MATHÉMATIQUES ET SCIENCES HUMAINES

OVE FRANK

Composition and structure of social networks

Mathématiques et sciences humaines, tome 137 (1997), p. 11-23

<http://www.numdam.org/item?id=MSH_1997__137__11_0>

Math. Inf. Sci. hum. (35e année, n°137, *1997*, pp.11-23)

COMPOSITION AND STRUCTURE OF SOCIAL NETWORKS[1]

Ove FRANK[2]

RÉSUMÉ — Composition et structure des réseaux sociaux

Les réseaux sociaux représentent une ou plusieurs relation entre des individus, et des informations sur ces individus eux-mêmes. Les réseaux sociaux montrent à la fois la structure du réseau et des informations sur les individus. Des modèles probabilistes peuvent être utilisés pour analyser les interrelations entre les variables structurelles et les variables individuelles ; par exemple pour expliquer comment la structure peut être interprétée par les informations dont on dispose sur les individus ou comment la composition de celles-ci peut être interprétée par la structure. L'auteur discute différents modèles et utilise diverses méthodes statistiques pour illustrer les interrelations entre des données concernant un réseau.

SUMMARY — *Social networks representing one or more relationships between individuals and one or more categorical characteristics of the individuals exhibit both structure and composition. Probabilistic models of such networks can be used for analyzing the interrelations between structural and compositional variables, for instance in order to find how structure can be explained by composition or how structure explains composition. Different models are discussed and different statistical methods are employed to illustrate such interrelationships in network data.*

1. INTRODUCTION

Social networks are composed of individuals, various individual attributes, interindividual relationships, and various attributes of these relationships. The statistical description and analysis of compositional and structural data on social networks can benefit from the use of probabilistic models that formalize and separate composition and structure in various ways. The purpose of this article is to review and extend some of the most common social network models and illustrate how composition and structure can be reflected in the probabilistic assumptions. For other reviews and expositions of social network models, reference is given to the books by Knoke and Kuklinski (1982), Pattison (1993), and Wasserman and Faust (1994) and to the articles by Frank (1981, 1988a).

Section 2 introduces the concept of a colored multigraph in order to represent a social network with attributes attached to individuals and relationships. A few examples of the use of colored multigraphs are discussed to show the generality and flexibility of the concept. In particular, colored multigraphs comprise simple graphs and digraphs as well as graphs with both directed and undirected edges.

The simplest probabilistic models of colored multigraphs have independent dyads (induced subgraphs of order two). Sections 3 and 4 give some results for simple graphs and digraphs

---

[1] This article presents an extended version of a talk given at the International Sunbelt Social Network Conference in Charleston, SC, 1996.

[2] Department of Statistics, Stockholm University, Sweden.

with independent dyads in order to settle terminology and notation and to provide an adequate background for more general models.

Sections 5 and 6 generalize the results of Sections 3 and 4 to simple graphs and digraphs with possible dependence between incident dyads and conditional independence between non-incident dyads, i.e. Markov dependence for dyads. Section 7 gives some results for graphs and digraphs with general dependence between the dyads, and Section 8 treats the extension to colored multigraphs. The interplay between compositional and structural modeling is discussed in Section 9 together with a description of a few broad classes of network models. Section 10 gives a brief introduction to data analytic methods in social network analysis.

2. SOCIAL NETWORK DATA REPRESENTATIONS

To analyse simultaneous distributions of several attributes it is often convenient to categorize or recategorize each attribute to two or a few categories only. All attributes in the social networks are here assumed to be categorical. A general specification of a social network on n individuals can be given by a p-dimensional vector variable x defined on the individuals, a q-dimensional vector variable y defined on the unordered pairs of individuals, and an r-dimensional vector variable z defined on the ordered pairs of individuals. Thus network data consist of individual data vectors x_i for $i=1,...,n$ and pairwise data vectors $y_{ij}=y_{ji}$ and z_{ij} for $i=1,...,n$ and $j=1,...,n$ with $i\neq j$.

If the p components of x have $a_1,...,a_p$ categories, the q components of y have $b_1,...,b_q$ categories, and the r components of z have $c_1,...,c_r$ categories, there is a possible total of $a=a_1,...,a_p$ combined categories of individual attributes, $b=b_1,...,b_q$ combined categories of attributes of unordered pairs of individuals, and $c=c_1,...,c_r$ combined categories of attributes of ordered pairs of individuals. If the network is represented as a graph on n vertices with $\binom{n}{2}$ undirected edges and $n(n\text{-}1)$ directed edges, each vertex is given one of a distinct colors, each undirected edge is given one of b distinct colors, and each directed edge is given one of c distinct colors. In total there is a possibility of $a^n b^{\binom{n}{2}} c^{n(n-1)}$ distinct colored labeled multigraphs.

