Research Report

Department of Statistics

ML-ESTIMATION OF THE CLUSTERING PARAMETER IN A MARKOV GRAPH MODEL

by

Karin Dahmström   Per Dahmström

1993:4

STOCKHOLM UNIVERSITY   S-106 91 STOCKHOLM   SWEDEN

ML-ESTIMATION OF THE CLUSTERING PARAMETER IN A MARKOV GRAPH MODEL

by

Karin Dahmström   Per Dahmström

1993:4

Stockholm University
Department of Statistics

ISBN 91-7474-140-3
ISSN 0280-7564

February 1993

ML-ESTIMATION OF THE CLUSTERING PARAMETER IN A MARKOV GRAPH MODEL

Karin Dahmström Per Dahmström
Department of Statistics
Stockholm University

Abstract: We consider a special case of Markov graph models with a clustering parameter σ. ML-estimation of this parameter is performed by using simulation of Markov graphs. More specifically an expansion of the expected number of two-stars is done and a starting solution with the exact values of the first three cumulants when $\sigma = 0$ is used. Furthermore, the cumulants up to the 4:th order are estimated for successively better estimates of σ. A special computer program is written to perform the whole stepwise estimation procedure. The previously suggested method of pseudolikelihood estimation by the use of logistic regression is shown to be unsatisfactory in many situations. All methods considered are compared with the exact ML-estimates for complete enumerations.

Keywords: Markov graph models. Clustering parameter. Markov simulations. Logistic regression. Cumulants.

1 INTRODUCTION

During the last two decades the use of graph theory in social networks has been developed from being a descriptive method to an analytical statistical method. The theoretical progresses include for instance different model assumptions for random graphs, survey designs and inference for the graphs and other relations to traditional statistical theory. For a review, see Frank (1980, 1988, 1991) and Karonski (1982).

The random "element" in a graph concerns the realisation or not of an edge between two vertices. Then it is possible to define different graph models according to the choice of probability distribution for the edge indicators.

It is well-known that many of the models can be defined as log-linear models for which the parameters in principle can be estimated by means of methods for log-linear models in contingency tables. To this category the Holland-Leinhardt p_1-model, see Holland and Leinhardt (1981), and Markov graph models belong. Another classification is whether the dyads are independent or not. A model with independent dyads is the Holland-Leinhardt p_1-model. However, it is often more realistic to consider dependent dyad models such as Markov graph models.

In this paper, attention is given to a special case of a Markov graph model, the pure clustering model. The methods suggested in the literature for ML-estimation of the clustering parameter σ, a graphical method and the use of logistic regression respectively, are compared with a stepwise method with simulations of Markov graphs and estimation of cumulants. The starting value is based upon the first three exact cumulants when $\sigma=0$. A special designed computer program, in which all the steps are made automatically, is written.

2 NOTATION AND DEFINITIONS

Let us consider a random graph G with n vertices and r edges, say an (n,r)-graph. For any two vertices i and j, we define the edge indicator

$$Y_{ij} = \begin{cases} 1 \text{ if the edge between } i \text{ and } j \text{ exists} \\ 0 \text{ otherwise} \end{cases} \qquad i,j=1,2,\ldots,n$$

In general, the graph is directed, but here we will consider undirected, symmetric graphs. The n-by-n matrix with the variables Y_{ij} is called the adjacency matrix. In this matrix, the diagonal elements are so called structural zeros, since by definition, $Y_{ii}\equiv 0$ for $i=1,2,\ldots,n$. For an undirected graph, $Y_{ij}=Y_{ji}$, and the matrix is shortly given by its "upper triangle", $i,j=1,2,\ldots,n$.

The variables Y_{ij} are assumed to follow some probability distribution. In the simplest case, the edge indicators Y_{ij} are independent and identically distributed with $P(Y_{ij}=1) = p$, $i,j=1,2,\ldots,n$. This will give us a *Bernoulli graph*; see for instance Frank (1985). However, in many applications it is considered more realistic to assume some sort of dependence structure between the "actors", and this is attained in ***Markov graph models***. For the notation and definitions of a Markov graph model, see chapter 3.

3 MARKOV GRAPH MODELS

3.1 Introduction

Let us consider Markov graph models as they are defined in Frank and Strauss (1986) and Strauss (1986). A graph is said to be a Markov graph if non-incident edges are conditionally independent. For an undirected Markov graph, it can be shown that the stars and triangles are sufficient statistics. In the following, we will restrict the study to homogeneous Markov graphs, for which all isomorphic graphs have the same probability; we are assuming that the vertices are a priori indistinguishable.

In this paper, we are considering a special case of a Markov graph model, the pure clustering model with the parameter σ. In section 3.4, the ML-estimate of this parameter is theoretically derived, and the suggestions in the literature to numerically find this estimate are described. More specifically, we consider the disadvantages and the problems with the methods suggested up to now. The results will bring us to a new, stepwise estimation procedure described in chapter 4 and in the rest of the paper.

