Scan

A6472
etc

Frank & Svensson

add to 3 segs

Statist. Comput. Simul., 1981, Vol. 12, pp. 121–131
0094-9655/81/1202-0121 $06.50/0

On Probability Distributions of Single-Linkage Dendrograms

OVE FRANK and KLAS SVENSSON

Department of Statistics, Lund University, S-220 07 Lund, Sweden

(Received March 13, 1980)

There are $\binom{N}{2}!$ ways to order the pairwise similarities between N objects, assuming no ties. According to single linkage (SL) clustering, each such order determines a dendrogram for the N objects. We give an algorithm for calculating the number of different SL-dendrograms on N objects. We also give an algorithm for calculating the probability distribution of the SL-dendrograms under pure randomness, i.e. assuming that all the similarity orders are equally probable. The results are used to illustrate the statistical risks for small values of N, when SL-dendrograms are used to test cluster structure hypotheses.

KEY WORDS: Partition, cluster, similarity, graph inference, stochastic dendrogram.

1. INTRODUCTION

Methods for cluster analysis are usually developed as tools for exploratory data analysis, and statistical inference based on dendrograms and other kinds of output data from cluster analyses has not been investigated much in the literature. A discussion of the need for model-based approaches to cluster analysis is included in the book by Hartigan (1975, Section 1.4) and in the survey article by Cormack (1971). Ling (1973) and Ling and Killough (1976) have used probabilistic models for cluster analysis. A recent bibliography is given by Naus (1979).

Our purpose here is to show that it is possible, by an algorithmic approach, to enumerate all possible single-linkage (SL) dendrograms for N objects and find their probabilities under a particular randomization model for the similarities between the objects. For small values of N we shall also consider some alternative models of cluster structure and investigate the usefulness of SL-dendrograms for testing cluster structure against pure randomness.

We shall use graph concepts to describe cluster structure and formulate our problem as a graph inference problem with dendrogra data. Frank (1978a, b, 1979) has investigated similar graph inferen problems with other kinds of data obtained from sampling ai measurement error models.

The next section defines the basic concepts we will need, and Section describes a randomization model. The algorithms for calculating t! number of SL-dendrograms and their probabilities are given in Section Finally, some statistical applications are discussed in Section 5.

2. CLUSTER STRUCTURE AND DENDROGRAMS

Consider a set V of N objects and a cluster structure in V defined as partition of V into K parts (non-empty disjoint subsets with union V This cluster structure can also be considered as an equivalence relation V having K equivalence classes or as a transitive graph G having vert set V and K complete components. When we use graph concepts we sh in general follow the terminology of Harary (1969).

The number of non-isomorphic transitive graphs of order N is equal the number of partitions of N. Denote this number by A^N; A_N is equal 1, 2, 3, 5, 7, 11 for $N=1,\ldots,6$. The number of transitive labeled graphs order N is given by the so called Bell number B_N which can obtained from the recurrence relation

$$B_{N+1}=\sum_{n=0}^{N}\binom{N}{n}B_n$$

for $N=0,1,\ldots,$ where $B_0=1$ (see, for instance, Riordan (1968)). The fir values of B_N are 1, 2, 5, 15, 52, 203 for $N=1,\ldots,6$.

We define a *dendrogram* for N objects as a sequence of hierarchic partitions of the object set starting with the partition into N one-obje parts and successively merging two parts $N-1$ times, so that the fin partition consists of one N-object part. It follows that there are

$$\binom{N}{2}\binom{N-1}{2}\cdots\binom{2}{2}=N!(N-1)!/2^{N-1}$$

labeled dendrograms. This number can also be determined for $N=5$ indicated in Figure 1; the numbers at the arcs are the numbers of paths reach the next partition by merging two parts, and the total number paths from the initial to the final partition is equal to the number dendrograms. This number is obtained by calculating successively fro below the number of paths to the final partition. For unlabele dendrograms, note that even though objects are not distinguishable, a

