This is to certify that the

thesis entitled

CLUSTER VALIDITY AND INTRINSIC DIMENSIONALITY

presented by

Thomas Anderson Bailey, Junior

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Computer Science

Richard C. Dubes

Major professor

Date Aug. 18, 1978

[redacted] 7-84 348

A 005

[redacted] 270

[redacted] 311

JUN 03 '87

[redacted]

[redacted]

CLUSTER VALIDITY AND INTRINSIC DIMENSIONALITY

By

Thomas Anderson Bailey, Junior

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

1978

ABSTRACT

CLUSTER VALIDITY AND INTRINSIC DIMENSIONALITY

By

Thomas Anderson Bailey, Junior

Cluster analysis, one type of exploratory data analysis, is a crucial part of a pattern recognition study. Although many methods for discovering clusters in a data set have been developed, few techniques for objectively evaluating clusters are available. One approach to this question of cluster validity uses a null hypothesis of "no clustering" based on random graph theory and applies when the proximities between data points have ordinal scale, as is assumed for several popular clustering techniques.

A new random-graph-based technique for validating clusters, called the cluster profile, is developed and analyzed in this dissertation. Simple indices of cluster compactness and isolation are defined, and measures of validity are developed from probability distributions of the indices over sample populations selected from subsets of nodes in random graphs. The measures are inexpensive to calculate and may be applied to any subset of nodes, whether discovered by clustering methods or defined a priori. Thus the measures may be used to judge the validity of arbitrary clusters and are not limited to

clusters found by a specific clustering method.

Random-graph-based techniques for validating clusters are limited to a particular null hypothesis. This dissertation studies the extent to which techniques for judging cluster validity based on random graph theory are applicable to data sets produced under an alternative random model. Monte Carlo simulation is used to create data sets of points from a uniform density in a hypercube. Distributions of the validity indices, obtained from these data sets, are compared to the distributions under the random graph null hypothesis. The dissertation shows that as the dimensionality of the hypercube decreases, the distributions of validity indices obtained under this uniform hypercube null hypothesis consistently shift away from those obtained under the random graph null hypothesis. Because of this shift, almost all data sets produced under the uniform hypercube model with low dimensionality contain "valid" clusters when tested against a random graph null hypothesis.

One resolution of this difficulty is suggested by a simulation of the placement of points at random in a hypercube. The ratio of the shortest to longest interpoint distances among five points approaches one as the dimensionality of the hypercube becomes much larger than five. This result implies that, at very high

dimensionalities, the distributions of validity indices under a uniform hypercube null hypothesis should be approximately the same as those under the random graph null hypothesis.

The cluster profile technique is applied to a speaker recognition problem to illustrate its applicability. This technique not only extends existing validity measures, but also applies to clusters formed by any method, which is a primary advantage over previously defined validity tests.

ACKNOWLEDGMENTS

I would like to acknowledge the guidance provided by my thesis director, Dr. Richard Dubes. During my tenure at Michigan State University he has served as advisor, teacher, research director, and finally as editor of this dissertation. His continuing support and direction were essential to the completion of my doctoral program.

I thank Dr. Carl Page, Dr. Anil Jain, Dr. John Forsyth and Dr. Edgar Palmer for serving on my doctoral committee. Two fellow students, Karl Pettis and Richard Bolz, provided emotional support and often served as a captive audience for the presentation and clarification of ideas.

Thanks are also due to the Division of Engineering Research at Michigan State University and to the National Science Foundation (Grant No. ENG 76-11936) for financial support during the last two years.

Finally, a special thanks to my wife, Carolyn. Her support through many trying moments, her gentle (and sometimes not-so-gentle) prodding, and her sacrifice of a comfortable home, made this study possible.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Clustering is used as a tool of data analysis in areas as diverse as sociology [BRE75] and medicine [WON77]. Clustering may be used as a means of classifying a large mass of data [MSH77] or to suggest a causative relationship between variables [BOR75]. The aim of clustering is to find natural groupings present in the data. Ideally, these natural groups are compact and well isolated from one another. Unfortunately, the clustering process may impose cluster structure even if it is not appropriate.

1.1. The Cluster Validity Problem

Many clustering methods are currently available and are being used for data analysis and data description. The data generally comprise a set of points, which may be objects of any kind, such as subjects, features, samples, journals, species, etc., plus information about the points, which may take the form of a set of features for each point or a proximity value for each pair of points. The proximity value indicates the extent to which the two data points are alike or close together. (See [AND73] or [BLA77] for a thorough presentation of proximity measures and clustering techniques and algorithms.) Though these methods provide information about the data, it is unclear in many cases whether the information is inherent in the

data, or is being imposed by the clustering method, especially since different clustering methods often give different results [DUB76]. Thus, the question of validity arises. Should the results, the clusters or partitions or hierarchies produced by a clustering method, be accepted as a representation of the data, or should they be viewed as a form which has been imposed on the data?

We can distinguish several types of cluster validity questions. The first type asks whether the data exhibit a tendency to cluster. There are two models of the mechanism which produces the tendency to cluster. The usual model assumes the data are drawn, by independent trials, from a distribution characterized by the clustering observed in the data set. This is the model assumed by almost all users of clustering. An alternative model, used by Strauss [STR75, KEL76], assumes the data points would have a non-clustered distribution if they could be drawn by independent trials, but the presence of an interaction mechanism causes the set of observed points to occur in clusters. Under Strauss' model, it is also possible to produce "anti-clustered" data sets in which the interaction mechanism causes the observed points to be more evenly distributed than expected for independent trials. Whichever model is appropriate, tests for the presence of clustering tendency should be the first step in an analysis of clustering structures. They can be

performed before clustering algorithms are applied to the data [DUB77].

The second type of cluster validity question asks whether the relationship between clusters represents the true structure of the data set. Most clustering methods produce a set of clusters and an indication of the relationship between clusters. The membership of the clusters and the relationship between clusters defines the structure of the data set. Under Strauss' model such a relationship does not exist in the underlying distribution, but under the more usual model this data structure indicates the structure of the population from which the data were drawn. For example, if a clustering method has forced a data set into an inappropriate hierarchical structure, we would like to be able to detect this error.

The third type of question concentrates on the individual cluster. Given the data set in which the cluster is located, the cluster may exhibit characteristics, such as compactness, long lifetime, small variance about the clustering center, or large separation from other clusters, which lead to the conclusion that the cluster forms a valid subgroup which should be treated as a single entity.

The validity questions can be asked on several levels. The observed structure can be compared with other

similar structures on the same data set. It can be compared with a postulated ideal structure. Or it can be compared with similar structures over the set of all similar data sets. In the same manner, a cluster may be compared with all possible clusters in the same data set, with all possible clusters in all similar data sets, or with the best cluster in all similar data sets.

This thesis examines techniques for answering the question: Is this cluster valid? More specifically, is it unusual to find a cluster which is as compact and isolated as the observed cluster? The approach is probabalistic in nature and is based on the theory of random graphs.

The remainder of Chapter 1 provides some necessary definitions and a literature review. Chapter 2 describes and analyzes a new tool called a Cluster Profile. Appropriate indices of cluster isolation and compactness are defined. Probability distributions which can be used to test cluster validity are developed and measures of cluster isolation and compactness based on these distributions are defined. Chapter 3 describes a study which considers one relaxation of the "no clustering" assumption used in Chapter 2. Chapter 4 reports an application of the Cluster Profile technique to a speaker recognition problem and Chapter 5 draws conclusions, identifies the contributions of the thesis and outlines further work.

1.2. Non-Graph-Theoretic Approaches

Many different approaches to the cluster validity questions have appeared. One approach is to consider only whether the results of a clustering algorithm, either the clusters or the structure, make sense. The user must try to explain the results and no validity checks based on the distribution of the data points are used. Anderberg [AND73, pp.18-19] poses three cases in response to the question "How do you know when you have a good set of clusters?" In each case the question is answered without reference to the distribution of the data points. In Case One clustering is used to provide summary statistics. In this case the validity question is irrelevant because the only question is the accuracy of the calculation. In Case Two the clustering technique is defined in such a way that any clusters found must have the desired properties. Again the validity of the results is not in question. Any clusters found are, by definition, valid.

In Case Three the clustering technique is used as an exploratory tool. Anderberg takes the position that all the results should be evaluated by attempting to explain them and validity tests based on distributions of the data are of no value. If the results cannot be explained, then validity tests will not save them. If an explanation is forthcoming, then the validity test results are not relevant to the explanation.

In contrast, Rapoport and Fillenbaum [RAP72] state that "Safeguards of various sorts (such as stress values less than critical cutoff points, significant clusterings, significant graph results, etc.) are obviously necessary to guard against elaborate interpretation of randomly generated data." The increasing use of clustering in many different fields and continuing work on the question of cluster validity (see [DUB77] for a complete review) shows that the view of Rapoport and Fillenbaum is widely accepted at present.

Several existing approaches to cluster validity that do not use the graph-theoretical models on which this thesis is based, but which establish a background for the thesis, are outlined in the subsequent sections.

1.2.1. The Restricted Definition Approach

The work of McQuitty [MQU61, MQU67] is an early example of the use of a strict definition of cluster to ensure that any clusters which are found will be valid. Hubert [HUB74a] uses the term "perfect cluster" for subsets which satisfy some strict definition of cluster.

Definitions of perfect clusters usually compare an index of compactness with an index of isolation. A cluster is compact if its points have a high degree of similarity, thus forming a cohesive set. A cluster is isolated if it is well separated from other clusters, so

that points in the cluster are very dissimilar from points not in the cluster. The significance of an index of compactness or isolation depends on the distribution of the index across the data set. In the following definitions a cluster with strong compactness needs less isolation to be considered perfect than does a cluster with weak compactness.

McQuitty defines a "comprehensive type" as a subset of points for which each point in the subset is more like every other point in the subset than it is like any point not in the subset. McQuitty also provides another, more easily satisfied, definition for a perfect cluster. A "restricted type" is a subset of points for which each point in the subset is more like some other point in the subset than it is like any point outside the subset. The isolation criterion is the same for both definitions, the least similarity between the point in question and a point not in the subset. This isolation criterion provides, for each point in the subset, a standard outside point against which the compactness is compared. The point in question must be more similar to some point in the subset -- or to all points in the subset -- than to the standard outside point. The isolation index is used as a reference with which the compactness index is compared.

A less easily satified definition of perfect cluster, given by van Rijsbergen [RIJ70], requires that the

smallest similarity between points in the subset be greater than the largest similarity between some point in the subset and some point not in the subset. Hubert [HUB74a] gives several generalizations of these definitions.

The work of Day [DAY77] is an extension of this approach to cluster validity. His work applies to the "overlapping" case, in which a data point may belong to more than one cluster, which is not covered in this thesis. He defines two properties, consistency and authenticity, which a clustering method should exhibit and then investigates classes of clustering methods to see if they have the desired properties. Day also defines general indices of cluster cohesion (or compactness) and cluster attenuation (or isolation). The indices must be specialized to each clustering method and no expected values or distributions are given.

1.2.2. The Statistical Test Approach

Other work on cluster validity has approached the question via statistical tests. This approach requires a null hypothsis, or random distribution, against which cluster validity indices, derived from a data set or from the results of applying a clustering method to a data set, may be evaluated. This section describes two types of

random distributions which have been used as the "no clustering" hypothesis.

1.2.2.1. Tests Based on Random Patterns

Data sets in pattern recognition studies often consist of feature values measured on each object under study. The values can be used to form a pattern matrix, a matrix with rows representing the objects and columns representing the features. Each object may then be represented by a point in a multidimensional space, with each dimension corresponding to a feature. One type of randomness hypothesis assumes the data have been drawn from a known unimodal distribution in a multidimensional space. Typical distributions of this type are the uniform distribution inside a multidimensional sphere and the multidimensional normal distribution [ENG69, SNE77]. Both of these distributions seem natural as null hypotheses. They represent the cases of no clustering or only one cluster.

In one strategy for testing cluster validity we would like to know the null distribution of a cluster validity index over the best partition or the best cluster in each random data set. Tests based on such null distributions are difficult to devise because it is necessary to find the best cluster or partition for a general data set, and this is very difficult, especially in a space of many

dimensions. The requirement that only the best clusters or partitions in each random set be considered also prohibits the use of standard analysis of variance techniques to determine cluster validity.

1.2.2.2. Tests Based on Random Proximities

The information in the pattern matrix can be used to form a proximity matrix. A proximity matrix is a square matrix with row i and column i both representing object i. The proximity matrix is symmetric and represents either similarity measures between objects, such as correlation, or dissimilarity measures, such as distance. Sometimes, especially with psychometric data, the proximities are measured directly and there is no pattern matrix. A second type of randomness hypothesis assumes that the starting point for the clustering analysis is the proximity matrix. A model for the creation of the proximity matrix in the absence of clustering is developed and tests of cluster validity are based on random proximity matrices drawn from the model distribution.

Mountford [MOU70] offers a test of the difference between two clusters. The null hypothesis is that the proximities are drawn from a normal distribution with mean u and variance S. The covariance is c*S, where c is a constant, if the two proximities share a data point, ie., if both are on the same row or column of the proximity

matrix, otherwise the covariance is Ø. This hypothesis is tested against the alternative that the data points split into two groups with at least two items in each group. The test is conservative because of the bound which must be used to ensure that the best possible value of the statistic over all partitions of the random data set is found.

1.3. Graph Theoretic Approaches

Another model of the null hypothesis for statistical tests of cluster validity assumes that values in the proximity matrix have only ordinal significance. This is widely assumed for data collected in psychology and the social sciences [JOH67]. The information of interest is the order of the values in the matrix. Under this model there is a clear relationship between graph theory and the proximity matrix. Descriptions of clustering methods in terms of graph theory occur in many papers [HAR67, ZAH71, HUB74a, HUB74b, MAT77].

1.3.1. Graph Theory and Clustering

Two prominent hierarchical clustering methods, the single link method and the complete link method [JOH67], produce results which depend on the order of the proximities but not on the actual values. Single link clusters are subgraphs of minimum spanning trees [GOW7Ø].

Complete link clusters are related to the node colorability of graphs [BAK76].

Agglomerative algorithms for implementing these methods begin by considering each point in the data set to be a cluster. The set of proximities is searched to find the most similar pair of clusters, which are then joined to form a new cluster. If ties occur in the set of proximity values, the definition of these clusters is greatly complicated, especially in the case of complete link clusters. In this thesis, we assume that no ties occur. The process is continued until all the points have been joined into one cluster.

The single and complete link methods differ in the way in which new proximities are defined when a cluster forms. The single link method, also called the minimum distance method, defines the proximity of the new cluster to an old cluster to be the smallest of the two proximities between the parts of the new cluster and the old cluster. In contrast, the complete link method, also called the maximum distance method, defines the proximity of the new cluster to an old cluster to be the largest of the two proximities between the parts of the new cluster and the old cluster.

In both methods the clusters are determined solely by the order of the proximities, so these methods are appropriate for proximity matrices composed of ordinal

data. A survey [BLA77 (1) pg.13, Table 3] of applications in 122 research publications in 1973 showed that at least 58 of 162 applications of clustering involved clustering methods which required only ordinal information. The widespread use of these clustering methods implies that a description of "no clustering" based on the rank of entries in the proximity matrix rather than on the positions defined by the pattern matrix will be of value.

1.3.2. Random Graph Models

We define two graphs from a proximity matrix D (on an interval scale) as follows. Let a threshold c be given. A threshold graph, T(D,c), is a graph on n labeled nodes with two nodes i and j connected by an edge if the (i,j) entry in D is less than or equal to c. A different threshold graph is defined for each distinct entry in D. A rank graph, R(D,c), is an edge-weighted threshold graph at level c with an order imposed on the edges by the order of the entries in D. If we assume no ties among the proximities, the edges of R(D,c) may be labeled sequentially to indicate the order of the proximities. Each proximity matrix determines a set of threshold and rank graphs, one pair of graphs for each distinct value of c.

The representation of the proximity matrix D by a set of rank graphs is accompanied by a loss of information,

except in those cases where the entries in D are on an ordinal rather than an interval or ratio scale. The utility of the rank graph is seen in the wide use of clustering methods, such as single link and complete link, which require only ordinal information from the proximity matrix to form the sequences of clusterings. The clusters of the single link method are components of some threshold graph, and each component of a threshold graph is a single link cluster. The clusters of the complete link method are maximal complete subgraphs (cliques) of some threshold graph. However, not all cliques of a threshold graph are complete link clusters. In both methods, the clusters are determined by the order of the proximity values but do not depend on the actual values. Thus the proximity matrix may be replaced with the complete rank graph with no change in the sequence of clusters produced by these methods. Other graph theoretic concepts of connectedness, such as k-edge and k-node connectedness, have also been used to define clusters which depend only on the rank graph [HUB74a, MAT77].

Consider the set of all symmetric n by n matrices with zero entries on the diagonal and with the integer values 1 to n(n-1)/2 in the upper triangle. A random rank matrix is a matrix chosen at random from this set. An experiment in which the proximities of objects are ordered by random choice should give a "no clustering" result. In

fact, not only is there "no clustering", but other forms of structure will also be absent [LIN73a].

A random rank matrix may be represented by a complete rank graph. The weights on the edges of a random rank graph are integers showing the order of the edges. A random rank graph may be formed directly by randomly ordering the edges (or node pairs) of a complete graph. A random threshold graph with N edges is the subgraph of a random rank graph having the same node set as the random rank graph and having all the edges with rank less than or equal to N. In this thesis a "random graph" is a random threshold graph. The random graph formed by choosing a random rank graph on n nodes and then using the threshold rank N to form a threshold graph is equivalent to the random graph found by choosing a graph at random from the set of all labeled graphs with n nodes and N edges. This definition of random graph is used by Erdos and Renyi [ERD59], Ling [LIN73a] and Baker and Hubert [BAK76].

Several authors speak of the evolution of a random graph [ERD60]. If a random rank graph on n nodes is given, a different threshold graph is defined for each of the n(n-1)/2 distinct threshold ranks. The threshold graph at rank i+1 differs from the graph at rank i by a single edge. A random graph is evolved by starting with a graph of n nodes and no edges, and then repeatedly adding new edges one by one at random until the graph is

complete. If the edges are labeled as first, second, etc., as they are entered, then a random rank graph is created by this evolutionary process.

Because graphs provide a representation of the ordinal information in a proximity matrix, assuming all orderings of the proximities are equally likely is equivalent to assuming all rank graphs are equally likely. In turn, the concept of random graph evolution links random rank graphs and random threshold graphs. Any hypothesis of "no clustering" which is equivalent to assuming that all orderings of the proximities are equally likely will be called a random graph null hypothesis.

1.3.3. Validity Tests Based on Random Graphs

Erdos and Renyi [ERD59, ERD60, ERD61] list a number of asymptotic results in random graph theory. They are particularly interested in the asymptotic behavior of various graph properties as the number of nodes in the graph increases. In many cases they also give exact expressions for the probability of special subgraphs, such as cycles of order k, in random graphs. Erdos and Renyi define a random graph as a graph with n nodes and N edges where the edges have been chosen at random, without replacement, from among the n(n-1)/2 node pairs. The nodes are labeled. Thus there are three possible graphs with three nodes and one edge.

Abraham [ABR64] uses graph theoretic notions to define several different types of clusters. He also attempts to use random graph theory to determine the significance of clustering tendency. Several errors in his asymptotic expressions limit the usefulness of the results [LIN75].

