Journal of Classification 15:199-223 (1998)

Metric Models for Random Graphs

David Banks

National Institute of Standards and Technology

G. M. Constantine

University of Pittsburgh

Abstract: Many problems entail the analysis of data that are independent and identically distributed random graphs. Useful inference requires flexible probability models for such random graphs; these models should have interpretable location and scale parameters, and support the establishment of confidence regions, maximum likelihood estimates, goodness-of-fit tests, Bayesian inference, and an appropriate analogue of linear model theory. Banks and Carley (1994) develop a simple probability model and sketch some analyses; this paper extends that work so that analysts are able to choose models that reflect application-specific metrics on the set of graphs. The strategy applies to graphs, directed graphs, hypergraphs, and trees, and often extends to objects in countable metric spaces.

Keywords: Bernoulli graphs; Clustering; Gibbs distributions; Holland-Leinhardt models; Phylogeny; Trees.

1. Introduction

There are many situations which give rise to graph-valued random objects. The following two examples offer a partial indication of the range of application and the different types of graph structures that may be encountered.

Constantine's work was supported in part by a Fulbright grant.

Authors' addresses: David Banks, Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA. Gregory M. Constantine, Departments of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA.

1. *Trees with Unlabeled Interior Nodes.* Molecular biologists often build a phylogenetic tree from amino acid discrepancies in a specific protein sampled from many species. When multiple proteins are used, one has a random sample of descent trees, each with labelled terminal nodes (the species) and unlabelled interior nodes. Strategies for estimating the central descent tree and placing a confidence region around it are given in Trang and Speed (1992) and Felsenstein (1985). The same problem arises in cluster analysis — some practitioners apply many different clustering algorithms, obtain many trees, and then seek to estimate the central tree (cf. Fowlkes and Mallows 1983).
2. *Graphs with Labelled Nodes and Undirected Edges.* Social network theorists (e.g., Krackhardt 1987, and Banks and Carley 1994) often analyze samples of friendship graphs. Each respondent reports a graph in which nodes represent the group members and edges indicate their perception of a friendship relation between the corresponding members. Since each respondent has only a noisy understanding of the relationships, the researcher wants to estimate the central graph and then determine a confidence region around it. This problem may also be modelled as a graph inspection problem with random error (cf. Constantine 1991).

These examples are indicative, but not exhaustive. Another common random object is a graph with directed edges (cf. Holland and Leinhardt 1981).

Early methods for analyzing random graphs (not trees) were pioneered by Moreno (1934), Festinger (1949), and Katz (1947, 1953, 1955). Statistical investigation remains active; the usual perspective has been based on log-linear models, following the work of Holland and Leinhardt (1981) and subsequent research by Wasserman (1987), Frank and Strauss (1986), Wong (1987), and Strauss and Ikeda (1990). Their work has focused on the study of a single random graph, in which edges are random outcomes whose distribution depends upon features of the nodes. But Bloemena (1964), Capobianco (1970), and Frank (1976, 1988) treat problems that specifically involve random samples of graphs, which matches the interest of this paper.

Strauss and Freeman (1989) review stochastic models for graphs, and Fienberg, Meyer, and Wasserman (1985) survey statistical methods in social network analysis. In parallel, sociological literature focused on metrics for posets (cf. Boorman and Olivier 1973). Since some posets can be represented as binary trees, this work bears on the analysis of random trees that is central to this paper. Barthélemy, Leclerc, and Monjardet (1986) and Day (1986) offer general reviews of these methods.

In contrast, our research uses a metric to define distributions on sets of graphs, and thus automatically appropriates a well-studied arsenal of statistical techniques. This enables one to obtain maximum likelihood estimates of central tendency and spread, perform hypothesis tests, assess fit, and even develop an analogue of the linear model. The application of these methods is illustrated through the analysis of data examined by Fowlkes and Mallows (1983).

2. The Model

Let G be a finite set of graphs with elements g; depending on the application, these might be trees, directed graphs, networks, or other similar objects. Let $\boldsymbol{R}^+$ be the nonnegative reals, and denote by d: $G \times G \rightarrow \boldsymbol{R}^+$ an arbitrary metric on G. Given d, we mimic Mallows's method (1957) of setting probabilities on the set of permutations. For graphs, Mallows's approach yields the probability measure $H(g^*,\tau)$, defined by

$$P_{(g^*,\tau)}[g] = c(g^*,\tau)e^{-\tau d(g,g^*)} \quad \forall\, g \in G\,, \tag{1}$$

where $g^* \in G$ is the modal or central graph, τ is a concentration parameter, and $c(g^*,\tau)$ is a normalizing constant. Thus g^* and τ are analogous to the mean and precision of a normal distribution, respectively, and index the family of probability distributions $\{H(g^*,\tau)\}$. If $\tau = 0$, one has uniform measure on G, but $\tau >> 0$ implies concentration about g^*. These parameters index the family of probability distributions $\{H(g^*,\tau)\}$.

Other measures for random graphs have been proposed. Mathematicians use three basic families (cf. Bollobás 1985, Ch. 2), but the models are insufficiently rich for statistical applications. The chief deficiency is the lack of a location-scale interpretation; $\{H(g^*,\tau)\}$ automatically avoids this limitation.

One may characterize $\{H(g^*,\tau)\}$ according to entropy. In accordance with information-theoretic practice, define

$$e(p) = -\sum_{g \in G} p(g) \ln p(g)$$

as the entropy of the probability distribution that places probability $p(g)$ on $g \in G$. The distribution which maximizes $e(p)$ (the Gibbs distribution; see Geman and Geman 1975) provides the greatest sampling diversity.

Proposition 1: *Distinguish an element g^* of G. The probability distribution that maximizes the entropy $e(p)$ subject to the constraint that*

$$\sum_{g \in G} d(g, g^*) p(g) = v \tag{2}$$

satisfies $p(g) = c(g^*, \tau) e^{-\tau d(g,g^*)}$, *where*

$$c(g^*, \tau)^{-1} = \sum_{g \in G} e^{-\tau d(g,g^*)} ,$$

and τ *is the unique solution to*

$$\frac{d \ln c(g^*, \tau)}{d\tau} = v . \tag{3}$$

(A proof is in the Appendix.)

Proposition 1 implies that $\{H(g^*, \tau)\}$ is the maximum-entropy family (Gibbs family) over G under statistically natural constraints. Note that the constraint determines the dispersion of the distribution around the "central" element g^*, so that specifying v is equivalent to specifying τ. For example, $\tau = 0$ if and only if v is the arithmetic average of the values $d(g,g^*)$; this gives uniform measure on G. Bernardo and Smith (1994, 207-209) discuss data modeling through distributions that maximize entropy under moment constraints.

Alternative characterizations are possible in two special cases important to social network theory. Let G_m be the set of graphs on m distinct nodes with undirected (untagged, unweighted) edges and no loops. Frank and Strauss (1986) used the Hammersley-Clifford Theorem (Hammersley and Clifford 1971; see also Strauss 1983) to show that all probability measures on G_m can be written in the form

$$P_D [g] = c \exp \left[\sum_{A \subseteq g} \alpha_A \right] \quad \forall g \in G_m \tag{4}$$

where c is a normalizing constant and α_A is a nonzero constant if and only if A is a clique of the nonrandom dependence graph D. In Frank and Strauss's context, the vertices of D are all possible edges on the m nodes of g; a clique in D is a subset of the vertex set of D that is either a singleton set or has the property that all pairs of elements are connected by edges in D. The graph D determines the dependence structure between random edges in g; if D connects a specific pair of edges, then those edges in g are conditionally dependent given the other edges in g. Note that the models described by (4) strictly include those described in (1), because (1) does not allow bimodal distributions.

