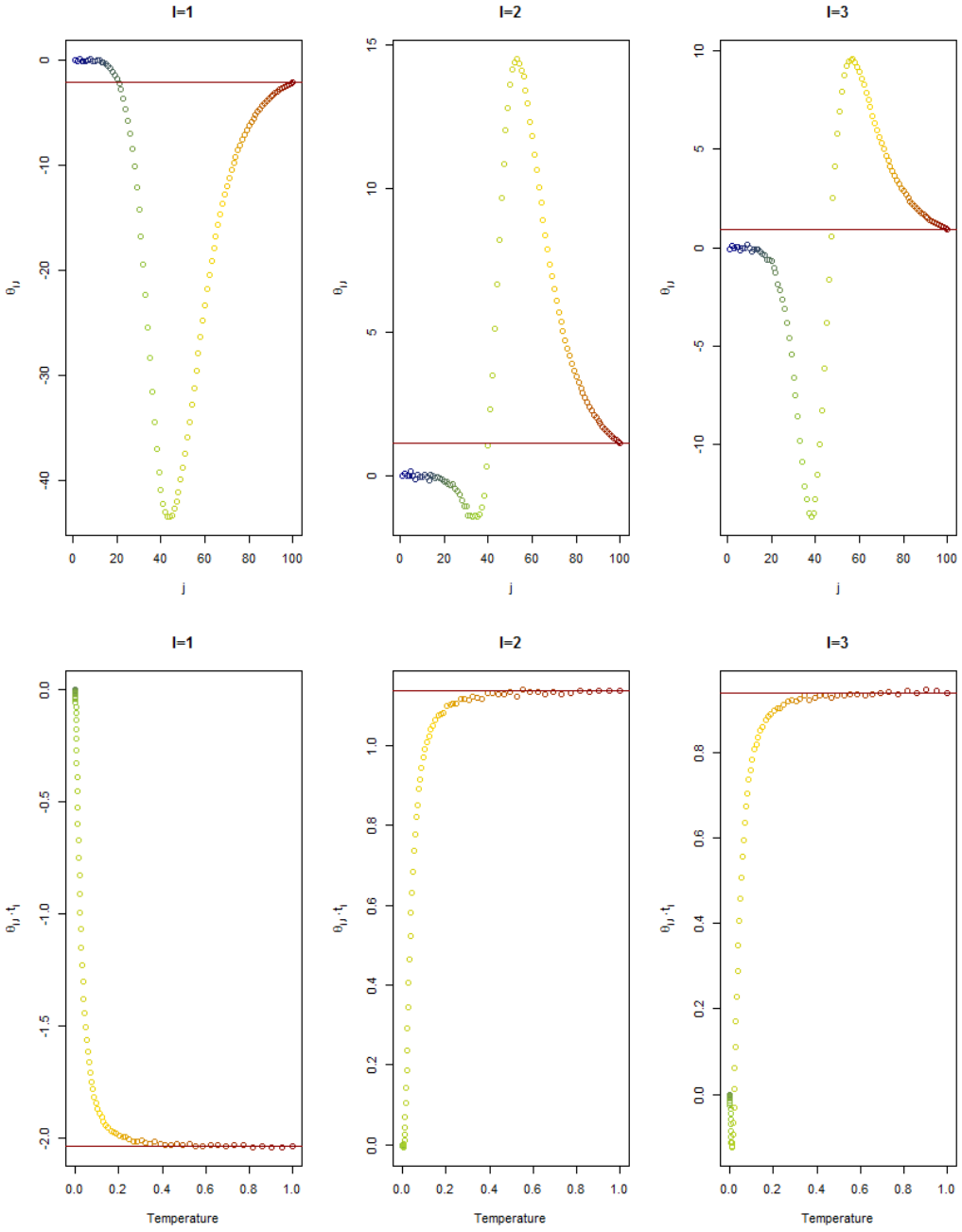

Figure D.4: PPEA Krackhardt's managers, m_1:
Upper panel: MCMC estimates of $\theta_{j,l}$. *Lower panel:* MCMC estimates of $\theta_{j,l}$ multiplied by t_j as a function of the temperature. The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t = 1$.

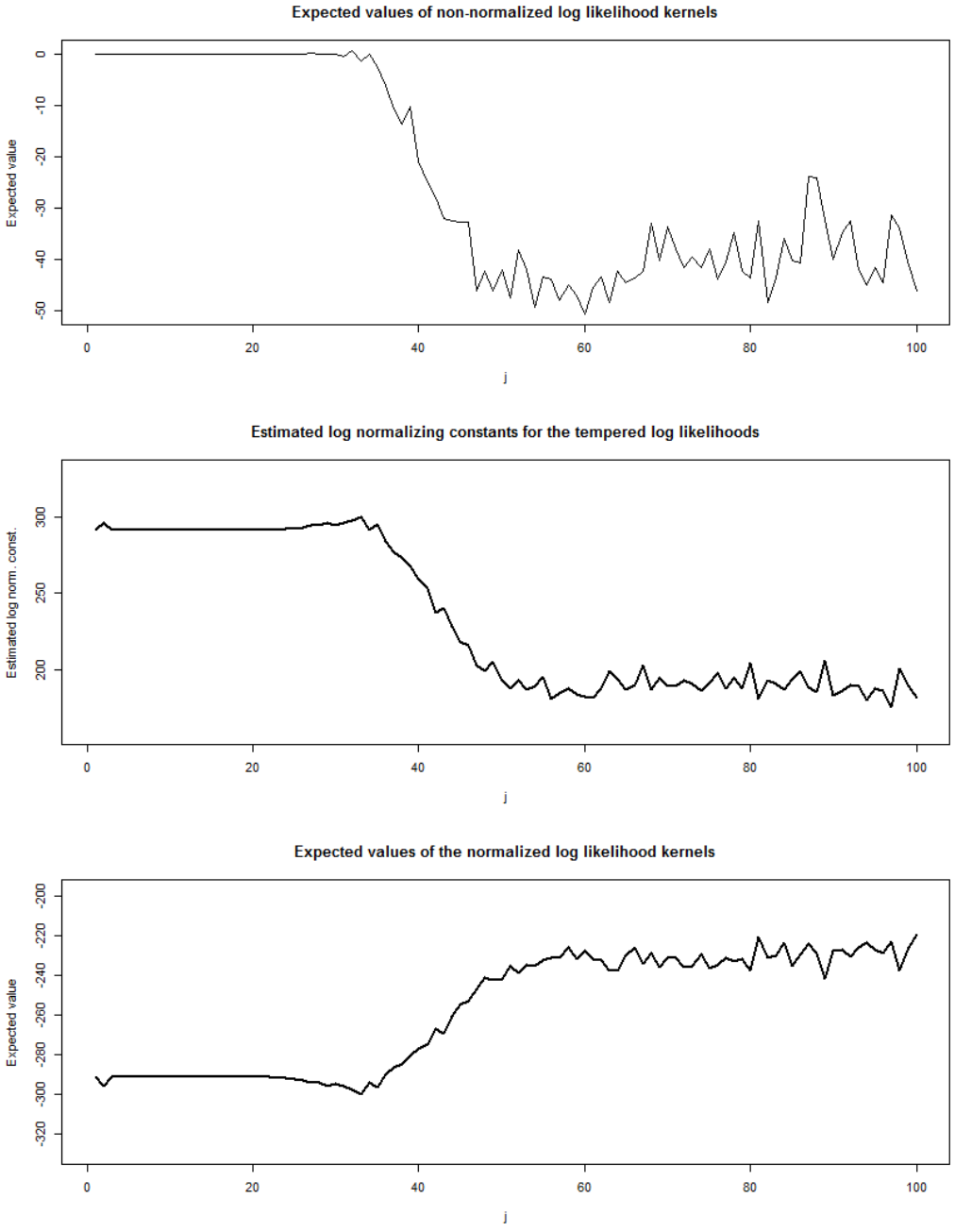

Figure D.5: PPEA Krackhardt's managers, m_1:
Top panel: Expected values of the non-normalized log likelihood kernels.
Middle panel: Importance sampling estimates of the normalizing constants of the tempered log likelihoods.
Lower panel: Expected values of the normalized tempered log likelihoods.

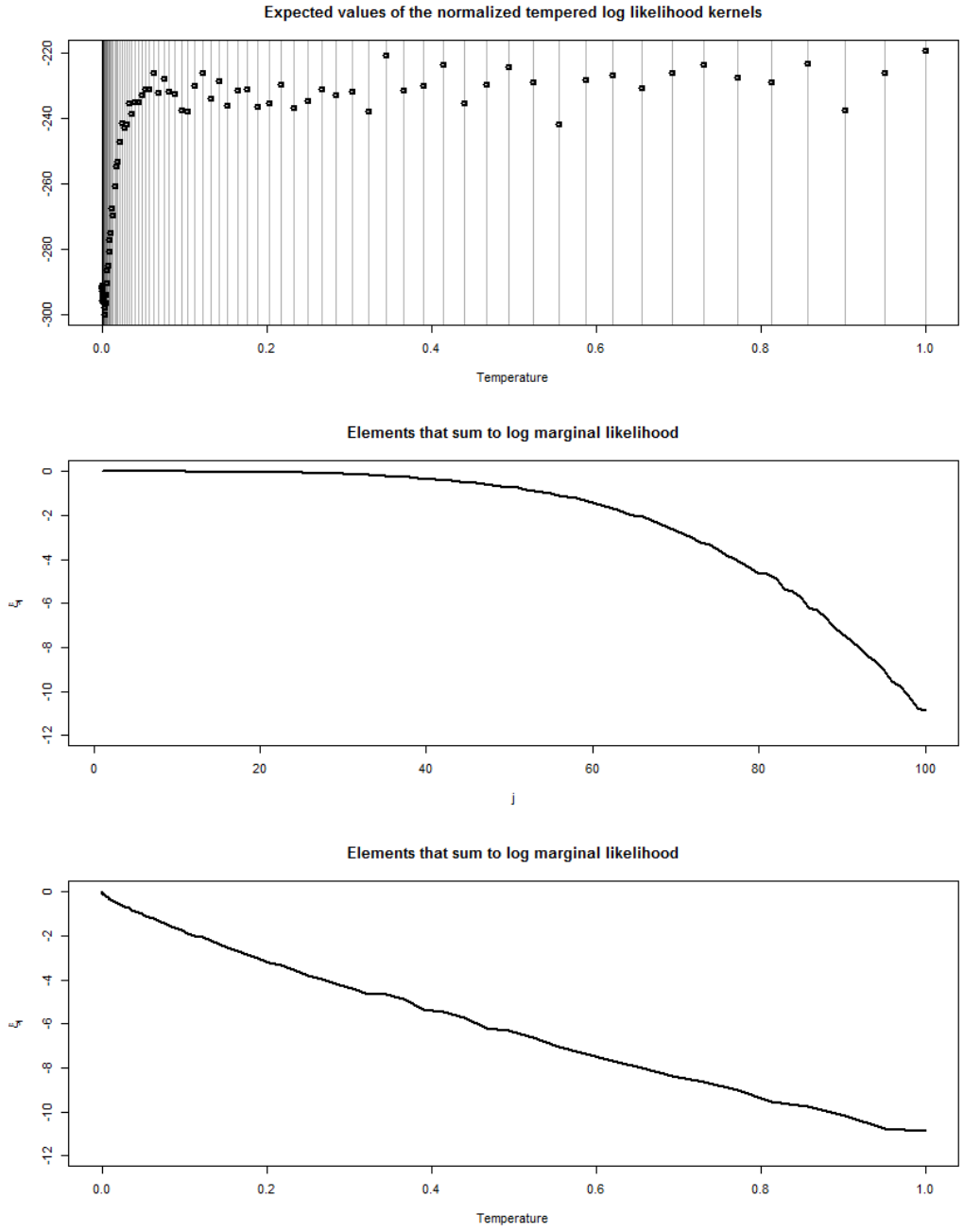

Figure D.6: PPEA Krackhardt's managers, m_1:
Top panel: Expected values of the normalized log likelihood kernels in relation to the temperature steps.
Middle panel: Summands of the trapezoidal approximation of the evidence.
Lower panel: Summands of the trapezoidal approximation in relation to the temperature.

D.4 PPEA Ghana

D.4.1 Model 1

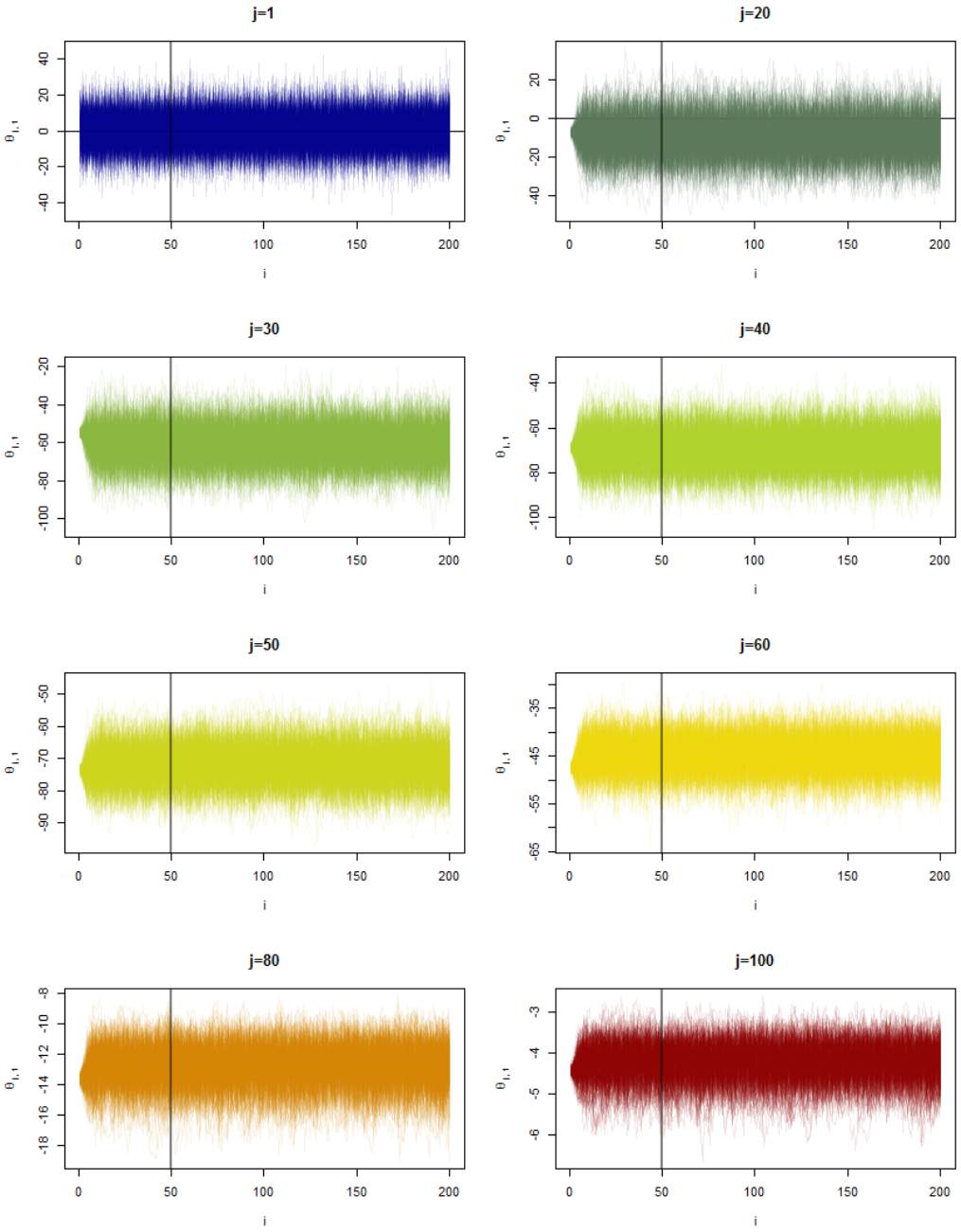

Figure D.7: PPEA Ghana, m_1:
Traceplots of the $c \cdot V = 320$ chains of the edge parameter $\theta_{j,l=1}$ at some temperatures j of the PPEA. The horizontal grey line indicates the burn-in period of $I_{burn} = 50$ iterations used.

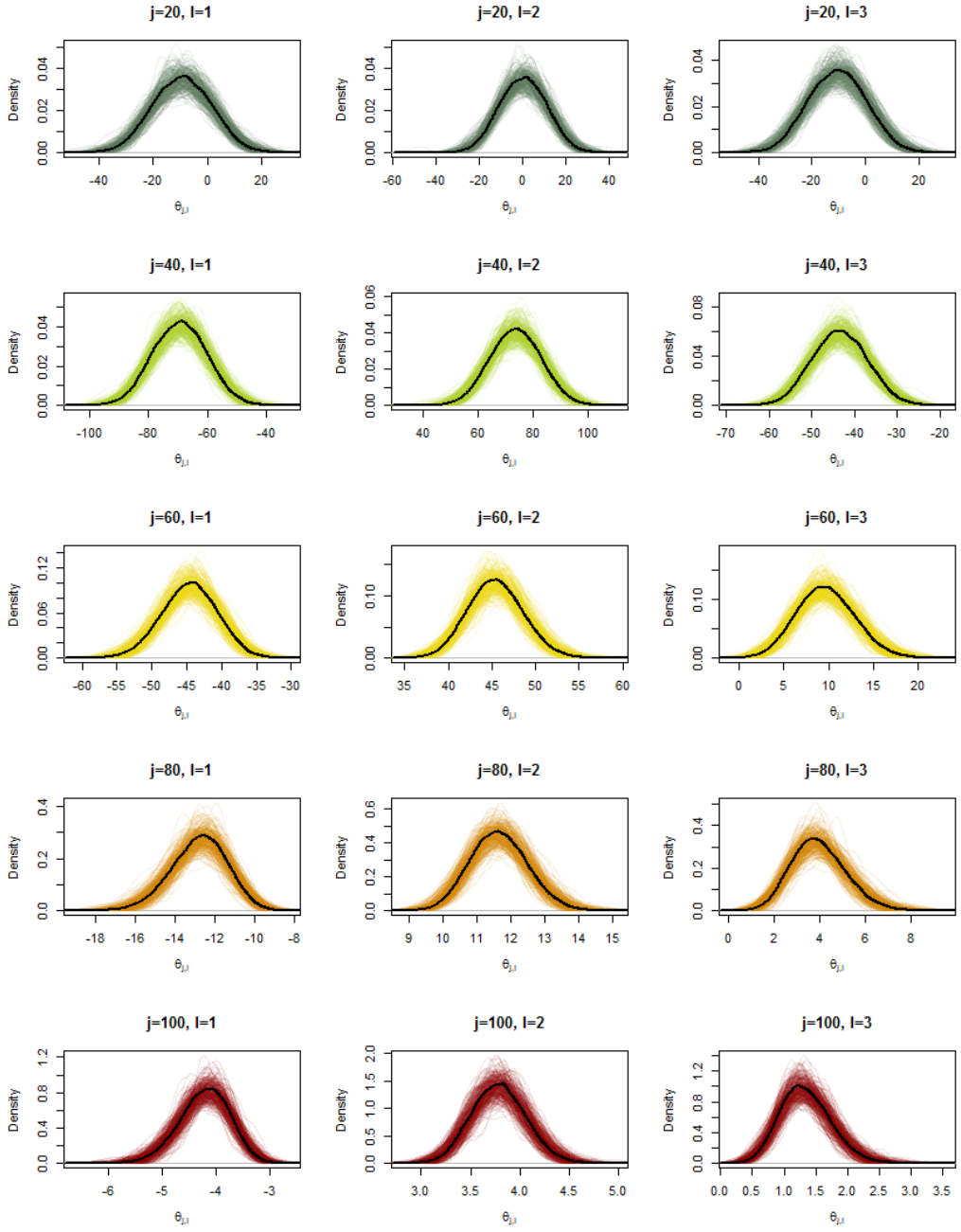

Figure D.8: PPEA Ghana, m_1:
Density plots of the $c \cdot V = 320$ chains of the parameters $\theta_{j,l}$ at some temperatures j of the PPEA. The black line indicates the density of the merged sample.

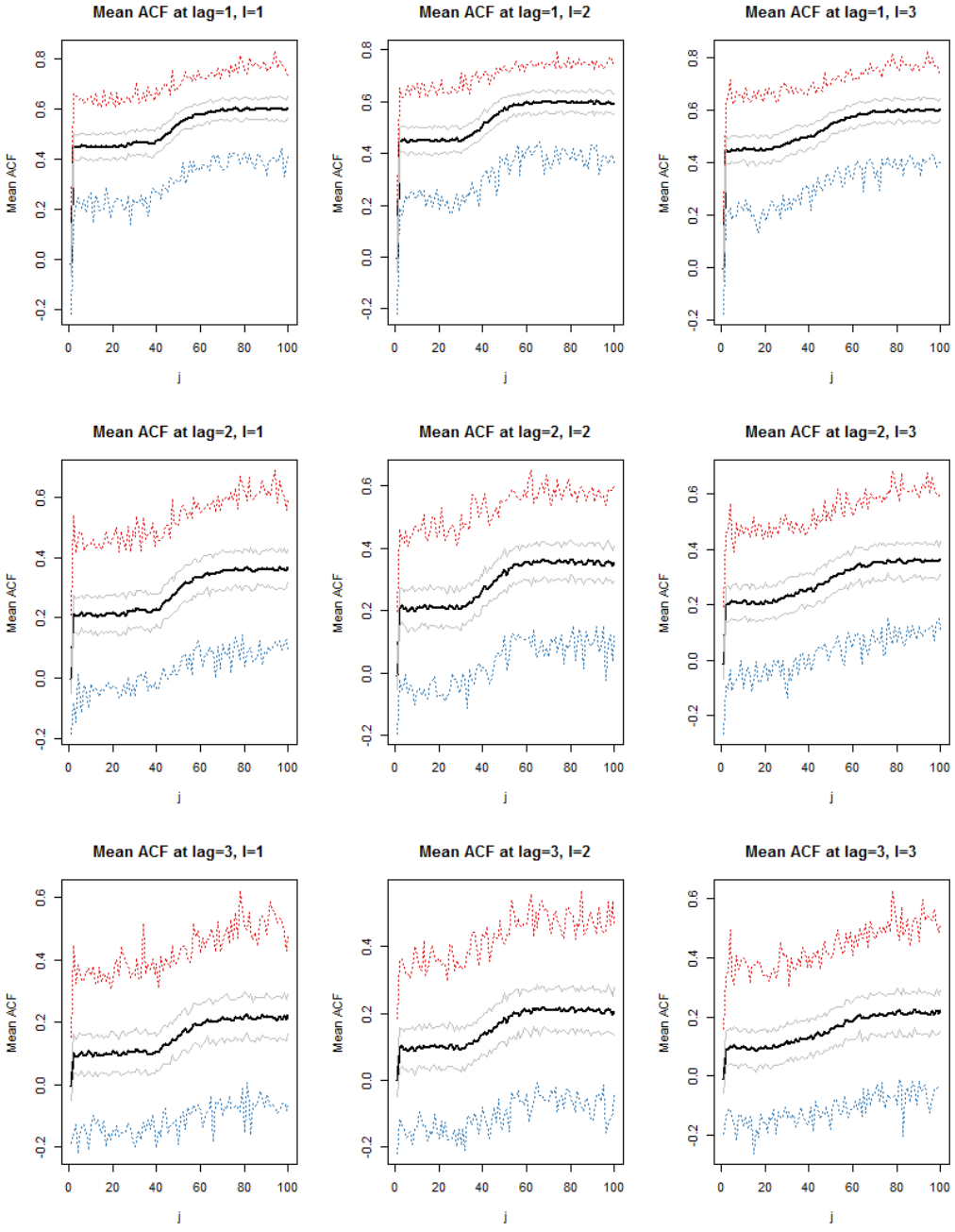

Figure D.9: PPEA Ghana, m_1:
Black line: mean value of the ACF of all $c \cdot V = 320$ PPEA chains of $\theta_{j,l}$. *Grey lines*: upper and lower quartile. *Red line*: maximum ACF. *Blue line*: minimum ACF.
Upper panel: lag=1. *Middle panel:* lag=2. *Lower panel:* lag=3.

