Stockholms universitet

Exponential Family Random Graph Models - A Survey

Mohammed Motaher Hossain

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2010:3
Matematisk statistik
Juni 2010

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2010:3**
http://www.math.su.se

Exponential Family Random Graph Models - A Survey

Mohammed Motaher Hossain *

Juni 2010

Abstract

Statistical modeling of social networks as complex systems has always been and remains a challenge for social scientists. We review a wide class of exponential family models for social networks, known as Exponential Random Graph Models (ERGMs), or p* models. They have been developed since the 1980s and are characterized by well-defined sufficient statistics that represent local network characteristics. However, due to the difficulty of dealing with the intractable normalizing constant, pseudo-likelihood estimation methods have been applied in most studies. Recently, simulation based MCMC maximum likelihood estimation techniques have been introduced to improve parameter estimation. An R-package statnet has been developed for ERGMs by Goodreau et al. (2008). We illustrate some of the functionality of statnet by analysing a friendship network of 1,461 adolescents. It turns out that several well-studied ERGMs do not fit this data set well, although the fit improves dramatically when the models include another recently developed geometrically weighted edgewise shared partner (GWESP) statistic. KEYWORDS: Social networks, exponential-family random graph models, goodness-of-fit, pseudolikelihood estimation, Markov chain Monte Carlo, generalized linear models, deviance.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:motahersweden@yahoo.com . Supervisor: Ola Hössjer.

Preface

This report constitutes a one year Master's thesis (15 ECTS credits), for the degree of Master of Science in Mathematical Statistics at Stockholm University.

Contents

1 Introduction

The uses of social network models are becoming important in a number of fields such as epidemiology (with the emergence of infectious diseases like AIDS and SARS), business (with the study "viral marketing") and political science (with the study of coalition formation dynamics).

Examples of different dependence assumptions and their associated models are Bernoulli, dyad-independent and Markov random graph models. In this paper, we study the more general class of exponential random graph models (ERGMs), introduced by Frank and Strauss (1986) and Wasserman and Pattison (1996) and extended e.g. in Robins et al. (1999). See also Robins et al. (2007) and Hunter et al. (2008) for overviews. We focus in particular on a systematic examination of a real network dataset using maximum likelihood estimation and a new goodness-of-fit procedure of Hunter et al. (2008) to evaluate how well fitted models match observed data. These techniques compare structural statistics of the observed network with the corresponding statistics on networks simulated from our fitted models.

Parameter estimation is in general straightforward for simpler random graphs models. In the case of ERGMs, however, the maximum likelihood estimates for model parameters are in general more complicated, utilizing Markov chain Monte Carlo (MCMC) procedures. Degeneracy in fact often prevents model estimation from converging on finite parameter estimates. The use of alternative approximation techniques such as maximum pseudolikelihood estimation for ERGMs (Strauss and Ikeda, 1990) does not solve this problem, but simply hides it. We argue that several well-studied models in the networks literature do not fit these data well, and we demonstrate that the fit improves dramatically when the models include the recently-developed geometrically weighted edgewise shared partner (GWESP) and related statistics, see Sneijder et al. (2006).

The primary contribution of this article is to give an overview of a systematic approach to the assessment of network ERGMs. It contains tools for accomplishing three important and interrelated tasks involving estimation, simulation and goodness of fit. The goal is to estimate model parameters of a given social network data set collected at a nationally representative sample of high schools in the United States, see Udry (2003) and Harris et al. (2003). Then we will evaluate how adequately the model represents the data. To this end, we use the R package statnet for social network data described in Goodreau et al. (2008). It has the capability of approximating a maximum likelihood estimator for an ERGM data set; simulating new network data sets from a fitted ERGM and assessing how well a fitted ERGM captures aspects of observed data, e.g. clusters and degree distributions.

We conclude that ERGMs with GWESP statistics capture aspects of the social structure of adolescent friendship relations not represented by previous models.

In Section 2 we give an overview of random graphs, including in particular exponential family random graph models (ERGMs). Inference for such models is described in Section 3, and an example data set is analyzed in Section 4. A final discussion can be found in Section 5, whereas details on generalized linear models are given in the appendix.

2 Random Graphs

2.1 Basic Definitions

A random graph G is a graph which is generated by some random procedure. It is usually denoted as $G = (V, E)$ where V is a set of n vertices, connected by m edges E. It is obtained by starting with the set of n vertices which is fixed and then adding edges at random according to some rule. Different random graph models produce different probability distribution on graphs.

My aim in using exponential random graph models (ERGMs) is to model the random behavior of the adjacency matrix

$$\boldsymbol{Y} = (\boldsymbol{Y}_{ij})_{i,j=1}^{n}.$$

This is a square matrix such that $Y_{ij} = 0$ if there is no edge between the pair of vertices (dyad) i and j and $Y_{ij} = 1$ if there is and edge between i and j. To each i we associate q attributes or covariates, represented within the vector $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iq})$. All these are gathered into a matrix

$$\boldsymbol{X} = (\boldsymbol{X}_i)_{i=1}^{n}$$

of attributes.

2.2 Exponential-Family Random Graph Models

For social networks, we argue that, due to very recent progress in the framework of exponential random graph models, we are now much closer to the goal of obtaining good statistical models for social networks than we have ever been before. For modelling and studying social netwoks, the concept of an exponential random graph model (ERGM) has become an important tool with interesting theoretical achievments since the 1980s. The ERGMs, also known as $p^{\star}$-models, are a class of stochastic models which use network local structures to model the formation of network ties for a network with a fixed number of nodes. Depending on the underlying neighbourhood assumptions, ERGM assigns probabilities to $\boldsymbol{Y}$ based on a set of counts of regular local configurations which are sufficient statistics for their parameters. The exponential family random graph model can be defined as

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X}) = \frac{1}{c} \exp\{\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})\boldsymbol{\eta}^T\}, \tag{1}$$

where

$$c = \sum_{\boldsymbol{y}} \exp\{\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})\boldsymbol{\eta}^T\} \tag{2}$$

is a normalizing constant,

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)$$

is the p-dimensional vector of parameters and $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})$ a row vector of network statistics of dimension p. It is helpful to introduce the change of the vector of statistics in $g(\cdot)$,

$$\Delta(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X}))_{ij} = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})|_{y_{ij}=1} - \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})|_{y_{ij}=0} \tag{3}$$

for all dyads (i, j). We can express the conditional distribution of Y_{ij} given

$$\boldsymbol{Y}^c_{ij} = \{Y_{kl};\ (k, l) \neq (i, j)\}$$

as

$$\text{logit}\left(P(Y_{ij} = 1 \mid \boldsymbol{Y}^c_{ij} = \boldsymbol{y}^c_{ij})\right) = \boldsymbol{\eta}^T \Delta(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X}))_{ij}. \tag{4}$$

2.3 Dyadic independence models

For some special cases of ERGMs (1) has a simpler structure, which facilitates exact estimation of model parameters $\boldsymbol{\eta}$. One such class of models is the dyadic independence models, for which

$$\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X}) = \sum_{i<j} \sum y_{ij}\ \boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{X}_j) \tag{5}$$

for some p-dimensional function $\boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{X}_j)$ of pairs of attribute vectors. It is easy to see that

$$\Delta(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X}))_{ij} = \boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{X}_j)$$

for dyadic independence model, and moreover, the probability (1) of observed data can be written as product

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X}) = \frac{1}{c} \prod_{i<j} \prod \exp\{y_{ij} \Delta(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X}))_{ij} \boldsymbol{\eta}^T\}, \tag{6}$$

over all dyads (i, j), explaining the name of this class of models.

