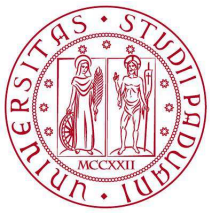

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIII

Topics in Statistical Models for Network Analysis

Direttore della Scuola: Ch.mo Prof. Alessandra Salvan

Supervisore: Ch.mo Prof. Ruggero Bellio

Co-supervisore: Ch.mo Prof. Susanna Zaccarin

Dottorando: Nicola Soriani

January 31, 2012

Acknowledgements

First, I would like to thank my parents, and dedicate this thesis to them, who during all these years in Padua have never failed to support me. They are the most important persons in my life, and I love them because their example helped me to become the person that I am.

A special thank goes to my supervisor Professor Ruggero Bellio and my co-supervisor Professor Susanna Zaccarin for their constant help in terms of presence, availability and experience in the development of this thesis. Of course also for their infinite patience in correcting my English prose.

I would also like to thank Professor Mark Handcock and the department of Statistics of the University of Washington, for their support during my visiting period in Seattle.

To Marina, Amedeo, Matteo and all the Italian people that I met in Seattle because with their friendships I learned that every place can be my home. They will have forever a special place in my heart.

To my flatmates for sharing with me the past 6 years and for introducing me to "Lost and Delirious".

Lastly, I would like to thank the comrades of the Department of Statistics of Padua and the XXIII cycle people, especially to my "colleagues" Michele and Slavica, for the time, the advices and laughter shared during the many coffee breaks and in room 122.

Abstract

Network Analysis is a set of statistical and mathematical techniques for the study of relational data arising from a system of connected entities. Most of the results for network data have been obtained in the field of Social Network Analysis (SNA), which mainly focuses on the relationships among a set of individual actors and organizations. The thesis considers some topics in statistical models for network data, with focus in particular on models used in SNA. The core of the thesis is represented by Chapters 3, 4 and 5. In Chapter 3, an alternative approach to estimate the Exponential Random Graph Models (ERGMs) is discussed. In Chapter 4, a comparison between ERGMs and Latent Space models in terms of goodness of fit is considered. In Chapter 5, alternative methods to estimate the p_2 class of models are proposed.

Sommario

La Network Analysis è un insieme di tecniche statistiche e matematiche per lo studio di dati relazionali per un sistema di entità interconnesse. Molti dei risultati per i dati di rete provengono dalla Social Network Analysis (SNA), incentrata principalmente sullo studio delle relazioni tra un insieme di individui e organizzazioni. La tesi tratta alcuni argomenti riguardanti la modellazione statistica per dati di rete, con particolare attenzione ai modelli utilizzati in SNA. Il nucleo centrale della tesi è rappresentato dai Capitoli 3, 4 e 5. Nel Capitolo 3, viene proposto un approccio alternativo per la stima dei modelli esponenziali per grafi casuali (Exponential Random Graph Models - ERGMs). Nel capitolo 4, l'approccio di modellazione ERGM e quello a Spazio Latente vengono confrontati in termini di bontà di adattamento. Nel capitolo 5, vengono proposti metodi alternativi per la stima della classe di modelli p_2.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Network Analysis is a set of statistical and mathematical techniques for the study of relational data arising from a system of connected entities.

The relationships may be any kind of irreducible property between two or more entities: economic, political, interactional or affective, to name just a few.

The main parts of the results for networks data have been obtained in the field of Social Network Analysis (SNA), that focus on implementation of theoretical and applicative methodologies to summarize, describe, visualize and analyze the social structures deriving mainly from human groups, communities or organizations.

In developing methods for the analysis of relational data, some good reasons motivate the statistical modeling of an observed social network. Social behavior is complex and it seems reasonable to suppose that social processes giving rise to network ties are stochastic, so statistical models are appropriate to understand if, and how, certain network characteristics are more or less observed in the network than expected by chance. In general, models, and especially statistical models that are estimable from data and explicitly recognize uncertainty, may help to understand the range of possible outcomes for processes on networks. Moreover, network models are especially useful to deal with data dependence induced by social relations. The leading assumption in modeling is that the observed network is generated by some (unknown) stochastic processes.

This thesis considers some topics in statistical models for network data, with focus in particular on models developed in SNA (Wasserman and Faust, 1994; Kolaczyk, 2009). The plan of the thesis is as follows.

The second chapter of the thesis presents a review of the basic concepts and the classes of statistical models for social network analysis. The approaches presented belong essentially to the classes of Exponential Random

Graph Models (ERGMs) (Wasserman and Robins, 2005), and of random effects models for graphs, distinguishing between Latent Space Models (LSMs) (Hoff et al., 2002), and p_2 models (Van Duijn et al., 2004). A review of the state of the art of these approaches is provided, including their theoretical bases and main features, estimation issues and possible variations of their basic formulation as proposed in Snijders et al. (2006); Handcock et al. (2007); Hunter (2007); Zijlstra et al. (2009); Krivitsky et al. (2009); Snijders (2011).

In the third chapter an alternative estimation approach for ERGMs is proposed. The fourth chapter focuses on the comparison between ERGMs and LSMs based on their performances in terms of goodness of fit. In the last chapter of the thesis we consider the existing methods to estimate the p_2 class of models, and we propose some alternative procedures.

1.2 Main Contributions of the Thesis

The core of the thesis is represented by Chapters 3, 4 and 5.

In Chapter 3, alternative approaches to estimate ERGMs are discussed. The existing procedures, based on simulated maximum likelihood methods, are computationally challenging, due to numerical difficulties to approximate the likelihood function. So they at times fail to converge as the likelihood approximation may degrade, especially for certain choices of the sufficient statistics of interest. A Monte Carlo Quasi-Newton algorithm for computing the maximum likelihood estimate for ERGM is introduced, borrowing some ideas from the method of maximization by parts. Two crucial aspects of the proposed method are the steplength determination based on a simulated likelihood function, and a suitable backtracking mechanism to deal with model near degeneracy. Comparisons with the existing methods (Hummel et al., 2011) are provided, both on real data from the literature and on simulations studies. The results show an improvement in terms of robustness of our algorithm with respect to the existing procedures included in the `R` package `ergm`. Our method, in fact, permits to obtain maximum likelihood estimate (MLE) also for Markov random graph models (Frank and Strauss, 1986), or when the models present near-degeneracy and instability problems. Furthermore, the proposed method is capable of estimating networks with missing data in an efficient and accurate way, with a clear improvement over the existing methods. Our procedure is included in a new `R` package `ergmQN`.

In Chapter 4, a comparison between ERGMs and Latent Space models in terms of goodness of fit is considered. In particular, the performance of the two model approaches to reproduce the dependence structure of the observed network (Hunter et al., 2008a), and their predictive power are seen as complementary features for describing their goodness of fit. In order to better evaluate the predictive power, different missing data cases are

considered. For both the approaches, the predictive procedures are built by combining routines implemented specifically for this goal, included in `ergmQN` package, with routines already implemented in the `R` library `statnet`. The comparisons are made both on real data from the literature and simulated networks.

In Chapter 5, alternative methods to estimate the p_2 class of models are proposed. The main feature of this model class is the correlated crossed structure of random effects to represent actor heterogeneity and within-dyad dependence. In the literature there are proposals to estimate the parameters of p_2 models either by joint maximization methods (Van Duijn et al., 2004) or employing MCMC methods in a Bayesian approach (Zijlstra et al., 2009). The methods proposed are based on the Laplace approximation approach for random arrays. First-order Laplace approximations and simulated maximum likelihood based on Laplace importance sampling are studied. These solutions represent good approximations to maximum likelihood estimation. Numerical comparisons with alternative approaches are provided based on simulations and real data analyses.

In the appendix to the thesis, some attempts to apply the composite likelihood approach for ERGM estimation are briefly presented.

Chapter 2

Introduction to Network Analysis and Modeling

This chapter reviews some basic concepts of Network Analysis, mainly derived from the Social Network Analysis (SNA) field.

In the first part, we present the mathematical concepts and definitions for network representation with emphasis on different type of network dependencies.

Then, we summarize the state of art of the main model approaches for relational data that will be considered in the next chapters.

2.1 Graphs and Matrices for Network Representation

One of the definitions that the *Oxford English Dictionary* provides for *Network* is "collection of connected things". In Social network Analysis (SNA) a network is a structure resulting from ties among entities (actors), according to any irreducible property (relationship) connecting them. The study of these ties (relational data) is the core of the SNA. In particular as stated in De Nooy et al. (2011) (p. 5) *"... the main goal of social network analysis is detecting and interpreting patterns of social ties among actors ..."*.

Two class of mathematical tools are used to represent relational data: *graphs* and *matrix*.

Graphs, known in SNA literature as *sociogram* (Moreno, 1946), are useful tools to visualize and summarize the network information, providing insights on network characteristics.

From a theoretical point of view (Wasserman and Faust, 1994; Kolaczyk, 2009) a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is composed of a set of nodes or vertex (actors), and a set of edges (ties, relations). The cardinalities $N = |\mathcal{N}|$ and $E = |\mathcal{E}|$, representing respectively the numbers of nodes and edges in the network, are called the *order* and the *size* of the graph $\mathcal{G}$. The size varies between

0 and $\binom{N}{2}$. A graph is called *empty* or *null* if the edge set is empty. If the size is equal to the maximum number of possible ties, the graph is called *complete*.

Relations can have or not a versus, in the sense that by definition the choice of one node to establish a relation with another may imply or not the vice versa. For example, if in a classroom we ask a child, "Mike", to nominate the names of his best friends, then the people nominated by him should not necessarily consider "Mike" as their best friend. Whereas, if we are studying, for example, the co-authorship between two academics on a set of published papers then their relation has not a versus because both the academics share the status of co-author.

If the relation has not a versus, the graph is said *undirected*. Otherwise, the graph is called *directed graph* or *digraph* and the edges in $\mathcal{E}$ are ordered by their vertex, that is $\{i, j\}$ is different from $\{j, i\}$ for $i, j \in \mathcal{N}$. Directed ties are represented by arrows.

Two nodes $i, j \in \mathcal{N}$ are *adjacent* if joined by an edge in $\mathcal{E}$. A node is *isolated* if it is not adjacent to any node in the graph. Similarly two edges $e_1, e_2 \in \mathcal{E}$ are adjacent if joined by a common end point in $\mathcal{N}$. A vertex $i \in \mathcal{N}$ is *incident* on a edge $e \in \mathcal{E}$ if i is an end point of e.

For a digraph, the relation between two node i and j is *mutual* if both the edges $\{i, j\}$ and $\{j, i\}$ are present. Otherwise, if the relation has only one versus it is called *asymmetric*. In undirected graphs any relation is mutual.

Two graphs, $\mathcal{G}$ and $\mathcal{G}^*$, are *isomorphic* if there is a one-to-one mapping from the nodes of $\mathcal{G}$ to the nodes of $\mathcal{G}^*$ that preserve the adjacency of the nodes. So the two graphs differ only from a switching of the node labels.

Considering a subset $\mathcal{N}_H \subseteq \mathcal{N}$, of order d, the subgraph $H = (\mathcal{N}_H, \mathcal{E}_H)$ generated from $\mathcal{N}_H$ is the graph that has $\mathcal{N}_H$ as nodes set and $\mathcal{E}_H = (\mathcal{N}_H \times \mathcal{N}_H) \cap \mathcal{E}$ as edges set. A complete subgraphs is called *clique*. *Dyad* and *triad* are respectively 2-subgraphs and 3-subgraphs of the graph $\mathcal{G}$. Every dyads and triads of the network could be counted with respect to the number of Mutual, Asymmetric and Null edges (MAN census).

A *walk* is a sequence of nodes and edges, in which each node is incident with the edges that follow and precede it in the sequence. The length of a walk between two nodes is given by the number of steps. A walk in which any node figures only one time is called *path*. Considering two nodes, many paths can connect them, the shortest one is called *geodesic* and its length *geodesic distance* or simply *distance*. Obviously, for a digraph we should consider the versus of the relation.

A graph is *connected* if there is a path for each pair of nodes in the graph. A maximal connected subgraph, i.e a subgraph with maximal order that preserves the property of "to be connected", is called *component* of the graph.

According to the above definition of adjacency, a square matrix y, called *adjacency matrix* or *sociomatrix*, can be associated to the graph. The ele-

ments of the sociomatrix y are defined as

$$y_{ij} = \begin{cases} 1, & \text{if } \{i,j\} \in \mathcal{E}, \\ 0, & \text{otherwise,} \end{cases}$$

with $i, j = 1, \ldots, N$. Each y_{ij} is a binary variable that assumes values 1 if there is a tie from i and j. A measure of association can be linked to each variable y_{ij}.

This measure can be simply a positive or negative measure of association (as liking or disliking), in this case the network is called *signed*. Or it can be a quantitative measure indicating the strength of ties (*valued network*). The traditional methods developed in the literature focus mainly on binary networks.

No self-ties are considered, so the elements on the diagonal of the matrix are set to structural zeros.

Starting from the sociomatrix, a set of summary statistics can be defined.

The first statistic we introduce is the *density*:

$$\Delta = \frac{\sum_{i,j=1}^{N} y_{ij}}{N(N-1)}, \tag{2.1}$$

that is the ratio between the size of the network (cardinality of the edges set) and the maximum number of edges, equal to $N(N-1)$ for a network of order N. In the undirected case, the sociomatrix is symmetric so the total number of edges reduces to $N(N-1)/2$.

The numbers of incidental edges on each node, obtained by summing with respect to one index of the matrix, are known as *out-degree*:

$$y_{i+} = \sum_{i} y_{ij}, \tag{2.2}$$

and *in-degree*:

$$y_{+j} = \sum_{j} y_{ij}, \tag{2.3}$$

statistics.

In the undirected case they are called simply *degree* statistics because the in-degree and the out-degree statistics of a node coincide. The sum of the degree statistics of the nodes returns the size of the network.

2.1.1 Network Dependencies

The structure of a network is the result of complex dependencies among nodes. Following Snijders (2011) the main forms are:

- *Reciprocation:* for directed ties is how y_{ij} influences y_{ji} and vice versa. Reciprocation is not limited to the study of pairs but also of dependences in larger cycles such as y_{ij},y_{jh},y_{hi}.

- *Homophily:* how the similarity between actors affects the tendency to relate to each other. This leads to a higher probability of ties being formed between actors with similar values on relevant covariates.
- *Transitivity:* it can be expressed by the statement "friends of my friends are my friends", that is sharing a partner increases the propensity of two actors to be in relation.
- *Degree differentials:* differences in degree statistics can explain unequal *productivity* - the tendency to create ties - and *popularity* - the tendency to attract ties - of the actors in the network.

Specific network statistics can provide insight on the presence of the different types of dependencies in the observed network structure.

The distribution of degree statistics and the geodesic distances provide information about the levels of socio activity and connectivity of the actors in the network. Low proportion of large degree statistics combined with small values of distances can indicate the presence of few central and very active actors. Otherwise, large degree statistics combined with small values of distances may indicate that all actors are equally active in the network.

A set of statistics can be useful to evaluate transitivity in the network. The first statistic is the *clustering coefficient*:

$$CC = \frac{\sum_{i,j,h} y_{ij} y_{jh} y_{hi}}{\sum_{i,j,h} y_{ij} y_{jh}}, \tag{2.4}$$

defined as the ratio of the number of transitive triangles (divide by 6 in the undirected case) and the number of triads.

The *edges-wise shared partners* (ESP) is given by

$$ESP_{ij} = y_{ij} \sum_{k=1;\, k\neq i,j}^{N} y_{jk} y_{ki}, \tag{2.5}$$

and the *dyad-wise shared partners* (DSP) is defined as

$$DSP_{ij} = \sum_{k=1;\, k\neq i,j}^{N} y_{jk} y_{ki}. \tag{2.6}$$

Both the statistics (2.5) and (2.6) count the total number of nodes (partners) with whom two fixed nodes share relations. The ESP statistic (2.5) is an explicit measure of transitivity because it counts the total number of transitive triangles that have the ties y_{ij} as basis (Figure 2.1).

The DSP statistic (2.6) is an implicit transitivity measure because it counts the number of potential closed triads that we should observe if the tie y_{ij} were present (Figure 2.2).

It is possible to evaluate the level and the effects of the transitivity in the network by comparing ESP and DSP distributions.

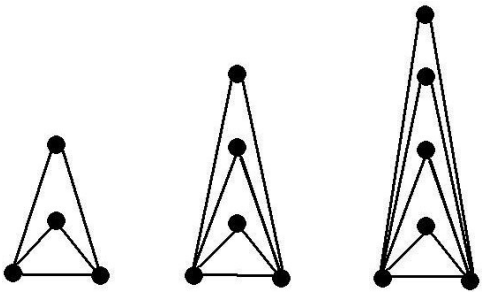

Figure 2.1: Examples of k-triangles alias edge-wise shared partners configurations (Robins et al., 2007b).

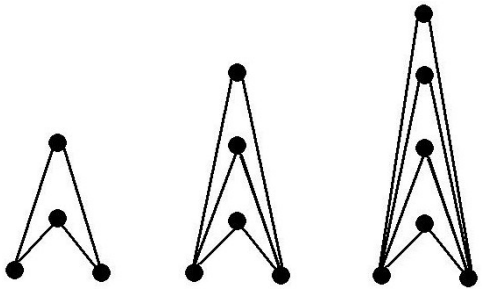

Figure 2.2: Examples k-two-paths alias dyad-wise shared partners configurations (Robins et al., 2007b).

2.2 Main Modeling Approaches in Network Analysis

Social behavior is complex and it seems reasonable to suppose that social processes giving rise to network ties are stochastic, so statistical models are appropriate to understand if, and how, certain network characteristics are more or less observed in the network than expected by chance. In general, models, and especially statistical models that are estimable from data and explicitly recognize uncertainty may help to understand the range of possible outcomes for processes on networks (Robins et al., 2007a). Moreover, network models are especially useful to deal with data dependence induced by social relations.

The leading assumption in modeling is that the observed network is generated by some (unknown) stochastic process (Wasserman and Faust, 1994; Wasserman et al., 2007; Kolaczyk, 2009; Snijders, 2011).

The statistical variables are represented by the ties between the actors. This means that a network of order N contain informations on $N(N-1)$ relations.

The model approaches that we will introduce are mainly taken from the SNA literature. All these approaches include terms to account for the dependence in the network and make different assumptions concerning the stochastic part.

In the following, the directed case is assumed, unless otherwise stated.

2.2.1 Homogeneous Bernoulli Graph Models

The easiest way to model a set of relational data is to consider independent ties. The approach known as *homogeneous Bernoulli graph*, or also *Erdös-Renyi* model (Erdös and Rényi, 1959), simply assumes that:

$$P(Y_{ij} = 1) = p,$$

for every $i, j = 1, \ldots, N$; $i \neq j$. Equivalently:

$$\text{logit}\{P(Y_{ij} = 1)\} = \eta = \log \frac{p}{1-p}.$$

This implies that the probability for the observed network is

$$\begin{aligned} P(Y = y) &= p^{y_{++}}(1-p)^{M-y_{++}} \\ &= \frac{e^{\theta y_{++}}}{k(\eta)} \end{aligned} \tag{2.7}$$

where $y_{++} = \sum_{i,j} y_{i,j}$ is the size of the network, $M = N(N-1)$ the number of possible edges and $k(\eta) = e^{\eta}/(1+e^{\eta})$ is a normalizing constant. It is

trivial to see that under the edges independence assumption, the Erdös-Renyi model is exactly a logistic model for binary data.

This approach represents a poor model, just slightly less poor compared to assuming that each tie is the result of a coin flip.

A first simple improvement is to assume independence of ties but conditioned to the valued assumed by a set of network covariates. The network covariates are often derived from specific actor attributes that are combined to obtain values for their ties.

Some of these covariates refer directly to the main effect of the actor attributes to the ties. For example, the network covariate $u(x_i, x_j)$, on the tie (i, j), relative to the actor attribute x is

$$u(x_i, x_j) = \begin{cases} x_i & \text{if relative to the productivity of the sender,} \\ x_j & \text{if relative to the popularity of the receiver,} \\ x_i + x_j & \text{if relative to the sociality of both.} \end{cases} \tag{2.8}$$

Other covariates refer to the homophily effect of some qualitative actor attributes, such as

$$u(x_i, x_j) = \begin{cases} 1 & \text{if } x_i = x_j, \\ 0 & \text{Otherwise.} \end{cases} \tag{2.9}$$

They could be defined on specific values s_0 assumed by the variables

$$u(x_i, x_j) = \begin{cases} 1 & \text{if } x_i = x_j = s_0, \\ 0 & \text{Otherwise.} \end{cases} \tag{2.10}$$

Alternatively, they could defined by the different conditions

$$u(x_i, x_j) = \begin{cases} 1 & \text{if } x_i = 1 \text{ and } x_j = 2, \\ \vdots & \vdots \\ s(s-1) & \text{if } x_i = s-1 \text{ and } x_j = s, \end{cases} \tag{2.11}$$

where s is the number of categories of the qualitative actor attribute.

After including q of these terms, the model (2.7) becomes:

$$P(Y = y) = \frac{\exp(\sum_{ij} \eta_{ij} y_{ij})}{k(\theta)},$$

where

$$\eta_{ij} = \text{logit}\{P(Y_{ij} = 1)\} = \sum_{k=1}^{q} \theta_k u_k(x),$$

and

$$k(\theta) = \log \left(1 + \exp \left\{ \sum_{k=1}^{p} \theta_k u_k \right) \right\},$$

The resulting model is still a logistic model as in (2.7).

2.2.2 p_1 Models

The p_1 model is due to Holland and Leinhardt (1981). It assumes dyad independence, after including terms for popularity, productivity of actors and the mutuality of the relation.

In particular:

$$\text{logit}\{P(Y_{ij} = y_{ij}, Y_{ji} = y_{ji})\} = \theta(y_{ij} + y_{ji}) + \rho y_{ij} y_{ji} + \alpha_i y_{ij} + \beta_j y_{ji}.$$

The probability mass function of the full network is:

$$\begin{aligned} P(Y = y) = & \quad \frac{\exp[\sum_{i \leq j} \{\theta(y_{ij} + y_{ji}) + \rho y_{ij} y_{ji} + \alpha_i y_{ij} + \beta_j y_{ij}\}]}{k(\mu, \rho, \alpha, \beta)} \\ = & \quad \frac{\exp(\mu y_{++} + \rho \sum_{i \leq j} y_{ij} y_{ji} + \sum_i \alpha_i y_{i+} + \sum_j \beta_j y_{+j})}{k(\mu, \rho, \alpha, \beta)}, \end{aligned} \tag{2.12}$$

where the normalizing constant is:

$$\begin{aligned} k(\mu, \rho, \alpha, \beta) \quad = \quad \prod_{i \leq j} \quad & \{1 + \exp(\mu + \alpha_i + \beta_j) + \exp(\mu + \alpha_j + \beta_i) \\ & + \exp(\mu + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho)\}. \end{aligned}$$

The sufficient statistics for model estimation are the total number of edges (equivalently we can use the density by rescaling the parameter μ), the numbers of mutual ties governed by the parameter ρ and the degree statistics (2.2) and (2.3) governed by the parameters α_i and β_j that measure the productivity and the attractiveness of the actors.

The dyad independence assumption makes the p_1 model a multinomial logit regression model.

2.2.3 Exponential Random Graph Models

The general class of Exponential Random Graph Models (ERGM), also known as p^* (Wasserman and Pattison, 1996), takes his name from the fact that it can be written as an element of the exponential family (Barndorff-Nielsen, 1978):

$$P_{\boldsymbol{\theta}}(Y = y) = \exp\{\boldsymbol{\theta}\, u(y, x) - \psi(\boldsymbol{\theta})\}\,, \tag{2.13}$$

where

$$\psi(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta}; \mathcal{Y}, x) = \log \sum_{y \in \mathcal{Y}} \exp\{\boldsymbol{\theta}\, u(y, x)\}$$

is the normalizing constant.

ERGMs include homogeneous Bernoulli graph and p_1 models as special cases, but it allows also for transitivity. Transitivity requires to study larger configurations of nodes, at least the triads. In order to study this kind of

configurations, Frank and Strauss (1986) adopted a first order Markovian assumption for which two edges are dependent only if they share a common node. This assumption is related to the concept of *Markov Graph*. To explain this concept we need to define what it is the *Dependence Graph* of an observed graph (Frank and Strauss, 1986).

Definition 1. *Given a graph $\mathcal{G}$ of order N, let $Y = (Y_1, Y_2, \ldots, Y_m)$ be the set of the $m = N(N-1)$ possible edges of the graph. And $M = \{1, \ldots, m\}$ is the index set.*

The Dependence Graph $\mathcal{D}$ of $\mathcal{G}$ is a non-random graph that specifies the dependence structure between the random variables Y_i, $i \in M$. The vertex of $\mathcal{D}$ are the edges of $\mathcal{G}$. The edges of $\mathcal{D}$ are the pairs of edges of $\mathcal{G}$ that are conditionally dependent.

In Figure 2.3 and 2.4 we can see an example for a graph with five nodes.

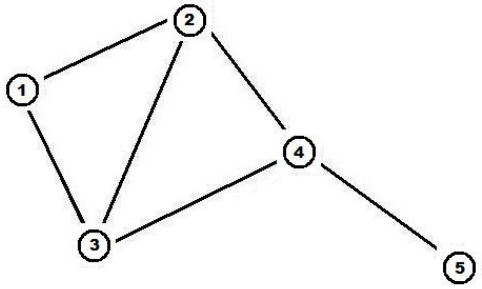

Figure 2.3: Example of a network of order five.

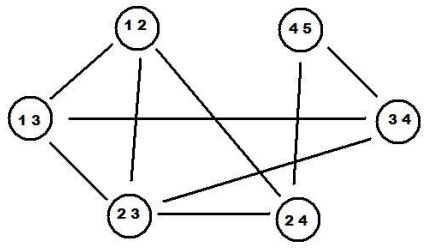

Figure 2.4: Dependence graph of the example in Figure 2.3.

Definition 2. *A graph $\mathcal{G}$ is a Markov graph if $\mathcal{D}$ contains no edges between disjoint sets in M. This means that nonincident edges in $\mathcal{G}$ are conditionally independent given the rest of the graph.*

A model for a Markov graph is obtained working on the cliques of the dependence graph. According to the Hammersley-Clifford theorem (see Besag, 1974), it can be proved that the probability function of the network can

be factorized as a function of only the cliques of the dependence graph. It can also be proved that any clique on the dependence graph $\mathcal{D}$ is identified on the original graph $\mathcal{G}$ as a triangle or a k-star. In an undirected graph, a k-star (Figure 2.5) represents a configuration in which k edges are expressed by one node.

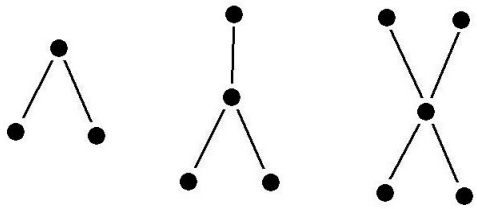

Figure 2.5: Examples of k-stars.

In Figure 2.6 we can see how the blue and the red cliques in the dependence graph identify a triangle and a 3-star on the network in Figure 2.3.

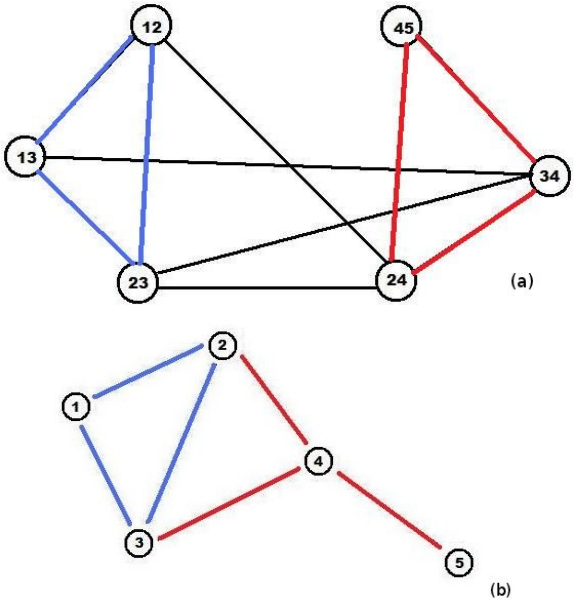

Figure 2.6: Relations between cliques in the dependence graph (a) and configurations in the original graph (b) for the network in Figure 2.3.

An ERGM model with Markovian property (Wasserman and Pattison, 1996) is

$$P(Y = y) = \frac{\exp\left\{\boldsymbol{\theta}_t T(y) + \boldsymbol{\theta}_{s_1} S_1(y) + \ldots + \boldsymbol{\theta}_{s_{n-1}} S_{n-1}(y)\right\}}{Z(\boldsymbol{\theta}; \mathcal{Y})}, \qquad (2.14)$$

where $T(y)$ is the number of triangle in the network, $S_k(y)$ for $k = 1, \ldots, n-1$ are the numbers of k-stars in the network, $S_1(y)$ is the number of 1-stars (that is the number of edges):

$$\text{number of edges:} \quad S_1(y) = \sum_{1 \le i \le j \le g} y_{ij}, \tag{2.15}$$

$$\text{number of } k\text{-stars:} \quad S_K(y) = \sum_{1 \le i \le g} \binom{y_{i+}}{k},\ k \ge 2 \tag{2.16}$$

$$\text{number of triangles:} \quad T(y) = \sum_{1 \le i \le j \le h \le g} y_{ij} y_{ih} y_{jh}. \tag{2.17}$$

The statistics in formulas (2.15), (2.16) and (2.17) represent the sufficient statistics of a ERGM, but in real applications is common to use models that include only the terms for triangle and up to 3-stars (Robins et al., 2007b).