For dyads (colored labeled multigraphs of order two) there are a^2bc^2 distincts versions, and for triads (colored labeled multigraphs of order three) there are $a^3b^3c^6$ distinct versions. In particular, $(a,b,c)=(1,2,1)$ yields 2 dyads (labeled graphs of order two) and 8 triads (labeled graphs of order three), and $(a,b,c)=(1,1,2)$ yields 4 dyads (labeled digraphs of order two) and 64 triads (labeled digraphs of order three). Structural properties of graphs are often defined as properties that are invariant under isomorphism, and the class of isomorphic labeled graphs can be represented by an unlabeled graph. There are 2 undirected dyads (unlabeled graphs of order two) and 4 undirected triads (unlabeled graphs of order three), and there are 3 directed dyads (unlabeled digraphs of order two) and 16 directed triads (unlabeled digraphs of order three). According to Frank (1988b) there are $\binom{ac+1}{2}b$ dyads and $\binom{abc^2+2}{3}-a^2b^2c^2\binom{c}{2}$ triads for unlabeled colored multigraphs. The counts of these dyads and triads among all induced subgraphs of order two and three, respectively, contained in a colored multigraph of order n are important structural statistics. In exploratory social network analysis the dyad and triad counts might be convenient summary statistics, and under special probabilistic assumptions they can be shown to be sufficient statistics. See Frank and Strauss (1986), Frank (1985, 1988a), and Frank and Novicki (1993).

In order to illustrate the generality and flexibility of the colored multigraph, consider first the case of a social network comprising individuals of two kinds and three symmetrical binary relations. This network can be represented by a colored multigraph with $(a,b,c)=(2,8,1)$. The 2 vertex colors correspond to the two kinds of individuals. The 8 undirected edge colors correspond to the possible combinations of occurrence or non-occurrence on each one of the three symmetrical relations. The single color on directed edges corresponds to the absence of any unsymmetrical binary relation.

As a second example consider the case of a social network comprising males and females of three age groups and two binary relations. Both the relations describe different kinds of pairwise contacts between the individuals, and contact intensities are reported as low, medium, or high from each individual to every other individual. Here the social network can be represented as a colored multigraph with $(a,b,c)=(6,1,9)$. The 6 vertex colors correspond to the combinations of gender and age group. The single undirected edge color means that there is no symmetrical relation. The 9 directed edge colors correspond to the combinations of contact intensities for the two kinds of contacts.

Finally, consider the case of a social network defined on individuals categorized as being presently employed or not, as having ever been employed or not, and as being healthy or not. There is information about kinship, about father-son relationships, and about brother and/or sister relationships. This example implies that a straightforward cross-classification of the attributes might lead to categorical combinations that are known apriori to be impossible (structural zeros among the cross-classification frequencies). It might be advantageous to avoid this kind of attribute redundancy by defining combined attributes so that the total number of combined categories is reduced. For instance, the straightforward approach is to define three binary individual attributes (indicating present employment, previous or present employment, and healthy condition) and three binary relationships corresponding to the two symmetrical kinship and brother and/or sister relationship, and one asymmetrical father-son relationship. This yields a colored multigraph with $(a,b,c)=(8,4,2)$. An alternative approach keeping the same information is the following. Introduce an employment attribute with three categories corresponding to never employed, previously employed only, and presently employed. Keep the health status attribute. Furthermore, introduce a kinship attribute with four categories corresponding to no kinship, father-son relationship, brother and/or sister relationship, and other kinds of kinship. This yields a colored multigraph with $(a,b,c)=(6,1,4)$. Thus, compared to the initial approach a and b have been reduced but not c. Using the dyad formula reported above it follows that the number of non-isomorphic dyads have been reduced from 544 to 300. A further reduction to 234 non-isomorphic dyads can be achieved by modifying the alternative approach so that the initial father-son relationship is kept and the initial two symmetrical relationship are replaced by one symmetrical relationship with three categories. This yields a colored multigraph with $(a,b,c)=(6,3,2)$.

Generally it is important to have mutually exclusive and exhaustive categories for vertices, undirected edges, and directed edges. As the last example illustrates, a reduction of b that implies an increase in c is guaranteed neither to reduce nor to increase the number of dyads.

3. INDEPENDENT DYAD MODELS FOR SIMPLE GRAPHS

Bollobas (1985) and Palmer (1985) in their extensive accounts of the theory of random graphs treat simple parametric models and uniform models which have influenced much research on graphical limit theorems, graphical evolution, and random graph processes. Such models have not got a sufficiently rich probabilistic structure for most applications and need to be extended to provide good fit to statistical network data. Extensions might be mixture

distributions of simple models or various kinds of generalizations allowing more complex specifications.

Even for the simplest network models the statistical aspects deserve some attentions. Network data offer numerous possibilities of varying the sampling and observation procedures, and this might result in unconventional statistical problems. For instance, vertex sampling and induced subgraph observation or different kinds of complete and partial snowball sampling have been discussed in the literature. See Frank (1988a) for references.

