MATHÉMATIQUES ET SCIENCES HUMAINES

OVE FRANK

Random sampling and social networks. A survey of various approaches

Mathématiques et sciences humaines, tome 104 (1988), p. 19-33

<http://www.numdam.org/item?id=MSH_1988__104__19_0>

Math. Inf. Sci. hum. (26^e année, n°104, *1988*, p.19-33).

RANDOM SAMPLING AND SOCIAL NETWORKS A SURVEY OF VARIOUS APPROACHES

Ove FRANK[1]

1. GRAPH SAMPLING

In statistics, the concept of random sampling is used to describe selection procedures that are governed by probability laws. Thus, two-stage sampling, cluster sampling, Bernoulli sampling, and simple random sampling are examples of selection procedures where the randomness is controlled by the design.

Random sampling is also used to describe selection schemes influenced by a series of different causes that are considered to be either irrelevant or impossible to control, but that are supposed to be sufficiently well described by some probabilistic model. Such selection procedures controlled by "nature" are often assumed to produce random samples from a probability distribution belonging to some specific parametric family of distributions.

Both design-controlled and the model-based random samples are encountered in all areas of applied statistics. Generally speaking, statistical analysis is not confined to designs or models but is applied to a combination thereof. The random sample consists of observations from a probability distribution that contains controllable design parameters as well as other parameters modelling the data.

The analysis of observations of social networks and other systems of interrelationships between some basic units requires random graphs and statistical probability distributions on sets of graphs. Random graph theory has developed rapidly since the appearance of the classical papers by Erdös and Rényi (1959, 1960). The monographs by Bollobas (1985) and Palmer (1985) give excellent accounts of this field of research.

Statistical questions related to graph sampling cannot be handled without access to parametric or non parametric families of distributions of random graphs. Uniform probability distributions over a set of graphs of fixed order and size are usually not appropriate for statistical applications. A Bernoulli graph with a common edge probability assigned to independent edge occurrences is usually not rich enough to model empirical data. A richer parametric family or an even richer nonparametric family of probability distributions is generally needed as an appropriate statistical graph model.

[1] Department of Statistics, University of Stockholm, S-106 91 Stockholm, Sweden.

To define a statistical graph model, simple random graphs can be used as components of mixture distributions and parts of latent structure combinations of various kinds. Some illustrations of this are given in the next section. More elaborate random colorations and multigraphs are described in Section 3, and several examples illustrate the use of randomly colored multigraphs. This model focuses interest on the simultaneous consideration of composition and structure of a network. The idea of separating composition and structure and distinguishing between whether or not they are random is used to classify various models. Section 4 gives a survey of some multiparametric random graphs introduced by Holland and Leinhardt (1981), Frank and Strauss (1986) and others. In particular, the conditional dyad independence assumption is compared to the Markov dependence assumption at some length. Section 5 is devoted to random graph models in which the emphasis is on sampling design. Pure design-based inference as well as Bayesian superpopulation approaches are discussed. Finally, Section 6 gives some concluding views on research trends and prospects in statistical graph theory.

2. SIMPLE RANDOM GRAPHS

This section describes some basic random graphs and gives a few illustrations of how to obtain more elaborate models by combining such random graphs to obtain mixtures and other latent structures.

A *simple graph* g on a vertex set V can be identified with its edge set, that is, a subset of the set V^2 of pairs of vertices. If loops are not allowed, g is a subset of the set $V^{(2)}$ of pairs of distinct vertices. If all edges are undirected, g is symmetric in the sense that $(i,j) \in g$ implies that $(j,i) \in g$. Undirected graphs g can also be considered as subsets of the union of V and the set $\binom{V}{2}$ of the unordered pairs of distinct vertices. If loops are not allowed, g is simply a subset of $\binom{V}{2}$. The cardinalities of V and g are called the order and size of the graph.

Let G be a family of simple graphs defined on V. Set $|V|=n$. If G is the set of undirected graphs of order n and size r having 1 loops, G contains

$$|G| = \binom{n}{1} \binom{\binom{n}{2}}{r-1}$$

graphs. If G is the set of directed graphs of order n and size r having 1 loops, G contains

$$|G| = \binom{n}{1} \binom{n(n-1)}{r-1}$$

graphs. A random graph Y in G is a graph chosen from G according to a probability distribution. The probability function of Y is denoted by

$$P(Y=y) = p(y) \text{ for } y \in G.$$

If $p(y) = 1/|G|$ for $y \in G$, Y is said to be *uniform* on G. For instance, a uniform random undirected graph of order n and size r having no loops has $p(y) = 1/\binom{n'}{r}$ where $n' = \binom{n}{2}$, and uniform random undirected graph of order n and size r having any number of loops has $p(y) = 1/\binom{n'}{r}$ where $n' = \binom{n+1}{2}$.

Let m denote the maximum number of graphs in G, no two of which are isomorphic. Then G is the disjoint union $G_1 \cup ... \cup G_m$ of m classes of isomorphic graphs. If Y has a probability distribution on G that is invariant under isomorphism, there are probabilities $p_1,...,p_m$ such that $p_1+...+p_m = 1$ and $p(y) = p_i / |G_i|$ for $y \in G_i$ and $i=1,...,m$. If furthermore $p_i = 1/m$ for $i=1,...,m$, Y is said to be *isomorphically uniform* on G. For instance, an isomorphically uniform random undirected graph of order 4 and size 3 having no loops has p(y) = 1/36 if y is a path, and p(y) = 1/12 otherwise. The corresponding uniform random graph has p(y) = 1/20 for the same graphs y. Table 1 illustrates this and some other comparisons between uniform and isomorphically uniform distributions. Generally, uniform and isomorphically uniform distributions on G are equal if and only if any two graphs in G have the same number of isomorphisms.