ERGMs and especially the Markovian models suffer problems such as:

- **degeneracy:** Model degeneracy occurs when the model places disproportionate probability mass on only a few of the possible graph configurations in $\mathcal{Y}$ (empty or complete graphs).
- **instability:** A random graph model is *instable* if small changes in the parameter values result in large changes in the probabilistic structure of the model.

The instability and degeneracy problems of ERGMs are related to the closeness of the observed sufficient statistics at the boundary of the convex hull of their possible values (Handcock et al., 2003; Rinaldo et al., 2009; Schweinberger, 2011).

In order to solve these problems and to improve the ability of the model to catch the dependence structure in the network, new specifications have been proposed (Snijders et al., 2006) including the partial dependence concept proposed by Pattison and Robins (2002) in addition to Markov dependence. These new specifications include in the model the **Alternating k-statistics** (Figure 2.7), as higher order measures of transitivity.

- **Alternating k-stars**

$$u = \sum_{k=2}^{n-1} (-1)^k \frac{S_k}{\lambda_s^{k-2}}; \tag{2.18}$$

- **Alternating k-triangles**

$$v = \sum_{k=2}^{n-1} (-1)^k \frac{T_k}{\lambda_t^{k-2}}; \tag{2.19}$$

- **Alternating independent k-two-paths**

$$w = \sum_{k=2}^{n-1} (-1)^k \frac{P_k}{\lambda_p^{k-2}}. \tag{2.20}$$

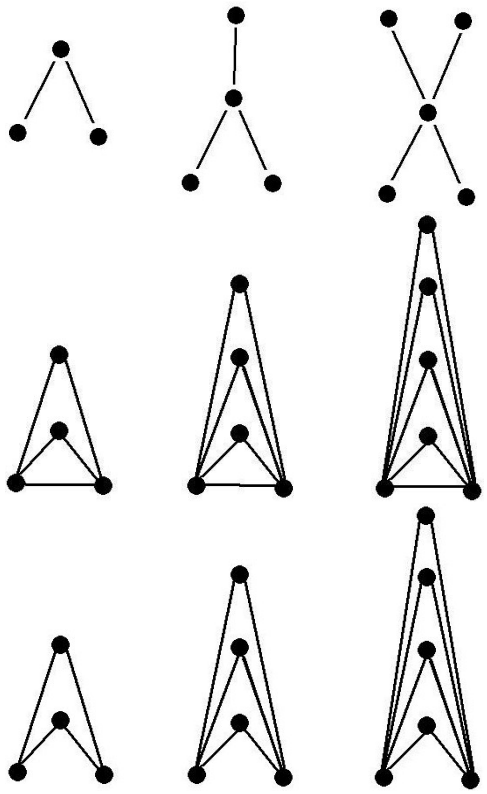

Figure 2.7: Examples of k-stars, k-triangles and k-two-paths (Robins et al., 2007b).

The parameters λ_s, λ_t and λ_t are respectively weights associated to the statistics.

It could be shown (Hunter, 2007) that these alternating k-statistics are strictly related to the distribution of the network statistics discussed in §2.1.1. Every k-star statistics can be written as a combination of the degree distribution

$$S_1(y) = \frac{1}{2}\sum_{i=1}^{n-1} iD_i(y), \qquad S_k = \sum_{i=k}^{n-1} \binom{i}{k} D_i(y). \tag{2.21}$$

Every k-triangle statistics can be written as a combination of the edge-wise shared partners statistic (2.5) distributions

$$T_1(y) = \frac{1}{3}\sum_{i=1}^{n-1} iEP_i(y), \qquad T_k = \sum_{i=k}^{n-2} \binom{i}{k} EP_i(y). \tag{2.22}$$

Finally every k-two-path statistics can be written as a combination of the dyad-wise shared partners statistic (2.6) distributions

$$P_2(y) = \frac{1}{2}\sum_{i=2}^{n-2} \binom{i}{2} DP_i(y), \qquad P_k = \sum_{i=k}^{n-k} \binom{i}{k} DP_i(y). \tag{2.23}$$

All these relations lead to a geometrically equivalent specification (Hunter, 2007):

- **Geometrically weighted degree (GWD)** is exactly equivalent to the *Alternating k-stars*;

$$u(y, \boldsymbol{\theta}_s) = e^{\boldsymbol{\theta}_s} \sum_{i=1}^{n-1} \left\{ 1 - \left(1 - e^{-\boldsymbol{\theta}_s}\right)^i \right\} D_i(y). \tag{2.24}$$

- **Geometrically weighted edge-wise shared partners (GWESP)** is exactly equivalent to the *Alternating k-triangles*;

$$v(y, \boldsymbol{\theta}_t) = e^{\boldsymbol{\theta}_t} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\boldsymbol{\theta}_t}\right)^i \right\} EP_i(y). \tag{2.25}$$

- **Geometrically weighted dyad-wise shared partners (GWDSP)** is exactly equivalent to the*Alternating independent two-paths.*

$$w(y, \boldsymbol{\theta}_p) = e^{\boldsymbol{\theta}_p} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\boldsymbol{\theta}_p}\right)^i \right\} DP_i(y). \tag{2.26}$$

Also in this case, the parameters $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t$ are weights associated to the statistics.

These equivalent parameterizations have mathematical elegance and enhance interpretation, because more strictly related to degree, edge-wise and dyad-wise shared partners statistic distributions, and they will be used in the applications reported in the next chapters.

Estimation of ERGMs

Maximum likelihood estimation for ERGMs is very challenging. Considering the ERGM (2.13)

$$P(Y = y) = \exp\left\{\boldsymbol{\theta}\, u(y, x) - \psi(\boldsymbol{\theta})\right\},$$

the main difficulty for computing the likelihood function concerns the normalizing constant

$$\psi(\boldsymbol{\theta}) = \log\left[\sum_{y \in \mathcal{Y}} \exp\{\boldsymbol{\theta}\, u(y, x)\}\right],$$

that is the result of the sum on all possible networks. For a directed network with g nodes we have $2^{g(g-1)}$ combinations, that can be computed in explicit form only for small size networks or in trivial cases.

The log-likelihood function in this case is

$$\ell(\boldsymbol{\theta}) = \log P(Y = y) = \boldsymbol{\theta} u(y, x) - \psi(\boldsymbol{\theta}).$$

The score function is

$$\ell_*(\theta) = \nabla \ell(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta} \ell(\boldsymbol{\theta}) = u(y, x) - \nabla \psi(\boldsymbol{\theta})$$

The results for exponential families (Barndorff-Nielsen, 1978) state that:

$$\widehat{\boldsymbol{\theta}}_{MLE}: \qquad \nabla \psi(\boldsymbol{\theta}) = u(y, x),$$

and the Fisher information matrix is:

$$i(\boldsymbol{\theta}) = \mathbb{E}_\theta\{-\nabla^2 \ell(\boldsymbol{\theta})\} = \mathbb{E}_\theta\{\nabla \psi(\boldsymbol{\theta})\} = \mathbb{V}\{u(Y, x)\}.$$

These results can not be used directly due to the impossibility to compute $\psi(\boldsymbol{\theta})$ in explicitly form, and alternative methods to estimate $\boldsymbol{\theta}$ have been proposed.

Given an edge y_{ij}, let Y_{ij}^C denote the adjacency matrix obtained excluding the edge (i, j).

Denoting by Y_{ij}^{+} and Y_{ij}^{-} respectively the adjacency matrix in which we set $y_{ij} = 1$ and $y_{ij} = 0$, it is easy to prove that (Strauss and Ikeda, 1990)

$$\begin{aligned} P(Y_{ij} = 1 | Y_{ij}^{C}; \boldsymbol{\theta}) &= \frac{P(Y_{ij}^{+}; \boldsymbol{\theta})}{P(Y_{ij}^{+}; \boldsymbol{\theta}) + P(Y_{ij}^{-}; \boldsymbol{\theta})} \\ &= \frac{\exp\{\boldsymbol{\theta}\, u(y_{ij}^{+}, x_{ij})\}}{\exp\{\boldsymbol{\theta}\, u(y_{ij}^{+}, x_{ij})\} + \exp\{\boldsymbol{\theta}\, u(y_{ij}^{-}, x_{ij})\}}. \end{aligned} \tag{2.27}$$

It follows that

$$\begin{aligned} \text{logit} P(Y_{ij} = 1 | Y_{ij}^{C}; \boldsymbol{\theta}) &= \boldsymbol{\theta}\, (u(y_{ij}^{+}, x_{ij}) - u(y_{ij}^{-}, x_{ij})) \\ &= \boldsymbol{\theta}\, \Delta u(y_{ij}, x_{ij}). \end{aligned} \tag{2.28}$$

This implies the conditional independence of the edges given their complementary graph. It can be noted that the expression (2.28) does not depend on $\psi(\theta)$, but only on the difference of the sufficient statistic, sometimes called *vector of change statistics* (Wasserman and Pattison, 1996). From these results model coefficients can be estimated in a simple way.

In particular the function

$$L_p(\theta; y) = \prod_{ij} P(Y_{ij} = y_{ij} | Y_{ij}^{C}; \boldsymbol{\theta}), \tag{2.29}$$

is called *pseudo-likelihood* function and the value $\widetilde{\boldsymbol{\theta}}$ that maximizes (2.29) is called maximum pseudo-likelihood estimate (MPLE) (Strauss and Ikeda, 1990). The MPLE is obtained under the working assumption of conditional link independence in the network and so it coincides with the logistic regression coefficients for the observed edges and the vector of change statistics $\Delta u(y_{ij}, x_{ij})$.

Unfortunately often the MPLE is not a good estimator (Geyer and Thompson, 1992; Van Duijn et al., 2009) because it does not consider the dependence between ties, that it is the most interesting aspect in network modeling.

An alternative estimation method for ERGMs is via simulate maximum likelihood estimation methods (Snijders, 2002; Hunter and Handcock, 2006).

These methods use Markov Chain Monte Carlo (MCMC) and exploit the possibility to simulate a sample of M networks $y_1^*, \ldots, y_M^*$ from a ERGM for a fixed value $\boldsymbol{\theta}_k$. The network simulations are obtained by some classical algorithms like Gibbs-sampling or Metropolis-Hastings. The goal is to achieve the value $\widetilde{\boldsymbol{\theta}}$ that maximize the approximation

$$\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_k) \simeq (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^T - \log \left[\frac{1}{M} \sum_{i=1}^{M} \exp\{(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^T u(y_i^*)\} \right]. \tag{2.30}$$

The updating procedure for $\boldsymbol{\theta}_k$ is obtained via the Newton-Raphson method, or its modifications (Hunter and Handcock, 2006; Snijders, 2002). Some of these methods are often inefficient and computer intensive, especially for Markovian models (2.14) where, as said before it is easy to fall in degeneracy and instability problems (Handcock et al., 2003). In Chapter 3, we will propose an alternative approach for obtaining maximum likelihood estimates of ERGM parameters.

2.2.4 Latent Space Models

Latent space models are derived from a different strategy to catch and explain the dependencies of the data (Lazarsfeld and Henry, 1968). They assume the existence of latent variables and the distribution of the data is simply computable given these latent variables. The models that belong the latent space family are centered on the concepts of "position", that takes different meanings in the social network literature (see Snijders, 2011).

Here we refer mainly to the approach proposed by Hoff et al. (2002). This approach assumes links independence given the positions of the nodes (actors) in a small dimensional latent space $\mathbb{R}^d$, often not larger than $d = 2, 3$. The conditional probability model for the adjacency matrix y is

$$\begin{aligned} P(Y = y|Z, X, \boldsymbol{\theta}) &= \prod_{i \neq j} P(Y_{ij} = y_{ij}|Z_i, Z_j, X_{ij}; \boldsymbol{\theta}), \qquad (2.31) \\ &= \prod_{i \neq j} \frac{y_{ij} e^{\eta_{ij}}}{1 + e^{\eta_{ij}}}, \end{aligned}$$

where

$$\eta_{ij} \equiv \text{logit} P(Y_{ij} = 1) = \alpha + \beta X_{ij} - d(Z_i, Z_j)$$

is the logit linear predictors with $\boldsymbol{\theta} = (\alpha, \beta)$, $d(Z_i, Z_j) = \|Z_i - Z_j\|$ is a distance in $\mathbb{R}^d$ between the positions Z of the nodes, and X is a set of dyad covariates defined on a set of actor attributes, see §2.2.1.

At this stage the model implies symmetry on the probabilities of the links between nodes. The symmetry assumption does not include different productivity and attractiveness effects for the nodes. Hoff (2005) solved the problem including in the model Gaussian random effects for productivity (δ) and attractiveness (γ).

In latest years, early versions have been quickly improved to handle many network observed characteristics. Often, the main interest focuses in identifying groups of similar nodes (White et al., 1976; Wang and Wong, 1987; Anderson et al., 2002). For this goal, clustering methods must be used to analyze explicitly the set of latent positions inferred by the latent space model. To allow joint inference on latent positions and clusters, Handcock et al. (2007) introduced an explicit clustering model in the latent space in the form of a mixture of a spherical Gaussian distributions:

$$\begin{cases} P(Y|Z,X,\boldsymbol{\theta}) &= \prod_{i\neq j} \frac{y_{ij}e^{\eta_{ij}}}{1+e^{\eta_{ij}}} \\ \eta_{ij} &= \alpha + \beta X_{ij} - \|Z_i - Z_j\| \\ Z &\sim \sum_k N_d(\mu_k, \sigma_k^2 I). \end{cases} \tag{2.32}$$

All these extensions were summarized in what it is called Latent Cluster Random effects model (LCRM) (Krivitsky et al., 2009):

$$\begin{cases} P(Y|Z,X,\boldsymbol{\theta},\delta,\gamma) &= \prod_{i\neq j} \frac{y_{ij}e^{\eta_{ij}}}{1+e^{\eta_{ij}}} \\ \eta_{ij} &= \alpha + \beta X_{ij} - \|Z_i - Z_j\| + \delta_i + \gamma_j \\ Z &\sim \sum_k N_d(\mu_k, \sigma_k^2 I), \\ \delta &\sim N(0, \sigma_\delta^2), \\ \gamma &\sim N(0, \sigma_\gamma^2). \end{cases} \tag{2.33}$$

In this formulation the main kind of dependencies of the data are included in the model. Transitivity, clustering and homophily are caught by closeness in the latent space. Differences in socio activity are explained by the random effects. Terms for dyad covariates defined on a set of actor attributes can also be considered to help the understanding of the underlying social structure.

The main task in the estimation of a LCRM (2.33) is to determine the latent positions of the actors in the latent space. The estimation is obtained under a Bayesian framework, via MCMC procedures (Krivitsky and Handcock, 2008; Krivitsky et al., 2009).

2.2.5 p_2 Models

This model approach has been proposed as an extension of the p_1 model and so it was called p_2 (Van Duijn et al., 2004). In particular, the p_2 model reduces the number of parameters of the p_1 model and uses random effects to simplify the way to model the difference in productivity and attractiveness of actors.

As shown in §2.2.2, p_1 model (2.12) assumes that the probability for a dyad (y_{ij}, y_{ji}) is given by

$$P(Y_{ij} = y_1, Y_{ji} = y_2) = \frac{\exp\{y_1(\mu_{ij} + \alpha_i + \beta_j) + y_2(\mu_{ij} + \alpha_j + \beta_i) + \rho_{ij} y_1 y_2\}}{1 + \exp(\mu_{ij} + \alpha_i + \beta_j) + \exp(\mu_{ij} + \alpha_j + \beta_i) + \exp(2\mu_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho_{ij})},$$

where α_i is the sender parameter of actor i, β_i is receiver parameter of actor i, whereas μ_{ij} and ρ_{ij} are the density and the reciprocity parameters respectively of the dyad (i, j).

In the p_2 model, these parameters are defined as function of some covariates and random effects, namely:

$$\alpha_i = X_{1i}\,\gamma_1 + a_i\,, \qquad \beta_i = X_{2i}\,\gamma_2 + b_i\,, \qquad \mu_{ij} = \mu + Z_{1ij}\,\delta_1\,, \qquad \rho_{ij} = \rho + Z_{2ij}\,\delta_2\,.$$

Here X_1 and X_2 are actor-specific design matrices, Z_1 and Z_2 contain dyad-specific covariates and $U_i = (a_i, b_i)^T$ are independent normally distributed random effects:

$$U_i \sim N_2(0, \Sigma)\,, \qquad \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}\,. \tag{2.34}$$

Notice that random effects model the ties sent or received by a given actor, that are thus assumed to be dependent. All the parameters of this model can be collected together in the vector θ:

$$\theta = (\gamma_1, \gamma_2, \mu, \delta_1, \rho, \delta_2, \sigma_A^2, \sigma_{AB}, \sigma_B^2)^T\,.$$

2.3 Real Data Examples

In the following we introduce the data sets that will be used in the next chapters. They are taken from the literature and, in some cases, they are available in the suite of software packages `statnet` for `R`.

2.3.1 Molecule Network

This synthetic 20-node graph, shown in Figure 2.8, resembles the chemical structure of a molecule.

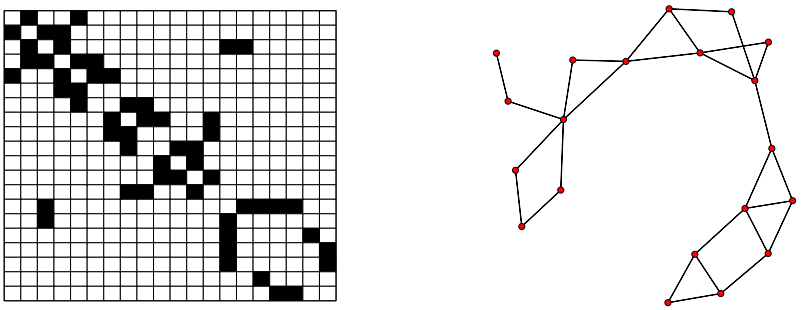

Figure 2.8: `molecule` data.

This data set is used in a recent paper on the Bayesian estimation for ERGM (Caimo and Friel, 2011).

2.3.2 Ecoli Network

This network data set (Salgado et al., 2001; Shen-Orr et al., 2002) is a biological network in which the nodes are operons in Escherichia Coli.

The analyses reported in the next chapters are based on the network object `ecoli2`, which is an undirected network with 418 nodes.

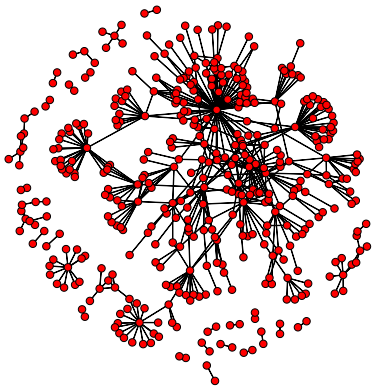

Figure 2.9: `ecoli2` data.

Applications on this data set are present in Hummel et al. (2011).

2.3.3 Kapferer's Tailor Shop Network

This data set (Kapferer, 1972) is about undirect interactions between 39 workers (`kapferer`) in a tailor shop in Zambia over a period of ten months. The focus was on the changing patterns of alliance among workers during extended negotiations for higher wages. The data contains an actor attribute (`highstatusjob`) refers to the prestige of the job, maybe who works in a strategical position.

This data set was considered in Snijders and Nowicki (1997) and Hummel et al. (2011).

2.3.4 Lazega's Lawyers Network

The data set refers on informal relationships between 71 lawyers of a corporate law firm in New England, collected by Lazega (2001). Networks for advice, collaboration and friendships relations are available. Additional attributes were recorded for each lawyer: `seniority` (i.e., the rank into the law firm); `gender`; `status` (i.e., if the lawyer is a partner or an associate); `office` location; `years` in the firm; `age`; type of `specialty` (i.e., litigation or corporate law); law `school` (i.e., where Yale = 1). For additional details

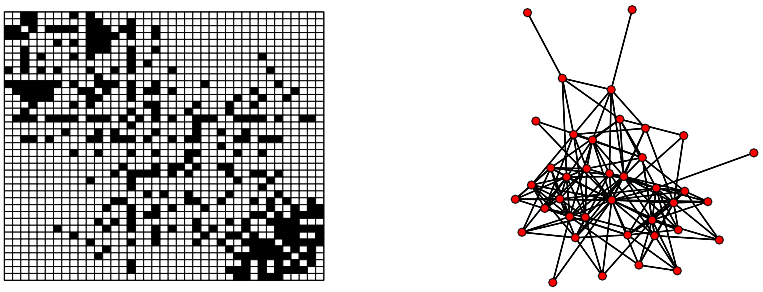

Figure 2.10: `kapferer` data.

see Lazega and Pattison (1999). In our analyses, we restricted the attention to the network of collaboration among partners on mutual relations, Figure 2.11, and on the network of the friendship among associates, Figure 2.12.

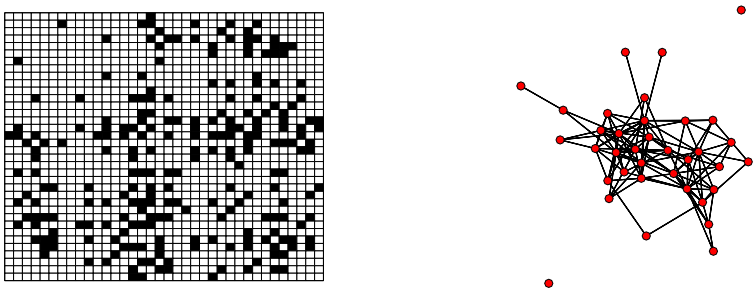

Figure 2.11: Lazega's Lawyers Data, collaboration among partner lawyers.

Applications on this data set can be found in Van Duijn et al. (2004), Hunter and Handcock (2006) and Kolaczyk (2009).

2.3.5 Sampson's Monks Network

In its work (Sampson, 1968), Sampson recorded the social interactions among a group of 18 monks. The study identifies three groups in the monastery (loyal, Turks and outcasts), added as an actor attribute (`group`).

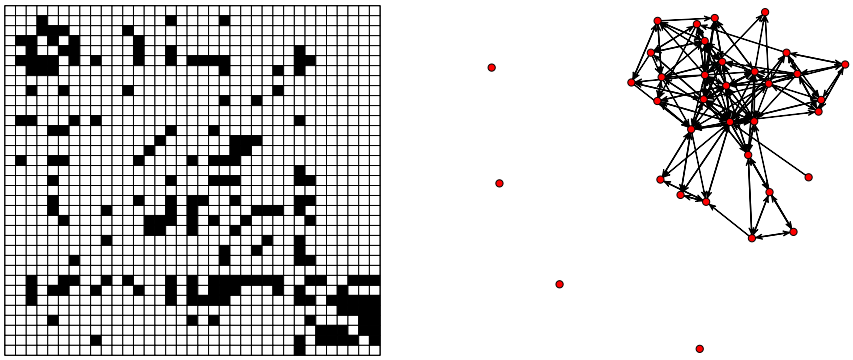

Figure 2.12: Lazega's Lawyers Data, friendship among associate lawyers.

What used in the following is the time-aggregated graph, `data(samplike)` in R. It is the cumulative tie for liking over three time periods. For this network, a tie from monk A to monk B exists if A nominated B as one of his three best friends at any of the three time points. The relation is thus directed.

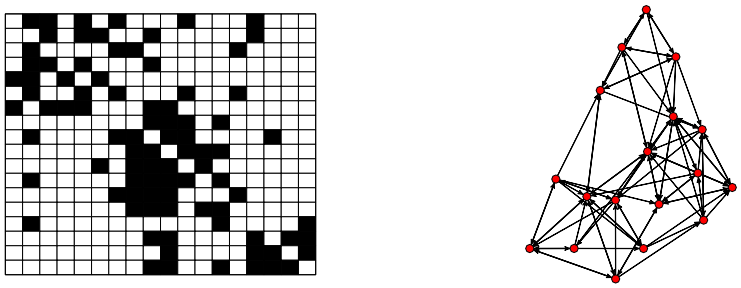

Figure 2.13: `samplike` Data.

An application for this data set, Figure 2.13, is given in Krivitsky et al. (2009).

2.3.6 Krackhardt's High-tech Managers Network

It is a one-mode network, with three relations measured on a set of people. The data were gathered by Krackhardt (1987) in a small manufacturing organization on the west coast of the U.S. The firm, producing high-tech machinery, employed approximately one hundred people, and had twenty-one manager. These twenty-one managers are the set of actors for this data set. The data set includes four actor attributes: `age`; length of time employed by the organization `tenure`; `level` in the corporate hierarchy; and the `department`. The first two are measured in years. There are four department in the firm. All but the president have a department attribute codes as an integer from 1 to 4. Thel `level` attribute is measured on an integer scale from 1 to 3: `1 = CEO, 2 = vice president, and 3 = manager`. In our analysis we focused only on the "friendship" relations. More about this data set can be found in Wasserman and Faust (1994).

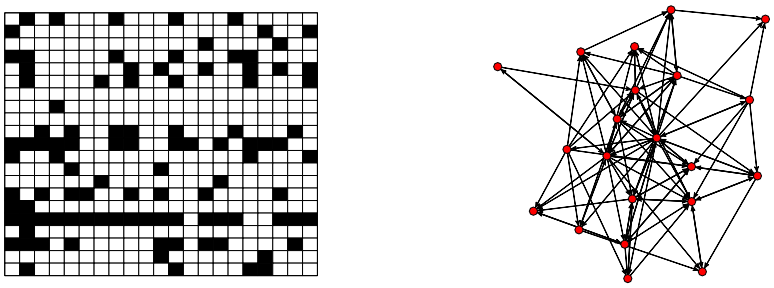

Figure 2.14: HighTech Managers Data.

2.3.7 Dutch Social Behavior Study

The data from the Dutch Social Behavior Study (Baerveldt and Snijders, 1994), already analysed in Zijlstra et al. (2005), are about the emotional support on two networks of high school pupils, of order 62 and 39 respectively. The available actor attributes are the with different `ethnic background` and `gender`. The networks are plotted in Figure.

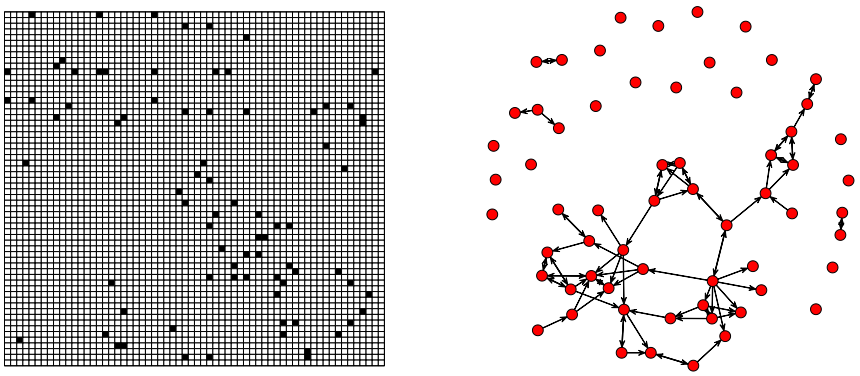

Figure 2.15: Dutch Social Behavior Study, school 1 data.

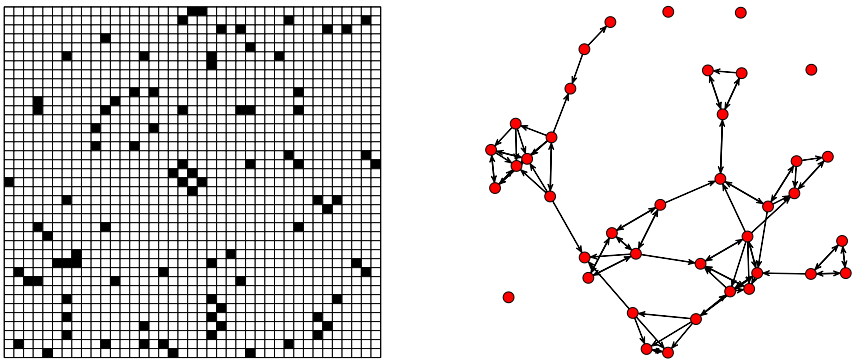

Figure 2.16: Dutch Social Behavior Study, school 2 data.

Chapter 3

Monte Carlo Quasi-Newton Estimation for ERGMs

This chapter presents a Monte Carlo Quasi-Newton algorithm for computing the maximum likelihood estimate of an exponential random graph model (ERGM).

The existing procedures to estimate the parameters of an ERGM, based on simulated maximum likelihood methods, are computationally challenging, due to numerical difficulties to approximate the likelihood function. So they at times fail to converge as the likelihood approximation may degrade, especially for certain choices of the sufficient statistics of interest.

The method is proposed to improve the quality of the estimated ERGM and reducing degeneracy and instability problems.

3.1 Introduction

We focus now on exponential random graph model (ERGM) for an observed network y of order g. The model, described in §2.2.3, has the general form

$$P_{\boldsymbol{\theta}}(y = y) = \exp\{\boldsymbol{\theta}^T u(y) - \psi(\boldsymbol{\theta})\}\,. \tag{3.1}$$

The main problem to work with this model concerns the intractability of the normalizing quantity $\psi(\boldsymbol{\theta})$, that does not permit to compute explicitly the likelihood and to use the classic results for exponential families (Barndorff-Nielsen, 1978).