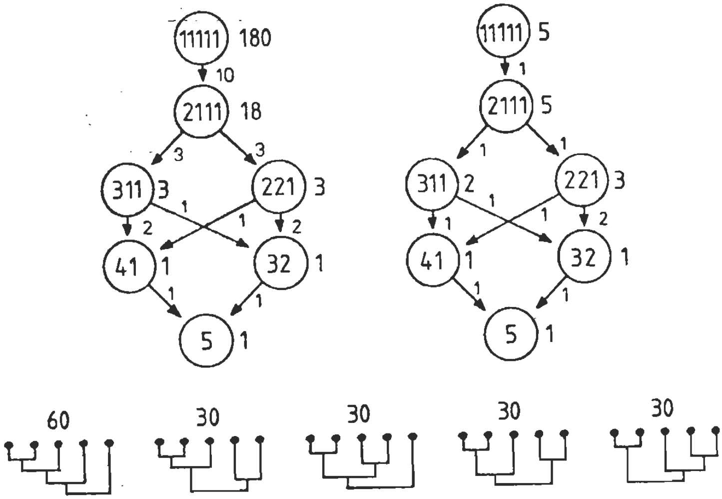

FIGURE 1. Counting the labeled and unlabeled dendrograms for five objects.

parts consisting of more than one object are distinguishable since they are created at different levels in the hierarchy. We find that there are 180 labeled dendrograms of five non-isomorphic types for five objects. Figure 1 shows these non-isomorphic dendrograms and their numbers of labeled isomorphic variants.

An *SL-dendrogram* for N objects is a dendrogram in which the successive merges are associated with integer levels which are obtained by using ranked similarities with no ties. The similarities between the pairs of objects are ranked by $1, \ldots, \binom{N}{2}$, so that rank 1 is assigned to the most similar pair and so forth. The similarity data can be considered as an edge-ranked labeled complete graph X. The number of ways to edge-rank a labeled complete graph of order N is $\binom{N}{2}!$. Since there are $N!$ ways to label the vertices and since distinct vertex labelings yield distinct edge-ranked graphs if $N > 2$. it follows that there are

$$\binom{N}{2}!/N! \qquad (3)$$

non-isomorphic edge-ranked complete graphs of order N for $N > 2$. These graphs are constructed for $N = 4$ along the paths from top to bottom in Figure 2. The numbers at the arcs are the numbers of ways to add another edge, and the product of these numbers along a path yields the number of isomorphic variants of the corresponding edge-ranked graph. There are 720 edge-ranked complete graphs of 30 non-isomorphic types for $N = 4$.

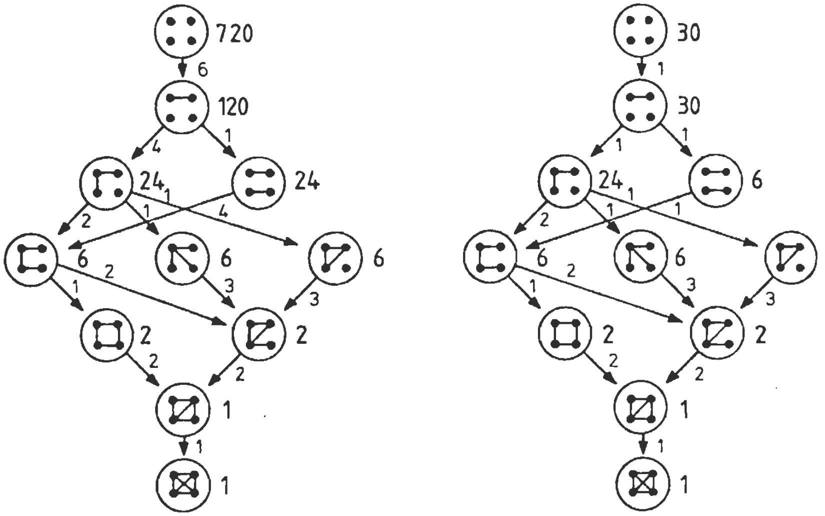

FIGURE 2. Counting the edge-ranked labeled and unlabeled complete graphs of order ·