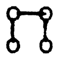
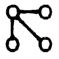
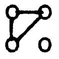
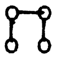
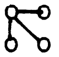
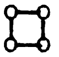
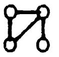
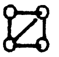
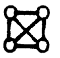
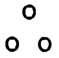
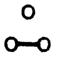
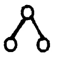
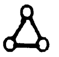
G_i	$\|G_i\|$	Uniform	Isomorphically uniform
	12	1/20	1/36
	4	1/20	1/12
	4	1/20	1/12
	12	1/38	1/72
	4	1/38	1/24
	3	1/38	1/18
	12	1/38	1/72
	6	1/38	1/36
	1	1/38	1/6
	1	1/8	1/4
	3	1/8	1/12
	3	1/8	1/12
	1	1/8	1/4

Table 1. Three examples of a uniform and an isomorphically uniform distribution.

Let g be a fixed graph on V, and let Y be obtained from g by keeping or deleting its edges independently with probabilities p and q = 1-p, respectively. Then Y is said to be a *Bernoulli* (g,p) *graph* or a *Bernoulli* (p) *subgraph* of g. Here

$$p(y) = p^{|y|}\,q^{|g|-|y|} \quad \text{for} \quad y \in 2^g$$

(2^g denotes the set of all subsets of g). In particular, a Bernoulli (p) subgraph of the complete undirected graph g = $\binom{V}{2}$ is one of the random graphs most frequently investigated ; with p = $r/\binom{n}{2}$, it is often used as an approximation to the uniform random undirected graph of order n and size r having no loops. For p = 1/2, the Bernoulli (p) subgraph of g is a uniform random graph on the set of subgraphs of g. The probability distribution of any Bernoulli graph is invariant under isomorphism.

If a directed random graph Y on V = {1,...,n} is obtained by selecting independently and randomly at each vertex i ∈ V a number a_i of vertices in V to be joined by edges from i, then Y is said to be a *random* $(a_1,...,a_n)$ - *mapping from* V. Analogously, a *random* $(b_1,...,b_n)$ - *mapping to* V is defined by selecting independently and randomly at each i ∈ V a number b_i of vertices in V to be joined by edges to i. In particular, a random (1,...,1) - mapping from V is simply called a *random mapping* on V.

Let $G(a_1,...,a_n)$ be the set of directed graphs on V having a_i vertices joined by edges from i for i=1,...,n. Thus a random $(a_1,...,a_n)$ - mapping Y from V is a uniform random graph on $G(a_1,...,a_n)$ having probability function

$$p(y) = 1/\binom{n}{a_1}\cdots\binom{n}{a_n} \quad \text{for } y \in G(a1,...,an) \; .$$

It is a isomorphically uniform if and only if all the a_i are equal. A random $(a_1,...,a_n)$ - mapping from V with no loops is analogously found to be a uniform random graph on the set $G_0(a_1,...,a_n)$ containing $\binom{n-1}{a_1}...\binom{n-1}{a_n}$ directed graphs on V having a_i vertices other than i joined by edges from i for i=1,...,n.

If Y is a random $(a_1,...,a_n)$ - mapping, and Y' is obtained from Y by reversing all the edges, the intersection graph Y ∩ Y' consists of all pairs of vertices that are joined both ways in Y, and the union graph Y ∪ Y' consists of all pairs of vertices that are joined in at least one direction in Y. The undirected graphs corresponding to the symmetric graphs Y ∩ Y' and Y ∪ Y' are called the *strongly and weakly symmetrized versions* of Y, respectively. In particular, a strongly symmetrized random mapping consists of a number of isolated loops and edges corresponding to the loops and 2-cycles in the random mapping.

Let $Y_1,...,Y_k$ be random graphs in G with probability functions

$$P(Y_i = y) = p_i(y) \quad \text{for } y \in G$$

where i=1,...,k. A random graph Y is said to be a mixture of $Y_1,...,Y_k$ with mixing probabilities $\theta_1,...,\theta_k$ if the probability function of Y is given by

$$P(Y=y) = \sum_{i=1}^{k} \theta_i \, p_i \, (y) \quad \text{for } y \in G.$$

Here $\theta_1+...+\theta_k = 1$. In particular, consider a mixture of the strongly and weakly symmetrized versions of a random mapping on $V = \{1,...,n\}$ with mixing probabilities θ and $1-\theta$. This is a random undirected graph consisting of connected components that are trees or single loops or cycles with or without trees attached to them. A large value of θ implies that isolated loops and edges are common, while a small value of θ implies that trees and attached trees are more frequent. A typical realization of such a mixture is shown in Figure 1.

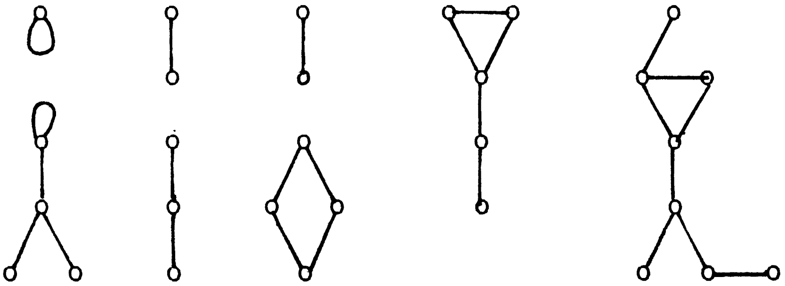


Figure 1. A typical realization of a mixture of the strongly and weakly symmetrized versions of a random mapping.

Identifiability problems for general mixtures are discussed by Titterington, Smith and Makov (1985). Frank (1986b) discusses conditions on the parameters that guarantee the identifiability of a mixture of Bernoulli graphs.

3. RANDOMLY COLORED MULTIGRAPHS

Simple graph structures are not rich enough for many statistical applications involving relational data. For instance, a relationship like kinship can be further specified as sibling, parent, etc. A relationship like similarity can be further specified by some quantitative measure of the degree of similarity or by some separation of different aspects of similarity. Communication networks may need to have capacities or distances attached to the edges. Sociograms and other contact patterns are often more interesting if some information about the individuals is also available.