The existing procedures to estimate $\boldsymbol{\theta}$ are generally based on simulated maximum likelihood (Snijders, 2002; Hunter, 2007). These methods take advantage of Markov Chain Monte Carlo methods to simulate a sample of M networks $y_1^*, \ldots, y_M^*$ from an ERGM (3.1) for a given parameter value $\boldsymbol{\theta}_0$.

Denoting by

$$\ell(\boldsymbol{\theta}) = \boldsymbol{\theta}^T u(y, x) - \log \psi(\boldsymbol{\theta}) \tag{3.2}$$

the log-likelihood function at $\boldsymbol{\theta}$, an approximation to $\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0)$ is (Hunter and Handcock, 2006)

$$\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_k) \simeq (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^T - \log \left[\frac{1}{M} \sum_{i=1}^{M} \exp\{(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^T u(y_i^*)\} \right]. \quad (3.3)$$

A well-known problem for simulated maximum likelihood is that if $\boldsymbol{\theta}_0$ is not close to MLE $\widehat{\boldsymbol{\theta}}$, the approximation (3.3) can be very poor (Caimo and Friel, 2011; Hummel et al., 2011). At times, the maximizer $\widetilde{\boldsymbol{\theta}}$ of (3.3) may not even be closer to $\widehat{\boldsymbol{\theta}}$ than $\boldsymbol{\theta}_0$, so that repeating the computation of $\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0)$ after setting $\boldsymbol{\theta}_0 = \widetilde{\boldsymbol{\theta}}$ and iterating may not give a convergent algorithm. For the special case of ERGMs, the determination of $\widehat{\boldsymbol{\theta}}$ is made even more difficult by possible model degeneracy that may occur for certain choices of $u(y)$, and actually the MLE may not even exist; see the discussion in Rinaldo et al. (2009) and the references therein. Especially for Markov random graphs, for certain parameter values there may be symptoms of *near degeneracy* (Handcock et al., 2003), although the MLE does exist. The recent work by Schweinberger (2011) illustrates some mathematical aspects of this issue, providing some partial explanations of it.

The *Steplength* algorithm (Hummel et al., 2011) represents a notable improvement over standard simulated maximum likelihood. It proceeds through a series of steps based on alternating between the canonical parameterization, $\boldsymbol{\theta}$, of the exponential family (3.1) and the mean-value parameterization (Handcock et al., 2003), where the parameter is given by $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\{u(Y)\}$.

Exploiting the mean-parametrization and some loose distributional assumptions, reasonable in many cases, a log-likelihood approximation alternative to (3.3) could be obtained noting that (Hummel et al., 2011)

$$\begin{aligned}
\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0) &= \log \left[\frac{\sum_{y \in \mathcal{Y}} \exp\{\boldsymbol{\theta} u(y)\}}{\sum_{y \in \mathcal{Y}} \exp\{\boldsymbol{\theta}_0 u(y)\}} \right] \\
&= \log \left[\sum_{y \in \mathcal{Y}} \frac{\exp\{\boldsymbol{\theta}_0 u(y)\}}{\sum_{y \in \mathcal{Y}} \exp\{\boldsymbol{\theta}_0 u(y)\}} \exp\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0) u(y)\} \right] \\
&= \log \left[\mathbb{E}_{\boldsymbol{\theta}_0} \left\{ e^{(\boldsymbol{\theta} - \boldsymbol{\theta}_0) u(Y)} \right\} \right] \\
&\simeq (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbb{E}_{\boldsymbol{\theta}_0} \{u(Y)\} + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbb{V}_{\boldsymbol{\theta}_0} \{u(Y)\} (\boldsymbol{\theta} - \boldsymbol{\theta}_0). \quad (3.4)
\end{aligned}$$

Then the alternative approximation is given by

$$\begin{aligned}
\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0) &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0) u(y) - \{\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)\} \\
&\simeq (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \{u(y) - \widehat{\boldsymbol{\mu}}_0\} - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \widehat{\boldsymbol{\Sigma}}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_0), \quad (3.5)
\end{aligned}$$

where

$$\widehat{\boldsymbol{\mu}}_0 = \frac{1}{M}\sum_{k=1}^{M} u(y_k^*),$$

and

$$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{M-1}\sum_{k=1}^{M}\{u(y_k^*) - \widehat{\boldsymbol{\mu}}_0\}\{u(y_k^*) - \widehat{\boldsymbol{\mu}}_0\}^T,$$

are respectively the usual estimator of mean and variance for the simulated sample of networks.

This approximation assumes that if $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are the true mean vector and covariance matrix of $u(Y)$ when $\boldsymbol{\theta}_0$ is the true parameter for the network distribution, and $Z = (\boldsymbol{\theta} - \boldsymbol{\theta}_k)^T u(y)$ is approximately normally distributed with mean $\boldsymbol{\mu} = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{\mu}_0$ and variance $\sigma^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, then $\exp(Z)$ is log-normally distributed and $\log \mathbb{E}_{\boldsymbol{\theta}}(\exp(Z)) = \boldsymbol{\mu} + \sigma^2/2$.

3.2 Monte Carlo BFGS Algorithm

Here we provide a further algorithm which is similar in spirit to the steplength algorithm, but designed to achieve a more robust convergence behavior.

The method is implemented using `R` and it is based on the procedures made available in the `R` suite package `statnet` (Handcock et al., 2008; Hunter et al., 2008b).

The algorithm that we propose is essentially a sort of BFGS algorithm (see e.g Fletcher, 1980; Dennis and Schnabel, 1996) based on Monte Carlo simulation. The essential theory is as follows. A Newton-Raphson step to maximize a function $f(x)$, with x multidimensional variable, is

$$x_{n+1} = x_n - [\nabla^2 f(x_n)]^{-1} \nabla f(x_n),$$

where $\nabla f(x_n)$ and $\nabla^2 f(x_n)$ represent the gradient vector and Hessian matrix, i.e the set of first and second order of partial derivatives of $f(x)$ evaluated on the point x_n.

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is a quasi-Newton method for solving nonlinear optimization problems useful in the case in which the Hessian matrix of $f(x)$ is difficult to be evaluated directly.

The BFGS algorithm updates, at the step k, the new value x_{k+1} as

$$x_{k+1} = x_k + \boldsymbol{\alpha}_k \boldsymbol{\Delta}_k,$$

along a direction $\boldsymbol{\Delta}_k$ such that

$$J_k \boldsymbol{\Delta}_k = -\nabla f(x_k). \tag{3.6}$$

The matrix J_k is an approximation to the Hessian matrix iteratively updated at each stage. A line search is requested to determine an acceptable value for the step size $\boldsymbol{\alpha}_k$ in the direction Δ_k.

Setting $w_k = \nabla f(x_k) - \nabla f(x_{k+1})$ and $s_k = \boldsymbol{\alpha}_k \Delta_k$, the approximate Hessian J_k is updated by

$$J_{k+1} = J_k + \frac{w_k w_k^T}{w_k^T s_k} - \frac{J_k s_k s_k^T J_k^T}{s_k^T J_k s_k}. \tag{3.7}$$

Back to the maximization of the log-likelihood of an ERGM given by (3.2), we have that

$$\nabla \ell(\boldsymbol{\theta}) = u(y) - \nabla \psi(\boldsymbol{\theta}) \tag{3.8}$$

and

$$\nabla^2 \ell(\boldsymbol{\theta}) = -\nabla^2 \psi(\boldsymbol{\theta}), \tag{3.9}$$

both involving $\psi(\boldsymbol{\theta})$ that can not be computed in explicit form.

From likelihood theory for exponential families (Barndorff-Nielsen, 1978) we know that the gradient of $\psi(\boldsymbol{\theta})$ is

$$\nabla \psi(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\{u(Y, x)\},$$

and the expected Fisher information is

$$i(\boldsymbol{\theta}) = \mathbb{E}\{-\nabla^2 \ell(\boldsymbol{\theta})\} = \mathbb{V}_{\boldsymbol{\theta}}\{u(Y, x)\}.$$

So the solution of (3.8) is the MLE that solves

$$\hat{\theta}: \quad \mathbb{E}_{\boldsymbol{\theta}}\{u(Y)\} = u(y).$$

Therefore, a BFGS algorithm for maximizing $\ell(\boldsymbol{\theta})$ has the following structure:

1. Choose a starting point $\boldsymbol{\theta}_0$, and simulate a sample $y_1^*, \ldots, y_M^*$ of networks from model (3.1) with this parameter value. A natural choice for $\boldsymbol{\theta}_0$ is given by MPLE $\widehat{\boldsymbol{\theta}}_P$ (2.29). A safer alternative, in case of near degeneracy at $\hat{\boldsymbol{\theta}}_P$ (Handcock et al., 2003), is to start from the null point $\boldsymbol{\theta}_0 = 0$ that corresponds to a binomial random graph.

2. Set $J_0 = -\nabla^2 L_p(\boldsymbol{\theta}_0; y)$ equal to minus the Hessian matrix from the pseudo-likelihood (2.29) based on the univariate conditional distribution of each y_{ij} evaluated at $\boldsymbol{\theta}_0$. This is inspired from the principle of maximization by parts (Song et al., 2005), that is a method that tries to compute the MLE using a pseudo-likelihood estimator.

3. Compute the sample mean vector $\widehat{\boldsymbol{\mu}}_0$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}_0$ of the sufficient network statistics $u(y_1^*), \ldots, u(y_M^*)$ for the simulated networks, together with the related convergence t-ratios t_0 (Snijders, 2002)

$$t_{0j} = \frac{u_j(y) - \widehat{\mu}_{0j}}{\widehat{\sigma}_{0j}}, \tag{3.10}$$

with $\widehat{\sigma}_{0j}^2 = (\widehat{\boldsymbol{\Sigma}}_0)_{jj}$, $j = 1, \ldots, d$.

4. Obtain the direction of the update, using equations (3.6) and (3.8)

$$\boldsymbol{\Delta}_0 = J_0^{-1}\,\{u(y) - \widehat{\boldsymbol{\mu}}_0\},$$

where $\nabla\psi(\boldsymbol{\theta}) \approx \widehat{\boldsymbol{\mu}}_0$. At the first step of the algorithm, this is actually a Monte Carlo approximation to the linearized direction of the update of the maximization by parts algorithm.

5. Compute the steplength $\boldsymbol{\alpha}_0$ as

$$\underset{\boldsymbol{\alpha}_0}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}_0 + \boldsymbol{\alpha}_0\Delta_0)$$

Exploiting the approximation (3.5) evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \boldsymbol{\alpha}_0\Delta_0$, the problem reduces to maximize

$$\boldsymbol{\alpha}_0\,\boldsymbol{\Delta}_0^T\,\{u(y) - \widehat{\boldsymbol{\mu}}_0\} - \frac{1}{2}\,\boldsymbol{\alpha}_0^2\boldsymbol{\Delta}_0^T\,\widehat{\boldsymbol{\Sigma}}_0\,\boldsymbol{\Delta}_0,$$

with respect to $\boldsymbol{\alpha}_0$. The maximum is at

$$\boldsymbol{\alpha}_0 = \frac{\boldsymbol{\Delta}_0^T\,\{u(y) - \widehat{\boldsymbol{\mu}}_0\}}{\boldsymbol{\Delta}_0^T\,\widehat{\boldsymbol{\Sigma}}_0\,\boldsymbol{\Delta}_0}\,, \tag{3.11}$$

which is then restricted to lie in $[0, 1]$.

6. The proposal for the first update is then

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \boldsymbol{\alpha}_0\,\boldsymbol{\Delta}_0.$$

The next thing to do should be to update J_0 to J_1 respect the formula (3.7). However, near degeneracy may occur at $\boldsymbol{\theta}_1$. Hence, we incorporate some sort of back-tracking mechanism on the algorithm, which is maintained for a few iterations. The solution proposed consists in generating $y_1^*, \ldots, y_M^*$ from (3.1) at $\boldsymbol{\theta}_1$, and then obtaining the sample statistics $\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\Sigma}}_1$. In case $\widehat{\boldsymbol{\Sigma}}_1$ is not of full rank, step-halving is implemented i.e. $\boldsymbol{\theta}_1$ is re-computed after halving the steplength $\boldsymbol{\alpha}_0$. The same operation is repeated until the convergence t-ratios t_1 is an improvement over t_0, and in any case up to a fixed number of times.

7. After moving from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_1$, the upgraded Jacobian J_1 is obtained from J_0 by the BFGS formula (3.7). In case the BFGS formula gives a singular Jacobian, set J_1 equal to minus the Hessian matrix from the pseudo-likelihood evaluated at $\boldsymbol{\theta}_1$ (as in Step 2). The algorithm is ready for another iteration from Step 4.

The algorithm is stopped using a convergence criterion e.g. when the absolute values of convergence t-ratios are all smaller than a certain threshold (say 0.05).

The crucial point is to incorporate some escaping mechanisms from parameter values associated to near degeneracy or instability, as done at Steps 6 and 7, that in the previously existing methods sometimes drove to poor or no convergent results.

3.3 Missing Data

The basic structure of the algorithm could be subject to several variations.

The main variation we implemented is represented by a procedure to estimate the model parameters in presence of missing data.

As in any other field of statistical analysis, the presence of missing data on a network can be the result of error of the sampling procedure. But in SNA missing data can be also the result of the sample design (Thompson et al., 1996; Handcock and Gile, 2010).

To this goal, suppose that the adjacency matrix $Y = y$ is partially observed so it is possible to split the data on a subset $Y^{obs} = y^{obs}$ of observed links and on a subset $Y^{miss} = y^{miss}$ of missing links.

In this case, our procedure implements the so-called *face-valued likelihood* (Handcock and Gile, 2010). In this procedure the inference is based only on the observed subset of the data

$$L(\boldsymbol{\theta}|Y^{obs} = y^{obs}) \quad \propto \sum_{v:y^{obs}+v\in\mathcal{Y}} P_\theta(Y = y^{obs} + v), \tag{3.12}$$

where v is one of the possible combinations of the missing part Y^{miss} of the network, and the expression $y^{obs} + v$ correspond to the event $(Y^{obs} = y^{obs} \cap Y^{miss} = v)$.

Note that the conditional distribution of Y^{miss} given Y^{obs} is (Handcock and Gile, 2010)

$$P_\theta(Y^{miss} = v|Y^{obs} = y^{obs}) = \exp\{\theta^T u(y^{obs} + v) - \psi(\theta|y^{obs})\} \qquad y \in \mathcal{Y}(y^{obs})$$

with

$$\mathcal{Y}(y^{obs}) = \{v : y^{obs} + v \in \mathcal{Y}\},$$

and

$$\psi(\theta|y^{obs}) = \log \sum_{v\in\mathcal{Y}(y^{obs})} \exp\{\theta^T u(y^{obs} + v)\}.$$

This result can be used to sample from the conditional distribution, and it is implemented in the `ergm` package.

The face-valued log-likelihood function is given by

$$\ell(\boldsymbol{\theta}|y^{obs}) \quad \propto \quad \psi(\boldsymbol{\theta}|y^{obs}) - \psi(\boldsymbol{\theta}). \tag{3.13}$$

Its gradient and Hessian matrix are respectively

$$\begin{aligned} \nabla\ell(\boldsymbol{\theta}|y^{obs}) &= \nabla\psi(\boldsymbol{\theta}|y^{obs}) - \nabla\psi(\boldsymbol{\theta}) \\ &= \mathbb{E}_{\boldsymbol{\theta}}\{u(Y)|Y^{obs}\} - \mathbb{E}_{\boldsymbol{\theta}}\{u(Y)\}, \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} \nabla^2\ell(\boldsymbol{\theta}|y^{obs}) &= \nabla^2\psi(\boldsymbol{\theta}|y^{obs}) - \nabla^2\psi(\boldsymbol{\theta})) \\ &= \mathbb{V}_{\boldsymbol{\theta}}\{u(Y)|Y^{obs}\} - \mathbb{V}_{\boldsymbol{\theta}}\{u(Y)\}. \end{aligned} \tag{3.15}$$

It is fundamental to note that the results for face-valued likelihood hold only in the cases in which the network data are missing at random (Rubin, 1976) or in general when there is amenability of the sample design (Handcock and Gile, 2010), namely in the cases in which the mechanism to generate missing data in the network is ignorable.

It also holds that

$$\begin{aligned} \ell(\boldsymbol{\theta}|y^{obs}) - \ell(\boldsymbol{\theta}_0|y^{obs}) &= \psi(\boldsymbol{\theta}|y^{obs}) - \psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0|y^{obs}) + \psi(\boldsymbol{\theta}_0) \\ &= \left\{\psi(\boldsymbol{\theta}|y^{obs}) - \psi(\boldsymbol{\theta}_0|y^{obs})\right\} - \left\{\psi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta}_0)\right\}. \end{aligned}$$

For $\psi(\boldsymbol{\theta}|y^{obs}) - \psi(\boldsymbol{\theta}_0|y^{obs})$ holds a result similar to (3.4)

$$\begin{aligned} \psi(\boldsymbol{\theta}|y^{obs}) - \psi(\boldsymbol{\theta}_0|y^{obs}) &\simeq (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbb{E}_{\boldsymbol{\theta}_0}\left\{u(Y)|y^{obs}\right\} + \\ &\quad + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbb{V}_{\boldsymbol{\theta}_0}\left\{u(Y)|y^{obs}\right\}(\boldsymbol{\theta} - \boldsymbol{\theta}_0). \end{aligned} \tag{3.16}$$

Then

$$\begin{aligned} \ell(\boldsymbol{\theta}|y^{obs}) - \ell(\boldsymbol{\theta}_0|y^{obs}) &\simeq (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T (\widehat{\boldsymbol{\mu}}^c - \widehat{\boldsymbol{\mu}}) + \\ &\quad - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left(\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^c\right)(\boldsymbol{\theta} - \boldsymbol{\theta}_0), \end{aligned} \tag{3.17}$$

in which $\widehat{\boldsymbol{\mu}}^c$ and $\widehat{\boldsymbol{\Sigma}}^c$ are respectively the mean vector and the variance matrix of an additional stage in which the network statistics $\{u(y_1^c), \ldots, u(y_M^c)\}$ are simulated conditional on the observed data.

In order to maximize the face-valued likelihood (3.13), the steps of the algorithm are changed as follows:

- Whenever required, an additional stage in which the networks are simulated conditionally on the observed data is added. And then their means and covariance matrices for the network statistics are computed.

- At Step 4, the value of $\boldsymbol{\Delta}_0$ becomes

$$\boldsymbol{\Delta}_0 = J_0^{-1}\{\widehat{\boldsymbol{\mu}}_0^c - \widehat{\boldsymbol{\mu}}_0\}.$$

- At Step 5, Formula (3.17) is used, instead of (3.5), to compute the steplength $\boldsymbol{\alpha}_0$ such that

$$\underset{\boldsymbol{\alpha}_0}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}_0 + \boldsymbol{\alpha}_0\Delta_0|y^{obs}).$$

  In this case $\boldsymbol{\alpha}_0$ is maximized by

$$\boldsymbol{\alpha}_0 = \frac{\boldsymbol{\Delta}_0^T\{\widehat{\boldsymbol{\mu}}_{kj}^c - \widehat{\boldsymbol{\mu}}_{kj}\}}{\boldsymbol{\Delta}_0^T\left\{\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^c\right\}\boldsymbol{\Delta}_0}, \tag{3.18}$$

  restricted to lie in $[0, 1]$.

- The generic t-ratio convergence criterion (3.10) becomes

$$t_{kj} = \frac{\widehat{\boldsymbol{\mu}}_{kj}^c - \widehat{\boldsymbol{\mu}}_{kj}}{\widehat{\sigma}_{kj} - \widehat{\sigma}_{kj}^c},$$

  with $\widehat{\sigma}_{kj} = \sqrt{(\widehat{\boldsymbol{\Sigma}}_k)_{jj}}$, $\widehat{\sigma}_{kj}^c = \sqrt{(\widehat{\boldsymbol{\Sigma}}_k^c)_{jj}}$, and $j = 1, \ldots, d$.

While in the complete observed case the procedure converges if the means of the statistics of the simulated networks are sufficient close to the observed networks statistics, here the convergence is reached if the constrained mean and unconstrained mean of the statistics of the two simulated sample networks are close enough.

3.4 Examples on Real Data

The algorithm presented in the previous section has been tested with several publicly available data sets, for a wide choice of sufficient statistics. The results for MPLE, Steplength algorithm (MC-MLE) and Monte Carlo BFGS (MC-BFGS) algorithms are compared.

We took advantage of the highly efficient implementation of MCMC methods to simulate from model (3.1) made available in the `R` package `ergm v2.4-3` (Hunter et al., 2008b).

3.4.1 Ecoli and Kapferer Data

The first two examples that we present are both considered in Hummel et al. (2011).

The model for `ecoli2` data (§2.3.2) includes the degree terms until the order five and the new specification term GWDegree to cover the remaining degrees of order larger that five.

The model for `kapferer` data (§2.3.3) has only new specification terms, in fact it includes the GWESP and GWDSP statistics.

As we can see in Table 3.1 and Table 3.2, in both cases the estimated values for MC-MLE and MC-BFGS are very close. This also because the models do not present neither near-degeneracy nor instability problems.

Notice that the standard errors for the MPLE are those obtained from the maximization of the pseudo-likelihood (2.29) through logistic regression, and they are not adjusted using estimating equations theory. This is relevant with what commonly done by practitioners, and it will be close also in the other examples of this chapter.

The good convergence properties of the MC-BFGS algorithm to actual MLE can be checked by the scatter plots in Figure 3.1 and in Figure 3.3, that show how the algorithm stops effectively when the sample means of the simulated network statistics reached the sufficient statistics of the observed networks. The plots in Figure 3.2 and in Figure 3.4 show the differences between the mean parametrization values in the algorithm iterations and the observed network statistics. In other words the lines represent the estimated log-likelihood gradients, given by the means of the network statistics for the simulated samples minus the observed network statistics. The plots confirm the good convergence property of MC-BFGS algorithm but also that MC-BFGS seems to converge slower than the Steplength algorithm. This is likely due to the different convergence criteria of the algorithms.

Table 3.1: Parameter estimates (s.e.) for `ecoli2` data.

Parameter	MPLE	MC-MLE	MC-BFGS
$\boldsymbol{\theta}_1$: edges	-5.35 (0.08)	-5.07 (0.04)	-5.07 (0.05)
$\boldsymbol{\theta}_2$: degree2	-2.58 (0.09)	-1.47 (0.14)	-1.46 (0.14)
$\boldsymbol{\theta}_3$: degree3	-3.06 (0.12)	-2.36 (0.19)	-2.35 (0.20)
$\boldsymbol{\theta}_4$: degree4	-2.39 (0.13)	-2.30 (0.23)	-2.30 (0.23)
$\boldsymbol{\theta}_5$: degree5	-1.85 (0.11)	-2.92 (0.42)	-2.91 (0.42)
$\boldsymbol{\theta}_6$: GWDegree	8.13 (0.33)	1.86 (0.27)	1.83 (0.32)

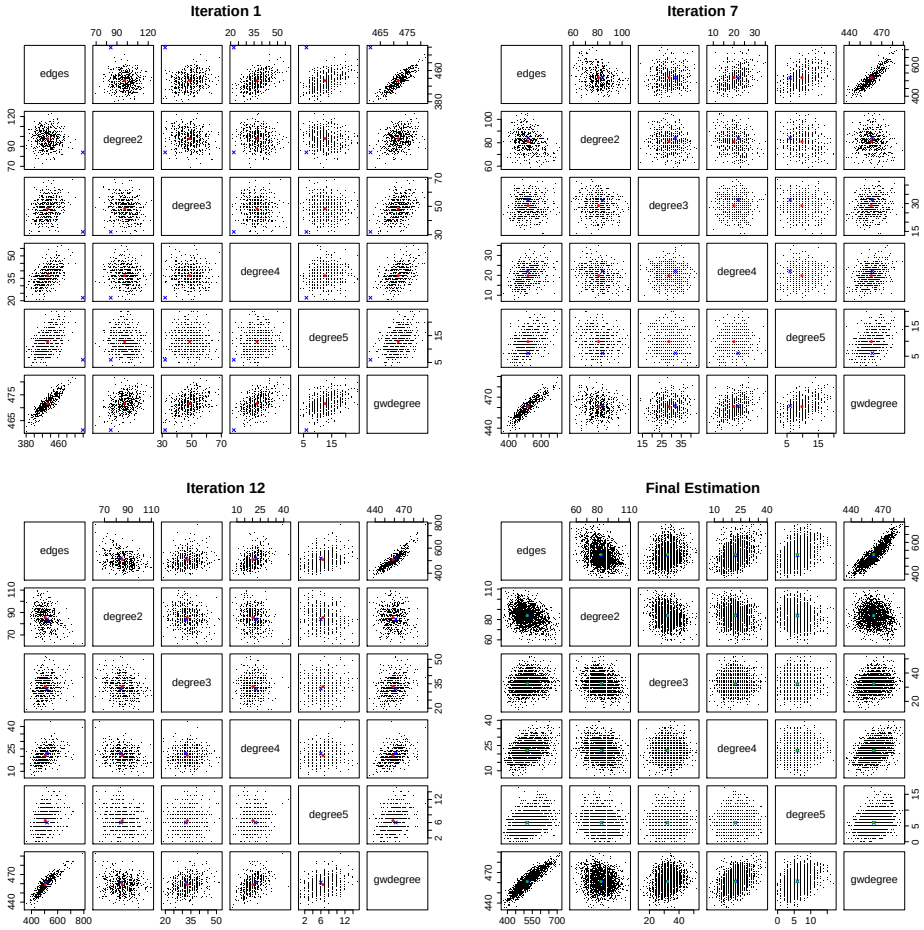

Figure 3.1: Some iteration steps of the MC-BFGS algorithm for `ecoli2` data. The values of the statistics for the networks simulated at each step are plotted, together with their sample mean (● and ● in the final step) and the observed values (×)

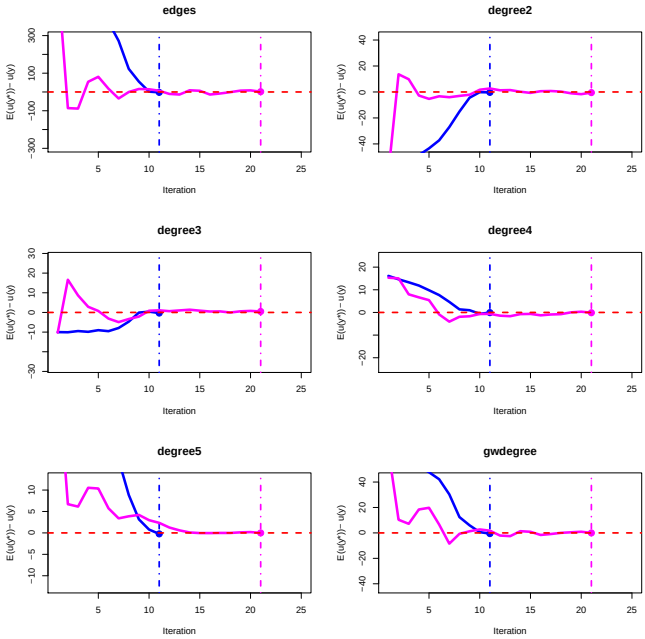

Figure 3.2: `ecoli2` data, differences between the coefficients expressed in mean parameterization in the iterations of MC-MLE (blue) and MC-BFGS (violet) algorithms and the observed sufficient statistics.

Table 3.2: Parameter estimates (s.e.) for `kapferer` data.