3.2 Notation

Generally, the probability of any homogeneous, undirected Markov graph G with n vertices can be defined as the log-linear model

$$P(G) = c^{-1}\exp(\tau t + \sum_{k=1}^{n-1} \sigma_k s_k)$$

with the parameters τ and σ_k, $k=1,2,\ldots,n-1$, and the normalizing constant c. Furthermore, the variable t is equal to the number of triangles in G and s_k is equal to the number of k-stars in G, $k=1,2,\ldots,n-1$. A simplified model is the so called *triad model*, for which

$$P(G) = c^{-1}\exp(\rho r + \sigma s + \tau t)$$

with the parameters $\rho = \sigma_1$, $\sigma = \sigma_2$ and τ. The sufficient graph statistics r, s and t correspond to the number of edges, i e $Y_{ij}=1$, the number of two-stars, i e combinations of edges such that $Y_{ij}Y_{ik}=1$, $j\neq k$, and the number of triangles, i e combinations of edges such that $Y_{ij}Y_{jk}Y_{ki}=1$, $i\neq j\neq k\neq i$, respectively. This model is also referred to as the $\rho\sigma\tau$-model.

The parameter σ can be considered as a measure of the clustering in the graph. If $\sigma > 0$, then the number of interacting edges is tending to increase, while the edges are repulsing if $\sigma < 0$, cf Strauss (1986). Furthermore, the parameter τ is a measure of the transitivity between the edges. A relation is said to be transitive if, whenever (i,j) and (j,k) are two related pairs of elements, then (i,k) is a related pair too. Thus, if $\tau > 0$, the tendency of complete connected elements is increasing and for $\tau < 0$, it is decreasing. Finally, the parameter ρ measures the overall density of the graph; it is also called a reciprocity parameter.

Let us here consider a special case of the Markov triad model, namely the pure clustering model

$$P(G) = c^{-1}\exp(\rho r + \sigma s)$$

Since this model generally is conditioned by the number of edges r, it can be written

$$P(G|r) = \exp(\sigma s)/c(\sigma)$$

The normalizing constant is explicitly

$$c(\sigma) = \Sigma \exp(\sigma s)$$

where the sum is over all the possible graphs given the numbers of vertices and edges.

It can be seen that the probability distribution of the number of two-stars may be very asymmetrical if σ is large in absolute values. When $\sigma=0$ however, all the Markov graphs have the same probability, i e the Markov graph is identical to a conditioned Bernoulli graph with a fixed number of edges r. In the next section we will consider some exact expressions for the moments and cumulants for the number of two-stars for this special case. The third cumulant is of importance for the starting value in the stepwise method described later.

3.3 Null cumulants for the number of two-stars

Consider the special case of the pure clustering model when $\sigma=0$. In that case, the expected value and the variance of s can be written explicitly, see Frank and Strauss (1986). Let us denote

$$v^{(k)} = v!/(v - k)!$$

and

$$m = \binom{n}{2}$$

Since n denotes the number of vertices, m is equal to the number of edges, actual and potential. Then we have

$$E_{\sigma=0}(s) = \frac{n^{(3)}\ r^{(2)}}{2m^{(2)}}$$

The first moment is identical to the first cumulant K_1.

An alternative algebraic expression for $\mathrm{Var}_{\sigma=0}(s)$ - compared with the article - is the following:

$$\mathrm{Var}_{\sigma=0}(s) = (2n - 5)\cdot\frac{n^{(3)}r^{(3)}}{m^{(3)}} + \frac{n^2 - 6}{4} \cdot \frac{n^{(4)}r^{(4)}}{m^{(4)}} - K_1(K_1 - 1)$$

The variance is equal to the second cumulant K_2.

As an example, the expected value of the number of two-stars in a graph with n=12 vertices and r=33 edges and for which $\sigma=0$, is equal to 162.46; cf Figure 5 in Frank and Strauss (1986). The variance of the number of two-stars is 29.46.

Here an exact expression for the third cumulant K_3 when $\sigma=0$ is derived. This is used later in the starting value of σ in the suggested stepwise method.

Let us consider the number of two-stars and let

$$A_i = \begin{cases} 1 \text{ if the i:th two-star exists} & \\ 0 \text{ otherwise} & i=1,2,\ldots,n(n-1)(n-2)/2 \end{cases}$$

The expected values of the sums over all two-stars, over all pairs of two-stars and over all triples of two-stars are denoted

$$E_1 = \Sigma\ E(A_i)$$
$$E_2 = \sum_{i<j}\Sigma\ E(A_iA_j)$$
$$E_3 = \sum_{i<j<k}\Sigma\Sigma\ E(A_iA_jA_k)$$

Furthermore, let the first three moments of the number of two-stars about 0 be denoted M_1, M_2 and M_3. Then

$$M_1 = E_1$$
$$M_2 = E_1 + 2E_2$$
$$M_3 = E_1 + 6E_2 + 6E_3$$

The three first null cumulants can be written

$$K_1 = M_1$$
$$K_2 = M_2 - M_1^{\ 2}$$
$$K_3 = M_3 - 3M_2 \cdot M_1 + 2M_1^{\ 3}$$

The expected value E_1 is of course the value $E_{\sigma=0}(s)$ above.