When the dependence graph D is edgeless, one has the Bernoulli graph model, in which the presence or absence of each edge is an independent Bernoulli trial (with possibly different probabilities). Bollobás (1985, Ch. 2)

describes the model, and it has been used by Banks and Carley (1994) to analyze friendship networks. To show that the Bernoulli graph model is a special case of $\{H(g^*,\tau)\}$, let $I\{e_{ij}: g_1,g_2\}$ be an indicator function taking the value 1 when the edge e_{ij} linking nodes i and j is present in exactly one of the graphs g_1 and g_2, but is otherwise zero. Thus the indicator function notices discrepancies between the edges in the two graphs. The Frank and Strauss representation enables the following result.

Proposition 2: *A probability measure on G_m is a Bernoulli graph model if and only if it can be written as (1) for a semimetric of the form*

$$d_B(g_1,g_2) = \sum_{i<j} a_{ij} I\{e_{ij}: g_1,g_2\} \tag{5}$$

where the $a_{ij} \geq 0$. (A proof is in the Appendix.)

As is apparent from the proof of Proposition 2, the a_{ij} are the log odds ratios of the edge probabilities, where the numerator of the ratio contains the least probable outcome. When $a_{ij} = 1$ for all i,j, this reduces to the Hamming metric (1950). The assumption of independent edges that characterizes this model is equivalent to a concept of distance based upon a weighted sum of edge discrepancies. This concept contrasts with metrics that imply edge dependencies, which may be more appropriate for certain datasets.

A similar result holds for the p_1 model proposed by Holland and Leinhardt (1981), which extends the Bernoulli graph model to the set G_m^* of loopless graphs with directed edges. The p_1 model imposes additional structure on the α_A terms; edges in the p_1 model are still independent, but with probabilities that depend on node characteristics.

Proposition 3: *A probability measure on G_m^* is a Holland-Leinhardt p_1 model if and only if it can be written as (1) for a metric of the form*

$$d_{HL}(g_1,g_2) = \sum_{i<j} | b_{ij}(g_1) - b_{ij}(g_2) | , \tag{6}$$

where the b_{ij} are specified in the proof in the Appendix, and the sum is over pairs of nodes.

These results provide some insight into the relationship between the models given by (1) and models previously proposed, and show that the Bernoulli graph model and the p_1 model are maximum entropy distributions. However, from a practical standpoint, one also wants models that are computable and interpretable. The next section addresses these points.

3. A Simple Case: Random Graphs

For arbitrary metrics d and general sets of graphs, applying the model in (1) requires numerical methods. But some choices reduce computation, and we use such a case to illustrate a range of inferential tools. This section considers the Hamming metric d_H on the set G_m, consisting of graphs on m labelled nodes having undirected edges and no loops. It is a prelude to Section 4, which treats trees.

Hamming (1950) introduced a metric that counts the number of edge discrepancies between two graphs. It has a geometric interpretation because the elements of G_m can be viewed as the vertices in an $r = \binom{m}{2}$ dimensional unit hypercube. A particular vertex with a given binary sequence indicates the graph with edges determined by the ones in the sequence and non-edges determined by the zeroes in the sequence. Vertices (graphs) that are adjacent to g are a single edge change from g. The distance between graphs is just the Hamming distance between the sequences of zeroes and ones that identify the corresponding hypercube vertices, or the shortest path in the hypercube between those vertices.

The symmetry of this hypercube geometry ensures that the normalizing constant cannot depend on g^*. In this case, the underlying binomial structure implies:

$$c(g^*,\tau)^{-1} = \sum_{g \in G_m} e^{-\tau d_H(g,g^*)} = \sum_{k=0}^{r} \binom{r}{k} e^{-\tau k} = (1 + e^{-\tau})^r . \qquad (7)$$

For other metrics, it is often still possible to find representations analogous to the hypercube geometry, so that the distance between objects is defined as the minimum path metric associated with a graph whose vertices are the objects and whose edges correspond to an elementary transformation between objects. If that representation has the property that structure appears the same from each vertex's perspective, then the normalizing constant is independent of g^*.

The maximum likelihood estimate of g^* based on the random sample $g_1, \ldots, g_n$ is

$$\hat{g}^* = \text{argmin}_{g \in G_m} \sum_{i=1}^{n} d_H(g_i,g) , \qquad (8)$$

where the argmin function gives the value in G_m that minimizes the argument. Barthélemy and Monjardet (1988) refer to the quantity $\Sigma\, d_H(g_i,g^*)$ as the remoteness function, and the solutions of (8) as medians. A straightforward argument shows the maximum likelihood estimate, or median, is found by majority rule; i.e., an edge is in $\hat{g}^*$ if and only if it is present in more than

half of the sample graphs (non-uniqueness may arise when n is even). Given $\hat{g}^*$, the maximum likelihood estimate of τ follows from differentiation; i.e.:

$$\hat{\tau} = -\ln \frac{(rn)^{-1}\Sigma_{i=1}^{n} d_H(g_i,\hat{g}^*)}{1-(rn)^{-1}\Sigma_{i=1}^{n} d_H(g_i,\hat{g}^*)}. \tag{9}$$

Thus we have estimates of the central graph and the concentration parameter.

After estimation, one wants to assess goodness-of-fit. The discreteness of the parameter space precludes the usual tests. Instead, we recommend a test in two steps: the first checks that the observed number of sample graphs at distance k from $\hat{g}^*$ conforms with the model, and the second checks that the sample is symmetrically distributed around $\hat{g}^*$.

1. The k-th orbit around g^* is the set of graphs that are distance k from the central graph g^*. From binomial probability, the proportion of sample observations expected in the k-th orbit is $\binom{r}{k}(1+e^{-\tau})^{-r}e^{-k\tau}$. A binomial plot, or a goodness-of-fit test, can discover deviations from this model. Of course, this procedure is not exact, since one must estimate τ and g^* from the data.
2. Let $\mathbf{X}_i$ be an r-component vector of zeroes and ones such that the ones denote which edges are discrepant between g_i and g^*; similarly, let $s_i = d(g_i,g^*)$. Then $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are independent, and under the hypothesized model, the conditional distribution of $\mathbf{X}_i$ given s_i is uniform on the set of vectors with exactly s_i non-zero components. Thus one can simulate the null distribution of $\mathbf{Y} = \Sigma\,\mathbf{X}_i$. If some sensible function of this statistic, such as $\max_j \{Y_j\}_{j=1}^{r}$, is significantly improbable when compared to its simulated value, this is evidence against the adequacy of the fit. Again, this is an approximate test, since one must estimate g^* from the data.

Using two tests, each of which assesses a qualitatively different aspect of fit, helps diagnose the kind of model failure that may occur. For example, in the social network context, conformity pressure may appear as underdispersion in the first test, whereas subgroups are signaled by large maxima in the second.

These two tests are complementary, in a sense analagous to the independence of the radius and angle when describing the location of a point in the plane. The first test summarizes the sample by its orbits; such classification is done entirely by "radius." The second test classifies the sample by "angle"; here it is the direction from the central graph, not the distance, that is used. Under the null hypothesis that model (1) holds, significance probabilities from these two tests are essentially independent (a minor dependence introduced by estimation of g^* diminishes with n). Thus one can combine P-values from these tests into a single assessment of fit

superior to that proposed in Banks and Carley (1994).