Rapoport and Fillenbaum [RAP72] use several results from Erdos and Renyi in tests of non-randomness for data gathered in studies of semantic structure. They test sets of trees using the distributions of the degree sequence of the nodes. They also test sets of graphs using degree sequence, occurrence of cycles of order 3 and 4, and the number of edges required to connect the graphs. They test clusters by using the difference between the mean rank of edges inside clusters and the mean rank of edges between clusters. Unfortunately, the distribution which they use to test the clusters is appropriate only if the cluster of nodes being tested is selected independently of the graph which determines the statistic. In their case, the cluster is chosen by selecting a good cluster on the basis of the graph, so the distribution on which the test is based is not the appropriate one.

Ling [LIN73a] defines an isolation index called lifetime for single link clusters. The lifetime is the number of edges in the graph in which the cluster is absorbed by creation of another cluster less the number of

edges in the graph in which the cluster is formed. Using the random graph null hypothesis, Ling determines the distribution of this index for a given cluster size and a given number of edges in the graph at formation of the cluster. The lifetime index is defined for clusters obtained by any hierarchical technique. However, the distribution is specific to the single link clustering technique.

Several authors have obtained results for the expected value of the number of edges needed to connect a random graph. Erdos and Renyi give an asymptotic form. Rapoport and Fillenbaum used this asymptotic form for small graphs. Schultz and Hubert [SCH73] later showed, using Monte Carlo simulation, that the asymptotic form was not accurate for small graphs. Ling [LIN75] and Ling and Killough [LIN76] used exact results due to Riddell and Ulenbeck [RID53] to produce expressions of greater accuracy for small graphs and tables of accurate results.

The number of edges needed to connect a graph is an index of the tendency toward clustering of the nodes in the graph. It is not an indication of compactness or isolation for a particular cluster. The number of edges needed to connect a graph measures the clustering tendency at only one rank in the evolution of the graph. A test which is applicable at all ranks uses the number of components in a graph. Expected values for this index are

given by Ling [LIN73b]. Since only the expected values and not the complete distribution are given, no test of significance can be based on this index.

Baker and Hubert [BAK75] use the random graph null hypothesis in a study of the power of a test of clustering. The test is based on the Goodman-Kruskal gamma statistic and is applied to single and complete link hierarchies. The gamma statistic is used to measure rank correlation between the actual proximity matrix and an ideal proximity matrix derived from the cluster hierarchy. The alternative hypotheses consist of proximity matrices which are "perfect" for a partition into three clusters, to which Gaussian noise is added. The study consists of Monte Carlo runs to determine the distributions of interest. The results can be used to test the fit of a proximity matrix to the hierarchy of clusters given by the single or complete link clustering technique. The required simulation results are given only for the 12 node case.

In another study Baker and Hubert [BAK76] used Monte Carlo simulation under the random graph null hypothesis to find the distribution of an isolation index defined for a partition into complete link clusters. The index is the number of extraneous edges, edges which are not internal to some complete link cluster. They propose a test of goodness-of-fit in which the observed number of extraneous

edges after each new complete link cluster forms is evaluated by reference to tables produced by their simulation. The published tables are for clusters on 8, 12 and 16 nodes.

Matula [MAT77] finds the distribution of the size of the largest clique (maximal complete subgraph) for a random edge graph. A random edge graph is a graph in which each node pair has probability p of being chosen as an edge. The number of edges is not specified. The expected number of edges for a random edge graph of n nodes is pn(n-1)/2. The distribution of largest clique size, or clique number, is quite peaked. Chance occurrence of a clique which is more than a few nodes larger than the expected size is quite unlikely.

1.3.4. Limitations of Present Validity Tests

The work to date in cluster validity based on random graphs is limited in two ways. First, with the exception of the clique number test of Matula, the tests of cluster validity which have been proposed are specific to particular clustering methods. They cannot be used to test clusters found by any of the many other proposed graph theory based clustering techniques. Second, with the exception of the cluster lifetime test of Ling and the clique number test of Matula, the tests of validity are based on distributions which must be obtained by

simulation. Since the simulations are specific to graphs of certain sizes, the results cannot be used to test experimental results unless the sizes happen to match. In most cases the experimenter would need to run a simulation for the graph size which matches his experiment in order to use the proposed test. The tests proposed in Chapter 2 can be applied to any subset of points and do not require simulation to determine the required distributions.

2. Cluster Profiles

Most random graph tests of cluster isolation and compactness are based on one or two ranks in the evolution of the complete graph. Ling [LIN73a] forms an index of the isolation of a single link cluster by noting the difference in the rank at formation and at absorption of the cluster. Two important ranks in the evolution of the graph are used, the lowest rank at which the nodes of the cluster are connected and the lowest rank at which some node in the cluster is connected to a node not in the cluster. An index of isolation proposed by Baker and Hubert [BAK76] is the number of edges between complete link clusters. All edges of a partition which are not within clusters are counted. The fewer edges between clusters, the more isolated they are, and the more valid is the partition. In this case one rank is considered for each new cluster, the rank at which the complete link cluster is formed.

By contrast, the Cluster Profile method proposed in this chapter looks at all ranks, or thresholds, in the data set. This requires compactness and isolation indices which are defined at each threshold. The proposed method has the advantage of being applicable to the results of any clustering method. Any ranks which are of special significance for a particular clustering method will be included.

Sections 2.1 and 2.2 define the concepts of "raw" and "probability" profiles. Indices of validity appropriate to profiles are defined and several probability distributions which are used to investigate cluster validity are developed. The probability distributions lead directly to the definition of several measures of cluster validity which comprise the probability profiles. Section 2.3 examines a special case of one probability distribution to determine the accuracy of a probability bound developed in Section 2.2.

2.1. Raw Profiles

A "raw profile" is a sequence of cluster validity indices. The sequence is formed by observing the indices in the threshold graphs for each distinct threshold. A raw profile provides a basic picture of the interaction of a cluster with other elements in the data set. This section develops the definition of the raw profile and defines indices of cluster isolation and compactness.

2.1.1. The Sequence of Threshold Graphs

A "rank graph", representing the order of edges in a proximity matrix, may be thought of as a sequence of threshold graphs, one threshold graph for each possible threshold. If the thresholds are distinct, then each threshold graph in the sequence has one more edge than the

preceeding graph. Suppose that a clustering method has been applied to the data and one or more clusters have been identified. We wish to test the proposed clusters for isolation and compactness.

A raw profile of each cluster is developed as follows. Indices of compactness and isolation of the proposed cluster are observed in each threshold graph. The sequence of values for these indices, one value for each rank, forms a profile of the cluster over the evolution of the rank graph. With properly chosen indices it may be possible to classify clusters as "isolated" or "compact" directly from these raw profiles, although the information may be hard to interpret. This motivates the probability profile.

2.1.2. Indices of Isolation and Compactness

In this section two indices of cluster validity are defined. Typically, indices of cluster validity are based on some definition of compactness or cohesiveness of a set of nodes or on some definition of isolation or uniqueness of the set of nodes. The indices proposed here are very simple. This simplicity has two advantages. First, the indices are easy to evaluate, which makes them computationally inexpensive. Second, their simple nature allows them to be applied to many different situations. More complex definitions of validity indices are often

limited to one clustering method. These simple indices can be applied to the clusters produced by any clustering method. A disadvantage is that it is difficult to compare clusters of different sizes or clusters from different data sets using these indices.

The indices are based on a threshold graph. Let D be a proximity matrix. Let A be a subset of the data points which has been proposed as a cluster. The subset A may be used to divide the proximities in the upper right triangle of D into three sets. The first set, D(A,in), is the set of proximities, d(i,j) with i<j, for which both data points i and j are in the subset A. The second set, D(A,out), is the set of proximities, d(i,j) with i<j, for which both data points i and j are not in A. The remaining proximities in the upper right triangle, D(A,betw), are those with one data point in A and one data point not in A.

For each possible threshold t, an index of cluster compactness can be defined as follows. Let e(t) be the number of proximities in D(A,in) which are less than or equal to t. Then, for each t, e(t) is an index of cluster compactness. If a cluster is very compact at level t it will have many pairs of points with dissimilarity less than t. If the cluster is not compact, it will have relatively few pairs with proximities below t.

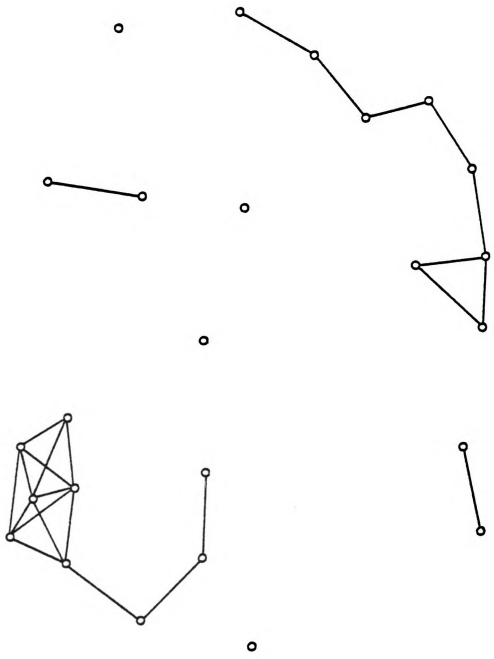
An index of cluster isolation can be defined in a

similar way. Let b(t) be the number of proximities in D(A,betw) which are less than or equal to t. Then b(t) is an index of cluster isolation. A very isolated subset at level t will have relativly few proximities below t, while a subset of points which is not isolated will have many points linking to points not in the subset with proximities less than t.

A cluster which is perfectly compact and isolated will have, for some threshold, all proximities in D(A,in) below the threshold and all proximities in D(A,betw) above the threshold. Such a cluster satifies van Rijsbergen's definition [RIJ70] of a perfect cluster.

2.1.3. An Example

Figure 2.1 is an artificial data set of 25 points in two dimensions. The dissimilarity measure is the Euclidean distance between points. The threshold graph, containing 25 edges, for a threshold of 1.00 inches is shown. The six points which are circled at the lower left are a proposed cluster. These six points are a single-link cluster and also a complete-link cluster. For this threshold graph with 25 edges the cluster has a compactness index of 12 and an isolation index of 1. This cluster does not satisify van Rijsbergen's definition of a perfect cluster. However, it is apparent that the edges of the threshold graph are concentrated in the cluster.



The threshold graph with 25 edges using distance as the proximity measure.

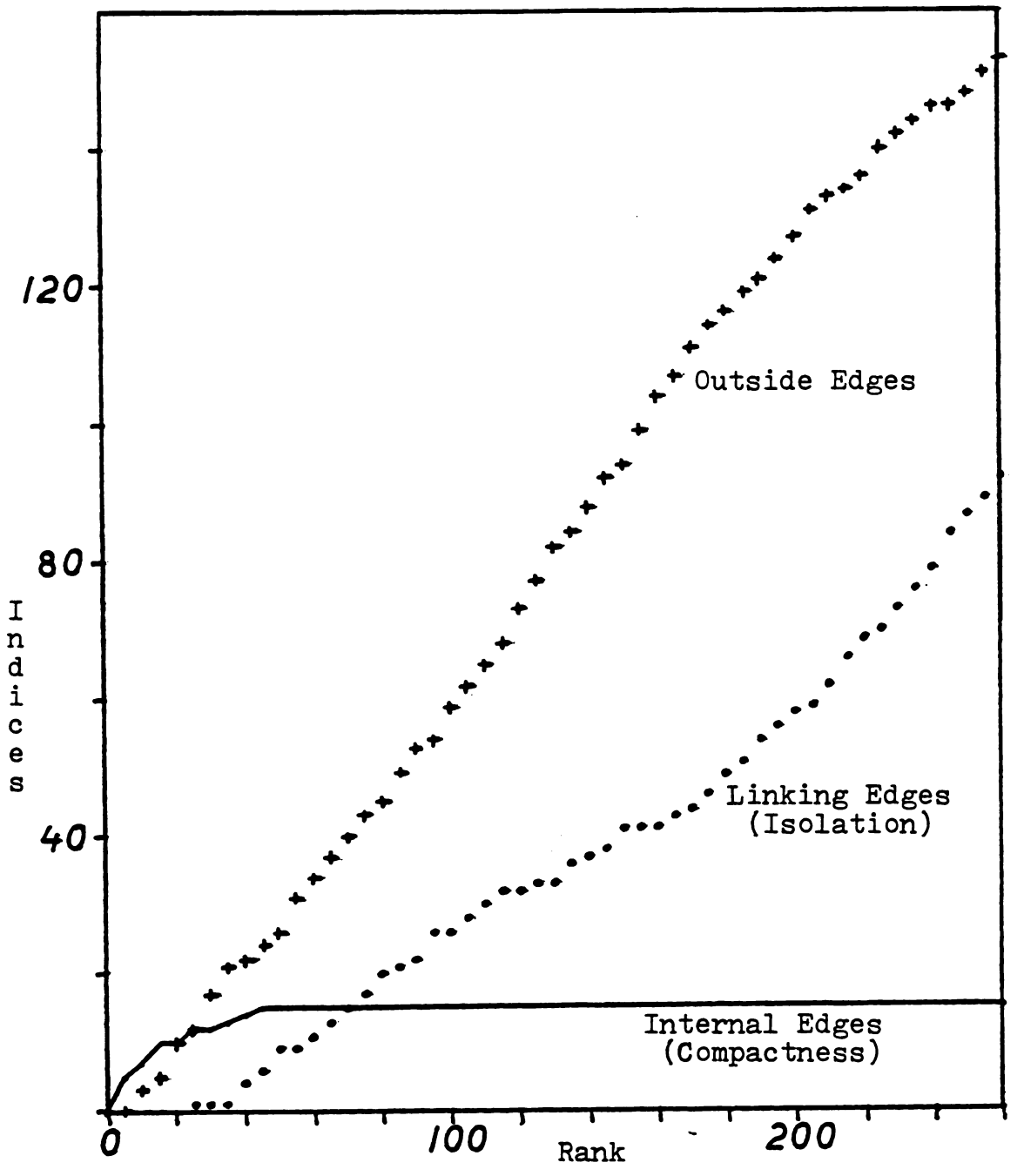
Figure 2.1. An Artificial 25-point Data Set

Figure 2.2 gives the raw profiles for the proposed cluster. The compactness index (internal edges), the isolation index (linking edges), and the edges which have no connection with the cluster (outside edges) are plotted for every fifth rank. The compactness index rises rapidly to its maximum value of 15, indicating a relatively compact cluster. The isolation index is quite small when the number of edges in the threshold graph is small, then rises at a fairly constant rate until the threshold graph is complete at 300 edges. Tests which quantify the significance of the rapid rise in the compactness index and the slow rise in the isolation index are developed in the next section.

2.2. Probability Profiles

A "probability profile" is a refinement of a raw profile and is formed by computing, for each rank, the probability, p, that the validity index for a subset of nodes in a random graph would be as good as the observed index. This significance level, p, is used as a measure of cluster validity at each rank. The sequence of measures forms the probability profile.

The validity measures proposed here are based on distributions over random graphs as defined in Section 1.3.2. Section 2.2.1 discusses several of these distributions. Sections 2.2.2 and 2.2.3 develop the



Plots of the two validity indices and the number of outside edges for the six-point cluster of Figure 2.1.

Figure 2.2. Raw Profiles of a Cluster

expressions for the probabilities required to compute validity measures. Section 2.2.4 discusses the application of the probability profiles and Section 2.2.5 applies probability profiles to an example.

2.2.1. Distributions Based on Random Graphs

One possible distribution is the set of validity indices found by letting the sampled population contain all possible subsets of nodes of fixed size in all random graphs. This "complete" distribution provides a proper evaluation of a cluster if the cluster has been defined a priori, without reference to the proximities or the features used to find the proximities. In most cases, this type of distribution is not useful because the cluster has been chosen to maximize the validity indices. Such a cluster is almost certain to have better validity indices than a randomly chosen subset of points.

The distribution used to test validity must somehow take into account the special way in which the cluster to be tested was formed. One way to do this is to restrict the sampled population to subsets of nodes of fixed size in the random rank graph which are recognized as clusters by the clustering method, CM, being used. By restricting the population in this manner, the clustering method is used to pick out good subsets of nodes from each random graph for inclusion in the null distribution. This

"CM-reachable" distribution is a "limited" distribution, a distribution computed on the assumption that only selected subsets of nodes in each random graph are included in the sampled population. Baker and Hubert [BAK76] use the "Complete-Link-reachable" distribution to develop their test of cluster validity. A test of cluster validity based on a reachable distribution is specific to clusters found by the clustering method used to determine the subsets of the random graph.

Another type of limited distribution assumes the validity index rather than the clustering method limits the sampled population. A "best case" distribution assumes the sampled population is restricted to a subset of nodes of fixed size in each random graph which produces the optimum value of the validity index. A best case distribution has the advantage of being applicable to clusters formed by any clustering method. In a sense, the best case distribution is the least upper bound on all the reachable distributions. For any random graph, a subset of points which is included in the best case sampled population has validity index at least as good as any subset included in a reachable sampled population. A test of cluster validity based on a best case distribution is a "general test" in that it can be applied to clusters formed by any clustering technique.

2.2.2. Distributions for the Isolation Index

In this section two probability distributions for the cluster isolation index defined in Section 2.1.2 are presented. The null hypotheses which form the bases for the distributions are defined using random graphs. One of the distributions is a best case distribution.

Let G be a labeled graph with n nodes and N edges (a labeled (n,N) graph). Let A be a k-node subset of the nodes of G which is to be tested for compactness and isolation. Let <A> be the subgraph of G induced by the node set A.

Let e be the number of edges in <A>. If e' is the number of edges in <-A> (where -A is the subset of nodes not in A), then let b = N-e-e' be the number of linking edges, or edges in G which join a node in A to a node in -A.

The following two probability distributions are used to evaluate the number of linking edges, b, as an index of isolation.

Theorem 1.

The probability, P(b|k), that a random labeled (n,N) graph has at least one subset of k nodes with b or fewer linking edges is bounded above by

$$P(b|k) <= \binom{n}{k} \frac{\sum_{B=0}^{b} \binom{k(n-k)}{B} \binom{n:2 - k(n-k)}{N - B}}{\binom{n:2}{N}} .$$

The notation i:j is used for the binomial coefficient, the number of ways of choosing j unordered items from a set of i distinguishable items. If i<j or j<∅ then i:j = ∅. This convention simplifies the notation.

Proof:

For a particular subset A of k nodes there are k(n-k) pairs of nodes in G at which linking edges can be placed. The number of ways in which B of these edges can be chosen is

$$\binom{k(n-k)}{B} .$$

Since the remaining N-B edges must be chosen from the non-linking possibilities, the number of labeled (n,N) graphs in which A may have b or fewer linking edges is

$$\sum_{B=0}^{b} \binom{k(n-k)}{B} \binom{n:2 - k(n-k)}{N - B} .$$

There are n:k ways to pick a particular subset of k nodes. If all such possible subsets are considered, the set of labeled (n,N) graphs contains

$$\binom{n}{k} \sum_{B=\emptyset}^{b} \binom{k(n-k)}{B} \binom{n:2 - k(n-k)}{N - B}$$

subsets of k nodes which have b or fewer linking edges.

The number of labeled (n,N) graphs which have at least one k node subset with b or fewer linking edges is bounded above by the number of k-node subsets with b or fewer linking edges which appear in all labeled (n,N) graphs. Thus, dividing by the number of labeled (n,N) graphs gives an upper bound on the probability that a random labeled (n,N) graph contains a subset of k nodes with b or fewer linking edges.