Parameter	MPLE	MC-MLE	MC-BFGS
$\boldsymbol{\theta}_1$: edges	-2.26 (0.24)	-3.05 (0.47)	-3.08 (0.58)
$\boldsymbol{\theta}_2$: GWESP	0.93 (0.14)	1.45 (0.33)	1.45 (0.32)
$\boldsymbol{\theta}_3$: GWDSP	-0.10 (0.04)	-0.12 (0.01)	-0.12 (0.06)

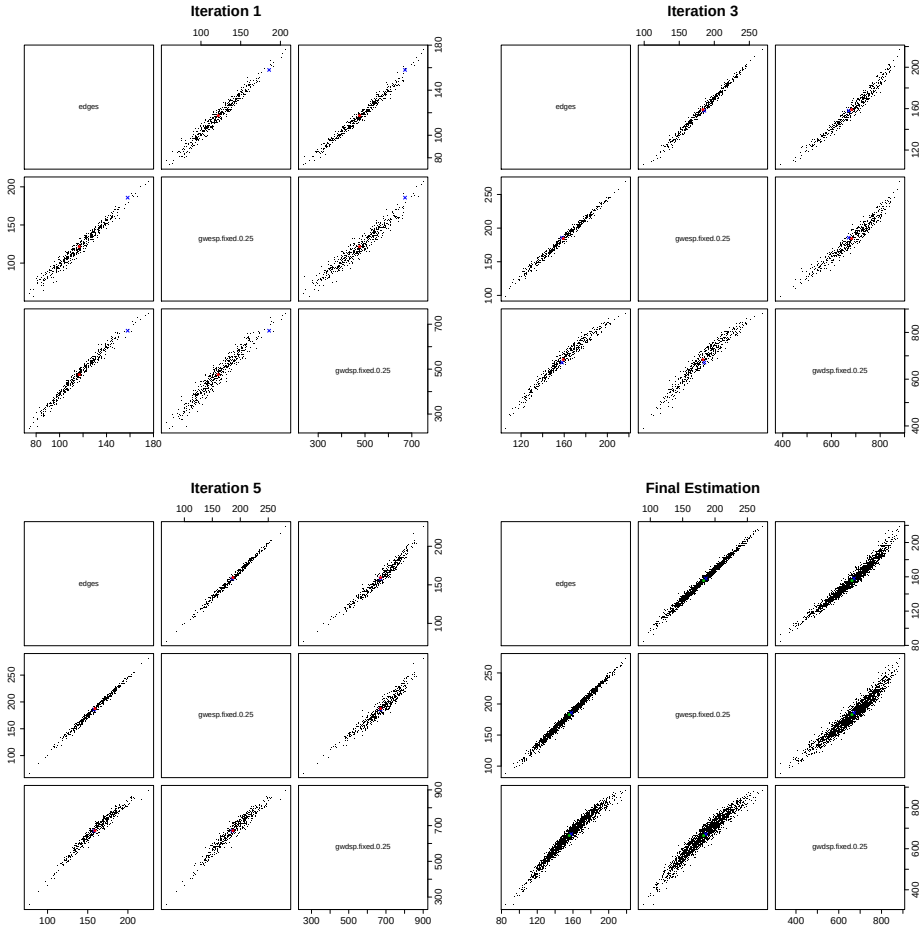

Figure 3.3: Some iteration steps of the MC-BFGS algorithm for `kapferer` data. The values of the statistics for the networks simulated at each step are plotted, together with their sample mean (● and ● in the final step) and the observed values (×)

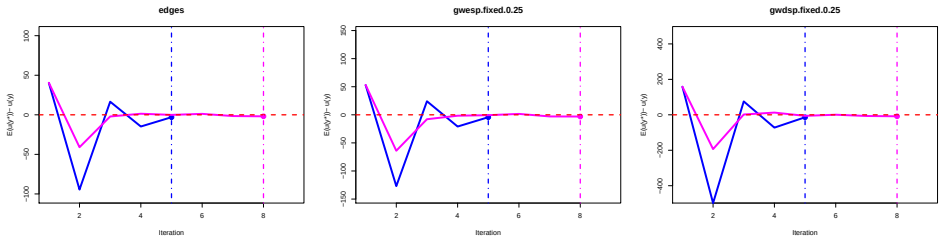

Figure 3.4: `kapferer` data, differences between the coefficients expressed in mean parameterization in the iterations of MC-MLE (blue) and MC-BFGS (violet) algorithms and the observed sufficient statistics.

3.4.2 Molecule

For the `molecule` data, used in Caimo and Friel (2011), we considered three different models.

A Markov random graph model (p^*) was fitted, and the results are reported in Table 3.3 (Model 1). Here the MPLE corresponds to near degeneracy, and none of MC-MLE algorithms implemented in `ergm` converged. The MC-BFGS instead converged to an estimate similar to the Bayesian results reported in Caimo and Friel (2011).

At any rate, there are surely problems with Model 1, as minor changes to the reported estimate may lead to near degeneracy. Figure 3.5 reports the differences between the observed sufficient statistics and the mean parametrization of the coefficients in the algorithm steps. The values of MC-MLE (blue line) were truncated after 200 algorithm iterations, this because it is clear that it was trapped in a, maybe infinite, loop. MC-BFGS (violet line) reached the convergence after 20 iterations. This result is mostly due to the effects of the back-tracking mechanisms included in the procedure. In the plots in Figure 3.6, we can see how the MC-BFGS algorithm is already close to convergence after the iteration 8, and the next iterations refine the convergence due to a small level of the threshold for the convergence t-ratios (3.10).

Table 3.3: Parameter estimates (s.e.) for `Molecule` data.

	Model 1			Model 2			Model 3		
Parameter	MPLE	MC-BFGS	MC-MLE	MPLE	MC-BFGS	MC-MLE	MPLE	MC-BFGS	MC-MLE
$\boldsymbol{\theta}_1$ edges	5.08 (NA)	2.98 (3.58)	NA	-1.05 (1.27)	-0.91 (2.02)	-0.96 (1.96)	-6.06 (1.07)	-6.07 (1.49)	-6.01 (1.42)
$\boldsymbol{\theta}_2$ 2-stars	-2.02 (NA)	-1.31 (1.23)	NA	-0.53 (0.24)	-0.57 (0.41)	-0.56 (0.39)	–	–	–
$\boldsymbol{\theta}_3$ 3-stars	0.52 (NA)	0.19 (0.59)	NA	–	–	–	–	–	–
$\boldsymbol{\theta}_4$ triangles	1.60 (NA)	1.71 (0.50)	NA	0.15 (0.48)	0.20 (0.68)	0.20 (0.69)	–	–	–
$\boldsymbol{\theta}_5$: GWDegree (0.8)	–	–	–	–	–	–	4.50 (1.58)	4.29 (2.43)	4.19 (2.37)
$\boldsymbol{\theta}_6$: GWESP (0.8)	–	–	–	–	–	–	0.25 (0.19)	0.21 (0.29)	0.18 (0.10)
$\boldsymbol{\theta}_7$: Main eff. Atomic type	–	–	–	3.12 (0.75)	3.16 (0.78)	3.16 (0.79)	2.70 (0.75)	2.94 (0.74)	2.96 (0.19)

Things are surely better by inserting a main effect term for atomic type node attribute in place of the 3-stars term (Model 2), and using the new ERGM specifications (Hunter, 2007) (Model 3). The results of these two

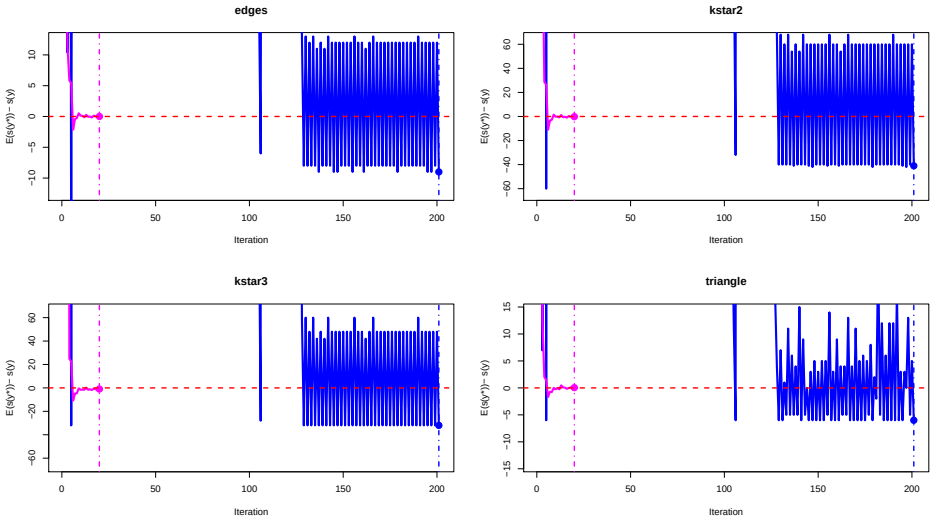

Figure 3.5: Model 1 of `Molecule` data, differences between the coefficients expressed in mean parameterization in the iterations of MC-MLE (blue) and MC-BFGS (violet) algorithms and the observed sufficient statistics.

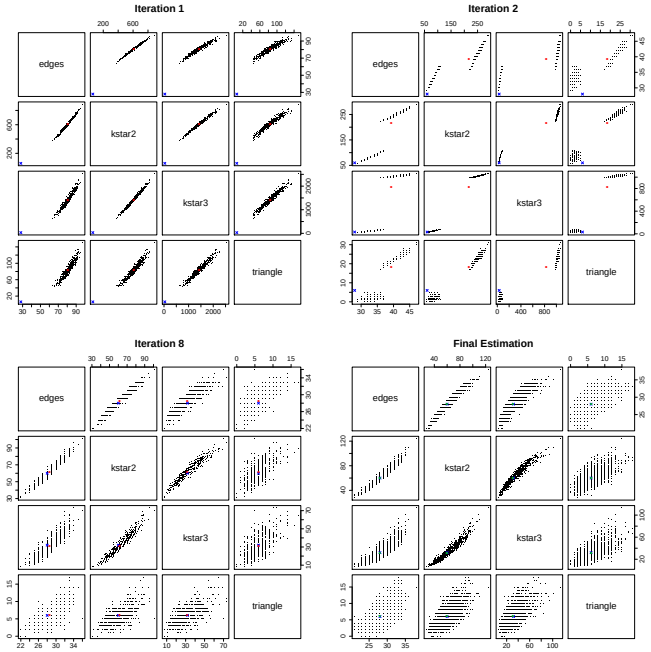

Figure 3.6: Some iterations of MC-BFGS algorithm for Model 1 of `Molecule data`. The values of the statistics for the networks simulated at each step are plotted, together with their sample mean (• and • in the final step) and the observed values (×)

further models are also reported in Table 3.3; the similarities between the two MLE estimates are apparent also observing Figures 3.7 and 3.8.

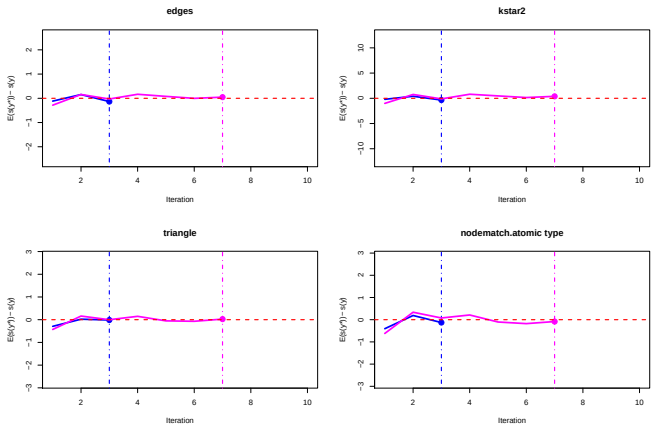

Figure 3.7: Model 2 of `Molecule` data, differences between the coefficients expressed in mean parameterization in the iterations of MC-MLE (blue) and MC-BFGS (violet) algorithms and the observed sufficient statistics.

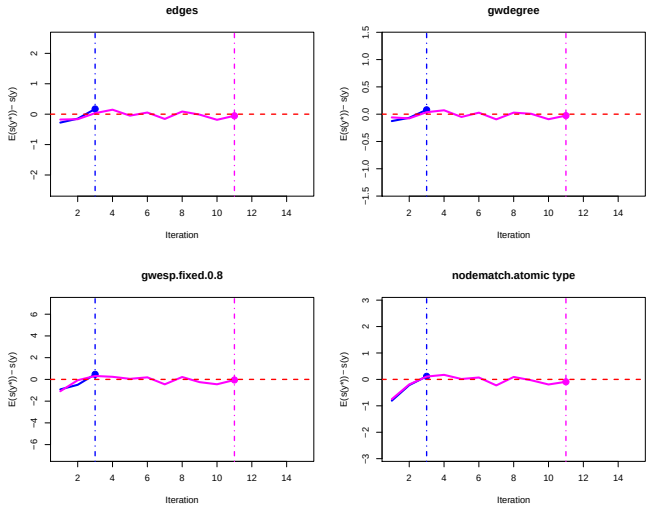

Figure 3.8: Model 3 of `Molecule` data, differences between the coefficients expressed in mean parameterization in the iterations of MC-MLE (blue) and MC-BFGS (violet) algorithms and the observed sufficient statistics.

3.5 Simulation Studies Based on Lazega's Lawyers Data

In this section, first we analyze the `Lawyers` data (Lazega and Pattison, 1999) for two models. Second, we replicate and extend the simulation study contained in Van Duijn et al. (2009).

Data Analysis

For the data analysis we considered a Markov random graph (p^*) model (Model 1) and a model including the new specification term GWESP. Both the models include also the same actor attribute terms, in according with the literature (Hunter and Handcock, 2006; Van Duijn et al., 2009; Kolaczyk, 2009).

In the first column of Table 3.4 we can see that the MC-MLE algorithm fails for Model 1. From Figure 3.9 we see that the trajectory of the terms diverged from the observed statistics, it entered in an infinite loop and the procedure was stopped after 100 iterations. Instead, MC-BFGS converged to the MLE, although slowly and maybe as a result of the back-tracking mechanisms.

For Model 2, there are no apparent differences between the estimates obtained by MC-MLE and MC-BFGS algorithms (Model 2 in Table 3.4). In Figure 3.10, we can see how both the algorithms converged quickly, probably because the MPLE is a good starting point. We note that there are, instead, some differences between the standard errors estimated by MC-MLE and MC-BFGS.

Table 3.4: Parameter estimates (s.e.) for `Lawyers` data.

	Model 1			Model 2		
Parameter	MPLE	MC-MLE	MC-BFGS	MPLE	MC-MLE	MC-BFGS
$\boldsymbol{\theta}_1$: edges	-8.10 (1.25)	NA	-6.97 (0.83)	-6.43 (0.81)	-6.47 (0.32)	-6.53 (0.60)
$\boldsymbol{\theta}_2$: 2stars	0.27 (0.11)	NA	0.20 (0.09)	–	–	–
$\boldsymbol{\theta}_3$: 3-stars	-0.02 (0.01)	NA	-0.03 (0.01)	–	–	–
$\boldsymbol{\theta}_4$: triangles	0.30 (0.12)	NA	0.34 (0.12)	–	–	–
$\boldsymbol{\theta}_5$: GWESP (0.7781)	–	–	–	0.90 (0.11)	0.89 (0.14)	0.89 (0.15)
$\boldsymbol{\theta}_6$: main eff. seniority	0.75 (0.37)	NA	1.04 (0.26)	0.88 (0.36)	0.85 (0.03)	0.87 (0.24)
$\boldsymbol{\theta}_7$: main eff. specialty	0.26 (0.19)	NA	0.44 (0.12)	0.38 (0.19)	0.42 (0.03)	0.41 (0.12)
$\boldsymbol{\theta}_8$: homoph. specialty	0.85 (0.26)	NA	0.82 (0.22)	0.72 (0.25)	0.75 (0.05)	0.77 (0.19)
$\boldsymbol{\theta}_9$: homoph. gender	0.70 (0.39)	NA	0.85 (0.27)	0.64 (0.39)	0.71 (0.03)	0.70 (0.25)
$\boldsymbol{\theta}_{10}$: homoph. office	1.58 (0.31)	NA	1.28 (0.24)	1.15 (0.28)	1.15 (0.04)	1.15 (0.20)

Simulation Studies

The simulation studies are designed as following. In all the cases, Model 2 is taken as reference. In the first case 500 networks are simulated exactly from Model 2, in the second case the network transitivity is increased as in

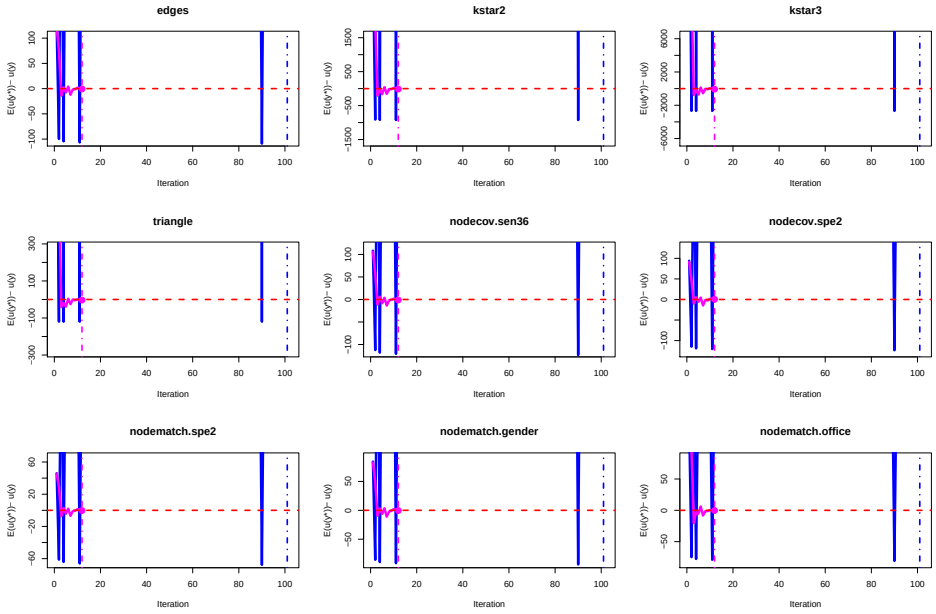

Figure 3.9: Model 1 of `Lawyers` data, differences between the coefficients expressed in mean parameterization in the iterations of MC-MLE (blue) and MC-BFGS (violet) algorithms and the observed sufficient statistics.

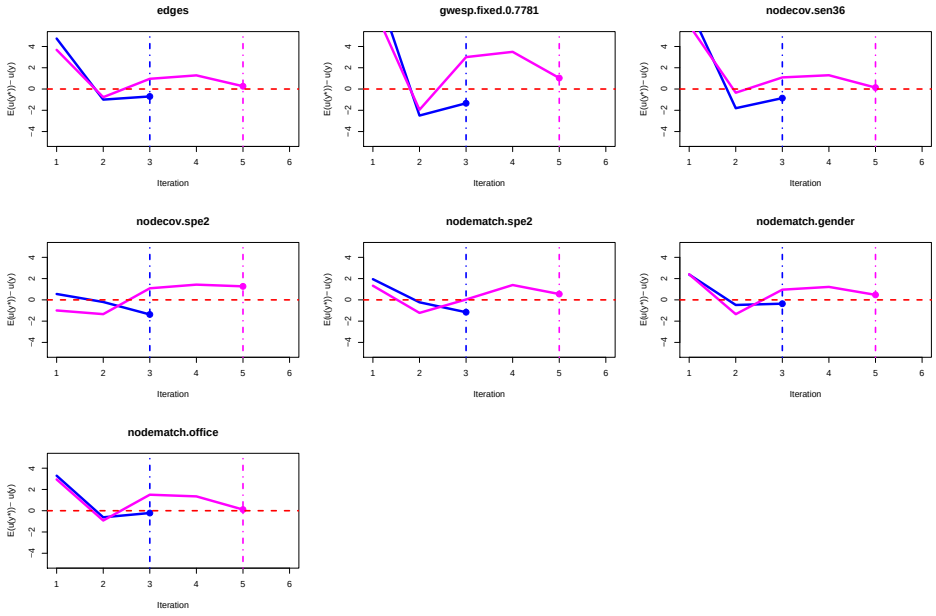

Figure 3.10: Model 2 of `Lawyers` data, differences between the coefficients expressed in mean parameterization in the iterations of MC-MLE (blue) and MC-BFGS (violet) algorithms and the observed sufficient statistics.

Van Duijn et al. (2009) in the 500 simulated networks, whereas in the last case the 500 networks are obtained by setting randomly a 10% of missing data in the networks simulated in the first case.

Table 3.5 reports the results relative to the first simulation case. The values are the mean and the standard deviation of the estimated coefficients for the simulated networks, expressed in natural and mean parametrization. The mean parametrization values are the mean of the statistics on further 200 simulate networks for each estimate.

The results for the first case do not show any substantial difference between the methods in the natural parametrization. However, in the mean parametrization, the estimates based on MC-BFGS look much closer to the mean parameter corresponding to the true parameter values $\boldsymbol{\mu}(\boldsymbol{\theta}_0)$. This is confirmed also in Figure 3.11, where the boxplots relative to the mean parametrization of the estimates obtained by the MC-BFGS method are centered at $\boldsymbol{\mu}(\boldsymbol{\theta}_0)$. Note that bias in the natural parametrization is finite sample bias; accuracy in mean parametrization reflects instead the capability of the method in computing correctly the MLE.

Table 3.5: Mean of parameter estimates (std. deviation), in natural and mean parametrization, for networks simulated from `Lawyers` data, with true parameter values $\boldsymbol{\theta}_0$.

	Natural Parametrization				Mean Parametrization			
Parameter	$\boldsymbol{\theta}_0$	MPLE	MC-MLE	MC-BFGS	$\boldsymbol{\mu}(\boldsymbol{\theta}_0)$	MPLE	MC-MLE	MC-BFGS
$\boldsymbol{\theta}_1$: edges	-6.51	-6.70 (0.66)	-6.63 (0.61)	-6.67 (0.61)	114.40	98.74 (34.29)	120.83 (30.56)	114.32 (16.60)
$\boldsymbol{\theta}_5$: GWEsp (0.7781)	0.90	0.93 (0.20)	0.86 (0.14)	0.85 (0.14)	189.09	160.61 (66.08)	203.96 (72.20)	188.91 (36.67)
$\boldsymbol{\theta}_6$: main eff. seniority	0.85	0.87 (0.27)	0.92 (0.26)	0.94 (0.26)	129.59	111.45 (39.40)	135.99 (31.84)	129.51 (19.52)
$\boldsymbol{\theta}_7$: main eff. specialty	0.41	0.42 (0.16)	0.47 (0.14)	0.47 (0.14)	128.33	109.44 (38.75)	133.69 (26.25)	128.23 (18.31)
$\boldsymbol{\theta}_8$: homoph. specialty	0.76	0.76 (0.22)	0.76 (0.21)	0.76 (0.22)	71.65	62.42 (21.30)	75.01 (15.44)	71.56 (11.00)
$\boldsymbol{\theta}_9$: homoph. gender	0.70	0.75 (0.35)	0.77 (0.30)	0.78 (0.30)	98.52	85.17 (29.88)	104.02 (26.26)	98.44 (15.11)
$\boldsymbol{\theta}_{10}$: homoph. office	1.15	1.18 (0.23)	1.20 (0.21)	1.21 (0.21)	84.33	73.04 (25.24)	87.61 (15.65)	84.28 (12.38)

We draw similar conclusions also observing the results relative to the increased transitivity case in Table 3.6 and in Figure 3.12. In fact, while for the natural parametrization both the algorithms are close to $\boldsymbol{\theta}_0$, in the mean parametrization MC-BFGS is much closer to the true values. A possible explanation to this results could be that MC-BFGS are more stable, as for the convergence of the MC-BFGS algorithm all the convergence statistics must to be under a threshold. This is more restrictive that the convergence criterion of the MC-MLE. This can explain also why the MC-BFGS are at times slower to converge.

For the missing data case, as we said in §3.3, if a method correctly maximizes the face-valued likelihood, the convergence is reached when the conditional and unconditional sample mean of the network statistics are close. As we can see in Table 3.7, only MC-BFGS satisfies this condition. The absolute values of the differences for MPLE and the MC-MLE are really huge, as it can be seen in Figure 3.13. These results are quite surprising for

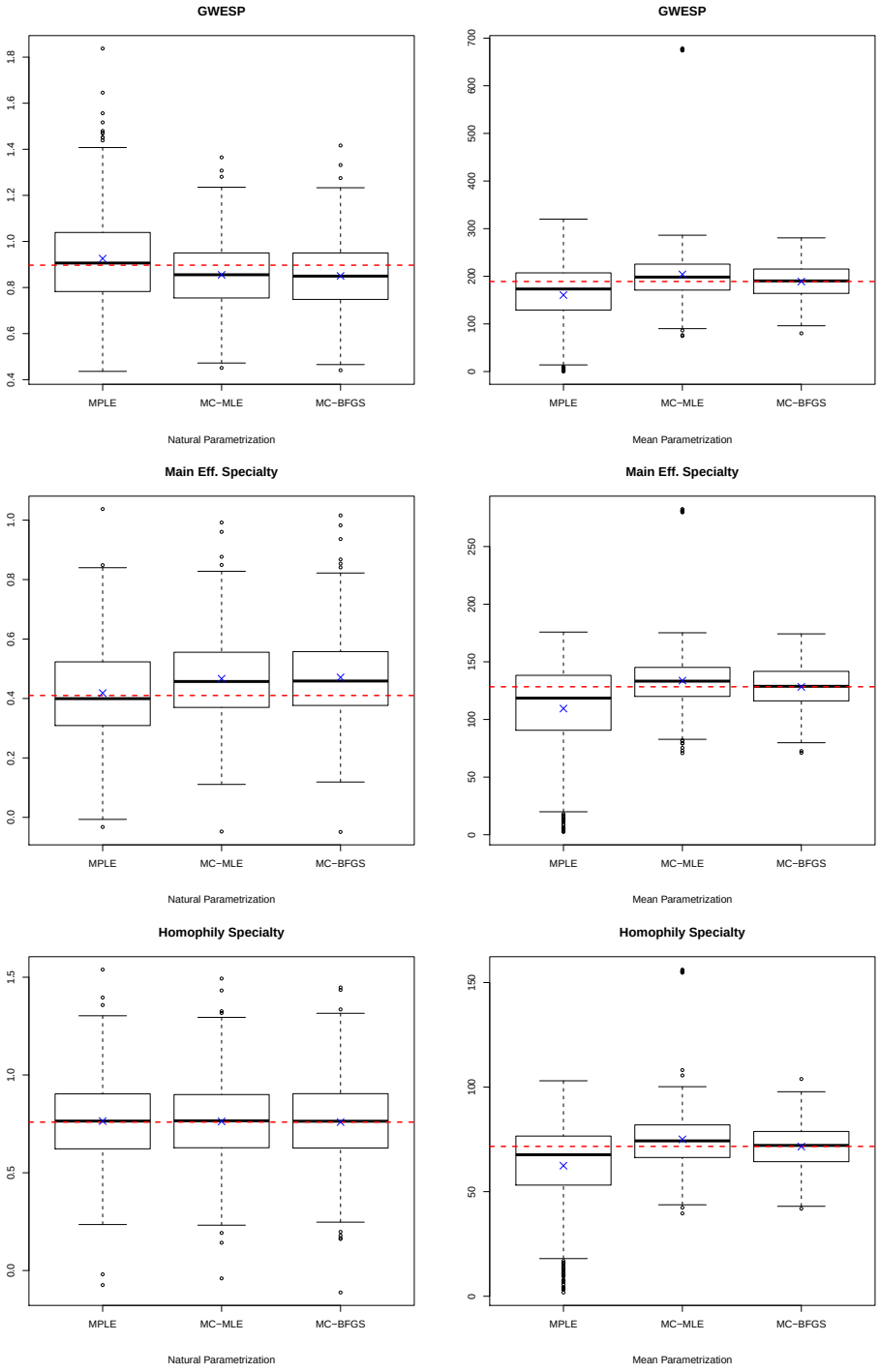

Figure 3.11: Boxplot of the estimates from the simulated networks for three network statistics in the natural and mean parametrization. The means of the estimates are represented by $\times$ and the horizontal lines correspond to the true parameter values.

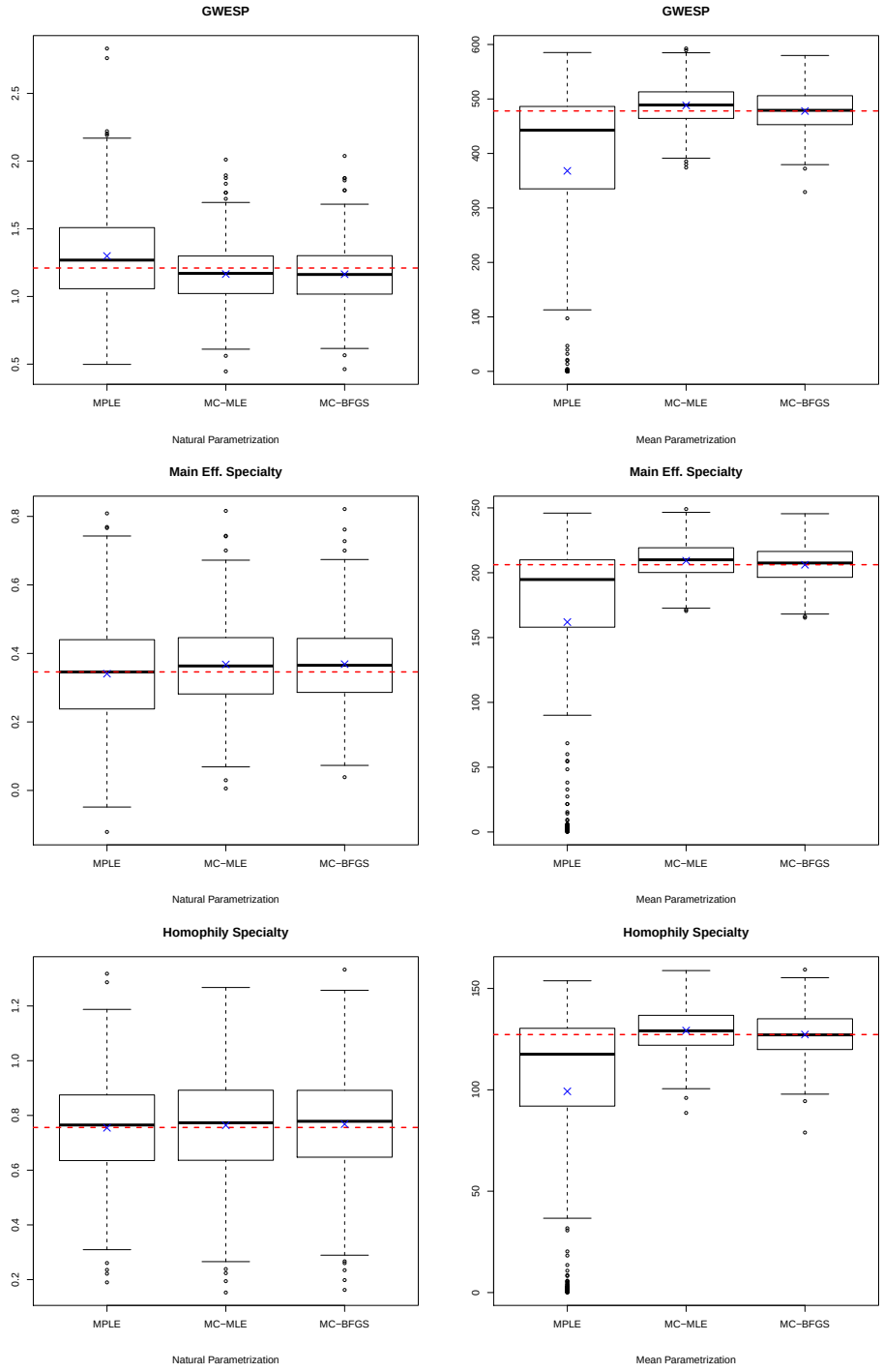

Figure 3.12: Boxplot of the estimates from the simulated networks with increased transitivity for three network statistics in the natural and mean parametrization. The means of the estimates are represented by $\times$ and the horizontal lines correspond to the true parameter values.