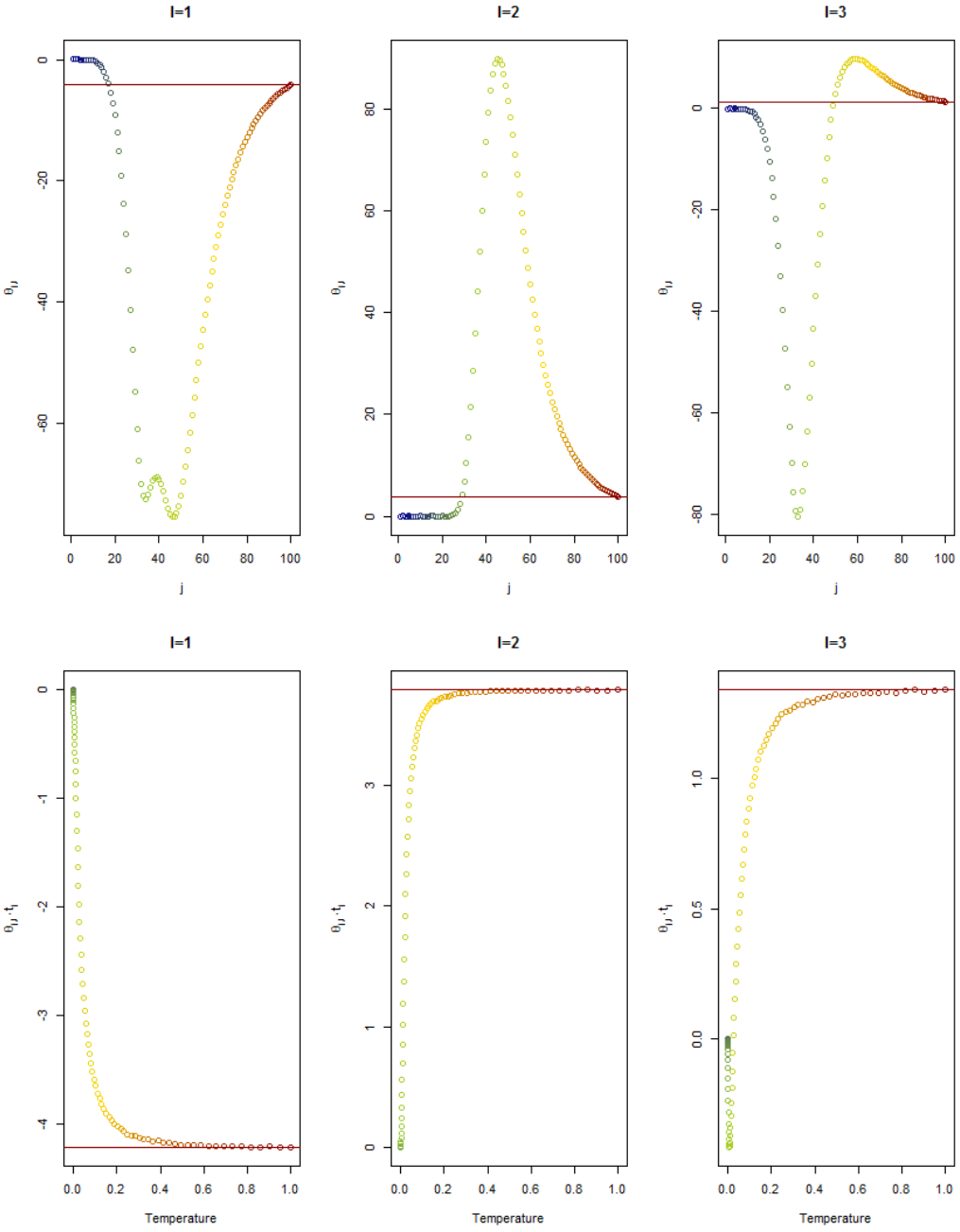

Figure D.10: PPEA Ghana, m_1:
Upper panel: Path of the MCMC estimates of $\theta_{j,l}$.
Lower panel: Path of the MCMC estimates of $\theta_{j,l}$ multiplied by t_j over the temperature. The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t_{j=J} = 1$.

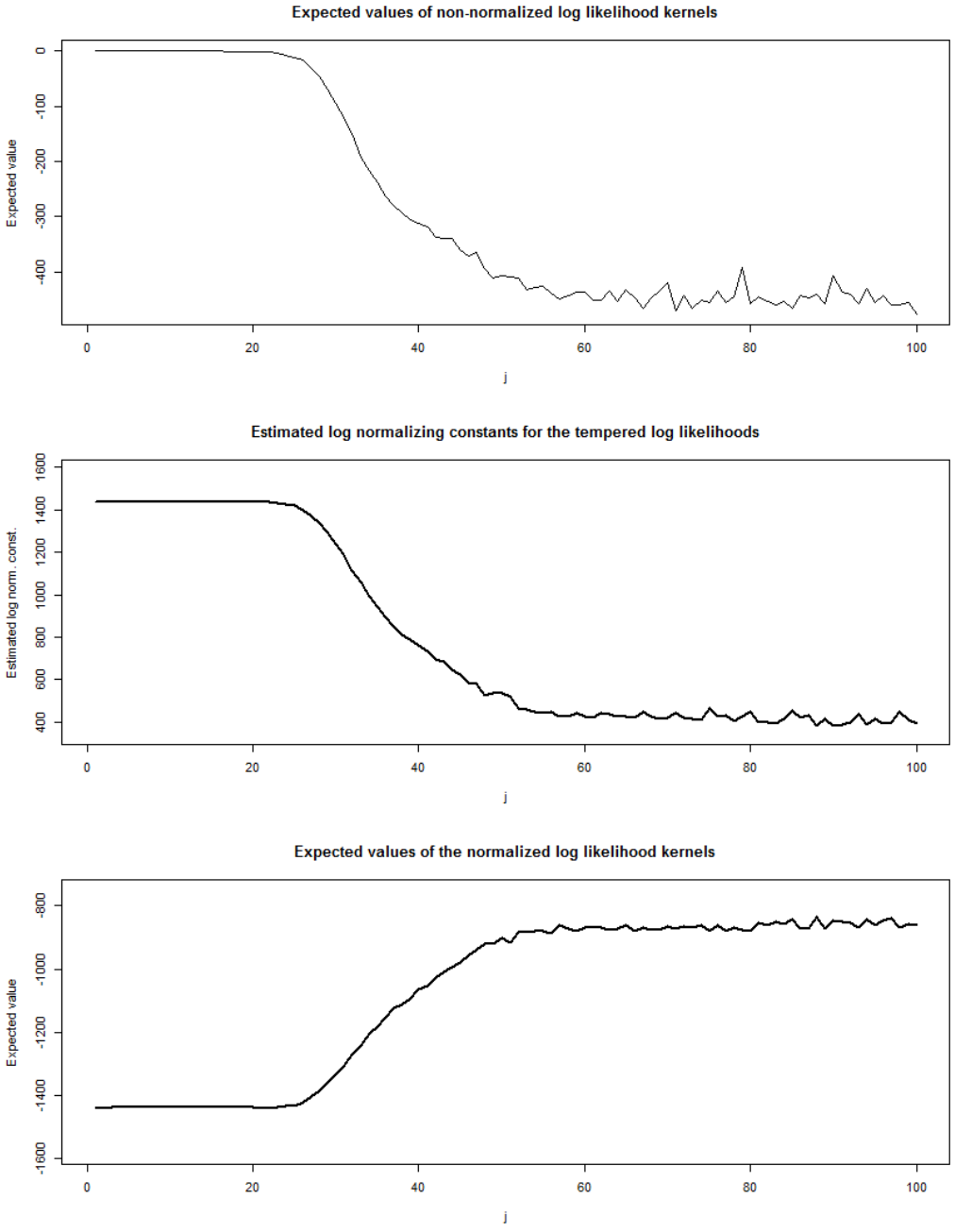

Figure D.11: PPEA Ghana, m_1:
Top panel: Expected values of the non-normalized log likelihood kernels.
Middle panel: Importance sampling estimates of the normalizing constants of the tempered log likelihoods.
Lower panel: Expected values of the normalized tempered log likelihoods.

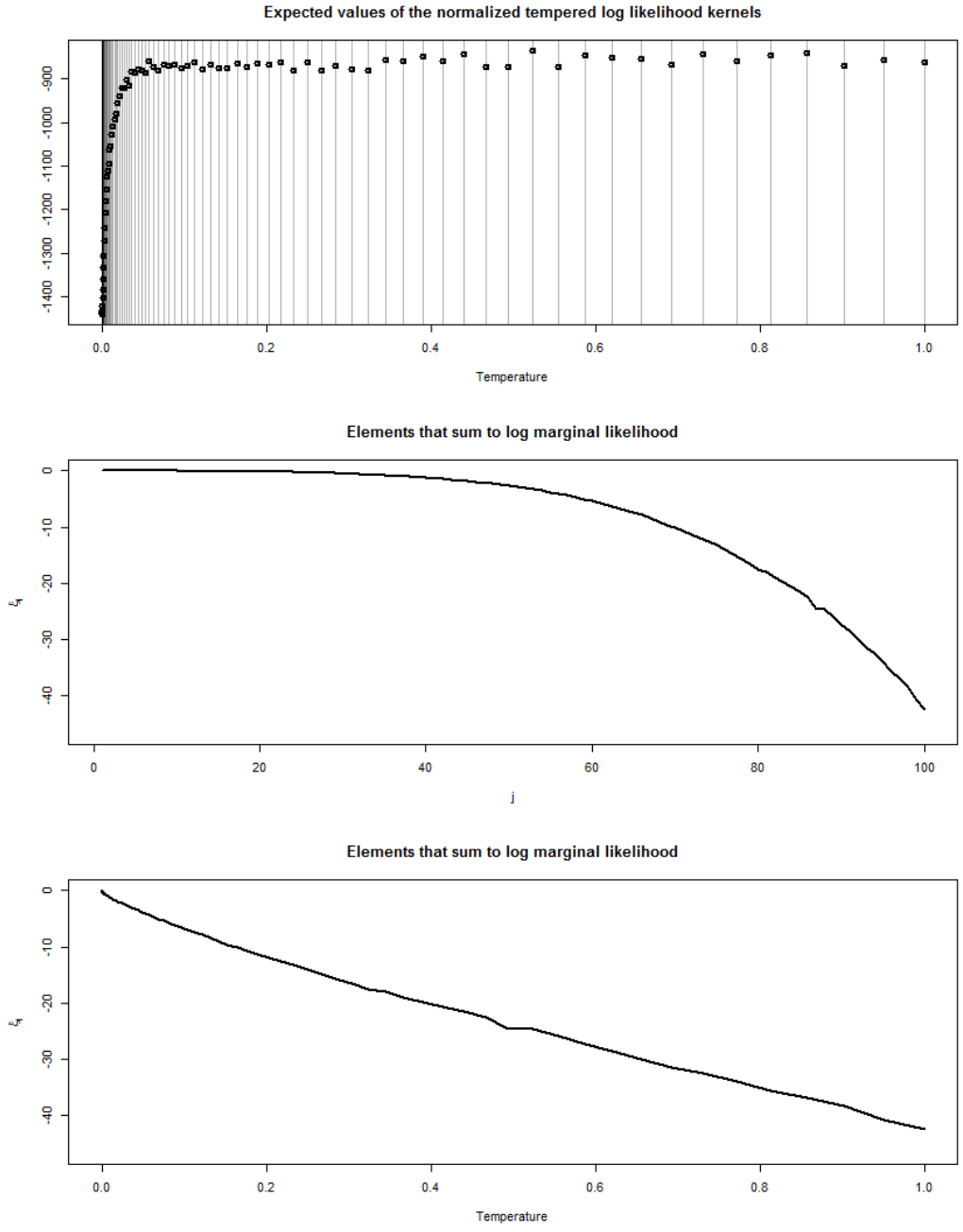

Figure D.12: PPEA Ghana, m_1:
Top panel: Expected values of the normalized log likelihood kernels in relation to the temperature steps.
Middle panel: Summands of the trapezoidal approximation of the evidence.
Lower panel: Summands of the trapezoidal approximation in relation to the temperature.

D.4.2 Model 2

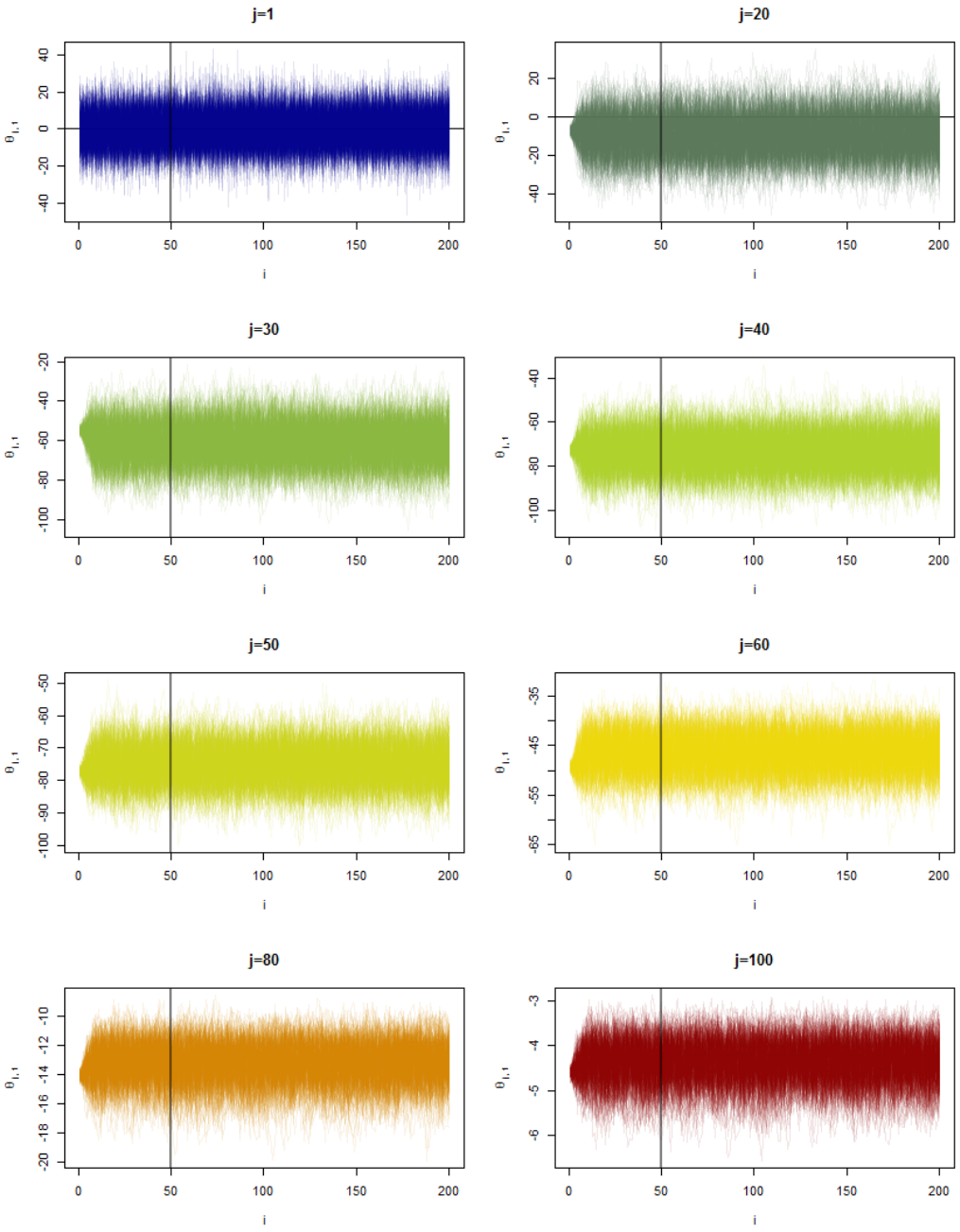

Figure D.13: PPEA Ghana, m_2:
Traceplots of the $c \cdot V = 320$ chains of the edge parameter $\theta_{j,l=1}$ at some temperatures j of the PPEA. The horizontal grey line indicates the burn-in period of $I_{burn} = 50$ iterations used.

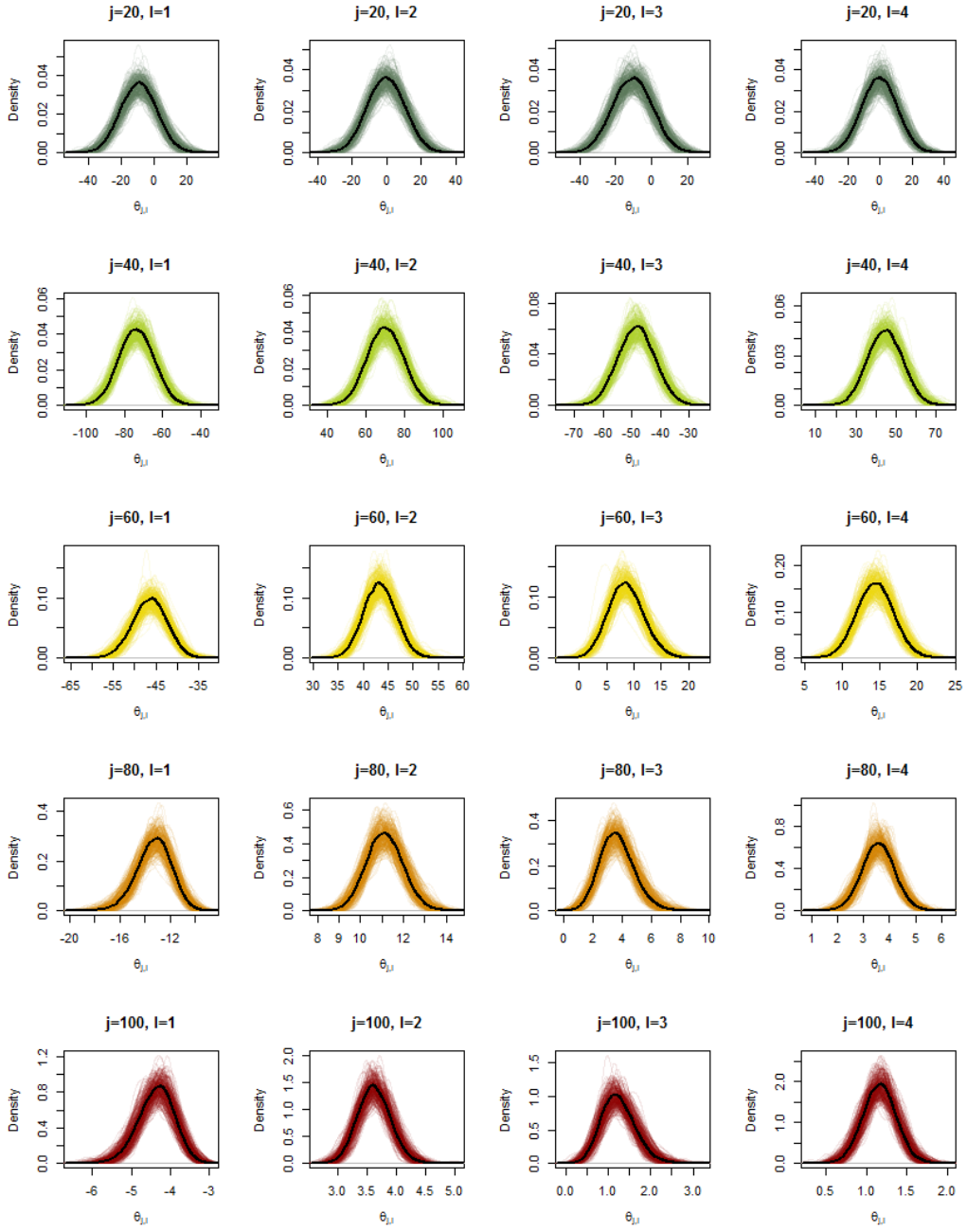

Figure D.14: PPEA Ghana, m_2:
Density plots of the $c \cdot V = 320$ chains of the parameters $\theta_{j,l}$ at some temperatures j of the PPEA. The black line indicates the density of the merged sample.