Consider several different relationships to be studied simultaneously and in conjunction with observations on both individual and relational variables. One or more individual attributes can be combined into a vertex variable, and one or more relational attributes can be combined into an edge variable. It is convenient to refer to the outcomes of these variables as colors and speak of vertex and edge colorations. Only finitely many colors are considered here. Symmetric and asymmetric relationships are distinguished so that the general structure is that of a colored multigraph consisting of a complete undirected graph $\binom{V}{2}$ and a complete directed graph $V^{(2)}$ defined on a common vertex set $V = \{1,...,n\}$. For convenience, directed edges are now called arcs. There are three random colorations : a vertex coloration that is a function X from V to A, an edge coloration that is a function Y from $\binom{V}{2}$ to B, and an arc coloration that is a function Z from $V^{(2)}$ to C. Here A,B and C are color sets of a,b and c colors, respectively. Consequently, the set G of all possible colored multigraphs contains

$$|G| = a^n \, b^{\binom{n}{2}} \, c^{n(n-1)}$$

outcomes of the randomly colored multigraph (X,Y,Z) on V. Three examples of sets G are shown in Figures 2-4. Figure 2 shows the graphs of order n=3 with a=1 vertex color, b=3 edge colors (solid, dotted, nothing), and c=1 arc color (nothing). Figure 3 shows the graphs of order n=3 with a=1, b=4 corresponding to the possible combinations of solid and dotted edges, and c=1. Figure 4 shows the graphs of order n=3 with a=2, b=2, and c=1. The number at each graph is the number of isomorphisms.

The probability function of (X,Y,Z) is denoted by

$$P((X,Y,Z) = (x,y,z)) = p(x,y,z)$$

where $x = (x_1,...,x_n)$, $y = (y_{ij} : 1 \leq i < j \leq n)$, $z = (z_{ij} : i \neq j)$ are elements of A^n, $B^{\binom{n}{2}}$, and $C^{n(n-1)}$, respectively. If the probability distribution is invariant under isomorphism, and if there are m isomorphism classes G_i, then there are probabilities p_i such that $p_1+...+p_m = 1$ and

$$p(x,y,z) = p_i / |G_i| \quad \text{for} \quad (x,y,z) \in G_i, \quad i=1,...,m.$$

Distributions that are invariant under isomorphism are of particular interest for modelling phenomena that depend not on the identities of the vertices but only on their colors. The combinatorial problem of counting the number m of non-isomorphic graphs can be solved by an application of Burnside's lemma (see Frank, 1986a) and for n=2 and n=3 the following formulae apply :

$$m = (a^2 bc^2 + abc)/2 = \binom{ac+1}{2} \, b,$$

$$m = (a^3 b^3 c^6 + 3a^2 b^2 c^3 + 2abc^2) / 6 = \binom{abc^2+2}{3} - a^2b^2c^2 \binom{c}{2} \cdot$$

In particular, the last formula checks with the number of graphs in Figures 2-4.

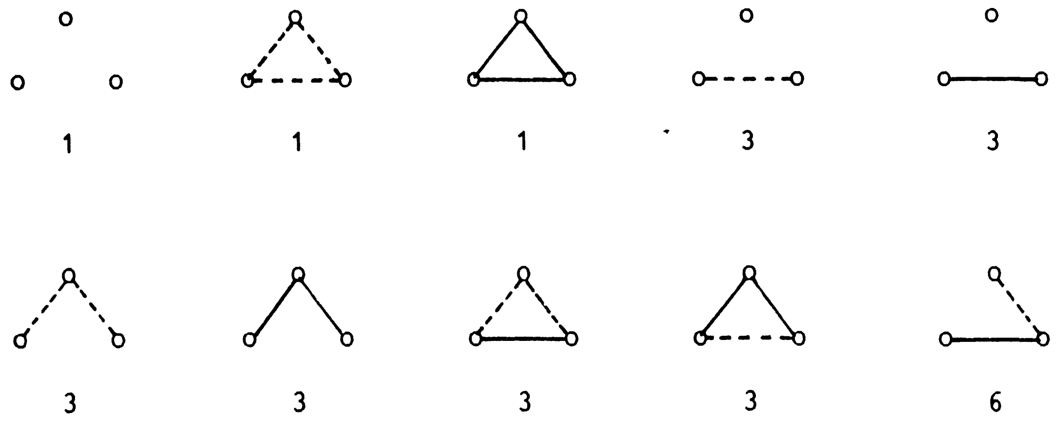


Figure 2. Graphs of order 3 with three edge colors.