The explicit expressions for E_2 and E_3 are

$$E_2 = \frac{2n-5}{2} \cdot \frac{n^{(3)}\ r^{(3)}}{m^{(3)}} + \frac{n^2-6}{8} \cdot \frac{n^{(4)}\ r^{(4)}}{m^{(4)}}$$

$$E_3 = \frac{n-2}{6} \cdot \frac{n^{(3)} r^{(3)}}{m^{(3)}} + \frac{16n-37}{6} \cdot \frac{n^{(4)} r^{(4)}}{m^{(4)}} + 4 \cdot \frac{n^{(4)} r^{(5)}}{m^{(5)}} +$$

$$+ \frac{2n^2+3n-8}{4} \cdot \frac{n^{(5)} r^{(5)}}{m^{(5)}} + \frac{n^3+6n^2+14n+60}{48} \cdot \frac{n^{(6)} r^{(6)}}{m^{(6)}} +$$

$$+ \frac{47}{6} \cdot \frac{n^{(5)} r^{(6)}}{m^{(6)}} + \frac{1}{3} \cdot \frac{n^{(4)} r^{(6)}}{m^{(6)}}$$

The third null cumulant for the number of two-stars in a (12,33)-graph is equal to 97.97.

3.4 ML-estimation of the clustering parameter

Let us assume that our observed graph has s_o two-stars and that we want to determine the ML-estimate of the clustering parameter σ. This is obtained by maximizing the likelihood function

$$L(\sigma|s_o) = N_{s_o}\cdot\exp(\sigma s_o)/\sum_i \exp(\sigma s_i)$$

where N_{s_o} is equal to the number of graphs with s_o two-stars, s_i is equal to the number of two-stars in the i:th graph, and the sum is over all the possible graphs. Differentiating with respect to σ, gives

$$\exp(\sigma s_o)[\Sigma(s_o - s_i)\exp(\sigma s_i)]/[\Sigma \exp(\sigma s_i)]^2 = 0$$

The equation can be written

$$s_o\Sigma \exp(\sigma s_i)/\Sigma \exp(\sigma s_i) = \Sigma s_i\exp(\sigma s_i)/\Sigma \exp(\sigma s_i)$$

$$s_o = \Sigma s_i\exp(\sigma s_i)/\Sigma \exp(\sigma s_i)$$

i e

$$s_o = E_\sigma(s)$$

Thus, the estimate $\hat{\sigma}_{ML}$ is that σ-value for which the *expected* value of the number of two-stars is equal to the *observed* number.

This equality can also be written

$$E_\sigma(s) = (d/d\sigma)[\log c(\sigma)] = s_o$$

which will be used later on.

It may first be noted that it is impossible in practice to obtain an exact numerical solution of this equation for graphs consisting of more than 6-7 vertices. It is only for small graphs that all the possible graphs and the distribution of the number of two-stars can be calculated exactly. For larger graphs however, two approximative methods are suggested in the literature. In this paper, a complete enumeration of the graphs will be given for some of these smaller graphs in order to make a comparison between our estimation results and the true ML-estimates, cf section 5.2.

The first approximative method is a simulation of Markov graphs for different values of σ such that a curve of $E_\sigma(s)$ against σ is obtained, where $E_\sigma(s)$ is estimated by the mean value $\bar{s}$. From this curve, the ML-estimate $\hat{\sigma}$ can be found graphically given the observed value s_o. This method is suggested by Frank and Strauss (1986) and by Strauss (1986). However, they say that "the construction of the $(\sigma, E_\sigma(s))$ curve seems too large a task for routine analysis...".

The second method is the use of pseudolikelihood estimates and logistic regression. In Strauss and Ikeda (1989), it is shown that a log-linear model

$$P(G) = c^{-1}(\Theta) \exp \Sigma \, \theta_m x_m(G)$$

where Θ is a vector of v parameters, $x_m(G)$, m=1,2,...,v, are different observable graph statistics, and $c(\Theta)$ is a normalizing constant, is compatible to a conditional logistic model. Let $G = \{Y_{ij}: i \neq j, 1 \le i,j \le n\}$ be a realization of an (n,r)-graph. Furthermore, let G_{ij}^{+} be the realization of the graph when $Y_{ij}=1$, and G_{ij}^{-} the realization when $Y_{ij}=0$ respectively. We will also consider the complement C_{ij} of the graph ("the rest" of the graph) to an observed value of Y_{ij}. This gives

$$P(Y_{ij}=1|C_{ij}) = P_{ij} = \frac{P(G_{ij}^{+})}{P(G_{ij}^{+}) + P(G_{ij}^{-})}$$

The log odds of the probability of an edge between the vertices i and j, conditional of the rest of the graph, can be written

$$\text{logit } P(Y_{ij}=1|C_{ij}) = \text{logit } P_{ij} = \Sigma\ \theta_m(\Delta_{ij})_m$$

where $(\Delta_{ij})_m$ is equal to the difference in the m:th graph statistic when $Y_{ij}=1$ and $Y_{ij}=0$ respectively, $m=1,2,\ldots,v$. For the general triad model, we get

$$\text{logit } P_{ij} = \rho\Delta r_{ij} + \sigma\Delta s_{ij} + \tau\Delta t_{ij}$$

where Δr_{ij}, Δs_{ij} and Δt_{ij} are the differences in the number of edges, two-stars and triangles respectively when $Y_{ij}=1$ instead of $Y_{ij}=0$. For the pure clustering model we get

$$\text{logit } P_{ij} = \rho + \sigma\Delta s_{ij}$$

since $\Delta r_{ij}\equiv 1$. The difference Δs_{ij} can explicitly be written, see for instance Frank (1991),

$$\sum_j Y_{ij} + \sum_i Y_{ij} - 2Y_{ij}$$

Furthermore, we define the pseudolikelihood (PL) function for the parameter vector Θ as a *conditional* likelihood function, where each probability is conditional on the rest of the data. We have

$$PL(\Theta) = \prod_{i\neq j} P(Y_{ij}|C_{ij}) = \prod_{i\neq j} P_{ij}$$

A maximum pseudolikelihood estimate is defined to be a value of Θ which maximizes the PL-function. Compared to the "ordinary" likelihood function, the observations are not independent and thus, the PL-estimates will generally differ from the true ML-estimates. It has been shown that the determination of the PL-estimates in the conditional logit model can be performed in the same way as the maximization of an ordinary likelihood function for logistic regression with *independent* observations Y_{ij}. This means that available statistical packages containing this procedure such as BMDP and SAS can be used.