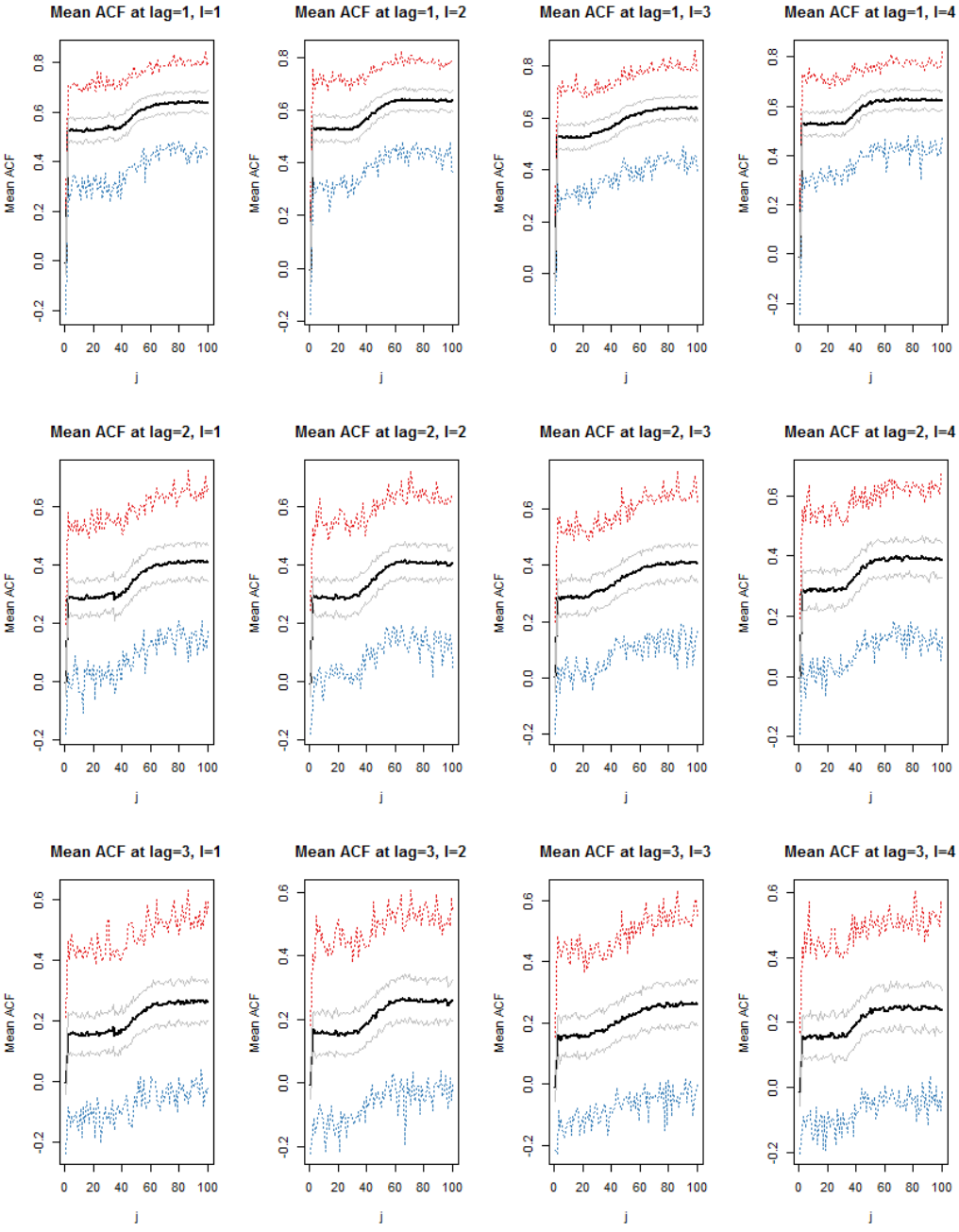

Figure D.15: PPEA Ghana, model 2:
Black line: mean value of the ACF of all $c \cdot V = 320$ PPEA chains of $\theta_{j,l}$. *Grey lines*: upper and lower quartile. *Red line*: maximum ACF. Blue line: minimum ACF.
Upper panel: lag=1. *Middle panel:* lag=2. *Lower panel:* lag=3.

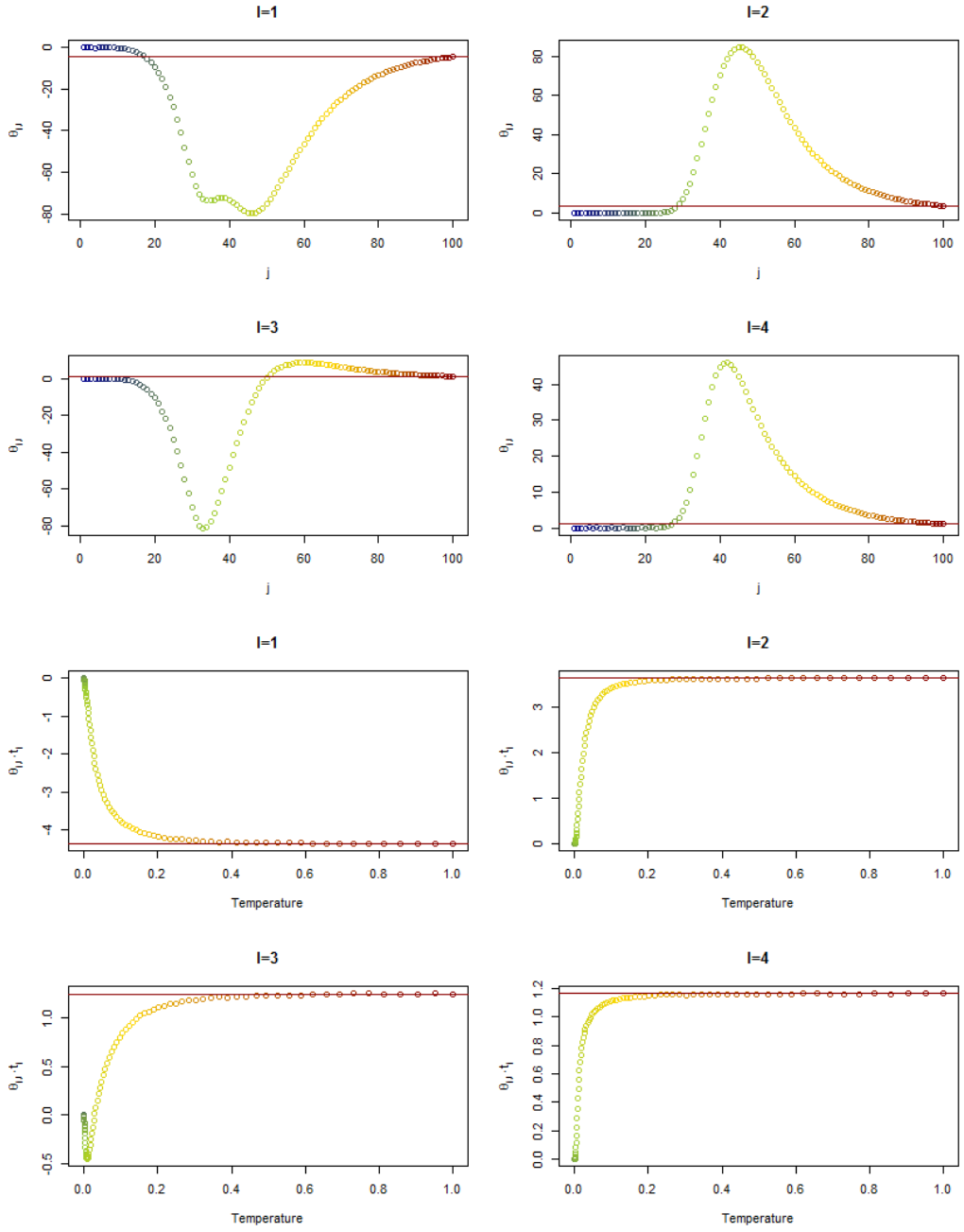

Figure D.16: PPEA Ghana, m_2:
Upper two panels: Path of the MCMC estimates of $\theta_{j,l}$.
Lower two panels: Path of the MCMC estimates of $\theta_{j,l}$ multiplied by t_j over the temperature.
The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t_{j=J} = 1$.

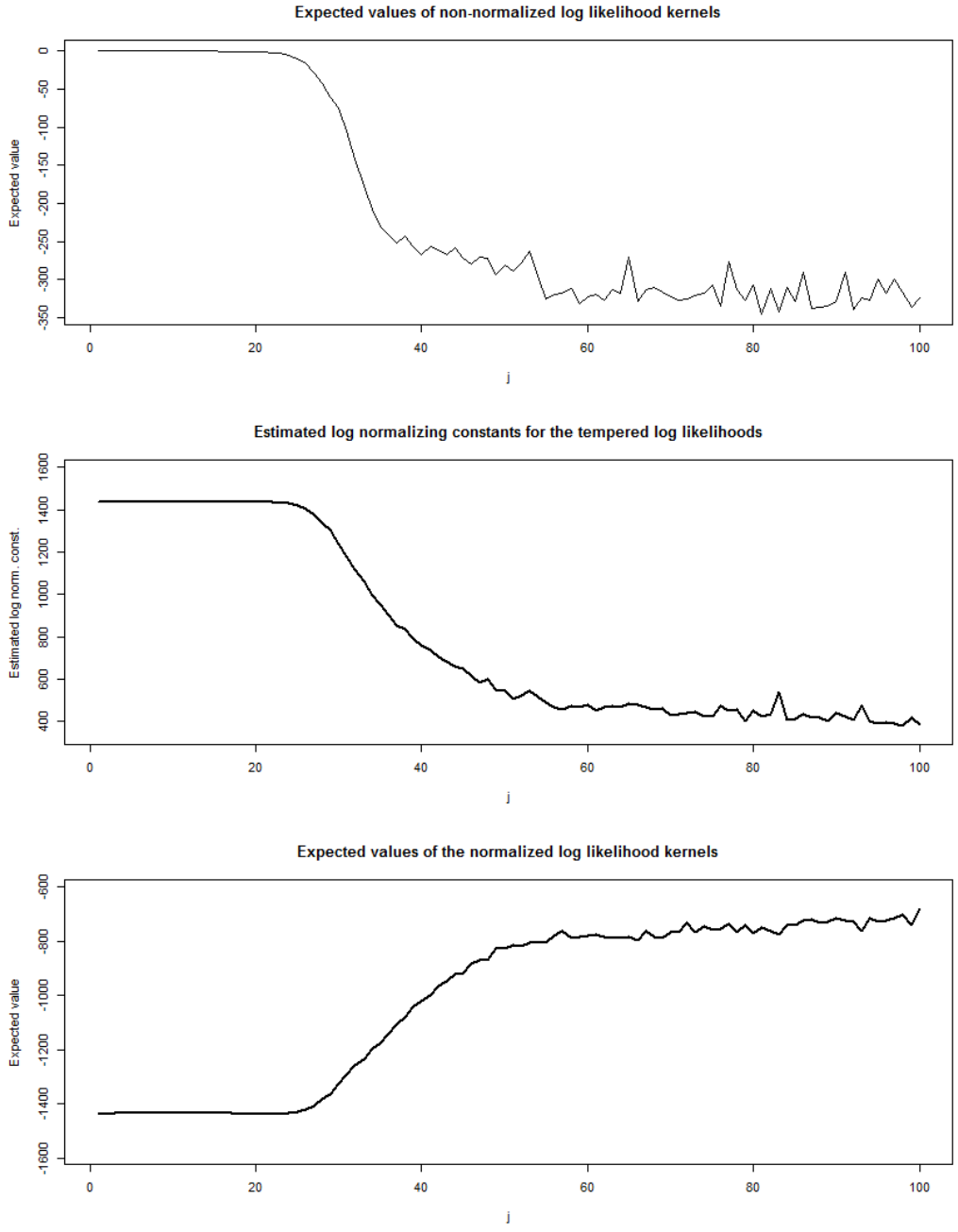

Figure D.17: PPEA Ghana, m_2:
Top panel: Expected values of the non-normalized log likelihood kernels.
Middle panel: Importance sampling estimates of the normalizing constants of the tempered log likelihoods.
Lower panel: Expected values of the normalized tempered log likelihoods.

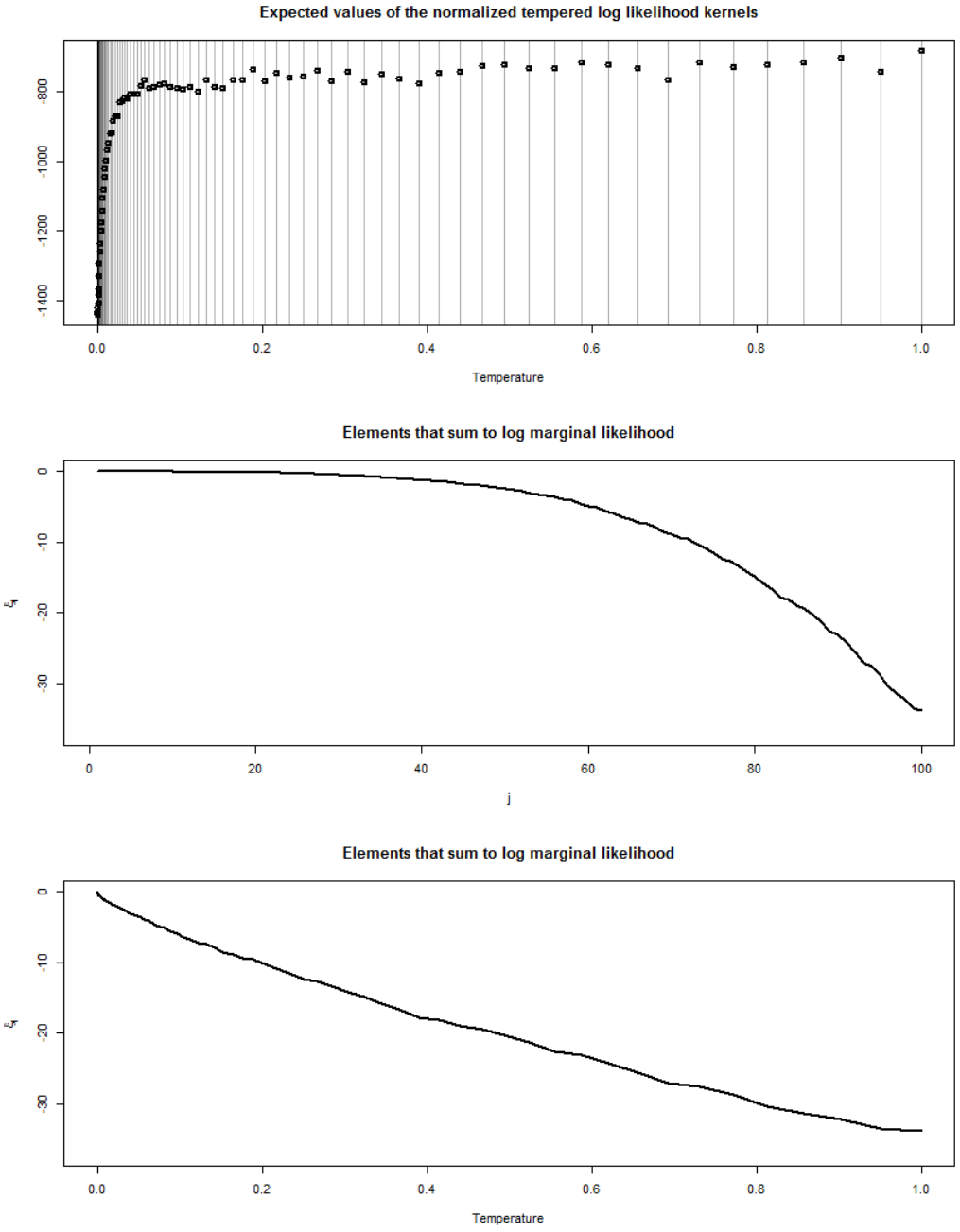

Figure D.18: PPEA Ghana, m_2:
Top panel: Expected values of the normalized log likelihood kernels in relation to the temperature steps.
Middle panel: Summands of the trapezoidal approximation of the evidence.
Lower panel: Summands of the trapezoidal approximation in relation to the temperature.

D.4.3 Model 3

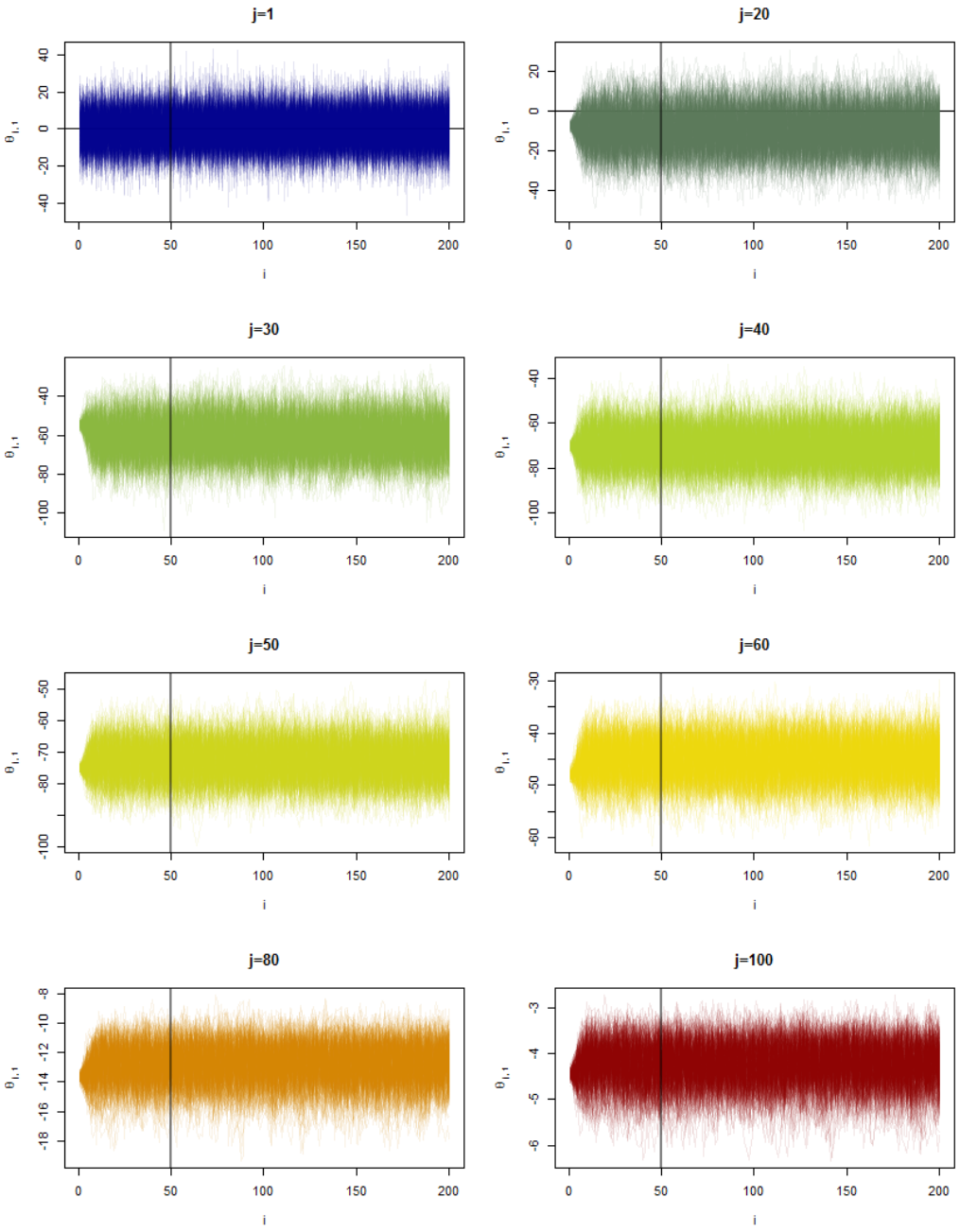

Figure D.19: PPEA Ghana, m_3:
Traceplots of the $c \cdot V = 320$ chains of the edge parameter $\theta_{j,l=1}$ at some temperatures j of the PPEA. The horizontal grey line indicates the burn-in period of $I_{burn} = 50$ iterations used.

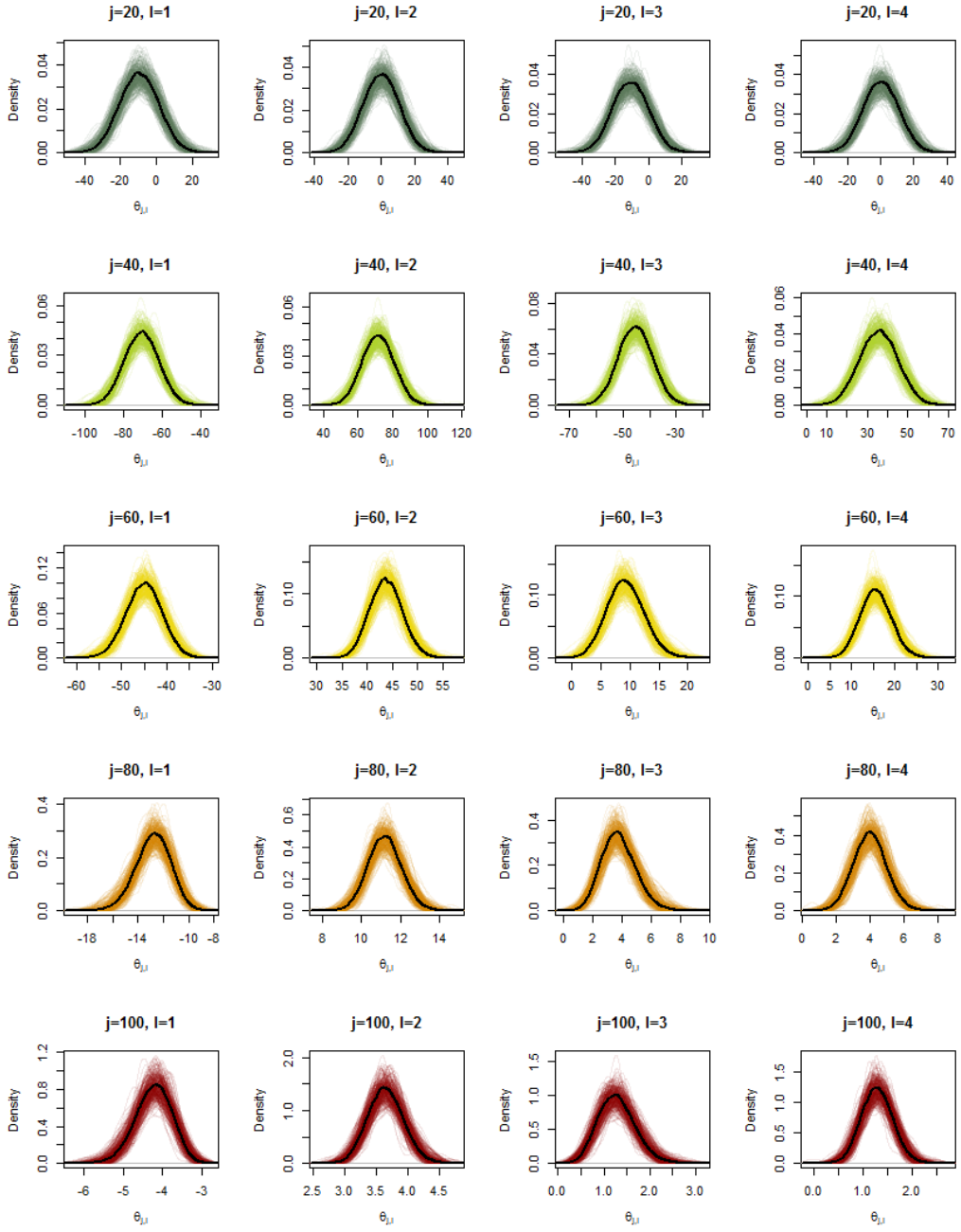

Figure D.20: PPEA Ghana, m_3:
Density plots of the $c \cdot V = 320$ chains of the parameters $\theta_{j,l}$ at some temperatures j of the PPEA. The black line indicates the density of the merged sample.