Here interest is not focused on sampling or other sources of explanation for the stochastic dependence prevailing in netwotk data. Models with different kinds of structure dependence are investigated without referring to whether sampling, measurement errors or other sources of variation explain the randomness. Dyad independence for simple graphs is a starting point.

Consider a finite vertex set $V=\{1,...,n\}$ and a simple random graph on V defined by its symmetric adjacency matrix $Y=(Y_{ij})$ with $Y_{ii}=0$. There are $2^{\binom{n}{2}}$ possible outcomes y of Y, and under dyad independence the probability function is given by

$$P(Y=y) = p(y) = \prod_{i<j} p_{ij}^{y_{ij}} \ (1-p_{ij})^{1-y_{ij}}$$

where p_{ij} is the probability of edge $\{i,j\}$. A convenient reparameterization is obtained by introducing the logodds $\alpha_{ij} = log[p_{ij} / (1-p_{ij})]$

so that $$p(y) = c^{-1} \ exp \sum_{i<j} \alpha_{ij} \ y_{ij}$$

where $$c = \prod_{i<j} (1+e^{\alpha_{ij}})$$

is a normalizing constant.

A particular case is the homogeneous model with all edge probabilities equal, $p_{ij}=p$, which is called a Bernoulli (V,p) model. Setting $q=1-p$ and $\alpha=log(p/q)$ implies that

$$p(y) = p^r q^{\binom{n}{2}-r} = (1+e^{\alpha})^{\binom{n}{2}} r^{\alpha r}$$

where $r = \sum_{i<j} y_{ij}$ is the edge frequency of y. Here the edge frequency $R = \sum_{i<j} y_{ij}$ of Y is a minimal sufficient statistic with a binomial $\left(\binom{n}{2},p\right)$-distribution, and the maximum likelihood estimator of p is given by the edge density

$$\hat{p} = R / \binom{n}{2}.$$

An alternative to homogeneity is given by the dyad independence model with a multiplicative edge probability decomposition according to $p_{ij} = p\beta_i\beta_j$ where $\beta_1,..., \beta_n$ are activity probabilities of the vertices and p is a latent edge probability. A manifest edge occurs if and only if the latent edge occurs and is supported by active vertices. Thus the $\binom{n}{2}$ edge probabilities are replaced by $n+1$ probabilities $p,\beta_1,...,\beta_n$ of latent edge and vertex acitivities. The probability function is given by

$$p(y) = p^{r}(\prod_{i=1}^{n} \beta_i^{y_{i.}})(\prod_{i<j} (1 - p\beta_i\beta_j)^{1-y_{ij}})$$

where r is the edge frequency as before and $y_{i.} = \sum_{j=1}^{n} y_{ij}$ is the degree of vertex i. The probability function is invariant to admissible parameter changes that leave $\sqrt{p}\,\beta_i$ invariant for $i=1,\ldots, n$. Identifiability of the parameters can be achieved by imposing the restriction $\sum_{i<j} \beta_i\beta_j = \binom{n}{2}$ This means that the expected numbers of manifest and latent edges are the same.

A similar dyad independence model is obtained by assuming an additive logodds decomposition according to $\alpha_{ij} = \alpha+\gamma_i+\gamma_j$. This implies that the probability function is equal to

$$p(y) = c^{-1}exp \sum_{i=1}^{n} (\gamma_i + \alpha/2)\, y_{i.}$$

where $y_{i.}$ is the degree of vertex i in y. The parameters $\alpha,\gamma_1,\ldots,\gamma_n$ restricted by $\sum_{i=1}^{n} \gamma_i = 0$ are identifiable and can be considered as overall and local specifications of Y. The degrees $Y_1,\ldots,Y_n$ of Y are minimal sufficient statistics. Thus this model might be preferable to the previous model with a multiplicative edge probability decomposition.

4. INDEPENDENT DYAD MODELS FOR SIMPLE DIGRAPHS

A simple directed graph on V is defined by its adjacency matrix $Z=(Z_{ij})$ with $Z_{ii}=0$. A dyad induced by i and j (in that order) is specified by (Z_{ij}, Z_{ji}). There are $2^{n(n-1)}$ outcomes z of Z, and under dyad independence the probability function is given by

$$P(Z=z) = p(z) = \prod_{i<j} p_{ij}\,(z_{ij}, z_{ji})$$

where $p_{ij}(0,0)$, $p_{ij}(0,1)$, $p_{ij}(1,0)$, and $p_{ij}(1,1)$ are the probabilities of a dyad with no edges between i and j, with an edge from j to i only, with an edge from i to j only, and with mutual edges between i and j. It is convenient to introduce dyad probabilities $p_{ij}(k,l)$ for all i and j and put $p_{ij}(k,l)=p_{ji}(l,k)$. By denoting $p_{ij}(1,0)=a_{ij}$ and $p_{ij}(1,1)=b_{ij}$ it follows that $p_{ij}(0,0)=1-a_{ij}-a_{ji}-b_{ij}$, $p_{ij}(0,1)=a_{ji}$, and $b_{ij}=b_{ji}$. Hence