The numerical comparisons in Frank and Strauss (1986) and in Strauss and Ikeda (1990) of the pseudolikelihood estimates with the ML-estimates are, just as the authors say, limited. More specifically, only four true σ-values are tested with five replicates for each value. These values were all situated in a narrow interval around 0; the values are -0.10, -0.05, 0, and 0.075. The value $\sigma=0$ means that the Markov graph is identical to a conditioned Bernoulli graph. The opinion of the authors concerning the results was that "The two methods appear to give estimators that are about equally good.". However, there is support for a more pessimistic view about the possibility of obtaining accurate estimates by means of the logistic regression.

3.5 Simulation of Bernoulli graphs

A third method of obtaining an unbiased estimate of $E_{\sigma}(s)$ is by simulating the simpler Bernoulli graphs instead of Markov graphs and then using a ratio estimator. This is according to a general method from importance sampling suggested in Snijders (1989), but described in a graph model context below. It has been performed in our work of finding a suitable estimation method.

Suppose that we want to estimate the expected value of a graph statistic $x_m(G)$ where the graph G follows a model M. The desired parameter might for instance be the expected number of two-stars in an (n,r)-graph which follows a Markov graph model. Generally, we want to estimate the parameter

$$\Theta = E_M[x_m(G)]$$

There is also an alternative graph model, say model B, such that for an observed graph G^*, the relation

$$P_M(G=G^*) > 0 \text{ gives } P_B(G=G^*) > 0$$

is valid. Thus, there is an absolute continuity of P_M with respect to P_B.

For a graph G, let us define the probability ratio

$$R_0(G) = P_M(G)/P_B(G)$$

Then

$$\Theta = E_M[x_m(G)] = E_B[x_m(G) \cdot R_0(G)]$$

That means that Θ can be estimated by the sample mean of $x_m(G) \cdot R_0(G)$ according to probabilities from model B. Given our estimation problem, the true model is the Markov clustering model, but simpler Bernoulli graphs could be simulated, all having the same probability.

Generally, a Bernoulli (n,r)-graph can be created by randomly choosing r edges among the $m = \binom{n}{2} = n(n-1)/2$ possible places in the adjacency matrix. This method is the same as choosing r elements from n(n-1)/2 elements without replacement. A sequential method for this is suggested by Cassel (1970).

Suppose that k Bernoulli graphs are simulated. For each graph, the number of two-stars s_i is computed. Let $P_M(G_i)$ and $P_B(G_i)$ be the probabilities for a graph according to a Markov and a Bernoulli model respectively, i=1,2,...,k. Furthermore, let g be equal to the number of possible graphs. In this context, $x_m(G) = s$, and we have

$$E\{(g/k)\sum_{i}^{k} s_i P_M(G_i)/P_B(G_i)\} = E(s)/P_B(G)$$

and

$$E\{(g/k)\sum_{i}^{k} P_M(G_i)/P_B(G_i)\} = 1/P_B(G)$$

where $P_B(G)$ is the constant probability for a Bernoulli graph. Therefore, a ratio estimator for $E_\sigma(s)$ is given by

$$\frac{(g/k)\cdot\Sigma\ s_i\cdot P_M(G_i)/P_B(G_i)}{(g/k)\cdot\Sigma\ P_M(G_i)/P_B(G_i)} = \frac{\Sigma\ s_i P_M(G_i)}{\Sigma\ P_M(G_i)} =$$

$$= \frac{\Sigma\ s_i\cdot\exp(\sigma s_i)/c(\sigma)}{\Sigma\ \exp(\sigma s_i)/c(\sigma)} = \frac{\Sigma\ s_i\cdot\exp(\sigma s_i)}{\Sigma\ \exp(\sigma s_i)}$$

where all the sums are over the k simulated Bernoulli graphs. A ratio estimator is generally biased, but from sampling theory it is known that its bias can be negligible for large values of k.

However, this method is not to be recommended when $|\sigma|$ is large. In that case, the estimated value of $E_\sigma(s)$ will be seriously biased even if the number of simulations is large. That is due to the fact that only graphs with a low probability according to a Markov model generally will be observed when Bernoulli graphs are simulated. As an example, we have noticed that in a simulation of 10^6 Bernoulli graphs, the observed maximum value of s for a (12,33)-graph was only 196, while the maximum value theoretically is 213.

4 A STEPWISE ESTIMATION METHOD

4.1 Introduction

Since the experiences of the estimation methods for the clustering parameter σ are rather unsatisfactory up to now, a method emanating from the earlier methods but combined with some new ideas is suggested. In these experiences we also include our own simulations of Bernoulli graphs which were described in section 3.5.