The identity (6) does not hold for general ERGMs. However, the right-hand side of this equation is often used as an approximation of the likelihood when parameters are estimated, the so called pseudo likelihood (see Section 3).

2.4 Example of ERGMs

2.4.1 Erdös-Renyi models

A major initial attempt of statistical modelling of social networks is the Bernoulli Random Graph Models proposed by Erdös and Renyi (1959). It can be viewed as a special case of a dyadic independence model (6) with $h(\boldsymbol{X}_i, \boldsymbol{X}_j) = 1$, so that

$$g(\boldsymbol{y}, \boldsymbol{X}) = R(\boldsymbol{y}) = \sum\sum_{i<j} y_{ij}$$

is the number of partnerships in the network, i.e. the total number of edges. The single parameter η can be interpreted as the common log-odds of the probability of partnership formation within any dyad. We find that

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{c} \exp\left(\eta \sum\sum_{i<j} y_{ij}\right),$$

where $\boldsymbol{c}$ can be given explicitly in ERG model.
Thus Y_{ij} are independent and identically distributed with success probabiliby $e^{\eta}/(1+e^{\eta})$.

2.4.2 Markov Random Graphs Models

The Markov neighbourhood assumption was introduced by Frank and Strauss (1986) in which all ties sharing a node are conditionally dependent on each other. This yields a Markov assumption, on which the Markov models are based. For an introduction to such models, see for instance Sundberg (2010).

The class of models considered by Frank and Strauss contains no covariates and has the form

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{c} \exp\left(\rho R(\boldsymbol{y}) + \sum_{l=2}^{p-1} \sigma_l S_l(\boldsymbol{y}) + \tau T(\boldsymbol{y})\right) \tag{7}$$

where

$$\boldsymbol{\eta} = (\rho, \sigma_2, \ldots, \sigma_{p-1}, \tau)$$

is the set parameters and

$$g(\boldsymbol{y}, \boldsymbol{X}) = (R(\boldsymbol{y}), S_2(\boldsymbol{y}), \ldots, S_{p-1}(\boldsymbol{y}), T(\boldsymbol{y}))$$

the network statistics. The new statistics compared to the Erdös-Rényi model are

$$S_l(\boldsymbol{y}) = \text{total number of } l\text{-stars} = \sum_i \sum_{j_1<\cdots<j_l} y_{ij_1} \cdot \ldots \cdot y_{ij_l},$$

$$T(\boldsymbol{y}) = \text{total number of triangles} = \sum_{i<j<l} y_{ij} y_{il} y_{jl}.$$

The normalizing constant of (7) is given by

$$c = c(\boldsymbol{\eta}) = \sum_{\boldsymbol{y}} \exp\left(\rho R(\boldsymbol{y}) + \sum_{l=2}^{p-1} \sigma_l S_l(\boldsymbol{y}) + \tau T(\boldsymbol{y})\right). \tag{8}$$

In particular, a Markov random graph model for a network with edges, two-star, three-star and triangle statistics is given by

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{c}\exp(\eta R(\boldsymbol{y}) + \sigma_2 S_2(\boldsymbol{y}) + \sigma_3 S_3(\boldsymbol{y}) + \tau T(\boldsymbol{y})).$$

2.4.3 Assortative Mixing Model

A special type of dyadic independence model is one that proposes a tendency for assortative mixing, that is, a greater or smaller probability of individuals to form edges with others having the same covariates. We model this using

$$\boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{X}_j) = (1, 1_{\{X_{i1}=X_{j1}\}}, \ldots, 1_{\{X_{iq}=X_{jq}\}}). \tag{9}$$

This model has $p = q + 1$ parameters with distribution

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{c}\exp\left(\eta_1 R(\boldsymbol{y}) + \sum_{l=1}^{q} \eta_{l+1} N_l(\boldsymbol{y})\right) \tag{10}$$

where $R(\boldsymbol{y})$ is the number of edges and $N_l(\boldsymbol{y})$ the number of dyads that have an edge and the same value of covariate l. Hence $\eta_{l+1} > 0$ indicates assortative mating with respect to covariate l.

2.4.4 Models with geometrically egdewise shared partner statistics

An k-triangle is a set of $k \in \{1, 2, \ldots, n-2\}$ distinct triangles that share a given edge. Let $T_k(\boldsymbol{y})$ denote the total number of k-triangles for network data $\boldsymbol{y}$. Since a 1-triangle is an ordinary triangle, we have that $T_1(\boldsymbol{y}) = T(\boldsymbol{y})$. For a fixed $\alpha \geq 0$, introduce the alternating k-triangle statistic

$$V(\boldsymbol{y}; \alpha) = 3T_1(\boldsymbol{y}) + \sum_{k=2}^{n-2} e^{-\alpha k}(-1)^{k-1} T_k(\boldsymbol{y})$$

of Sneijders et al. (1996). It is also called the Geometrically weighted edgewise shared partner (GWESP) statistic. If we add this statistic to the assortative mixing model (10), we get $p = q + 2$ parameters and

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{1}{c}\exp\left(\eta_1 R(\boldsymbol{y}) + \sum_{l=1}^{q} \eta_{l+1} N_l(\boldsymbol{y}) + \eta_p V(\boldsymbol{y}; \alpha)\right). \tag{11}$$

3 Inference

3.1 Parameter Estimation

3.1.1 Pseudolikelihood Maximization

Development of estimation methods for ERGMs has not kept pace with development of ERGMs themselves. To understand why, consider the sum of equation (2). A sample space consisting of all possible undirected graphs on n nodes contains $2^{n(n-1)/2}$ elements, an astronomically large number even for moderate n. Therefore, direct evaluation of the normalizing constant c in equation (2) is computationally infeasible for all but the smallest networks except in certain special cases such as the dyadic independence model of equation (5). For instance, most of the Markov models treated in Subsection 2.4.2 encounter computational difficulties. As a consequence, inference using maximum likelihood estimation is difficult.