Let $S=S(X)$ be the SL-dendrogram obtained from the edge-rank labeled complete graph X. The SL-dendrogram S can conveniently defined by means of the following sequence of subgraphs of X. For $=1,\ldots,\binom{N}{2}$, let X_r be the subgraph of X of order N and size r whi consists of the edges of ranks at most r. Further, let X_0 be the zero-si graph of order N. The SL-dendrogram S merges two parts at level r these two parts are the vertex sets of distinct components in X_{r-1} whi belong to a common component in X_r. For $N=4$, we find from Figure that the SL-dendrograms can be represented by the paths in Figure They can be counted by the algorithm described in Section 4. For $N=$ there are 30 labeled SL-dendrograms of three non-isomorphic types, shown in Figure 3.

3. A RANDOMIZATION MODEL

Consider the cluster structure given by a transitive graph G of order and size R. Assume that there are uncertain measurements of similari available for all pairs of objects, and that the R adjacent pairs in G have higher similarities than the other pairs. Let the first R ranks assigned at random to the edges in G and the next $\binom{N}{2}-R$ ranks assign at random to the non-adjacent vertex pairs in G. Then there are

$$R!\left[\binom{N}{2}-R\right]! \qquad ($$

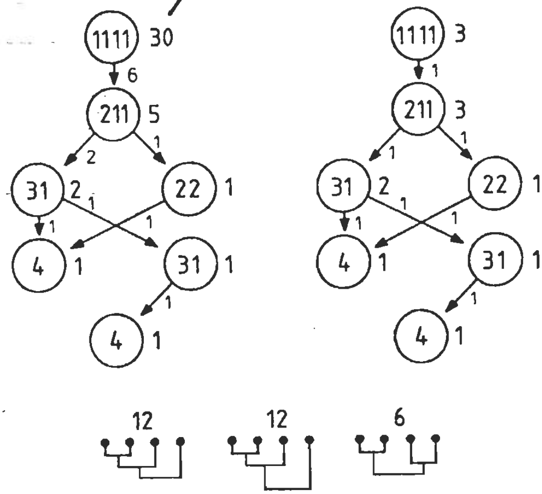

FIGURE 3. Counting the labeled and unlabeled SL-dendrograms for four objects.

equally probable outcomes of the edge-ranked complete graph X. The model corresponding to $R=0$ or $R=\binom{N}{2}$ will be referred to as *pure randomness*. Other values of R compatible with transitive graphs G correspond to cluster structure models. For instance, $N=4$ yields the possible values 1, 2 and 3 on R, and X has 120, 48 and 36 equally probable outcomes, respectively.

4. THE NUMBER OF SL-DENDROGRAMS AND THEIR PROBABILITIES UNDER PURE RANDOMNESS

In order to count the labeled SL-dendrograms for N objects, we shall apply a technique which is based on the numbers of ways of creating a coarser partition by merging two parts in an arbitrary partition.

Consider a partition of N into K parts of which K_n are equal to n for $n=1,2,\ldots$. Let R be the level of this partition, i.e., let the partition be preceded by exactly R merges. It can be shown that R satisfies the inequalities

$$N-K \leqq R \leqq \sum_n \binom{n}{2} K_n. \tag{5}$$

The number of ways to merge two distinct parts m and n is equal to

$$\begin{aligned} &K_m K_n \quad \text{if } m \neq n,\ K_m \geqq 1,\ K_n \geqq 1 \\ &\binom{K_n}{2} \quad \text{if } m=n,\ K_n \geqq 2. \end{aligned} \tag{6}$$

A partition can also remain at the next level if and only if corresponding graph is not transitive; i.e., no merge occurs if and only i

$$\sum_n \binom{n}{2} K_n > R.$$

The total number of ways to merge two parts is equal to

$$\sum_n \binom{K_n}{2} + \sum_{m<n}\sum K_m K_n = \binom{K}{2}.$$

By starting with a partition into N parts and applying the rules above, see that the last merge will occur at level $\binom{N-1}{2}$ and will lead to the fi one-part partition at level $\binom{N-1}{2}+1$.