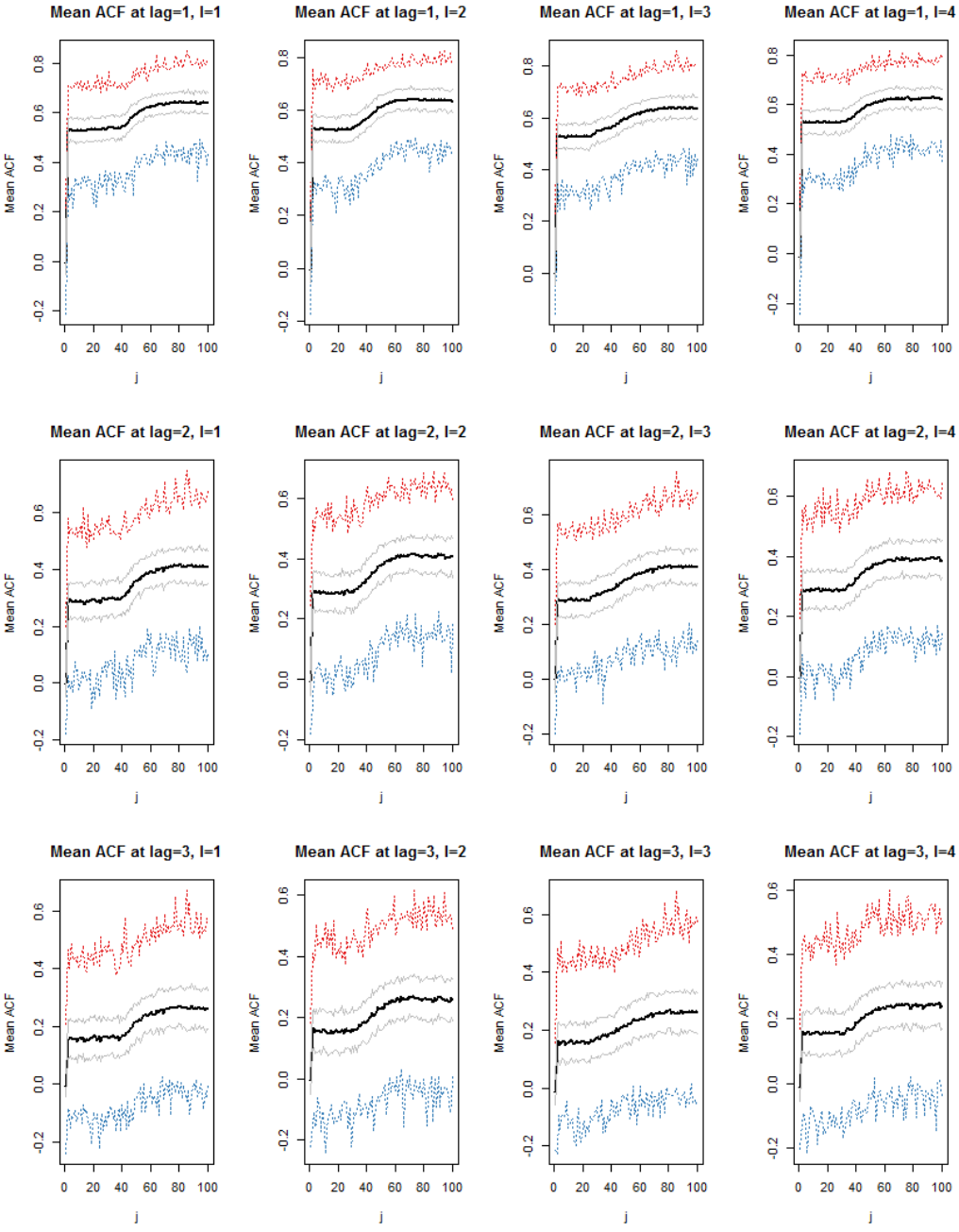

Figure D.21: PPEA Ghana, model 3:
Black line: mean value of the ACF of all $c \cdot V = 320$ PPEA chains of $\theta_{j,l}$. *Grey lines*: upper and lower quartile. *Red line*: maximum ACF. Blue line: minimum ACF.
Upper panel: lag=1. *Middle panel:* lag=2. *Lower panel:* lag=3.

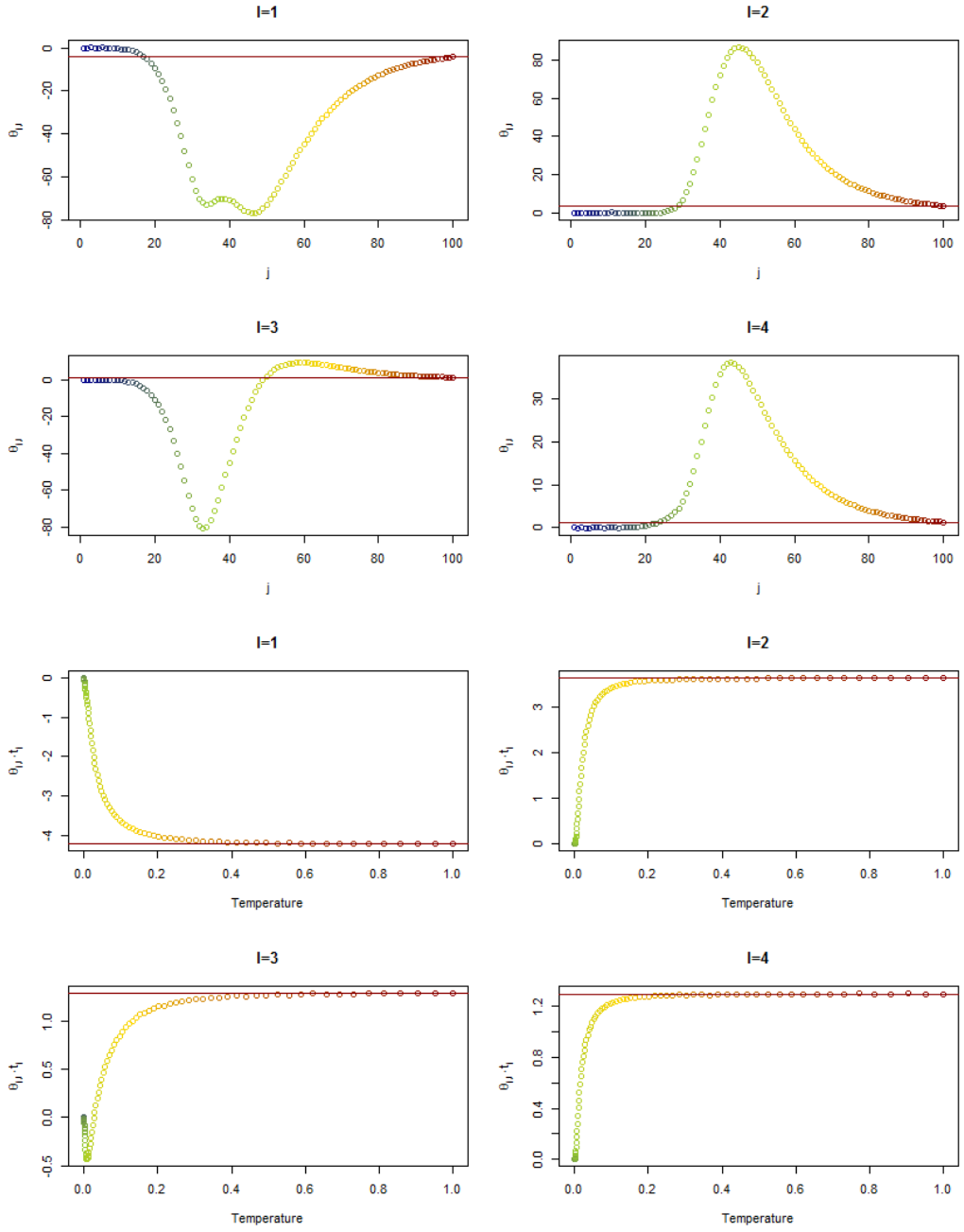

Figure D.22: PPEA Ghana, m_3:
Upper two panels: Path of the MCMC estimates of $\theta_{j,l}$.
Lower two panels: Path of the MCMC estimates of $\theta_{j,l}$ multiplied by t_j over the temperature.
The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t_{j=J} = 1$.

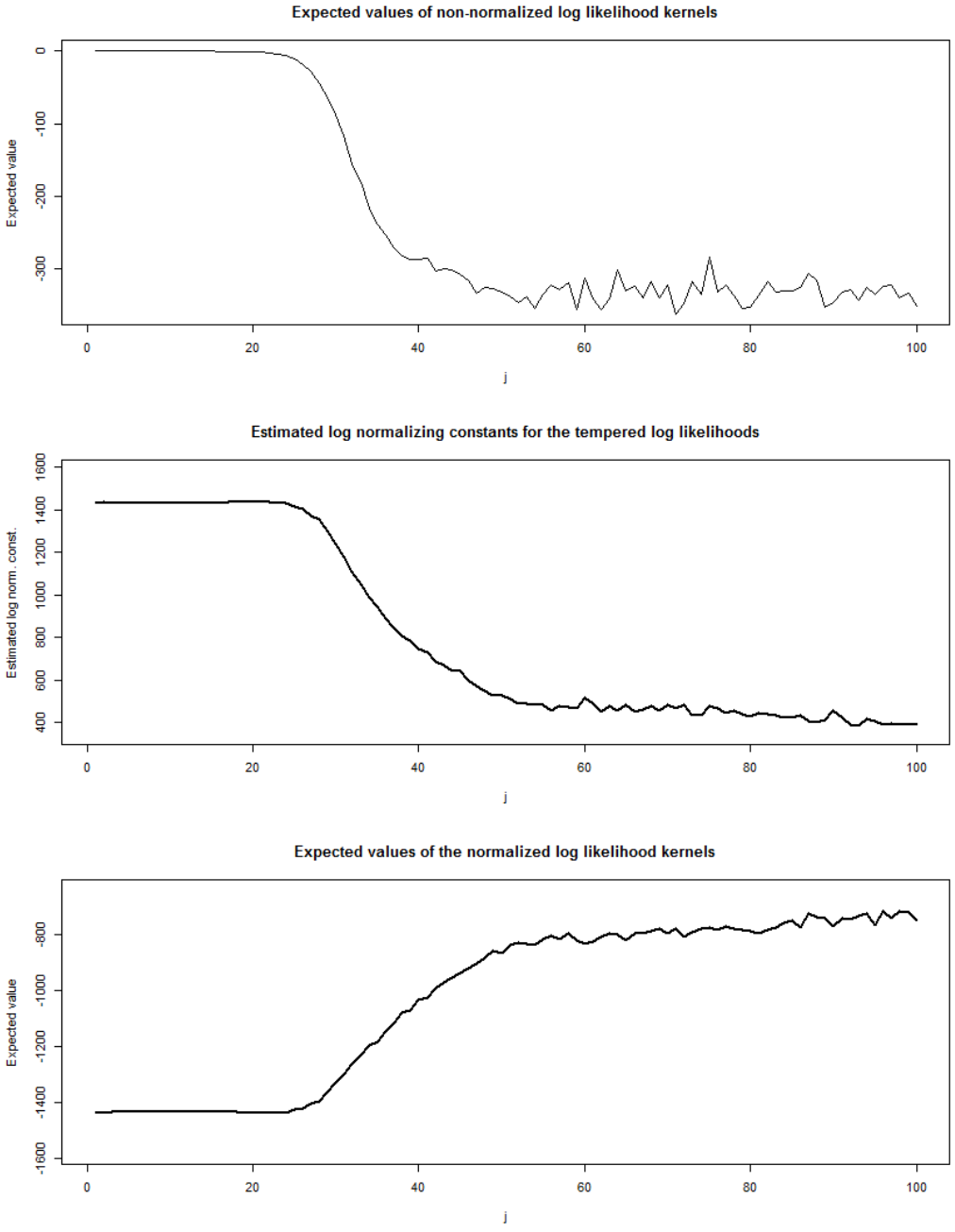

Figure D.23: PPEA Ghana, m_3:
Top panel: Expected values of the non-normalized log likelihood kernels.
Middle panel: Importance sampling estimates of the normalizing constants of the tempered log likelihoods.
Lower panel: Expected values of the normalized tempered log likelihoods.

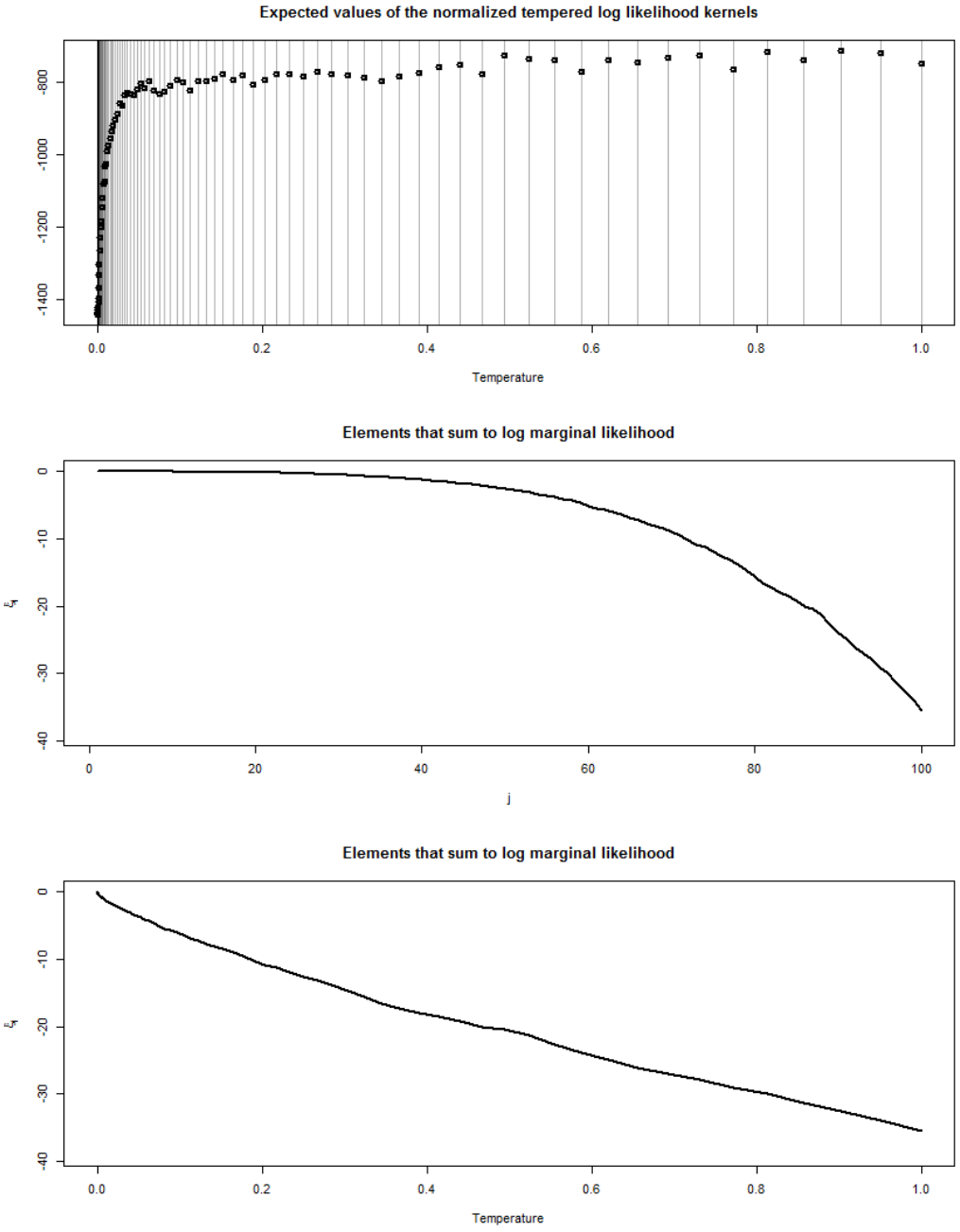

Figure D.24: PPEA Ghana, m_3:
Top panel: Expected values of the normalized log likelihood kernels in relation to the temperature steps.
Middle panel: Summands of the trapezoidal approximation of the evidence.
Lower panel: Summands of the trapezoidal approximation in relation to the temperature.

D.4.4 Model 4

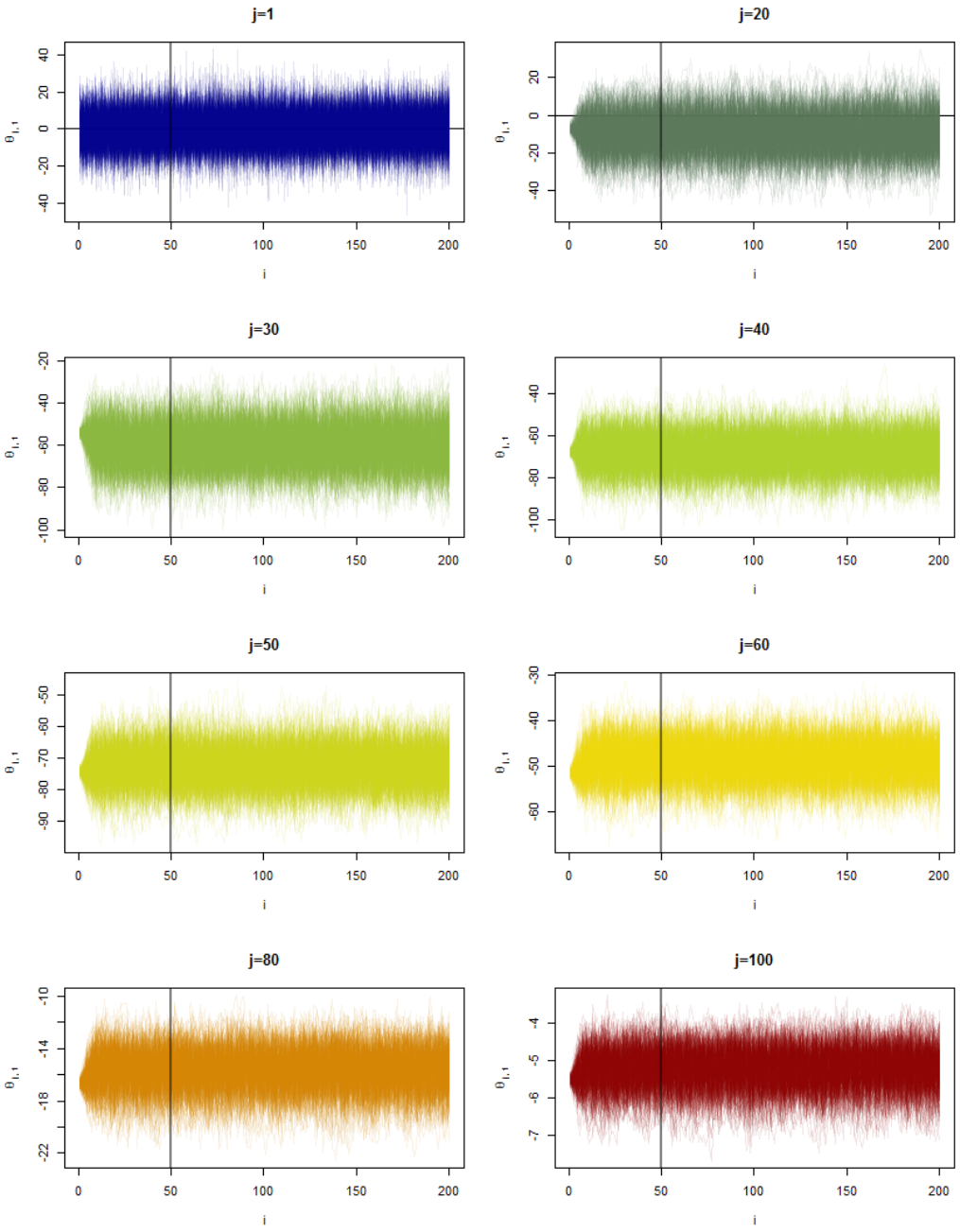

Figure D.25: PPEA Ghana, m_4:
Traceplots of the $c \cdot V = 320$ chains of the edge parameter $\theta_{j,l=1}$ at some temperatures j of the PPEA. The horizontal grey line indicates the burn-in period of $I_{burn} = 50$ iterations used.