In general for dyadic dependence models equation (6) does not hold, but the right-hand side is refered to as the pseudolikelihood. Now we can estimate $\boldsymbol{\eta}$ by pseudo likelihood estimation, as proposed by Strauss and Ikeda (1990). In the context of ERGMs, pseudo likelihood estimation is easy to carry through, even for complicatted models. A logit model is fitted for each edge indicator, given the rest of the graph,

$$\log \frac{P(Y_{ij} = 1 | \boldsymbol{Y}_{ij}^c = \boldsymbol{y}_{ji}^c)}{P(Y_{ij} = 0 | \boldsymbol{Y}_{ij}^c = \boldsymbol{y}_{ji}^c)} = \Delta(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X}))_{ij} \boldsymbol{\eta}^t.$$

For dyadic independence models (6) holds exactly, and then pseudolikelihood estimation coincides with maximum likelihood estimation. Indeed, we can then write (6) as

$$P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X}) = \prod_{ij} P_{ij}(Y_{ij} = y_{ij} | \boldsymbol{X}),$$

where P_{ij} is the marginal distribution of Y_{ij} given $\boldsymbol{X}$. In other words, we have that $\{Y_{ij}\}$ are conditionally independent given all covariates $\boldsymbol{X}$. This simplifies the likelihood, which is essentially a kind of logistic regression likelihood, a special case of the generalized linear model likelihood, as described in the appendix.

For dyadic dependence models $g(\boldsymbol{y}, \boldsymbol{X})$ typically has terms $y_{ij_1} y_{ij_2}$ and $y_{ij} y_{jk} y_{ik}$, and there is no linear expansion of g of the kind (5) and hence no independence between $\{Y_{ij}\}$ given $\boldsymbol{X}$.

3.1.2 Maximum Likelihood Estimation and Monte Carlo Markov Chain

From Equation(2) the log likelihood function is

$$l(\boldsymbol{\eta}) = \log P_{\boldsymbol{\eta}}(\boldsymbol{Y} = \boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X}) \boldsymbol{\eta}^t - \log(c(\boldsymbol{\eta})).$$

Maximimizing the likelihood with respect to $\boldsymbol{\eta}$ is equivalent to maximizing

$$l(\boldsymbol{\eta}) - l(\boldsymbol{\eta}_0) = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^t - \log(c(\boldsymbol{\eta})/c(\boldsymbol{\eta}_0)),$$

where $\boldsymbol{\eta}_0$ is an arbitrary fixed parameter vector. The difficult part is to estimate $c(\boldsymbol{\eta})/c(\boldsymbol{\eta}_0)$, and this can be accomplished by running a discrete-time Markov chain whose stationary distribtion is the distribution we wish to sample from. This is the Markov Chain Monte Carlo (MCMC) idea, see Geyer and Thompson (1992) and Snijders (2002). For fixed $\boldsymbol{\eta}_0$, we consider the identity

$$\begin{aligned} E_{\boldsymbol{\eta}_0}\left(\exp(\boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{X})(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^t)\right) &= \sum_y \exp(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})(\boldsymbol{\eta} - \boldsymbol{\eta}_0)) P_{\boldsymbol{\eta}_0}(\boldsymbol{Y} = \boldsymbol{y}) \\ &= \sum_y \exp(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^t) \frac{\exp(\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})\boldsymbol{\eta}_0^t)}{c(\boldsymbol{\eta}_0)} \\ &= \frac{c(\boldsymbol{\eta})}{c(\boldsymbol{\eta}_0)}. \end{aligned}$$

Thus, $c(\boldsymbol{\eta})/c(\boldsymbol{\eta}_0)$ is an expectation, where the symbol $E_{\boldsymbol{\eta}_0}$ denotes the expectation operator assuming $\boldsymbol{Y}$ is random from the ERGM with parameter $\boldsymbol{\eta}_0$.

The Law of large numbers suggests that we approximate an unknown population mean by a sample mean. Thus,

$$\begin{aligned} l(\boldsymbol{\eta}) - l(\boldsymbol{\eta}_0) &= \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^t - \log E_{\boldsymbol{\eta}_0}(\exp(\boldsymbol{g}(\boldsymbol{Y}, \boldsymbol{X})(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^t)) \\ &\approx \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{X})(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^t - \log\left(\frac{1}{I}\sum_{i=1}^{I} \exp(\boldsymbol{g}(\boldsymbol{y}_i, \boldsymbol{X})(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^t)\right), \end{aligned}$$

where $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$ is a random sample of networks from the distribution defined by the ERGM with parameter $\boldsymbol{\eta}_0$.

3.2 Model Selection

Model selection is the way of selecting a statistical model from a set of potential models, given data. Determining the principle that explains a series of observations is often linked directly to a mathematical model predicting those observations. We consider to model selection based on Akaike's information criterion (AIC) and the Bayesian information criterion (BIC).

AIC was developed by Hirotsugu Akaike under the name of "an information criterion" in 1971 and proposed in Akaike (1974) as a measure of the goodness of fit of an estimated statistical model. It is usually used for model selection. For a given model M it is defined as

$$\mathrm{AIC}(M) = 2p - 2l(\hat{\boldsymbol{\eta}}),$$

where $\hat{\boldsymbol{\eta}}$ is the maximum likelihood estimate and p the number of parameters of model M. The goal is to minimize AIC(M) as a function of M.

The BIC was defined by Schwarz (1978) and is also called Schwarz' Criterion. It is closely related to the Akaike information criterion, but has another penalty term for the number of parameters p of the model;

$$\mathrm{BIC}(M) = \log(N)p - 2l(\hat{\boldsymbol{\eta}}),$$

where N is the number of data points $\{Y_{ij}\}$, i.e. $N = n(n-1)/2$. Hence, for network models, the sample size N is not the same as the number of nodes n.

3.3 Goodness of fit

The deviance is a quality of fit statistic for a model that is usually used for statistical hypothesis testing. For a given model M the deviance is defined as

$$D(M) = -2\left(l(\hat{\boldsymbol{\eta}}) - l(\hat{\boldsymbol{\eta}}_{full})\right),$$

where $\hat{\boldsymbol{\eta}}$ is the ML-estimate for the given model M and $\hat{\boldsymbol{\eta}}_{full}$ the ML-estimate for a full model with one parameter per observation. The deviance is frequently used for GLMs, see McCullagh and Nelder (1989). The difference in deviance $D(M_1) - D(M_2)$ between two models M_1 and M_2 can be used for hypotheses testing.

As described in Hunter et al. (2008a), goodness of fit of an ERGM M can also be assesed by means of simulation. Various statistics from an observed network are compared with the corresponding distributions of the statistcs for simulated data. The símulated data sets are generated from the fitted parameters $\hat{\boldsymbol{\eta}}$. Good agreement between the observed statistics and the simulated distributions indicate a good fit.