The unlabeled SL-dendrograms can be counted by a similar techniq The only difference is that (6) for $m=1$ is replaced by

$$\begin{aligned} &K_n \ \text{ if } n \neq 1,\ K_1 \geqq 1,\ K_n \geqq 1 \\ &1 \ \text{ if } n=1,\ K_1 \geqq 2. \end{aligned}$$

Figures 3 and 4 illustrate the application of these rules to determine t numbers of labeled and unlabeled SL-dendrograms for four and f objects.

In order to find the probabilities of the SL-dendrograms under pı randomness, we shall make use of the fact that these probabilities can obtained by multiplying the probabilities of the successive merges.

Consider an arbitrary partition of $N=\Sigma_n nK_n$ at level $R \leqq \binom{N-1}{2}$. T next edge can be assigned to any of $\binom{N}{2}-R$ non-adjacent vertex pairs. T probability that two distinct parts m and n are merged is equal to

$$\begin{aligned} &mnK_mK_n \Big/ \left[\binom{N}{2}-R\right] \ \text{ if } m \neq n,\ K_m \geqq 1,\ K_n \geqq 1 \\ &n^2\binom{K_n}{2}\Big/\left[\binom{N}{2}-R\right] \ \text{ if } m=n,\ K_n \geqq 2. \end{aligned} \quad (1$$

and the probability of no merge is equal to

$$\left[\sum_n \binom{n}{2} K_n - R\right] \Big/ \left[\binom{N}{2}-R\right]. \quad (1$$

A259345(5)