$$p(z) = \left[\prod_{i<j} (1 - a_{ij} - a_{ji} - b_{ij})^{(1-z_{ij})(1-z_{ji})}\, a_{ji}^{(1-z_{ij})z_{ji}}\, a_{ij}^{z_{ij}(1-z_{ji})}\, b_{ij}^{\,z_{ij}z_{ji}} \right] =$$

$$= c^{-1}\, exp\left(\sum_{i\neq j} \alpha_{ij}z_{ij} + \prod_{i<j} \beta_{ij}z_{ij}z_{ji} \right)$$

where

$$\alpha_{ij} = log[a_{ij} / (1-a_{ij}-a_{ji}-b_{ij})],$$
$$\beta_{ij} = log[b_{ij}\,(1-a_{ij}-a_{ji}-b_{ij}) / a_{ij}\,a_{ji}],$$

and c is a normalizing constant given by

$$c = \prod_{i<j} (1 + e^{\alpha_{ij}} + e^{\alpha_{ji}} + e^{\alpha_{ij}+\alpha_{ji}+\beta_{ij}}).$$

A particular case is the homogeneous model with all dyad distributions equal, that is $a_{ij}=a$ and $b_{ij}=b$, or, equivalently, $\alpha_{ij}=\alpha$ and $\beta_{ij}=\beta$ where

$$\alpha = \log a - \log (1-2a-b),$$
$$\beta = \log b + \log (1-2a-b) - 2 \log a.$$

It follows that the frequencies of induced dyads in Z of size 0,1, and 2 are multinomial $\left(\binom{n}{2}; 1-2a-b, 2a, b\right)$-distributed. If these frequencies are denoted

$$N_0 = \sum_{i<j} (1 - Z_{ij})(1 - Z_{ji}),$$

$$N_1 = \sum_{i<j} (Z_{ij} + Z_{ji} - Z_{ij}Z_{ji}),$$

$$N_2 = \sum_{i<j} Z_{ij}Z_{ji},$$

the maximum likelihood estimators of a and b are given by $\hat{a} = N_1 / n(n-1)$ and $\hat{b} = N_2 / \binom{n}{2}$.

An alternative to homogeneity is the dyad independence model with partial homogeneity $\beta_{ij}=\beta$. With no restrictions on α_{ij} there are $n(n-1)+1$ parameters in the model. If an additive decomposition of α_{ij} is assumed according to $\alpha_{ij} = \lambda + \alpha_i + \beta_j$ with $\sum_{i=1}^{n} \alpha_i = 0$ and $\sum_{j=1}^{n} \beta_j = 0$, then there are $2n$ free parameters and sufficient statistics corresponding to in- and outdegrees and the total mutual adge frequency. This is the well known model introduced by Holland and Leinhardt (1981). An extension is obtained by relaxing the partial homogeneity $\beta_{ij}=\beta$ to an additive decomposition according to $\beta_{ij} = \mu + \gamma_i + \gamma_j$ with $\sum_{i=1}^{n} \gamma_i = 0$ This model has $3n-1$ free parameters and sufficient statistics corresponding to in-, out-, and mutual degrees at every vertex, that is

$$\left(\sum_{j=1}^{n} Z_{ij}, \sum_{j=1}^{n} Z_{ji}, \sum_{j=1}^{n} Z_{ij}Z_{ji} \right)$$

for $i=1,\ldots, n$.

5. MARKOV DYAD MODELS FOR SIMPLE GRAPHS

The assumption of stochastic independence between dyads might seem inappropriate, since a network is usually studied because there is an interest in the links and influences across several individuals in the network. Dyad independence means that all links and influences involving three or more individuals are the results of random effects governed by a set of fixed values on dyad parameters. Therefore the structural properties beyond those of dyads are only indirectly controlled and might fail to fit data. It should be preferable to have access to parameters reflecting dyad interactions.