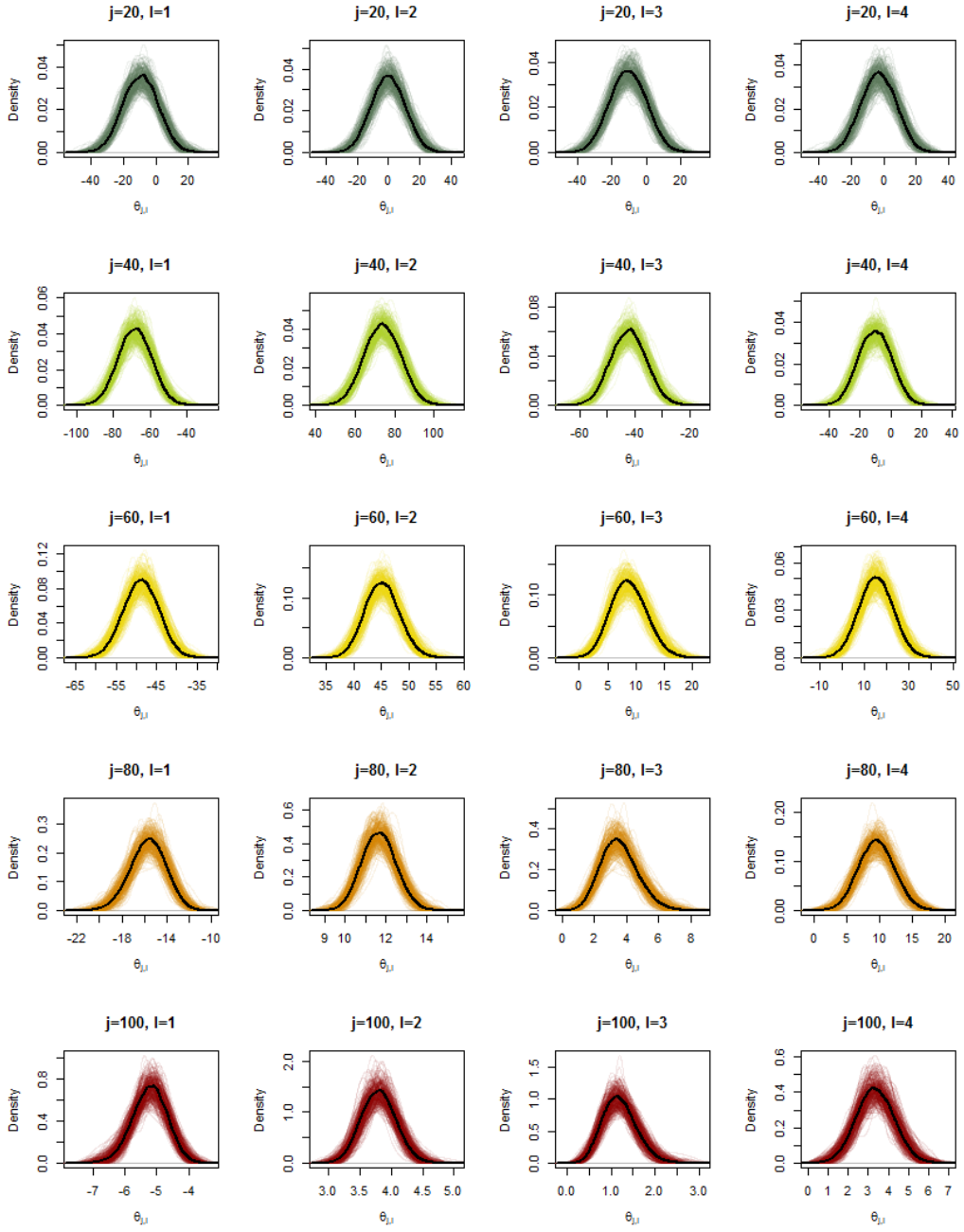

Figure D.26: PPEA Ghana, m_4:
Density plots of the $c \cdot V = 320$ chains of the parameters $\theta_{j,l}$ at some temperatures j of the PPEA. The black line indicates the density of the merged sample.

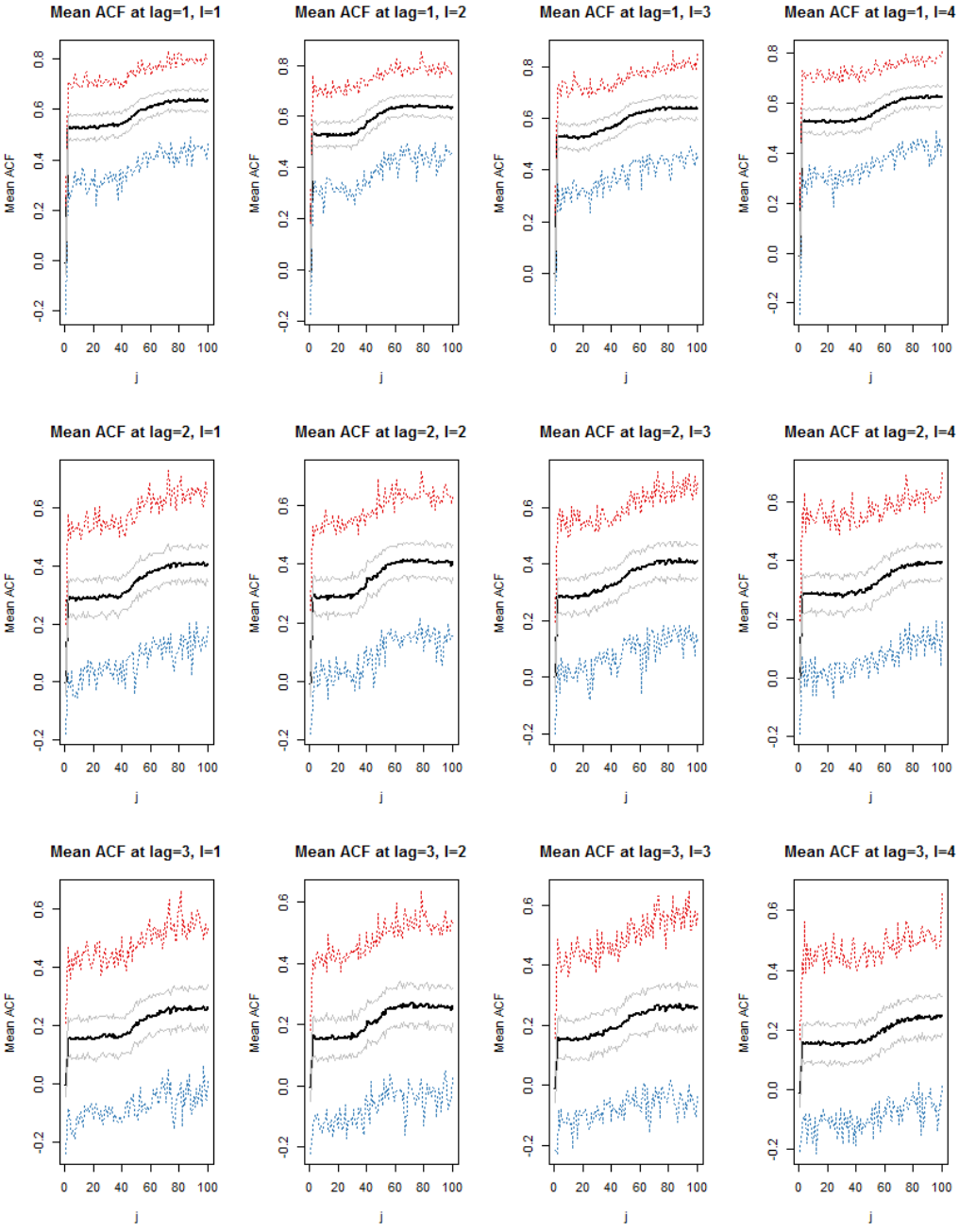

Figure D.27: PPEA Ghana, m_4:
Black line: mean value of the ACF of all $c \cdot V = 320$ PPEA chains of $\theta_{j,l}$. *Grey lines*: upper and lower quartile. *Red line*: maximum ACF. Blue line: minimum ACF.
Upper panel: lag=1. *Middle panel:* lag=2. *Lower panel:* lag=3.

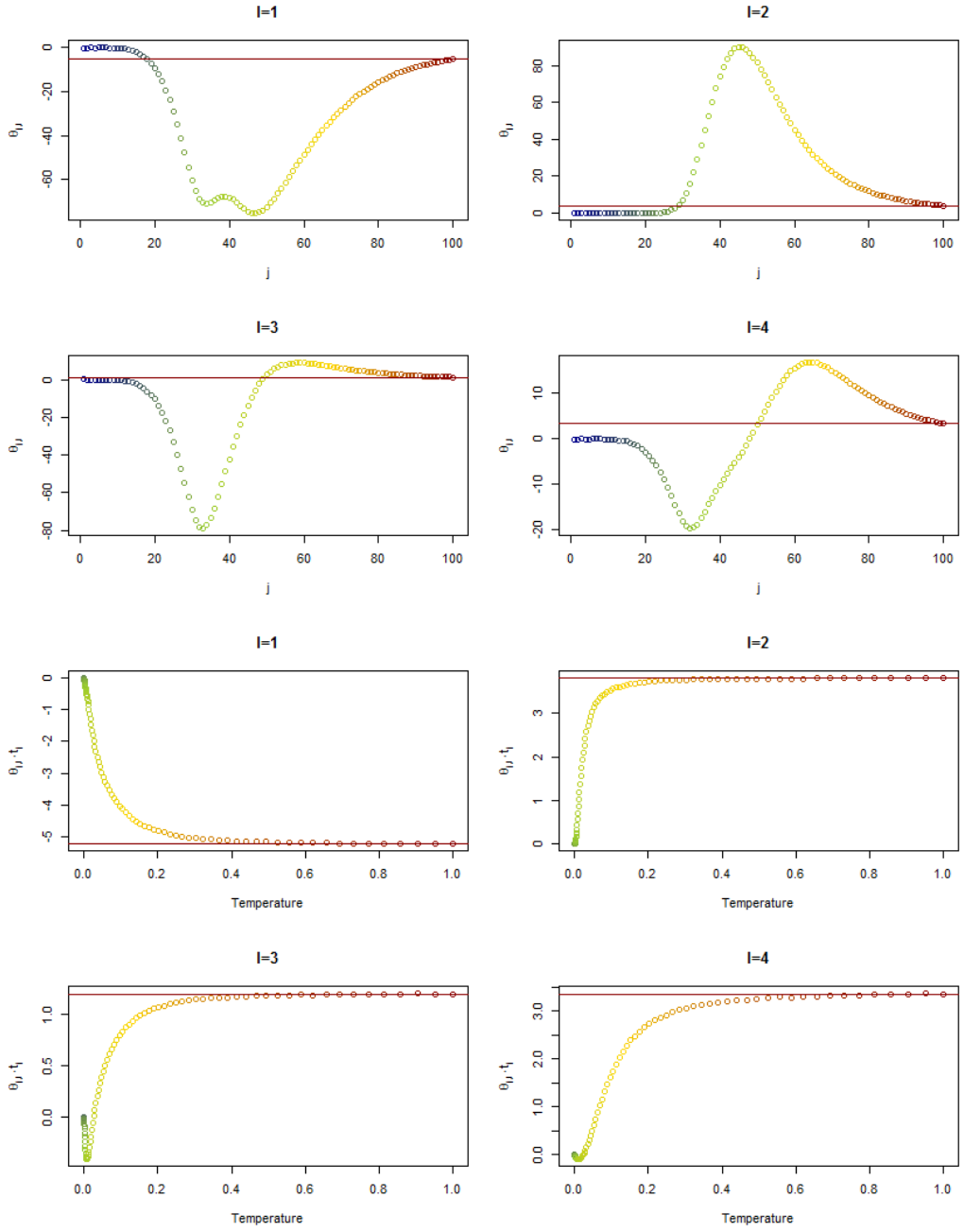

Figure D.28: PPEA Ghana, m_4:
Upper two panels: Path of the MCMC estimates of $\theta_{j,l}$.
Lower two panels: Path of the MCMC estimates of $\theta_{j,l}$ multiplied by t_j over the temperature.
The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t_{j=J} = 1$.

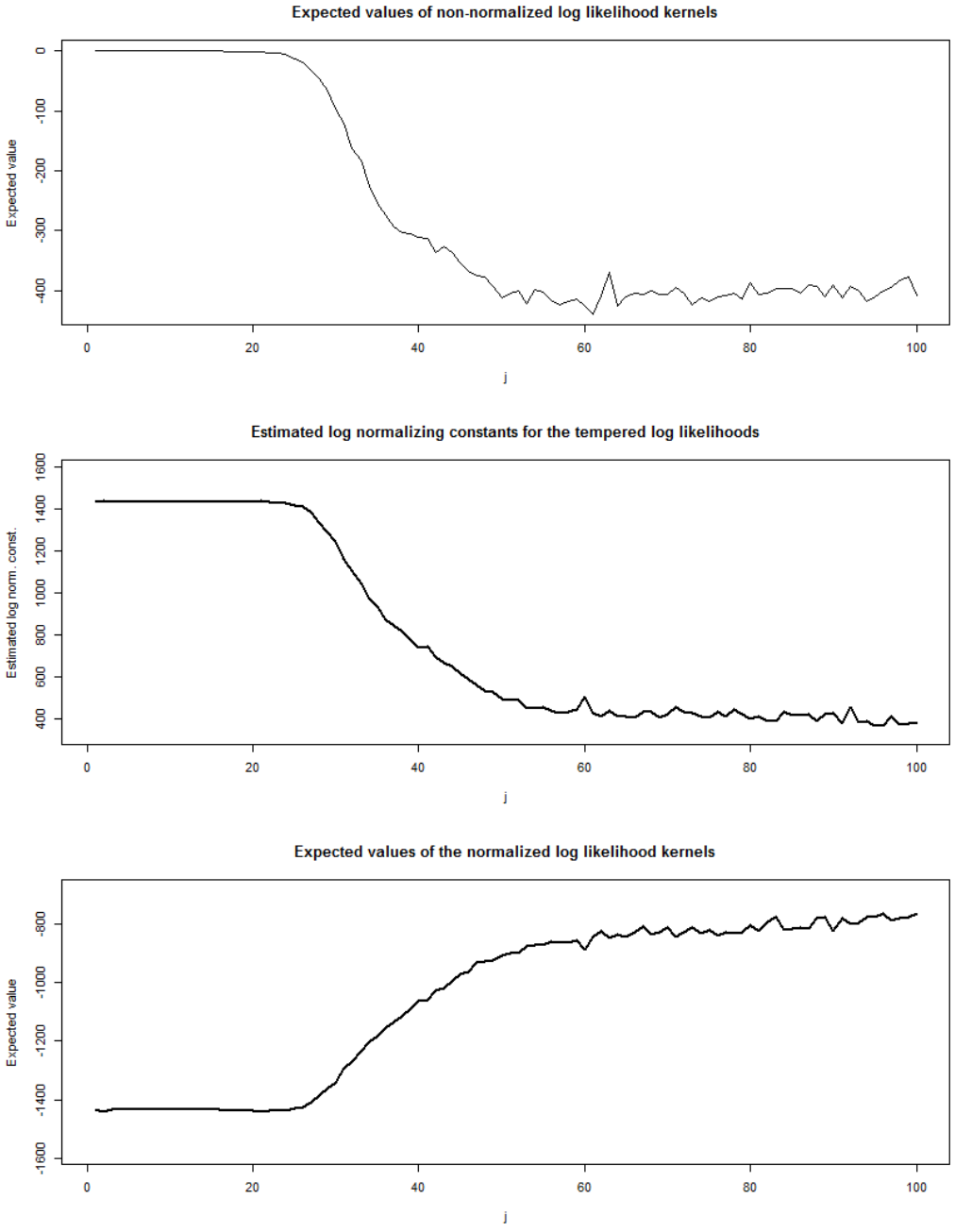

Figure D.29: PPEA Ghana, m_4:
Top panel: Expected values of the non-normalized log likelihood kernels.
Middle panel: Importance sampling estimates of the normalizing constants of the tempered log likelihoods.
Lower panel: Expected values of the normalized tempered log likelihoods.

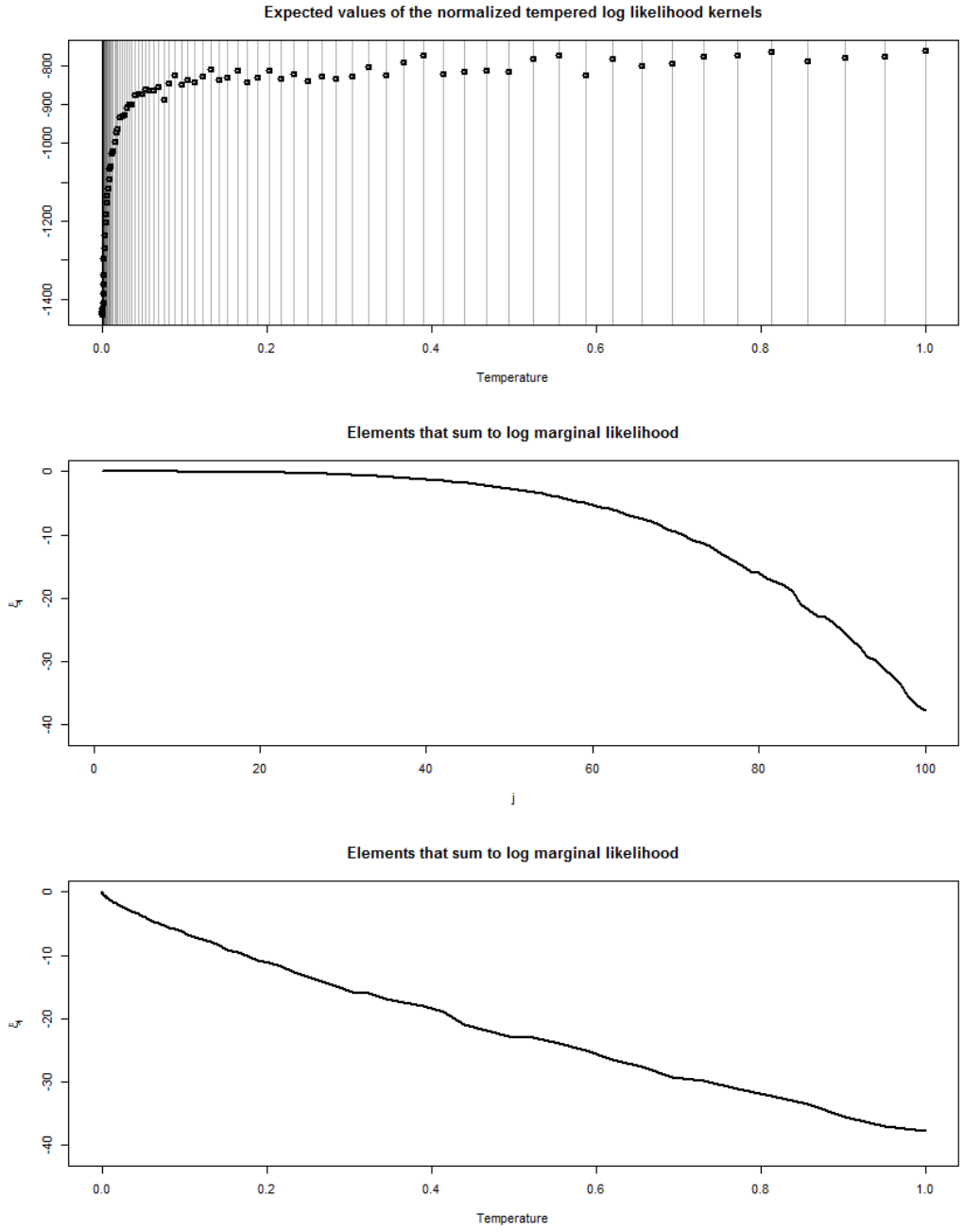

Figure D.30: PPEA Ghana, m_4:
Top panel: Expected values of the normalized log likelihood kernels in relation to the temperature steps.
Middle panel: Summands of the trapezoidal approximation of the evidence.
Lower panel: Summands of the trapezoidal approximation in relation to the temperature.

D.4.5 Model 5

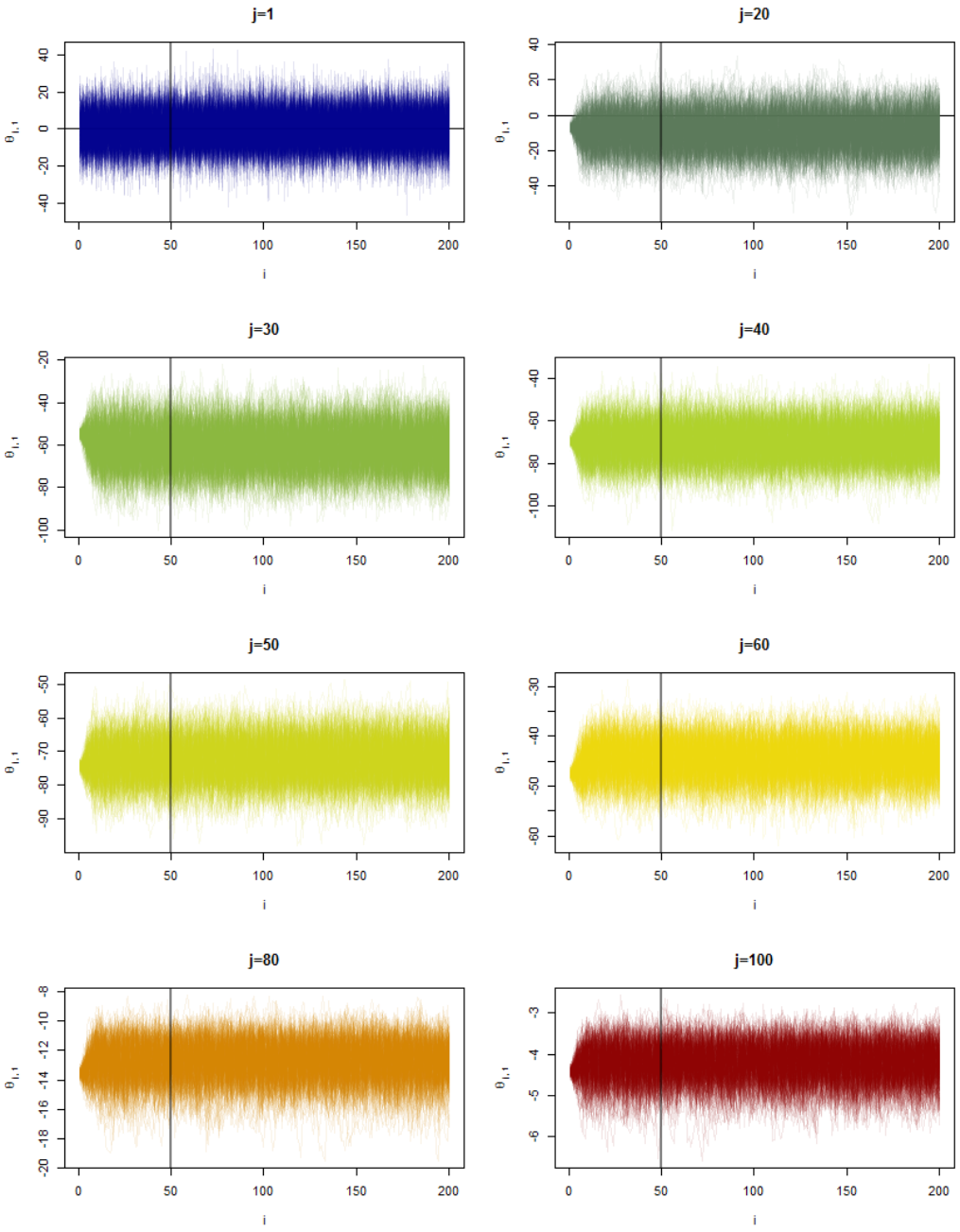

Figure D.31: PPEA Ghana, m_5:
Traceplots of the $c \cdot V = 320$ chains of the edge parameter $\theta_{j,l=1}$ at some temperatures j of the PPEA. The horizontal grey line indicates the burn-in period of $I_{burn} = 50$ iterations used.