4 An Example Data Set

4.1 Descriptive analysis

The data that we consider in this paper is taken from the National Longitudinal Study of Adolescent Health or Add Health, as described in Udry (2003) and Harris et al. (2003). This data set was analyzed by Goodreau et al. (2008) and our analysis closely parallels theirs. The data set is an undirected, one-mode friendship network of $n = 1461$ vertices $R(\boldsymbol{y}) = 974$ edges and $T(\boldsymbol{y}) = 169$ triangles.

Each node i represents a student and $Y_{ij} = 1$ indicates friendship between i and j. The $q = 3$ different attributes for each individual i are

$$\boldsymbol{X}_i = (\mathrm{Grade}_i, \mathrm{Race}_i, \mathrm{Sex}_i). \tag{12}$$

From Figure 1 we see that one large component appears and then smattering of many very small components, all of which are not visuable. The count of the component size

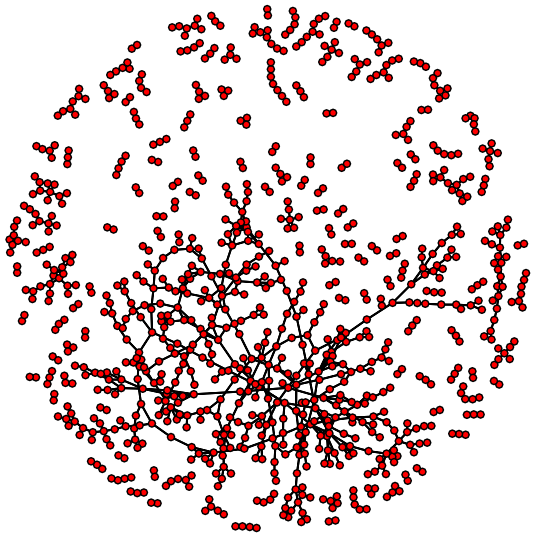

Figure 1: Faux Magnolia High, without isolates.

Table 1: Component size distribution.

Component size	1	2	3	4	5	6	7	8	9	11	12	19	23	439
Frequency	524	64	29	11	12	4	7	4	1	1	1	1	1	1

distribution in Table 1 shows that there are 524 isolates, one large component of 439 vertices and many components lie in between.
The distribution of vertex attributes is summarized in Table 2.

Table 2: Distribution of students (vertices) with respect to various attributes.

Vertex attribute	Class	Frequency
Race	Asian	48
	Black	261
	Hisp	68
	NatAm	24
	Other	7
	White	1053
Grade	7	185
	8	210
	9	317
	10	299
	11	257
	12	193
Sex	F	768
	M	693

It turns out that among the attributes, grade is the strongest determinat of social relations. We can visualize this by colouring vertices according to grade (Figure 2) or by considering the mixing matrix with respect to grade (Table 3), which is concentrated along the diagonal, with diagonal sum $N_{grade}(\boldsymbol{y}) = 820$. The corresponding mixing matrix for race (Table 4) is also concentrated along the diagonal, but somewhat less so than for grade, since $N_{race}(\boldsymbol{y}) = 787$.

The degree distribution of the data set is summarized in Table 4, for the whole population as well as for the female subpopulation:
Table 5 summarized the degree distribution for the whole and female subpopulation with edge prob

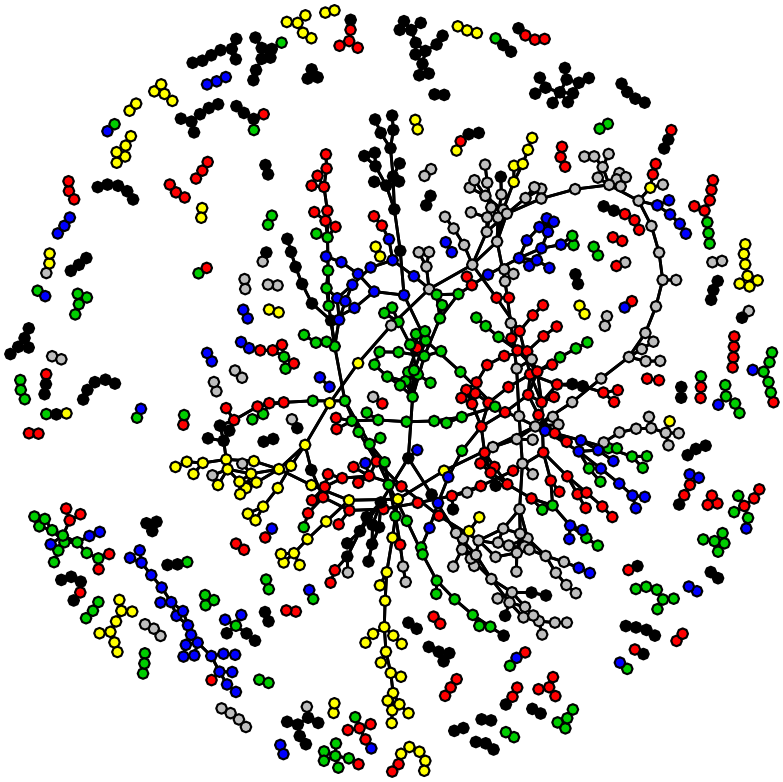

Figure 2: Faux Magnolia High, without isolates, coloured by grade.

Table 3: Mixing matrix with respect to grade.

Grade	7	8	9	10	11	12
7	110	11	3	3	0	0
8	11	165	9	7	0	2
9	3	9	152	24	10	4
10	3	7	24	151	38	11
11	0	0	10	38	152	32
12	0	2	4	11	32	90

Table 4: Mixing matrix with respect to race.

	Asian	Black	Hisp	NatAm	Other	White
Asian	7	4	0	1	0	35
Black	4	85	9	3	0	57
Hisp	0	9	1	0	0	48
NatAm	1	3	0	3	0	25
Other	0	0	0	0	0	5
White	35	57	48	25	5	691

4.2 Fitting ERGMs

We will fit data to four ERGMs, which are:

4.2.1 Model M_1 (Erdös-Rényi)

This model contains one single parameter η, the log odds of the edge probability. We may test M_1 versus the submodel $M_0 : \eta = 0$, for which all $2^{n(n-1)/2}$ edge configurations are equally likely. It can be seen from both Tables 6 and 7 that data strongly rejects M_0 in favour of M_1. The estimated edge probability is

$$\frac{e^{-6.998}}{1 + e^{-6.998}} \approx 0.000913.$$

4.2.2 Model M_2 (Assortative mixing)

This is the assortative mixing model based on attributes grade, race and sex, i.e. (9), with $\boldsymbol{X}_i$ as in (12).