Table 3.6: Mean of parameter estimates (std. deviation), in natural and mean parametrization, for networks simulated from `Lawyers` data, with true parameter values $\boldsymbol{\theta}_0$, and increased transitivity.

	Natural Parametrization				Mean Parametrization			
Parameter	$\boldsymbol{\theta}_0$	MPLE	MC-MLE	MC-BFGS	$\boldsymbol{\mu}(\boldsymbol{\theta}_0)$	MPLE	MC-MLE	MC-BFGS
$\boldsymbol{\theta}_1$: edges	-6.96	-7.24 (0.93)	-6.93 (0.66)	-6.98 (0.66)	212.64	164.81 (79.14)	216.48 (13.70)	212.61 (14.34)
$\boldsymbol{\theta}_5$: GWEsp (0.9075)	1.21	1.30 (0.35)	1.17 (0.22)	1.16 (0.22)	478.01	368.03 (181.06)	488.24 (37.62)	477.93 (38.63)
$\boldsymbol{\theta}_6$: main eff. seniority	0.78	0.79 (0.26)	0.85 (0.23)	0.86 (0.23)	236.83	183.24 (88.24)	240.47 (14.56)	236.78 (15.17)
$\boldsymbol{\theta}_7$: main eff. specialty	0.35	0.34 (0.15)	0.37 (0.12)	0.37 (0.12)	206.19	161.94 (76.53)	209.29 (14.18)	206.13 (14.71)
$\boldsymbol{\theta}_8$: homoph. specialty	0.76	0.75 (0.18)	0.76 (0.19)	0.77 (0.19)	127.25	99.25 (47.15)	129.25 (10.44)	127.28 (10.84)
$\boldsymbol{\theta}_9$: homoph. gender	0.66	0.66 (0.28)	0.68 (0.26)	0.70 (0.25)	183.41	141.99 (68.78)	186.26 (12.95)	183.36 (13.41)
$\boldsymbol{\theta}_{10}$: homoph. office	1.08	1.09 (0.19)	1.14 (0.16)	1.13 (0.16)	145.45	113.50 (53.93)	147.84 (9.54)	145.42 (9.48)

MC-MLE that should already use face-valued likelihood in its estimation procedure. A likely explanation for the latter fact is the presence of some bugs in the `ergm` estimation procedure for networks with missing values.

Table 3.7: Mean of parameter estimates (std. deviation), in natural and mean parametrization, for networks simulated from `Lawyers` data, with true parameter values $\boldsymbol{\theta}_0$, and 10% of missing data. For mean parametrization, the values are the absolute values of the difference between the unconditional and conditional means.

	Natural Parametrization				Mean Parametrization		
Parameter	$\boldsymbol{\theta}_0$	MPLE	MC-MLE	MC-BFGS	MPLE	MC-MLE	MC-BFGS
$\boldsymbol{\theta}_1$: edges	-6.51	-6.53 (0.75)	-6.42 (0.67)	-6.67 (0.65)	14.10 (10.46)	9.83 (3.64)	1.03 (0.84)
$\boldsymbol{\theta}_5$: GWEsp (0.7781)	0.90	0.70 (0.18)	0.66 (0.15)	0.84 (0.16)	27.05 (23.06)	29.44 (9.99)	2.36 (1.89)
$\boldsymbol{\theta}_6$: main eff. seniority	0.85	1.03 (0.33)	0.99 (0.29)	0.95 (0.28)	15.92 (12.07)	11.22 (4.20)	1.18 (1.00)
$\boldsymbol{\theta}_7$: main eff. specialty	0.41	0.50 (0.18)	0.50 (0.15)	0.48 (0.15)	11.19 (10.13)	11.04 (4.27)	1.15 (0.86)
$\boldsymbol{\theta}_8$: homoph. specialty	0.76	0.78 (0.24)	0.74 (0.23)	0.76 (0.23)	8.45 (6.51)	6.12 (2.44)	0.64 (0.51)
$\boldsymbol{\theta}_9$: homoph. gender	0.70	0.86 (0.39)	0.81 (0.35)	0.78 (0.32)	12.60 (9.39)	8.40 (3.29)	0.97 (0.75)
$\boldsymbol{\theta}_{10}$: homoph. office	1.15	1.28 (0.26)	1.23 (0.24)	1.21 (0.22)	10.12 (7.80)	7.27 (2.82)	0.80 (0.63)

An other kind of comparison between the three methods can be done by comparing the standard deviation of the estimated parameters in the natural parameterization with the mean standard errors estimated by the methods. If a method estimates the standard error in the right way, this two values should be very close. But the results in Tables 3.8, 3.9 and 3.10 show how MPLE and MC-MLE often tend to estimate badly the standard errors. The MC-MLE, in particular, exhibits the tendency to underestimate the estimation variability.

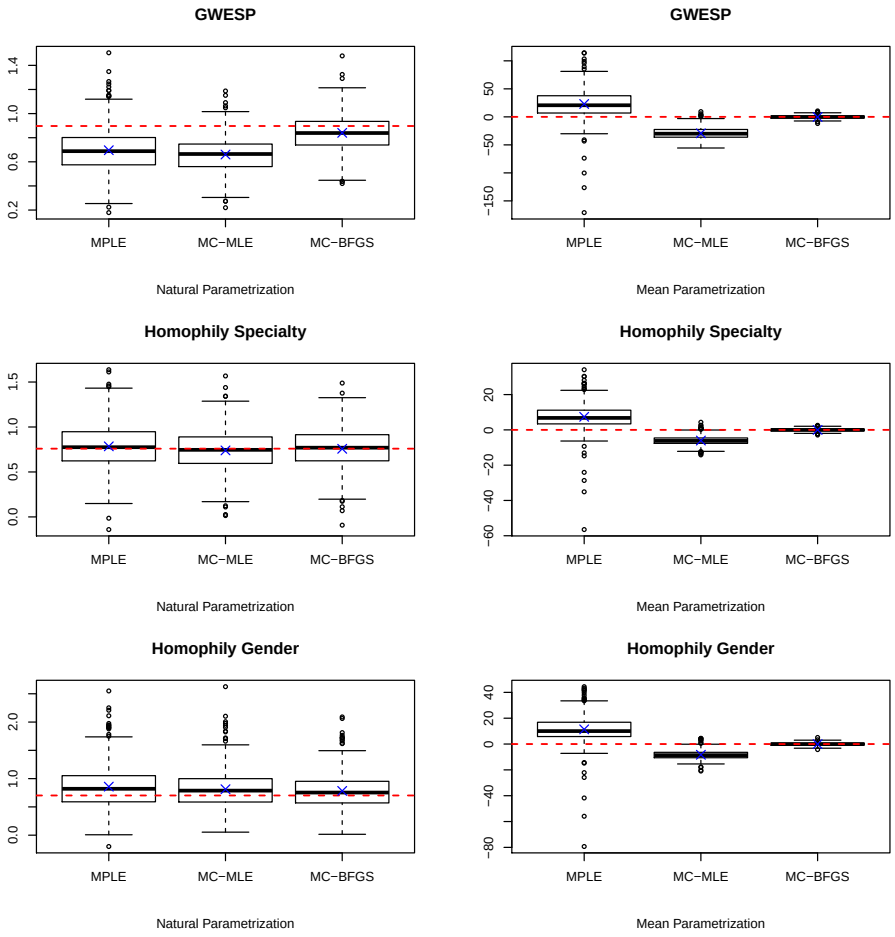

Figure 3.13: Boxplot of the estimates from the simulated networks for three network statistics with with 10% of missing data. In the mean parameterization the values plotted are the differences between the means of the network statistics for the unconstrained and constrained simulated sample networks. The means of the estimates, and the means of differences, are represented by $\times$ and the horizontal lines correspond to the true parameter values.

Table 3.8: Standard deviation of the estimates (s.d) and mean of simulated standard errors (s.e) for the parameter estimates of Table 3.5

Parameter	MPLE		MC-MLE		MC-BFGS	
	s.d	s.e	s.d	s.e	s.d	s.e
$\boldsymbol{\theta}_1$: edges	0.66	0.86	0.61	0.31	0.61	0.62
$\boldsymbol{\theta}_5$: GWEsp (0.7781)	0.20	0.12	0.15	0.14	0.15	0.15
$\boldsymbol{\theta}_6$: main eff. seniority	0.28	0.38	0.26	0.03	0.27	0.26
$\boldsymbol{\theta}_7$: main eff. specialty	0.16	0.20	0.14	0.03	0.14	0.13
$\boldsymbol{\theta}_8$: homoph. specialty	0.22	0.26	0.21	0.05	0.22	0.21
$\boldsymbol{\theta}_9$: homoph. gender	0.34	0.41	0.30	0.04	0.30	0.28
$\boldsymbol{\theta}_{10}$: homoph. office	0.23	0.29	0.21	0.04	0.21	0.21

Table 3.9: Standard deviation of the estimates (s.d) and mean of simulated standard errors (s.e) for the parameter estimates of Table 3.6

Parameter	MPLE		MC-MLE		MC-BFGS	
	s.d	s.e	s.d	s.e	s.d	s.e
$\boldsymbol{\theta}_1$: edges	0.93	0.76	0.66	0.52	0.66	0.62
$\boldsymbol{\theta}_5$: GWEsp (0.7781)	0.35	0.18	0.22	0.19	0.22	0.21
$\boldsymbol{\theta}_6$: main eff. seniority	0.26	0.30	0.23	0.04	0.23	0.23
$\boldsymbol{\theta}_7$: main eff. specialty	0.15	0.15	0.12	0.04	0.12	0.11
$\boldsymbol{\theta}_8$: homoph. specialty	0.18	0.20	0.19	0.05	0.19	0.17
$\boldsymbol{\theta}_9$: homoph. gender	0.28	0.31	0.26	0.04	0.25	0.24
$\boldsymbol{\theta}_{10}$: homoph. office	0.19	0.22	0.16	0.05	0.16	0.17

3.6 The Package ergmQN

For the MC-BFGS method presented in the previous section, we developed a new `R` package called `ergmQN`, implementing Quasi-Newton methods for ERGMs estimation.

The backbone of the package is the function **ergmQN** that implements the MC-BFGS algorithm also in presence of missing data in the network.

As we already said in the previous sections, ergmQN is based on the procedures made available in the `R` suite package `statnet` (Handcock et al., 2008), and especially in `ergm` (Hunter et al., 2008b). The `ergmQN` package can be considered as an add-on extension of the `ergm` package.

In fact, some support functions have been included in the package in order to make the procedures and the output of `ergmQN` compatible and similar to the procedures for diagnostics, plotting and goodness-of-fit analysis already included in the `ergm` package.

ergmQN includes the possibility to choose between other two Quasi-Newton methods. In fact, Symmetric Rank 1 (`rank1`) and Davidon-Fletcher-Powell (`dfp`) methods are also implemented in the function. They are alternative ways to update the approximated Hessian matrix (J) in formula (3.7).

Table 3.10: Standard deviation of the estimates (s.d) and mean of simulated standard errors (s.e) for the parameter estimates of Table 3.7

Parameter	MPLE		MC-MLE		MC-BFGS	
	s.d	s.e	s.d	s.e	s.d	s.e
$\boldsymbol{\theta}_1$: edges	0.75	0.76	0.67	0.52	0.65	0.62
$\boldsymbol{\theta}_5$: GWEsp (0.7781)	0.18	0.18	0.15	0.19	0.16	0.21
$\boldsymbol{\theta}_6$: main eff. seniority	0.33	0.30	0.29	0.04	0.28	0.23
$\boldsymbol{\theta}_7$: main eff. specialty	0.18	0.15	0.15	0.04	0.15	0.11
$\boldsymbol{\theta}_8$: homoph. specialty	0.24	0.20	0.23	0.05	0.23	0.17
$\boldsymbol{\theta}_9$: homoph. gender	0.39	0.31	0.35	0.04	0.32	0.24
$\boldsymbol{\theta}_{10}$: homoph. office	0.26	0.22	0.24	0.05	0.22	0.17

The `ergmQN` package includes also a procedure (**predict.ergmQN**) to predict the probabilities of the network ties based on an estimated ERGM. This procedure will be useful for the results discussed in the next chapter.

3.7 Discussion

In this chapter we proposed an alternative simulate maximum likelihood method for ERGM based on BFGS algorithm that has a robust convergence behavior. The reported results confirm how the back-tracking mechanism in the method permits to escape from near-degeneracy and instability cases in which the other methods often do not converge. The convergence criteria adopted for MC-BFGS provides a good control on the behavior of the method. The estimated coefficients expressed in mean value parametrization are often closer to the statistics of the observed network then those obtained with the Steplength algorithm (Hummel et al., 2011).

The BFGS method implemented in our proposal should not be confused with the usage of the BFGS algorithm to maximize the log-likelihood approximation (3.5) available in the `ergm` package. Indeed, in our case the BFGS updating is applied directly to the target true log-likelihood function, and simulations are used to approximate the score function.

The method performs also model estimation in presence of missing data, and this is a useful feature that will be used in the next chapter.

Chapter 4

Comparison between ERGMs and Latent Space Models

This chapter proposes a comparison between ERGMs and Latent Space models in terms of goodness of fit. In particular the performance of the two model approaches to reproduce the dependence structure of the observed network (Hunter et al., 2008a), and their predictive power are seen as complementary features for describing their goodness of fit.

The two approaches are compared on real social network data from the literature, already presented in §2.3.

Furthermore, experiments with simulated data are carried out in order to compare the behavior of the two models when the networks are, by construction, strongly in favor of one of the two approaches.

The results of this chapter are somehow preliminary, as computational constraints prevented a thorough assessment via experiments with simulated data. Some suggestions do emerge, although further work would be needed to reach more complete conclusions.

4.1 Goodness-of-Fit Procedure and Prediction Power

The **Goodness-of-Fit** (GOF) procedure (Hunter et al., 2008a) assesses the model ability to catch and rebuild the dependence structure of the observed network.

According to an estimated model, the GOF procedure is able to derive, for a large number of simulated networks, a set of network statistic distributions (as degree, minimum geodesic distance, edge-wise and dyad-wise shared partners). It is possible to evaluate if the estimated model reproduces the dependence structure of the observed network by comparing graphically the empirical distribution of the observed statistics with the corresponding

statistics for the simulated networks. For each element of the distribution of a network statistic, the GOF procedure computes a set of p-values.

Due to the different scale in the plots and the large number of elements of the network statistic distributions, the comparison between the GOF for two different models is not an easy task. Therefore, we propose to obtain a synthetic index by summarizing all the p-values of a distribution with their weighted mean

$$Pv_i = \frac{\sum_{k=1}^{n_i}(b_k + 1)\, pv_k}{\sum_{k=1}^{n_i}(b_k + 1)}, \tag{4.1}$$

where n_i is the number of elements of the i-th network distribution, b_k and pv_k are respectively the observed frequency and the p-value of the k-th element of the distribution. Note that in order to consider the p-values for the elements of the distribution that have observed frequency equal to 0, we used $(b_k + 1)$ instead of b_k. This synthetic index does not correspond to any statistical test, but it is useful to figure out an immediate idea of the behavior of the model.

The GOF procedure measures somehow the ability of the estimated model to globally reproduce the dependence structure of the observed network, but it does not provide any information on the existence of the single ties. This different aspect of goodness of fit can be evaluated by considering the prediction ability (Kolaczyk, 2009, Ch. 7) of the model.

By the nature of relational data, it is easy to think of an estimated model as a binary classifier, no matter which is the estimation method used for it (i.e. maximum likelihood estimation or Bayesian methods). The idea is to evaluate the model in term of true-positive and false-positive rates on the predicted ties.

The link prediction has an useful meaning only if the predictions are made on ties that have not contributed to the model estimation. A solution to this point can be to consider it as a missing data problem (Kolaczyk, 2009), where the missing data do not contribute to model estimation but they can be useful to assess the prediction power of the model. Each method has been evaluated considering two different cases of missing data.

- Missing edges: We randomly set some edges as missing. The predictions are made on this subset of edges.
- New nodes: We randomly chose a subset of nodes, setting as missing all the elements on their rows and columns. These could be thought as a set of new nodes that enter the network, or nodes that were not sampled (Handcock and Gile, 2010). The prediction power is evaluated on the predicted links on the set of the excluded nodes.

Both model approaches (ERGMs and LCRMs) can be estimated in presence of missing data.

For an ERGM, an estimate of the link probabilities for the missing data can be made as follows:

1. Estimate the parameters of the ERGM via face-valued likelihood as presented in §3.3.

2. From the estimated model, we simulated a large number (1,000 for example) of networks conditional on the observed part of the data.

3. The estimated link probabilities are given by the proportion of time that the ties are present.

For LCRMs, the models are estimated assuming the dyad conditional independence (§2.2.4). This implies that the missing data are removed from model estimation. For the missing edges case, the models are still estimable even thought the presence of partial data can lead to some bias both on fixed and random parts of the model. In the new node case, instead, removing the missing part of the data is more problematic, because no information is then available for the missing node, and we have to rely on the distributional assumptions made for the latent part of the model.

For the missing edges case, the link probabilities are obtained using the predictive distribution (see for example Tanner, 1996) approximated as the mean on the estimated MCMC sample, with a large size (at least 4,000) and after a large number of burn-in steps (at least 5,000).

$$\begin{aligned}
\hat{P}(y_{ij}=1|Y) &= \int p(y_{ij}=1|x_{ij},Z_i,Z_j,\beta,\delta,\gamma)\pi(Z,\beta,\delta,\gamma|Y)dZd\beta d\delta d\gamma \\
&\approx \frac{1}{S}\sum_{s=1}^{S} p(y_{ijs}|x_{ij},Z_{is},Z_{js},\beta_s,\delta_s,\gamma_s) \\
&= \frac{1}{S}\sum_{s=1}^{S} \text{logit}^{-1}(\beta_s x_{ij} - ||Z_{is}-Z_{js}|| + \delta_s + \gamma_s), \qquad (4.2)
\end{aligned}$$

where $||Z_{is}-Z_{js}||$ is the euclidean distance between the actors i and j, β_s are the model coefficients, and δ_s and γ_s the sender/receiver random effects for the s-th MCMC sample.

In the new node case, the random parts (positions in the latent space and sender/receiver random effects) of the new nodes are simulated for the s-th MCMC sample as follows

$$Z_s^m \sim \sum_{k=1}^{G} \lambda_g^s \, MVN_d(\mu_k^s, \Sigma_k^s),$$

$$\begin{pmatrix} \delta_s^m \\ \gamma_s^m \end{pmatrix} \sim N_2\left(0, \begin{pmatrix} \sigma_\delta^s & 0 \\ 0 & \sigma_\gamma^s \end{pmatrix}\right),$$

where again μ_k^s, Σ_k^s, σ_δ^s and σ_γ^s are simulated parameters for the s-th MCMC sample. The predictive probabilities are obtained as in formula (4.2) after considering together the estimated terms and the new simulated parts.

4.1.1 Evaluation of the prediction power

A common procedure to compare two classifiers is by looking at Receiver Operating Characteristic (**ROC**) curves (see for example Hanley and McNeil, 1982). These curves represent the true-positive rate (y-axis) and the false-positive rate (x-axis) in the model predictions, switching the cut-off value. The Area Under the ROC Curve (AUC) gives a simple index for the link prediction power of the models, it is maximized at 1 in the case of a perfect classifier, and it is expected to be 0.5 in the case of random guessing. See Hanley and McNeil (1982), Hanley and McNeil (1983) and Cortes and Mohri (2005) for further details.

In order to reduce the trade-off between model estimation burden and evaluation of prediction accuracy we adopted a K-fold cross-validation procedure (Kolaczyk, 2009, Ch. 7). For both missing edges and new node cases we randomly partitioned the original sets (edges or nodes) in subsets. A single subset was retained as validation data for testing the model, and the remaining subsets were used as training data. The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as validation data. The results of the cross-validation processes were averaged to produce a single estimate.

4.2 Data Examples

In the following, we report the main results of the comparisons between ERGMs and LCRMs based on four known network datasets taken from the literature (Wasserman and Faust, 1994; Hunter and Handcock, 2006; Snijders and Nowicki, 1997). They were already presented in §2.3 and their main characteristics are summarized in Table 4.1.

ERGMs are estimated by the Quasi-Newton procedure introduced in the previous chapter and included in the `ergmQN` packages. LCRMs and all the GOF procedure are obtained using the tools in the packages that compose the `statnet` suite in R.

In every case, we compared the best models in terms of BIC (for ERGM) and Overall BIC (for LCRM) information. To choose the models we checked also, by MCMC diagnostics, that in the estimated models there were no degeneracy, good mixing of the chains, and that the estimated clusters in the latent space were not nested.

For the prediction power assessment, the number of K-fold considered for the cross-validation in the missing edges and the new nodes cases sometimes differed to maintain balanced the proportions of missing data in the two

cases. At any rate, K was always in the range 5-10, that seemed a sensible choice.

Table 4.1: Main characteristics of the datasets.

Data set	Relation type	Size	Actor Attributes
Kapferer tailor shop	Undirected	39	No
Lazega Lawyers	Undirected	36	Yes
Sampson Monks	Directed	18	Yes
Krackhardt High-Tech Managers	Directed	21	Yes

4.2.1 Kapferer's tailor shop

For the `kapferer` data (§2.3.3), we compared the following models.

Model 1: ERGM with the terms for edges, number of Nonedgewise shared partners of order 2 and nodal attribute mixing effects for the vertex attribute `highstatusjob`.

Model 2: LCRM considering euclidean distances in a two dimensional latent space with three clusters.

Table 4.2: Estimated parameters (s.e.) for `kapferer` data.

	ERGM	LCRM
edges	−1.696 (0.108)	1.520 (0.492)
nsp(2)	−0.227 (0.061)	
mix.highstatusjob.0.0	0.581 (0.190)	
mix.highstatusjob.1.1	1.331 (0.225)	

Table 4.2 reports coefficients and standard errors (posterior standard deviation for the LCRM) for the estimated models.

The coefficients of the ERGM suggest low density in the network and the tendency to local transitivity. This can confirm the presence of clusters in the network.

The LCRM finds three clusters in the data as reported in Figure 4.1.

For the GOF comparison, the plots in Figure 4.2 show that the networks simulated from the ERGM are closer to the observed one. The synthetic indexes of Table 4.3 confirm that the ERGM gives a better fit except for the distribution of the edge-wise shared partners.

For the link prediction comparison, plots in Figure 4.3, LCRM provides a better prediction in the missing edges case. While the two methods provide a comparable accuracy in new node case.

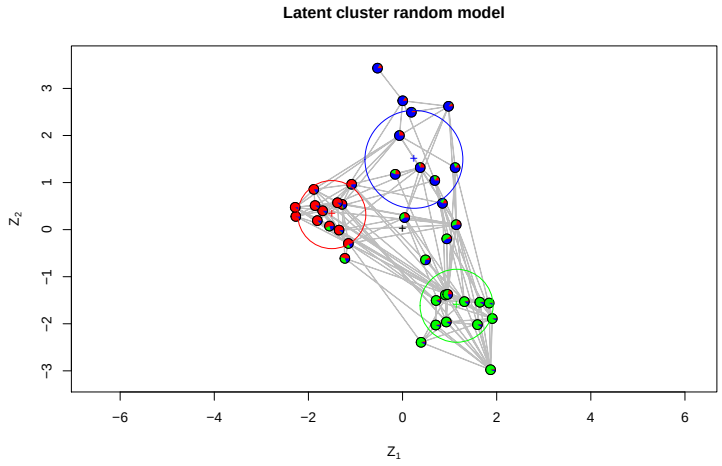

Figure 4.1: Estimated latent positions for `kapferer` data

Figure 4.2: Goodness-of-Fit plots for `kapferer` data

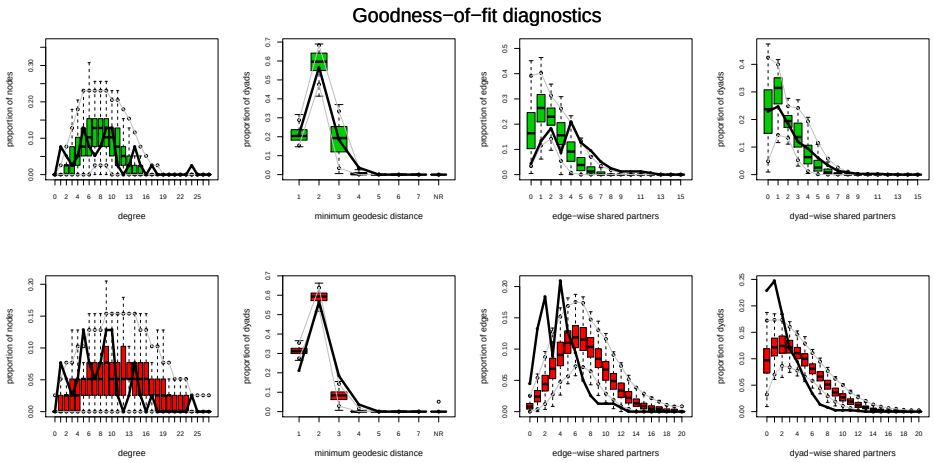

Table 4.3: Mean p-values for GOF terms for `kapferer` data

Model	degree	distance	espartners	dspartners
ERGM	0.711	0.755	0.305	0.665
LCRM	0.605	0.263	0.232	0.163

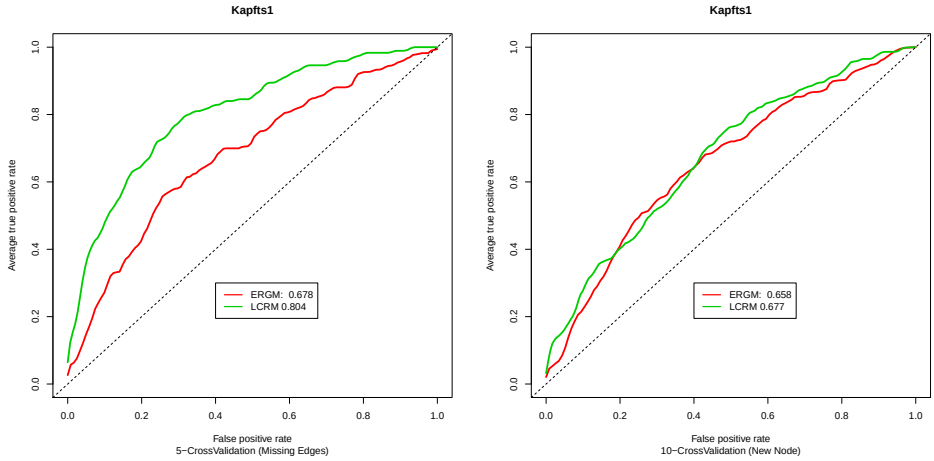

Figure 4.3: ROC curves in Missing Edges and New Node cases for `kapferer` data. The black curve is for the ERGM and the green curve is for the LCRM. AUC values are also reported.

4.2.2 Lazega Law Firm

Four models were considered for the collaborative network among the 36 partner lawyers (§2.3.4). The results are summarized in Table 4.4.

Model 1: ERGM with the terms for edges, geometrical weight edge-wise shared partners and geometrical weighted dyad-wise shared partners, both with weights equal to 0.45.

Model 2: ERGM with the terms for edges and the weight for the geometrical edge-wise shared partners fixed to 0.778. Main effects for seniority, specialty, age and office were added, plus homophily terms for specialty, gender and office.

Model 3: LCRM setting two clusters in a two dimensional latent space.

Model 4: LCRM considering no clusters and the main effects for specialty, age, office, and the same homophily terms included in Model 2.

For the estimated ERGMs, we can say that almost all the attribute terms have positive effects on the formation of the ties. In both cases, there is a negative effect for the edges and a positive effect for the shared partner terms, pointing to networks with low density and transitive between actors.