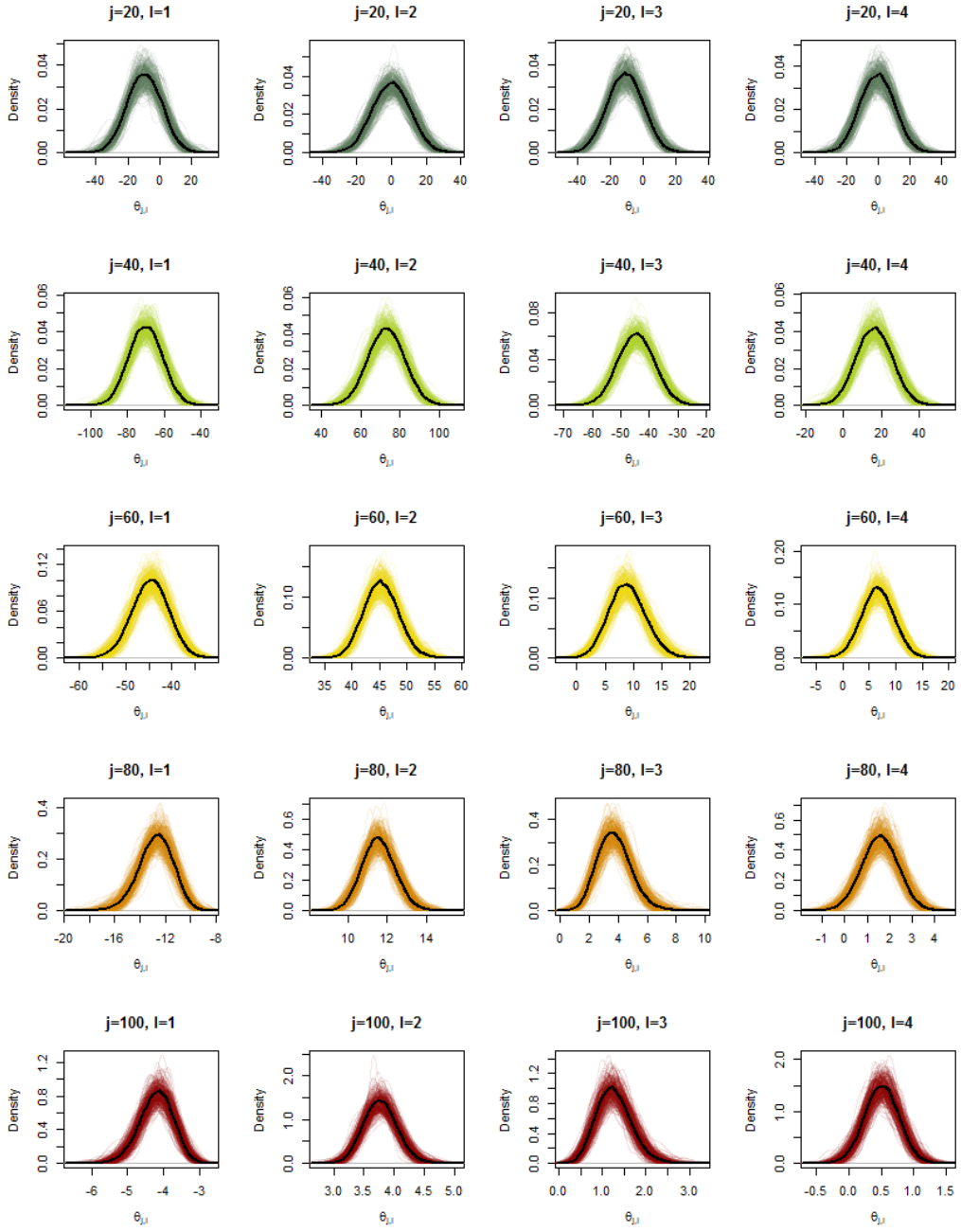

Figure D.32: PPEA Ghana, m_5:
Density plots of the $c \cdot V = 320$ chains of the parameters $\theta_{j,l}$ at some temperatures j of the PPEA. The black line indicates the density of the merged sample.

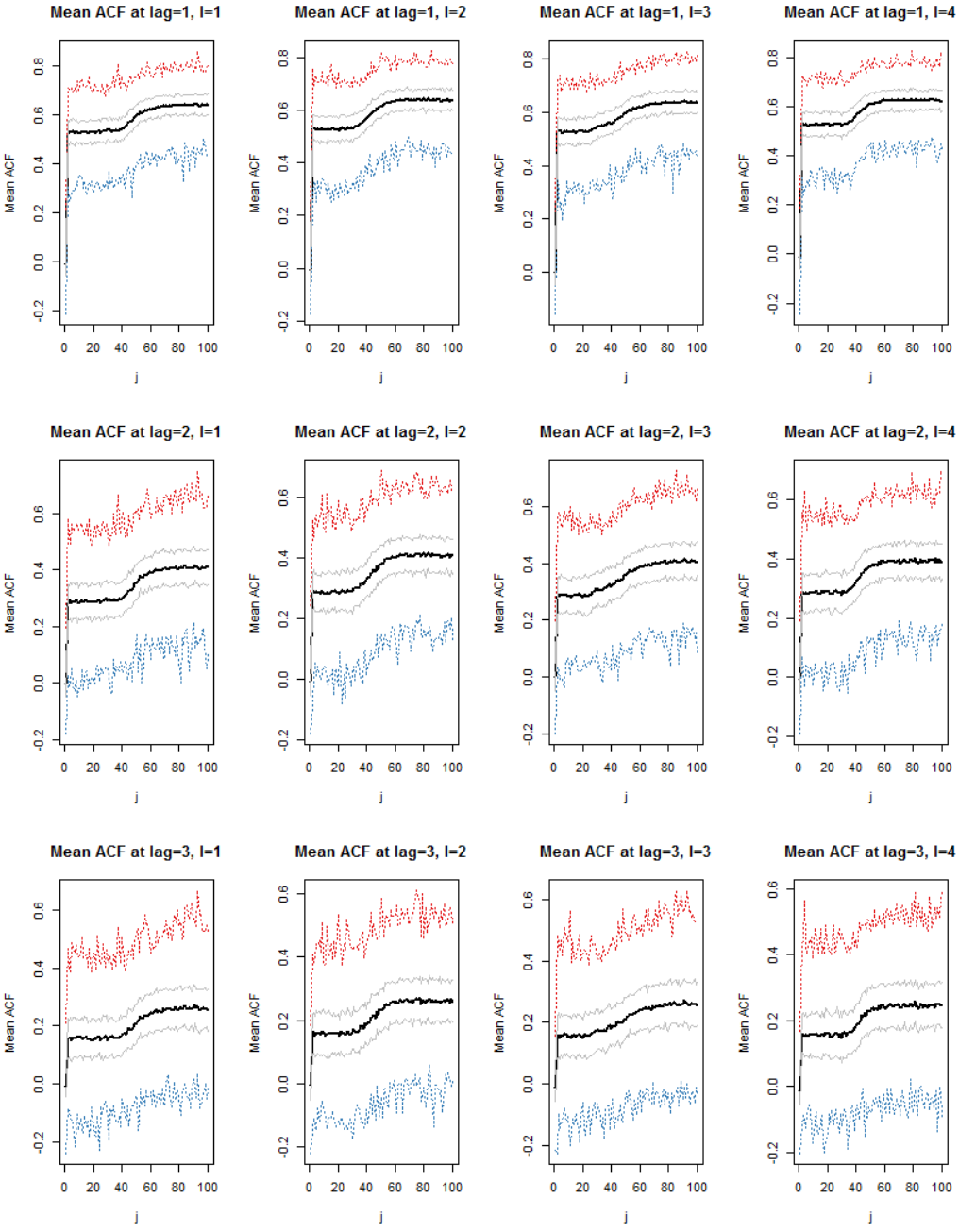

Figure D.33: PPEA Ghana, model 5:
Black line: mean value of the ACF of all $c \cdot V = 320$ PPEA chains of $\theta_{j,l}$. *Grey lines*: upper and lower quartile. *Red line*: maximum ACF. Blue line: minimum ACF.
Upper panel: lag=1. *Middle panel:* lag=2. *Lower panel:* lag=3.

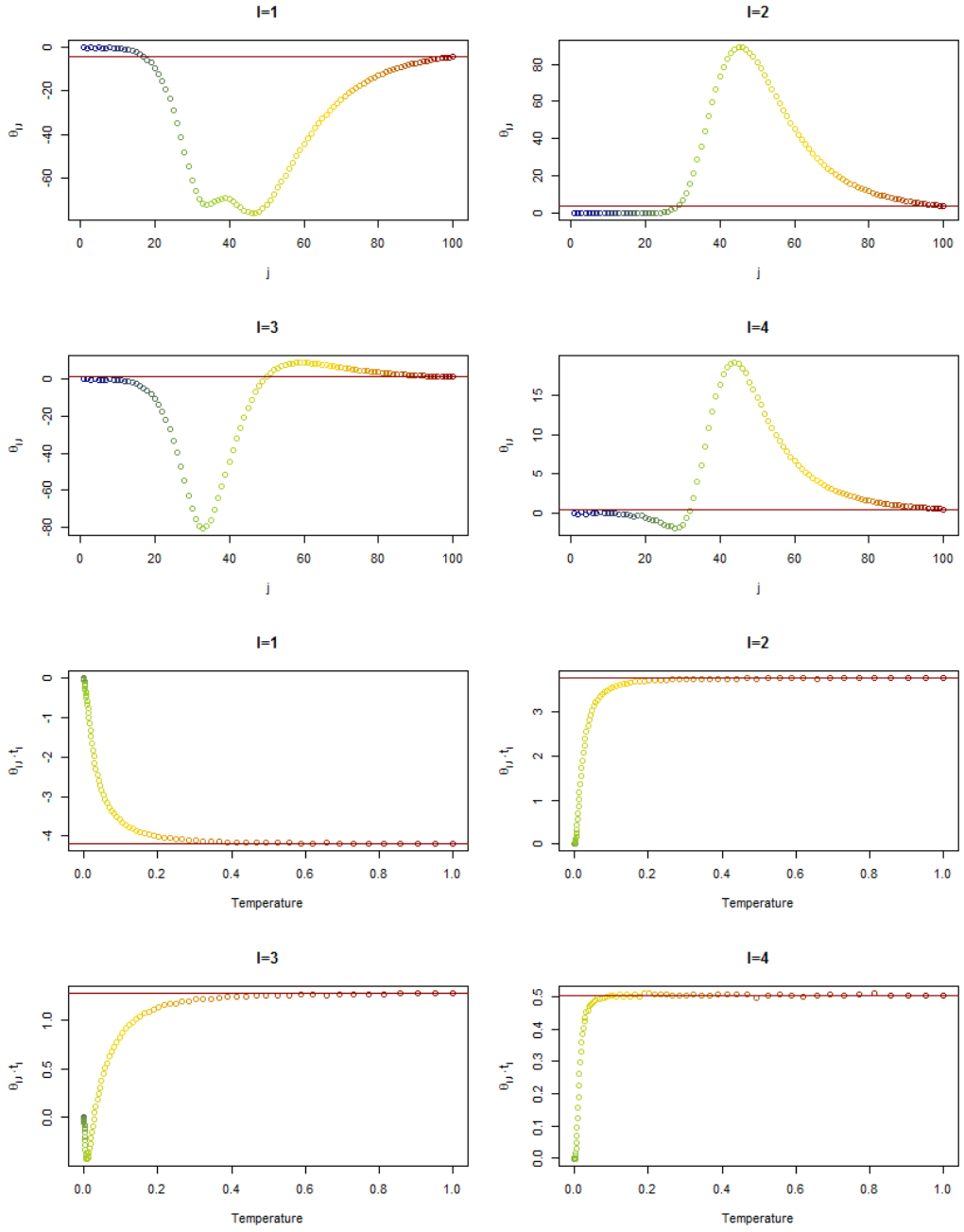

Figure D.34: PPEA Ghana, m_5:
Upper two panels: Path of the MCMC estimates of $\theta_{j,l}$.
Lower two panels: Path of the MCMC estimates of $\theta_{j,l}$ multiplied by t_j over the temperature.
The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t_{j=J} = 1$.

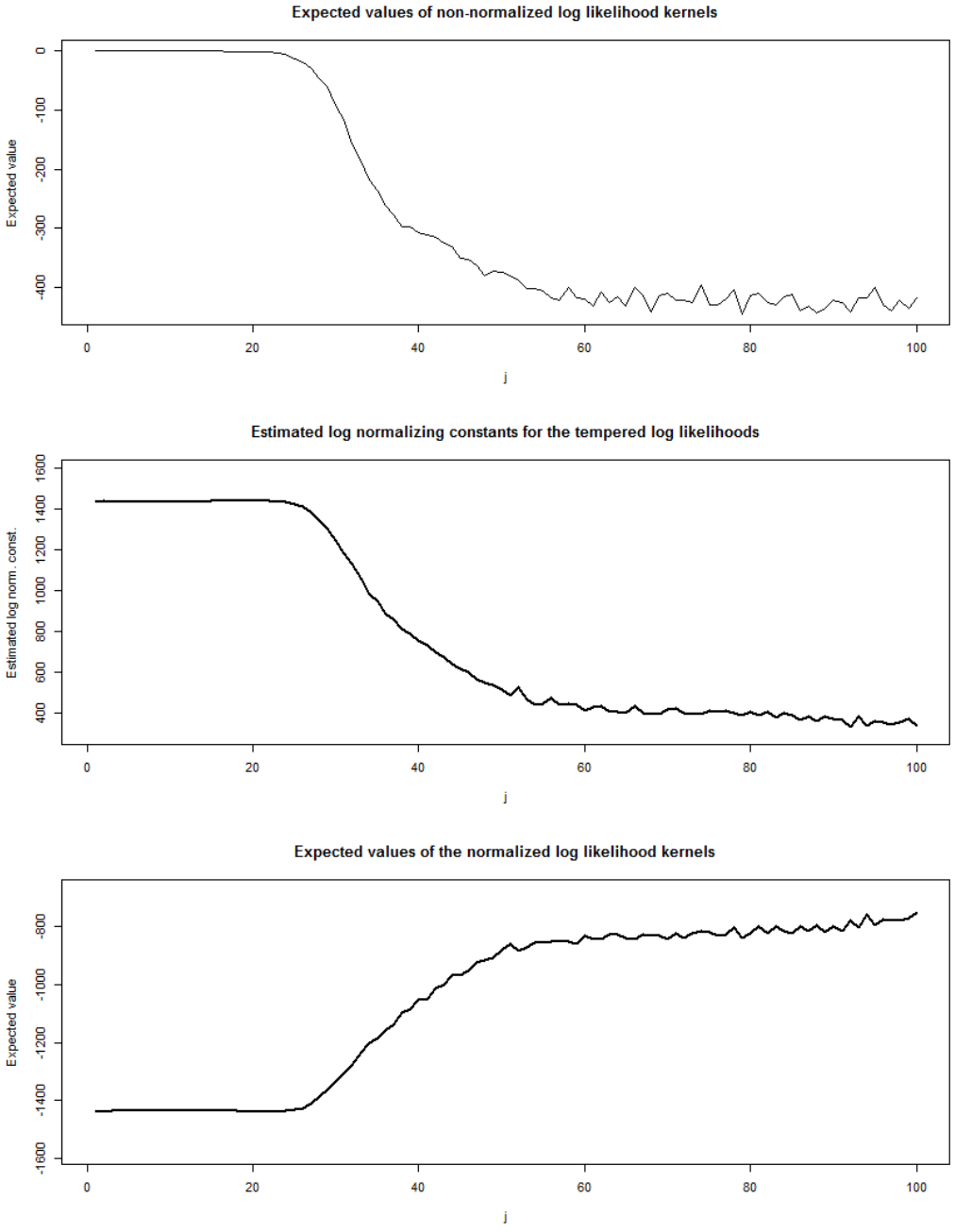

Figure D.35: PPEA Ghana, m_5:
Top panel: Expected values of the non-normalized log likelihood kernels.
Middle panel: Importance sampling estimates of the normalizing constants of the tempered log likelihoods.
Lower panel: Expected values of the normalized tempered log likelihoods.

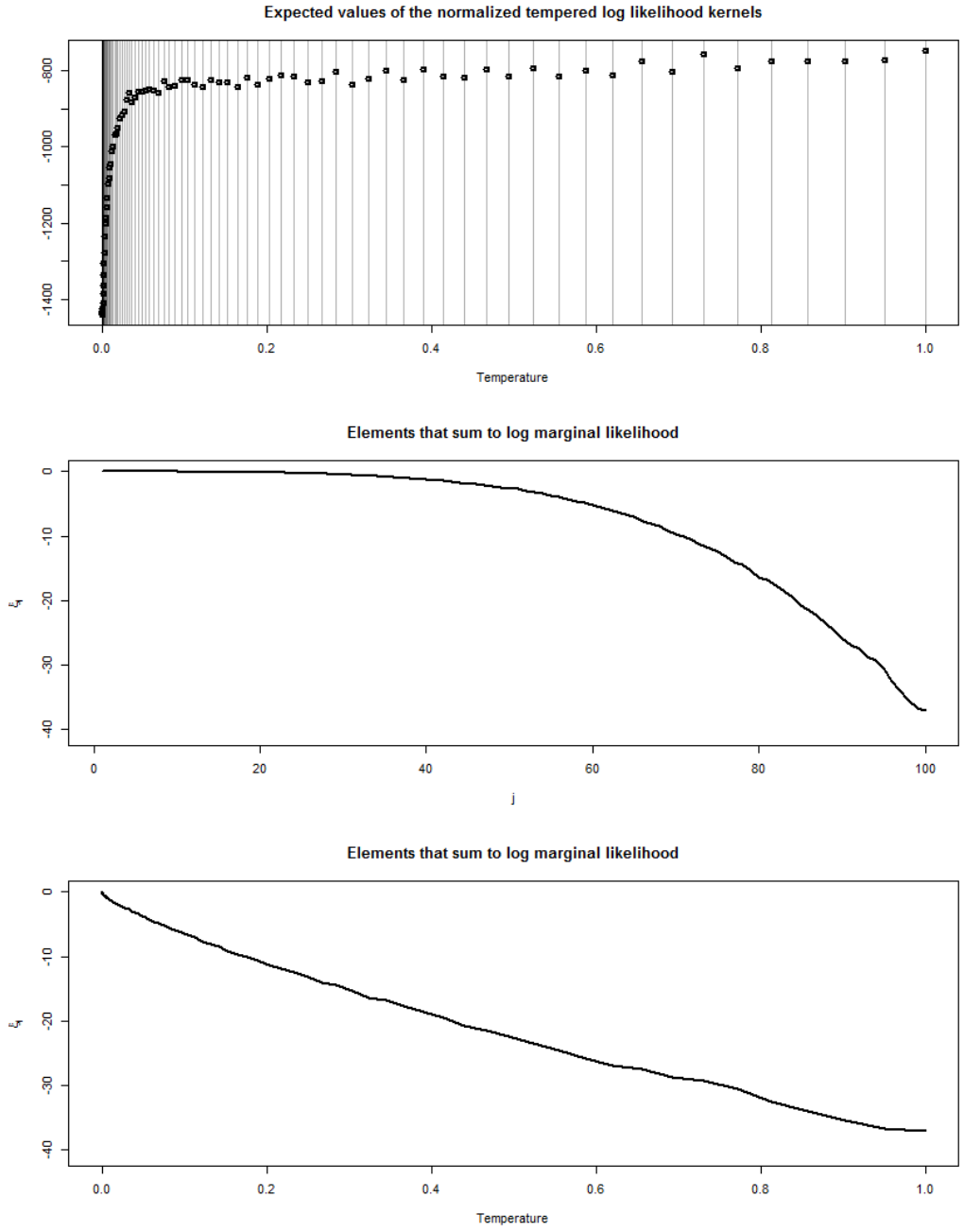

Figure D.36: PPEA Ghana, m_5:
Top panel: Expected values of the normalized log likelihood kernels in relation to the temperature steps.
Middle panel: Summands of the trapezoidal approximation of the evidence.
Lower panel: Summands of the trapezoidal approximation in relation to the temperature.