To set a confidence region for g^*, one can use the fact that $\Sigma_{i=1}^{n} d_H(g_i,g^*)$ has binomial distribution Bin(nr,θ) with $\theta = e^{-\tau}/(1+e^{-\tau})$. Then, for the maximum likelihood estimate $\hat{\theta}$ obtained from $\hat{\tau}$, one finds k^* such that

$$1-\alpha \approx \sum_{j=0}^{k^*} \binom{nr}{j} \theta^j (1-\theta)^{nr-j}$$

and sets the approximate $100(1-\alpha)\%$ confidence region on $\hat{g}^*$ having the form

$$\{g \in G_m: \sum_{i=1}^{n} d_H(g_i,g) \le k^*\}.$$

One could slightly improve the accuracy of the nominal confidence level by taking account of the uncertainty in $\hat{\theta}$, but this requires simulation. The confidence region is not symmetric around $\hat{g}^*$, as are those obtained from the most practicable bootstrapping procedure recommended in Banks and Carley (1994); rather, it is pulled towards the sample graphs, enabling smaller confidence regions.

Bayesian inference is possible if one has a joint prior over g^* and τ. A natural choice is formed as the product of the uniform prior on g^* and the exponential prior on τ with parameter λ. Then the joint posterior is

$$\pi_1(g^*,\tau) = k(1+e^{-\tau})^{-m} e^{-\tau(\lambda+\Sigma_{i=1}^{n} d_H(g_i,g^*))}$$

where

$$k^{-1} = \sum_{g \in G_m} \int_0^\infty (1+e^{-\eta})^{-m} e^{-\eta(\lambda+\Sigma_{i=1}^{n} d_H(g_i,g))}\, d\eta .$$

The solution is clearly numerical, as is usual in modern Bayesian analysis. If the cardinality of G_m is large, one must resort to approximations.

Last, we give an analogue of the linear model decomposition for graph-valued data. For (G_m,d_H), this might arise if one asked first graders to report their perception of the class's friendship network. Boys would know best about friendship patterns among boys, and similarly for girls. The researcher can decompose an individual's response into a common knowledge graph and sex-specific knowledge graph according to the following model:

$$g_i^s = (g^\mu \oplus g^s) \otimes g_i^\varepsilon \qquad (10)$$

where g_i^s is the graph generated by student i with sex s, g^μ is the common knowledge graph, g^s is the graph common to members of sex s, and g_i^ε is a random error graph that describes how the i-th student deviates from the

expected graph. Here $\oplus$ denotes the operation of edge union, and $\otimes$ denotes the operation of addition modulo 2 on the entries of the adjacency matrices. Thus a child's graph is the union of the common knowledge and sex-specific graphs, corrupted by random error. Assuming that g_i^{ε} is distributed according to (1) with g^* the edgeless graph, then the previous discussion enables maximum likelihood estimation and hypothesis testing regarding g^{μ}, g^{s}, and the dispersion of the error term g_i^{ε}.

4. A Harder Case: Random Trees

One would like to use the methods of Section 3 for trees, rather than the set G_m, and to employ more realistic metrics than d_H. Let $\tilde{G}_m$ denote the set of binary (or phylogenetic) trees with m distinct terminal nodes. A graph is a binary tree if it is a tree, has m labeled terminal nodes, one labeled root, and all interior nodes are unlabeled with degree 3 (some authors call these "binary trees"). When there is a label-preserving isomorphism between two trees, they are considered identical; thus if two terminal nodes are dependent from the same internal node, it does not matter which is on the left and which is on the right.

To analyze a random sample of trees $g_1, \ldots, g_n$ from $\tilde{G}_m$, assume a model of the form (1), with a suitably chosen metric d. Using $c(g^*,\tau) = [\Sigma_g \exp(-\tau d(g,g^*))]^{-1}$, one finds the log-likelihood function as

$$\ln L(g^*,\tau) = -n \ln \Big[\sum_{g \in \tilde{G}_m} e^{-\tau d(g,g^*)} \Big] - \tau \sum_{i=1}^{n} d(g_i,g^*) . \tag{11}$$

We seek the value $(\hat{g}^*,\hat{\tau})$ that maximizes (11). Qualitatively, the behavior of this function depends upon the value of τ.

For large values of τ (the typical application, with substantial concentration of probability around g^*) the second term in (11) is dominant. Thus an approximate method of moments estimator for g^* is the sample median g^{**}, where

$$g^{**} = \operatorname{argmin}_{g \in \tilde{G}_m} \sum_{i=1}^{n} d(g_i,g) ,$$

but this can be difficult to calculate. One could examine elements in $\tilde{G}_m$ that are near to g^{**}, and it is unlikely that one need search far to find the global maximum.

For small τ, there is limited applicability to the methods pursued in this paper. As $\tau \to 0$, the data become less relevant to the problem, and (11) reflects this fact in that the first term dominates the second. Examination of the term $-n \ln [\Sigma_g \exp(-\tau d(g,g^*))]$ shows that it tends to select an estimate

g^{***} of g^{*} that makes the sum of the distances $d(g,g^{***})$ as large as possible. Any search for a good initial guess of g^{***} should exploit the asymmetries in the metric space, as indicated by a matrix of inter-tree distances.

Solving (11) shows that the maximum likelihood estimates must satisfy:

$$\frac{1}{n}\sum_{i=1}^{n} d(g_i,\hat{g}^{*}) = \frac{\Sigma_{g\in\tilde{G}_m} d(g,\hat{g}^{*}) e^{-\hat{\tau}d(g,\hat{g}^{*})}}{\Sigma_{g\in\tilde{G}_m} e^{-\hat{\tau}d(g,\hat{g}^{*})}}, \tag{12}$$

$$\text{where } \hat{g}^{*} = \text{argmin}_{g^{*}\in\tilde{G}_m} n \ln \sum_{g\in\tilde{G}_m} e^{-\hat{\tau}d(g,g^{*})} + \hat{\tau}\sum_{i=1}^{n} d(g_i,g^{*}).$$

Solution of (12) requires enumeration of the elements of $\tilde{G}_m$. In view of the following result, solution is difficult for even moderate m.

Proposition 4. (Schröder, 1870.) *There are* $(1)(3)(5)\cdots(2m-3)$ *distinct trees in* $\tilde{G}_m$, *where* $m \geq 3$. *(A proof is in the Appendix.)*

Realistically, solution of (12) cannot attempt the sum over all elements of $\tilde{G}_m$. This problem can be avoided by truncating the sum (whose terms rapidly become small as $d(g,g^{*})$ increases) at some convenient point on any interim evaluation, and letting the truncation point grow as one converges toward the solution.

Also, because (11) is the sum of n unimodal log-likelihoods, all having the same form, then the number of local maxima in the likelihood surface cannot be greater than n. Following the logic outlined in an analogous calculation by Reeds (1985) for the estimation of the location parameter with an i.i.d. Cauchy sample, each local maximum must occur near at least one of the observations. This fact suggests a maximization strategy that starts n steepest ascent searches, each taking a sample point as the initial guess at the central tree. One would alternate estimation between the location and dispersion parameters, first treating the current estimate of τ as fixed in order to estimate g^{*}, and then treating the current estimate of g^{*} as fixed in order to estimate τ.

In practice one always wants to assess the fit of the model. A simple method for the $\{H(g^{*},\tau)\}$ model uses the Pearson chi-squared test. But the parameter space $\tilde{G}_m$ is discrete, and thus conventional asymptotic theory is unworkable — Fienberg, Meyer, and Wasserman (1985) identify this problem as a key area for research. This discreteness prevented Frank and Strauss (1986) from examining the fit of their Markov graph model, and for the p_1 model's fit, Holland and Leinhardt (1981) had to rely upon an ad hoc test based on triad counts (which found their example data had very bad fit).