The method is built upon simulations of Markov graphs with a starting value of σ which is more accurate than earlier suggested estimates. It is followed by an expansion of $E_\sigma(s)$ around that σ-value using the cumulants up to the 4:th order which are estimated from the simulations. Then a new, better approximation of $\hat{\sigma}_{ML}$ can be obtained for which new Markov simulations are performed. The procedure is successively repeated with an increasing number of simulations for each approximative value until the difference between two σ-approximations is small enough. No curve (σ, $E_\sigma(s)$) is obtained with this method and no graphical finding is necessary either. The result is *one* numerical value, the ML-estimate of σ.

4.2 The stepwise method - an overview

The suggested method is shortly described below. However, some of the procedures are given a more detailed description in the following sections.

The input to the computer program is the number of vertices, the number of edges and the number of observed two-stars s_o in the graph respectively.

1. The minimum and maximum possible values of s are computed. If s_o equals s_{min} or s_{max}, then the model is degenerate in the sense that there is no clustering and the clustering is complete respectively; the estimation procedure is finished. Whatever the value of σ, $E_\sigma(s)$ cannot take the value s_{min} or s_{max}.

2. The number of observed two-stars is compared with the expected value of s when $\sigma=0$, cf section 3.2. If $E_{\sigma=0}(s) = s_o$, then $\hat{\sigma}_{ML}=0$. If $s_o < E_{\sigma=0}(s)$, then $\hat{\sigma}_{ML} < 0$, otherwise $\hat{\sigma}_{ML} > 0$. This apriori-knowledge will decrease the calculations to about half the time if the goal is to find a graphical solution.

3. An approximative ML-estimate can be obtained by solving the equation

$$s_o = K_1 + \sum_{j=2}^{h} K_j \cdot \frac{\sigma^{j-1}}{(j-1)!} \qquad h=2,3,\ldots$$

where K_j is the j:th cumulant for the number of two-stars when $\sigma=0$ and h is the number of added terms in the infinite sum. Frank and Strauss (1986) consider the case h=2, which gives the estimator

$$\sigma^* = (s_o - K_1)/K_2$$

However, with the help of an exact expression for K_3 derived in section 3.3, the equation

$$s_o = K_1 + K_2\sigma + K_3 \cdot \frac{\sigma^2}{2!} \qquad (4.1)$$

can be solved. When a solution exists, it is used as the starting value $\tilde{\sigma}$ with still better accuracy for the successive approximations in the suggested estimation method. The formulas above are described in more detail in section 4.3.

4. A starting graph for the Markov simulations is given by a simulated Bernoulli graph. To stabilize the Markov simulations, a "round" with 1000 simulations is initially performed. Then the simulations continue with 100000 graphs given the starting value $\tilde{\sigma}$. From these graphs, estimates of the cumulants up to the 4:th order can be computed. According to the expansion of the expected value of s given an arbitrary value of σ, an increment $\Delta\sigma$ to the old value of σ, say σ_0, is obtained from the solution of the equation

$$s_o - K_1(\sigma_0) - K_2(\sigma_0)\cdot\Delta\sigma - K_3(\sigma_0)\cdot\frac{(\Delta\sigma)^2}{2} - K_4(\sigma_0)\cdot\frac{(\Delta\sigma)^3}{6} = 0$$

A more detailed description of the use of estimated cumulants and Markov simulations is given in section 4.3 and 4.4 respectively.

5. A new simulation "round" is performed unless the absolute difference between the old and the new σ-value is less than 0.01.

An example of the output of the computer program is given on the next page. The name of the program is ESTSIGMA. The input is the number of vertices and edges in a (12,33)-graph with 190 two-stars, cf the example in Frank and Strauss (1986) p 837.

```
                 **  ESTSIGMA  **

   MAXIMUM-LIKELIHOOD ESTIMATION
   OF THE CLUSTERING PARAMETER σ
   IN THE PROBABILITY FOR A MARKOV-GRAPH
   BASED ON THE NUMBER OF TWO-STARS IN AN OBSERVED GRAPH

   FOR THE OBSERVED GRAPH:
   NUMBER OF VERTICES   12
   NUMBER OF EDGES      33
   NUMBER OF TWO-STARS 190

   FOR A GRAPH WITH  12 VERTICES AND  33 EDGES
   MINIMUM NUMBER OF TWO-STARS   150
   MAXIMUM NUMBER OF TWO-STARS   213
   EXPECTED NUMBER OF TWO-STARS WHEN σ = 0   162.4615

   APPROXIMATION NO  1   .507

   APPROXIMATION NO  2   .530

   APPROXIMATION NO  3   .532

   APPROXIMATION NO  4   .533
```

Approximation no 1 is the starting value obtained from the solution of eq (4.1).

4.3 The starting value and the estimated cumulants

In the derivation of the ML-equation in section 3.4, it was noticed that

$$E_\sigma(s) = \frac{d}{d\sigma}[\log c(\sigma)] \qquad (4.2)$$

where the normalizing constant c(σ) was defined as

$$c(\sigma) = \sum^{g} e^{\sigma s}$$

with the sum taken over all possible graphs.