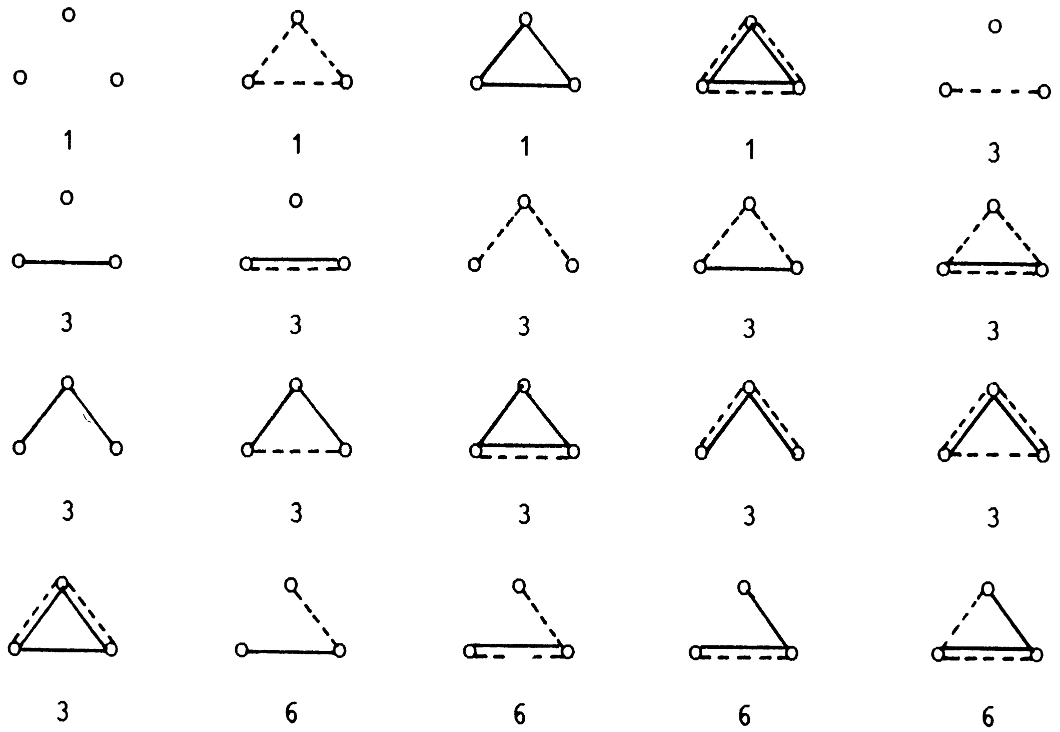


Figure 3. Graphs of order 3 with four edge colors.

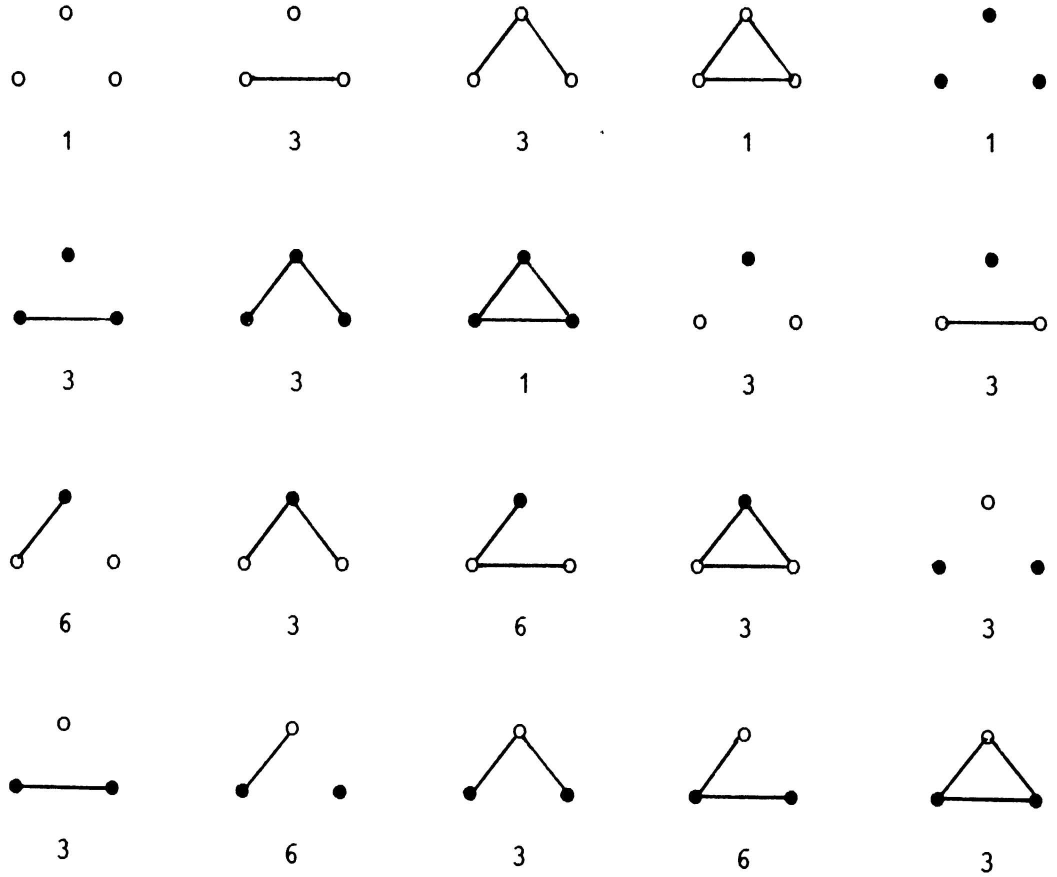


Figure 4. Graphs of order 3 with two vertex colors.

Assume now that the probability distribution of (X,Y,Z) is invariant under isomorphism. Let the colors be labeled by integers so that $A = \{1,\ldots,a\}$, $B = \{1,\ldots,b\}$ and $C = \{1,\ldots,c\}$. Set N_i equal to the number of vertices $v \in V$ having $X_v = i$, set R_{ijk} equal to the number of edges $\{u,v\} \in \binom{V}{2}$ having $X_u = i$, $X_v = j$, $Y_{uv} = k$ (where $Y_{uv} = Y_{vu}$ by definition), and set S_{ijkl} equal to the number of arcs $(u,v) \in V^{(2)}$ having $X_u = i$, $X_v = j$, $Z_{uv} = k$, $Z_{vu} = l$. The vertex color frequencies $(N_1,\ldots,N_a)$ are summary statistics of the *composition* of the graph, and the edge and arc color frequencies $(R_{ijk} : k = 1,\ldots,b)$ and $(S_{ijkl} : k = 1,\ldots,c \,;\, l = 1,\ldots c)$ between and within different vertex colors $1 \leq i \leq j \leq a$ are summary statistics of the *structure* of the graph conditional on the composition. Here

$$\sum_{i=1}^{a} N_i = n\,,\quad \sum_{k=1}^{b} R_{ijk} = N_{ij}\,,\quad \sum_{k=1}^{c}\sum_{l=1}^{c} S_{ijkl} = N_{ij}$$

where $N_{ii} = \binom{N_i}{2}$ and $N_{ij} = N_i N_j$ for $i \neq j$. The composition and structure summary statistics are sufficient statistics if the edge and arc colors (Y_{uv}, Z_{uv}, Z_{vu}) are independent for distinct pairs $\{u,v\} \in \binom{V}{2}$ conditional on the vertex colors $(X_1,\ldots,X_n)$. An example of such a model is described in the next section.