The chi-squared test bins the data, then finds the expected counts in each bin under the $\{H(g^*,\tau)\}$ model with maximum likelihood estimates for g^* and τ. Banks and Carley (1994) give details for the case in the previous section, and the method extends directly, but with greater computation, to more complex cases. Banks and Carley recommend referring the test statistic to a chi-squared distribution with two fewer degrees of freedom than the number of bins. This subtraction accounts for constraints imposed by the estimation of τ and the fact that probabilities must sum to one, but is conservative in that it does not reduce the degrees of freedom for estimation of the problematic discrete parameter g^*.

To set confidence regions, we recommend the use of the bootstrap. Details for the case in the previous section are laid out by Banks and Carley (1994); the approach generalizes to harder cases. The next section demonstrates this extension for tree-valued data using regions of the form $\{g \in \tilde{G}_m: d(g,\hat{g}^*) \leq \delta\}$, where δ is determined by bootstrap resampling and the desired coverage probability. Also, the duality between confidence regions and hypothesis tests enables bootstrap methodology to address testing problems. Fisher and Hall (1990) describe the appropriate strategy for inverting bootstrap confidence regions in order to perform statistical tests. Constantine (1991) shows that the choice of metric can strongly affect the power function.

5. Example: Consensus Among Binary Trees

Tree problems are more difficult than the random graphs considered in Section 3 because there is no reasonable metric that imposes a neighborhood structure which is the same for all trees. Previously, the hypercube representation ensured that the number of graphs a fixed distance from the central graph did not depend upon the central graph, but this fails for trees. Consequently, the normalizing constant is a function of both the central parameter and the dispersion, making analytical maximization of the likelihood function intractable.

The large size of $\tilde{G}_m$ poses enumerational difficulties, but there are theoretical and computational devices that extend the range of practical solutions beyond that possible from direct enumeration. To illustrate these, we examine a sample of nine binary trees built from various clustering algorithms using letter frequency/authorship data reported in Fowlkes and Mallows (1983). We chose this dataset because (a) it was used in the major previous work on consensus for binary trees in the statistics literature, (b) the consensus problem is pertinent to both phylogeny and cluster analysis, and (c) results from slightly different metrics point up an interesting contrast between these two application domains.

Table 1

Books and Authors Used to Produce the Random Binary Trees

1.	*The Three Daughters of Madame Liang*	Pearl S. Buck
2.	*The Drifters*	James Michener
3.	*The Lost Worlds of 2001*	Arthur C. Clarke
4.	*East Wind, West Wind*	Pearl S. Buck
5.	*A Farewell to Arms*	Ernest Hemmingway
6.	*The Sound and the Fury* (Part I)	Willam Faulkner
7.	*The Sound and the Fury* (Part II)	William Faulkner
8.	*Profiles of the Future*	Arthur C. Clarke
9.	*Islands in the Stream*	Ernest Hemmingway
10.	*Bride of Pendorric* (Part I)	Victoria Holt
11.	*The Voice of Asia*	James Michener
12.	*Bride of Pendorric* (Part II)	Victoria Holt

In our example, the raw data consist of the frequencies with which each letter of the English alphabet is used in 12 different novels (see Table 3 in Fowlkes and Mallows (1983) for the data). Thus there are 12 vectors in $\boldsymbol{R}^{26}$, corresponding to the authors and novels in Table 1. The natural speculation is that books by the same author tend to cluster together. A more tenuous speculation is that authors with similar styles will also cluster together. To facilitate comparisons we have retained Fowlkes and Mallows's numbering system for the cases, which is why the order in Table 1 appears haphazard.

We applied nine different clustering algorithms to the data, trying each of the linkage options available in SAS (those considered by Fowlkes and Mallows (1985), as well as all other SAS procedures except density linkage and two-stage density linkage, which require substantial domain-specific judgment). The algorithms used were average linkage (AVE), centroid method (CEN), complete linkage (COM), estimated maximum likelihood method (EML), flexible-beta method (FLE), McQuitty's similarity analysis

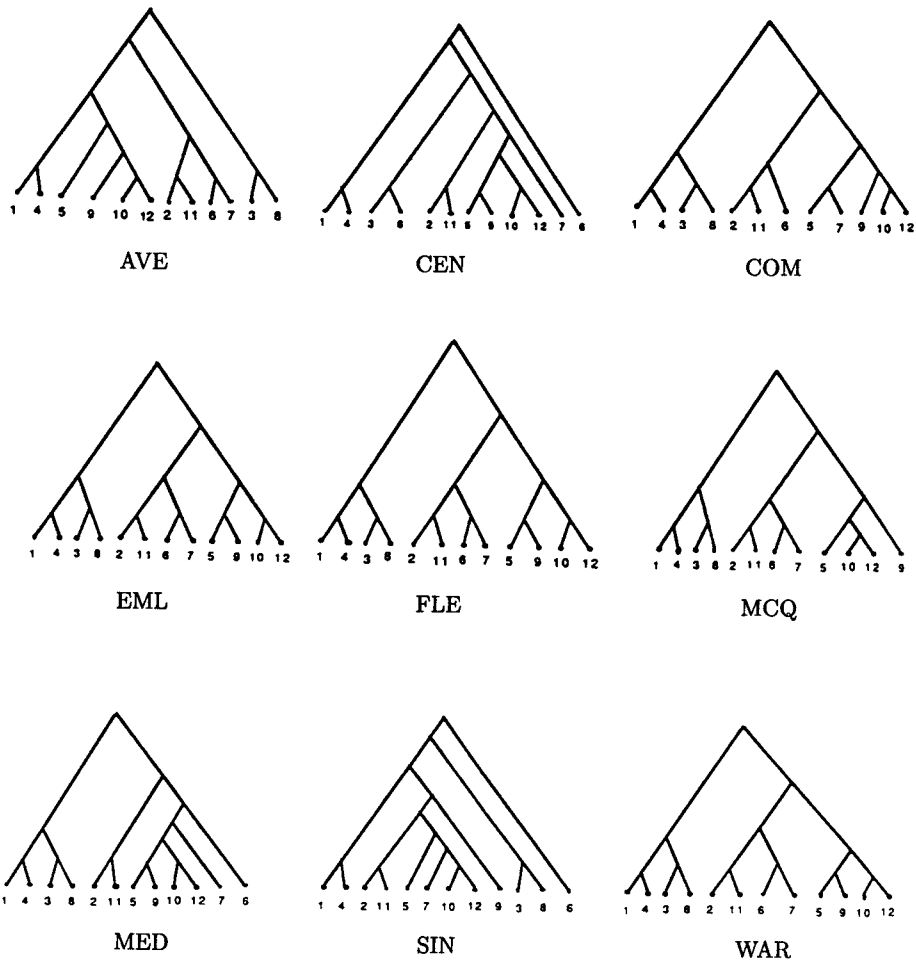

Figure 1. The nine binary trees produced by the application of different SAS algorithms for agglomerative clustering to the cases described in Table 1, using data given in Fowlkes and Mallows (1983).

(MCQ), median method (MED), single linkage (SIN), and Ward's minimum variance method (WAR). Details and references on these algorithms appear in Chapter 18 of *SAS/STAT User's Guide*, (1990).

In producing the nine trees shown in Figure 1, we used SAS program defaults throughout, varying only the specification of the linkage algorithm. The representations do not show the root nodes, which may be imagined as appended to the top of each tree.

Our analysis ignores information on the "height" of the tree when a particular split occurs, focusing instead on the topology of the tree. This simplification emphasizes the more stable features of the sample trees, and corresponds to practical usage. An alternative analysis that captures height information is possible, and work of this kind has been undertaken by Hendy and Penny (1993).