According to standard theory, see e g Kendall-Stuart (1963, volume 1), the moment generating function can be written

$$M_S(t) = E(e^{ts}) = \sum^{g} e^{ts} \cdot P_{\sigma}(G) = \sum^{g} e^{ts} \cdot \frac{e^{\sigma s}}{\sum e^{\sigma s}} =$$

$$= c(\sigma+t)/c(\sigma)$$

Thus, the cumulant generating function is

$$C_s(t) = \log M_s(t) = \log c(\sigma+t) - \log c(\sigma)$$

Furthermore, the cumulants are defined as the coefficients K_i, $i=1,2,\ldots,\infty$, in the identity

$$C_s(t) = \sum_{i}^{\infty} K_i(\sigma) \cdot \frac{t^i}{i!}$$

Then we have

$$\log[c(\sigma+t)] = \log c(\sigma) + \sum K_i(\sigma) \cdot \frac{t^i}{i!} \qquad (4.3)$$

Let t be a difference between two σ-values, say $\Delta\sigma$. Then (4.3) is

$$\log[c(\sigma+\Delta\sigma)] = \log c(\sigma) + \sum K_i(\sigma) \cdot \frac{(\Delta\sigma)^i}{i!} \qquad (4.4)$$

After differentiation with respect to $\Delta\sigma$, the left side of (4.4) becomes

$$\frac{d}{d\Delta\sigma}\log[c(\sigma+\Delta\sigma)] = \frac{c'(\sigma+\Delta\sigma)}{c(\sigma+\Delta\sigma)} = \frac{\sum s e^{(\sigma+\Delta\sigma)s}}{\sum e^{(\sigma+\Delta\sigma)s}} = E_{\sigma+\Delta\sigma}(s)$$

and the right side is

$$\frac{d}{d\Delta\sigma}[\Sigma\ K_i(\sigma)\cdot\frac{(\Delta\sigma)^i}{i!}] = \Sigma\ K_i(\sigma)\cdot\ \frac{(\Delta\sigma)^{i-1}}{(i-1)!}$$

and we have

$$E_{\sigma+\Delta\sigma}(s) = \sum_{i=1}^{\infty} K_i(\sigma)\cdot\frac{(\Delta\sigma)^{i-1}}{(i-1)!} \qquad (4.5)$$

Thus, the expected value of s given an arbitrary value of σ, can be written as an expansion of the cumulants.

If we consider the special case $\sigma=0$, then $\Delta\sigma$ could be interpreted as the difference between an arbitrary σ-value and 0, i e $\Delta\sigma$ could be written as σ itself. Then we have

$$E_\sigma(s) = K_1(0) + K_2(0)\cdot\frac{\sigma^1}{1!} + K_3(0)\cdot\frac{\sigma^2}{2!} + K_4(0)\cdot\frac{\sigma^3}{3!} + \ldots.$$

where the cumulants are calculated for $\sigma=0$. If $E_\sigma(s)$ is replaced by the observed number of two-stars s_o, then the equation has the ML-estimate of σ as its solution. If the expansion is terminated after the second term, we get an approximative estimate of σ by using

$$\sigma^* = \frac{s_o - K_1}{K_2}$$

where s_o is the observed number of two-stars. This estimate is suggested by Frank and Strauss (1986).

The estimate σ^* was first used as a starting value in the stepwise method suggested in this paper. However, a still better estimate could be obtained by using the third cumulant K_3 in the expansion above. In section 3.3 an exact expression for K_3 when $\sigma=0$ is derived, and the value $\tilde{\sigma}$ obtained as the solution - when it exists - of the equation

$$s_o = K_1(0) + K_2(0)\cdot\sigma + K_3(0)\cdot\frac{\sigma^2}{2!}$$

is used as a starting value.

The relation (4.5) could be used in a stepwise method to get a successively better estimate of σ in the simulation of Markov graphs. From the starting value of σ, a new value of σ could be created by solving $\Delta\sigma$ with Newton-Raphson's method. The expected value of s is replaced by the observed number of two-stars s_o. If we terminate the expansion after the fourth cumulant and write an "old" σ-value generally by σ_0, the equation can be written

$$s_o - K_1(\sigma_0) - K_2(\sigma_0)\cdot\Delta\sigma - K_3(\sigma_0)\cdot\frac{(\Delta\sigma)^2}{2} - K_4(\sigma_0)\cdot\frac{(\Delta\sigma)^3}{6} = 0$$

Then a new value σ_1 is given by $\sigma_0 + \Delta\sigma$.

The cumulants are estimated for the actual σ-value from the distribution of two-stars in the simulations of Markov graphs. In each "round", the central moments μ are first estimated. The cumulant K_r of order r is related to the central moments of the same and lower orders in the following way:

$$\begin{aligned} K_1 &= \mu \\ K_2 &= \mu_2 = \sigma^2 \\ K_3 &= \mu_3 \\ K_4 &= \mu_4 - 3\mu_2^2 \end{aligned}$$

It may be argued that cumulants up to an higher order, say the 10:th order, should be used instead of stopping "already" at the 4:th cumulant. In fact, the estimation method has, in an earlier stage, included these cumulants of higher order, but numerical studies showed that they had a very large variation. Furthermore, no better accuracy of the ML-estimates was obtained with these higher cumulants included.

4.4 Markov simulations

The Markov simulations are performed by using a method described in Strauss (1986), which is based on the so called Metropolis method, cf Hammersley and Handscomb (1964).