The Markov dyad models introduced by Frank and Strauss (1986) assume that non-incident dyads are conditionally independent but incident dyads might be dependent. They show that this implies that there are

$$n2^{n-1} + \binom{n}{3} - \binom{n+1}{2}$$

parameters and sufficient statistics corresponding to triangles (3-cycles) and stars. The probability function is given by

$$p(y) = c^{-1} \exp\left(\sum_{i<j<k} \tau_{ijk}\, y_{ij}\, y_{jk}\, y_{ki} + \sum_{m=1}^{n-1} \frac{1}{m!} \sum_{i_0,\ldots,i_m} \sigma_{i_0,\ldots,i_m}\, y_{i_0 i_1} \cdots y_{i_0,\ldots,i_m} \right),$$

where the last sum is over distinct vertices, and there are $\binom{n}{3}$ triangle parameters τ_{ijk}, $n\binom{n-1}{m}$ star parameters $\sigma_{i_0,\ldots,i_m}$ for $m=2,\ldots, n-1$, and $\binom{n}{2}$ edge parameters $\sigma_{i_0 i_1}$.. The normalizing constant c is a function of the parameters determinated so that the $2^{\binom{n}{2}}$ probabilities $p(y)$ sum to unity.

Under homogeneity all triangle parameters are equal and the star parameters depend only on the order of the star: $\tau_{ijk} = \tau$ and $\sigma_{i_0,\ldots,i_m} = \sigma_m$ for $m=1,\ldots, n-1$. This implies that isomorphic graphs y get the same probability

$$p(y) = c^{-1} \exp\left(\tau t + \sum_{m=1}^{n-1} \sigma_m s_m\right)$$

where

$$t = \sum_{i<j<k} y_{ij} y_{jk}\, y_{ki}$$

is the number of triangles in y,

$$s_m = \frac{1}{m!} \sum_{i_0,\ldots,i_m} y_{i_0 i_1} \cdots y_{i_0,\ldots,i_m}$$

(with distinct vertices in the sum) is the number of m-stars (stars of size m) in y for $m=2,\ldots, n-1$, and $s_1 = 2 \sum_{i<j} y_{ij}$ is twice the number of edges in y.

A simplified version of the homogeneity model assumes $\sigma_m=0$ for $m>2$. If the remaining star parameters are denoted $\sigma_1=\rho/2$ and $\sigma_2=\sigma$, it follows that

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

where r,s,t are the frequencies of edges, 2-stars, and triangles in y. This is a simple model for a random graph with dependence between incident edges. Inference for this model is discussed by Frank and Strauss (1986), Frank (1991), and Frank and Nowicki (1993).

Without assuming homogeneity but assuming that the star parameters are 0 for all stars of size 3 or more, the model has $\binom{n}{2}$ parameters σ_{ij} for $i<j$, $n\binom{n-1}{2}$ parameters σ_{ijk} for

$j<k$, and $\binom{n}{3}$ parameters τ_{ijk} for $i<j<k$. One way of simplifying this model is to make additive decomposition assumptions according to

$$\begin{aligned}\sigma_{i...j} &= \rho + \alpha_i + \alpha_j \\ \sigma_{ijk} &= \sigma_i + \beta_j + \beta_k \\ \tau_{ijk} &= \tau + \gamma_i + \gamma_j + \gamma_k\end{aligned}$$

with restrictions $\Sigma\alpha_i = \Sigma\beta_i = \Sigma\gamma_i = 0$. This model has $4n$-1 parameters. Further feasible simplifications are $\sigma_i = \sigma$, $\alpha_i = \beta_i = \gamma_i$, $\alpha_i = 0$, $\beta_i = 0$, or $\gamma_i = 0$. The vertex parameters α_i, β_i, γ_i, σ_i can be considered as controlling for edges, 2-paths, 3-cycles, and 2-stars at vertex i. (A 2-star at i with edges to j and k is also a 2-path at j and a 2-path at k.)

6. MARKOV DYAD MODELS FOR SIMPLE DIGRAPHS

In the directed case, conditional independence for non-incident dyads implies that a random digraph with adjacency matrix Z has a probability function

$$p(z) = c^{-1} exp \sum_{a \leq z} \theta(a)$$

where the sum is over all adjacency matrices of subgraphs of z and $\theta(a)=0$ unless a is an adjacency matrix of any of the following graphs on V : a single edge, a 2-cycle, a star of order 3 or more, a triangle. The parameter $\theta(a)$ is denoted $\rho_{ij}^{(1)}$ for a single edge, $\rho_{ij}^{(2)}$ for a 2-cycle, $\sigma_{i_0,\ldots,i_m}^{(k,l)}$ for a star of order m+1 with center i_0 and edges from i_0 to the first k vertices among $i_1,\ldots, i_m$ and edges to i_0 from the last l vertices among $i_1,\ldots, i_m$, and $\tau_{ijk}^{(m)}$ for a triangle of type m. There are triangles of 7 types. Types 1 and 2 have size 3. Types 3, 4, and 5 have size 4. Type 6 has size 5 and Type 7 has size 6. Type 2 is a 3-cycle. Type 3 has a vertex with two outedges and Type 4 has a vertex with two inedges. In total there are $n(n\text{-}1)$ edges, $\binom{n}{2}$ 2-cycles, $n\binom{n\text{-}1}{m}3^m$ m-stars (stars of order m+1) for m=2,..., n-1, and $27\binom{n}{3}$ triangles. The 27 triangles on each fixed set of three vertices consist of 6+2+3+3+6+6+1 triangles of the 7 types.