End of proof.

The probability that a random (n,N) graph contains at least one subset of order k with b or fewer linking edges is equal to the probability that the subset of order k in a random (n,N) graph with the fewest linking edges has b or fewer linking edges. Thus, the probability we have bounded is from a best case distribution. This probability bound, which we call measure I1, is a measure of the isolation of a cluster with isolation index b.

A test of cluster validity may be based on measure I1. Under the null hypothesis for the test, the isolation index has the best case distribution derived above. Measure I1 is an upper bound on the size of the test for

rejection of this null hypothesis. The technique for determining cluster validity proposed in this thesis involves judging the validity on the basis of the sequence of measures, or test sizes, over the evolution of the rank graph. Thus we do not propose a significance level for this test at one rank. At a particular rank, small values of Il are evidence of a valid cluster, while large values are not.

The form of the above result suggests a close connection to the hypergeometric distribution. We now develop the bound on P(b|k) using the hypergeometric model.

In the hypergeometric model [BRO65] we have a population of R elements, of which D are defective. We draw a sample of size W, without replacement. The probability that our sample contains exactly X defectives is

$$p(X) = \frac{\binom{D}{X}\binom{R-D}{W-X}}{\binom{R}{W}} .$$

Consider a specific k-node subset of a graph with n nodes. The R elements in our population are the n:2 node pairs of the graph. The D defectives are the k(n-k) linking node pairs for the subset. The sample of size W is the set of N node pairs chosen as edges in the random

graph. The probability that the sample contains exactly B defectives, or linking edges, is then

$$p(B) = \frac{\binom{k(n-k)}{B}\binom{n:2 - k(n-k)}{N - B}}{\binom{n:2}{N}} .$$

The probability that at least one of the k-node subsets in the random graph has b or fewer linking edges is bounded above by the probability that our specific subgraph has b or fewer linking edges times the number of ways in which the specific subgraph could be chosen,

$$P(b|k) <= \binom{n}{k} \sum_{B=\emptyset}^{b} p(B) .$$

This is the same as our previous expression.

An exact expression for the probability P(b|k) would be much more satifactory than the present result. However, the problem is made very difficult by the high degree of interaction among the various subgraphs of each possible random graph. An exact expression for the desired probability is not known, and we must settle for the upper bound obtained, which ignores interactions among the subgraphs.

Consider k-node subsets of the nodes of labeled (n,N) graphs which have exactly e internal edges (call these

(k;e) subsets of the graphs). We compute the probability, P(b|k,e), that a subset chosen at random from all (k;e) subsets of all labeled (n,N) graphs has b or fewer linking edges.

Theorem 2.

A labeled (n,N) graph with a (k;e) subset has N-e edges to distribute between linking edges and external edges. Thus

$$P(b|k,e) = \frac{\sum_{B=0}^{b} \binom{k(n-k)}{B} \binom{(n-k):2}{N-e-B}}{\binom{k(n-k)+(n-k):2}{N-e}} .$$

The sum is over the first b+1 terms of the probability mass function of the hypergeometric distribution. This time, in contrast to the result for P(b|k), the expression for the probability is exact. However, a test based on this result is not a general test since the sampled population is not limited to the best subset of nodes in each random graph. For this "fixed compactness index" distribution, the sampled population includes all k-node subsets with exactly e internal edges and is a third way of limiting the sampled population. This probability, which we call measure I2, is also a measure of the isolation of a cluster with b linking

edges. Again, we may base a test of cluster validity on this measure. Under the null hypothesis, the isolation index has the fixed compactness index distribution given above. Measure I2 is the size of the test for rejection of this null hypothesis.

2.2.3. Distributions for the Compactness Index

The following two probability distributions concern e, the number of internal edges.

Theorem 3.

The probability, P(e|k), that a random labeled (n,N) graph has at least one subset of k nodes with e or more internal edges is bounded above by

$$P(e|k) \;<=\; \binom{n}{k} \; \frac{\sum_{E=e}^{k:2} \binom{k:2}{E} \binom{n:2 - k:2}{N - E}}{\binom{n:2}{N}} \; .$$

Proof:

Consider a particular subset of k nodes. It has k:2 pairs of nodes. The number of ways in which E of these pairs can be chosen as internal edges is

$$\binom{k:2}{E} .$$

The number of ways in which the remaining N-E edges can be placed in the graph as linking or external edges is

$$\binom{n:2 - k:2}{N - E} .$$

Thus

$$\sum_{E=e}^{k:2} \binom{k:2}{E} \binom{n:2 - k:2}{N - E}$$

is the number of ways in which a particular subset of k nodes may have e or more internal edges. If all possible subsets of k nodes are considered, the set of all labeled (n,N) graphs contains

$$\binom{n}{k} \sum_{E=e}^{k:2} \binom{k:2}{E} \binom{n:2 - k:2}{N - E}$$

subsets of k nodes which have e or more internal edges.

The number of labeled (n,N) graphs which have at least one k node subset with e or more internal edges is bounded above by the number of k-node subsets with e or more internal edges which appear in all labeled (n,N) graphs.

End of proof.

The same result can be obtained using the hypergeometric probability model by an arguement similar to that used in Section 2.2.2. The probability that a k-node subset with e or more internal edges occurs in a random (n,N) graph is equal to the probability that the

number of edges in the k-node subset with the largest number of internal edges is e or more. Thus, the probability we have bounded is from a best case distribution. This probability bound, which we call measure C1, is a measure of the compactness of a cluster with compactness index e.

The corresponding test of cluster validity uses the null hypothesis which assumes the compactness measure has the best case distribution given above. Measure C1 is an upper bound on the size of the test for rejection of this null hypothesis. If a cluster has small measure C1 at rank N, we may conclude that the cluster is compact at that rank.

Consider subsets of k nodes which have exactly b linking edges (call these (k;b) subsets). We compute the probability, P(e|k,b), that a subset chosen at random from all (k;b) subsets that exist in all labeled (n,N) graphs has e or more internal edges.

Theorem 4.

A labeled (n,N) graph with a (k;b) subset A has N-b edges to distribute between internal edges of A and external edges of A (internal edges of -A). Thus

$$P(e|k,b) = \sum_{E=e}^{k:2} \frac{\binom{k:2}{E}\binom{(n-k):2}{N-b-E}}{\binom{k:2 + (n-k):2}{N - b}} .$$

Again, the sum is over terms of the probability mass function of the hypergeometric distribution and we have an exact probability rather than an upper bound. However, we do not have a best case distribution since the sampled population includes many subsets from some labeled (n,N) graphs and no subsets from others. This "fixed isolation index" distribution uses a sampled population which is limited to all k-node subsets with exactly b linking edges. This probability, measure C2, is also a measure of the compactness of a cluster with e internal edges. Measure C2 is the size of a test for rejection of the null hypothesis that the compactness index has the fixed isolation index distribution given above.

2.2.4. Application of Probability Profiles

The measures developed in Sections 2.2.2 and 2.2.3 comprise the probability profiles of a cluster. At each rank in the sequence of threshold graphs defined from the proximity matrix, the indices of isolation and compactness are evaluated. The probability that the value obtained, or some better value, would occur in a random graph is calculated using the best case and fixed index

distributions. Thus, at each rank, we form four measures (I1, I2, C1 and C2) of the validity of the cluster. The measures are the sizes of tests of cluster compactness and isolation under the random graph null hypothesis. If the null probability of an index at least as good as the observed index is large, we have evidence that the cluster is not valid at the rank tested. If the probability is low, we have evidence that the cluster is unusual, and therefore valid, at the rank tested. If the probability is low over a span of many ranks, we conclude that the cluster is valid.

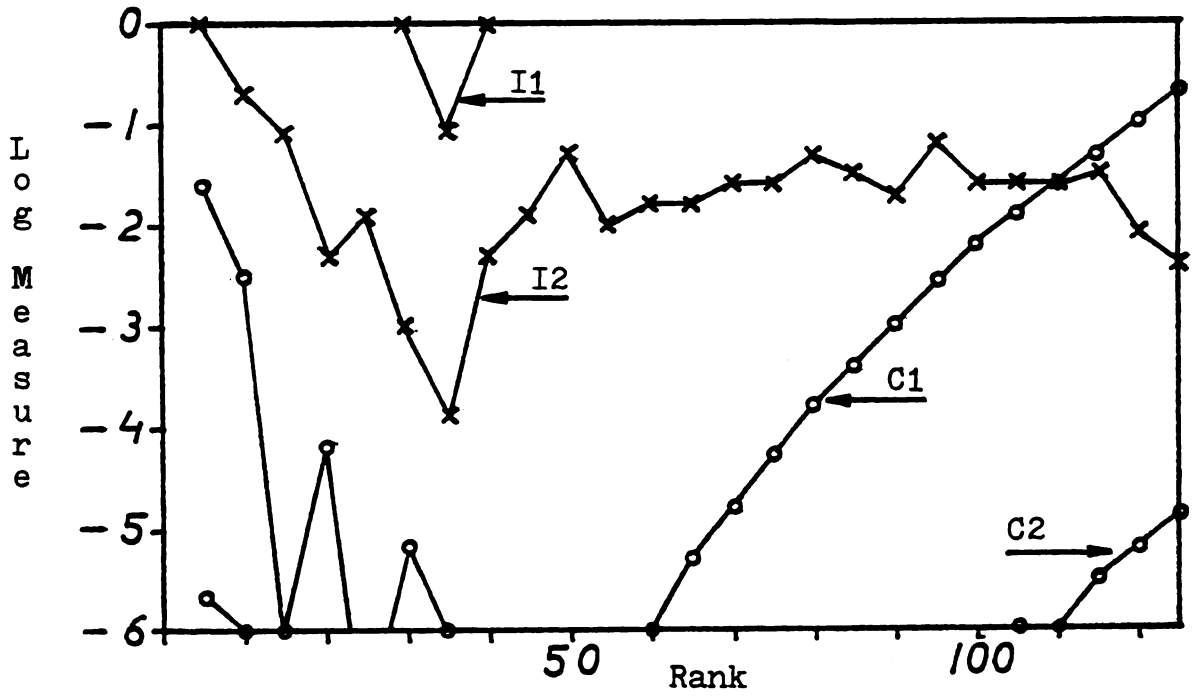
The most favorable results are the simultaneous occurrence of low values, say less than 10 ** (-3), for both measure I1 and measure C1 over a wide range of ranks. However, it may happen that a subset of points has low measure for one validity index, say compactness, but not for the other, isolation, when using the best case distributions. In this case we consider the fixed index distributions developed in Sections 2.2.2 and 2.2.3. If a subset is compact with respect to the best case distribution, we ask whether it is isolated with respect to the fixed compactness index distribution. That is, we ask if both measure C1 and measure I2 are small over a wide range of ranks. If so, we may conclude that the subset is compact and, for a subset with its compactness, it is isolated. Similarly, if a subset is isolated with

respect to the best case distribution, we ask whether it is compact with respect to the fixed isolation index distribution. If it is, we may conclude that the subset is isolated and, for a subset with its isolation, it is compact. In other words, the subset of points forms a valid cluster.

If both measure I1 and measure C1 are large, the subset is neither isolated nor compact with respect to the best case distributions, and we have no justification for using the fixed index distributions. In this case we conclude that the subset of points does not form a valid cluster.

2.2.5. An Example

Figure 2.3 shows a logarithmic plot of isolation measures I1 and I2 and compactness measures C1 and C2 for the 6-node subset of Figure 2.1. The four measures are plotted for every fifth rank of the sequence of threshold graphs. The probability profiles show that this cluster is unusually compact, but not isolated, when compared against the most compact (measure C1) and most isolated (measure I1) subsets of a random graph. When compared with subsets of the same compactness, this cluster is unusually isolated (measure I2). From this observation, we conclude that, under the random graph null hypothesis, this subset of 6 points forms a valid cluster.



Logarithmic plots of the four validity measures for the six-point cluster of Figure 2.1.

Figure 2.3. Probability Profiles of a Cluster

2.3. Accuracy of Bounds for P(e|k) when e = k:2

In Section 2.2 we developed upper bounds on the best case distributions for the indices of isolation and compactness. The upper bounds are useful measures only if they are close to the actual values. If the upper bounds are much larger than the actual probabilities, tests of compactness and isolation based on them will be very conservative and the probability profiles will show very few low values. In the extreme, an upper bound of 1.00 is always available, but it is of no use in judging cluster validity.

In this section we investigate the accuracy of the upper bound for a special case of the compactness index. Several results on the probability p(k) that a random labeled (n,N) graph contains at least one complete subgraph of order k are presented. This is a special case of the probability P(e|k), studied under Theorem 1 in Section 2.2.3, that a random labeled (n,N) graph contains a subset of k nodes with e or more internal edges. In this case e = k:2 and the subset has all its internal edges.

2.3.1. Upper and Lower Bounds

The probability p(k) that a randomly chosen labeled (n,N) graph contains at least one complete subgraph of order k is simply the number C(k) of labeled (n,N) graphs

which contain at least one complete subgraph of order k divided by the total number of labeled (n,N) graphs. Thus

$$p(k) = C(k) \; / \; Q$$

where

$$Q = \left(\begin{matrix} n:2 \\ N \end{matrix} \right) .$$

The quantity C(k) may be expressed in terms of the clique number of a graph. A clique is a maximal complete subgraph and the clique number of a graph is the order of the largest clique in the graph. Thus C(k) is the number of labeled (n,N) graphs with clique numbers greater than or equal to k.

Let C(k,r) be the number of labeled (n,N) graphs which contain exactly r distinct complete subgraphs of order k. These r subgraphs may overlap.

Note that

$$Q = \sum_{r=\emptyset}^{R} C(k,r)$$

and

$$C(k) = \sum_{r=1}^{R} C(k,r)$$

where R is the maximum number of complete subgraphs of order k which can occur in a graph with N edges.

Bounds on C(k) will first be expressed in terms of the number of complete subgraphs of order k and the number

of pairs (both from the same graph) of complete subgraphs of order k in the set of all labeled (n,N) graphs.

Let S(k,1) be the number of complete subgraphs of order k in the set of all labeled (n,N) graphs.

$$S(k,1) = \sum_{r=1}^{R} r\, C(k,r)$$

Comparing S(k,1) and C(k) gives

$$S(k,1) - C(k) = \sum_{r=1}^{R} r\, C(k,r) - \sum_{r=1}^{R} C(k,r)$$

$$= \sum_{r=2}^{R} (r-1)\, C(k,r) >= \emptyset \quad .$$

Thus S(k,1) is an upper bound on C(k) and

$$p(k) <= S(k,1) / Q \quad .$$

We proceed to find a lower bound. If a labeled (n,N) graph contains r complete subgraphs of order k, then there are r:2 ways to choose a pair of complete subgraphs of order k from the graph. The total number of pairs (from the same graph) of complete subgraphs of order k which occur among all labeled (n,N) graphs is

$$S(k,2) = \sum_{r=2}^{R} (r:2)\, C(k,r) \quad .$$

Comparing C(k) and the difference S(k,1) - S(k,2) we see that

$$C(k) - S(k,1) + S(k,2)$$

$$= \sum_{r=2}^{R} (r:2)\, C(k,r) - \sum_{r=2}^{R} (r-1)\, C(k,r)$$

$$= \sum_{r=3}^{R} ((r-1):2)\, C(k,r) >= 0 \quad .$$

Thus S(k,1) - S(k,2) is a lower bound on C(k) and

$$p(k) >= (S(k,1) - S(k,2)) / Q \ .$$

We could proceed in this manner, counting triples, quadruples, etc., of complete subgraphs of order k and developing tighter bounds on C(k). However, the value of this development depends on being able to calculate the bounds.

The number S(k,1) may be found by considering each specific subset of k nodes in turn. Let A be a fixed subset of k nodes. The number of labeled (n,N) graphs in which A induces a complete subgraph is

$$\left(\begin{matrix} n:2 - k:2 \\ N - k:2 \end{matrix} \right)$$

since k:2 edges are used to form the complete subgraph of order k and the remaining edges may be used to join any of the remaining pairs of nodes.

Since there are n:k such subsets of k nodes, we have

$$S(k,1) = \binom{n}{k} \left(\begin{matrix} n:2 - k:2 \\ N - k:2 \end{matrix} \right) .$$

Thus

$$p(k) <= \frac{\left(\begin{matrix} n \\ k \end{matrix} \right) \left(\begin{matrix} n:2 - k:2 \\ N - k:2 \end{matrix} \right)}{\left\langle \begin{matrix} n:2 \\ N \end{matrix} \right\rangle} .$$

This is a special case of the upper bound on P(e|k) given in Section 2.2.3.

The derivation of an expression for S(k,2) is similar to that for S(k,1), though the details are more complex. Let A and B be different subsets of k nodes. The number of labeled (n,N) graphs in which both A and B induce complete subgraphs depends on the overlap between the two subsets. Let m be the number of nodes the two sets share. Then u = k-m is the number of unshared nodes in each subset. Note that $\emptyset <= m < k$.

As in the case of S(k,1), we first find the number of labeled (n,N) graphs in which a specific subset of nodes, A union B, induces a pair of complete subgraphs of order k. Since the induced graph has (2*(k:2) - m:2) edges,

the result we need is

$$\left(\begin{matrix} n:2 - 2\,(k:2) + m:2 \\ N \;\; - 2\,(k:2) + m:2 \end{matrix} \right) .$$

The number of ways in which the set of 2k-m nodes may be chosen can be developed in several ways. For example, first choose the m nodes to be shared. This may be done in n:m ways. From the remaining nodes choose u unshared nodes to complete subset A, which can be done in (n-m):u ways. Finally choose another u nodes to complete B, in (n-m-u):u ways. Since the order in which the two sets are completed is not significant, we divide by two. The result is

$$\binom{n}{m}\binom{n-m}{u}\binom{n-m-u}{u}\Big/ 2 = \binom{n}{m,\,u,\,u}\Big/ 2 .$$

Thus,

$$S(k,2) = \sum_{m=0}^{k-1} \frac{\binom{n}{m,\,u,\,u}}{2} \left(\begin{matrix} n:2 - 2\,(k:2) + m:2 \\ N \;\; - 2\,(k:2) + m:2 \end{matrix} \right) .$$

We could proceed to calculate the number of triples of complete subgraphs of order k in the set of (n,N) graphs. However, instead of one parameter to define the overlap, we would now have four, implying that the number of cases to be considered would increase as the fourth power of k. Thus, consideration of triples appears to be computationally infeasible.

2.3.2. Asymptotic Forms for the Bounds

This section examines the asymptotic behavoir of the bounds, developed in Section 2.3.1, on the probability p(k) that a random labeled (n,N) graph contains at least one complete subgraph of order k. The results indicate that p(k) tends to be close to the upper bound whenever the upper bound is small.

In the limit of large graphs the upper and lower bounds on p(k) take on simple forms. We will show that the upper bound takes on the form

$$u(k) = \frac{S(k,1)}{Q} \quad \text{---->} \quad s = \frac{n^k \, a^{k:2}}{k!}$$

where

$$a = N \, / \, n:2 \quad .$$

We will present evidence that the difference between the upper and lower bounds takes on the form

$$r(k) = \frac{S(k,2)}{Q} \quad \text{---->} \quad 1/2 \; s^2 \quad .$$

We consider the asymptotic case where n, N and k all grow without bound with a and s held constant. The requirement that a be held constant ensures that the number of edges (N) in the random graph increases as the number of nodes (n) increases so as to keep the proportion of edges which are present constant. The requirement that s be held constant means that the size (k) of the subset

of nodes under consideration also increases as n increases, though not nearly as rapidly as n. By holding s constant, k increases in such a way that the probability of occurrence of a complete subgraph of order k approaches s.