Table 8 summarizes ML parameter estimates for M_2. For each parameter η_i we also give p-values for testing $\eta_i = 0$ versus $\eta_i \neq 0$. We find that all four terms, that is, edges,

Table 5: Degree distribution for the whole and female subpopulation.

Degree	0	1	2	3	4	5	6	7	8
Frequency (whole pop)	524	403	271	128	85	30	13	5	2
Frequency (female subpop)	226	226	160	80	44	18	7	0	0

Table 6: ML estimation for M_1 and hypothesis testing for M_1 versus M_0.

$\hat{\eta}$	Std. error	p-value
-6.99760	0.03205	$<$0.0001

grade, race and sex are significant and also that the likelihood increased dramatically relative to Model 1.

We notice from Table 8 that the log-odds of a tie which is completely heterogeneious is -10.01, the log-odds of a tie that is homogeneious by race is $-10.01 + 1.20 = -8.82$ and the log-odds of a tie when all the three attributes are homogeneous is $-10.01 + 3.23 + 1.20 + 0.88 = -4.70$.

We assessed goodness of fit for M_2 in Figure 3 by checking if the degree distribution for one simulated data set, generated from M_2 with estimated parameters as in Table 8, matches the observed network's degree distribution. The two distributions match in the upper tail but not towards the lower end of the distribution. Most notably in the relative proportion of vertices with degree 0 and 1. The corresponding race mixing matrix for the simulated data set is given in Table 9, with diagonal sum $N_{race}(\boldsymbol{y}_{sim}) = 753$. Comparing this with Table 4 we see again that the model fit for M_2 could be improved.

4.2.3 Model M_3 (Triangle Model)

This is a Markov model with egde and triangle counts, i.e. $p = 2$ in (7). Clearly this is a dyadic dependence model, hence the fitting algorithm draws on MCMC and is stochastic. After 29 Newton-Raphson iterations of the pseudo likelihood we get the initial estimates $\hat{\eta}_1 = -7.254$ for the edge parameter and $\hat{\eta}_2 = 4.558$ for the triangle parameter. These are used as starting values for the MCMC estimation procedure. After iterating the Markov chain 10 000 times, the final Monte Carlo ML estimates were $\hat{\eta}_1 = -7.310$ and $\hat{\eta}_2 = 4.584$ for edges and triangles respectively.

Recall that the data set has $R(\boldsymbol{y}) = 974$ edges and $T(\boldsymbol{y}) = 169$ triangels. In order to asses possible model degeneracy of M_3, the middle column of Table 10 shows the distribution of $E_{\hat{\boldsymbol{\eta}}}(R(\boldsymbol{Y})) - R(\boldsymbol{y})$ for the 10000 simulated values of $\hat{\boldsymbol{\eta}}$ during one iteration of the Markov chain. Analogolsly, the right hand column gives the distribution of $E_{\hat{\boldsymbol{\eta}}}(T(\boldsymbol{Y})) - T(\boldsymbol{y})$.

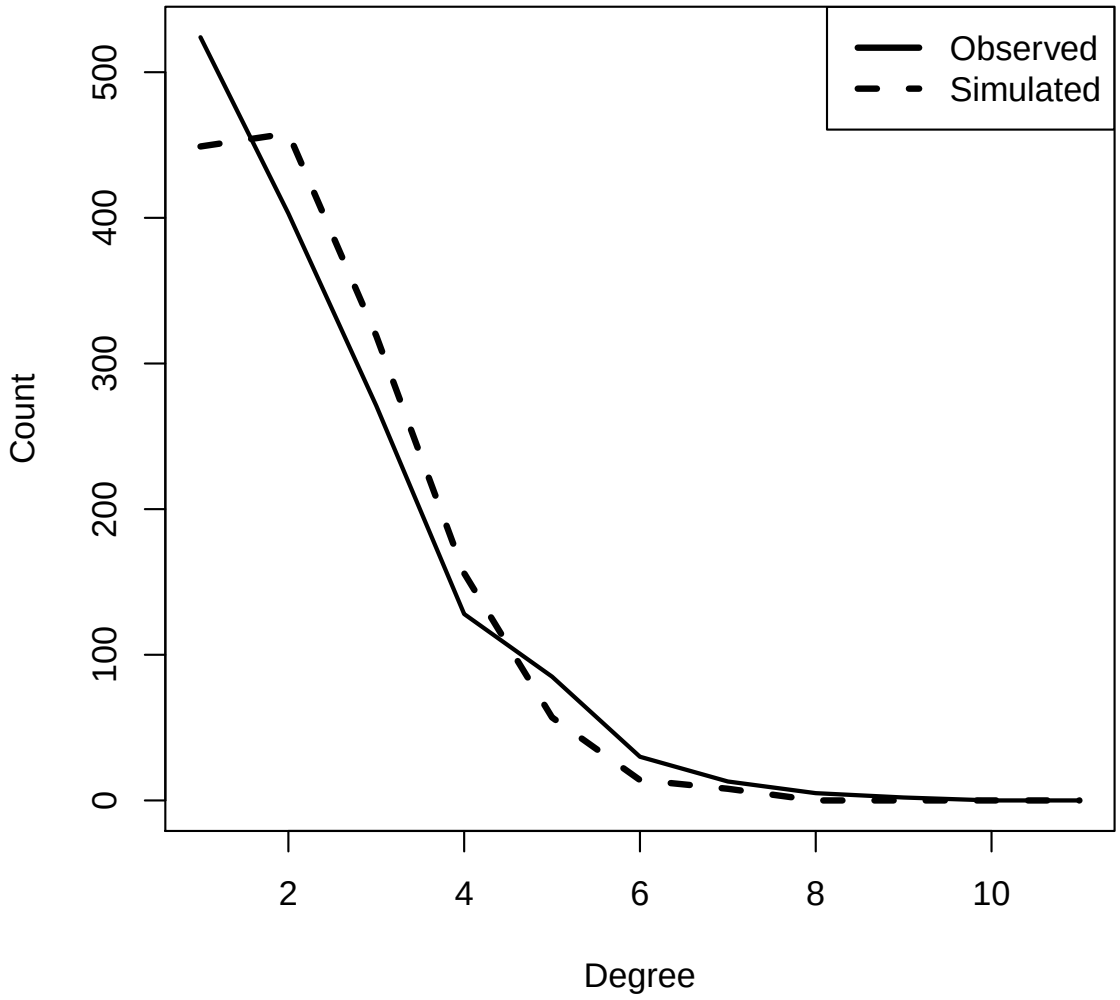

Figure 3: Degree distributions for observed data and simulated data from fitted model M_2.

Table 7: Deviances for M_0 and M_1.