One approach to analyzing this data is to try to micromodel both letter frequency dependencies and the effects of the clustering algorithms. Since this level of detail is impractically difficult, we propose a simpler analysis that regards the outcomes of the different algorithms as a random sample from a distribution of the form in (1) on $\tilde{G}_{12}$. This formulation treats the data as fixed and the algorithms as random, which is unusual in statistics; the motivation is more natural in consensus theory (cf. Margush and Neumann 1983), where one often combines information from different trees built from a common dataset. Our approach is equivalent to assuming that if the data contain true central structure, then the observed trees consist of the central tree corrupted by errors capturing independent differences in the mechanics of the clustering algorithms. And the Gibbs distribution is especially appropriate here because the clustering algorithms, either by original design or evolutionary selection, are not trivially duplicative, and thus tend to increase the dispersion in the data.

Our objective is to estimate the central binary tree and to place a confidence region around it. Note that there are three identical trees: FLE, EML, and WAR. These should lie at or close to the center of our procedure's confidence region.

Choosing the Metric

Two broad strategies exist for defining a suitable metric: one strategy counts and weights the number of elementary operations needed to transform one tree into another, whereas the other strategy maps the trees into alternative mathematical structures for which natural metrics already exist. We describe two metrics for building probability models according to (1): an extension of the Hamming metric, and the Robinson crossover metric (1971). Related work on metrics for trees is described in Boorman and Oliver (1973), Margush (1982), and Barthélemy and Guénoche (1991).

One way to extend the Hamming metric strategy on graphs to a metric on binary trees is based on hypergraphs. The hypergraph generalizes traditional graphs by admitting edges that link more than two nodes (cf. Berge 1989). Our concern is hypergraphs that correspond to binary trees (an edge of this hypergraph is a cluster of cardinality greater than one in the hierarchy associated with the binary tree). To illustrate the correspondence, Figure 2

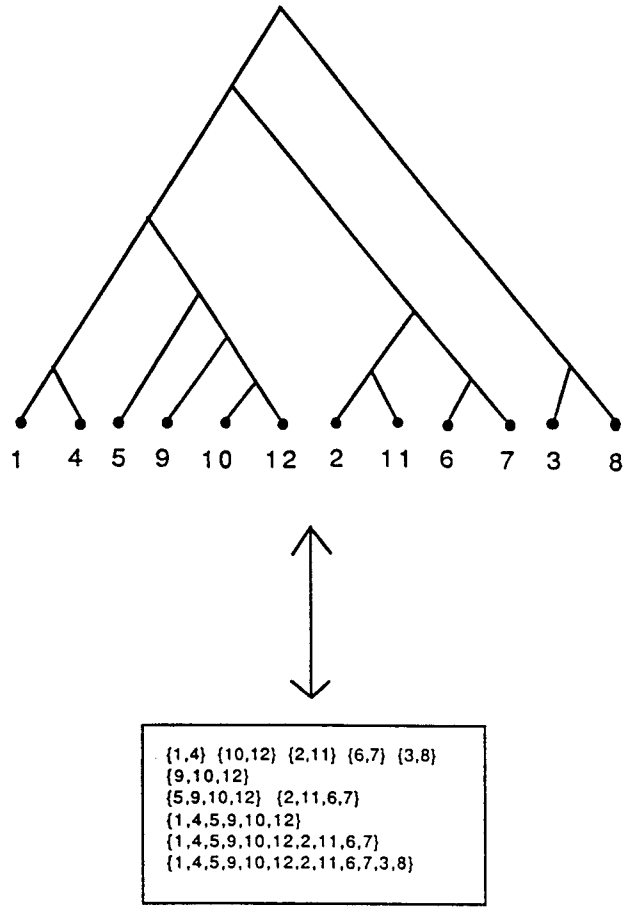

Figure 2. This figure illustrates the hypergraph representation of a binary tree. Sets in the lower square that contain r elements correspond to r-edges in the hypergraph, and the elements of the sets indicate which nodes are joined by the r-edge.

shows a binary tree and its equivalent hypergraph representation. We refer to an edge that connects r nodes as an r-edge. The binary tree in Figure 2 has five 2-edges, one 3-edge, two 4-edges, one 6-edge, one 10-edge, and one 12-edge.

In this framework, the extension of the Hamming strategy is to count and weight the number of discrepancies in each type of edge. Specifically, set

$$d_h(g_1,g_2) = \frac{1}{2}\sum_{r=2}^{m} \alpha_r \mid \{\text{discrepant } r\text{–edges between } g_1 \text{ and } g_2\} \mid , \quad (13)$$

where $|\cdot|$ denotes the cardinality of the argument set. The weights α_r must be positive, but may be chosen to reflect the practitioner's sense of an appropriate distance for a specific problem. In most cases it seems reasonable that α_r should increase with r, in order to impose a small penalty upon discrepancies in the twigs (r-edges with small r), but greater penalties for discrepancies among the branches (r-edges with large r).

The other metric we use was introduced by Robinson (1971); it counts the number of 'crossover' operations needed to convert one tree into another. To define the crossover operation, let e be any interior edge in a given tree $g \in \tilde{G}_m$; i.e., e links two nonterminal nodes, say v and w. Then e divides g into four subtrees; suppose A and B are the subtrees joined at v, and C and D are the subtrees that join at w. By exchanging subtrees, one can create two new trees g_1 and g_2 which are also in $\tilde{G}_m$; g_1 joins A and C at v, while g_2 joins A and D at v. In either case, the remaining subtrees connect at w. Figure 3 shows the two trees that arise from this crossover operation.

Robinson shows that any tree $g_1 \in \tilde{G}_m$ can be converted into any other tree $g_2 \in \tilde{G}_m$ through a sequence of crossover operations. Define the metric

$$d_R(g_1,g_2) = \min_k \{k : g_1 \text{ can be transformed to } g_2 \text{ in } k \text{ crossover operations}\} .$$

To see that d_R defines a metric on $\tilde{G}_m$, construct the graph S whose node set consists of the elements of $\tilde{G}_m$ and whose edges link nodes that are one crossover operation away. Because S is connected, there is a shortest path between any pair of nodes. This shortest path is necessarily a metric on the node set, and it exactly agrees with d_R.

The Analysis

For the nine trees in our sample, we employed the metric d_h in (13). Because edges in the hypergraph representation define elements in the nested partitions of the terminal nodes defined by the tree, the Hamming hypergraph metric may be viewed as a particular partition metric (cf. Boorman and Oliver 1973).

In using d_h, one must specify the $\{\alpha_r\}$ that weight discrepancies according to the number of nodes linked by the r-edge. In phylogeny, one imagines that discrepancies corresponding to large values of r are more significant than discrepancies corresponding to small values. For example, disagreements at the level of family or class are more serious than disagreements at the level of genus or species. In cluster analysis, the situation is less clear; for our data, the natural presumption is that books cluster according to