We derive the first result as follows. With a and s held constant as n, N and k increase, we have

$$u(k) = \frac{S(k,1)}{Q} = \frac{\binom{n}{k}\binom{n:2 - k:2}{N - k:2}}{\binom{n:2}{N}}$$

$$= \binom{n}{k} \frac{(n:2 - k:2)!\ N!}{(n:2)!\ (N - k:2)!} .$$

Using

$$\frac{t!}{(t-v)!} \longrightarrow t^{v} \quad \text{as} \quad t/v \longrightarrow \text{infinity},$$

and the definitions of a and s,

$$u(k) \longrightarrow \binom{n}{k} \frac{N^{k:2}}{(n:2)^{k:2}} = \binom{n}{k} a^{k:2}$$

$$\longrightarrow \frac{n^{k}\, a^{k:2}}{k!} = s .$$

The definition of s specifies the relationship between n and k as they become very large. Starting with this expression we may write k as a function of n as follows. Start with

$$s = \frac{n^k \, a^{k:2}}{k!} \quad .$$

Taking the logarithm of both sides and expanding ln k!, we obtain

$$\ln s \sim k \ln n + k:2 \ln a - k \ln k + k - \ln (2 * 3.14159...) / 2 - \ln (k) / 2 \; .$$

Now divide by k and drop terms which go to zero as n (and k) get very large. The remaining terms are

$$\emptyset \sim \ln n + (k-1) \ln (a) / 2 - \ln k + 1.$$

Now let a' = - 2 / ln a and multiply by a' to obtain

$$\emptyset \sim a' \ln n - k + 1 - a' \ln k + a' \quad .$$

Thus, since k << n,

$$k \dashrightarrow a' \ln n - a' \ln (a' \ln n) + a' + 1 \quad .$$

The terms which are ignored are of order (ln (ln n) / ln n) or (ln k / k). This expression for k is (a' ln 2) larger than the result obtained by Matula [MAT77] for the expected value of the largest clique size in a random edge graph with edge probability a. Note that s does not appear in the final result. The important point is that s is constant.

A partial proof of the second result has been found. Confidence in the correctness of the result is enhanced by

observing the behavior of r(k) / u(k) ** 2 as k increases. The plots in Figure 2.4 show, for several different values of s and a, the convergence to a value of 1/2. The curves were calculated by fixing the values of a and s, then for each subgraph size (k) calculating the required number of nodes (n) and edges (N). The exact upper and lower bounds on the probability of a complete subgraph of order k in a random labeled (n,N) graph were then calculated and used to determine the desired ratio. Convergence to 1/2 occurs for relatively small values of k.

Using this result, the asymptotic difference between the bounds is one-half the square of the upper bound. Thus, for the special case e = k:2, the upper bound on P(e|k) = p(k) is quite close to the actual value if the upper bound is small and k is large enough. For u(k) = .1, the lower bound approaches .1 - (.1)(.1)/2 = .Ø95 for large k. The relationship of the upper and lower bounds is exactly the relationship expected if the objects of interest, the complete subgraphs of order k, are distributed at random among the set of all labeled (n,N) graphs. We may hope that the objects of interest in the calculations of P(e|k) and P(b|k) are also distributed at random for large k, so that the upper bounds are tight bounds whenever they are small.

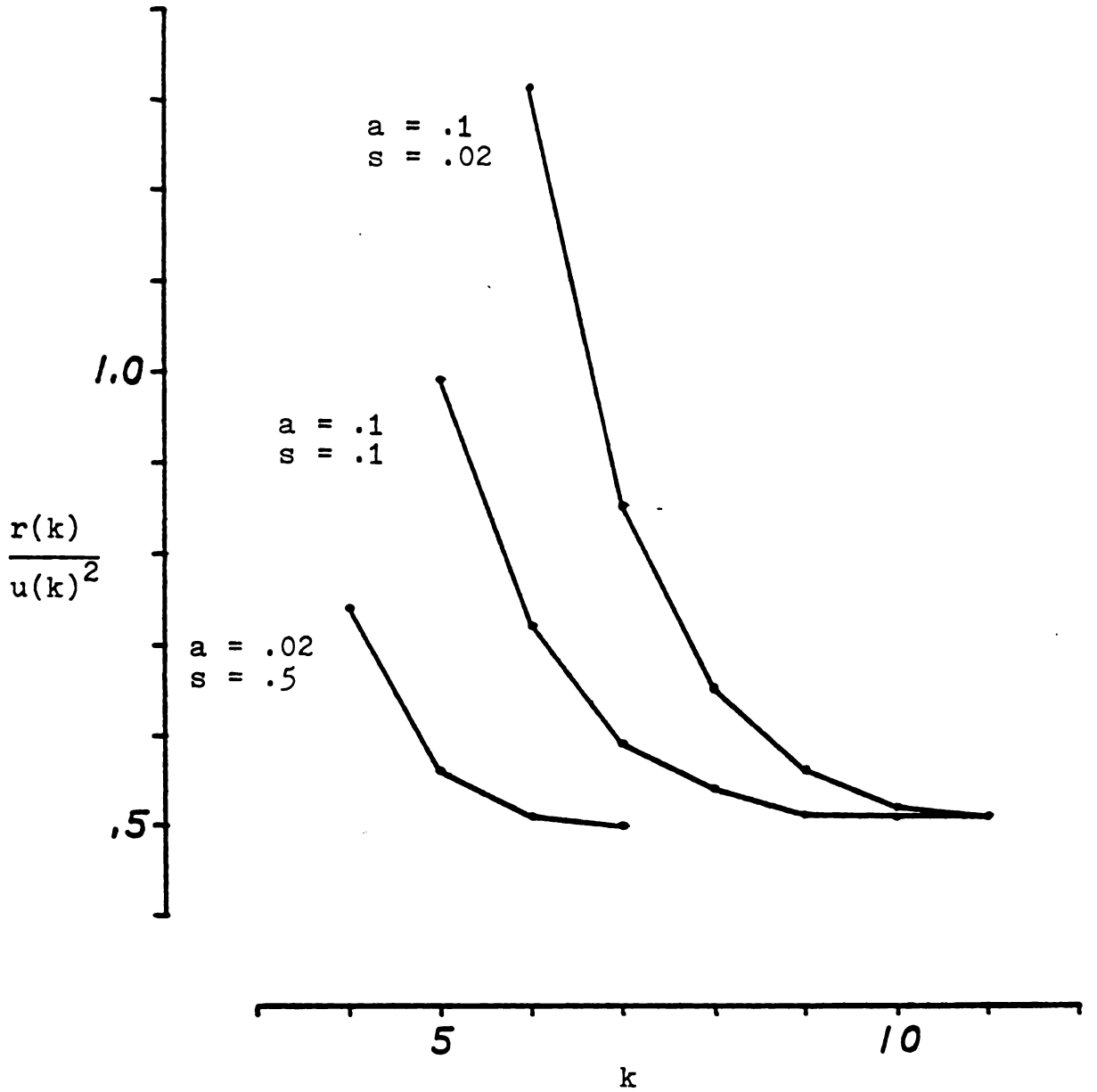


Figure 2.4. Asymptotic Behavior of r(k) / u(k) ** 2

2.4. Summary

This chapter develops two Cluster Profiles, defined from a proximity matrix using a sequence of threshold graphs, which describe the interaction of a subset of points with its environment. The Raw Profile of a cluster is the sequence of cluster validity indices observed for the sequence of threshold graphs. Computationally inexpensive indices of compactness and isolation for any subset of nodes in a threshold graph are defined. The Probability Profile of a cluster is the sequence of measures -- the test sizes for a cluster validity test applied at all ranks -- for the sequence of threshold graphs. Validity tests under the Random Graph Null Hypothesis for the indices of compactness and isolation are developed. We argue that the distributions based on the best case sampled populations provide the most useful distributions for validity tests. Using these distributions, upper bounds on the cumulative distribution functions for the validity indices are developed. These upper bounds form two measures of cluster validity.

The cumulative distribution functions for the validity indices using a second sampled population are also developed. This sampled population includes all subsets of points with a specified validity index, either isolation or compactness. We argue that this distribution is useful for testing one validity index when the other

index is known to be valid under the best case distribution. The probabilities given by these two cumulative distribution functions form two additional measures of cluster validity.

The final section of the chapter explores the asymptotic accuracy of the upper bound on the best case distribution of the compactness index. For the special case in which the compactness index takes on its maximum value, we develop a lower bound on the cumulative distribution function of the index and show that the difference between the upper and lower bounds asymptotically approaches one-half the square of the upper bound. We conclude that the upper bound will be useful for testing cluster validity.

3. Intrinsic Dimensionality and Cluster Validity

This chapter explores some limitations of validity tests based on random graphs when the proximity matrix is derived from a pattern matrix. We investigate the effect this pattern matrix starting point has on distributions of indices of cluster validity.

3.1. Introduction

Ling [LIN73a] has applied his random-graph-based test of the lifetime of single link clusters to clusters found in a star map of sixty bright stars in the neighborhood of Polaris. He found several clusters with unusually long lifetimes and concluded that the clusters were valid. Such results, on data sets which are presumably random, have led several investigators, including Ling [LIN76, pg.294] and Matula [MAT77, pg.126], to warn against using cluster validity tests based on a null hypothesis of a random graph.

This chapter studies the relationship between the dimensionality of a set of data points and the distributions of indices of compactness and isolation used for testing cluster validity. We ask whether a prescribed dimensionality for the data set automatically precludes the use of tests based on random graphs.

Suppose that, instead of using random graphs as the null hypothesis of "no clustering", we assume that the

patterns themselves are randomly chosen points from a uniform distribution in a hypercube and compute the dissimilarity matrix with some distance metric. This random experiment provides a null hypothesis, called the "uniform hypercube null hypothesis," which may be more appropriate than the random graph null hypothesis if our data are patterns in a feature space. Other distributions from which the points could be chosen include the uniform distribution in a hypersphere and the multidimensional normal distribution. The uniform distribution in the hypercube is used here because it is the cheapest to simulate.

Suppose we use hypothesis tests based on the random graph null hypothesis to test clusters from data sets created by choosing points from a uniform distribution in a hypercube. It may be that 10% of the data sets exhibit clusters which are valid at the 10% level, or perhaps there are valid clusters in 40% of the data sets. Perhaps there are never any valid clusters by our test. Considerations such as these motivate the investigation below.

Section 3.2 derives analytic results for the case of four points placed at random on the unit interval and presents simulation results for five points placed at random in a unit hypercube. Section 3.3 examines a phenomenon discovered in the five point simulation of

Section 3.2 and relates it to an issue in the determination of the intrinsic dimensionality of a data set. Section 3.4 presents and discusses the results of a simulation study of the relationship between the random graph null hypothesis and the uniform hypercube null hypothesis for data sets of medium size. The results of this study provide empirical evidence which supports the note of caution suggested by Ling and Matula, namely, tests based on the random graph null hypothesis cannot be indiscriminately applied to data sets best represented as points in a multidimensional space.

3.2. Very Small Data Sets

It is easy to show that the uniform hypercube null hypothesis will generate distributions of clustering statistics different from those computed under the random graph null hypothesis for the special case when four points are chosen at random from a uniform distribution on the unit interval. Form a threshold graph by using distance as proximity and including only the edges joining the three closest pairs of points. The distributions of statistics based on the resulting unlabeled (4,3) graphs are certain to be different from those based on random labeled (4,3) graphs since one of the graphs which occurs for the random graph case, the graph where one of the nodes is adjacent to each of the other three nodes, cannot

occur when the four points lie on a line.

The distribution over the possible rank orders of interpoint distances for four points placed at random on the interval (Ø,1) is equivalent to that obtained when two points are placed at random on the interval and the end points, Ø and 1, are used as the other two points. Thus, the problem of finding distributions based on the possible rank orders of the proximities reduces to a two variable problem. Counting distributions over threshold graphs may be derived by considering a unit square representing all possible values of the two interior points. The square is divided into regions corresponding to the various possible rank orderings of the interpoint distances and the areas are calculated. The counting distributions over threshold graphs with various numbers of edges for four points chosen at random in a unit interval and for four-node random graphs are given in Table 3.1. For example, there are twenty possible random graphs with four nodes and three edges and sixteen of them are connected. However, four of the connected graphs cannot occur under the one dimensional hypothesis, and each of the twelve which can occur have only one-third the probability of occurrance of one of the unconnected graphs. Whenever there is more than one possible threshold graph the distributions under the two null hypotheses are different. The change in the distribution over threshold graphs affects the

Table 3.1 Frequency Distributions over (4,N) Graphs

Edges	Graph	Probability: Random Graph Null Hypothesis	Probability: One Dimensional Uniform Hypercube Null Hypothesis
0		1.	1.
1		1.	1.
2		.80	.67
		.20	.33
3		.60	.50
		.20	.50
		.20	0.
4		.80	1.
		.20	0.
5		1.	1.
6		1.	1.

distribution of the indices of compactness and isolation defined in Section 2.1.2. The best case distributions of the indices for three-node subgraphs are shown in Table 3.2. The distributions are different under the two hypotheses.

This analysis is possible for four points on a line because the problem reduces to a two-variable problem. With more points, or points in higher dimensions, the increase in the number of variables creates an intractable problem. Some insight into the five point case may be gained by noting which five-node graphs can occur if the points are restricted to a line. A theoretical analysis of the probabilities in this case is very tedious.

The approach taken here is to simulate the random placement of points in an interval and observe the counting distribution over the various graphs. This approach is easily extended to higher dimensions. Table 3.3 shows several counting distributions over the six possible threshold graphs with five nodes and five edges. The distributions were obtained by Monte Carlo simulation of the placement of five points in hypercubes with various dimensionalities of one through two hundred. For each dimensionality, 1000 sets of five points were obtained. For each set of five points the ten interpoint distances were calculated and the five smallest distances were used to define a (5,5) threshold graph. The random graph case

Table 3.2 Distributions of Validity Indices

e: Maximum number of internal edges for a 3-node subset
b: Minimum number of linking edges for a 3-node subset

Edges	e	b	Probability: Random Graph Null Hypothesis	Probability: One Dimensional Uniform Hypercube Null Hypothesis
2	1	1	.20	.33
	2	0	.80	.67
3	2	1	.80	.50
	3	0	.20	.50
4	2	2	.20	0.

Table 3.3 Counting Distributions over (5,5) Graphs

	1	2	3	4	5	6	T Statistic
Graph:							
Random Graph Null Hypothesis (Theoretical)							
	238.1	238.1	238.1	119.0	119.0	47.6	
Random Graph Null Hypothesis (Simulation)							
	251	233	218	131	118	49	3.76
Uniform Hypercube Null Hypothesis (Simulation)							
D = 1	560	87	0	353	0	0	1395.
D = 2	433	177	22	282	79	7	643.
D = 3	337	256	45	232	114	16	328.
D = 4	365	216	43	245	109	22	378.
D = 10	245	248	98	239	148	22	225.
D = 20	228	284	111	186	150	41	124.
D = 100	211	232	154	206	176	21	139.
D = 200	202	243	159	171	200	25	120.

was simulated by randomly choosing five of the ten possible node pairs as edges in the (5,5) threshold graph.

The observed simulation results, {O(j), j=1,6}, may be compared with the results expected under the random graph null hypothesis, {E(j), j=1,6}, using the statistic

$$T = \sum_{j=1}^{c} \frac{(O(j) - E(j))^2}{E(j)} ,$$

where c=6 is the number of classes. For large sample size and under the null hypothesis that the observed values are drawn from the distribution given by the expected values, T has the chi-squared distribution with five degrees of freedom. The mean of this distribution is 5.0 and the variance is 10.0. With one exception, the difference between the simulation distribution and the distribution expected under the random graph null hypothesis is extremely significant, with $p << .001$. The exception, as expected, is the direct simulation of the random graph case, for which the value T = 3.74 falls between the 40th and 50th percentiles.

The best case distributions for the indices of compactness and isolation are easily determined from the counting distributions over the threshold graphs. The optimum value for each measure on each possible graph is determined by inspection. For four-node subsets, graphs number 1, 2, 3 and 5 in Table 3.3 have a best case

compactness index of 4 and a best case isolation index of 1, while graph number 4 has best case compactness and isolation indices of 5 and 0, respectively, and graph number 6 has indices of 4 and 2. The counts for the various graphs are combined to produce the distributions for the indices. For example, for the three dimensional hypercube simulation, the counts for graphs 1, 2, 3 and 5 are added to give 752 graphs with a best case isolation index of 1. The results for four-node subsets are given in Table 3.4. The statistic, T (with c=3), is again used to test the goodness-of-fit of the simulation results to the expected distribution under the random graph null hypothesis. Under the null hypothesis and for large samples, T has a chi-squared distribution with two degrees of freedom, a mean of 2.0 and a variance of 4.0. Again, with the exception of the random graph simulation, all of the simulation distributions are significantly different from the expected distribution under the random graph null hypothesis. This provides additional support for the contention that cluster validity measures have significantly different distributions under the two hypotheses.

3.3. High Dimensionality in Small Data Sets

An interesting phenomenon appears in Table 3.3. Note that the distribution continues to change as the

Table 3.4 Distributions of Validity Indices

The counts are the numbers of (5,5) graphs (see Table 3.3) with the indicated best case validity indices on 4-node subsets.

Compactness Isolation	5 0	4 1	3 2	T Statistic
Random Graph Null Hypothesis (Theoretical)				
	119.0	833.3	47.6	
Random Graph Null Hypothesis (Simulation)				
	131	820	49	1.46
Uniform Hypercube Null Hypothesis (Simulation)				
D = 1	353	647	0	549.
D = 2	282	711	7	276.
D = 3	232	752	16	136.
D = 4	245	733	22	159.
D = 10	239	739	22	145.
D = 20	186	773	41	43.0
D = 100	206	773	21	82.8
D = 200	171	804	25	34.5

dimensionality of the hypercube is increased beyond four. This is counter-intuitive. Since any set of five points in Euclidean space can be used to determine a four dimensional space in which interpoint distances are maintained, intuition dictates that the distributions for all dimensions beyond three should be the same. The observed changes raise an interesting question concerning the nature of intrinsic dimensionality, which is developed below.

3.3.1. Definitions of Intrinsic Dimensionality

The intrinsic dimensionality of a set of points can be viewed in two ways. The first is that the intrinsic dimensionality is the order of the lowest dimensional space in which the data points can be embedded without changing the rank order of the interpoint distances [KRU64]. Given a matrix of interpoint distances for a set of n points, it is always possible to embed the n points in a space of n-2 dimensions without changing the order of the interpoint distances. In other words, all possible orderings of interpoint distances can be achieved in a space of n-2 dimensions using Euclidean distance as the proximity measure. Several embedding methods which attempt to decrease the dimensionality while maintaining the order of the interpoint distances, at least locally, have been proposed [BEN69, CHE74]. Methods based on

finding principal axes also attempt to find a small number of dimensions in which most of the information contained in the interpoint distances is retained. For n points in a high dimensional Euclidean space, it is always possible to embed the points in a space of n-1 dimensions while maintaining the interpoint distances. From this viewpoint, the intrinsic dimensionality of our five point data set cannot be larger than four. However, since the higher dimensional hypercubes do provide different distributions of graphs, either the viewpoint must be changed, or some additional information must be permitted in the description of the data.