D.4.6 Model 6

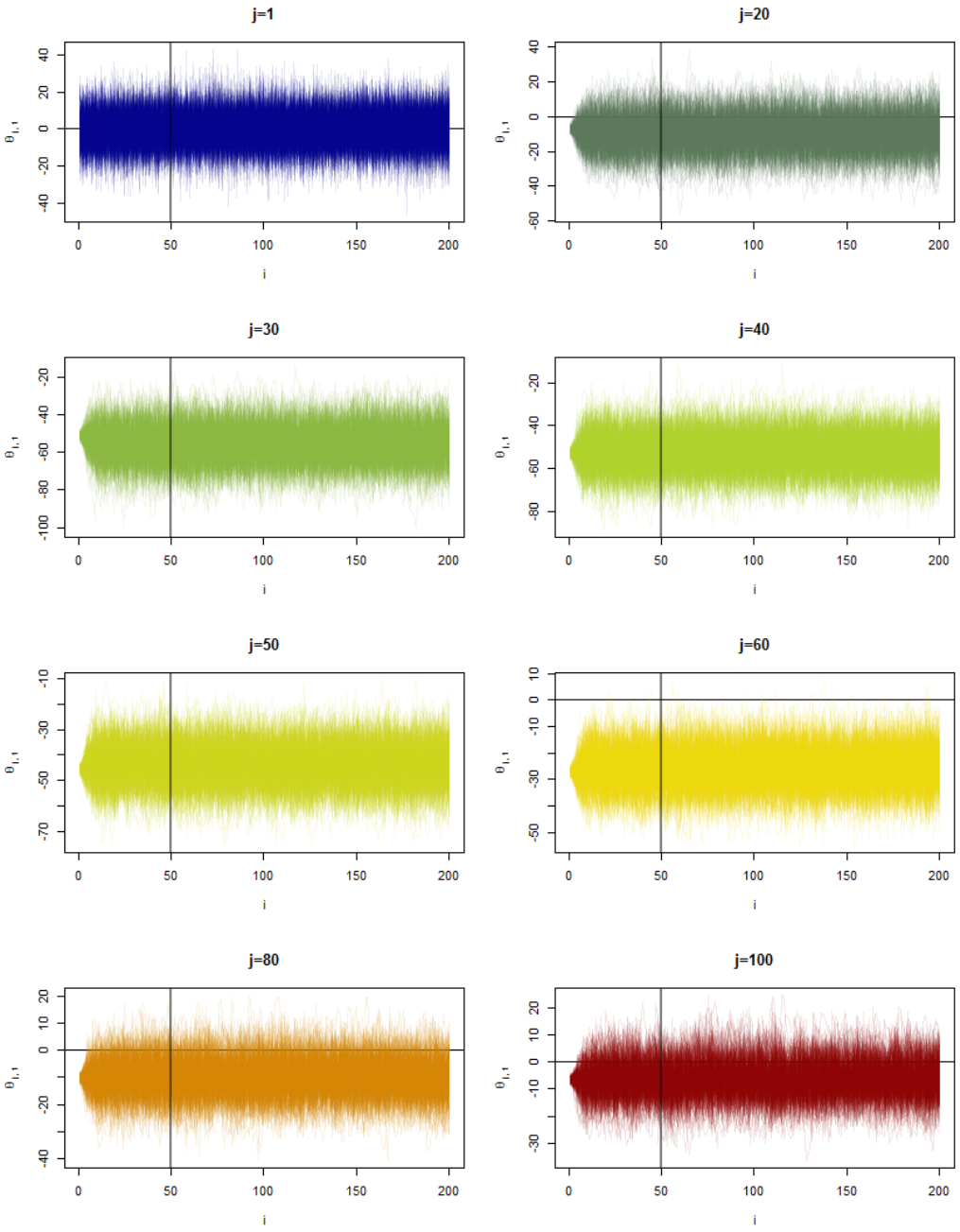

Figure D.37: PPEA Ghana, m_6:
Traceplots of the $c \cdot V = 320$ chains of the edge parameter $\theta_{j,l=1}$ at some temperatures j of the PPEA. The horizontal grey line indicates the burn-in period of $I_{burn} = 50$ iterations used.

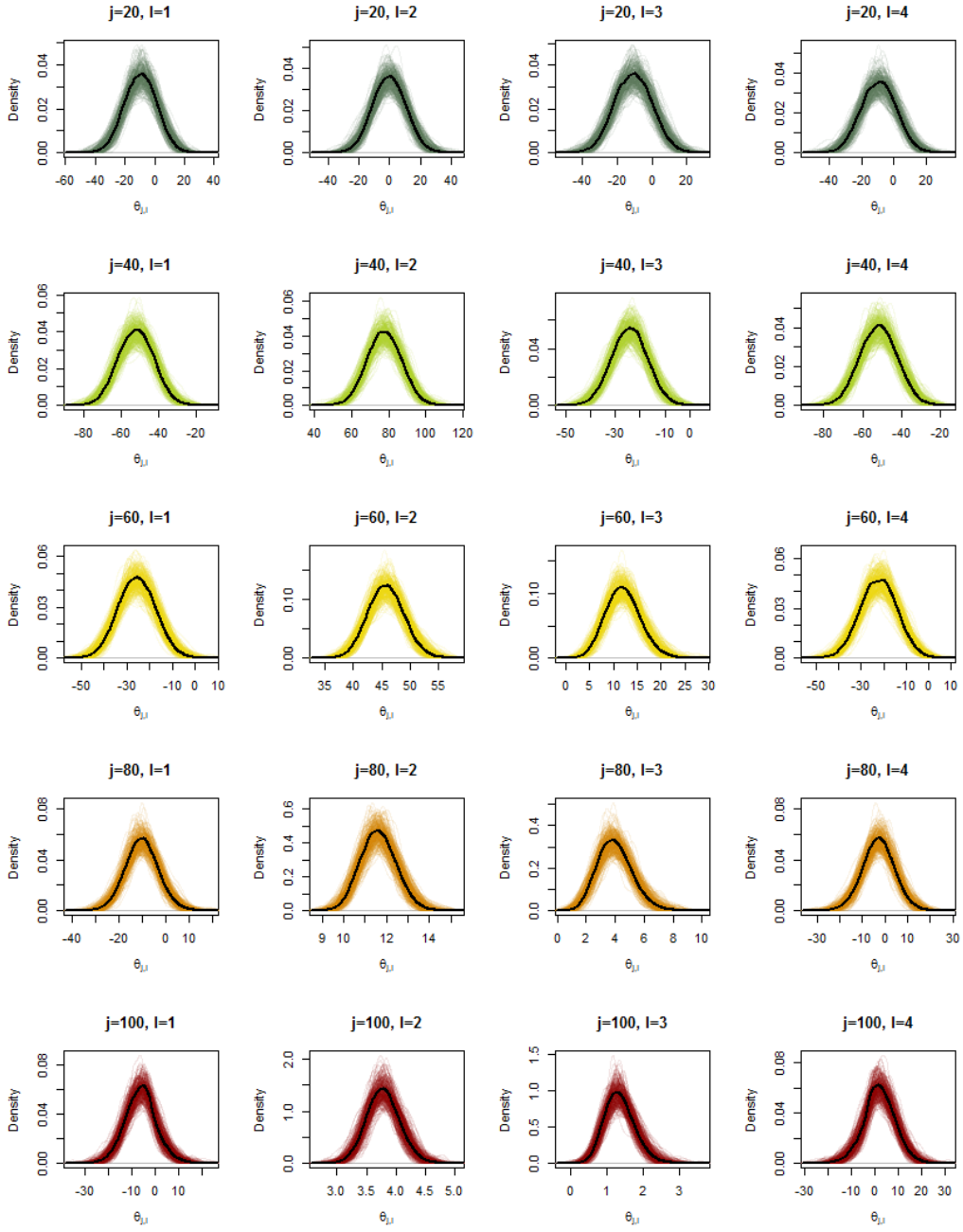

Figure D.38: PPEA Ghana, m_6:
Density plots of the $c \cdot V = 320$ chains of the parameters $\theta_{j,l}$ at some temperatures j of the PPEA. The black line indicates the density of the merged sample.

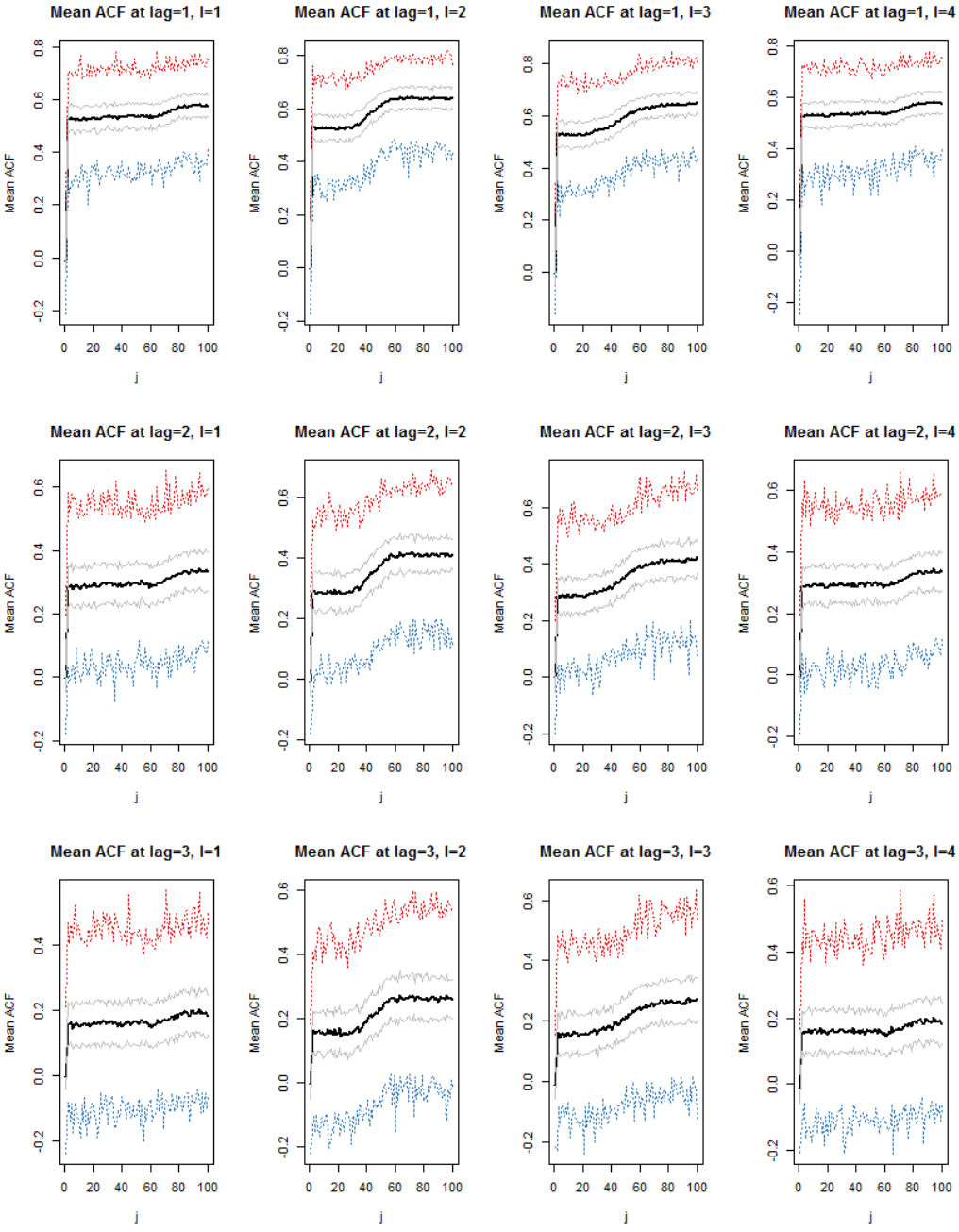

Figure D.39: PPEA Ghana, m_6:
Black line: mean value of the ACF of all $c \cdot V = 320$ PPEA chains of $\theta_{j,l}$. *Grey lines*: upper and lower quartile. *Red line*: maximum ACF. Blue line: minimum ACF.
Upper panel: lag=1. *Middle panel:* lag=2. *Lower panel:* lag=3.

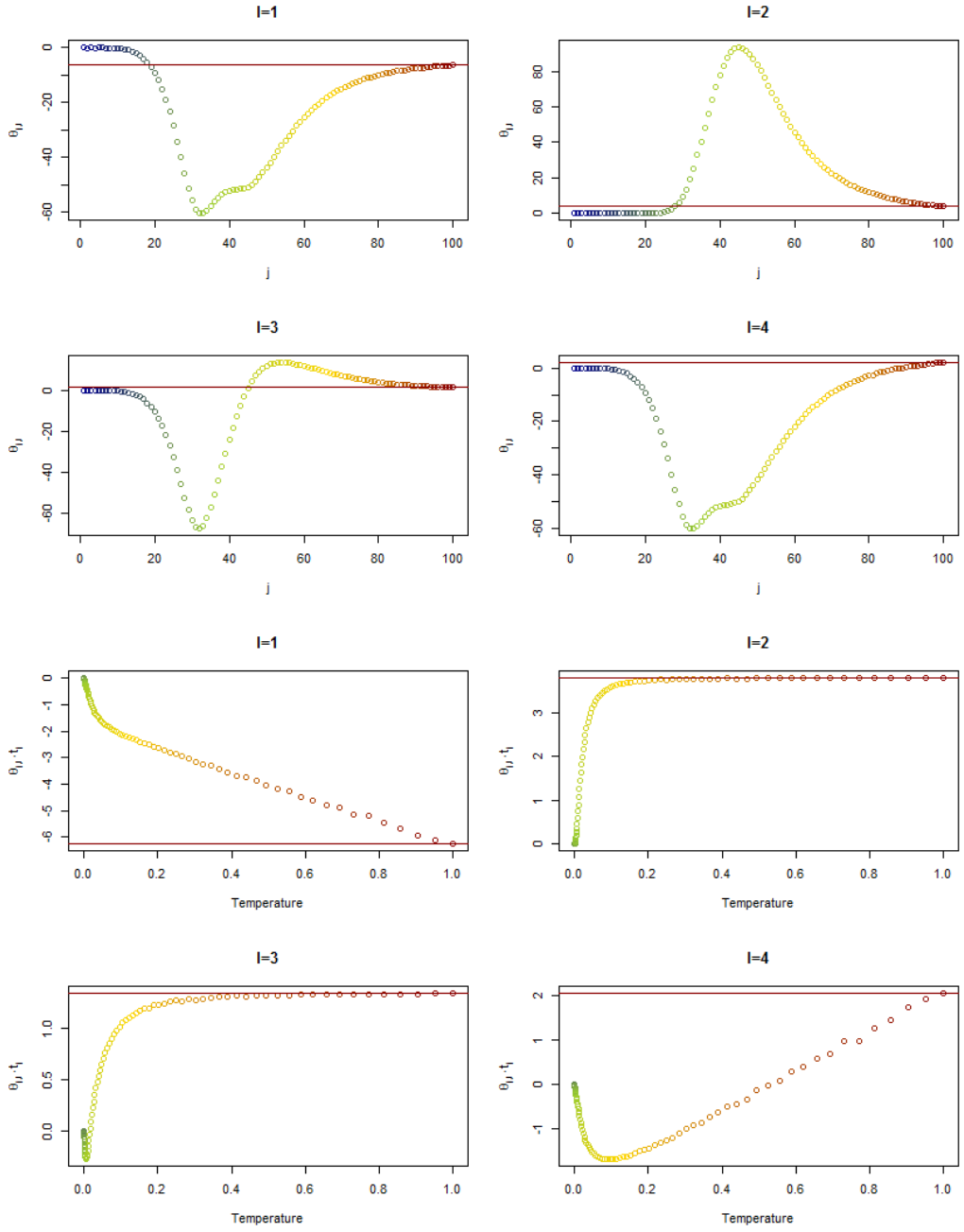

Figure D.40: PPEA Ghana, m_6:
Upper two panels: Path of the MCMC estimates of $\theta_{j,l}$.
Lower two panels: Path of the MCMC estimates of $\theta_{j,l}$ multiplied by t_j over the temperature.
The horizontal dark red line indicates the MCMC parameter estimate of the target posterior at $t_{j=J} = 1$.

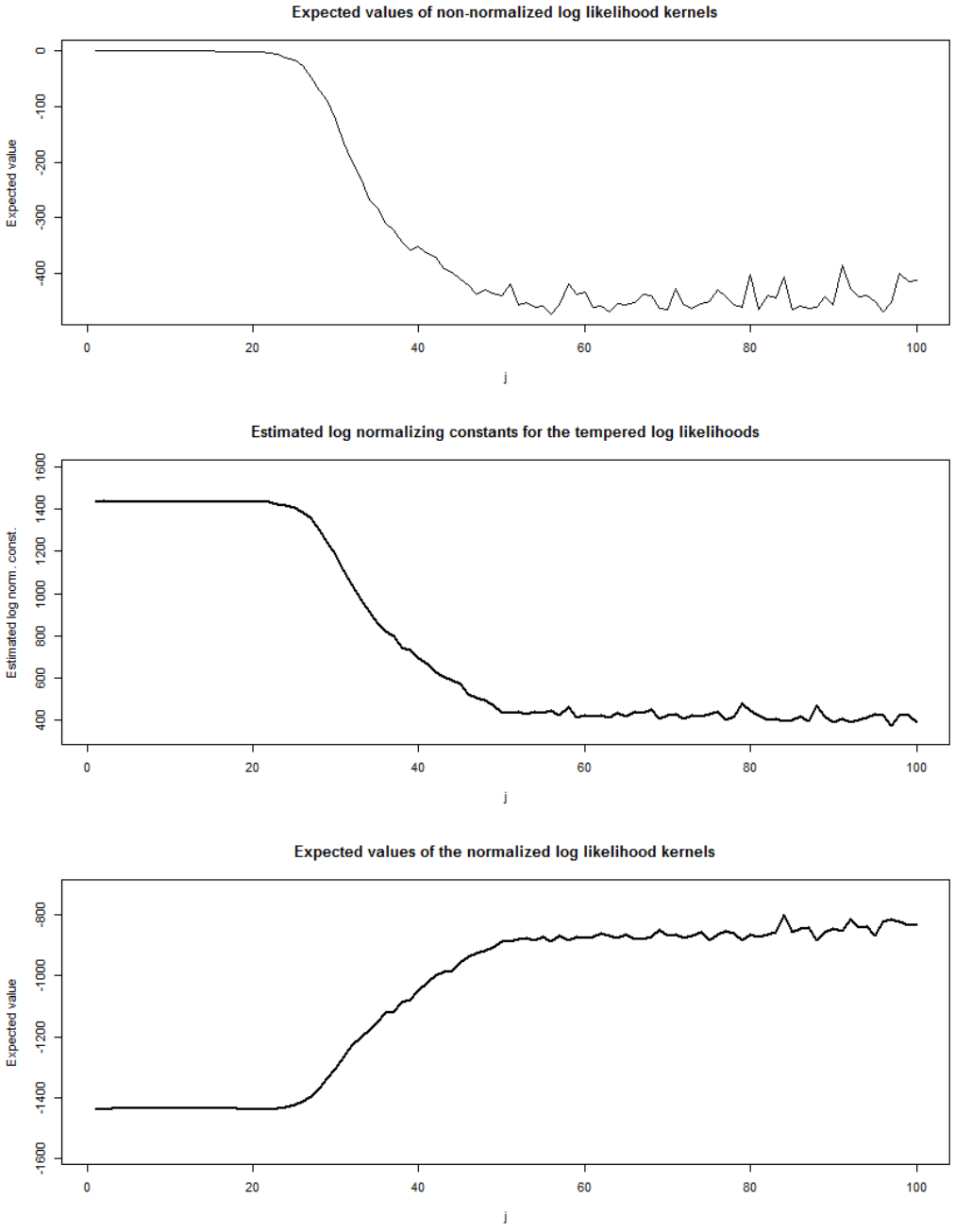

Figure D.41: PPEA Ghana, m_6:
Top panel: Expected values of the non-normalized log likelihood kernels.
Middle panel: Importance sampling estimates of the normalizing constants of the tempered log likelihoods.
Lower panel: Expected values of the normalized tempered log likelihoods.

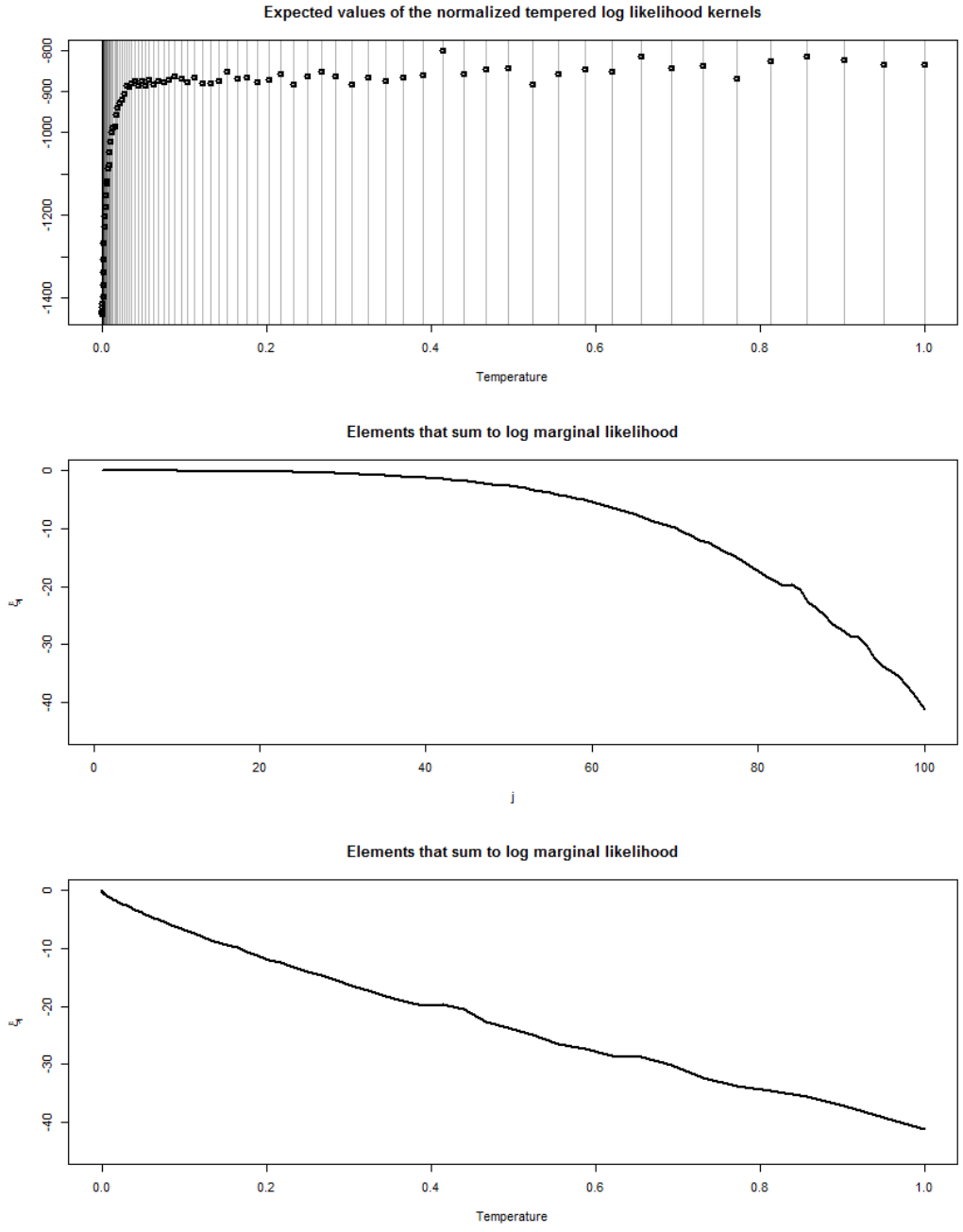

Figure D.42: PPEA Ghana, m_6:
Top panel: Expected values of the normalized log likelihood kernels in relation to the temperature steps.
Middle panel: Summands of the trapezoidal approximation of the evidence.
Lower panel: Summands of the trapezoidal approximation in relation to the temperature.