Let us start with a Bernoulli graph of n vertices and r edges and the starting value $\tilde{\sigma}$ of σ, (4.1). The following steps are performed to simulate a sequence of Markov graphs G_k, k=1,2,....

1. Consider the upper triangle of the adjacency matrix in the starting graph G_1. Choose randomly one existing edge no I and one potential edge no J. The graph G* = G-I+J is constructed such that edge nr I is excluded and a new edge, no J, is included; the total number of edges is unchanged.

2. Compute the difference Δs of two-stars in the two graphs, i e $\Delta s = \#s(G_1) - \#s(G^*)$. If $\tilde{\sigma}$ denotes the latest estimate of σ, then if $\tilde{\sigma}\Delta s<0$, set $G_2= G^*$. If $\tilde{\sigma}\Delta s > 0$, set $G_2 = G^*$ with probability $\exp(-\tilde{\sigma}\Delta s)$, otherwise $G_2 = G_1$, i e the new graph is equal to the old one.

3. Simulate a new Markov graph by starting again from step no 1.

In order to stabilize the distribution of the number of two-stars, it is recommended that, say, 1000 graphs are simulated in an initial sequence. These first graphs are not used in the later computations. Furthermore, it is very important that the simulations are not too time-consuming, for the method to be used in routine analysis.

5 RESULTS

5.1 Introduction

The results obtained have been very promising. The most positive result is that the ML-estimate of σ has been obtained, not by any graphical method, but as one numerical result. Furthermore, it has been possible to compare the results with the true ML-estimates for some graphs by means of complete enumeration. It can also be concluded that the pseudo-likelihood estimates using logistic regression differ considerably from the true values. For larger graphs, comparisons have been made by doing an extremely large number of Markov simulations and by comparing the results found graphically from the curve $(\sigma, E_\sigma(s))$ in the literature.

5.2 Comparisons with complete enumeration

It is considered not practical to obtain a complete enumeration of the graphs when the number of vertices is larger than 7. Here a graph with n=7 vertices and r=12 edges is studied. The total number of graphs of that size is given by $\binom{\binom{n}{2}}{r} = \binom{\binom{7}{2}}{12} = \binom{21}{12} = 293930$.

The number of two-stars s varies between 30 and 39, see Table 5.1. The mean value of s when σ=0 is 33.

No of two-stars	No of graphs
30	19 355
31	45 360
32	71 190
33	48 055
34	51 030
35	27 090
36	18 585
37	8 190
38	4 200
39	875
Total	293 930

Table 5.1:
The exact distribution of two-stars for a (7,12)-graph.

By computing $E_\sigma(s)$ for σ=-2.50(0.01)2.50, the true ML-estimates can be obtained. These are shown in Table 5.2.

No of two-stars	ML-estimate of σ
30	Not defined
31	-0.978
32	-0.350
33	0
34	0.257
35	0.486
36	0.727
37	1.038
38	1.597
39	Not defined

Table 5.2:
True ML-estimates of σ for a (7,12)-graph given different number of two-stars.

The relation between σ and $E_\sigma(s)$ is shown in Diagram 5.1, which is the same curve from which the ML-estimate of σ can be found by the graphical method suggested by for instance Frank and Strauss (1986).

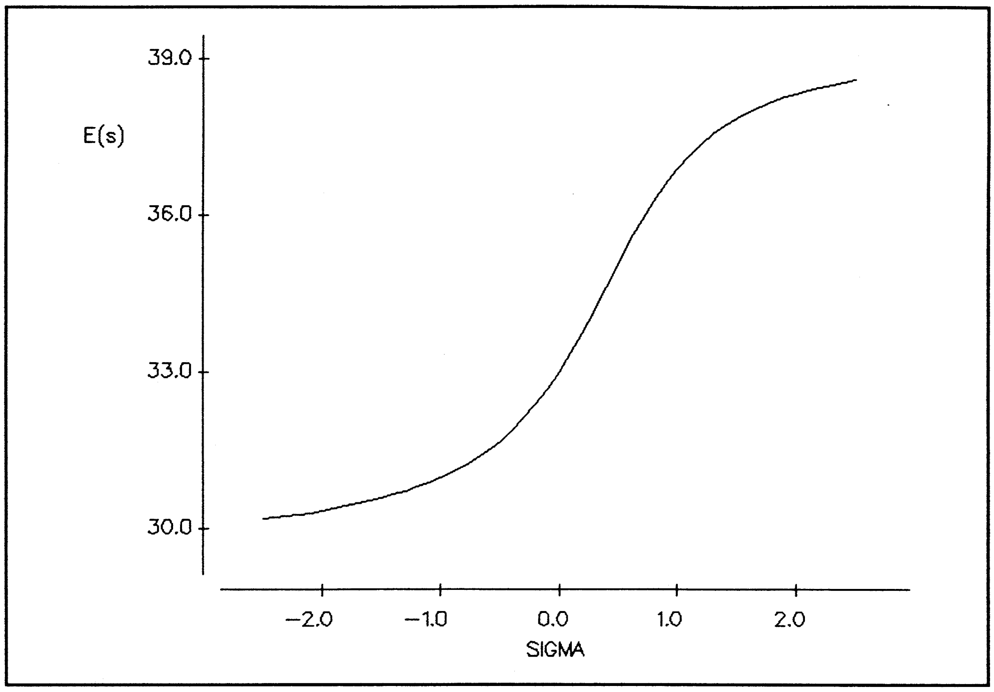

Diagram 5.1:
The relation between σ and $E_{\sigma}(s)$ for a (7,12)-graph.