Type of deviance	Value	Degrees of freedom $N-p$
Null $D(M_0)$	1478525	1066530
Residual $D(M_1)$	15580	1066529
Difference $D(M_0)-D(M_1)$	1462944	1

Table 8: Inference for M_2.

Parameter	Estimate	Std. error	p-value
η_1 (edges)	-10.01277	0.11526	<0.0001
η_2 (grade assortative mixing)	3.23105	0.08788	<0.0001
η_3 (race assortative mixing)	1.19646	0.08147	<0.0001
η_4 (sex assortative mixing)	0.88438	0.07057	<0.0001

From this simulation we see that the mean of the edge statistic is off by an average of 468 and the triangle statistic by an average of 784. The smallast simulated edge and triangle statistics are larger than the observed increased by 78 and 144 respectively. This clearly indicates degeneracy of M_3 in terms of fitting this data set.

4.2.4 Model M_4 (Model with assortative mixing and GWESP statistic)

This is an extention of M_2 where a geometrically weighted edgewise share partner (GWESP) statistic is added, to explore the clustering of the network. In other words, we use (11) with $q=3$ and $p=5$. This model has a GWESP parameter η_5 as well as the usual parameters $\eta_1,\ldots,\eta_4$ from Model M_2. There is also a fixed non-negative scaling parameter α included in the GWESP statistic. When $\alpha=0$ this statistic is equivalent to the number of edges that belongs to at least one triangle. To reduce degeneracy we chose small values of α, close to zero, and then increased α. For $\alpha=0$ and $\alpha=0.2$ we obtain the parameter estimates shown in Table 11.

When $\alpha=0$ we see that the log-odds is -9.83 for two arbitrary individuals to have a common friends. But if they share at least one common friend the log-odds increases to $-9.83+1.80=-8.03$.

To evaluate goodness-of-fit, we simulate a number of data set for the fitted M_4 with $\alpha=0.2$. In particular, we will investigate whether the fitted M_4 captures the observed triangle count distribution well. Figure 4 gives the result based on 100 MCMC iterations of 1e+05 steps each. It turns out that the observed triangle count lies outside the central 95% interval of the simulated distribution. However, it can be shown that M_4 is still an improvement over M_2 in terms of model fit for triangle counts.

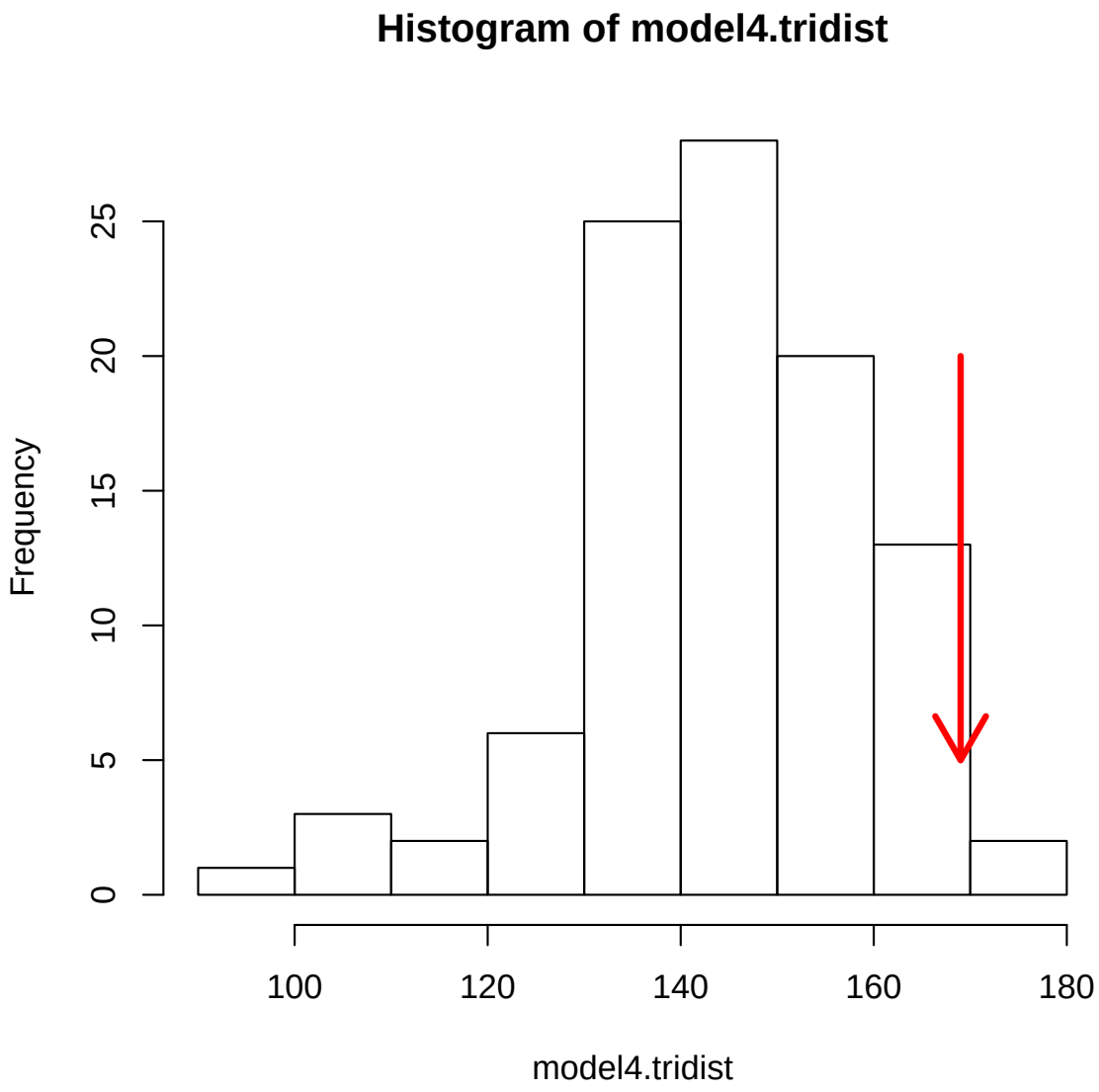

Figure 4: Triangle counts for observed data and 100 simulateded data sets from fitted M_4 with $\alpha = 0.2$.