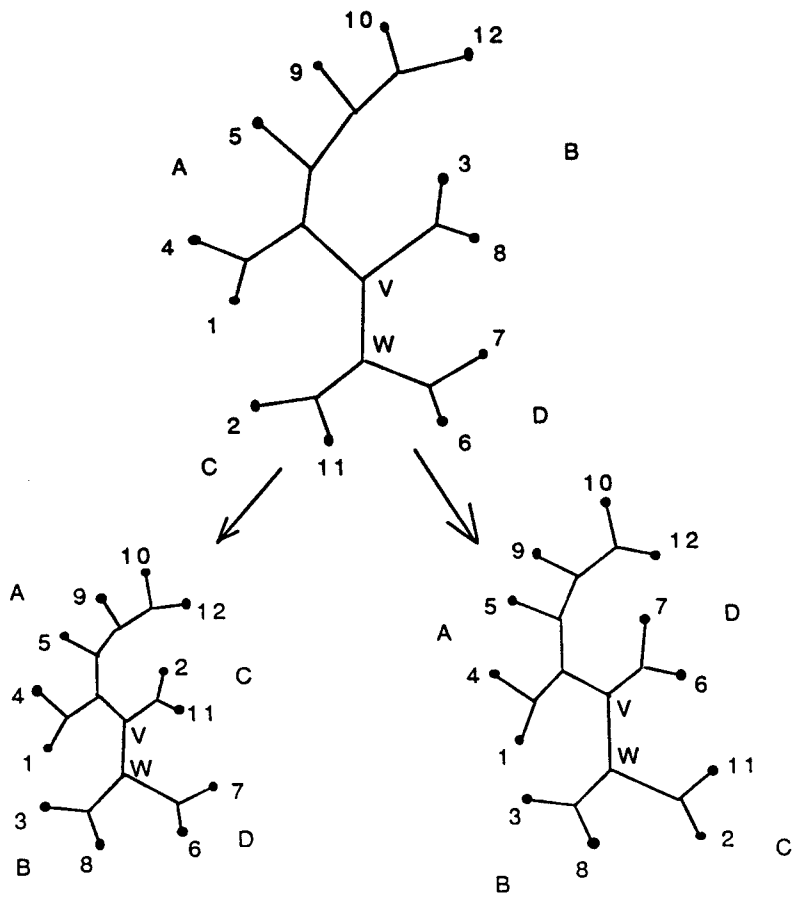

Figure 3. This figure illustrates the Robinson crossover; each interior edge determines two possible neighbor trees. In this case, the interior edge between V and W generates the lefthand neighbor tree by interchanging subtrees B and C, and generates the righthand neighbor tree by interchanging subtrees B and D.

author, and thus discrepancies for small values of r are most interesting. We examined three versions of d_h, with $\alpha_r = r$ (appropriate for phylogeny), $\alpha_r = r^{-1}$ (appropriate for the data reported in Fowlkes and Mallows (1983)), and $\alpha_r = 1$ (appropriate when all discrepancies are equally important). For our dataset, all variants of d_h showed essential agreement; this consensus reflects the fact that our sample of trees shows little dispersion.

There is a tension between the easy calculation of distances between two arbitrary trees and the easy determination of the neighbors of a specified tree. In our case, the Hamming hypergraph metric accomplishes the first goal

but encounters difficulty with the second. For the Robinson metric, the situation is reversed. This point is crucial, because maximizing the likelihood function requires both capabilities. Rapid calculation of the distances between arbitrary graphs is necessary when finding the sum of the distances between a candidate central tree and sample trees in (12). But to find a new candidate central tree that may achieve a larger likelihood, one wants to search the near neighbors of the current candidate.

To resolve this problem, our inference uses d_h, to make calculation of distances easy, but the algorithm searches for new candidate maxima among the Robinson neighborhood rather than the d_h neighborhood. To see that the Robinson neighborhood is a close approximation to the d_h neighborhood, note that the Robinson crossover preserves partitions below crossover points, and thus the number of r-edge discrepancies tends to be small, especially for small values of r. Also, we used progressively less truncated summation and multiple ascent searches started at the sample trees, as suggested in Section 4.

Specifically, the likelihood search went as follows:

1. At initialization, designate a sample observation as $\hat{g}_0^*$, the candidate central tree.
2. Conditional on $\hat{g}_0^*$ being the central tree, do numerical search to estimate τ according to the first equation in (12) (truncated summation is used to estimate the ratio). For this maximizing estimate of τ, calculate the loglikelihood.
3. Generate a random element g from the Robinson neighborhood of $\hat{g}_i^*$.
4. Conditional on g being the central tree, do numerical search as before to estimate τ and calculate the loglikelihood.
5. If the new loglikelihood is greater than the old, set $\hat{g}_{i+1}^* = g$ and go to Step 3. If the new loglikelihood is less than the old, go to Step 3. If the new loglikelihood is equal to the old and the number of consecutive changes in which equality has occurred is less than 100, set $\hat{g}_{i+1}^* = g$ and go to Step 3. If the new loglikelihood is equal to the old and the number of consecutive changes in which equality has occurred equals 100, or if there have been 200 returns to Step 3 without finding an improving candidate, return to Step 1, and select a new sample observation as the starting point. If all observations have been tried, end the program and report the candidate tree and the associated estimate of τ that gave the largest likelihood over all n searches.

This algorithm changes the candidate tree whenever it finds an increase in the likelihood, rather than searching the entire neighborhood to find the change that offers the largest increase. Thus it is an ascent search, but not steepest ascent.

For data with sufficiently small dispersion, the likelihood function is unimodal on $\tilde{G}_{12} \times R^+$. The performance of our algorithm indicates unimodality in this application because each restart at a sample graph led to the same answer.

To find confidence regions, we used the nonparametric bootstrap, repeated 200 times, and reran the search to find a 95% confidence region on the central tree. The entire program (estimation and bootstrapping) took approximately three hours to run using unoptimized code and IMSL routines on a Hewlett-Packard Apollo workstation, model 715/75.

For our dataset, all three d_h metrics identified the same tree as the central tree. That tree appears three times in the sample; it is generated by the EML, FLE, and WAR clustering algorithms. Moreover, the sample trees are all close to the estimated central tree, because the estimated values of τ were so large they produced underflow during the Monte Carlo evaluation of the ratio in (12), despite use of double precision. From that standpoint, we can only say that the best estimate of τ is larger than 9.4. For this application, imprecision in $\hat{\tau}$ is not problematic; its role in the likelihood function is influential only when data are dispersed, implying that τ is small and hence computable.

Naturally, the confidence regions generated by the three versions of d_h are different. When using $\alpha_r = r$, the bootstrap 95% confidence region consists of all binary trees that are within distance 17.50 of the estimated central tree. When using $\alpha_r = 1$, the 95% confidence region includes all trees within distance 3.00, and for $\alpha_r = r^{-1}$, the 95% confidence region includes all trees within distance .42. Thus, for the last case, the confidence region includes the EML, FLE, WAR, and, just barely, MCQ trees from the sample (there are other trees, not in the sample, that are also within the confidence region). This result strongly supports the view that books by the same author cluster together, and corroborates the conclusions reached by Fowlkes and Mallows (1983). In contrast, for the version of the metric more appropriate to phylogeny, the confidence region contains all of the sample trees except CEN and SIN, and would offer less support for the traditional interpretation of this data.

6. Conclusions

This paper presents a unified strategy for handling graph-valued random objects. For simple kinds of graphs with simple metrics, much can be worked out analytically, and there are interesting relationships between social network models and maximum-entropy models. Many of the standard tools of conventional statistics, such as Bayesian inference, hypothesis testing, and the linear model, are available in this unconventional setting. For more complex situations, the analysis is computer-intensive, but this obstacle is not insuperable.

The example with Fowlkes and Mallows's data shows how the method can be applied to a difficult problem arising in both cluster analysis and phylogeny. It demonstrates the importance of the choice of metric, and shows that a reasonable choice finds, for the cluster analysis of letter-frequency data, that authors show surprisingly similar profiles across books. Moreover, the confidence region obtained on those data is relatively small.

However, the metric used for cluster analysis is not as appropriate in phylogentic inference, underscoring the need for the involvement of domain experts in selecting metrics used in these problems. Also, the volume of the confidence region is sensitive to the metric employed, which is another reason for care in determining this aspect of the model.

More generally, the example points out computational difficulties that arise in analyzing general graph-valued random variables. These difficulties are solvable, and this paper gives computational methods and search heuristics that enable practical solutions.

7. Appendix

Proof of Proposition 1.