However, the stepwise method gives direct the *numerical* values of the ML-estimate for different values of s_o as is shown in Table 5.3 together with the starting values using the exact third cumulant. The pseudolikelihood estimates obtained from the logistic regression approach and the true ML-estimates are also given for a comparison.

No of two-stars	ML-estimate of σ	Starting value	Stepwise method	Pseudo-estimate
30	Not defined	Not defined	Not defined	(-)
31	-0.978	-0.970	-0.979	(-1.6)
32	-0.350	-0.352	-0.355	(-)
33	0	0	0	-0.390
34	0.257	0.255	0.258	-0.077
35	0.486	0.465	0.487	0.182
36	0.727	0.648	0.730	0.433
37	1.038	0.812	1.034	0.736
38	1.597	0.962	1.612	1.182
39	Not defined	Not defined	Not defined	(-)

Table 5.3:
Comparisons between different estimates of σ.

It can be seen that the pseudolikelihood estimates are far away from the true ML-estimates, even in the central part of the curve. Furthermore, it may happen that the estimation does not converge. The independent variable in the logistic model will be the difference Δs_{ij} of the number of two-stars when $Y_{ij}=1$ instead of $Y_{ij}=0$, cf section 3.4. If there exists a Δs_{ij}-value such that for all Δs_{ij}-values less than this value, the edge indicators Y_{ij} all take one specific value, e g 1, while for Δs_{ij}-values greater than this value, the indicators all take the opposite value (here 0), then according to Haberman (1974) p 315, this is a necessary and sufficient condition for the ML-estimates of the logistic model not to exist.

It is worth to be noticed the very close agreement between the estimates from the stepwise method and the true values. This will be true in the whole σ-interval. Furthermore, for most values of the number of two-stars, the starting value itself will be very close to the true value. It will even be better than the estimates from the logistic regression approach.

The computer runs have been performed on a PC 386 with a mathprocessor. The running time for the estimation is dependent on the number of steps to obtain the necessary accuracy. For the first approximative value besides the starting value, 100000 simulations are performed. For the following values, the number of simulations increases with 50000 in each round.

The logistic regression has been performed by using the program LR (Stepwise Logistic Regression) in BMDP and a specially written program for only two parameters in the logistic model.

The computer program for the stepwise method is written in Fortran 77. All graphs up to 30 vertices and with an arbitrary number of edges can be handled by the program. Furthermore, graphs with 31-40 vertices can be handled unless the number of edges is not too extreme.

6 DISCUSSION

The estimation of the only parameter σ in the pure clustering model has an analogue to the parameter τ in the pure transitivity model

$$P(G|r) = \exp(\tau t)/\Sigma \exp(\tau t)$$

where t is equal to the number of triangles in the graph. The estimation methods for τ suggested in the literature are the same as for the clustering parameter, namely simulations of Markov graphs to obtain a curve $E_\tau(t)$ against τ and use of pseudolikelihood estimates from a logistic regression respectively. It can be expected that the methods suggested here can be applied for the transitivity parameter too. This will be a task for the future research.

7 REFERENCES

Cassel, P G (1970). *Statistisk databehandling.* Studentlitteratur, Lund.

Frank, O (1980). A Survey of Statistical Methods for Graph Analysis. In *Sociological Methodology 1981*, ed S Leinhardt, San Fransisco: Jossey-Bass, 110-155.

Frank, O (1985). Random Sets and Random Graphs. In *Contributions to Probability and Statistics in Honour of Gunnar Blom*, ed J Lanke and G Lindgren, Lund, 113-120.

Frank, O (1988). Random Sampling and Sampling Networks. A Survey of Various Approaches. *Math Inf Sci hum*, 26, 19-33.

Frank, O (1991). Statistical Analysis of Change in Networks. *Statistica Neerlandica*, 45, 283-293.

Frank, O and Strauss, D (1986). Markov Graphs. *J Amer Statist Assoc*. 81, 832-842.

Haberman, S J (1974). *The Analysis of Frequency Data*. The University of Chicago Press, Chicago.

Hammersley, J M and Handscomb, D C (1964). *Monte Carlo Methods*, London: Methuen.

Holland, P W and Leinhardt, S (1981). An Exponential Family of Probability Distributions for Directed Graphs (with discussion). *J Amer Statist Assoc*, 76, 33-65.

Karonski, M (1982). A Review of Random Graphs. *J Graph Theory*, 6, 349-389.

Kendall, M G and Stuart, A (1963). *The Advanced Theory of Statistics*, Volume 1. Second Edition. Charles Griffin and Company, London.

Snijders, T A B (1989). Enumeration and Simulation Methods for 0-1 Matrices with given Marginals. Research Report for the Stockholm Conference on Random Graphs and Applications. Department of Statistics, University of Stockholm.

Strauss, D (1986). On a General Class of Models for Interaction. *SIAM Review*, 28, 513-527.

Strauss, D and Ikeda, M (1990). Pseudolikelihood Estimation for Social Networks. *J Amer Statist Assoc*. 85, 204-212.