A second viewpoint is that intrinsic dimensionality is the minimum number of variables needed to determine the spatial position of a point in the data set. This viewpoint was Bennett's [BEN69] motivation for his dimension-reducing algorithm. He wanted to find the number of free system parameters needed to generate a set of signals. A method for estimating the intrinsic dimensionality which is not based on finding an embedding has been proposed by Pettis, Bailey, Jain and Dubes [PET79]. If data points are known to lie along a curve in three dimensions, two of the coordinates of a data point may be determined from the third using the equations which define the curve. It takes only one variable to specify the position of a member of this data set, which is

intrinsically one dimensional. From this viewpoint, our two hundred dimensional hypercube defines a set of data points with intrinsic dimensionality of two hundred. Even though the five points are on a four dimensional hyperplane, each distinct set of five points establishes a different hyperplane. Thus the equations which define a fixed hyperplane cannot be used to determine additional coordinates of a data point once the four coordinates which specify its position in the hyperplane are known. All two hundred coordinates must be given.

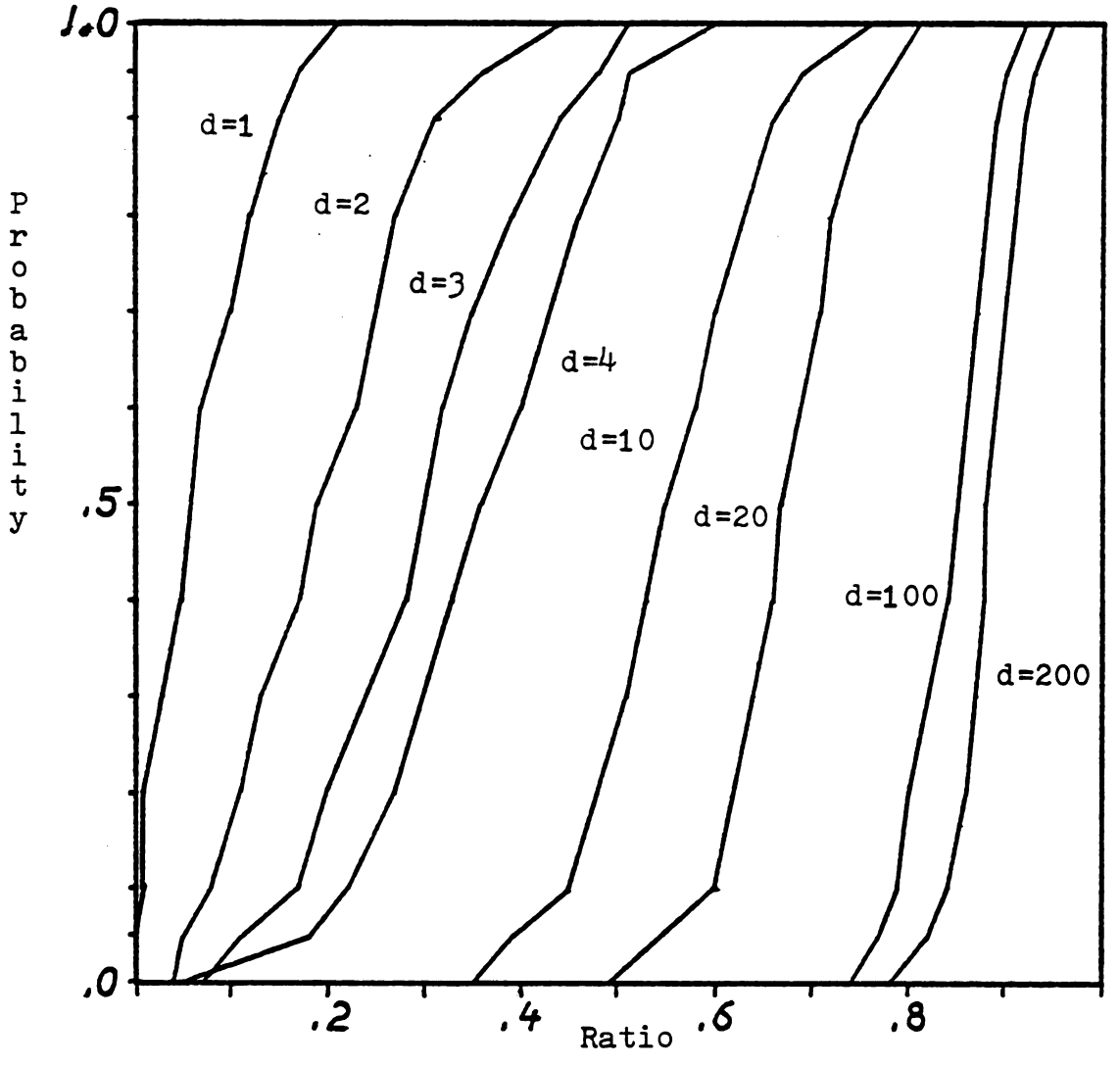
The first viewpoint of intrinsic dimensionality is useful if the problem at hand is to represent the data in a space of low dimensions. However, valuable information concerning the data is lost if the ability to represent the data using few dimensions is mistaken for the ability to determine the spatial positions of the data points by specifying only a few variables. If the data are a sample from some target population, the ability to embed the sample in a space of low dimensions does not mean the target population is of low dimensionality. On the other hand, the number of variables needed to determine the spatial position of a point in the sample should accurately represent the dimensionality of the target population. The second viewpoint is the more useful one if a basic description of the data is the goal. One

consequence of large dimensionality for a small number of points is developed in the next section.

3.3.2. Five Points in a Hypercube

One effect of intrinsic dimensionality higher than four on a set of five random points is seen in the set of curves in Figure 3.1. Each curve is computed from one hundred sets of five points. Each point consists of d numbers chosen at random from a uniform distribution on the unit interval. These d numbers are used as the d coordinates of a point in a hypercube. The ten interpoint distances are calculated and the ratio of the smallest to the largest is found. Graphed are the empirical cumulative distribution functions of this random variable. As the dimensionality increases the ratio becomes closer to one. In other words, it becomes more likely that all ten interpoint distances are close to the longest interpoint distance. As the dimensionality is increased to extremely large values, the distribution of the five random points approaches the state in which all interpoint distances are approximately the same.

Bennett [BEN69] notes a similar result for the distribution of points in a unit hypersphere. As the dimensionality of the hypersphere increases the distribution of interpoint distances approaches a delta



Cumulative distributions of the ratio of shortest to longest interpoint distances for 100 sets of five random points in a d-dimensional hypercube.

Figure 3.1. Interpoint Distance Ratios for Five Points

function located at square root (2.Ø). That is, almost all pairs of data points are the same distance apart.

The distribution of interpoint distances depends on the underlying distribution of the data points. Since all interpoint distances are approximately equal for both the uniform hypersphere and the uniform hypercube, we conjecture that any uniform distribution bounded by a convex hull will exhibit the same effect. We also conjecture that the same effect occurs for the multidimensional Gaussian distribution, and perhaps for any distribution which is unimodal.

When all the interpoint distances are approximately the same, restrictions on the occurrence of some threshold graphs, which are so evident in the one dimensional hypercube case, disappear. The lengths of the edges are randomly ordered and all possible threshold graphs are equally likely. Thus, at extremely large dimensionalities, the uniform hypercube null hypothesis produces distributions over threshold graphs which are approximately equivalent to those under the random graph null hypothesis. If the distributions were equal at very low dimensionality, then validity tests based on the random graph null hypothesis could be applied in situations where a null hypothesis based on random pattern distributions is appropriate.

The simulation reported in this section indicates

that in the asymptotic case of very high dimensionality the uniform hypercube null hypothesis and the random graph null hypothesis produce identical distributions over threshold graphs. This asymptotic identity does not apply in practical situations, where the dimensionality is typically much less than 100. For data sets of five points it is apparent that the random graph null hypothesis should not be used to check the validity of clusters in a data set for which a uniform distribution of points in a hypercube of low dimensionality is an appropriate null hypothesis.

3.4. Data Sets of Medium Size

The evidence concerning data sets with four and five points presented in Section 3.2, while certainly suggestive, does not rule out the possibility that the random graph and uniform hypercube null hypotheses may, for larger data sets, be more or less equivalent with respect to the indices and measures of validity defined in Chapter 2. The following simulation experiments are designed to shed some light on this question. We simulate the creation of random graphs under the random graph and uniform hypercube null hypotheses and find best case distributions for the indices of compactness and isolation defined in Chapter 2. We must resort to simulation because calculation of the needed distributions under the

uniform hypercube null hypothesis by analytic methods is intractable for more than four points. The inspection technique used in Section 3.2 to find the best case must also be abandoned, both because the number of possible graphs increases rapidly with the number of nodes and because finding the best case by inspection becomes impossible for large graphs.

3.4.1. Simulation of the Distributions

The simulations which find empirical best case distributions for the uniform hypercube model are performed as follows. Choose a set of n points from a uniform distribution over a d-dimensional hypercube and determine the N closest pairs of points. These N pairs define an (n,N) threshold graph. Find the optimal values of the indices of compactness and isolation over all k-node subsets in this threshold graph. Record the indices. Repeating this many times will build up the probability distributions for the indices.

Bias is introduced into the distributions because it is not computationally possible to find the optimal values for the validity indices over all k-node subsets in each threshold graph. The huge number of k-node subsets in a graph of n nodes, (n:k), precludes an exhaustive search. The simulations reported here use a gradient ascent technique to find subsets of nodes with good values for

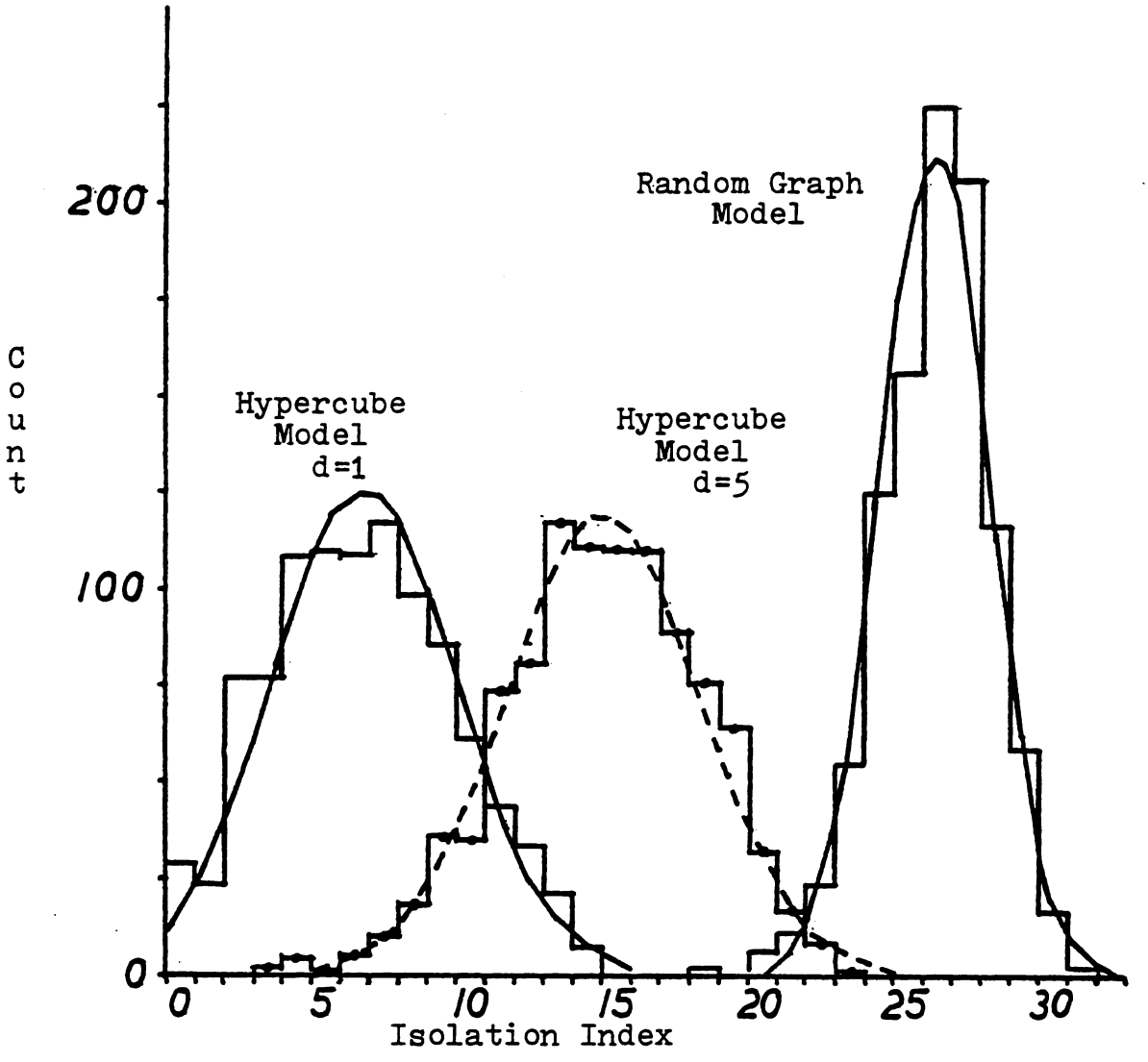
each index of validity. The technique is explained in the Appendix.

Simulations which find empirical best case distributions for the random graph model are run in a similar fashion. For these simulations the threshold graph is created by randomly choosing as edges N pairs of points from the n:2 equally likely possible pairs. Simulation of the random graph model can be used to check the accuracy of the technique for finding the best case k-node subset and the accuracy of the upper bound given by the theory of Chapter 2. The empirical best case distributions for the random graph model will match the upper bounds on the best case distributions, developed in Sections 2.2.2 and 2.2.3 under the random graph null hypothesis, only if both are accurate.

Simulations to find the best case distributions for the cluster validity indices were run for a variety of graph sizes, ratios of edges to node pairs, subset sizes and dimensionalities. The graphs are of medium size, the smallest having 20 nodes and the largest, 40. The 40 node size was chosen because the data studied in Chapter 4 consists of 40 samples. The first set of simulations was used to find the shapes of the empirical distributions of the validity indices. The main feature of the first set is the large number of random data sets used for each run of the simulation. Best case distributions for the

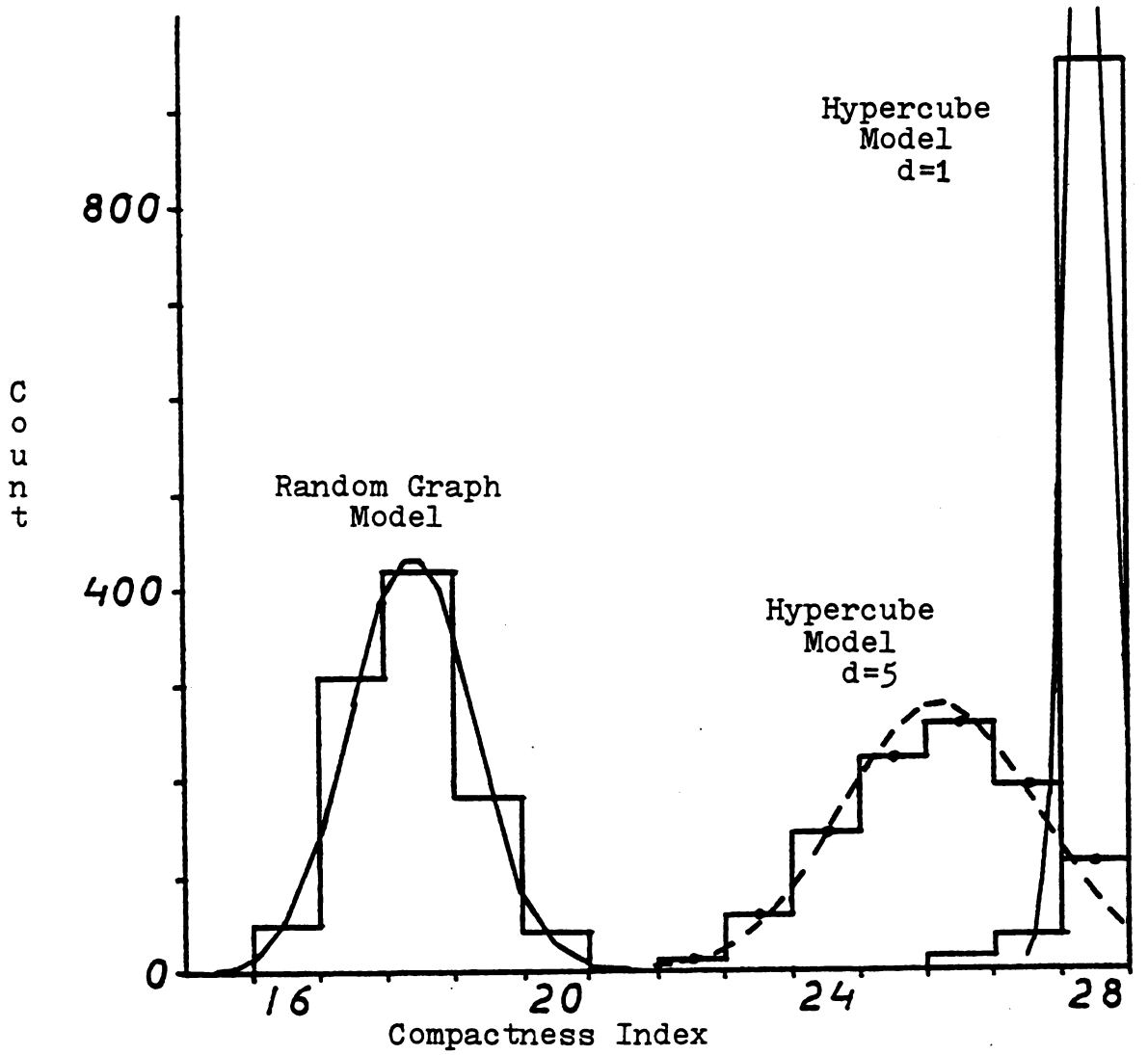
compactness and isolation indices were obtained for graphs with 40 nodes and 156 edges, subsets with 8 nodes and dimensionalities of one and five for the uniform hypercube model. The same graph and subset sizes were used to find best case distributions for the random graph model. The 156-edge case was chosen because both the compactness and isolation indices have extended ranges for this edge number. Each simulation was repeated for 1000 random data sets. Histograms of the results are given in Figures 3.2a,b. We note that the random graph and one dimensional histograms do not overlap for either index. Gaussian distributions with the same mean and variance as each distribution are also plotted in Figures 3.2a,b. The histograms appear Gaussian in shape, except for the one dimensional compactness histogram which is strongly affected by the upper limit of 28 on the compactness index. With more edges in the graph, distributions for the compactness index are strongly affected by the upper limit for dimensionalities higher than one. With fewer edges in the graph, distributions for the isolation index are affected by the lower limit of 0.

The main feature of the second set of simulations is the large assortment of different dimensionalities used. Runs were made for the uniform hypercube model with dimensionalities of 1, 2, 3, 5, 10 and 20 and for the random graph model. Simulation runs of 25 data sets each



Histograms and fitted Gaussian curves for best case isolation indices of 8-node subsets in 1000 random (40,156) graphs.