Under homogeneity the parameters are denoted

$$\rho_{ij}^{(1)} = \rho_1,\ \rho_{ij}^{(2)} = \rho_2,\ \sigma_{i_0,\ldots,i_m}^{(k,l)} = \sigma_{klm},\ \tau_{ijk}^{(m)} = \tau_m$$

and it follows that

$$p(z) = c^{-1} \exp\left(\rho_1 r_1 + \rho_2 r_2 + \sum_{m=2}^{n-1} \sum_{k+l \leq m} \sigma_{klm}\, s_{klm} + \sum_{m=1}^{7} \tau_m t_m\right)$$

where r_1 and r_2 are the frequencies of edges and 2-cycles in z, s_{klm} is the frequency of m-stars having k outedges and l inedges in z, and t_m is the frequency of triangles of Type m in z. The number of parameters and sufficient statistics is $5+\binom{n+2}{3}$. Only 15 parameters remain if star parameters are set to 0 for m>2. In this case the sufficient statistics are the

numbers of edges, 2-cycles, 2-stars of six kinds, and triangles of seven kinds. An equivalent set of sufficient statistics is the set of triad counts. This set contains 16 kinds of triads, and their counts sum to $\binom{n}{3}$. Thus the triad counts are sufficient statistics if we assume homogeneous Markov dyads with no parameters for stars of order 4 or more.

By dropping homogeneity and assuming additive decompositions of the parameters it is possible to obtain models that might be of interest as alternatives to the Holland-Leinhardt model. Assuming

$$\rho_{ij}^{(1)} = \lambda + \alpha_i + \beta_j$$

with $\Sigma\alpha_i = \Sigma\beta_j = 0$ together with partial homogeneity or zero values on other parameters should give interesting alternatives.

7. CONDITIONAL INDEPENDENCE MODELS FOR SIMPLE GRAPHS AND DIGRAPHS

Consider first the adjacency matrix $Y=(Y_{ij})$ of a simple random graph on V. Define d_{ijkl} equal to 0 or 1 according to whether or not Y_{ij} and Y_{kl} are stochastically independent conditional on the rest of Y, that is conditional on all elements of Y except Y_{ij}, Y_{ji}, Y_{kl}, Y_{lk}. Obviously $d_{ijkl}=d_{klij}$ for all i, j, k, l. Moreover, $d_{ijkl}=0$ if $i=j$ or $k=l$, and $d_{ijij}=1$ if $i\neq j$. Let $a=(a_{ij})$ be the adjacency matrix of a simple graph on V. There are $2^{\binom{n}{2}}$ such matrices. Define a function $\theta(a)$ on the class of adjacency matrices in such a way that $\theta(a)=0$ unless

$$a_{ij}\, a_{kl} \leq d_{ijkl}$$

for all i, j, k, l. Then the probability function of Y can be shown to be given by

$$P(Y=y) = p(y) = c^{-1} \exp \sum_{a \leq y} \theta(a)$$

where c is a normalizing constant determined so that the $p(y)$ sum to 1 for all adjacency matrices y of simple graphs on V. There are terms in the exponent for all subgraphs a of y having $\theta(a)\neq 0$. These terms corresponds to subgraphs a that are restricted by the numbers in $d=(d_{ijkl})$. We can consider d as an adjacency matrix of a graph on V^2 with loops at all $(i,j)\in V^2$ with $i\neq j$. This graph is called the dependence graph of Y. If $a=(a_{ij})$ is considered as an indicator of a subset of V^2, the restriction on a in terms of d means that a should correspond to a clique of the dependence graph d.

Consider now the adjacency matrix $Z=(Z_{ij})$ of a simple random digraph on V. Define d_{ijkl} equal to 0 or 1 according to whether or not the dyads (Z_{ij},Z_{ji}) and (Z_{kl},Z_{lk}) are stochastically independent conditional on the rest of Z. Proceeding as above we define a function $\theta(a)$ for adjacency matrices $a=(a_{ij})$ of simple digraphs on V, such that $\theta(a)=0$ unless

$$a_{ij}\, a_{kl} \leq d_{ijkl}$$

for all i, j, k, l. It follows that

$$P(Z=z) = p(z) = c^{-1} \exp \sum_{a \leq z} \theta(a)$$

for any adjacency matrix z of a simple digraph on V. Again the relevant parameters correspond to subsets of V^2 that are cliques of the dependence graph d.

8. CONDITIONAL INDEPENDENCE MODELS FOR COLORED MULTIGRAPHS

The conditional independence models for simple graphs and digraphs considered in the previous three sections can be extended to colored multigraphs by applying conditional independence specifications to more general multivariate random variables. The books by Whittaker (1990) and Edwards (1995) discuss conditional independence modelling in a general context. The field is known as graphical modelling not because network data is of concern but because graphs are used to represent multivariate models.