For the estimated LCRMs, the model without attribute terms (Model 3) founds two clusters in the network (first plot in Figure 4.4). These clusters could not be explained by a single actor effect, but by a mix of them, as it can be seen in the second plot in Figure 4.4 where the introduction of

Table 4.4: Estimated parameters (s.e.) of the parameters of the models for Lawyers data

	ERGM	ERGM Cov	LCRM	LCRM Cov
edges	-3.936 (0.638)	-1.762 (2.756)	1.080 (0.587)	-0.08 (1.516)
gwdsp.fixed.0.45	1.603 (0.266)			
gwesp.fixed.0.45	-0.063 (0.065)			
gwesp.fixed.0.778		0.865 (0.157)		
Main Effects:				
Seniority		-0.016 (0.017)		
Specialty		0.330 (0.117)		0.829 (0.280)
Age		-0.046 (0.021)		-0.060 (0.014)
Office		0.264 (0.099)		0.611 (0.204)
Homophily Effects:				
Specialty		0.744 (0.199)		0.981 (0.300)
Gender		0.404 (0.282)		0.887 (0.545)
Office		1.377 (0.215)		2.469 (0.350)

Table 4.5: Mean p-values for GOF terms for Lawyers data

Model	degree	distance	espartners	dspartners
ERGM	0.304	0.045	0.144	0.036
ERGM Cov	0.736	0.809	0.654	0.783
LCRM	0.612	0.083	0.408	0.119
LCRM Cov	0.595	0.077	0.478	0.128

actor effects renders the clusters not necessary any longer. All the attribute terms, but age, have positive effects on the ties formation.

The GOF comparison, based on the plots in Figure 4.5 and the values in Table 4.5, points to the ERGM with covariates as the model that seems to catch better the network statistic distributions.

The plots in Figure 4.6 show that for the missing edges case, all the models have a comparable behavior in terms of AUC. For the new node case, the best performance is observed for the LCRM with covariates, but also the ERGM with covariates has a good performance. In both cases, the introduction of covariates increases the AUC values.

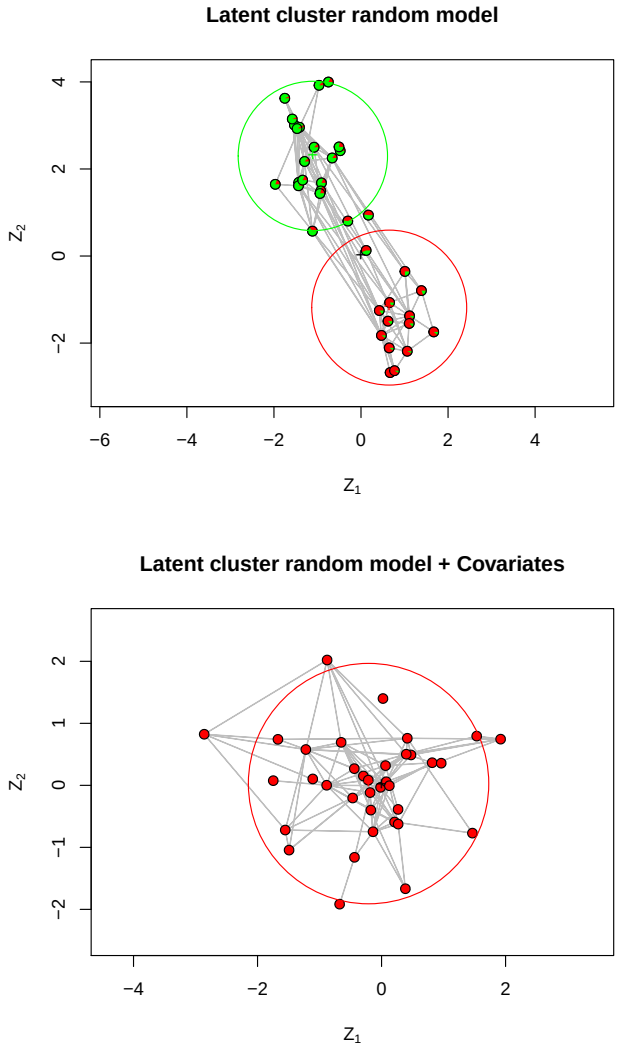

Figure 4.4: Estimated latent positions of Model 3 and Model 4 for Lawyers data

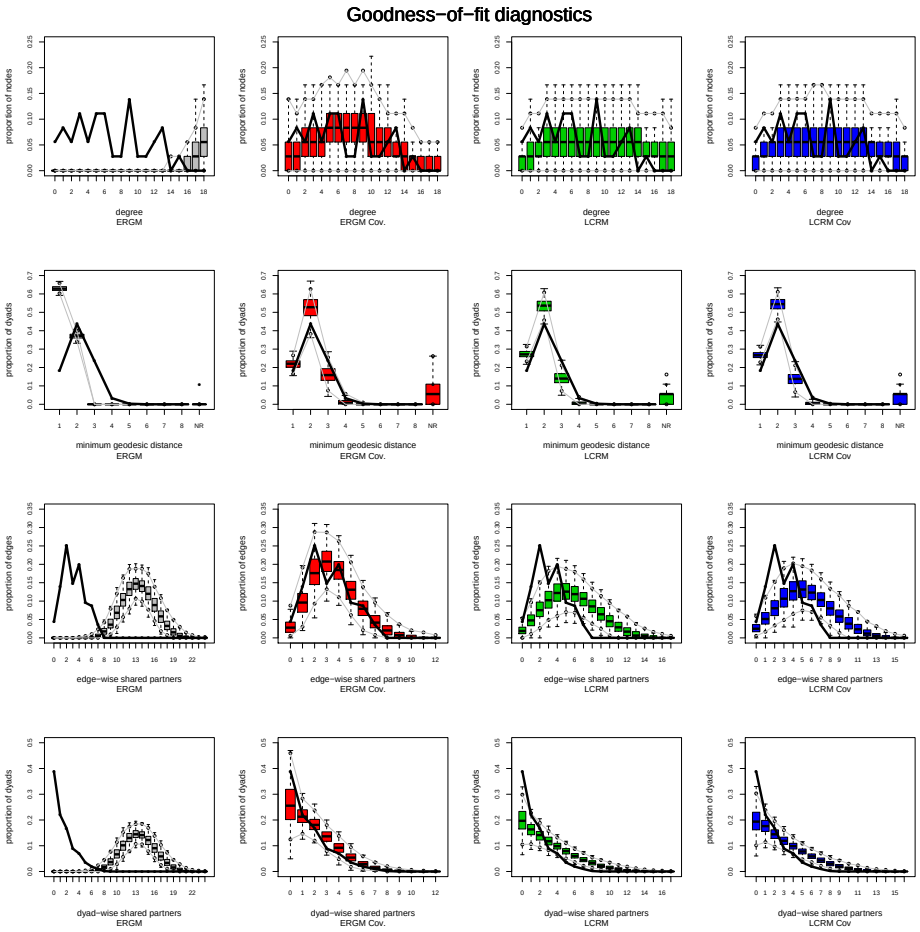

Figure 4.5: Goodness-of-Fit plots for Lawyers data

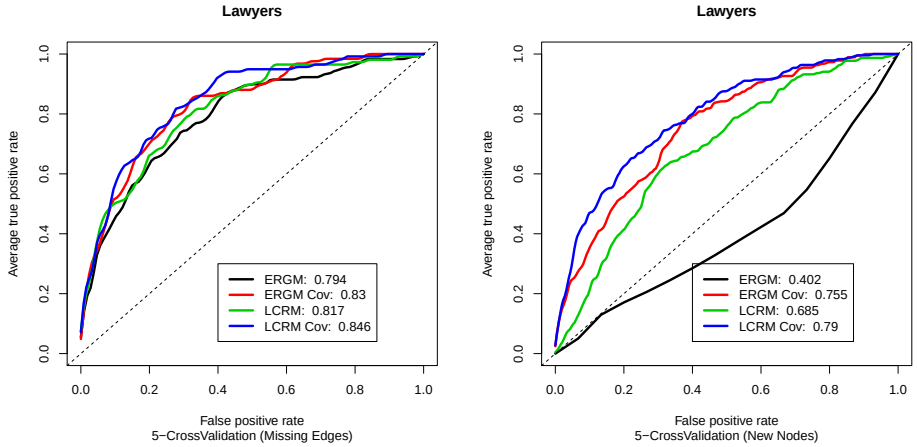

Figure 4.6: ROC curves in missing edges and new node cases for Lawyers data. The black curve is for the ERGM, green curve is for the LCRM, the red and blue curves are for the ERGM and the LCRM, both with the covariates. AUC values are also reported.

4.2.3 Sampson's Monastery Study

For the Monks data (§2.3.5), we considered two models, Table 4.6

Model 1: ERGM with the terms for edges, geometrical weight in and out degree, geometrical weight edge-wise shared partners, geometrical weight dyad-wise shared partners, with weights equal to 0.45.

Model 2: LCRM in a two dimensional latent space with three clusters.

Table 4.6: Estimated parameters (s.e.) of the parameters of the models for Monks data

	ERGM	LCRM
edges	-2.006 (0.605)	2.202 (0.507)
gwidegree.fixed.0.45	8.528 (3.679)	
gwodegree.fixed.0.45	85.231 (28.897)	
gwdsp.fixed.0.45	0.644 (0.200)	
gwesp.fixed.0.45	-0.479 (0.088)	

From the plots in Figure 4.8 and the values in Table 4.7, we can see that the ERGM performs better in terms of GOF.

From the ROC curve in Figure 4.9, LCRM provides a better prediction in the missing edges case. In the new node case, both the estimated models practically perform a sort of random guessing.

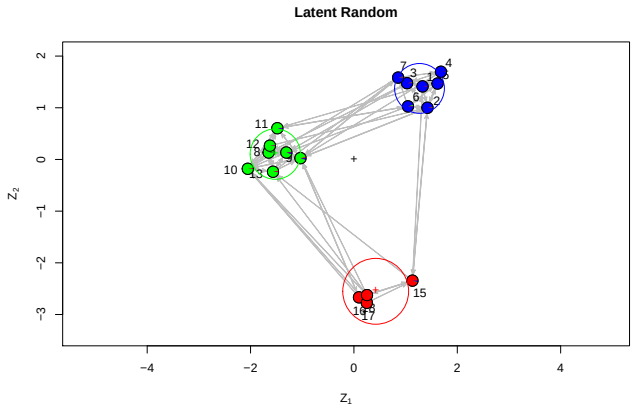

Figure 4.7: Latent position for Monks data

Table 4.7: Mean p-values for GOF terms for Monks data

Model	idegree	odegree	distance	espartners	dspartners
ERGM	0.839	0.785	0.784	0.546	0.931
LCRM	0.772	0.555	0.592	0.660	0.782

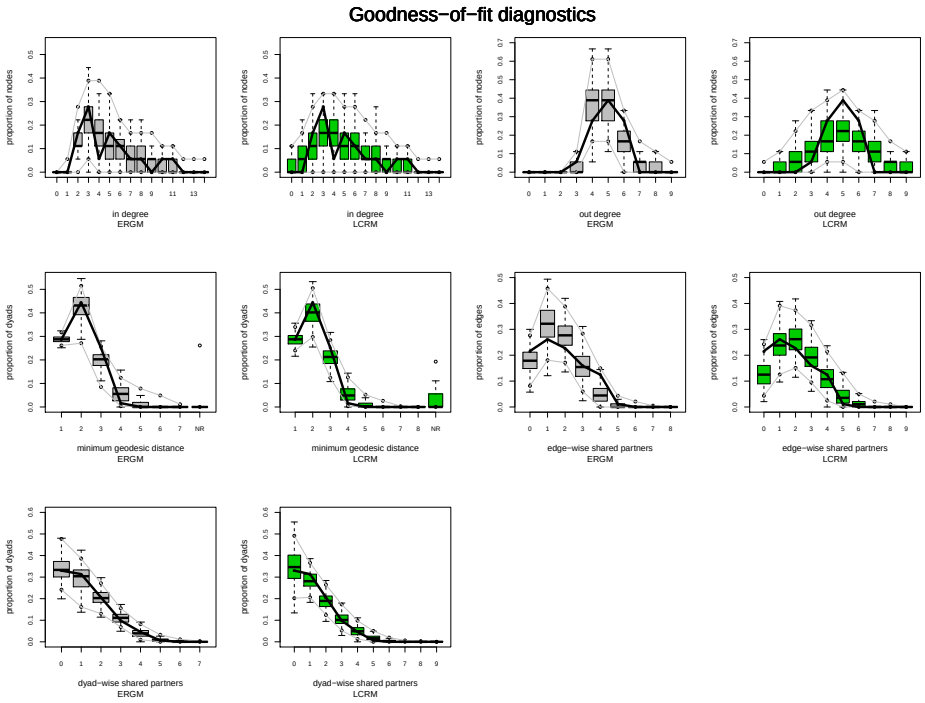

Figure 4.8: Goodness-of-Fit plots for Monks data

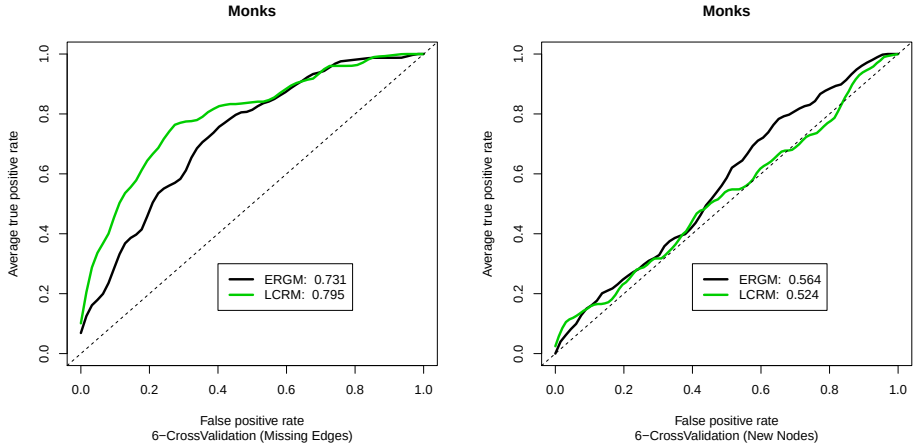

Figure 4.9: ROC curves in missing edges and new node cases for Monks data. The black curve is for the ERGM and green curve is for the LCRM. AUC values are also reported.

4.2.4 Krackhardt's High-tech Managers

For Hightech data (§2.3.6), Table 4.8 reports the estimated models that were compared.

Model 1: ERGM with the terms for edges, geometrical weighted edge-wise shared partners and geometrical weighted dyad-wise shared partners, with weights equal to 0.25.

Model 2: ERGM where the weight for the geometrical edge-wise shared partners is fixed to 0.35. It has also terms of input and output main effects for Age, Tenure, Department, Level, plus homophily terms for Department and Level.

Model 3: LCRM in a two dimensional latent space, with two clusters.

Model 4: LCRM in a two dimensional latent space without clusters and considering some of the terms for actor attributes of Model 2.

For ERGMs, in both the two models the term for the number of edges is negative and the terms for the shared partners are positive except the one for the edge-wise shared partners in Model 1. These facts are due to the low density of the network and maybe they point to the presence of clusters. The two homophily terms in Model 2 have positive effects on tie creation.

For LCRMs, we can see that Model 3, first plot in Figure 4.10, finds two clusters in the network. In the Model 4, second plot in Figure 4.10, the cluster information is explained in the model by the attribute terms but it is not possible to identify which homophily term represents the clusters.

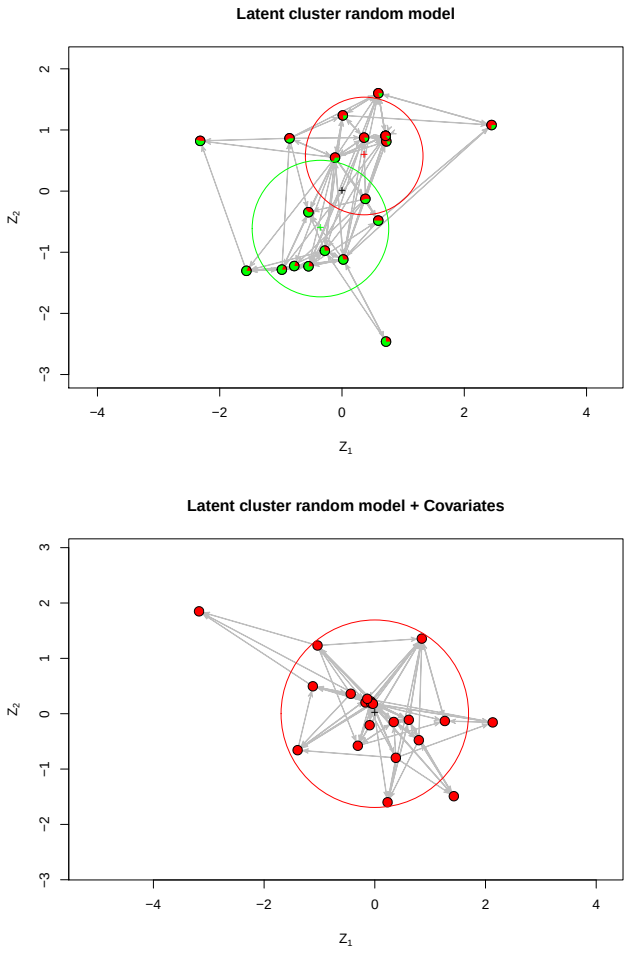

Figure 4.10: Estimated positions in the latent spaces for HighTech Managers data.

Table 4.8: Estimated parameters (s.e.) of the parameters of the models for HighTech Managers data

	ERGM	ERGM Cov	LCRM	LCRM Cov
edges	-1.967 (0.376)	-5.462 (1.757)	0.794 (0.660)	-0.652 (1.477)
gwdsp.fixed.0.25	1.027 (0.210)			
gwesp.fixed.0.25	-0.221 (0.057)			
gwesp.fixed.0.35		0.628 (0.249)		
Main in-effects:				
Age		-0.002 (0.016)		-0.092 (0.023)
Tenure		-0.001 (0.020)		0.143 (0.031)
Department		0.204 (0.115)		-0.165 (0.161)
Level		-0.147 (0.275)		1.209 (0.360)
Main Out-effects:				
Age		-0.064 (0.023)		-0.031 (0.020)
Tenure		0.081 (0.026)		0.065 (0.026)
Department		-0.049 (0.110)		
Level		1.688 (0.426)		
Homophily Effects:				
Department		1.202 (0.542)		1.647 (0.414)
Level		2.246 (0.686)		1.114 (0.427)

The plots in Figure 4.11 and the values in Table 4.9 suggest that the ERGM with covariates provides best fit between the considered models. Including the attribute effects in the ERGM specification increases considerably the capability of reproducing the observed network statistic distributions. The LCRMs have a similar behavior and the effect of actor attributes on the GOF of the model is not clear.

Table 4.9: Mean p-values for GOF terms for HighTech Managers data

Model	idegree	odegree	distance	espartners	dspartners
ERGM	0.581	0.600	0.324	0.482	0.166
ERGM Cov	0.721	0.760	0.779	0.559	0.547
LCRM	0.714	0.773	0.842	0.557	0.623
LCRM Cov	0.670	0.573	0.637	0.552	0.448

The curves in Figure 4.12 support Model 2 (ERGM with Covariates) and Model 3 (LCRM) as the best models respectively in the missing edges and new node cases. In both the cases, the LCRM without covariates performs better the ties prediction respect to the LCRM with covariates. Instead, for the two ERGMs the inclusion of the covariates increases the link prediction performance.

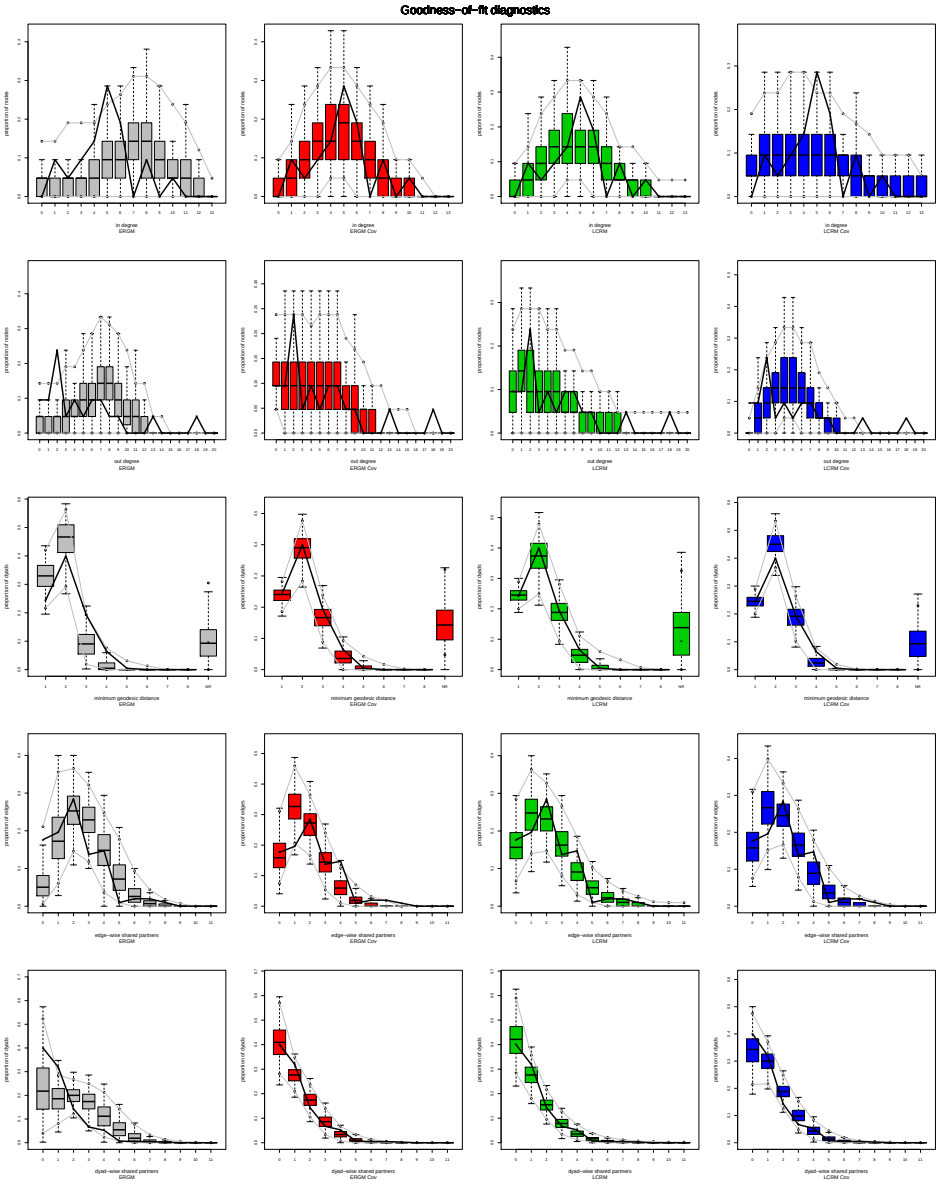

Figure 4.11: Goodness-of-Fit plots for HighTech Managers data

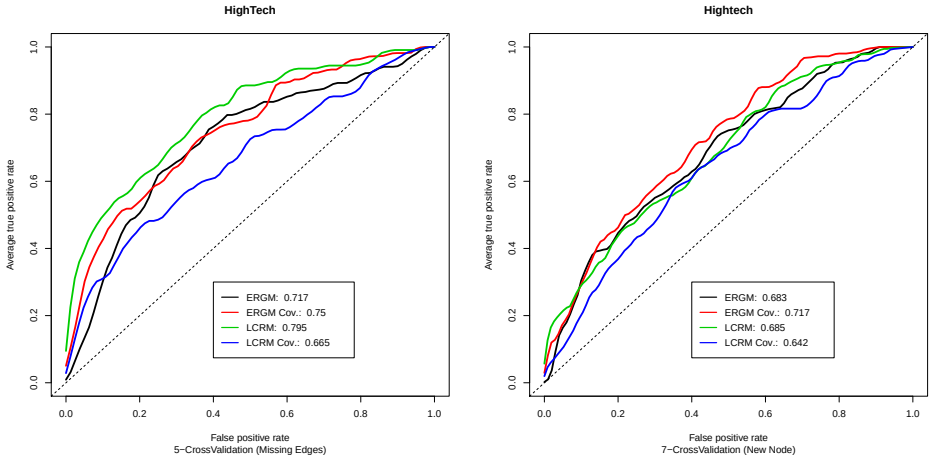

Figure 4.12: ROC curves in missing edges and new node cases for HighTech Managers data. The black curve is for the ERGM without covariates, the green curve is for the LCRM without covariates, the red and blue curves are for the ERGM and LCRM both with covariates. AUC values are also reported.

4.3 Experiments with Simulated Data

The results about the real data examples did not point clearly to one of the two model approaches. In order to deepen the study on the behavior of the two methods, we performed some experiments with simulated data. In fact, with simulated data we have the important advantage of knowing which is the true model.

Here we limit the analysis to simulated networks from a simple ERGM. The ERGM specification includes only the edges and the edge-wise shared partners terms. The setups of the two studies differ only in the coefficient values specified in order to obtain networks of order 50 with medium (around 0.25) and low (around 0.1) densities.

As this kind of analysis is computationally intensive, the number of simulated networks was limited just to 10, a small number that required nonetheless many days of computing time. In fact, each simulated network requires to estimate 16 models: the model on the full network, used for the GOF evaluation; five models (with missing data) for the link comparison in the missing edges case and ten models (with missing data) for the new nodes case. Although limited, such a small sample provides some information about the model comparison. At any rate, this study should be taken as merely of explorative nature.

Figure 4.13 shows the boxplots of the density for the ten networks simulated in each simulation study setup.

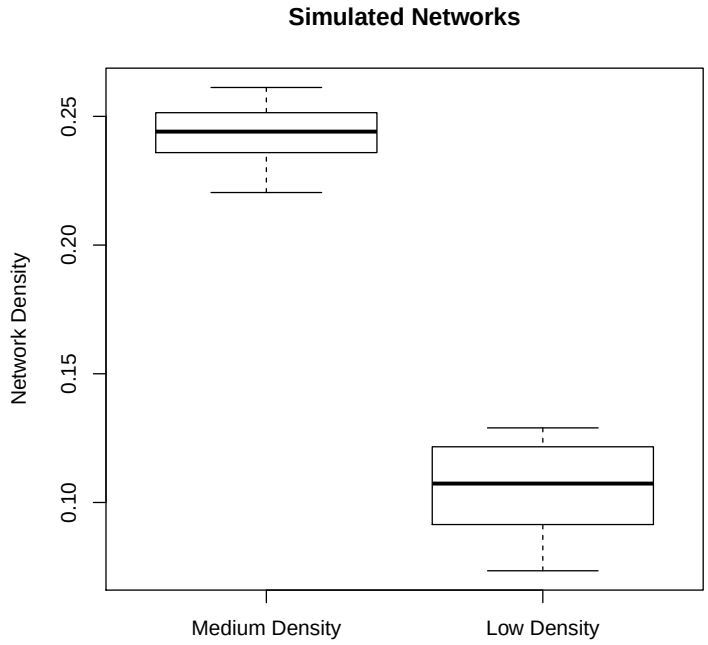

Figure 4.13: Empirical distribution of density statistics for the two setups of simulated networks.

The models compared in the studies are a LCRM versus the same ERGM used in the simulations. For the link prediction comparison, the Wilcoxon signed-rank test is used for comparing the distribution of the AUC values for the two models.

Note that we do not include instead the results of the dual experiments with networks generated from some specified LCRMs. In fact, whereas it is simple to fit a LCRM with a fixed number of clusters for a network generated from an ERGM, the automatic determination of an ERGM specification in the dual case is not as simple. Therefore, the results obtained from data simulated from LCRMs would be strongly in favor of the LCRMs, unless a fine tuning were performed about a sensible ERGM specification.

4.3.1 Medium density case

The networks simulated in this study have a density around 0.25, and they are obtained from an ERGM with the coefficients for the edge and GWESP(0.5) terms equal to -2.25 and 0.5, respectively.

The boxplots in Figure 4.14 display the synthetic index of formula (4.1) computed on the GOF procedures for the simulated networks. The ERGM

approach is the best in terms of GOF comparison.

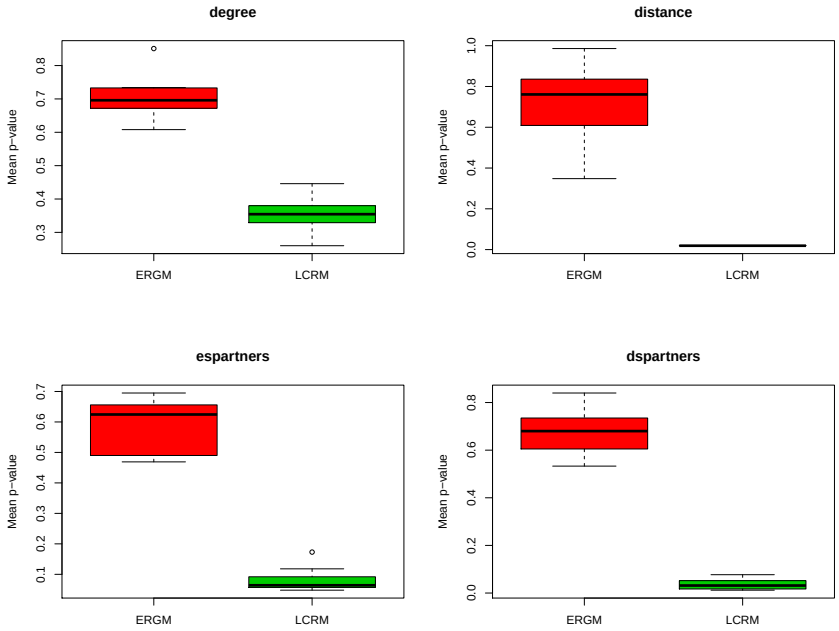

Figure 4.14: Empirical distributions of the synthetic index of formula 4.1 for the GOF procedures for medium density simulated networks.