Bibliography

Adam, Silke and Hanspeter Kriesi, "The Network approach," in Paul A. Sabatier, ed., *Theories of the policy process*, Boulder: Westview Press, 2007, pp. 129–154.

Ahn, T. K., Robert Huckfeldt, Alexander K. Mayer, and John Barry Ryan, "Expertise and Bias in Political Communication Networks," *American Journal of Political Science*, 2013, *57* (2), 357–373.

Akaike, Hirotogu, "Information Theory and an Extension of the Maximum Likelihood Principle," in Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, eds., *Selected Papers of Hirotugu Akaike*, Springer series in statistics, New York, NY: Springer New York, 1998, pp. 199–213.

Alquier, Pierre, Nial Friel, Richard G. Everitt, and Aidan Boland, "Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels," *Statistics and Computing*, 2016, *26*, 29–47.

Ando, Tomohiro, *Bayesian Model Selection and Statistical Modeling* Statistics: Textbooks and Monographs, Boca Raton: CRC Press, 2010.

Austen-Smith, David, "Information and Influence: Lobbying for Agendas and Votes," *American Journal of Political Science*, 1993, *37* (3), 799–833.

Ball, Richard, "Interest groups, influence and welfare," *Economics & Politics*, 1995, *7* (2), 119–146.

Behrens, Gundula, Nial Friel, and Merrilee Hurn, "Tuning tempered transitions," *Statistics and Computing*, 2012, *22* (1), 65–78.

Berardo, Ramiro and John T. Scholz, "Self-Organizing Policy Networks: Risk, Partner Selection, and Cooperation in Estuaries," *American Journal of Political Science*, 2010, *54* (3), 632–649.

Besag, Julian, "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1974, *36* (2), 192–236.

Bratton, Michael, "Formal versus Informal Institutions in Africa," *Journal of Democracy*, 2007, *18* (3), 96–110.

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

_ **and** _ , "Bayesian model selection for exponential random graph models," *Social Networks*, 2013, *35* (1), 11–24.

_ **and** _ , "Bergm: Bayesian Exponential Random Graphs in R," *Journal of Statistical Software*, 2014, *61* (2), 1–25.

Calderhead, Ben and Mark Girolami, "Estimating Bayes factors via thermodynamic integration and population MCMC," *Computational Statistics & Data Analysis*, 2009, *53* (12), 4028–4045.

Carpenter, Daniel P., Kevin M. Esterling, and David M. J. Lazer, "Friends, brokers, and transitivity: Who informs whom in Washington politics?," *Journal of Politics*, 2004, *66* (1), 224–246.

Chen, Ming-Hui, Qi-Man Shao, and Joseph G. Ibrahim, *Monte Carlo Methods in Bayesian Computation* Springer series in statistics, New York: Springer, 2002.

Chib, Siddhartha, "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 1995, *90* (432), 1313–1321.

_ **and Edward Greenberg**, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 1995, *49* (4), 327–335.

_ **and Ivan Jeliazkov**, "Marginal Likelihood From the Metropolis–Hastings Output," *Journal of the American Statistical Association*, 2001, *96* (453), 270–281.

Daudin, J.-J., F. Picard, and S. Robin, "A mixture model for random graphs," *Statistics and Computing*, 2008, *18* (2), 173–183.

Davis, James A., "Clustering and Hierarchy in Interpersonal Relations: Testing Two Graph Theoretical Models on 742 Sociomatrices," *American Sociological Review*, 1970, *35* (5), 843–851.

Del Moral, Pierre, Arnaud Doucet, and Ajay Jasra, "Sequential Monte Carlo samplers," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006, *68* (3), 411–436.

Diciccio, Thomas J., Robert E. Kass, Adrian Raftery, and Larry Wasserman, "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 1997, *92* (439), 903–915.

Erdös, P. and A. Renyi, "On Random Graphs," *Publicationes Mathematicae*, 1959, *1* (6), 290–297.

Everitt, Richard G., "Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks," *Journal of Computational and Graphical Statistics*, 2012, *21* (4), 940–960.

_ , Adam M. Johansen, Ellen Rowing, and Melina Evdemon-Hogan, "Bayesian model comparison with un-normalised likelihoods," *Statistics and Computing*, 2016, (2), 1–20.

Festinger, Leon, "A Theory of Social Comparison Processes," *Human Relations*, 1954, *7* (2), 117–140.

Fienberg, Stephen E. and Stanley S. Wasserman, "Categorical Data Analysis of Single Sociometric Relations," *Sociological Methodology*, 1981, *12*, 156.

Fosdick, Bailey K. and Peter D. Hoff, "Testing and Modeling Dependencies Between a Network and Nodal Attributes," *Journal of the American Statistical Association*, 2015, *110* (511), 1047–1056.

Frank, Ove and David Strauss, "Markov Graphs," *Journal of the American Statistical Association*, 1986, *81* (395), 832.

Friel, Nial, "Evidence and Bayes Factor Estimation for Gibbs Random Fields," *Journal of Computational and Graphical Statistics*, 2013, *22* (3), 518–532.

_ and Anthony N. Pettitt, "Marginal Likelihood Estimation via Power Posteriors," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2008, *70* (3), 589–607.

_ and Jason Wyse, "Estimating the model evidence - a review," *Statistica Neerlandica*, 2012, *66* (3), 288–308.

_ , **Antonietta Mira, and Chris J. Oates**, "Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods," *Bayesian Analysis*, 2016, *11* (1), 215–245.

_ , **Merrilee Hurn, and Jason Wyse**, "Improving power posterior estimation of statistical evidence," *Statistics and Computing*, 2014, *24* (5), 709–723.

Fruchterman, Thomas M. J. and Edward M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and Experience*, 1991, *21* (11), 1129–1164.

Frühwirth-Schnatter, Sylvia, "Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques," *The Econometrics Journal*, 2004, *7* (1), 143–167.

Gelman, Andrew and Xiao-Li Meng, "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling," *Statistical Science*, 1998, *13* (2), 163–185.

_ , **John B. Carlin, Hal Steven Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin**, *Bayesian data analysis* Texts in Statistical Science, 3rd edition ed., Baca Raton: CRC Press, 2013.

Geweke, John, "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 1989, *57* (6), 1317–1339.

_ , "Using simulation methods for Bayesian econometric models: Inference, development,and communication," *Econometric Reviews*, 1999, *18* (1), 1–73.

Geyer, Charles J. and Elizabeth A. Thompson, "Constrained Monte Carlo Maximum Likelihood for Dependent Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1992, *54* (3), 657–699.

Gilks, W. R. and G. O. Roberts, "Convergence of Adaptive Direction Sampling," *Journal of Multivariate Analysis*, 1994, *49* (2), 287–298.

_ , _ , **and E. I. George**, "Adaptive Direction Sampling," *The Statistician*, 1994, *43* (1), 179.

Goodreau, Steven M., James A. Kitts, and Martina Morris, "Birds of a Feather, or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks," *Demography*, 2009, *46* (1), 103–125.

Green, Peter J., "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 1995, *82* (4), 711–732.

Handcock, Mark S., "Assessing degeneracy in statistical models of social networks," *University of Washington (Center for Statistics and the Social Sciences) Working Paper*, 2003, (39).

_ , "Statistical Models for Social Networks: Inference and Degeneracy," in Kathleen M. Carley, Philippa E. Pattison, and Ronald L. Breiger, eds., *Dynamic Social Network Modeling and Analysis*, Washington, D.C.: National Academies Press, 2003.

_ **, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris**, "statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data," *Journal of Statistical Software*, 2008, *24* (1), 1548–7660.

Hankin, R. K. S., "Very large numbers in R: Introducing package Brobdingnag," *R News*, 2007, *7* (3).

Hanneke, Steve, Wenjie Fu, and Eric P. Xing, "Discrete temporal models of social networks," *Electronic Journal of Statistics*, 2010, *4* (0), 585–605.

Hastie, David I. and Peter J. Green, "Model choice using reversible jump Markov chain Monte Carlo," *Statistica Neerlandica*, 2012, *66* (3), 309–338.

Hastings, W. K., "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 1970, *57* (1), 97.

Henry, A. D., M. Lubell, and M. McCoy, "Belief Systems and Social Capital as Drivers of Policy Network Structure: The Case of California Regional Planning," *Journal of Public Administration Research and Theory*, 2011, *21* (3), 419–444.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky, "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 1999, *14* (4), 382–401.

Hoff, Peter D., "Bilinear Mixed-Effects Models for Dyadic Data," *Journal of the American Statistical Association*, 2005, *100* (469), 286–295.

_ , "Multiplicative latent factor models for description and prediction of social networks," *Computational and Mathematical Organization Theory*, 2009, *15* (4), 261–272.

_ , "Separable covariance arrays via the Tucker product, with applications to multivariate relational data," *Bayesian Analysis*, 2011, *6* (2), 179–196.

_ , "Multilinear tensor regression for longitudinal relational data," *The Annals of Applied Statistics*, 2015, *9* (3), 1169–1193.

Holland, Paul W. and Samuel Leinhardt, "Transitivity in Structural Models of Small Groups," *Comparative Group Studies*, 1971, *2* (2), 107–124.

_ **and** _ , "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association*, 1981, *76* (373), 33.

Huckfeldt, R. Robert and John D. Sprague, *Citizens, politics, and social communication: Information and influence in an election campaign* Cambridge Studies in Political Psychology and Public Opinion, Cambridge: Cambridge University Press, 1995.

Hug, Sabine, Michael Schwarzfischer, Jan Hasenauer, Carsten Marr, and Fabian J. Theis, "An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using Simpson's rule," *Statistics and Computing*, 2016, *26* (3), 663–677.

Hunter, David R., "Curved Exponential Family Models for Social Networks," *Social Networks*, 2007, *29* (2), 216–230.

_ **and Mark S. Handcock**, "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics*, 2006, *15* (3), 565–583.

_ , _ , **Carter T. Butts, Steven M. Goodreau, and Martina Morris**, "ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks," *Journal of statistical software*, 2008, *24* (3), 1–29.

_ , **Steven M. Goodreau, and Mark S. Handcock**, "Goodness of Fit of Social Network Models," *Journal of the American Statistical Association*, 2008, *103* (481), 248–258.

_ , _ , **and** _ , "ergm.userterms: A Template Package for Extending statnet," *Journal of statistical software*, 2013, *52* (2), 1–25.

Jackson, Matthew O., *Social and economic networks*, Princeton, N.J. and Woodstock: Princeton University Press, 2008.

Janning, Frank, Philip Leifeld, Thomas Malang, and Volker Schneider, "Diskursnetzwerkanalyse. Überlegungen zur Theoriebildung und Methodik," in Volker Schneider and Frank Janning, eds., *Politiknetzwerke*, Lehrbuch, Wiesbaden: VS Verlag für Sozialwissenschaften, 2009, pp. 59–92.

Kass, Robert E. and Adrian E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, 1995, *90* (430), 773–795.

Knoke, David, *Comparing policy networks: Labor politics in the U.S., Germany, and Japan* Cambridge studies in comparative politics, digital printing ed., Cambridge: Cambridge University Press, 1996.

Kolaczyk, Eric D., *Statistical Analysis of Network Data: Methods and Models* Springer Series in Statistics, New York: Springer, 2009.

Koskinen, Johan and Galina Daraganova, "Dependence Graphs and Sufficient Statistics," in Dean Lusher, Johan Koskinen, Garry Robins, Dean Lusher, Johan Koskinen, and Garry Robins, eds., *Exponential Random Graph Models for Social Networks*, Cambridge: Cambridge University Press, 2012, pp. 77–90.

Koskinen, Johan H., "The Linked Importance Sampler Auxiliary Variable Metropolis Hastings Algorithm for Distributions with Intractable Normalising Constants," *University of Melbourne (Department of Psychology, School of Behavioural Science) Working Paper*, 2008.

_ , Garry L. Robins, and Philippa E. Pattison, "Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation," *Statistical Methodology*, 2010, *7* (3), 366–384.

_ , _ , Peng Wang, and Philippa E. Pattison, "Bayesian analysis for partially observed network data, missing ties, attributes and actors," *Social Networks*, 2013, *35* (4), 514–527.

Krackhardt, David, "Cognitive social structures," *Social Networks*, 1987, *9* (2), 109–134.

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

Kullback, Solomon, *Information theory and statistics* Dover books on mathematics, 2nd ed., corrected and rev ed., Mineola, N.Y.: Dover Publications, 1968.

Lartillot, Nicolas and Herve Philippe, "Computing Bayes factors using thermodynamic integration," *Systematic biology*, 2006, *55* (2), 195–207.

Lee, Youngmi, In Won Lee, and Richard C. Feiock, "Interorganizational Collaboration Networks in Economic Development Policy: An Exponential Random Graph Model Analysis," *Policy Studies Journal*, 2012, *40* (3), 547–573.

Lefebvre, Geneviève, Russell Steele, Alain C. Vandal, Sridar Narayanan, and Douglas L. Arnold, "Path Sampling to Compute Integrated Likelihoods: An Adaptive Approach," *Journal of Computational and Graphical Statistics*, 2009, *18* (2), 415–437.

Lehmann, Erich L. and George Casella, *Theory of point estimation* Springer texts in statistics, 2. ed. ed., New York: Springer, 1998.

Leifeld, Philip and Volker Schneider, "Information exchange in policy networks," *American Journal of Political Science*, 2012, *56* (3), 731–744.

Liang, Faming, Ick H. Jin, Qifan Song, and Jun S. Liu, "An Adaptive Exchange Algorithm for Sampling From Distributions With Intractable Normalizing Constants," *Journal of the American Statistical Association*, 2016, *111* (513), 377–393.

Lijphart, Arend, *Patterns of democracy: Government forms and performance in thirty-six countries*, New Haven: Yale University Press, 1999.

Lohmann, Susanne, "A Signaling Model of Informative and Manipulative Political Action," *American Political Science Review*, 1993, *87* (2), 319–333.

Lubell, Mark, Adam Douglas Henry, and Mike McCoy, "Collaborative Institutions in an Ecology of Games," *American Journal of Political Science*, 2010, *54* (2), 287–300.

Lusher, Dean, Johan Koskinen, Garry Robins, Dean Lusher, Johan Koskinen, and Garry Robins, eds, *Exponential Random Graph Models for Social Networks*, Cambridge: Cambridge University Press, 2012.

Madigan, David and Adrian E. Raftery, "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 1994, *89* (428), 1535–1546.

Marin, Jean-Michel, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder, "Approximate Bayesian computational methods," *Statistics and Computing*, 2012, *22* (6), 1167–1180.

McCulloch, Robert and Peter E. Rossi, "A Bayesian approach to testing the arbitrage pricing theory," *Journal of Econometrics*, 1991, *49* (1-2), 141–168.

Meng, Xiao-Li and Wing Hung Wong, "Simulating ratios of normalizing constants via a simple identity: A theoretical exploration," *Statistica Sinica*, 1996, *6* (4), 831–860.

Møller, Jesper, Anthony N. Pettitt, Robert Reeves, and Kasper K. Berthelsen, "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants," *Biometrika*, 2006, *93* (2), 451–458.

Morris, Martina, Mark S. Handcock, and David R. Hunter, "Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects," *Journal of statistical software*, 2008, *24* (4), 1548–7660.

Murray, Iain, "Advances in Markov chain Monte Carlo methods." Phd thesis, University College London 2007.

_ , Zoubin Ghahramani, and David J.C. MacKay, "MCMC for doubly-intractable distributions," in "Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)" AUAI Press 2006, pp. 359–366.

Neal, Radford M., "Annealed importance sampling," *Statistics and Computing*, 2001, *11* (2), 125–139.

Newton, Michael A. and Adrian E. Raftery, "Approximate Bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1994, pp. 3–48.

Nowicki, Krzysztof and Tom A. B. Snijders, "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 2001, *96* (455), 1077–1087.

Oates, Chris J., Mark Girolami, and Nicolas Chopin, "Control functionals for Monte Carlo integration," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016, *97* (3), 695–718.

Ogata, Yosihiko, "A Monte Carlo method for high dimensional integration," *Numerische Mathematik*, 1989, *55* (2), 137–157.

Pappi, Franz Urban, Thomas König, and David Knoke, *Entscheidungsprozesse in der Arbeits- und Sozialpolitik: Der Zugang der Interessengruppen zum Regierungssystem über Politikfeldnetze: ein deutsch-amerikanischer Vergleich*, Frankfurt: Campus, 1995.

Pattison, Philippa and Garry Robins, "Neighborhood-Based Models for Social Networks," *Sociological Methodology*, 2002, *32* (1), 301–337.

_ **and Stanley Wasserman**, "Logit models and logistic regressions for social networks: II. Multivariate relations," *British Journal of Mathematical and Statistical Psychology*, 1999, *52* (2), 169–193.

R Development Core Team, *R: A Language and Environment for Statistical Computing*, Vienna: the R Foundation for Statistical Computing, 2011.

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting, "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 1997, *92* (437), 179–191.

Randall, Ian, "Guidelines for Non State Actor Participation in CAADP Processes," *NEPAD Working Paper*, 2011.

Rinaldo, Alessandro, Stephen E. Fienberg, and Yi Zhou, "On the geometry of discrete exponential families with application to exponential random graph models," *Electronic Journal of Statistics*, 2009, *3*, 446–484.

Ripley, Brian D., *Statistical inference for spatial processes: An essay awarded the Adams Prize*, Cambridge: Cambridge University Press, 1991.

Robins, Garry and Philippa Pattison, "Interdependencies and social processes: Dependence graphs and generalized dependence structures," in Peter J. Carrington, John Scott, and Stanley Wasserman, eds., *Models and methods in social network analysis*, Vol. 28 of *Structural analysis in the social sciences*, Cambridge: Cambridge University Press, 2005, pp. 192–212.

_ , **Jenny M. Lewis, and Peng Wang**, "Statistical network analysis for analyzing policy networks," *Policy Studies Journal*, 2012, *40* (3), 375–401.

Rubin, Donald B., "Inference and missing data," *Biometrika*, 1976, *63* (3), 581–592.

Sabatier, Paul A., "Knowledge, Policy-Oriented Learning, and Policy Change: An Advocacy Coalition Framework," *Science Communication*, 1987, *8* (4), 649–692.

_ **and Christopher M. Weible**, "The Advocacy Coalition Framework," in Paul A. Sabatier, ed., *Theories of the policy process*, Boulder: Westview Press, 2007, pp. 649–692.

Schwarz, Gideon, "Estimating the Dimension of a Model," *The Annals of Statistics*, 1978, *6* (2), 461–464.

Schweinberger, Michael, "Instability, Sensitivity, and Degeneracy of Discrete Exponential Families," *Journal of the American Statistical Association*, 2011, *106* (496), 1361–1370.

Shepsle, Kenneth A. and Barry R. Weingast, "The Institutional Foundations of Committee Power," *American Political Science Review*, 1987, *81* (1), 85.

Snijders, Tom A. B., "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure*, 2002, *3* (2).

_ , "Statistical Models for Social Networks," *Annual Review of Sociology*, 2011, *37* (1), 131–153.

_ **, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock**, "New Specifications for Exponential Random Graph Models," *Sociological Methodology*, 2006, *36* (1), 99–153.

Snijders, Tom and Marijtje van Duijn, "Simulation for Statistical Inference in Dynamic Network Models," in G. Fandel, W. Trockel, Rosaria Conte, Rainer Hegselmann, and Pietro Terna, eds., *Simulating social phenomena*, Vol. 456 of *Lecture Notes in Economics and Mathematical Systems*, Berlin: Springer, 1997, pp. 493–512.

Strauss, David and Michael Ikeda, "Pseudolikelihood Estimation for Social Networks," *Journal of the American Statistical Association*, 1990, *85* (409), 204.

Tavaré, Simon, David J. Balding, R. C. Griffiths, and Peter Donnelly, "Inferring Coalescence Times From DNA Sequence Data," *Genetics*, 1997, *145* (2), 505–518.

Ter Braak, Cajo J. F., "A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: Easy Bayesian computing for real parameter spaces," *Statistics and Computing*, 2006, *16* (3), 239–249.

_ **and Jasper A. Vrugt**, "Differential Evolution Markov Chain with snooker updater and fewer chains," *Statistics and Computing*, 2008, *18* (4), 435–446.

Thiemichen, Stephanie, Nial Friel, Alberto Caimo, and Göran Kauermann, "Bayesian exponential random graph models with nodal random effects," *Social Networks*, 2016, *46*, 11–28.

Tierney, Luke and Joseph B. Kadane, "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 1986, *81* (393), 82–86.

van der Walle, Nicolas, "Presidentialism and Clientelism in Africa's Emerging Party Systems," *Journal of Modern African Studies*, 2003, *41* (2), 297–321.

Wang, Cheng, Carter T. Butts, John R. Hipp, Rupa Jose, and Cynthia M. Lakon, "Multiple Imputation for Missing Edge Data: A Predictive Evaluation Method with Application to Add Health," *Social Networks*, 2016, *45*, 89–98.

Wang, Peng, Ken Sharpe, Garry L. Robins, and Philippa E. Pattison, "Exponential random graph () models for affiliation networks," *Social Networks*, 2009, *31* (1), 12–25.

Wasserman, Stanley and Katherine Faust, *Social network analysis: Methods and applications*, 19. printing ed., Vol. 8 of *Structural analysis in the social sciences*, Cambridge: Cambridge Univ. Press, 1994.

_ **and Philippa E. Pattison**, "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*," *Psychometrika*, 1996, *61* (3), 401–425.

Weible, Christopher M. and Paul A. Sabatier, "Comparing Policy Networks: Marine Protected Areas in California," *Policy Studies Journal*, 2005, *33* (2), 181–201.

_ , **Andrew Pattison, and Paul A. Sabatier**, "Harnessing expert-based information for learning and the sustainable management of complex socio-ecological systems," *Environmental Science & Policy*, 2010, *13* (6), 522–534.

Xie, Wangang, Paul O. Lewis, Yu Fan, Lynn Kuo, and Ming-Hui Chen, "Improving marginal likelihood estimation for Bayesian phylogenetic model selection," *Systematic biology*, 2011, *60* (2), 150–160.