It is possible to classify general random graph models by distinguishing between random and non-random composition and structure. Table 2 illustrates this typology. If both the composition and the structure are non-random, the graph is deterministic and not of concern here. If the composition is random but the structure is non-random, the graph can be considered as a model for a random process on the vertices of a fixed graph. Random percolation processes on a lattice medium and random adhesion processes on the surface of a polyhedron are of this type ; Markov chains, Ising models and more general Markov fields are other examples. Such models are typically concerned with some probabilistic process going on in some fixed environment having a specific neighborhood structure.

If the composition is non-random but the structure is random, the graph can be a classical uniform random graph or a Bernoulli graph. Also more elaborate statistical graph models like the ones in the next section belong to this type as long as there are no vertex colors or only deterministic vertex colors. Such models are typically concerned with probabilistic interrelationships or interactions between the units in a fixed set.

If there is some kind of probabilistic interdependence between the development of a process and the characteristics of the environment in which it arises, we need a model in which both composition and structure are random. Several of the models in the next section are of this type.

	Fixed structure	Random structure
Fixed composition	Deterministic models	Uniform random graphs Bernoulli graphs Dyad independence models Markov graphs
Random composition	Markov chains Ising models Markov fields	Conditional dyad independence models Randomly colored multigraphs

Table 2. Classification of random graphs according to composition and structure.

4. MULTIPARAMETRIC MODELS

Various attempts have been made to find a multiparametric model for social network data that is of sufficient generality to fit reasonably well in many different contexts ; see, for instance, Burt (1982) and Knoke and Kuklinski (1982). An implicit or explicit constraint in some methodological papers seems to be that the model should also be easy to handle, preferably via standard packages for linear statistical models or other easily accessible programs. This double goal of generality and readiness is good if it brings empirical findings and feedback from widespread applications. Otherwise it may just be too limiting and more of an obstacle to further development of specific models.

This section surveys some of the approaches in the literature which are mainly exploratory in purpose and aim at providing general data analytic tools for network data. A few more restricted models are also mentioned.

Holland and Leinhardt (1981) investigate a class of probability distributions for networks that belongs to the exponential family and contains parameters that can be interpreted as contact intensity, individual attraction, individual activity, and mutual contact intensity. The basic assumption is that contacts are independent between different pairs of individuals. More specifically, let Y be a random graph in the set G of directed graphs of order n having no loops or multiple edges. The vertex set is $V = \{1,...,n\}$, and Y is considered either as a random subset of $V^{(2)}$ or as a random adjacency matrix of edge indicators Y_{uv} for $(u,v) \in V^{(2)}$. The $\binom{n}{2}$ random dyads (Y_{uv}, Y_{vu}) for $1 \leq u < v \leq n$ are independent, and the probability function of Y is accordingly given by a product

$$p(y) = \prod_{u<v} p_{uv}(y_{uv}, y_{vu}) .$$

Its logarithm is equal to

$$\log p(y) = \sum_{u<v} [(1-y_{uv})(1-y_{vu}) \log p_{uv}(0,0) + y_{uv} y_{vu} \log p_{uv}(1,1)] +$$
$$+ \sum_{u \neq v} y_{uv}(1-y_{vu}) \log p_{uv}(1,0) .$$

Here $p_{uv}(1,0) = p_{vu}(0,1)$ for $u > v$. The most general model of this type would require $3 \binom{n}{2}$ parameters, but the Holland-Leinhardt model reduces this number to 2n by assuming

$$\log p_{uv}(0,0) = \lambda_{uv}$$
$$\log p_{uv}(1,0) = \lambda_{uv} + \alpha_u + \beta_v$$
$$\log p_{uv}(1,1) = \lambda_{uv} + \alpha_u + \beta_u + \alpha_v + \beta_v + \gamma$$

where $\sum \alpha_v = \sum \beta_v$, and λ_{uv} is a normalizing constant.

Holland and Leinhardt (1981) and Fienberg, Meyer and Wasserman (1985) extend the previous model of pure structure to situations in which composition data are also available ; that is, when the vertices are colored to distinguish between various categories of vertices. Let X_v be the color of vertex v and Y_{uv} an indicator of edge (u,v), as before. The probability function of the graph (X,Y) having composition $X = (X_1,...,X_n)$ and structure $Y = (Y_{uv} : u \neq v)$ can be given as

$$p(x,y) = \prod_{v=1}^{n} p_v(x_v) \prod_{u<v} P(y_{uv}, y_{vu} \mid x_u, x_v)$$

if it is assumed that the vertices are colored independently and that, conditional on the vertex colors, the dyads have probabilities that depend only on the colors of the two vertices in the pair, and dyads are independent for different pairs of vertices. Now, the number of parameters can be restricted by suitable parameterization so that it increases with the number of colors only and not with the number of vertices. This is a convenient way to avoid the complications due to increasing degrees of freedom that make it difficult to evaluate asymptotic distributions for large graph orders n.

White, Boorman and Breiger (1976), Arabie, Boorman and Levitt (1978) and others have analyzed block models that aim at an integrated representation of composition and structure. The vertices can be colored to distinguish between different blocks of vertices. The block-modelling approach is combinatorial and not probabilistic, and its purpose is to find a block composition that in a combinatorial sense gives an optimal structure between and within the blocks. Stochastic block-modelling is similar to the modelling of randomly colored graphs. An early reference is Holland, Laskey and Leinhardt (1983) and a more recent one is Wang and Wong (1987).