Consider a random colored multigraph given by a vector $X=(X_i)$ having a^n outcomes of vertex colors, a symmetrical matrix $Y=(Y_{ij})$ having $b^{\binom{n}{2}}$ outcomes of colors of undirected edges, and a matrix $Z=(Z_{ij})$ having $c^{n(n-1)}$ outcomes of colors of directed edges. There are various possibilities of specifying the probabilistic structure of (X, Y, Z).

One way is to condition on X and consider the $\binom{n}{2}$ dyad variables (Y_{ij}, Z_{ij}, Z_{ji}) for $i<$j conditional on X. These variables are chosen to be the basic variables for which a conditional dependence structure needs to be defined. An alternative is to consider the $3\binom{n}{2}$ variables Y_{ij}, Z_{ij}, Z_{ji} for $i<j$ separately conditional on X as the basic variables.

Another way is to consider the $\binom{n}{2}$ dyad variables $(X_i, X_j, Y_{ij}, Z_{ij}, Z_{ji})$ for $i<j$ as the basic variables with a conditional dependence structure. Here all incident dyads are certainly dependent via their common vertex colors. An alternative is to consider all $n+3\binom{n}{2}$ variables separately as the basic variables with a conditional dependence structure. However, the symmetry of the approach with dyad variables might by useful.

A convenient choice in any particular application should be a way that offers a simple conditional dependence structure. In addition to specifying a conditional dependence structure, there is also need for a number of parameters. This number depends not only on the number of basic variables but also on the number of outcomes of these variables.

In general, the conditional dependence structure of a set of random categorical variables $W=(W_{ij})$ on V^2 is given by a dependence graph on V^2 with adjacency matrix $d=(d_{ijkl})$. Let A be a subset of V^2 and $a=(a_{ij})$ the corresponding matrix of indicators a_{ij} that are 1 or 0 according to whether or not $(i,j) \in A$. Such a subset A is a clique of the dependence graph if and only if $a_{ij}a_{kl} \leq d_{ijkl}$ for all i, j, k, l. The probability function of W can be given by

$$P(W=w) = c^{-1} \exp \sum_{A \subseteq V^2} \theta_A (w)$$

where c is a normalizing constant, and the functions θ_A are identically zero for subsets A that are not cliques of the dependence graph. Moreover, $\theta_A(w)$ depends on w_{ij} for $(i,j) \in A$ only, and $\sum_{w_{ij}} \theta_A(w) = 0$ for all A, w, and $(i,j) \in A$. By introducing the submatrix $w_A=(w_{ij}\, a_{ij})$ which is w with w_{ij} replaced by 0 if $(i,j) \notin A$, it follows that $\theta_A(w)=\theta_A(w_A)$.

9. NETWORK MODELLING

Network modelling can help in understanding or predicting network behaviour. Like in all modelling, prior knowledge is confronted with empirical facts, and it is not always clear whether observed discrepancies between model and reality should call for a minor model modification or a major change to another class of models. Since no model is perfect and there always are statistics that do not follow the model pattern, it is good practice to be prepared for future model building by collecting also such general descriptive statistics that are not needed for estimating the model in use. Compositional and relational structure should be reflected among such general network statistics. Composition statistics are mainly various subgraph counts: vertex counts, dyad counts, triad counts, etc. Structure statistics are for instance distances, reachability, and connectivity of various kinds. Composition refers to how many elements of different kinds that make up the network, and structure refers to macro properties involving more than local properties. The interplay between micro and macro, between composition and structure is the object of network modelling.

It is possible to distinguish a few broad classes of network models that can be described as follows.

Loglinear models are models obtained for instance by conditional independence assumptions as illustrated in this paper. It is also possible to formally apply the loglinear methods for contingency table analysis to categorical counts of vertices, undirected edges, and directed edges even if appropriate independence assumptions are not met. It is tempting to use easily available statistical computer packages for this kind of analysis even if it is not yet theoretically justified or discredited.

Mixture models are models that express the probability function as a weighted average of a family of simple probability functions with unknown mixing weights. Usually the number of components in the mixing distribution is also unknown. Mixture models are often hard to estimate, and mixture models for networks should be no exception. See Frank (1989).

Block models refer to network models with structure parameters that depend on individuals through some kind of individual categories only. Thus the vertex set is partitioned into disjoint and exhaustive categories, and there is a partial homogeneity within and between the vertex categories. Block models with random categories can be considered as a special kind of mixture models. For block model testing, see for instance Wellman *et al.* (1991).