Figure 3.2a. Histograms of Validity Index Distributions



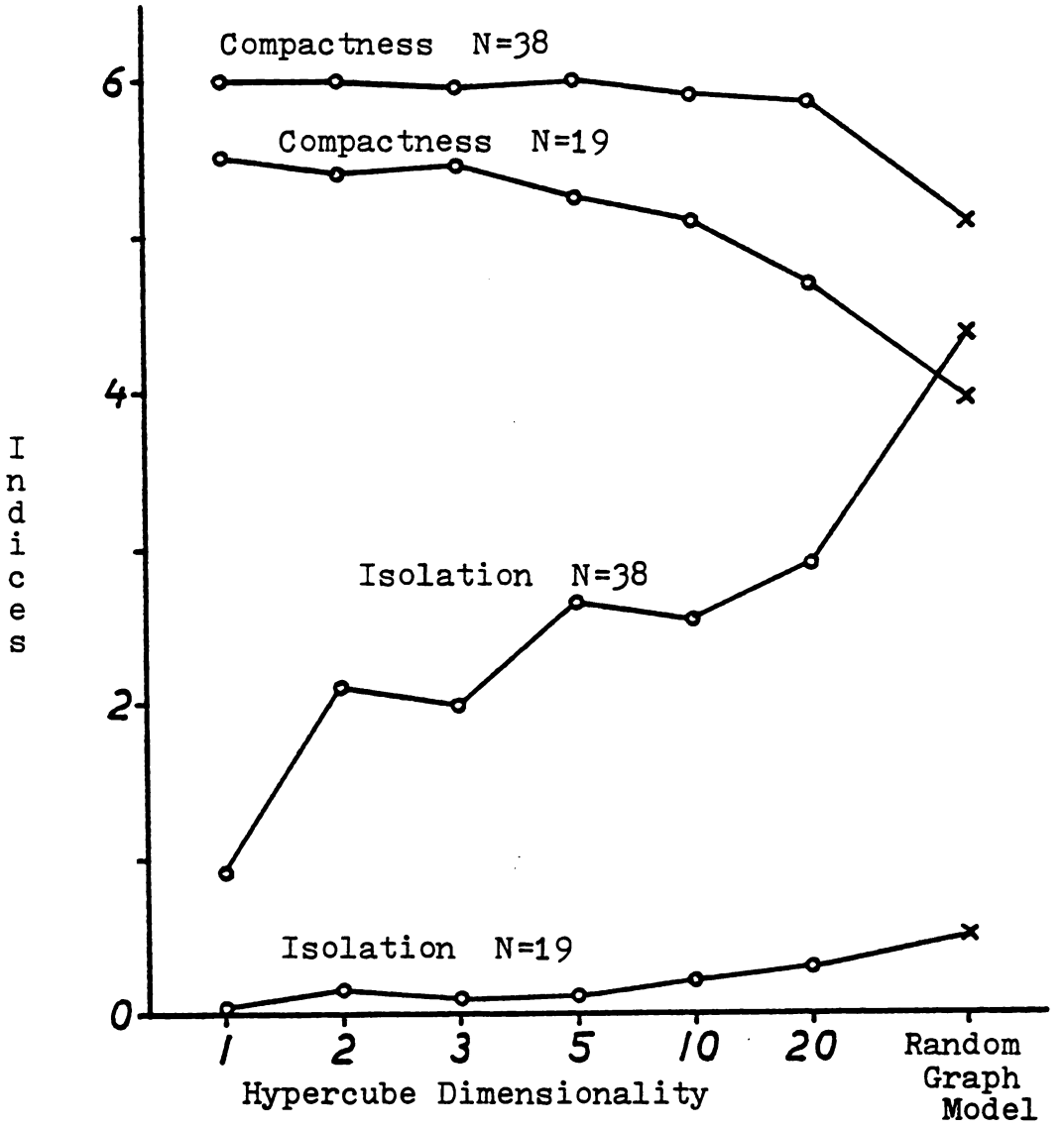
Histograms and fitted Gaussian curves for best case compactness indices of 8-node subsets in 1000 random (40,156) graphs.

Figure 3.2b. Histograms of Validity Index Distributions

were made for threshold graphs of 40 nodes with 8-node subsets and for threshold graphs of 20 nodes with 4-node subsets. Each run was repeated with ten and twenty percent of the node pairs present as edges. Figures 3.3a,b are plots of the means of the best case compactness and isolation distributions for the various runs. Figures 3.4a,b show the empirical cumulative distribution functions for the (40,156) threshold graphs for the various runs. The upper bounds on the cumulative distributions derived in Sections 2.2.2 and 2.2.3 under the random graph null hypothesis are also plotted.

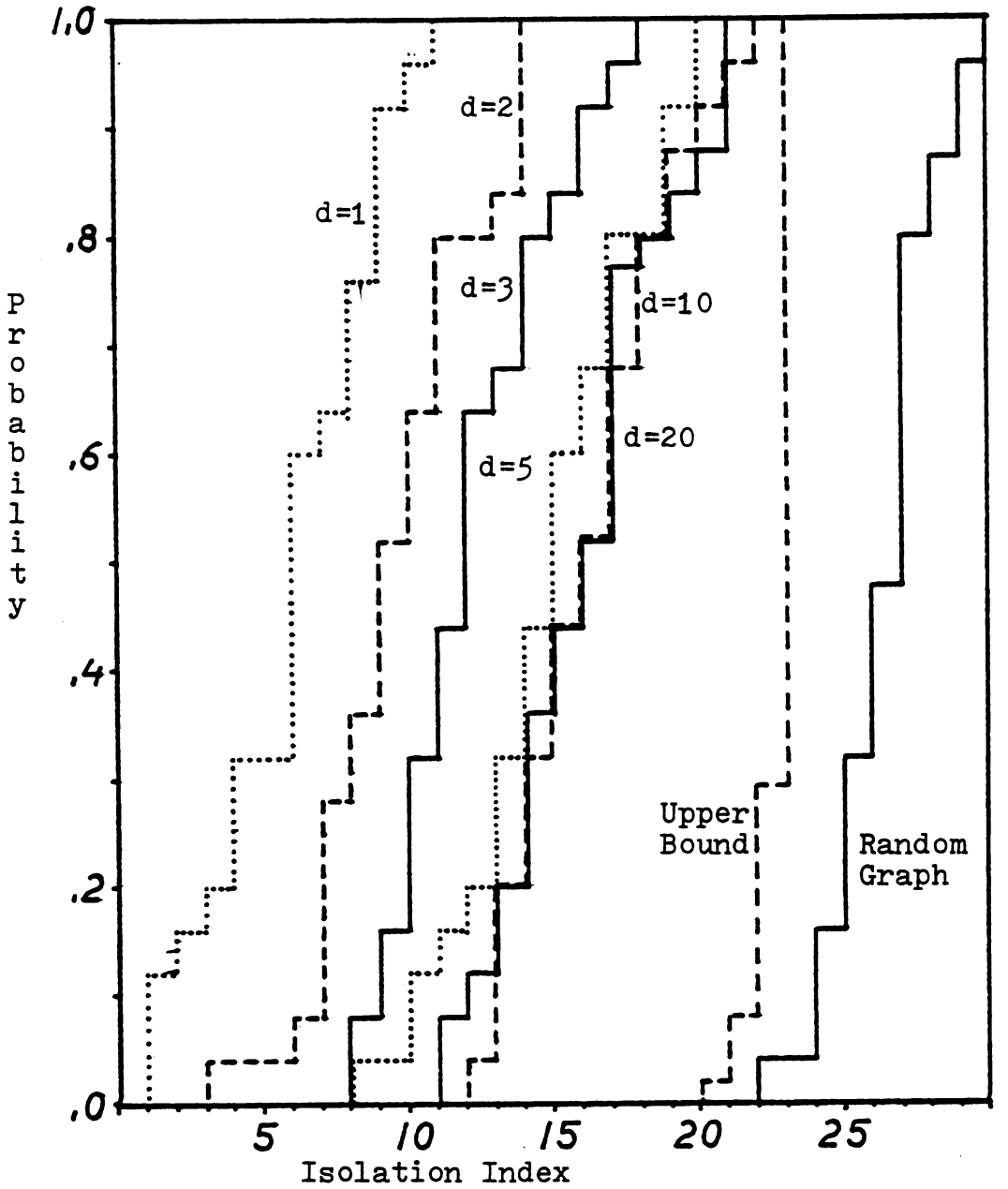
The third set of simulations uses the random graph model, and can be used to check the accuracy of the theoretical upper bounds and the best case approximation algorithm in the simulation. Plots of the empirical cumulative best case distributions of the compactness and isolation indices are presented in Figures 3.5a,b. Each plot represents 250 random data sets of 25 nodes. The distributions were determined for a 6-node cluster in graphs with 25, 50, 100, and 200 edges. Also plotted are the theoretical upper bounds on the best case distributions calculated from the expressions in Sections 2.2.2 and 2.2.3.

The results of the fourth set of simulations can be used to evaluate the cluster, in Figure 2.1, used as the example of Chapter 2. This simulation uses the two



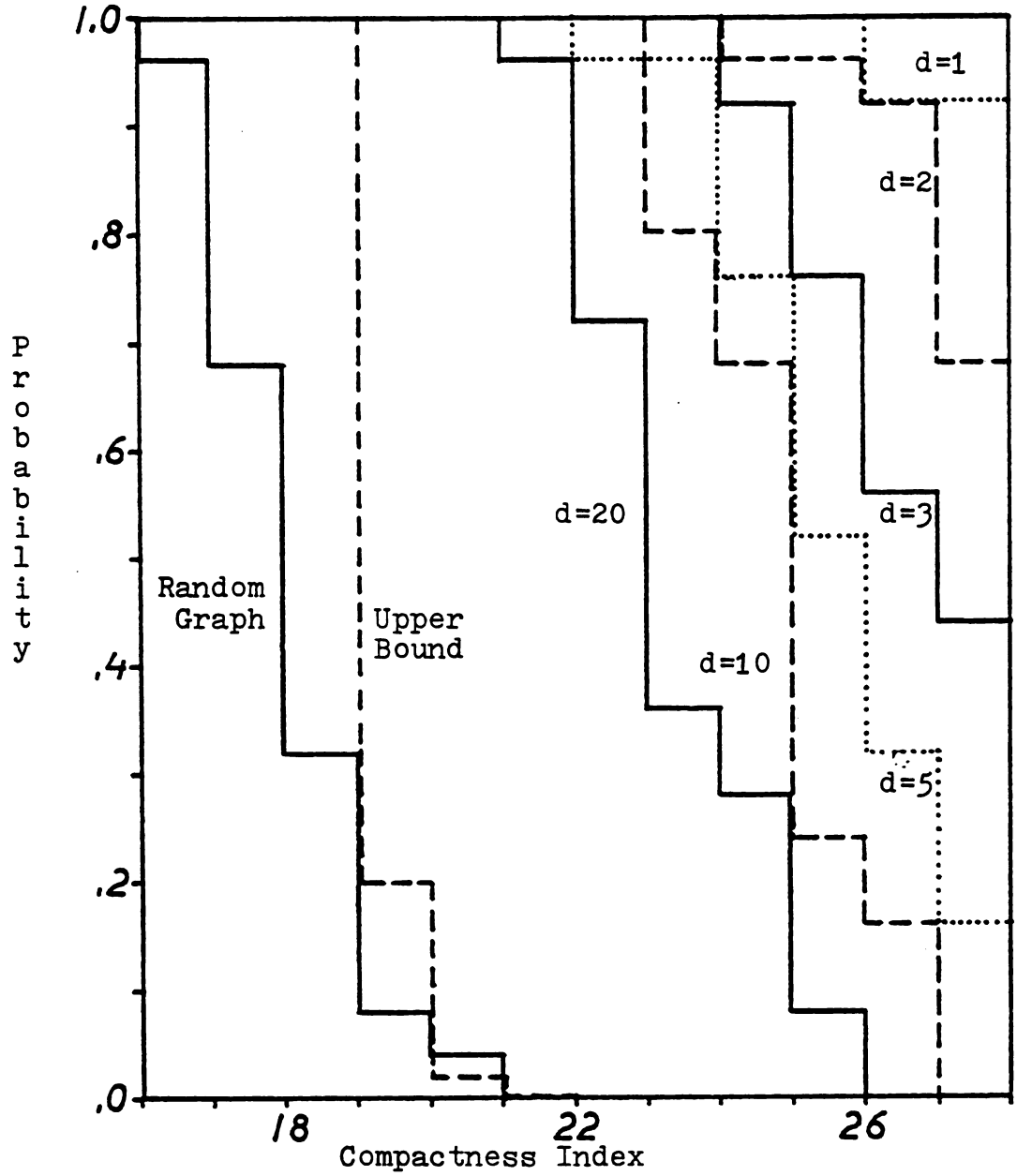
Means of the best case validity indices of 4-node subsets in 25 random (20,N) graphs.

Figure 3.3a. Means of Validity Index Distributions



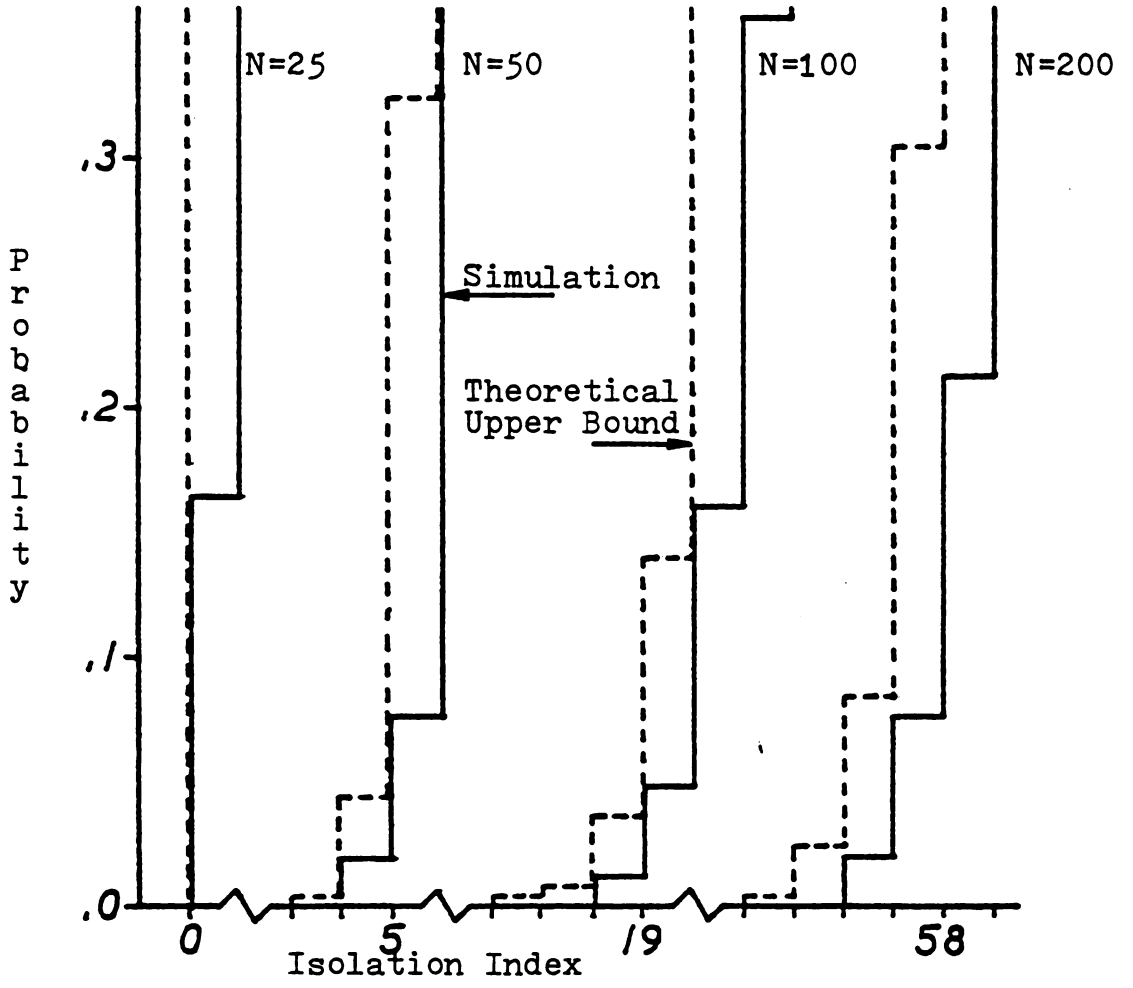
Cumulative distribution functions for the best case isolation indices of 8-node subsets in 25 random (40,156) graphs.

Figure 3.4a. CDF's of Validity Indices



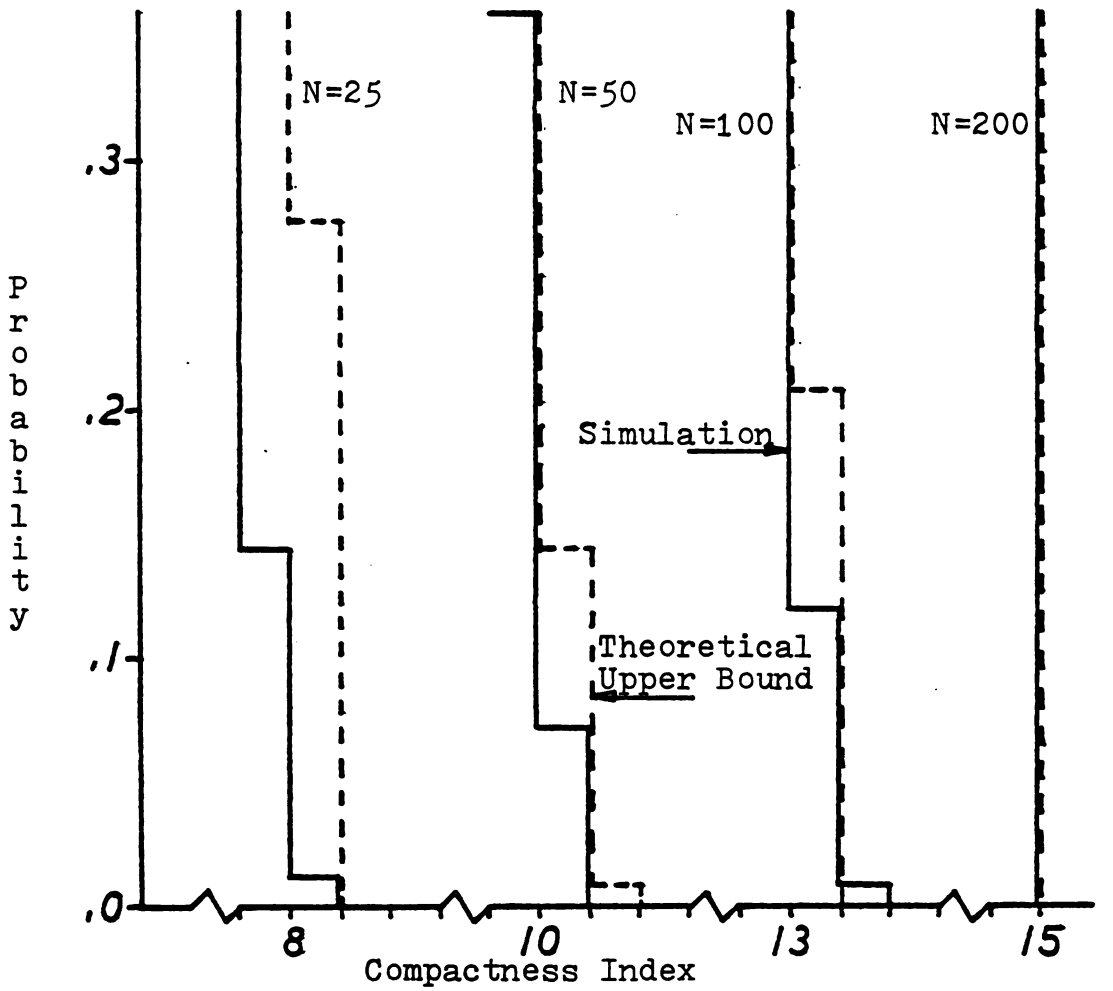
Cumulative distribution functions for the best case compactness indices of 8-node subsets in 25 random (40,156) graphs.