The approach of Frank et al. (1986) and Wellman et al. (1988) considers colored vertices and colored edges and allows the order of the graph to be random. The typical application is to an individual's social support network consisting of intimate kins and friends and some kinds of interrelationships between them. In order to include various attributes of the individuals, relationships, and networks in a study, we need models of the interdependence between composition and structure. The order of the network is taken as a truncated Poisson variable. Conditional on the order, the individual attributes are supposed to be independent for different members of the network. Conditional on the individual composition, the attributes of relationships are supposed to be independent for different dyads. This leads to a loglinear model that can be fitted by a stepwise testing procedure.

Another exploratory technique for investigating network data has been applied by Frank, Hallinan, and Nowicki (1985) and Frank, Komanska, and Widaman (1985). The main idea here is to use cluster analysis to reduce the number of dyad distributions between and within vertex categories. This typically leads to a substantial reduction of the initial number of parameters needed when all the dyad distributions are distinct. An advantage with this approach is that no assumptions are needed about the parametric form of the dyad distributions. The final model may contain a few quite general dyad distributions that refer to a coarser classification into vertex categories.

As an example of a more specific model of the interdependence between composition and structure of a network, let us now consider the following model investigated by Frank and Harary (1982). The vertices in $V = \{1,...,n\}$ are independently colored according to a probability distribution $p_1,...,p_r$ on r colors. An initial graph is formed by joining by edges all the vertices of the same color. Edges between vertices of the same color can be deleted with a probability α, and new edges between vertices of different colors can be inserted with a probability β. All the deletions and insertions are mutually independent and independent of the vertex coloration. The final graph of remaining initial edges and inserted new edges is observed. This graph Y can be described as the union of a graph that is Bernoulli $(X, 1-\alpha)$ and a graph that is Bernouilli $(\overline{X}, \beta)$ where X is the graph on V obtained by joining vertices of the same color and $\overline{X}$ is its complement.

This model can be considered appropriate for a set of n objects exhibiting a latent clustering into r clusters of similar objects. Similarity cannot, however, be observed without error. There is a probability α of observing a false dissimilarity and a probability β of observing a false similarity. If there is an unknown number r of equally likely clusters so that $p_i = 1/r$, then there are parameters r, α,β that must be estimated to fit this model to data. Frank and Harary (1982) discuss this and similar estimation problems. Further related material can also be found in Frank (1978a).

The models discussed so far have a common assumption of independence between dyads conditional on the vertex colors. Thus, in general there can be dependence between edge colors, but this dependence is then explained by the vertex colors. This means that structural dependence not explicable by the composition cannot be handled by these models. A kind of models governed by a more general structural dependence is the Markov graphs investigated by Frank (1985) and Frank and Strauss (1986). In a *Markov graph*, the colors of any pair of non-incident edges are independent conditional on the colors of all other edges, but the colors of a pair of incident edges could be conditionally dependent. A Markov graph with a probability function that is invariant under isomorphism is called a *homogeneous Markov graph*. The minimal sufficient statistics of a homogeneous Markov graph Y are given by the star and triangle counts, that is by the frequencies of subgraphs of Y that are isomorphic to various distinct stars and triangles. A *star* at center $v \in V$ is specified by the colors of all the vertices adjacent to v and of all the edges incident to v. A *triangle* at $\{u,v,w\} \in \binom{V}{3}$ is specified by the colors of these vertices and all edges joining them. Stars and triangles are called *sufficient subgraphs* of a Markov graph. Figure 5 shows the sufficient subgraphs of simple undirected Markov graphs of order 5 with two vertex colors and two edge colors. Inference problems for Markov graphs are not easy. Frank and Strauss (1986) have suggested a method of estimating Markov graph parameters which is based on simulations. The computational complexity involved seems to be prohibitive for extensive applications.