Figure 4.15 shows the boxplots for the area under the ROC curves computed in the missing edges and new node cases on the simulated networks. It seems that the prediction performance of the ERGM is slightly better than the performance of the LCRM in the missing edges case. This is confirmed also by the p-values of the Wilcoxon signed-rank test reported in the plots.

4.3.2 Low density case

The networks simulated here have a density around 0.1, and they are obtained from an ERGM with the coefficients for the edge and GWESP(0.5) terms equal to -2.85 and 0.5.

The GOF comparison based on the boxplots in Figure 4.16 points again to the ERGM approach as the best one. The behavior of the LCRM in this case seems better than the medium density case.

The plots in Figure 4.17 show that the prediction performance of the ERGM are better than the performance of the LCRM both in the missing edges and new node case, but only slightly so. This is also confirmed by Wilcoxon signed-rank test.

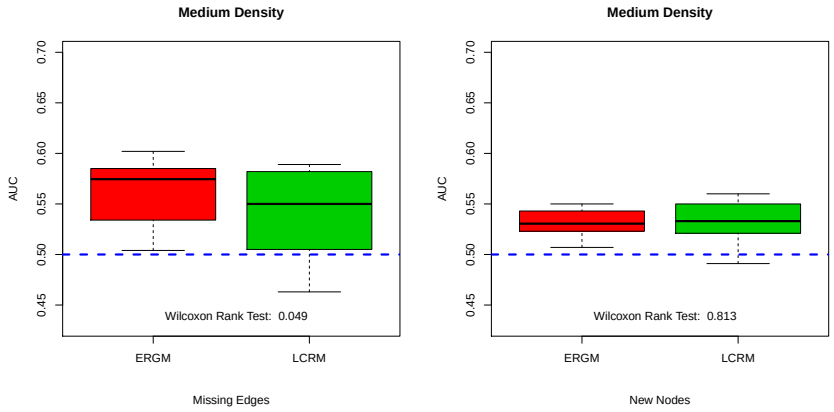

Figure 4.15: Empirical distributions of the AUC values for medium density simulated networks in the missing edges and the new node cases. The blue dashed line represents the AUC for a random guessing. The plots reports also the p-values for the Wilcoxon signed-rank test.

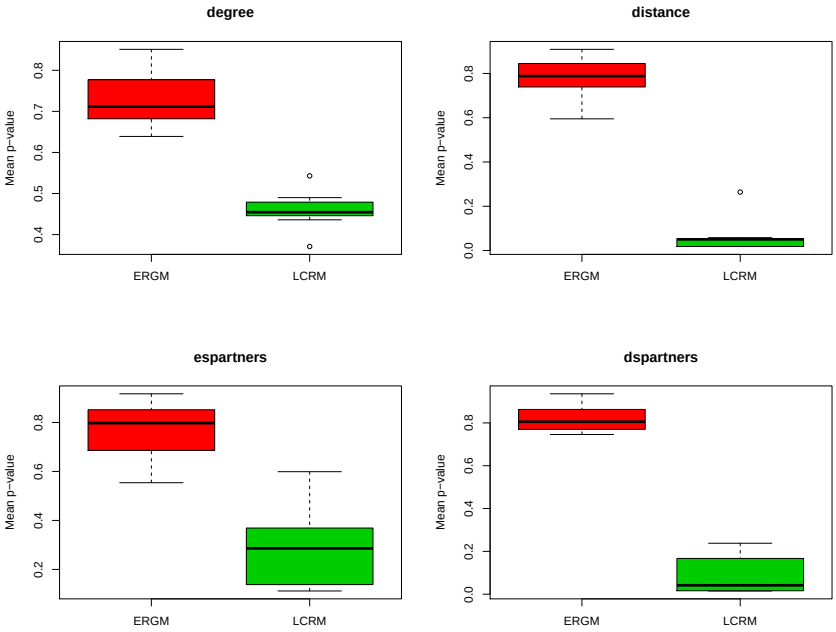

Figure 4.16: Empirical distributions of the synthetic index of formula 4.1 for the GOF procedures for medium density simulated networks.

4.4 Discussion

The results on the real data sets and on the experiments with simulated data show that there is not an overall best approach.

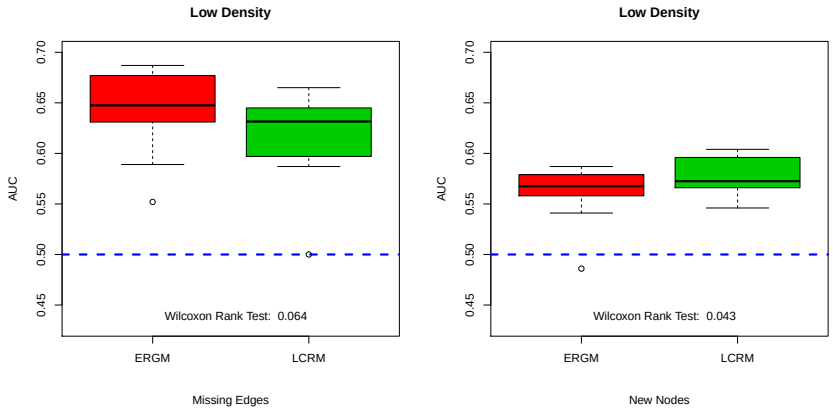

Figure 4.17: Empirical distributions of the AUC values for low density simulated networks in the missing edges and the new node cases. The blue dashed line represents the AUC for a random guessing. The plots reports also the p-values for the Wilcoxon signed-rank test.

There is some moderate evidence that ERGMs may provide better GOF than LCRMs, though this is not a general result. However, this fact does not apply to link prediction in the same manner. Indeed, for real data, the best-performing LCRMs are never worse than the best-performing ERGMs for link prediction, and sometimes they are better. On the other hand, for simulated networks ERGMs were preferable, but they corresponded to the model used as the data generating process.

For the link prediction comparison, we note that part of the difference found may be due to the prediction procedure we used. In fact, the predicted values for ERGMs were obtained by an estimative distribution (Young and Smith, 2005, Ch. 10), whereas for LCRMs a Bayesian predictive distribution was adopted. It could be interesting to compare the prediction power for ERGMs based on a Bayesian predictive distribution. In that case the predicted tie probabilities would be more comparable, and perhaps with more similar performances. Some results in this direction can be found in Koskinen et al. (2010).

Finally, the results of this chapter show also that the inclusion of covariates in the model always improves both the GOF and the prediction performance of ERGMs. For LCRMs, the introduction of the covariates may lead to estimated latent space with only one cluster, with an overall worse performance of the resulting model. This fact supports the key role of the latent space to model homophily and transitivity effects in the network (Krivitsky et al., 2009).

Chapter 5

A Laplace Approximation Approach for p_2 Network Regression Models with Crossed Random Effects

The class of p_2 models can be used for the study of binary relational data with covariates, typical of social network analysis. Such models have been somewhat underused in empirical applications, maybe due to the computational difficulties linked with the crossed structure of their random effects. At any rate, they represent a useful tool, capable of being extended in various directions, therefore their study appears fully justified.

In the literature there are proposals to estimate the parameters of p_2 models either by joint maximization methods (such as MQL estimation) or following a Bayesian approach and employing MCMC methods.

In this chapter we propose a further possibility, based on the Laplace approximation approach coupled with tilted importance sampling. This solution provides a good approximation to maximum likelihood estimation. Its implementation requires some care, but it can be performed efficiently. Numerical examples for real data and simulation studies are reported.

5.1 Introduction

The p_2 model (§2.2.5), assumes that the probability for a dyad (y_{ij}, y_{ji}) is given by

$$P(Y_{ij} = y_1, Y_{ji} = y_2)$$
$$= \frac{\exp\{y_1(\mu_{ij} + \alpha_i + \beta_j) + y_2(\mu_{ij} + \alpha_j + \beta_i) + \rho_{ij}\, y_1 y_2\}}{1 + \exp\{(\mu_{ij} + \alpha_i + \beta_j)\} + \exp\{(\mu_{ij} + \alpha_j + \beta_i)\} + \exp\{(2\,\mu_{ij} + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho_{ij})\}},$$

where α_i is the sender parameter of actor i, the β_i is receiver parameter of actor i, whereas μ_{ij} and ρ_{ij} are the density and the reciprocity parameters respectively of dyad (i, j). These parameters usually depend on some covariates and random effects, namely

$$\alpha_i = X_{1i}\,\gamma_1 + a_i\,, \qquad \beta_i = X_{2i}\,\gamma_2 + b_i\,, \qquad \mu_{ij} = \mu + Z_{1ij}\,\delta_1\,, \qquad \rho_{ij} = \rho + Z_{2ij}\,\delta_2\,.$$

Here X_1 and X_2 are actor-specific design matrices, Z_1 and Z_2 contain dyad-specific covariates and $u_i = (a_i, b_i)^T$ are normally distributed random effects. Namely, we assume the random effects U_i are normally distributed independent random variables,

$$U_i \sim N_2(0, \Sigma)\,, \qquad \Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}. \tag{5.1}$$

Notice that random effects model the ties sent or received by a given actor, that are thus assumed to be dependent. All the parameters of this model can be collected together in the vector θ

$$\theta = (\gamma_1, \gamma_2, \mu, \delta_1, \rho, \delta_2, \sigma_A^2, \sigma_{AB}, \sigma_B^2)^T\,.$$

The p_2 model is conveniently formulated as a multinomial regression model with random effects, as done in Van Duijn et al. (2004) and Zijlstra et al. (2009). In fact, for each dyad there are four possible outcomes: (0,0), (1,0), (0,1), and (1,1). Hence, taking (0,0) as the reference category, the response data Y can be represented by $g\,(g-1)/2$ stacked three-dimensional vectors d_{ij}

$$d_{ij} = \begin{pmatrix} d_{1ij} \\ d_{2ij} \\ d_{3ij} \end{pmatrix} = \begin{pmatrix} y_{ij}\,(1 - y_{ji}) \\ y_{ji}\,(1 - y_{ij}) \\ y_{ij}\,y_{ji} \end{pmatrix}.$$

Using this representation, the distribution of the response given the random effects has the exponential family form

$$P(D = d|u) = \exp\left\{\xi^T d - b(\xi)\right\}, \tag{5.2}$$

where the vector of linear predictors is

$$\xi_{ij} = \begin{pmatrix} \xi_{1ij} \\ \xi_{2ij} \\ \xi_{3ij} \end{pmatrix} = \begin{pmatrix} X_{1i}\,\gamma_1 + a_i + X_{2j}\,\gamma_2 + b_j + \mu + Z_{1ij}\,\delta_1 \\ X_{1j}\,\gamma_1 + a_j + X_{2i}\,\gamma_2 + b_i + \mu + Z_{1ij}\,\delta_1 \\ \xi_{1ij} + \xi_{2ij} + \rho + Z_{2ij}\,\delta_2 \end{pmatrix},$$

and $b(\xi) = \sum_{i<j}^{g} \log\{1 + \exp(\xi_{1ij}) + \exp(\xi_{2ij}) + \exp(\xi_{3ij})\}$.

In the paper that first introduced the p_2 model, Van Duijn et al. (2004) proposed to estimate θ by a MQL approach (Breslow and Clayton, 1993), further extended in Zijlstra et al. (2009). There is broad consensus in the random effects literature about the fact that both MQL and PQL approaches perform generally poorly for nonlinear models and discrete data (Molenberghs and Verbeke, 2005, Ch. 14), and actually this is also apparent in the simulation studies reported in Zijlstra et al. (2009). For this reason, Zijlstra et al. (2009) proposed a Bayesian approach, sampling from the posterior distribution by MCMC. In particular, they adopted a slightly informative prior for model parameters, and explored the performance of several sampling algorithms. The reported results were somewhat good, even when evaluated under a frequentist perspective. A maximum likelihood approach may be appealing nonetheless, avoiding the need to specify a prior distribution for the parameter and to perform diagnostic checking on the convergence of the MCMC algorithm. Furthermore, a maximum likelihood approach may have better scalability, with some computational efficiency gain for larger networks.

5.2 Approximate Maximum Likelihood Estimation

The likelihood function for the model defined by (5.1) and (5.2) is given by the following integral over the random effects

$$L(\theta) = \int_{\mathbb{R}^{2\,g}} \exp\left\{\xi^T d - b(\xi)\right\} \left\{\prod_{i=1}^{g} \phi_2(u_i; 0, \Sigma)\right\} du\,, \tag{5.3}$$

where $\phi_2(\cdot)$ is the bivariate normal density (5.1).

The high-dimensional integral (5.3) has to be evaluated numerically, as the correlated random effects u_i have a crossed structure, allowing no reduction in the size of the integral.

A possible resolution is via the Laplace's method of integration, as done in Skaug (2002). If

$$h(u; \theta, y) = \xi^T d - b(\xi) + \sum_{i=1}^{g} \log \phi_2(u_i; 0, \Sigma)\,,$$

the standard Laplace approximation to $L(\theta)$ is given by

$$L^*(\theta) = \exp\{h(\widehat{u}_\theta; \theta, y)\}\ |H(\theta)|^{-1/2}\ , \tag{5.4}$$

with $\widehat{u}_\theta = \underset{u}{\operatorname{argmax}}\ h(u; \theta, y)$ and

$$H(\theta) = -\left.\frac{\partial^2 h(u; \theta, y)}{\partial u\, \partial u^T}\right|_{u=\widehat{u}_\theta}.$$

The results in Shun and McCullagh (1995) imply that the error in the standard Laplace approximation (5.4) is of order $O(1)$ when $g \to \infty$, therefore for large g the maximization of $L^*(\theta)$ provides an estimator $\widehat{\theta}^*$ of θ asymptotically equivalent to the maximum likelihood estimator.

Yet it might be advisable to improve on the approximation to (5.3) by importance sampling. In particular, here we adopted tilted importance sampling, as already used by Skaug (2002) and endorsed by Brinch (2012). The method approximates $L(\theta)$ by using as the importance sampling distribution a normal distribution with mean vector $\widehat{u}_\theta$ and covariance matrix $H(\theta)^{-1}$. If $u^{(1)}, , \ldots, u^{(M)}$ is random sample of size M from such a distribution, the resulting approximation is given by

$$L^\dagger(\theta) = \frac{1}{M} \sum_{j=1}^{M} \frac{\exp\{h(u^{(j)}; \theta, y)\}}{\phi_{2\,g}\{u^{(j)}; \widehat{u}_\theta, H(\theta)^{-1}\}} \,. \tag{5.5}$$

Skaug (2002) gave some evidence that both $L^*(\theta)$ and $L^\dagger(\theta)$ provided satisfactory estimators in a broad array of mixed models with crossed random effects structures. This was further substantiated by Brinch (2012), who provided guidelines for practical implementation and further possible improvements.

For what concerns our implementation of the method, we obtained $\widehat{u}_\theta$ by a highly-reliable trust region algorithm, as implemented in the `R` routine `nlminb` when both the gradient and Hessian of the objective function are provided. We then coded the first derivative of $\log L^*(\theta)$, obtaining $\widehat{\theta}^*$ by a quasi-Newton algorithm, whereas $L^\dagger(\theta)$ was maximized by means of a derivative-free optimizer. We note in passing that more sophisticated methods making use of automatic differentiation could be employed (Skaug, 2002; Skaug and Fournier, 2006). The random draws $u^{(j)}$ were generated as

$$u^{(j)} = \widehat{u}_\theta + C(\theta)\, v^{(j)} \,,$$

where $C(\theta)$ denotes the Cholesky factor of $H(\theta)^{-1}$ and $v^{(j)}$ is a vector of independent standard normal variates. Following Skaug (2002), in order to facilitate the maximization of $L^\dagger(\theta)$ the same set of random draws $v^{(1)}, \ldots, v^{(M)}$ was used to generate $u^{(1)}, \ldots, u^{(M)}$, for all values of θ. It is then advisable to repeat the maximization of $\log L^\dagger(\theta)$ for various choices of M, to verify that the resulting estimates become stable for a sufficiently large M.

5.3 Data Examples

5.3.1 Dutch Social Behavior Study

We consider the data from the Dutch Social Behavior Study (Baerveldt and Snijders, 1994), already analysed in Zijlstra et al. (2005), and introduced in §2.3.7.

Zijlstra et al. (2005) used Bayesian approach, taking the first network of 62 pupils as calibration sample to be used to obtain prior distributions for the analysis sample, the second network of 39 pupils. Namely, they fitted a Bayesian model by MCMC with diffuse priors for the calibration sample (school 1) and then used the results to define moderately informative priors for the analysis sample (school 2). In particular, they performed model selection for the analysis sample using Bayes factors, ending up with a preferred model ('Model 4') among a set of five possible models. Here we replicate their analysis using Laplace approaches but focusing only on the analysis sample. The results showed that importance sampling performed

Table 5.1: Maximized log-likelihood values (AIC values) for five models of interest.

Method	Full model	Model 2	Model 3	Model 4	Empty model
$\ell^*(\theta)$	-252.8	-253.9	-263.8	-256.3	-271.8
	(539.6)	(531.8)	(543.6)	(530.6)	(553.5)
$\ell^\dagger(\theta),\ \ M = 5,000$	-253.6	-254.6	-264.5	-257.0	-272.6
	(541.1)	(533.3)	(545.0)	(532.1)	(555.3)
$\ell^\dagger(\theta),\ \ M = 20,000$	-253.5	-254.6	-264.5	-257.0	-272.6
	(541.1)	(533.3)	(545.0)	(532.1)	(555.2)

a modest adjustment to the standard Laplace approximation, especially for what concerns estimation of variance components. Approximated maximum likelihood estimates obtained from $L^\dagger(\theta)$ stabilized very quickly with the value of M, and actually we found little variation in both the estimates and the maximized likelihoods obtained with M in the range 5,000-50,000. Table 5.1 reports the maximized log-likelihood values and AIC values for the five models defined in Zijlstra et al. (2005). It is somewhat reassuring that likelihood-based model selection pinpoints the same model selected by the Bayesian procedure used by Zijlstra et al. (2005), as Model 4 is the one with the lowest AIC value with either $\ell^*(\theta)$ or $\ell^\dagger(\theta)$.

Table 5.2 reports the estimation results for Model 4 for the various methods, along with the results reported in Zijlstra et al. (2005). The only actor-specific covariate is the gender dummy variable, where boys have code one and girls code zero, while the dyadic covariates are dummy variables which are equal to 1 in case when two pupils have the same gender or ethnic background.

We notice that whereas the likelihood-based results are very similar, there are some differences with respect to the Bayesian results of Zijlstra et al. (2005). To some extent, this has to be expected as the latter used also some information from the calibration sample. Broadly speaking, the resulting inference is quite the same with either approach, and we conclude

Table 5.2: Estimation results for Model 4.

		$\ell^*(\theta)$	$\ell^\dagger(\theta)$, $M = 20,000$	Posterior
Effect	Covariate	Estimate (s.e.)	Estimate (s.e.)	Mean (s.d.)
Sender	Gender	1.10 (0.45)	1.11 (0.45)	1.00 (0.42)
Receiver	Gender	-0.89 (0.42)	-0.90 (0.42)	-0.90 (0.39)
Density		-5.18 (0.39)	-5.16 (0.39)	-4.23 (0.32)
	Gender	0.76 (0.23)	0.77 (0.23)	0.83 (0.23)
	Ethnic background	0.74 (0.19)	0.74 (0.19)	1.08 (0.21)
Reciprocity		5.20 (0.65)	5.16 (0.67)	4.70 (0.59)
Sender Variance	σ_A^2	0.87 (0.42)	0.84 (0.42)	0.82 (0.42)
Receiver Variance	σ_B^2	0.64 (0.34)	0.60 (0.34)	0.58 (0.35)
Covariance	σ_{AB}	-0.75 (0.34)	-0.71 (0.35)	-0.45 (0.34)

that boys more often than girls report having received emotional support, while emotional support is less often reported to come from boys. More received emotional support is reported from pupils with the same gender and ethnic background.

5.3.2 Lazega Lawyers

For the Lazega's associates friendship network (§2.3.4), we considered the three models analyzed in Van Duijn et al. (2004). In particular Model 0 is an "empty" model composed by μ, ρ and the variance terms, Model 1 and Model 2 contain also some terms for density and reciprocity effects. Moreover, the networks on advice and collaboration are set as covariates for the density parameter in Model 2. More details on the model specification can be found in Van Duijn et al. (2004).

Table 5.3 reports the MQL estimates of the original paper (Table 1, Van Duijn et al., 2004), and the estimates obtained with the Laplace methods. The importance sampling result is obtained with $M = 10,000$. Note that for Model 1 one density covariate has been dropped due to collinearity problems, therefore the comparisons with the results of Van Duijn et al. (2004) has to be taken with some care.

We note that MQL estimates are generally attenuated with respect to the other two methods. The results for the Laplace and Laplace Importance Sampling methods are instead quite similar.

5.4 Simulation Study

We investigate the properties of the proposed methodology by a simulation study. In particular, we replicate closely the simulation study given in Zijlstra et al. (2009). These authors considered three model setups for two network orders (20 and 40 nodes). Following their description, Model

Table 5.3: Estimation results for Lazega's associates friendship network.

		MQL	Laplace	Laplace IS
Effect	Covariate	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Model 0				
Density	μ	-2.70 (0.23)	-3.43 (0.30)	-3.43 (0.31)
Reciprocity	ρ	3.29 (0.31)	4.04 (0.46)	4.02 (0.45)
Sender Variance	σ_A^2	1.08 (0.25)	1.11 (0.41)	1.15 (0.43)
Receiver Variance	σ_B^2	0.75 (0.19)	0.69 (0.30)	0.71 (0.31)
Covariance	σ_{AB}	-0.33 (0.16)	-0.15 (0.27)	-0.13 (0.28)
Model 1				
Density	μ	-0.64 (0.35)	-1.09 (0.40)	-1.08 (0.41)
	Office	-2.33 (0.43)	-2.76 (0.50)	-2.76 (0.50)
	Seniority	-0.58 (0.09)	-0.62 (0.09)	-0.62 (0.09)
	Gender	-0.55 (0.17)	-0.70 (0.19)	-0.70 (0.19)
	Specialty	-0.51 (0.17)	-0.59 (0.19)	-0.59 (0.19)
Reciprocity	ρ	2.21 (0.36)	3.00 (0.44)	2.97 (0.48)
	Office	1.72 (0.94)	1.67 (1.01)	1.68 (1.01)
Sender Variance	σ_A^2	1.19 (0.27)	1.47 (0.54)	1.49 (0.55)
Receiver Variance	σ_B^2	0.63 (0.17)	0.68 (0.32)	0.68 (0.32)
Covariance	σ_{AB}	-0.01 (0.16)	0.003 (0.20)	0.03 (0.31)
Model 2				
Density	μ	-1.62 (0.37)	-1.98 (0.42)	-1.97 (0.42)
	Location	-1.45 (0.31)	-2.08 (0.36)	-2.10 (0.36)
	Seniority	-0.49 (0.09)	-0.53 (0.10)	-0.53 (0.10)
	Gender	-0.58 (0.18)	-0.83 (0.21)	-0.83 (0.21)
	Advise	1.50 (0.21)	1.44 (0.28)	1.44 (0.28)
	Cowork	0.53 (0.24)	1.05 (0.26)	1.06 (0.26)
Reciprocity	ρ	2.22 (0.35)	2.90 (0.49)	2.87 (0.49)
Sender Variance	σ_A^2	1.36 (0.30)	1.70 (0.62)	1.73 (0.64)
Receiver Variance	σ_B^2	0.65 (0.18)	0.86 (0.40)	0.86 (0.41)
Covariance	σ_{AB}	0.16 (0.17)	0.05 (0.36)	0.08 (0.37)

1 is an empty model with density and reciprocity parameters respectively $\mu = -2$ and $\rho = 2$ and independent standardized random effects. Model 2 has also a dyadic covariate for the density Z_1 and one sender covariate X_1. The density covariate Z_1, with regression parameter 0.5, is a network (*net1*) generated from Model 1. The sender covariate X_1 has a regression value 0.05 and it is the actor's rank number $(1, \ldots, g)$. Model 3 has a receiver covariates X_2, two density covariates Z_1, and one reciprocity covariate Z_2. The receiver covariate X_2 has coefficient -0.1 and it is a binary variable (0,1) drawn from a coin flip. The first component of Z_1 is *net1*, as in Model 2, and it has coefficient 0.5. The second density covariate (*fc*) contains the absolute differences of an actor covariate obtained as a multinomial distribution

with 5 equal probabilities; this variable is also used as reciprocity covariate. The coefficient values for the *fc* covariate are equal to 0.2 and 0.05, in the density and reciprocity case respectively. The random effects in Model 3 are negatively correlated ($\sigma_{AB} = -0.5$), with the sender variance $\sigma^2_A = 1.5$, and receiver variance $\sigma^2_B = 0.75$.

The results for the simulation studies refer to the Laplace-based methods, and an MQL ("RIGLS-3") and a Bayesian ("Random Walk") algorithm taken from Zijlstra et al. (2009).

Table 5.4 reports the sample mean of parameter estimates and the mean standard errors (standard deviation for the Bayesian method) over 1,000 replications. Estimated bias and root mean squared errors of the various estimators are reported in Table 5.5. Furthermore, Table 5.6 summarizes the sample mean of standard errors and the standard deviation for the estimates obtained with the two methods proposed here.

The standard errors for the two approximate maximum likelihood estimates were computed using the observed information matrix.

From these results we can say that the MQL approach produces estimates with very large bias, that renders such method totally unappealing for practical use. This is not surprising, and relevant with a vast body of literature on random effects modeling. Both the Bayesian approach and our proposals seem instead to perform well, with no appreciable differences between the Laplace and the Laplace Importance Sampling methods. At times the two approximated maximum likelihood estimation methods appear to be slightly more efficient than the Bayesian method, but in general we can say that they are largely comparable. Finally, the standard errors estimated using the observed information matrix for the two Laplace-based methods appear quite accurate, although for quite a few simulated networks such matrix was not positive definite and the standard errors could not be computed.

Table 5.4: Sample mean and standard deviation of parameter estimates in 1,000 simulated data sets.

Method	MQL		Bayesian		Laplace		Laplace IS		MQL		Bayesian		Laplace		Laplace IS	
	Mean	s.d	Mean	s.d	Mean	s.d	Mean	s.d	Mean	s.d	Mean	s.d	Mean	s.d	Mean	s.d
	20 Nodes								40 Nodes							
Model 1																
$\sigma_A^2 = 1$	0.86	0.27	1.01	0.49	1.00	0.52	1.04	0.53	0.72	0.14	1.00	0.28	0.99	0.29	1.00	0.29
$\sigma_B^2 = 1$	0.86	0.27	1.01	0.45	0.99	0.51	1.03	0.53	0.73	0.14	1.02	0.30	0.98	0.29	0.99	0.29
$\sigma_{AB} = 0$	-0.17	0.19	0.03	0.35	-0.02	0.40	-0.01	0.40	-0.14	0.10	-0.00	0.21	-0.02	0.21	-0.01	0.21
$\mu = -2$	-1.61	0.32	-1.97	0.39	-2.02	0.40	-2.02	0.41	-1.61	0.19	-1.99	0.26	-2.02	0.25	-2.02	0.25
$\rho = 2$	1.80	0.37	1.93	0.50	2.02	0.55	2.01	0.54	1.84	0.18	1.99	0.26	2.02	0.26	2.02	0.26
Model 2																
$\sigma_A^2 = 1$	0.83	0.26	1.06	0.50	0.91	0.48	0.94	0.47	0.70	0.14	1.03	0.29	0.96	0.28	0.98	0.28
$\sigma_B^2 = 1$	0.81	0.25	1.06	0.48	0.99	0.53	1.02	0.49	0.71	0.14	1.03	0.29	1.01	0.28	1.02	0.28
$\sigma_{AB} = 0$	-0.15	0.18	0.02	0.36	-0.03	0.37	-0.01	0.37	-0.12	0.10	-0.00	0.21	-0.03	0.20	-0.03	0.21
$\mu = -2$	-1.61	0.54	-1.97	0.67	-2.02	0.61	-2.02	0.62	-1.63	0.33	-1.99	0.42	-1.99	0.40	-1.99	0.40
$\rho = 2$	1.78	0.35	1.95	0.47	2.00	0.49	1.99	0.49	1.82	0.18	2.00	0.23	2.01	0.23	2.00	0.23
$\gamma_1 = 0.05$	0.04	0.04	0.05	0.05	0.05	0.04	0.05	0.04	0.04	0.01	0.05	0.02	0.05	0.01	0.05	0.01
$\delta_1 = 0.5$	0.36	0.27	0.50	0.31	0.50	0.27	0.50	0.27	0.04	0.15	0.51	0.15	0.50	0.14	0.50	0.14
Model 3																
$\sigma_A^2 = 1.5$	1.06	0.30	1.61	0.77	1.50	0.68	1.53	0.69	0.93	0.17	1.52	0.41	1.48	0.40	1.49	0.40
$\sigma_B^2 = 0.75$	0.64	0.21	0.85	0.41	0.73	0.40	0.75	0.41	0.51	0.10	0.78	0.23	0.73	0.22	0.74	0.22
$\sigma_{AB} = -0.5$	-0.25	0.18	-0.46	0.45	-0.54	0.43	-0.53	0.44	-0.25	0.10	-0.49	0.24	-0.52	0.24	-0.52	0.24
$\mu = -2$	-1.48	0.45	-2.01	0.52	-2.02	0.54	-2.02	0.55	-1.42	0.25	-2.00	0.30	-1.99	0.29	-1.99	0.29
$\rho = 2$	1.26	0.56	1.90	0.71	2.02	0.71	2.00	0.72	1.29	0.27	1.98	0.35	1.99	0.34	1.99	0.34
$\gamma_2 = -0.1$	-0.10	0.40	-0.10	0.46	-0.09	0.45	-0.09	0.45	-0.08	0.23	-0.10	0.28	-0.10	0.27	-0.10	0.27
$\delta_1(fc) = 0.2$	0.16	0.17	0.21	0.19	0.21	0.16	0.21	0.16	0.15	0.08	0.20	0.09	0.20	0.07	0.20	0.07
$\delta_1(net1) = 0.5$	0.39	0.26	0.52	0.32	0.49	0.26	0.49	0.26	0.38	0.12	0.50	0.15	0.51	0.13	0.51	0.13
$\delta_2(fc) = 0.05$	0.02	0.30	0.03	0.03	0.04	0.29	0.04	0.29	0.02	0.14	0.02	0.16	0.06	0.14	0.06	0.14

Table 5.5: Bias and root mean squared errors (RMSEs) from 1,000 simulated data sets.