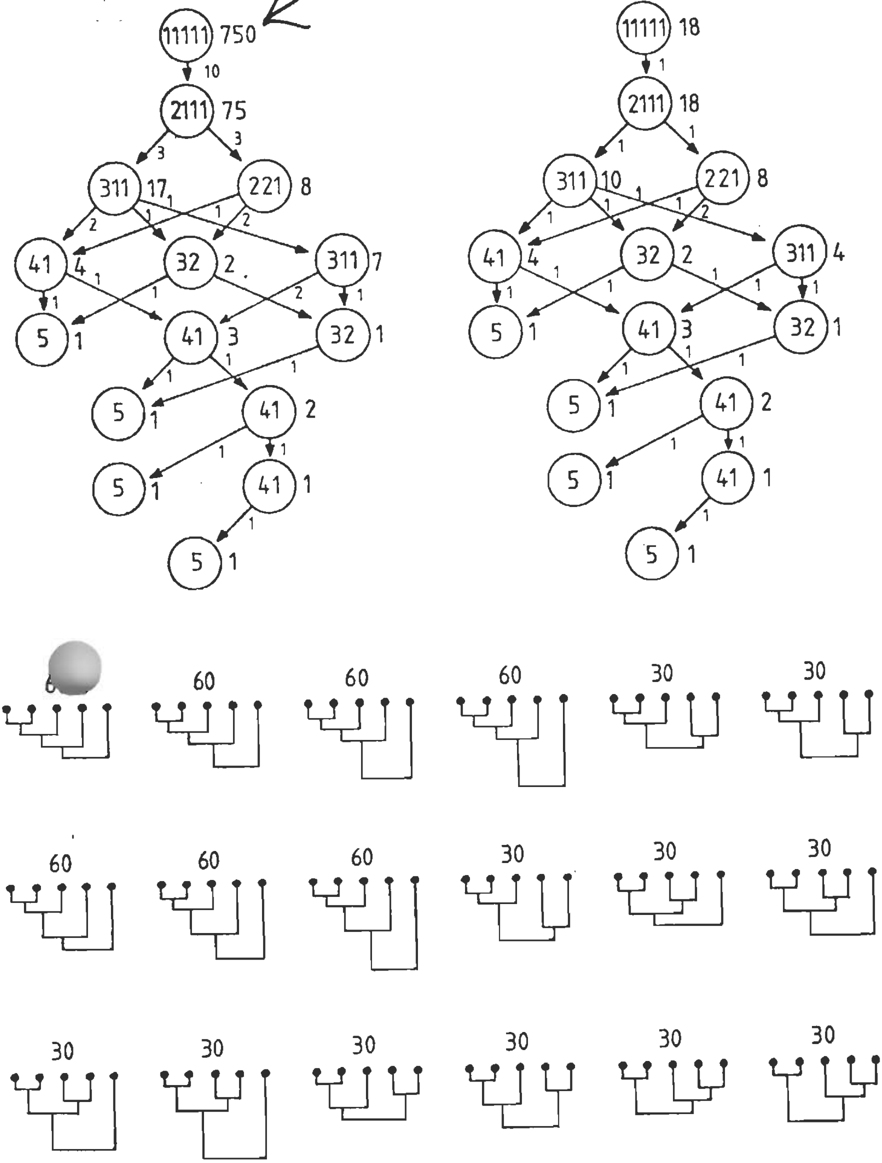

FIGURE 4. Counting the labeled and unlabeled SL-dendrograms for five objects.

These transition probabilities are shown in Figures 5 and 6 for four ar five objects. By multiplication along the paths we find the probabilities the SL-dendrograms. The SL-dendrograms for $\Lambda = 4$ in the ord displayed in Figure 3 have the probabilities 3/5, 1/5 and 1/5. The S

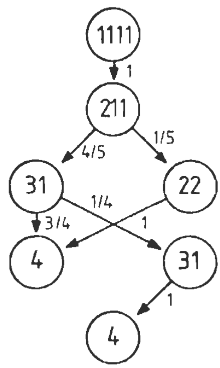

FIGURE 5. Transition probabilities for SL-partitions of four objects according to pu randomness.

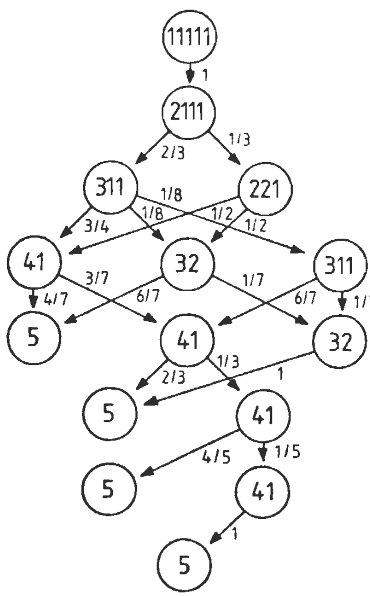

FIGURE 6. Transition probabilities for SL-partitions of five objects according to pur randomness.

dendrograms for $N=5$ in the order displayed in Figure 4 have the probabilities

120/420,	60/420,	24/420,	6/420,	30/420,	5/420,
20/420,	8/420,	2/420,	5/420,	40/420,	20/420,
8/420,	2/420,	30/420,	5/420,	30/420,	5/420.

5. TESTING CLUSTER STRUCTURE

Assume that a population V of N objects has an unknown cluster structure given by a transitive graph G, and that available information consists of an SL-dendrogram S generated according to the randomization model in Section 3. The empty and complete graphs G correspond to the hypothesis of pure randomness, and the other A_N-2 graphs correspond to cluster structure hypotheses. A partition $N=\Sigma_n nK_n$ corresponds to

$$N!/\prod_n K_n!(n!)^{K_n} \tag{12}$$

partitions of N labeled objects, and in total there are B_N-2 cluster structure models for N labeled objects, besides the degenerate model of pure randomness.

Consider $N=4$ and let $V=\{a,b,c,d\}$. The transitive unlabeled graphs will be denoted by partitions 4, 31, 22, 211 and 1111, and the transitive labeled graphs by partitions abc|d, ab|cd, ab|c|d, and so forth. The SL-dendrograms of Figure 3 will be denoted S_1, S_2, S_3 for unlabeled objects. For labeled objects, S_1(cd) and S_2(cd) denote S_1 and S_2, where a and b are merged first, c next and d last. S_3(cd) denotes S_3, where a and b are merged first and c and d next.