Figure 3.4b. CDF's of Validity Indices



Cumulative distribution functions of the best case isolation indices for 6-node subsets in 250 random (25,N) graphs for the Random Graph model, along with the theoretical upper bounds under the Random Graph Null Hypothesis.

Figure 3.5a. CDF's with Theoretical Upper Bounds



Cumulative distribution functions of the best case compactness indices for 6-node subsets in 250 random (25,N) graphs for the Random Graph model, along with the theoretical upper bounds under the Random Graph Null Hypothesis.

Figure 3.5b. CDF's with Theoretical Upper Bounds

dimensional uniform hypercube model with 250 data sets for each run. Runs were made for a data set of 25 nodes and a 6-node cluster with 25, 50, 75, 100 and 200 edges. Table 3.5 lists the numbers of linking and internal edges observed in the profiles of Figure 2.3 along with measures I1 and C1, the calculated upper bounds on the probabilities under the random graph null hypothesis, and the equivalent measures obtained from this simulation of the best case distributions under the two dimensional uniform hypercube null hypothesis.

3.4.2. Evaluation of the Results

We draw several observations from the results. First, the first and second sets of simulations, Figures 3.2, 3.3 and 3.4, show that the effect of low dimensionality is to shift the distributions of the indices of compactness and isolation away from the distributions under the random graph null hypothesis. This effect was observed in every simulation which was run. The direction of the shift, to higher compactness values and lower isolation values, is such that tests based on the random graph null hypothesis will give results which indicate that almost any set of points chosen at random from a uniform distribution in a hypercube of low dimensionality contains "valid" clusters.

Table 3.5 Evaluation of the Cluster of Figure 2.1

Edges	Observed Index	Measure 1: Random Graph (Theoretical Upper Bound)	Measure 1: Two Dimensional Uniform Hypercube (Empirical)
Compactness		C1	"C1"
25	12	< 1.00E-6	21/250 = .084
50	15	< 1.00E-6	62/250 = .248
75	15	5.3 E-5	232/250 = .928
100	15	5.84E-3	250/250 = 1.000
200	15	> 1.00	250/250 = 1.000
Isolation		I1	"I1"
25	1	> 1.00	250/250 = 1.000
50	9	> 1.00	250/250 = 1.000
75	17	> 1.00	250/250 = 1.000
100	26	> 1.00	250/250 = 1.000

If the bounds based on random graph theory are used to test the significance of clusters, too many significantly unusual values will be found.

As Figure 3.4 illustrates, the upper bounds developed in Chapter 2 do not bound the cumulative distributions of validity measures for points chosen from a hypercube of low dimensionality. For the lower dimensionalities under the uniform hypercube model, the entire set of indices obtained by simulation lies beyond the .01 probability level in the left tail of the upper bound.

The third set of simulations, Figure 3.5, shows that the cumulative distribution functions of the validity indices for the random graph model are close to the calculated upper bounds for the left tail of the cumulative probability. The discrepancies seen in Figure 3.5 are greatest for the poorer values of the indices, the situation in which the approximation to the optimum values of the validity indices is most likely to be inaccurate. Thus the theoretical upper bounds appear to be fairly close to the actual values of the cumulative probabilities, supporting the conclusions concerning the accuracy of the bounds reached in Section 2.3.

Finally, the two dimensional 25-node example of Sections 2.1.3 and 2.2.4 can be reevaluated using the results of the fourth simulation given in Table 3.5. The results for 25 and 50 edges cover the region in the

probability profiles (Figure 2.3) for which the measure Cl takes on its lowest values. The best case compactness measures under the two dimensional uniform hypercube null hypothesis are much larger than those found under the random graph null hypothesis. Under the random graph null hypothesis, the probability profiles of Figure 2.3, discussed in Section 2.2.4, indicate that the six-point subset is significantly compact at 25, 50 and 75 edges. Under the two dimensional uniform hypercube null hypothesis, the compactness measure is only low for 25 edges, and at p=.084 it is not significantly low. The best case measures for the observed isolation index remain high. Since the best case compactness measure and the best case isolation measure are both high, we do not need to look at the fixed index measures. Using the reasonable null hypothesis of points chosen at random from a uniform distribution in a unit square, this six-point subset is neither compact nor isolated.

3.5. Conclusions

The cumulative distributions of the two validity indices under the uniform hypercube null hypothesis do not match those obtained under the random graph null hypothesis. Figures 3.4a,b illustrate the fact that the upper bounds developed in Chapter 2 do not bound the distributions of the validity indices for points chosen

dimensionality. Thus, the probability profiles, developed in Section 2.3 using the random graph null hypothesis, should not be used to validate clusters if the patterns can be thought of as a set of points in Euclidean space. This is shown in Section 3.2 for small data sets of four and five points and in Section 3.4 for data sets of medium size, 20 to 40 points. Even if the space has dimensionality as high as 20, the structure imposed by embedding a moderate number of points in space shifts the distributions of these validity indices away from the distributions given by random graph theory.

The measures defined using random graph theory do provide a convenient base for evaluation of clusters drawn from a fixed experiment. The dimensionality of the data will be consistent from one cluster to another. Although the absolute significance of the measures will not be known, tentative conclusions regarding the relative validity of different clusters can still be drawn from the probability profiles. Experience gained on one data set can be carried over to other data sets representing the same type of data, since the probability structure of the measure calculations will account for changes in cluster size and for changes in the size of the data set.

4. Analysis of a Data Set

This chapter brings together the concepts developed in Chapters 2 and 3 in the analysis of a data set. The data set is clustered using several hierarchical clustering algorithms. The intrinsic dimensionality of the data set is determined using two different methods. The clustering tendency is tested. The validities of several clusters are determined using the lifetime measure of Ling and the profiles developed in Chapter 2.

4.1. Description of the Data

The data were obtained from Professor Oscar Tosi of the Department of Audiology and Speech Science at Michigan State University and consist of 40 samples of speech, 10 samples each from four different male subjects. Each subject read five different pieces of material while being recorded in two ways. The sound was recorded directly and, simultaneously, was transmitted over telephone lines and recorded. For each subject we have five direct and five phone samples. The recordings were then translated into choral speech [TOS75] and Fourier analyzed to determine the energy in each of 2048 frequency bands. Each sample was normalized so that the maximum feature value for that sample is 1.00. All 2048 features were used to calculate a proximity matrix with Manhattan distance, viz:

$$d(i,j) = \sum_{m=1}^{2048} \text{ABS}\ (\ F(i,m) - F(j,m)\ )\ ,$$

where F(i,m) is the m-th feature for the i-th sample. The data set for this study consists of the resulting 40 by 40 dissimilarity matrix.

4.2. Intrinsic Dimensionality of the Data Set

Two different methods were used to determine the intrinsic dimensionality of the data set. Kruskal's [KRU64] multidimensional scaling program (MDSCAL) was used to find a configuration of 40 points in a low dimensional space which preserved the order of the proximities. The two dimensional MDSCAL configuration is shown in Figure 4.1. The marked clusters will be discussed in Section 4.5. The ten samples for each subject are numbered consecutively with the direct samples first. For example, point 22 is the second direct sample for subject 3. The stress for this configuration, which is a measure of the amount of distortion introduced by embedding the data in a space of two dimensions, is .140 so we conclude that Figure 4.1 is a good representation of the proximity matrix.

The second method used to determine the intrinsic dimensionality of the data set is due to Pettis, Bailey, Jain and Dubes [PET79]. This method assumes that points

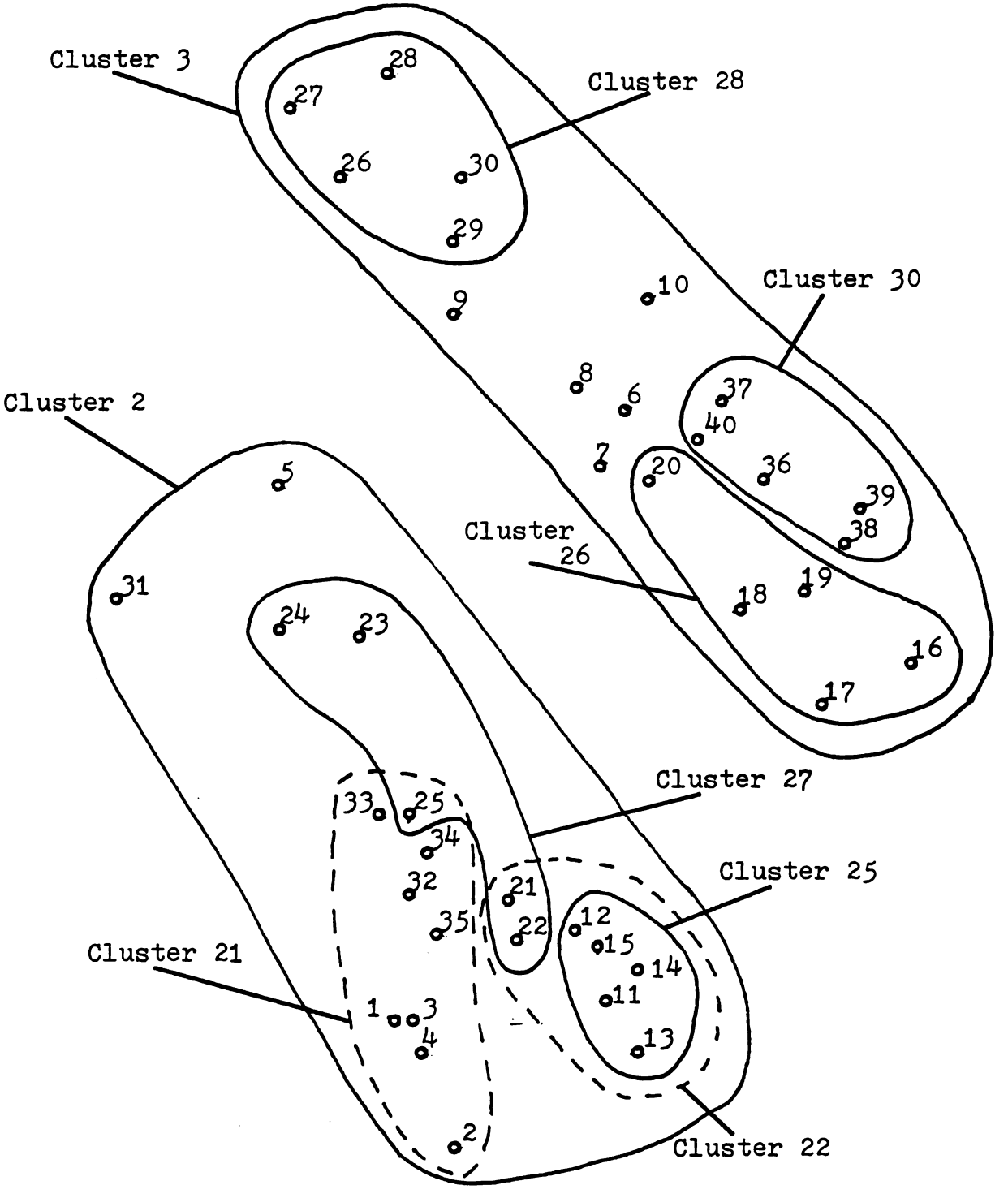


Figure 4.1. MDSCAL Configuration of the Tosi Data

are randomly chosen from a locally uniform distribution and uses nearest neighbor distances to estimate the dimensionality of the distribution. Using the two nearest and four nearest neighbors of each point, the estimates of the dimensionality are 38.6 and 31.4. With more neighbors the estimates become much smaller, dropping to 14.5 for sixteen neighbors. Since we have reason to believe this data set is organized into eight clusters of five points each, use of more than four nearest neighbors should cause the clustering structure to interfere with the estimate of the dimensionality. We therefore assume that the intrinisic dimensionality is best estimated by the nearest two to four neighbors.

The great discrepancy between the estimates of the intrinsic dimensionality by these two methods may be due to several factors. The methods differ in their assumptions regarding the scale of the proximities, which may affect the results. The MDSCAL program assumes only ordinal scale for the proximities. The Pettis technique, on the other hand, assumes the proximities have ratio scale.

A second interpretation of the discrepancy is that the two techniques are finding different things. The MDSCAL program searches for a configuration of points in a space of low dimensionality which is an accurate representation of the order of the proximities. The

Pettis technique estimates the number of variables necessary to specify the position of a data point in a local region. In the following analysis of cluster validity we assume the high value of the estimate of the intrinsic dimensionality found by the Pettis technique is correct. We further assume that the high intrinsic dimensionality permits use of the random graph null hypothesis as our hypothesis of "no clustering."

As Figure 3.3 shows, even with a dimensionality of 30 to 40, the distributions of the indices of compactness and isolation obtained using the Uniform Hypercube model are shifted away from those obtained using the Random Graph model. In the following analysis, the effects of this shift are partially offset by requiring rather low significance levels, 10**(-3) to 10**(-5), for our judgement of cluster validity. An alternative procedure would be to run the simulations for, say, a 38 dimensional Uniform Hypercube model. Since the simulation results would be needed for many cluster sizes and many ranks for each cluster size, the computing requirement is impractical.

4.3. Hierarchical Clusters of the Data Set

Three different clustering techniques were used to study the data set. The single link and complete link cluster hierarchies are shown in Figures 4.2 and 4.3.

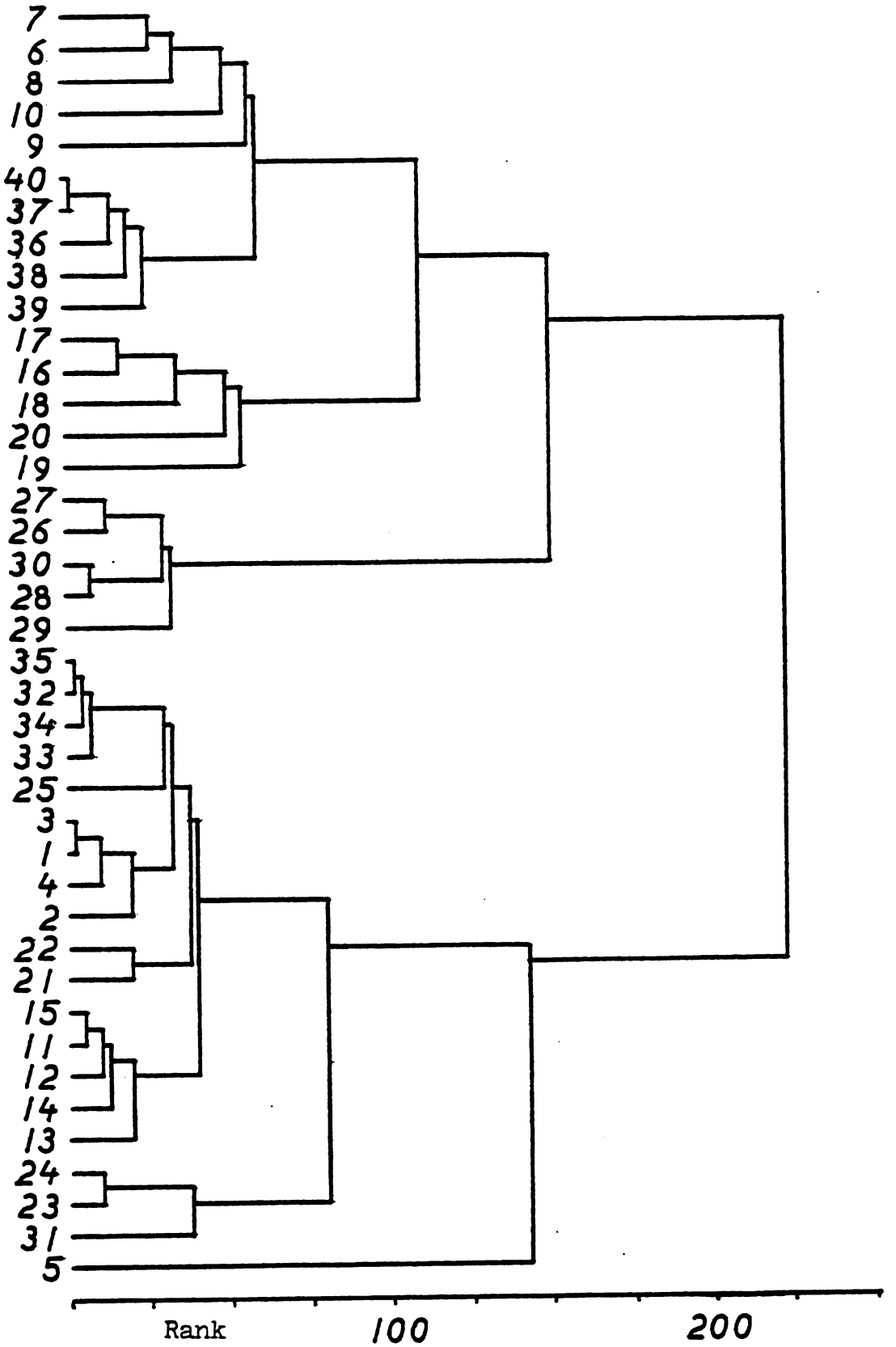


Figure 4.2. Single Link Dendrogram, Tosi Data

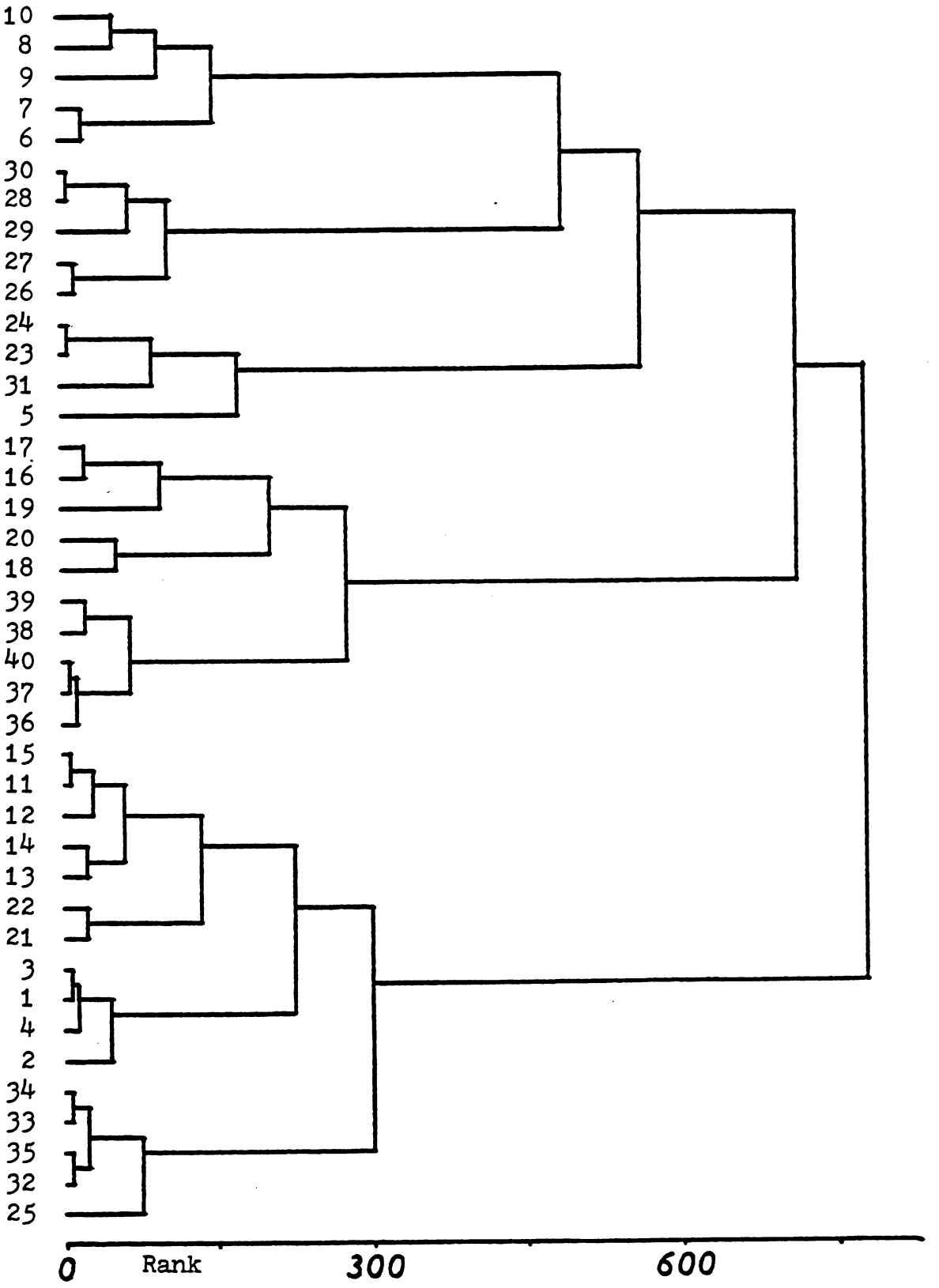


Figure 4.3. Complete Link Dendrogram, Tosi Data

Several groupings of five points representing single speakers are evident in the results. A clustering method defined by Ling [LIN72] was also applied to the proximity matrix. This method requires that each cluster be a connected subgraph with minimum degree k. The k = 1 clusters are identical to single link clusters. For values of k greater than one, k-clusters require a stronger internal connectedness than do single link clusters. The k-cluster hierarchy for k=4 is shown in Figure 4.4. Again, meaningful groupings of five points are evident in the results.