Metric models for random colored multigraphs have a distance defined on the set of outcomes. The probabilities of the outcomes have a maximum at a certain central graph and decrease with increasing distance from this graph. Such models are especially appropriate if the random variation is due to measurement or observation errors, and there is a fixed unknown colored multigraph to be estimated. See Banks and Carley (1994).

10. DATA ANALYTIC METHODS

Network modelling is often simplified if homogeneity can be assumed. A primary concern in exploratory network analysis is therefore to decompose the network into more homogeneous subnetworks. One way of doing this is by using local dyad counts, that is the number of dyads of different kinds at each vertex. The local dyad counts in colored multigraphs can be subjected to a cluster analysis in order to find out the similarities and dissimilarities between the vertices. In favourable cases, only a few different kinds of vertices need to be distinguished, and separate modelling can be tried out within and between the clusters. For further discussion and illustration of this method reference is given to Frank, Komanska, and Widaman (1985) and Frank, Hallinan, and Nowicki (1985).

Regression methods are useful to explain relationships between variables. With categorical variables regression and classification trees might be more appropriate than methods based on numerical scales. A general reference is Breiman *et al.* (1984), and an application to graph data is given by Frank (1986).

Ordering dyads, triads, and higher order induced subgraphs according to decreasing frequencies might give some insight into the network structure. Finding the locally most common dyads, triads, etc. can for instance reveal that some individuals are central or special in other respects. This can be useful if parameters varying with the individuals are relevant in the model. Dyad and triad counts are also useful for comparing networks and for detecting structural changes. See, for instance, Frank (1987).

BIBLIOGRAPHY

BANKS, D. and CARLEY, K., (1994), "Metric inference for social networks", *Journal of Classification* 11, 121-149.

BOLLOBAS, B., (1985), *Random Graphs*, London, Academic Press.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J., (1984), *Classification and Regression Trees.*, Belmont, Wadsworth.

EDWARDS, D., (1995), *Introduction to Graphical Modelling*, New York, Springer-Verlag.

FRANK, O., (1981), "A survey of statistical methods for graph analysis", *Sociological Methodology*, 1981, edited by S. Leinhardt, San Francisco, Jossey-Bass, 110-155.

FRANK, O., (1985), "Random sets and random graphs", *Contributions to Probability and Statistics in Honour of Gunnar Blom*, edited by J. Lanke and G. Lindgren, Lund, 113-120.

FRANK O., (1986), "Growing classification and regression trees on network data", *Classification as a Tool of Research*, edited by W. Gaul and M. Schader, Amsterdam, North-Holland, 137-143.

FRANK, O., (1987), "Multiple relation data analysis", *Operations Research Proceedings 1986*, edited by H. Isermann et al, Berlin-Heidelberg, Springer-Verlag, 455-460.

FRANK, O., (1988a), "Random sampling and social networks - a survey of various approaches", invited paper presented at Centre International de Recontres Mathématiques, Marseille-Luminy, 1987, *Mathématiques, Informatique et Sciences humaines* , 26, 104, 19-33.

FRANK, O., (1988b), "Triad count statistics", *Discrete Mathematics* , 72, 141-149.

FRANK, O., (1989), "Random graph mixture", *Annals of the New York Academy of Sciences* , 576, 192-199.

FRANK, O., (1991), "Statistical analysis of change in networks", *Statistica Neerlandica*, 45, 283-293.

FRANK, O., HALLINAN, M.., and NOWICKI, K., (1985), "Clustering of dyad distributions as a tool in network modelling", *Journal of Mathematical Sociology* , 11, 47-64.
FRANK, O., KOMANSKA, H.., and WIDAMAN, K., (1985), "Cluster analysis of dyad distributions in networks", *Journal of Classification* , 2, 219-238.
FRANK, O. and NOWICKI, K., (1993), "Exploratory statistical analysis of networks", *Annals of Discrete Mathematics* , 55, 349-366.
FRANK, O. and STRAUSS, D., (1986), "Markov graphs", *Journal of the American Statistical Association* , 81, 832-842.
HOLLAND, P. and LEINHARDT, S., (1981), "An exponential family of probability distributions for directed graphs", *Journal of the American Statistical Association* , 76, 33-50.
KNOKE, D, and KUKLINSKI, J.H., (1982), *Network Analysis*, Newbury Park, Sage.
PALMER, E., (1985), *Graphical Evolution*, New York, Wiley.
PATTISON, P., (1993), *Algebraic Models for Social Networks*, Cambridge, Cambridge University Press.
WASSERMAN, S. and FAUST, K., (1994), *Social Network Analysis: Methods and Applications*. Cambridge, Cambridge University Press.
WELLMAN, B., FRANK, O., ESPINOZA, V., LUNDQUIST, S. and WILSON, C., (1991), "Integrating individual, relational and structural analysis", *Social Networks* , 13, 223-249.
WHITTAKER, J., (1990), *Graphical Models in Applied Multivariate Statistics*, Chichester, Wiley.