We can write

$$e(p) - \theta - \tau v = -\sum_{g \in G} p(g) \ln p(g) - \theta \sum_{g \in G} p(g) - \tau \sum_{g \in G} d(g,g^*) p(g)$$

$$= \sum_{g \in G} p(g) \ln \left(\frac{1}{p(g)} e^{-\theta - \tau d(g,g^*)} \right)$$

$$\leq \sum_{g \in G} p(g) \left(-1 + \frac{1}{p(g)} e^{-\theta - \tau d(g,g^*)} \right)$$

$$= -1 + \sum_{g \in G} e^{-\theta - \tau d(g,g^*)} .$$

The inequality reflects the geometric fact that the graph of $\ln x$ lies below its tangent line at the point $x = 1$; i.e., $\ln x \leq x - 1$ for all $x > 0$, with equality if and only if $x = 1$.

Equality in the previous calculation therefore occurs if and only if $p(g) = e^{-\theta - \tau d(g,g^*)}$. This choice of $p(g)$ maximizes the entropy. Since $\Sigma_g p(g) = 1$, substituting these values of $p(g)$ yields

$$c(g^*, \tau) = e^{-\theta} = \left[\sum_{g \in G} e^{-\tau d(g,g^*)} \right]^{-1} .$$

Similarly, the constraint $\Sigma_{g \in G} d(g,g^*) p(g) = v$ yields

$$\frac{dc(g^*,\tau)^{-1}}{d\tau} = -\nu c(g^*,\tau)^{-1},$$

implying $\frac{d \ln c(g^*,\tau)}{d\tau} = \nu$. Monotonicity of the function $c(g^*,\tau)$ in τ ensures a unique solution for τ in the last equation. •

Proof of Proposition 2.

Assume the model in (1) holds. As d_B is clearly a semimetric, we proceed to show that the dependence graph D is edgeless, implying the Bernoulli graph model.

Let e_{st} and e_{uv} be distinct edges, possibly sharing a common node. Let $I\{e:g\}$ be an indicator function taking the value 1 if and only if edge e is in g. Conditional on all the other edges and nonedges in a random graph g, the probability of both e_{st} and e_{uv} in g is

$$\frac{\exp[-a_{st}(1-I\{e_{st}:g^*\}) - a_{uv}(1-I\{e_{uv}:g^*\}) - c]}{\exp[-a_{st}-a_{uv}-c] + \exp[-a_{st}-c] + \exp[-a_{uv}-c] + \exp[-c]}, \tag{14}$$

where c is a constant reflecting all of the edges and nonedges upon which we condition. Similarly, the probability of e_{st} conditional on all edges/nonedges except e_{uv} is

$$\frac{\exp[-a_{st}(1-I\{e_{st}:g^*\}) - c] + \exp[-a_{st}I\{e_{st}:g^*\} - a_{uv} - c]}{\exp[-a_{st}-a_{uv}-c] + \exp[-a_{st}-c] + \exp[-a_{uv}-c] + \exp[-c]}, \tag{15}$$

and the analogous expression holds for the conditional probability of e_{uv}. Multiplication of (15) with its analogue produces (14), so e_{st} and e_{uv} are conditionally independent, D is edgeless, and the Bernoulli graph model holds.

Going the other way, assume the Bernoulli graph model holds. If p_{ij} is the probability of an edge linking nodes i and j, then

$$P[g] = \prod_{i<j} p_{ij}^{I(e_{ij}:g)} (1-p_{ij})^{1-I(e_{ij}:g)}. \tag{16}$$

For the model in (1), the case when $g = g^*$ shows that the normalizing constant is

$$c(g^*,\tau) = P[g^*] = \prod_{i<j} \max\{p_{ij}, 1-p_{ij}\}.$$

Letting

$$a_{ij} = \begin{cases} -\frac{1}{\tau}\ln\frac{p_{ij}}{1-p_{ij}} & \text{if } p_{ij} \le .5 \\ -\frac{1}{\tau}\ln\frac{1-p_{ij}}{p_{ij}} & \text{if } p_{ij} > .5, \end{cases}$$

the Bernoulli graph model in (16) can be rewritten as

$$P[g] = \prod_{i<j} \max\{p_{ij}, 1-p_{ij}\} \times \prod_{e_{ij}\text{ing},\text{not}g^*} \frac{p_{ij}}{1-p_{ij}}$$

$$\times \prod_{e_{ij}\text{ing}^*,\text{not}g} \frac{1-p_{ij}}{p_{ij}} = c(g^*,\tau) \times \prod_{e_{ij}\text{ing},\text{not}g^*} e^{-\tau a_{ij}}$$

$$\times \prod_{e_{ij}\text{ing}^*,\text{not}g} e^{-\tau a_{ij}} = c(g^*,\tau) e^{-\tau d_B(g,g^*)},$$

which has the form of (1). Notice that τ is not identifiable; it can be subsumed by defining new edge probabilities $p'_{ij} = (p_{ij})^{1/\tau} / ((p_{ij})^{1/\tau} + (1-p_{ij})^{1/\tau})$. Thus we take $\tau = 1$, so that the a_{ij} are interpretable in terms of odds ratios. •

Proof of Proposition 3

Under the Holland-Leinhardt p_1 model, and letting i,j index the node set, it is straightforward to write the probability of a graph g as the product of the probabilities of each type of dyadic relationship between all possible pairs of nodes. Thus $P[g] = \prod_{i<j} p_{ij}[h_{ij}(g)]$ for

$$h_{ij}(g) = \begin{cases} 1 & \text{if no edges link node } i \text{ to node } j \text{ in } g \\ 2 & \text{if } i,j \text{ are linked by a single edge from } i \text{ to } j \text{ in } g \\ 3 & \text{if } i,j \text{ are linked by a single edge from } j \text{ to } i \text{ in } g \\ 4 & \text{if edges link nodes } i \text{ and } j \text{ in both directions in } g \end{cases}$$

and

$$p_{ij}[1] = = 1/k_{ij}$$

$$p_{ij}[2] = \exp[\theta + \alpha_i + \beta_j]/k_{ij}$$

$$p_{ij}[3] = \exp[\theta + \alpha_j + \beta_i]/k_{ij}$$

$$p_{ij}[4] = \exp[\rho + 2\theta + \alpha_i + \alpha_j + \beta_i + \beta_j]/k_{ij} \qquad (17)$$

where

$$k_{ij} = 1 + e^{\theta+\alpha_i+\beta_j} + e^{\theta+\alpha_j+\beta_i} + e^{\rho+2\theta+\alpha_i+\alpha_j+\beta_i+\beta_j}$$

and the α, β, θ, and ρ terms in (17) are as in Holland and Leinhardt (1981).

Define $p^*_{ij} = \max\{p_{ij}[1], \ldots, p_{ij}[4]\}$. Then $P[g^*] = \prod_{i<j} p^*_{ij}$, and so

$$\begin{aligned} P[g] &= P[g^*] \times \prod_{i<j} \frac{p_{ij}[h_{ij}(g)]}{p^*_{ij}} \\ &= P[g^*] \times \prod_{i<j} \exp\{-(\ln p^*_{ij} - \ln p_{ij}[h_{ij}(g)])\} \\ &= P[g^*] \times \exp\{-\sum_{i<j} (\ln p^*_{ij} - \ln p_{ij}[h_{ij}(g)])\} \\ &= P[g^*] \times \exp\{-d(g,g^*)\} . \end{aligned} \tag{18}$$

Since $P[g^*] = c(g^*,\tau)$, it remains to show that d is a metric. The concentration parameter τ is not an issue, since metricity is preserved under multiplication by a positive constant.