Method	MQL		Bayesian		Laplace		Laplace IS		MQL		Bayesian		Laplace		Laplace IS	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
	20 Nodes								40 Nodes							
Model 1																
$\sigma_A^2 = 1$	-0.140	0.315	0.011	0.494	0.002	0.574	0.042	0.595	-0.275	0.316	-0.002	0.279	-0.010	0.296	0.005	0.300
$\sigma_B^2 = 1$	-0.140	0.302	0.006	0.453	-0.010	0.550	0.029	0.571	-0.268	0.313	0.017	0.298	-0.020	0.284	-0.005	0.287
$\sigma_{AB} = 0$	-0.165	0.212	0.030	0.349	-0.023	0.419	-0.006	0.419	-0.140	0.158	-0.001	0.214	-0.019	0.215	-0.014	0.216
$\mu = -2$	0.388	0.490	0.033	0.389	-0.020	0.426	-0.021	0.425	0.389	0.436	0.008	0.259	-0.020	0.248	-0.020	0.248
$\rho = 2$	-0.198	0.499	-0.074	0.504	0.019	0.564	0.007	0.555	-0.163	3.050	-0.015	0.257	0.022	0.261	0.018	0.259
Model 2																
$\sigma_A^2 = 1$	-0.170	0.300	0.056	0.505	-0.088	0.517	-0.059	0.526	-0.300	0.336	0.027	0.295	-0.035	0.291	-0.024	0.292
$\sigma_B^2 = 1$	-0.187	0.298	0.065	0.482	-0.008	0.481	0.022	0.493	-0.292	0.332	0.034	0.293	0.006	0.283	0.016	0.285
$\sigma_{AB} = 0$	-0.150	0.202	0.021	0.363	-0.026	0.376	-0.015	0.376	-0.116	0.138	-0.003	0.214	-0.030	0.221	-0.026	0.221
$\mu = -2$	0.392	0.638	0.035	0.667	-0.018	0.663	-0.017	0.663	0.368	0.476	0.009	0.424	0.009	0.413	0.010	0.413
$\rho = 2$	-0.219	0.480	-0.047	0.474	-0.004	0.479	-0.012	0.473	-0.175	0.307	-0.005	0.230	0.008	0.238	0.005	0.237
$\gamma_1 = 0.05$	-0.015	0.032	-0.001	0.045	0.002	0.046	0.002	0.046	-0.015	0.018	0.000	0.016	0.000	0.015	0.000	0.015
$\delta_1 = 0.5$	-0.141	0.274	0.000	0.308	-0.002	0.275	-0.002	0.275	-0.463	0.182	0.008	0.152	-0.001	0.138	-0.001	0.138
Model 3																
$\sigma_A^2 = 1.5$	-0.439	0.522	0.107	0.776	0.003	0.720	0.034	0.734	-0.571	0.593	0.020	0.411	-0.018	0.407	-0.007	0.409
$\sigma_B^2 = 0.75$	-0.111	0.230	0.101	0.427	-0.017	0.415	0.002	0.419	-0.242	0.263	0.028	0.230	-0.018	0.224	-0.012	0.224
$\sigma_{AB} = -0.5$	0.254	0.285	0.036	0.446	-0.037	0.457	-0.028	0.456	0.247	0.255	0.006	0.243	-0.019	0.254	-0.017	0.254
$\mu = -2$	0.518	0.706	-0.007	0.522	-0.021	0.577	-0.018	0.578	0.585	0.636	0.002	0.302	0.010	0.302	0.011	0.302
$\rho = 2$	-0.739	0.994	-0.098	0.720	0.016	0.730	0.001	0.729	-0.710	0.792	-0.019	0.349	-0.008	0.350	-0.013	0.350
$\gamma_2 = -0.1$	0.005	0.349	-0.002	0.458	0.011	0.484	0.011	0.484	0.017	0.212	-0.001	0.283	-0.002	0.281	-0.002	0.281
$\delta_1(fc) = 0.2$	-0.036	0.167	0.013	0.190	0.013	0.165	0.013	0.165	-0.045	0.089	0.002	0.089	-0.002	0.076	-0.003	0.076
$\delta_1(net1) = 0.5$	-0.111	0.274	0.017	0.318	-0.009	0.268	-0.008	0.268	-0.117	0.164	0.004	0.148	0.006	0.138	0.005	0.138
$\delta_2(fc) = 0.05$	-0.027	0.321	-0.022	0.332	-0.015	0.307	-0.015	0.308	-0.031	0.155	-0.027	0.161	0.007	0.138	0.007	0.138

Table 5.6: Sample mean of standard errors and standard deviation for the two Laplace-based methods.

Method	Laplace		Laplace IS		Laplace		Laplace IS	
	Mean s.e	s.d	Mean s.e	s.d	Mean s.e	s.d	Mean s.e	s.d
	20 Nodes				40 Nodes			
Model 1								
$\sigma_A^2 = 1$	0.515	0.575	0.534	0.594	0.289	0.296	0.294	0.300
$\sigma_B^2 = 1$	0.510	0.550	0.529	0.570	0.287	0.284	0.292	0.287
$\sigma_{AB} = 0$	0.395	0.419	0.402	0.419	0.210	0.215	0.213	0.215
$\mu = -2$	0.402	0.426	0.408	0.425	0.250	0.247	0.252	0.248
$\rho = 2$	0.548	0.564	0.544	0.555	0.255	0.260	0.255	0.259
Model 2								
$\sigma_A^2 = 1$	0.477	0.510	0.467	0.523	0.276	0.289	0.276	0.291
$\sigma_B^2 = 1$	0.534	0.482	0.492	0.493	0.276	0.283	0.284	0.285
$\sigma_{AB} = 0$	0.374	0.375	0.373	0.375	0.205	0.219	0.207	0.219
$\mu = -2$	0.609	0.663	0.617	0.663	0.398	0.413	0.400	0.413
$\rho = 2$	0.492	0.480	0.489	0.474	0.234	0.238	0.234	0.237
$\gamma_1 = 0.05$	0.043	0.046	0.043	0.046	0.015	0.015	0.015	0.015
$\delta_1 = 0.5$	0.265	0.275	0.266	0.275	0.139	0.138	0.139	0.138
Model 3								
$\sigma_A^2 = 1.5$	0.676	0.720	0.691	0.733	0.399	0.407	0.403	0.409
$\sigma_B^2 = 0.75$	0.402	0.415	0.410	0.419	0.220	0.223	0.222	0.224
$\sigma_{AB} = -0.5$	0.433	0.455	0.437	0.456	0.240	0.253	0.241	0.254
$\mu = -2$	0.541	0.577	0.549	0.578	0.291	0.302	0.293	0.302
$\rho = 2$	0.713	0.730	0.716	0.729	0.345	0.350	0.345	0.349
$\gamma_2 = -0.1$	0.445	0.484	0.452	0.484	0.267	0.281	0.269	0.281
$\delta_1(fc) = 0.2$	0.156	0.165	0.157	0.165	0.074	0.076	0.075	0.076
$\delta_1(net1) = 0.5$	0.260	0.268	0.262	0.268	0.133	0.138	0.133	0.138
$\delta_2(fc) = 0.05$	0.289	0.307	0.291	0.307	0.135	0.138	0.135	0.138

5.4.1 Discussion

The results obtained with the approximate maximum likelihood estimation methods based on the Laplace approximation for the class of p_2 models are rather encouraging. Indeed, even the simple simulation-free approach given by the Laplace approximation seems to perform well, correcting the drawbacks of MQL-type estimation. The real advantage of our proposal with respect to the Bayesian approach is mainly the simplicity of usage, with no need of MCMC tuning. On the other hand, with the Bayesian approach is simpler to incorporate some sort of prior information when this is available, and the use of slightly-informative priors may help in those cases when the estimated matrix of random effects is close to singularity.

For practical implementation, Bayesian methods can be implemented using some publicly-available software, such as the BUGS engine (see Lunn

et al., 2000). Notice, however, that also the methods proposed here can be simply implemented using the freely-available ADMB software (see Fournier et al., 2011). Both the Bayesian approach and the Laplace-based ones can be extended to more complex data structures, such as the multilevel data set studied in Vermeij et al. (2009). We remark, however, that the required computational burden appears less severe for the frequentist approach proposed here.

Appendix A

Composite Likelihood Estimation for ERGMs

In this appendix, we briefly report some simple attempts to obtain simulation-free procedures for ERGM estimation. All these attempts are based on a composite likelihood approach and they extend the pseudolikelihood estimator. Here we focus only on the undirected case, with an illustrative example.

Consider an n-dimensional random variable $Y = (Y_1, \ldots, Y_n)$ with a joint density function $f(y; \boldsymbol{\theta})$ for some parameter $\theta \in \Theta \subseteq \mathcal{R}^p$ that is supposed unknown. The composite likelihood approach provides consistent estimation of θ in the case when for some reason $f(y; \boldsymbol{\theta})$ is not so easy to manage, but computing the likelihood function for a subset of y it is possible (Besag, 1974; Lindsay, 1988). The following definition is taken from Varin (2008).

Definition: *Consider a parametric statistical model* $\{f(y; \boldsymbol{\theta}), y \in \mathcal{Y} \subseteq \mathbb{R}^n, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$ *and a set of measurable events* $\{A_i; i = 1, \ldots, m\}$*. Then, a composite likelihood (CL) is the weighted product of the likelihoods corresponding to each single event,*

$$L_C(\boldsymbol{\theta}; y) = \prod_{i=1}^{m} f(y \in A_i; \boldsymbol{\theta})^{w_i}, \tag{A.1}$$

where w_i*,* $i = 1, \ldots, m$ *are positive weights. The associated composite log-likelihood is* $\ell_c(\boldsymbol{\theta}; y) = log\, L_C(\boldsymbol{\theta}; y)$ *and its maximum, if unique, is the maximum composite likelihood estimator (MCL).*

It is then possible to construct a pseudolikelihood by combining such likelihood objects and use it as a surrogate for the ordinary likelihood. These simplifications can be done using marginal or conditional distributions of the subsets of data (Varin, 2008).

For an ERGM

$$P(Y = y; \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}\, u(y) - \psi(\boldsymbol{\theta})\right\},$$

composite likelihood approaches can be obtained relaxing the Markovian dependence constrain that involve the dyads.

The maximum pseudolikelihood estimator (MPLE) (Strauss and Ikeda, 1990), presented in §2.2.3, is already an example of composite likelihood estimator, where each dyad value defines one of the events A_i.

It is possible to obtain others composite likelihood approaches increasing the number of the elements in A_i. These approaches are called *Maximum block-pseudolikelihood estimation (MBPLE)* (Rydén and Titterington, 1998; Friel et al., 2009). They request to split the dyad set in B disjoint subsets of size b. The resulting expression for the composite likelihood function is

$$L_{BP}(\boldsymbol{\theta}; y) = \prod_{b=1}^{B} P(Y_b = y_b | y_{-b}; \boldsymbol{\theta}), \tag{A.2}$$

where y_{-b} is the complementary network, i.e. the part of the observed network excluding the element contained in y_b. The y_b should be chose such that $\bigcup_{b=1}^{B} y_b = y$ and $\bigcap_{b=1}^{B} y_b = \varnothing$.

Our idea

The MBPLE approach attempts to capture larger interactions within the graph, but it requires, first, that the number of the dyads should be a multiple of the block size b. Second, the blocks must be chosen such that they belong to disjoint sets of elements. Third, there may exist several configurations that respect the block-disjoint condition and not all of them could be equally informative for inference on θ.

Our proposal is to relax the block-disjoint condition and to consider only blocks of dyads (pairs, triplets, quartets) in which the dyads share at least one node. So we focus only on blocks that should be inferentially informative. The expression for the composite likelihood function is the same as formula (A.2), the only difference is the number B that can be much larger.

The main feature for the implementation of this kind of composite likelihood method is how to compute a single term

$$P(Y_b = y_b | y_{-b}; \boldsymbol{\theta}).$$

Starting from the case where $b = 2$, the informative pair of dyads in the graph correspond to the set of the two stars.

The single element

$$P(Y_b = y_b | y_{-b}; \boldsymbol{\theta}) = \frac{P(y; \boldsymbol{\theta})}{P(y_b^{++}; \boldsymbol{\theta}) + P(y_b^{+-}; \boldsymbol{\theta}) + P(y_b^{-+}; \boldsymbol{\theta}) + P(y_b^{--}; \boldsymbol{\theta})} \tag{A.3}$$

corresponds to a multinomial distribution, where y is the observed network, $y_b = (y_{ij}, y_{ik})$ for $1 \leq i < j < k \leq n$ is the considered block, and y_b^{++}, y_b^{+-}, y_b^{-+}, y_b^{--} are the networks where the elements of the block y_b are set to 1 if + or to 0 if − (e.g: y_{ijk}^{+-} correspond to the block with the dyads $y_{ij} = 1$ and $y_{ik} = 0$).

To obtain an explicit form for (A.3) consider the following expression

$$\frac{P(y;\boldsymbol{\theta})}{P(y_b^{--};\boldsymbol{\theta})} = \exp\{\Delta(y)\},$$

where

$$\Delta(y) = \begin{cases} 0 & \text{if } y = y_b^{--} \\ u(y_b^{+-}) - u(y_b^{--}) & \text{if } y = y_b^{+-} \\ u(y_b^{-+}) - u(y_b^{--}) & \text{if } y = y_b^{-+} \\ u(y_b^{++}) - u(y_b^{--}) & \text{if } y = y_b^{++} \end{cases} \tag{A.4}$$

is equivalent to the network change statistic in formula (2.28).

The final result is that

$$\begin{aligned} P(Y_b = y_b | y_{-b};\boldsymbol{\theta}) &= \frac{P(y;\boldsymbol{\theta})}{P(y_b^{--};\boldsymbol{\theta})} \frac{P(y_b^{--};\boldsymbol{\theta})}{P(y_b^{++};\boldsymbol{\theta}) + P(y_b^{+-};\boldsymbol{\theta}) + P(y_b^{-+};\boldsymbol{\theta}) + P(y_b^{--};\boldsymbol{\theta})} \\ &= \frac{\exp\{\Delta(y)\}}{\sum_{y^* \in \mathcal{Y}_b} \exp\{\Delta(y^*)\}}, \end{aligned} \tag{A.5}$$

where $\mathcal{Y}_b$ is the set of all the networks that differ only for the value of the block y_b.

Formula (A.5) gives a general rule to compute the composite likelihood function elements for larger block size ($b = 3, 4, \ldots$); see also Asuncion et al. (2010).

Once obtained a general method to compute the composite likelihood function terms, we need to determine which blocks are informative for inference on θ. We considered a block as informative if the dyads that it contains share at least one node. In Figure A.1 the informative blocks that we considered are presented.

For $b = 2$, we chose the set of all the possible 2-stars (or equivalently the 2-paths).

For $b = 3$ and $1 \leq i < j < k < h \leq g$, we considered as informative blocks the set of the triads (y_{ij}, y_{ik}, y_{jk}),

For $b = 4$ and $1 \leq i < j < k < h \leq g$, we considered as informative the blocks $(y_{ij}, y_{jk}, y_{kh}, y_{hi})$ that form a 4-cycle.

Example

Here we report the results for a simple model, already presented in §3.5 (Model 1), for Lazega's Lawyers data on the network of collaborative partners. Table A.1 shows estimates and standard errors obtained by the composite likelihood method. The table reports also the MLE values of the

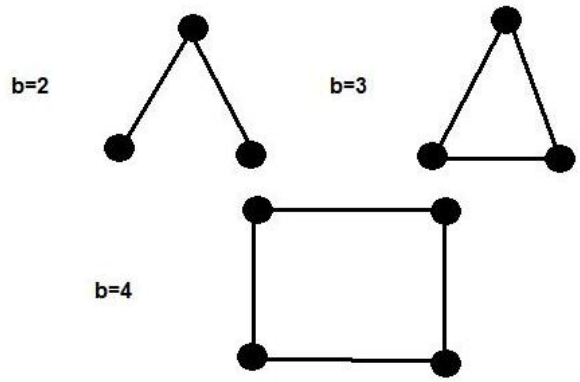

Figure A.1: Structures considered as informative blocks for $b = 2, 3, 4$

model, given by the estimated values obtained by the MC-BFGS method of §3, and already reported in Table 3.4.

In the example, we considered the estimates for informative blocks size b=2, 3, and 4 (CL-2, 3, and 4). The standard errors reported are corrected using sandwich estimator computed by Monte Carlo simulation.

The results of the various composite likelihood estimation are very similar, and not close enough to the MLE. There is a slightly improvement in terms of efficiency only for the CL-4 estimates, but the standard errors of the MLE are considerably smaller.

There is the possibility that larger blocks would provide a more precise estimation method. At any rate, increasing the block size would require a substantial amount of additional work, also considering that it would be difficult to select an informative block configuration for block size larger than 4. For these reasons this approach does not appear as a practical solution for the estimation of ERGMs parameters.

Table A.1: Parameters estimation and standard errors for composite methods on collaboration partner network of Lazega's Lawyers.

Method	MLE		MPLE		CL-2		CL-3		CL-4	
	Estimation	s.e	Estimation	s.e	Estimation	s.e	Estimation	s.e	Estimation	s.e
edges	-6.97	0.83	-8.10	2.20	-8.10	2.25	-8.10	2.24	-8.39	2.11
kstar2	0.20	0.09	0.27	0.19	0.27	0.16	0.27	0.16	0.26	0.14
kstar3	-0.03	0.01	-0.02	0.02	-0.02	0.02	-0.02	0.02	-0.02	0.01
triangle	0.34	0.12	0.30	0.10	0.30	0.08	0.30	0.10	0.28	0.09
nodecov.sen36	1.04	0.26	0.75	0.66	0.75	0.67	0.75	0.66	0.84	0.56
nodecov.specialty	0.44	0.12	0.26	0.33	0.26	0.33	0.26	0.33	0.31	0.29
nodematch.specialty	0.82	0.22	0.85	0.53	0.85	0.51	0.85	0.47	0.89	0.42
nodematch.gender	0.85	0.27	0.70	0.67	0.70	0.68	0.70	0.63	0.82	0.60
nodematch.office	1.28	0.24	1.58	0.81	1.58	0.67	1.58	0.77	1.62	0.70

Bibliography

Anderson, C., Wasserman, S. and Faust, K. (2002). Building stochastic blockmodels. *Social Networks: Critical Concepts in Sociology* **14**, 227.

Asuncion, A., Liu, Q., Ihler, A. and Smyth, P. (2010). Learning with blocks: Composite likelihood and contrastive divergence. In *Conference on Uncertainty in Artificial Intelligence (AISTATS)*.

Baerveldt, C. and Snijders, T. (1994). Influences on and from the segmentation of networks: Hypotheses and tests. *Social Networks* **16**, 213–232.

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

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

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Brinch, C. (2012). Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling. *Computational Statistics* **27**, 13–28.

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

Cortes, C. and Mohri, M. (2005). Confidence intervals for the area under the roc curve. *Analysis* **17**, 305–312.

De Nooy, W., Mrvar, A. and Batagelj, V. (2011). *Exploratory social network analysis with Pajek*. Cambridge University Press, New York.

Dennis, J. and Schnabel, R. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. Society for Industrial Mathematics, Philadelphia.

Erdös, P. and Rényi, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6**, 290–297.

Fletcher, R. (1980). *Practical methods of optimization. Vol. 1.* John Wiley & Sons Ltd., Chichester. Unconstrained optimization, A Wiley-Interscience Publication.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A. and Sibert, J. (2011). Ad model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **0**, 1–17.

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

Friel, N., Pettitt, A., Reeves, R. and Wit, E. (2009). Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics* **18**, 243–261.

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

Handcock, M. and Gile, K. (2010). Modeling social networks from sampled data. *The Annals of Applied Statistics* **4**, 5–25.

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

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M. and Morris, M. (2008). statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software* **24**, 1–11.

Handcock, M. S., Robins, G., Snijders, T., Moody, J. and Besag, J. (2003). Assessing degeneracy in statistical models of social networks. *Journal of the American Statistical Association* **76**, 33–50.

Hanley, J. and McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Hanley, J. and McNeil, B. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843.

Hoff, P. (2005). Bilinear mixed effects models for dyadic data. *Journal of the American Statistical Association* **100**, 286–295.

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

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76**, 33–65. (with discussion).

Hummel, R., Hunter, D. and Handcock, M. (2011). Improving simulation-based algorithms for fitting ergms. *Journal of Computational and Graphical Statistics* (In press).

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

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

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

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M. (2008b). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software* **24**, 1–29.

Kapferer, B. (1972). *Strategy and Transaction in an African Factory: African Workers and Indian Management in a Zambian Town*. Manchester University Press, Manchester.

Kolaczyk, E. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Verlag, New York.

Koskinen, J., Robins, G. and Pattison, P. (2010). Analysing exponential random graph (p-star) models with missing data using bayesian data augmentation. *Statistical Methodology* **7**, 366–384.

Krackhardt, D. (1987). Cognitive social structures. *Social Networks* **9**, 109–134.

Krivitsky, P., Handcock, M., Raftery, A. and Hoff, P. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* **31**, 204–213.

Krivitsky, P. N. and Handcock, M. S. (2008). Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software* **24**, 1–23.

Lazarsfeld, P. and Henry, N. (1968). *Latent structure analysis*. Houghton Mifflin Co., New York.

Lazega, E. (2001). *The Collegial Phenomenon: The Social Mechanisms of Cooperation among Peers in a Corporate Law Partnership*. Oxford University Press, Oxford.

Lazega, E. and Pattison, P. (1999). Multiplexity, generalized exchange and cooperation in organizations: a case study. *Social Networks* **21**, 67–90.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221–239.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer series in statistics. Springer.

Moreno, J. (1946). Sociogram and sociomatrix. *Sociometry* **9**, 348–349.

Pattison, P. and Robins, G. (2002). Neighborhoodbased models for social networks. *Sociological Methodology* **32**, 301–337.

Rinaldo, A., Fienberg, S. and Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electronic Journal of Statistics* **3**, 446–484.

Robins, G., Pattison, P., Kalish, Y. and Lusher, D. (2007a). An introduction to exponential random graph (p*) models for social networks. *Social Networks* **29**, 173 – 191.

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

Rubin, D. (1976). Inference and missing data. *Biometrika* **63**, 581.

Rydén, T. and Titterington, D. (1998). Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics* **7**, 194–211.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millán-Zárate, D., Díaz-Peredo, E., Sánchez-Solano, F., Pérez-Rueda, E., Bonavides-Martínez, C. and Collado-Vides, J. (2001). Regulondb (version 3.2): transcriptional regulation and operon organization in escherichia coli k-12. *Nucleic acids research* **29**, 72–74.

Sampson, S. (1968). A novitiate in a period of change. *An Experimental and Case Study of Social Relationships (PhD thesis). Cornell University, Ithaca* .

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association* (In press).

Shen-Orr, S., Milo, R., Mangan, S. and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics* **31**, 64–68.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B: Methodological* **57**, 749–760.

Skaug, H. (2002). Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. *Journal of Computational and Graphical Statistics* **11**, 458–470.

Skaug, H. and Fournier, D. (2006). Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. *Computational Statistics & Data Analysis* **51**, 699–709.

Snijders, T. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* **3**, 1–40.

Snijders, T. (2011). Statistical models for social networks. *Annual Review of Sociology* **37**, 129–151.

Snijders, T. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification* **14**, 75–100.

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

Song, P., Fan, Y. and Kalbfleisch, J. (2005). Maximization by parts in likelihood inference. *Journal of the American Statistical Association* **100**, 1145–1158.

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

Tanner, M. (1996). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions.* Springer Verlag, New York.

Thompson, S., Seber, G. et al. (1996). *Adaptive sampling.* Wiley, New York.

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

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

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis* **92**, 1–28.

Vermeij, L., Van Duijn, M. and Baerveldt, C. (2009). Ethnic segregation in context: Social discrimination among native dutch pupils and their ethnic minority classmates. *Social Networks* **31**, 230–239.

Wang, Y. and Wong, G. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association* **82**, 8–19.

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

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

Wasserman, S. and Robins, G. (2005). An introduction to random graphs, dependence graphs, and p*. In *Models and Methods in Social Network Analysis*, P. Carrington, J. Scott and S. Wasserman, eds. Cambridge University Press, New York, pp. 148–161.

Wasserman, S., Robins, G. and Steinley, D. (2007). Statistical models for networks: A brief review of some recent research. In *Statistical Network Analysis: Models, Issues, and New Directions*, E. Airoldi, D. Blei, S. Fienberg, A. Goldenberg, E. Xing and A. Zheng, eds. Springer-Verlag, New York, pp. 45–56.

White, H., Boorman, S. and Breiger, R. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology* **81**, 730–780.

Young, G. A. and Smith, R. L. (2005). *Essentials of statistical inference*. Cambridge University Press, Cambridge.

Zijlstra, B., Duijn, M. and Snijders, T. (2005). Model selection in random effects models for directed graphs using approximated bayes factors. *Statistica Neerlandica* **59**, 107–118.

Zijlstra, B., Duijn, M. and Snijders, T. (2009). MCMC estimation for the p_2 network regression model with crossed random effects. *British Journal of Mathematical and Statistical Psychology* **62**, 143–166.

Nicola Soriani

CURRICULUM VITAE

Personal Details

Date of Birth: February 22, 1982
Place of Birth: Badia Polesine (Rovigo), Italy
Nationality: Italian

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.
Tel. +39 049 827 4147
e-mail: soria@stat.unipd.it; nsoriani@yahoo.it

Current Position

Since September 2011;
Research Fellow, University of Trieste
Project title: "I metodi statistici per l'analisi rischi-benefici nella valutazione di strategie nutrizionali".
Supervisor: Prof. Nicola Torelli

Research interests

- Social Network Analysis
- Multivariate Analysis

Education

October 2004 – October 2007
Master (*laurea magistrale*) degree in Statistics and Informatics.
University of Padua, Faculty of Statistical Sciences
Title of dissertation: "La distribuzione t asimmetrica: analisi discriminante e regioni di tollerenza "

Supervisor: Prof. Adelchi Azzalini
Final mark: 108/110

October 2001 – October 2004
Bachelor degree (*laurea triennale***) in Statistics and Computing Technology**.
University of Padua, Faculty of Statistical Sciences
Title of dissertation: "Miglioramenti asintotici del log-rapporto di verosimiglianza con parametro di interesse multidimensionale "
Supervisor: Prof. Laura Ventura
Final mark: 110/110.

Visiting periods

January 2004 – July 2004
University of Aarhus, Department of Mathematics
Aarhus, Denmark.
Supervisor: Prof. Ole E. Barndorff-Nielsen

June 2009 – October 2009
University of Washington, Department of Statistics
Seeatle (WA), USA.
Supervisor: Prof. Mark Handcock

Awards and Scholarship

January 2008 - December 2010
Italian Ministry of University and Scientific Research: Three-year scholarship for Ph.d. studies at the University of Padova.

Computer skills

- R
- Linux
- LaTex
- SQL/PHP
- HTLM
- European Computer Driving Licence (ECDL)

Language skills

Italian: native; English: moderate; French: basic.

Conference presentations

Soriani, N., Handcock, M., (2010). ERGMs vs Latent Space Models: Comparing their goodness-of-fit for Kapferer's tailor shop network. (contributed) *Sunbelt XXX*, Riva del Garda (TN), Italy, June 29 – July 04, 2010.

Bellio, R., Soriani, N., (2011). A robust algorithm for maximum likelihood estimation of Exponential Random Graph Models, (contributed) *CLADAG 2011*, Pavia, Italy, September 7–9, 2011.

Bellio, R., Soriani, N., (2011). A Laplace approximation approach for p_2 network regretion models with crossed random effects. (contributed) *S.CO 2011*, Padova, Italy, September 21–23, 2011.

References

Prof. Ruggero Bellio
University of Udine
Department of Economics and Statistics
Via Treppo, 18
33100 Udine
Phone: +39 0432 249574
Fax: +39 0432 249595
e-mail: ruggero.bellio @ uniud.it

Prof. Susanna Zaccarin
University of Trieste
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche
'B. de Finetti'
P.le Europa 1
34127 Trieste
Phone: 040 5587021
e-mail: susannaz@econ.univ.trieste.it