The randomization models of types 31, 22 and 211 can be handled by the same counting techniques as the model of pure randomness. Table I shows the distribution of S for each G. From Table I we find that the maximum-likelihood decision $\hat{G}$ is of type 211, 31 and 22 for $S=S_1$, S_2 and S_3, respectively. Table II shows the distribution of $\hat{G}$ for each G. In particular, the risk of not finding a true cluster structure is 0 for structures of type 31 and 22 and 2/5 for structures of type 211; the risks of deciding upon various cluster structures under pure randomness are equal to 1/20, 1/15 and 1/10 for any labeled structure of type 31, 22 and 211, respectively. We also note that under pure randomness the labeled decisions of type 31, 22 and 211 have together a probability of 1/5, 1/5 and 3/5, respectively. The maximum-likelihood decision never rejects cluster structure among four objects.

TABLE I

Distribution of S for each G.

S \ G	abcd	abc|d	abd|c	acd|b	bcd|a	ab|cd	ac|bd	ad|bc	ab|c|d	ac|b|d	ad|b|c	bc|a|d	bd|a|c	cd|a|b
S_1 (ab)	36													3
S_1 (ac)	36												36	
S_1 (ad)	36											36		
S_1 (ba)	36													3
S_1 (bc)	36										36			
S_1 (bd)	36									36				
S_1 (ca)	36												36	
S_1 (cb)	36										36			
S_1 (cd)	36								36					
S_1 (da)	36											36		
S_1 (db)	36									36				
S_1 (dc)	36								36					
S_2 (ab)	12			12										1
S_2 (ac)	12		12										12	
S_2 (ad)	12	12										12		
S_2 (ba)	12				12									1
S_2 (bc)	12		12								12			
S_2 (bd)	12	12								12				
S_2 (ca)	12				12								12	
S_2 (cb)	12			12							12			
S_2 (cd)	12	12							12					
S_2 (da)	12				12							12		
S_2 (db)	12			12						12				
S_2 (dc)	12		12						12					
S_3 (ab)	24					24								2
S_3 (ac)	24						24						24	
S_3 (ad)	24							24				24		
S_3 (bc)	24							24			24			
S_3 (bd)	24						24			24				
S_3 (cd)	24					24			24					
	720	36	36	36	36	48	48	48	120	120	120	120	120	1

TABLE II

Distribution of $\hat{G}$ for each G.

$\hat{G}$ \ G	abcd	abc|d	abd|c	acd|b	bcd|a	ab|cd	ac|bd	ad|bc	ab|c|d	ac|b|d	ad|b|c	bc|a|d	bd|a|c	cd|a|b
abcd														
abc|d	36	36							12	12		12		
abd|c	36		36						12		12		12	
acd|b	36			36						12	12			12
bcd|a	36				36							12	12	12
ab|cd	48					48			24					24
ac|bd	48						48			24			24	
ad|bc	48							48			24	24		
ab|c|d	72								72					
ac|b|d	72									72				
ad|b|c	72										72			
bc|a|d	72											72		
bd|a|c	72												72	
cd|a|b	72													72
	720	36	36	36	36	48	48	48	120	120	120	120	120	120

References

Cormack, R. M. (1971) A review of classification. *J. Royal Statist. Soc.* A134, 321–367.

Frank, O. (1978a). Estimation of the number of connected components in a graph by using a sampled subgraph. *Scand. J. Statist.* 5, 177–188.

Frank, O. (1978b). Inferences concerning cluster structures. *Proceedings in computational statistics*, edited by L. C. A. Corsten and J. Hermans, Physica Verlag, Wien. 259–265.

Frank, O. (1979). Estimating a graph from triad counts. *J. Statist. Comput. Simul.* 9. 31–46.

Harary, F. (1969). *Graph Theory*. Addison-Wesley, Reading, Massachusetts.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley, New York.

Ling, R. F. (1973). A probability theory of cluster analysis. *J. Amer. Statist. Assoc.* 68, 159–164.

Ling, R. F. and Killough, G. G. (1976). Probability tables for cluster analysis based on a theory of random graphs *J. Amer. Statist. Assoc* 71, 293–300.

Naus, J. I. (1979). An indexed bibliography of clusters, clumps and coincidences. *Int. Statist. Rev.* 47, 47–78.

Riordan, J. (1958). *An Introduction to Combinatorial Analysis*. Wiley, New York.