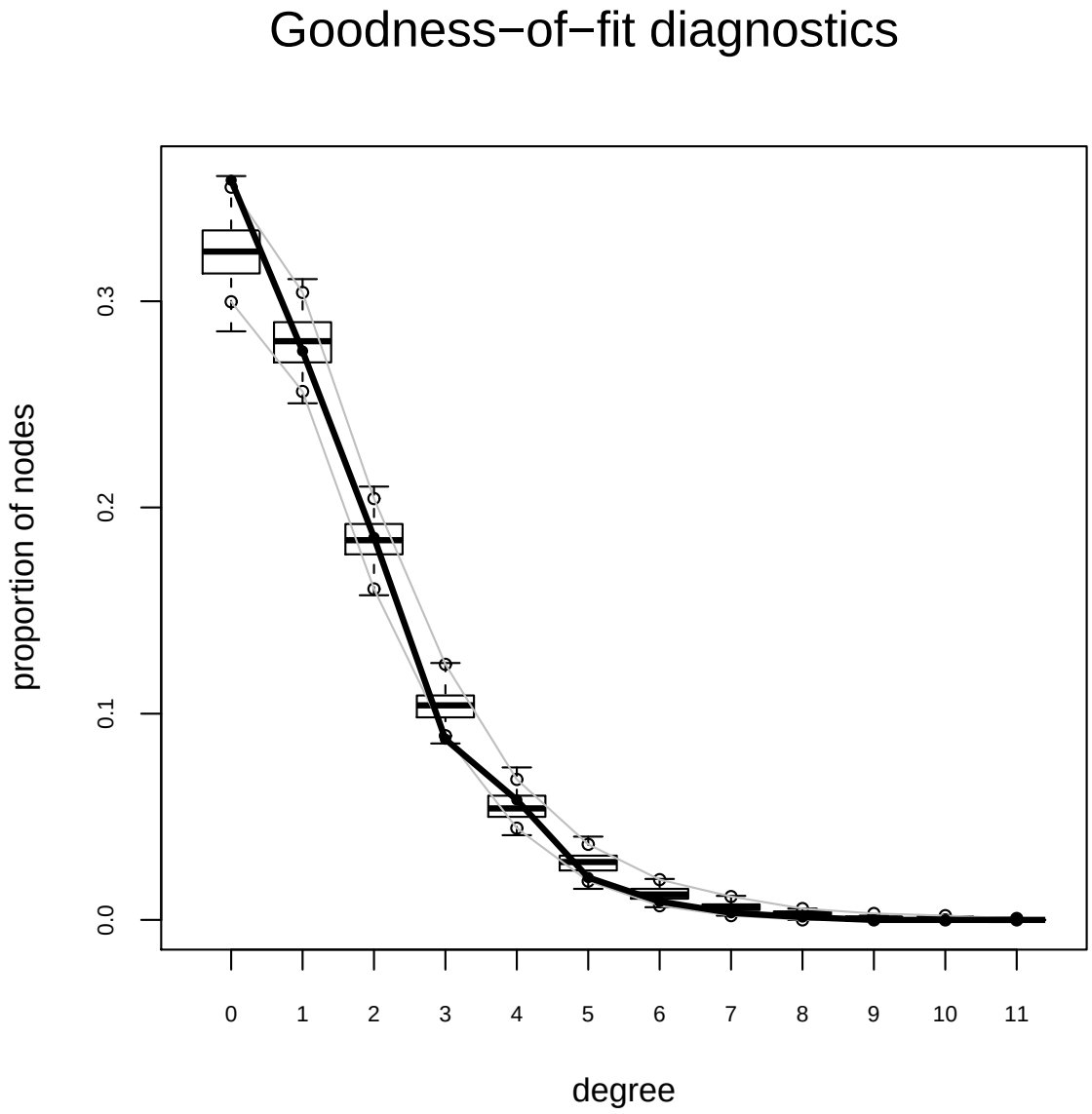

Figure 5: Observed degree distribution (solid line) and simulated degree distributions (box plots) for fitted M_4 with $\alpha = 0.2$ based on 100 MCMC simulations.

Table 9: Mixing matrix by race for data set simulated from fitted model M_2.

	Asian	Black	Hisp	NatAm	Other	White
Asian	0	6	1	3	0	16
Black	6	39	9	3	2	116
Hisp	1	9	4	1	0	35
NatAm	3	3	1	0	1	8
Other	0	2	0	0	0	4
White	16	116	35	8	4	708

Table 10: Summary of simulated average edge and triangle counts for M_3 based on a Markov chain with 10 000 steps, relative to the corresponding observed network statistics.

	Edges	Triangles
Min.	78	144
1st Quantile	235	314
Median	337	568
Mean	468	784
p-value dev.	0.00000	0.00000
Std. dev.	307.6	602.4

We next consider in Figure 5 the degree distributions for 100 MCMC iterations of 1e+05 steps each. For each degree a box plot is drawn, showing how the proportion of nodes with that degree varies over simulated data sets. The observed degree proportions are displayed as well. For each degree, the intervals between the soft lines contain 95 % of the simulated proportions. Thus, in principle, the p-values can be computed for each node degree based on such a plot. It is clear that the fitted M_4 captures the observed degree distribution for this network quite well.

Similarly, one may consider box plot curves based on e.g. edgewise shared partner distributions and geodesic distributions.

4.3 Model selection

The overall calculations of Model 1, Model 2 and Model 4 are summarized in Table 12. We see that M_4 indeed improves the likelihood considerably, and is also selected by either AIC or BIC. Estimates of all assortative mixing parameters decline in face of the new triangle GWESP statistic, while the edge parameter estimate increases.

Table 11: Inference for M_4.

Parameter	Estimate ($\alpha = 0$)	Estimate ($\alpha = 0.2$)
η_1 (edges)	-9.8297890	-9.7915376
η_2 (grade assortative mixing)	2.7904821	2.7557927
η_3 (race assortative mixing)	0.9511466	0.9184486
η_4 (sex assortative mixing)	0.7877837	0.7664999
η_5 (GWESP)	1.8031250	1.8150806

Table 12: Comparison between models M_1, M_2 and M_4.

Quantity to compare	M_1	M_2	M_4 ($\alpha = 0.2$)
p	1	4	5
$\hat{\eta}_1$ (edges)	-7.00	-10.01	-9.79
$\hat{\eta}_2$ (grade assortative mixing)		3.23	2.76
$\hat{\eta}_3$ (race assortative mixing)		1.20	0.92
$\hat{\eta}_4$ (sex assortative mixing)		0.88	0.77
$\hat{\eta}_5$ (GWESP, $\alpha = 0.2$)			1.82
$l(\hat{\boldsymbol{\eta}})$ (log likelihood)	-7790.1	-6528.71	-5502.28
AIC	15582.2	13065.4	11014.6
BIC	15593.9	13115.0	11074.1

5 Conclusions

The Exponential Family Random Graph Models are appropriate for modelling complex social networks. We reviewed the theory of EGRMs, including parameter estimation, model selection and goodness-of-fit. As an illustration, we analyzed a given social friendship data set using the R package statnet.