Consider $d_{ij}(g_1,g_2) = |\ln p_{ij}[h_{ij}(g_1)] - \ln p_{ij}[h_{ij}(g_2)]|$. This is clearly a semimetric on G^*_m since the absolute value function is a metric on $\boldsymbol{R}$ (non-identical graphs may have d_{ij} distance zero if they disagree on edges other than those between nodes i and j, so d_{ij} is only a semimetric). Defining $d_{HL}(g_1,g_2) = \Sigma_{i<j}\, d_{ij}(g_1,g_2)$ ensures that d_{HL} is a semimetric, since sums of semimetrics are semimetrics. More strongly, it is clear that d is in fact a metric, since $d(g_1,g_2) = 0$ if and only if $g_1 = g_2$.

This definition of d_{HL} satisfies the requirements for d in (18), and thus the Holland-Leinhardt model can be written in the form of (1). Thus, in Proposition 3, one has $b_{ij}(g) = \ln p_{ij}[h_{ij}(g)]$. We note that this proof took no account of the special structure on the $p_{ij}(1), \ldots, p_{ij}(4)$, and thus the more general models for directed graphs that are discussed in Holland and Leinhardt (1981), which require only independence, are also maximum-entropy distributions. •

Proof of Proposition 4

Consider a tree $g \in \tilde{G}_{m-1}$ having one root and $m-1$ labelled terminal nodes. Then there are $2m-3$ edges in g, and a new terminal node, labelled "m" may be inserted on any of these to create a tree in $\tilde{G}_m$. The insertion places an unlabelled interior node on the chosen edge, and depends node m from that.

Each tree in $\tilde{G}_m$ is created in this way from exactly one tree in $\tilde{G}_{m-1}$; insertion on different edges produces different trees. So if $\tilde{G}_{m-1}$ has

cardinality c_{m-1}, then $\tilde{G}_m$ has cardinality $(2m-3)c_{m-1}$. The result follows by noting that $c_3 = 3$. •

References

BANKS, D. L., and CARLEY, K. (1994), "Metric Inference for Social Networks," *Journal of Classification, 11*, 121-149.

BARTHÉLEMY, J. P., and GUÉNOCHE, A. (1991), *Trees and Proximity Representations*, New York: Wiley.

BARTHÉLEMY, J. P., and MONJARDET, B. (1988), "The Median Procedure in Data Analysis: New Results and Open Problems," in *Classification and Related Methods of Data Analysis*, Ed., H.H. Bock, North-Holland: Elsevier, 309-316.

BARTHÉLEMY, J. P., LECLERC, B., and MONJARDET, B. (1986), "Ordered Sets in Problems of Comparison and Consensus," *Journal of Classification, 3*, 187-224.

BERGE, C. (1989), *Hypergraphs: Combinatorics of Finite Sets*, New York: North-Holland.

BERNARDO, J. M., and SMITH, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

BLOEMENA, A. R. (1964), *Sampling from a Graph*, Amsterdam: Mathematisch Centrum.

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

BOORMAN, S. A., and OLIVIER, D. (1973), "Metrics on Spaces of Finite Trees," *Journal of Mathematical Psychology, 10*, 26-59.

CAPOBIANCO, M. (1970), "Statistical Inference in Finite Populations Having Structure," *Transactions of the New York Academy of Sciences, 32*, 401-413.

CONSTANTINE, G. (1991), "Graph Identification," *Proceedings of the 6th Caribbean Conference on Combinatorics and Computing*, 92-117.

DAY, W. H. E. (1986), "Comparison and Consensus of Classifications," *Journal of Classification, 3*, 183-186.

FELSENSTEIN, J. (1985), "Confidence Limits on Phylogenies, an Approach Using the Bootstrap," *Evolution, 39*, 783-791.

FESTINGER, L. (1949), "The Analysis of Sociograms Using Matrix Algebra," *Human Relations, 2*, 153-158.

FIENBERG, S. E., MEYER, M. M., and WASSERMAN, S. S. (1985), "Statistical Analysis of Multiple Sociometric Relations," *Journal of the American Statistical Association*, 80, 51-67.

FISHER, N. I., and HALL, P. (1990), "On Bootstrap Hypothesis Testing," *Australian Journal of Statistics, 32*, 177-190.

FOWLKES, E. B., and MALLOWS, C. L. (1983), "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association, 78*, 553-568; Rejoinder, 584.

FRANK, O. (1976), *Statistical Inference in Graphs*, Stockholm: Swedish Research Institute of National Defense.

FRANK, O. (1988), "Random Sampling and Social Networks: A Survey of Various Approaches," *Mathématiques, Informatique and Sciences Humaines*, 26, 19-33.

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

GEMAN, S., and GEMAN, D. (1975), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741.

HAMMERSLEY, J. M., and CLIFFORD, P. (1971), "Markov Fields on Finite Graphs and Lattices," unpublished manuscript.

HAMMING, R. (1950), "Error Detecting and Error Correcting Codes," *Bell System Technical Journal, 29*, 147-160.

HENDY, M. D., and PENNY, D. (1993), "Spectral Analysis of Phylogenetic Data," *Journal of Classification, 10*, 5-24.

HOLLAND, P. W., and LEINHARDT, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association, 76*, 33-65.

KATZ, L. (1947), "On the Matrix Analysis of Sociometric Data," *Sociometry, 10*, 233-241.

KATZ, L. (1953), "A New Status Index Derived from Sociometric Analysis," *Psychometrika, 18*, 39-43.

KATZ, L. (1955), "Measurement of the Tendency Towards Reciprocation of Choice," *Sociometry and the Science of Man, 18*, 659-665.

KRACKHARDT, D. (1987), "Cognitive Social Structures," *Social Networks, 9*, 109-134.

MALLOWS, C. (1957), "Non-Null Ranking Models I," *Biometrika, 44*, 114-130.

MARGUSH, T. (1982), "Distances Between Trees," *Discrete Applied Mathematics, 4*, 281-290.

MCMORRIS, F. R., and NEUMANN, D. A. (1983), "Consensus ; Functions Defined on Trees," *Mathematical Social Sciences, 4*, 131-136.

MORENO, J. L. (1934), *Who Shall Survive?* Washington, D.C.: Nervous and Mental Disease Publishing.

REEDS, J. A. (1985), "Asymptotic Number of Roots of Cauchy Location Likelihood Equations," *Annals of Statistics, 13*, 775-784.

ROBINSON, D. F. (1971), "Comparison of Labelled Trees with Valency Three," *Journal of Combinatorial Theory, 11*, 105-119.

SAS INSTITUTE, INC. (1990), *SAS/STAT User's Guide*, Volume 1, Version 6, Cary, NC: SAS Institute, Inc.

SCHRÖDER, E. (1870), "Vier Kombinatorische Probleme," *Zeitschrift fuer Mathematische und Physikalische, 15*, 361-376.

STRAUSS, D. (1983), "Hammersley-Clifford Theorem," in *Encyclopedia of Statistical Sciences, 3*., Eds., S. Kotz, N. Johnson and C. Read, New York, N.Y.: Wiley, 570-572.

STRAUSS, D., and FREEMAN, L. C. (1989), "Stochastic Modelling and the Analysis of Structural Data," in *Research Methods in Social Network Analysis*, Eds., L. Freeman, A. Romney and D. White, Fairfax, VA: George Mason University Press, 135-183.

STRAUSS, D., and IKEDA, M. (1990), "Pseudolikelihood Estimation for Social Networks," *Journal of the American Statistical Association, 85*, 204-212.

WASSERMAN, S. (1987), "Conformity of Two Sociometric Relations," *Psychometrika, 52*, 3-18.

WONG, G. Y. (1987), "Bayesian Models for Directed Graphs," *Journal of the American Statistical Association, 82*, 140-148.