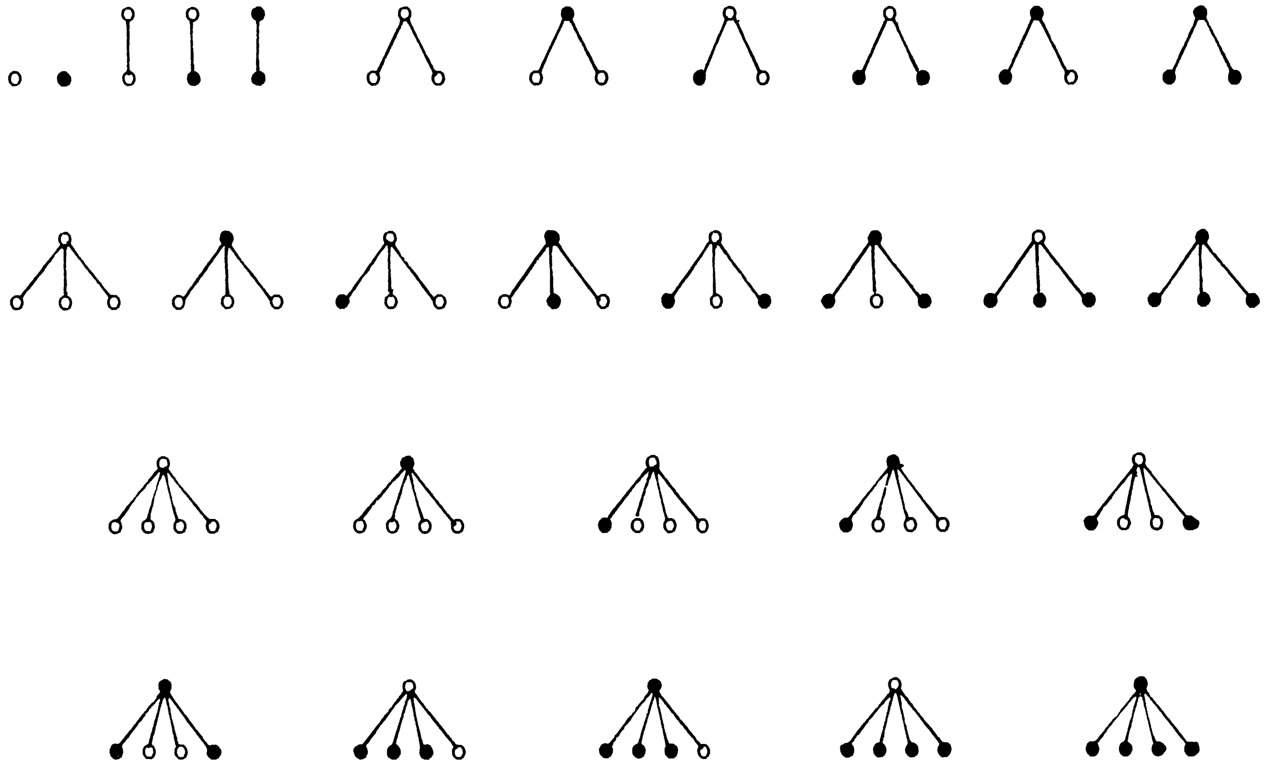


Figure 5. Sufficient subgraphs of undirected Markov graphs of order 5 with two vertex colors.

5. SUBGRAPH SAMPLING

A particular class of graph-sampling problems arises when a fixed population graph is investigated by sampling and observing only a part of it. This section describes various subgraph sampling designs that have been considered by Bloemena (1964), Capobianco (1970, 1974) and Frank (1969, 1977a,b,c, 1978b).

Let $V = \{1,...,N\}$ be a population of N units and $g = (x,y,z)$ a colored multigraph defined on V. Various population parameters of interest can be the color distributions of the vertices, edges, and arcs, the number of induced subgraphs of different kinds, etc. Sometimes the totals Σx_i, Σy_{ij}, Σz_{ij} can be interpreted as quantities of particular interest. When only a simple graph y is defined on V, its size and number of connected components are examples of population parameters that can be given as totals.

Let S be a subset of V selected according to a specific random sampling design having inclusion probabilities

$$P(i \in S) = \pi_i, \quad P(i \in S, j \in S) = \pi_{ij} .$$

In particular, simple random sampling of n units implies that $\pi_i = n/N$, $\pi_{ij} = n(n-1) / N(N-1)$ for $i \neq j$, and Bernoulli (p) sampling implies that $\pi_i = p$, $\pi_{ij} = p^2$ for $i \neq j$. If S is a vertex sample, *induced subgraph sampling* means that the subgraph of the population graph g induced by S is observed. Denote this subgraph by g(S). Thus, colors are observed of the vertices in the sample and of the edges and arcs between pairs of vertices in the sample. If g = y, then $g(S) = y \cap \binom{S}{2}$, and a total $T = \Sigma\, y_{ij}$ can be estimated by

$$\hat{T} = \Sigma y_{ij}\, S_i\, S_j / \pi_{ij} ,$$

where S_i indicates $i \in S$. This and other Horvitz-Thompson estimators for graph totals are investigated in Frank (1977a,b,c).

An alternative subgraph sampling procedure based on a vertex sample S is *star sampling*, in which colors are observed of vertices, edges, and arcs adjacent and incident to the sample. Star sampling based on S and induced subgraph sampling based on the complement $\bar{S} = V\text{-}S$ are complementary ; that is, $\bar{g}(\bar{S})$ denotes the star sample based on S. This relationship can be used to derive results for star sampling procedures from corresponding results for induced subgraph sampling procedures. Capobianco and Frank (1982) have compared different estimators based on these subgraph sampling procedures.

A generalization of star sampling is *snowball sampling* which is a successive extension of an initial vertex sample to its star, of the vertex set of this star to its star, etc. Goodman (1961) and Frank (1979) give further results on snowball sampling.

Other sampling procedures in graphs are based on edge sampling designs. For instance, *incident subgraph sampling* based on an edge sample $S \subseteq \binom{V}{2}$ means that the colors are observed of the vertices and edges incident to at least one edge in S. An application of this subgraph sampling procedure is the following. Let the vertices be people and the edges telephone calls during a specific period of time. Some calls are sampled, and for each sampled call the speakers are asked to report specific information about the people they have been calling and the calls they have made during the time period.

The subgraph sampling procedures discussed so far have referred to a fixed population graph. As in ordinary survey sampling, it is sometimes useful to introduce a superpopulation model by which the population graph is considered as sampled from some family (superpopulation) of population graphs. For instance, the introduction of a Bernoulli population graph makes it possible to use Bayesian methods for estimating the size of a population graph. The exploration of Bayesian and empirical Bayesian methods in graph sampling has hardly begun.

6. PROSPECTS

After the appearance of the pioneering papers by Erdös and Rényi (1959, 1960) random graph theory has developed rapidly. The review article by Karonski (1982) contains about 250 references, and the textbook by Bollobas (1985) has more than 750 references. Random graph evolution and limit theorems for random graphs continue to inspire much research and produce interesting results.

Computer science is one of the most important sources for new and interesting random graph problems. Random graphs also find interesting applications in theoretical chemistry and biology, and these fields are likely to have increasing influence on applied and theoretical graph theory in the future.

Families of random graphs or parametric graph distributions appropriate for statistical modelling of graph data are not yet available to any large extent. Important contributions to the development have mainly been initiated by applications in the social and behavioral sciences. Other areas in which statistical graph theory can be expected to find future applications are in pattern and image modelling. Spatial statistics and random field theory are expanding branches of statistics that may also be of importance for the development of statistical graph theory. See, for instance, Ripley (1981), Adler (1981), and Vanmarcke (1983).