The class of ERGMs includes the simplest Bernoulli (Erdös-Rényi) model with just one sufficient statistic; the total number of edges. The Markov models goes a step further and includes sufficient statistics like triangle and k-star counts. However, inference typically has degeneracy problems which makes maximum likelihood estimation very difficult even for small sized networks. Including geometrically weighted edgewise shared partner statistics into the model makes ML estimation possible even for large networks, as evidenced by converging MCMC-approximations of ML estimates and improved goodness-of-fit.

Appendix. Generalized linear models

A.1 Link Functions

Generalized linear models (GLMs) were introduced by Nelder and Wedderburn (1972) as a theoretical framework that incorporates linear regression, logistic regression and Poisson regression models. It connects a linear exponential dispersion family through a linear predictor and link function. A standard reference is McCullagh and Nelder (1989).

The multiple linear regression model assumes that the conditional expectation of the $N \times 1$ vector $\boldsymbol{Y} = (Y_k)_{k=1}^N$ of dependent or response variables is a linear combination of predictor variables or covariates. This can be expressed as

$$E(\boldsymbol{Y}|\boldsymbol{Z}) = \boldsymbol{Z}\boldsymbol{\eta}^t,$$

where $\boldsymbol{Z} = (\boldsymbol{Z}_k)_{k=1}^N$ is the $N \times p$ design matrix of design vectors $\boldsymbol{Z}_k$ as rows and $\boldsymbol{\eta}$ the p-dimensional vector of regression parameters. Typically Y_k are independent random variables given $\boldsymbol{Z}$. We can rewrite (5) as

$$E(Y_k) = \boldsymbol{Z}_k\boldsymbol{\eta}^t \tag{A.1}$$

for study units $k = 1, \ldots, N$

The link function g_{link} of a GLM provides a relationship between the linear predictor and the expected valaue of the response variable. It thus generalizes (A.1) in that

$$g_{link}(E(Y_k)) = \boldsymbol{Z}_k\boldsymbol{\eta}^t$$

There are many commonly used link functions, and their choice can be somewhat arbitrary. For Bernoulli Y_k, for example, any smooth cdf can be used. Typical links are the logistic and standard normal (Gaussian) cdfs which lead to logit and probit models, respectively. A further alternative for Bernoulli Y_k is the complementary log log link.

We will apply GLMs to dyadic independence random graph models with n vertices, $N = n(n-1)/2$, $k = (i,j)$ a dyad, $Y_k = Y_{ij}$ a binary edge indicator for dyad (i,j) and $\boldsymbol{Z}_k = \boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{X}_j)$ the contribution of (i,j) to the vector $g(\boldsymbol{Y}, \boldsymbol{X})$ of network statistics. We will use a logistic link function

$$g_{\text{link}}(y) = \log \frac{y}{1-y},$$

which gives logistic regression.

A.2 Exponential dispersion models

The distribution of Y_k for an exponential dispersion family of models is either discrete or continuous. It has the form

$$f(y_k, \theta_k, \psi) = \exp\left\{\frac{y_k\theta_k - b(\theta_k)}{a(\psi)} + c(y_k, \psi)\right\} \tag{A.2}$$

where θ_k is the canonical parameter, allowed to depend on k, and ψ is a dispersion parameter. In a GLM with canonical parametrization, the θ_k depends linearly on the predictor variables;

$$\theta_k = \boldsymbol{Z}_k \boldsymbol{\eta}^t.$$

The likelihood function of data for a GLM is

$$P_{\boldsymbol{\eta},\psi}(\boldsymbol{Y} = \boldsymbol{y}) = \prod_{k=1}^{N} f(y_k, \boldsymbol{Z}_k \boldsymbol{\eta}^t, \psi).$$

A.3 Parameter estimation

Now we are going to estimate the parameters of a GLM using maximum liklihood estimation. The maximum likelihood estimate (MLE) maximizes the log likelihood

$$l(\boldsymbol{\eta}) = \sum_k \left(\frac{y_k \boldsymbol{Z}_k \boldsymbol{\eta}^t - b(\boldsymbol{Z}_k \boldsymbol{\eta}^t)}{a(\psi)} + c(y_k, \psi) \right)$$

with respect to $\boldsymbol{\eta}$, treating ψ as a nuisance parameter.

For most GLMs, computation of the MLE requires some iterative numerical approximation algorithm. For instance, each step of the iteration can be given by a weighted least squares fit. Since the weights are varying during the iteration the likelihood is optimized by an iteratively reweighted least squares algorithm.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** (6), 716Ű723.

Erdös, P. and Rényi, A. (1959). On random graphs. *Publications Mathematicae* **6**, 290.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81**, 831-841.

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. B* **54**, 657-699.

Goodreau, S.M., handcock, M.S., Hunter, D.R., Butts, C.T. and Morris, M. (2008). A statnet tutorial. *Journal of Statistical Software* **24**(9).

Harris, K.M., Florey, F., Tabor, J., Bearman, P.S., Jones, J. and Udry, J.R. (2003). The national longitudinal study of adolescent health: research design. Technical report, University of North Carolina.

Hunter, D.R. (2007). Curved exponential family models for social networks. *Social Networks* **29**(2), 216-230.

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

Hunter, D.R., Handcock, M.S., Butts C.T., Goodreau S.M. and Morris, M. (2008b). egrm: A packeage to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software* **24**(3): nihpa54860.

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models*, 2nd edition, Chapman and Hall, Boca Raton.

Morris, M., Handcock M.S. and Hunter, D.R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software* **24**(4), 1548Ũ7660.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. R. Statist. Soc. A* **135**, 370-384.

Robbins, G., Pattison, P. and Wasserman, S. (1999). Logit models and logistic regressions for social networks III. Valued relations. *Psychometrika* **64**(3), 371-394.

Robbins, G., Pattison, P., Kalish, Y. and Lusher, D. (2007). An introduction to exponential random graph ($p^\star$) models for social networks. *Social Networks* **29**(2), 173-191.

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**(2), 461Ũ464.

Snijders, T.A.B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure* **3**. Available at www.cmu.edu/joss/content/articles/volume3/Snijders.pdf.

Snijders, T. A. B., Pattison, P.E., Robins, G.L. and Handcock, M.S. (2006). New specifications for exponential random graph models. *Sociological Methodology* **36**, 99-163.

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

Sundberg, R. (2010). *Statistical modelling by exponential families*. Lecture notes, Division of Mathematical Statistics, Stockholm University.

Udry, J.R. (2003). The national longitudinal study of adolescent health (Add Health)„ Wavew I and II, 1994-1996; Wave III, 2001-2002. Techical report, Carolina Population Center, Univeristy of North Carolina at Chapel Hill.

Wasserman, S.S. and Pattison, P. (1996). Logit models and logistic regression for social networks: I. An introduction to Markov graphs and $p^\star$. *Psychometrika* **61**(3), 401-425.