Furthermore, the rapid development of discrete mathematics and combinatorics may also bring with it an increasing interest in combinatorial configurations other than graphs. Random hypergraphs, random permutations, random partitions, and random tesselations are a few examples of such objects that have appeared already. See, for instance, Berge (1973), Lovasz (1979), and Ahuja and Schachter (1983).

REFERENCES

ADLER, R., *The Geometry of Random Fields*, New York, Wiley, 1981.

AHUJA, N., and SCHACHTER, B., *Pattern Models*, New York, Wiley, 1983.

ARABIE, P., BOORMAN, S.A., and LEVITT, P.R., "Constructing block models : how and why", *Journal of Mathematical Psychology* 17, 1978, 21-63.

BERGE, C., *Graphs and Hypergraphs*, Amsterdam, North-Holland, 1973.

BLOEMENA, A.R., *Sampling from a Graph*, Mathematical Centre Tracts, Amsterdam, 1964.

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

BURT, R., *Toward a Structural Theory of Action*, New York, Academic Press, 1982.

CAPOBIANCO, M., "Statistical inference in finite populations having structure", *Transactions of the New York Academy of Sciences* 32, 1970, 401-413.

CAPOBIANCO, M., "Recent progress in stagraphics", *Annals of the New York Academy of Sciences*, 231, 1974, 139-141.

CAPOBIANCO, M., and FRANK, O., "Comparison of statistical graph-size estimators", *Journal of Statistical Planning and Inference* 6, 1982, 87-97.

ERDOS, P., and RENYI, A., "On random graphs I", *Publ. Math. Debrecen* 6, 1959, 290-297.

ERDOS, P., and RENYI, A., "On the evolution of random graphs", *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 1960, 17-61.

FIENBERG, S.E., MEYER, M.M., and WASSERMAN, S., "Statistical analysis of multiple sociometric relations", *Journal of the American Statistical Association*, 80, 1985, 51-67.

FRANK, O., "Structure inference and stochastic graphs", *FOA-Reports*, 3:2, 1969, 1-8.

FRANK, O., "Estimation of graph totals", *Scandinavian Journal of Statistics*, 4, 1977a, 81-89.

FRANK, O., "A note on Bernoulli sampling in graphs and Horvitz-Thompson estimations", *Scandinavian Journal of Statistics*, 4, 1977b, 178-180.

FRANK, O., "Survey sampling in graphs", *Journal of Statistical Planning and Inference*, 1, 1977c, 235-264.

FRANK, O., "Inferences concerning cluster structure", *Proceedings in Computational Statistics*, edited by L.C.A. Corsten and J. Hermans, Vienne, Physica-Verlag, 1978a, 259-265.

FRANK, O., "Estimation of the number of connected components in a graph by using a sampled subgraph", *Scandinavian Journal of Statistics*, 5, 1978b, 177-188.

FRANK, O., "Estimation of population totals by use of snowball samples", *Perspectives on Social Network Research*, edited by P. Holland and S. Leinhardt, New York, Academic Press, 1979, 319-347.

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

FRANK, O., "Triad count statistics", Department of Statistics, University of Stockholm, 1986a. To appear in *Discrete Mathematics*.

FRANK, O., "Random graph mixtures", Department of Statistics, University of Stockholm, 1986a. To appear in the *Annals of the New York Academy of Sciences*.

FRANK, O., HALLINAN, M., and NOWICKI, K., "Clustering of dyad distributions as a tool in network modelling", *Journal of Mathematical Sociology*, 11, 1985, 47-64.

FRANK, O., and HARARY, F., "Cluster inference by using transitivity indices in empirical graphs", *Journal of the American Statistical Association*, 77, 1982, 835-840.

FRANK, O., KOMANSKA, H., and WIDAMAN, K., "Cluster analysis of dyad distributions in networks", *Journal of Classification*, 2, 1985, 219-238.

FRANK, O., LUNDQUIST, S., WELLMAN, B., and WILSON, C., "Analysis of composition and structure of social networks", Department of Statistics, University of Stockholm, 1986.

FRANK, O., and STRAUSS, D., "Markov graphs", *Journal of the American Statistical Association*, 81, 1986, 832-842.

GOODMAN, L.A., "Snowball sampling", *Annals of Mathematical Statistics*, 32, 1961, 148-170.

HOLLAND, P.W., LASKEY, K.B., and LEINHARDT, S., "Stochastic blockmodels, First steps", *Social Networks*, 5, 1983, 109-138.

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

KARONSKI, M., "A review of random graphs", *Journal of Graph Theory*, 6, 1982, 349-389.

KNOKE, D., and KUKLINSKI, J.H., *Network Analysis*, Beverly Hills, Sage Publications, Ca., 1982.

LOVASZ, L., *Combinatorial Problems and Exercices*, Amsterdam, North-Holland, 1979.

PALMER, E., *Graphical Evolution*, New York, Wiley, 1985.

RIPLEY, B., *Spatial Statistics*, New York, Wiley, 1981.

TITTERINGTON, D.M., SMITH, A.F.M., and MAKOV, U.E., *Statistical Analysis of Finite Mixture Distributions*, New York, Wiley, 1985.

VANMARCKE, E., *Random Fields*, Cambridge, The MIT Press, 1983.

WANG, Y.J., and WONG, G.Y., "Stochastic blockmodels for directed graphs", *Journal of the American Statistical Association*, 82, 1987, 8-19.

WELLMAN, B., FRANK, O., ESPINOZA, V., LUNDQUIST, S., and WILSON, C., "Integrating individual, relational and structural analysis", Department of Sociology, University of Toronto, 1988.

WHITE, H.C., BOORMAN, S.A., and BREIGER, R.L., "Social structure from multiple networks", *American Journal of Sociology*, 81, 1976, 730-780.