In addition to the striking appearance of five-point clusters in several of the clustering results, we note the split into two clusters of 20 points each which occurs in the single link hierarchy. An even split of this type is unusual when using the single link clustering method. The single link method tends to form one large cluster which then gradually absorbs the remaining points singly or in small clusters [BAK75]. In the next two sections tests of cluster validity based on the random graph null hypothesis will be applied to many of these clusters.

4.4. Connected Graph and Cluster Lifetime Tests

The first step in an analysis of the validity of clusters is a test of the clustering tendency of the data set. A statistic which can be used for this test is the

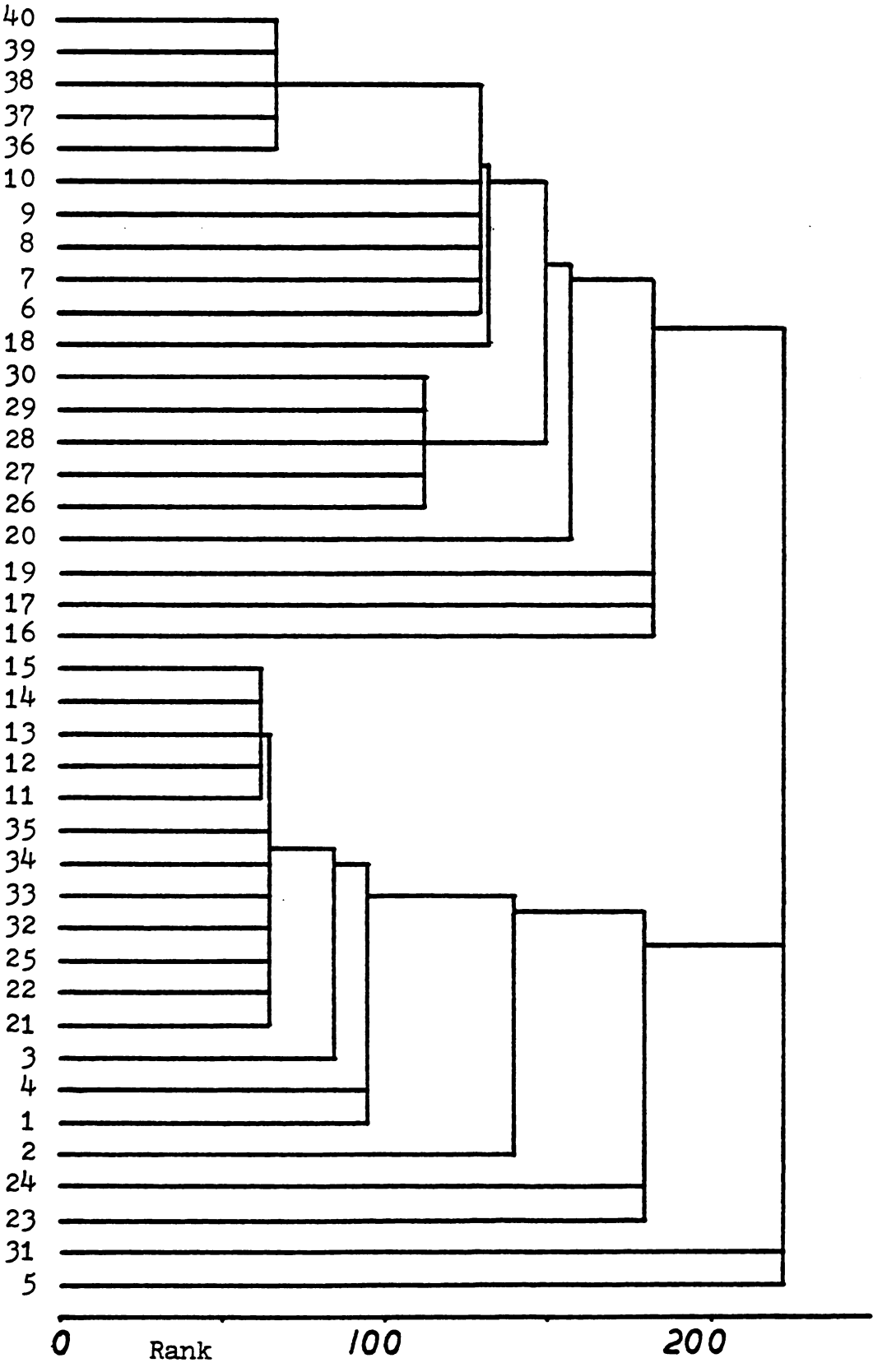


Figure 4.4. k-Clustering, k=4, Tosi Data

number of edges, V, required to form a connected graph. For this data set, V = 223. The tables published by Ling and Killough [LIN76] for the cumulative probability of this statistic under the random graph null hypothesis yield Prob (V <= 145) >= 0.99.

There is less than one chance in one hundred that a 40-node graph would require more than 145 edges to become connected under the random graph null hypothesis. Expressions given by Ling and Killough may be used to calculate the probability that a 40-node random graph with 223 edges is connected, since only two terms contribute to the required sums. The result is

Prob (V <= 223) = 0.999950.

We conclude that the data set is not a random data set under the random graph null hypothesis.

The validity of single link clusters may be tested using the cluster lifetime statistic developed by Ling [LIN73a]. Table 4.1 lists the number of nodes, rank at formation, value l of the random variable L (cluster lifetime) and Prob (L >= l) under the random graph null hypothesis for 26 single link clusters. The composition of each cluster is listed in Table 4.2. Twelve of the clusters, marked "*" in Table 4.1, have Prob (L >= l) <= 0.05 and, thus, are "real" clusters by this test. Several of these clusters are identified in Figure 4.1.

As expected, subclusters of five samples from single

Table 4.1 Single Link Cluster Lifetime Test

	Cluster Number	Size	Birth Rank	Lifetime l	Prob (L>=l)
*	2	20	144	79	small
*	3	20	150	73	small
*	4	19	82	62	small
*	6	16	41	41	small
*	8	15	110	40	small
	13	11	39	2	.57
*	19	10	60	50	small
	21	9	32	7	.60E-1
	24	5	57	3	.57
*	25	5	19	22	.38E-2
*	26	5	55	55	.17E-6
*	28	5	31	119	small
*	30	5	24	36	.78E-4
	31	5	29	3	.59
	32	4	20	12	.98E-1
	34	4	51	6	.33
	35	4	12	7	.29
	37	4	50	5	.41
	38	4	30	1	1.00
*	39	4	5	24	.82E-2
	40	4	21	3	.66
	41	3	10	10	.24
	42	3	35	16	.87E-1
	43	3	34	16	.87E-1
*	44	3	38	44	.75E-3
	45	3	16	5	.53

Cluster Number refers to clusters listed in Table 4.2.

small indicates the probability is less than 10 ** (-10).

.41E-2 indicates the probability is .41 x 10 ** (-2).

Table 4.2 Validity Measures of Potential Clusters

The entries are the smallest values for Cl and Il in the profiles of potential clusters from the Tosi data.

Cluster Number	Size	Samples in the Cluster	Cl	Il
1	24	5-10,16-20,23-24,26-31,36-40	.63E-10	small
2	20	1-5,11-15,21-25,31-35	small	small
3	20	6-10,16-20,26-30,36-40	.27E- 8	small
4	19	1-4,11-15,21-25,31-35	small	small
5	18	1-4,11-15,21-25,32-35	small	small
6	16	1-4,11-15,21-22,25,32-35	small	small
7	15	1,3-4,11-15,21-22,25,32,35	small	small
8	15	6-10,16-20,36-40	small	small
9	14	5-10,23-24,26-31	.19E- 3	small
10	14	6-10,16-19,36-40	.6 E-14	small
11	12	11-15,21-22,25,32-35	small	.12E- 4
12	11	1-4,11-15,21-22	small	small
13	11	1-4,21-22,25,32-35	small	.16E- 7
14	10	1-10	1.0	.15
15	10	11-20	1.0	.40E- 2
16	10	21-30	1.0	1.0
17	10	31-40	1.0	1.0
18	10	6-10,26-30	.21	small
19	10	6-10,36-40	.16E- 7	small
20	10	16-20,36-40	.70E-12	small
21	9	1-4,25,32-35	.36E- 8	.59E- 1
22	7	11-15,21-22	.68E- 9	1.0
23	5	1-5	1.0	1.0
24	5	6-10	.69E- 1	.27E- 3
25	5	11-15	.62E- 5	1.0
26	5	16-20	1.0	.79E- 6
27	5	21-25	1.0	1.0
28	5	26-30	.35E- 2	.42E-12
29	5	31-35	.24	1.0
30	5	36-40	.12- 4	.38- 2

Table 4.2 (Continued)

Cluster Number	Size	Samples in the Cluster	Cl	Il
31	5	25,32-35	.50E- 4	1.0
32	4	1-4	.47E- 2	1.0
33	4	5,23-24,31	1.0	.89E- 7
34	4	6-8,10	1.0	1.0
35	4	11-12,14-15	.11E- 2	1.0
36	4	16-19	.36	.12E- 3
37	4	16-18,20	1.0	.60E- 1
38	4	26-28,30	1.0	.70E- 9
39	4	32-35	.18E- 3	1.0
40	4	36-38,40	.39E- 1	1.0
41	3	1,3-4	.14	1.0
42	3	6-8	1.0	1.0
43	3	16-18	1.0	.38
44	3	23-24,31	1.0	.44E- 2
45	3	36-37,40	.14	1.0

small indicates the smallest value in the profile is less than 10 ** (-15)

.44E- 2 indicates the smallest value in the profile is .44 x 10 ** (-2)

subjects are important components of the data set. Four of the twelve clusters with significantly long lifetimes have five points and four others have multiples of five points. Furthermore, these eight clusters all consist of combinations of complete five point groupings, where each five point grouping contains all the samples from a single speaker using one mode of recording. The two twenty-point clusters consist of the direct and phone recording mode samples.

4.5. Cluster Profiles

Raw profiles, sequences of the compactness and isolation indices, and probability profiles, sequences of the four measures I1, I2, C1 and C2, were obtained for all single link and complete link clusters with four or more nodes and for all clusters which appeared in more than one k-clustering. In addition, profiles were calculated for all five-point subsets corresponding to a single subject and single mode of recording, and all ten-point subsets corresponding to a single subject. The indices and measures were calculated for every tenth rank from 10 to 780.

The probability profiles for four of the five-point subsets for a single subject and single recording mode are shown in Figure 4.5. These four subsets are circled in Figure 4.1. The profiles vary greatly from one subset to

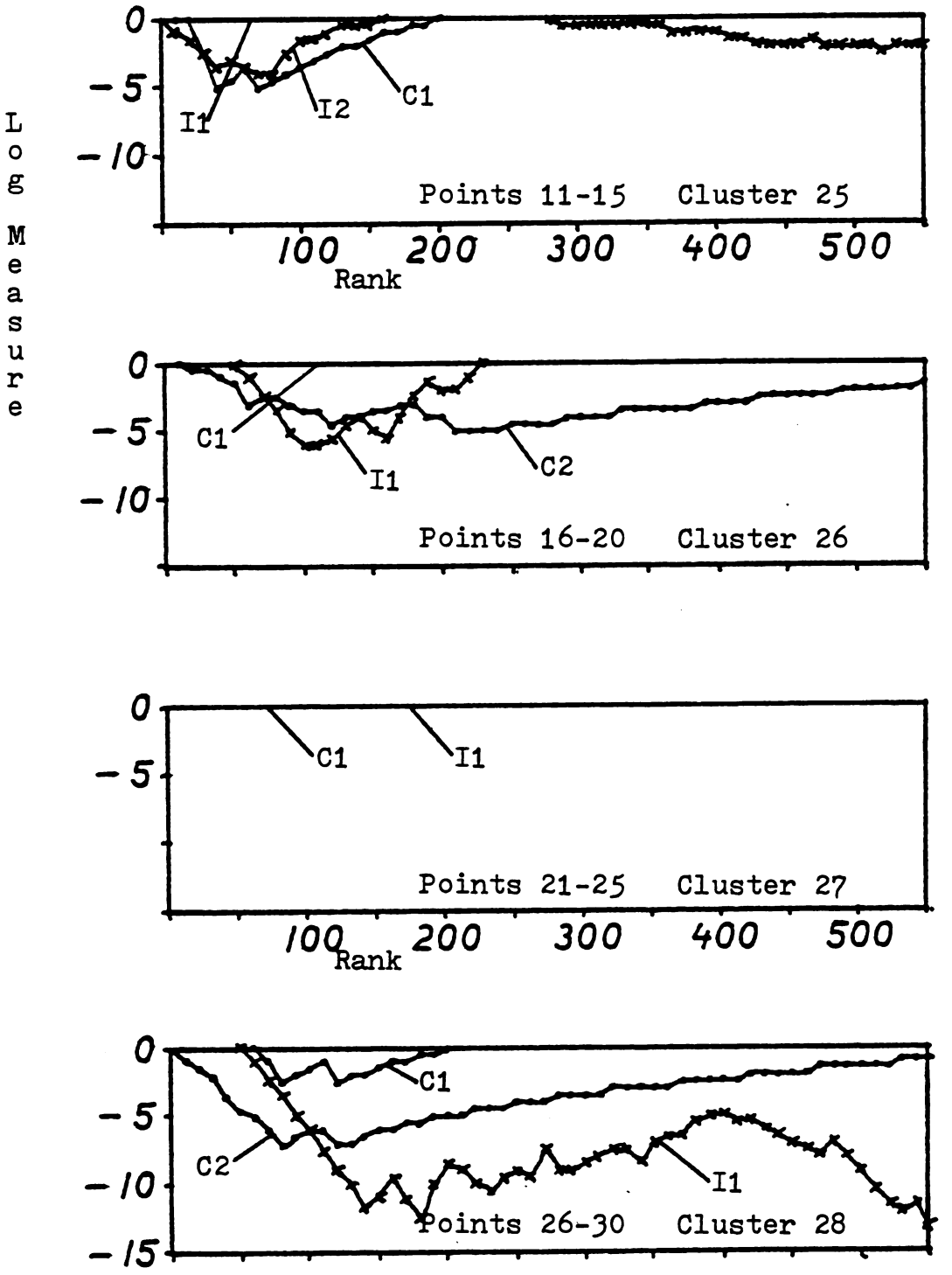


Figure 4.5. Probability Profiles of Clusters 25-28

the next. Cluster 25, points 11-15, shows good evidence of compactness, with measure C1 below .001 for ranks 40 to 100. The number of linking edges is not unusually low by measure I1, but I2 is below .001 for ranks 40 to 70, so this subset is compact and for a subset this compact, it is somewhat isolated. Cluster 26, points 16-20, exhibits strong isolation, with measure I1 below .001 for ranks 80 to 170, but weak compactness by measure C1. Since measure C2 is below .001 for ranks 100 to 160, this subset is isolated, and compact for a subset this isolated.

Cluster 27, points 21-25, is neither compact nor isolated. Since measures I1 and C1 are never small, we do not consider measures I2 or C2. Cluster 28, points 26-30, has lower isolation measure I1 than cluster 26. In addition, this cluster also has several low values for the compactness measure C1, but the evidence for compactness is not as strong as it was for cluster 25. The compactness measure C2 is quite low, so the cluster is compact among clusters with the same isolation. Of these four clusters, number 25 is the most compact and number 28 is the most isolated. Cluster 26 is also isolated, while cluster 27 is neither compact nor isolated.

Inspection of the two dimensional configuration in Figure 4.1 provides support for the conclusions reached by analyzing the probability profiles. An analysis of the single link and complete link hierarchies regarding the

four clusters also supports the conclusions reached above. Clusters 25, 26 and 28 are both single and complete link clusters. Cluster 27 is neither a single link nor a complete link cluster, which supports the conclusion that it is neither compact nor isolated. If we use rank at birth as a measure of compactness, Table 4.1 shows that cluster 25 is the most compact and cluster 26 is the least compact, with cluster 28 falling in between. If we use the lifetime as a measure of isolation, then cluster 28 is the most isolated, cluster 25 is the least isolated and cluster 26 falls in between.

From our knowledge of the data collection process, we might expect the data set to be organized as two clusters of twenty points each (telephone and direct) or as four clusters of ten points each (by speaker). The profiles for the four appropriate ten-point clusters, clusters 14, 15, 16 and 17 in Table 4.2, and the two appropriate twenty-point clusters, clusters 2 and 3, are given in Figures 4.6a,b. It is immediately apparent that the subsets of ten points in Figure 4.6a do not form valid clusters. None of them has a compactness measure C1 below 0.1 and only cluster 15 has an isolation measure I1 below 0.01 so none of these subsets are compact or isolated. Figure 4.6b shows that clusters 2 and 3 are compact and isolated. The isolation measures I1 are below 10**-15 for a wide range of ranks. The compactness measure C1 is

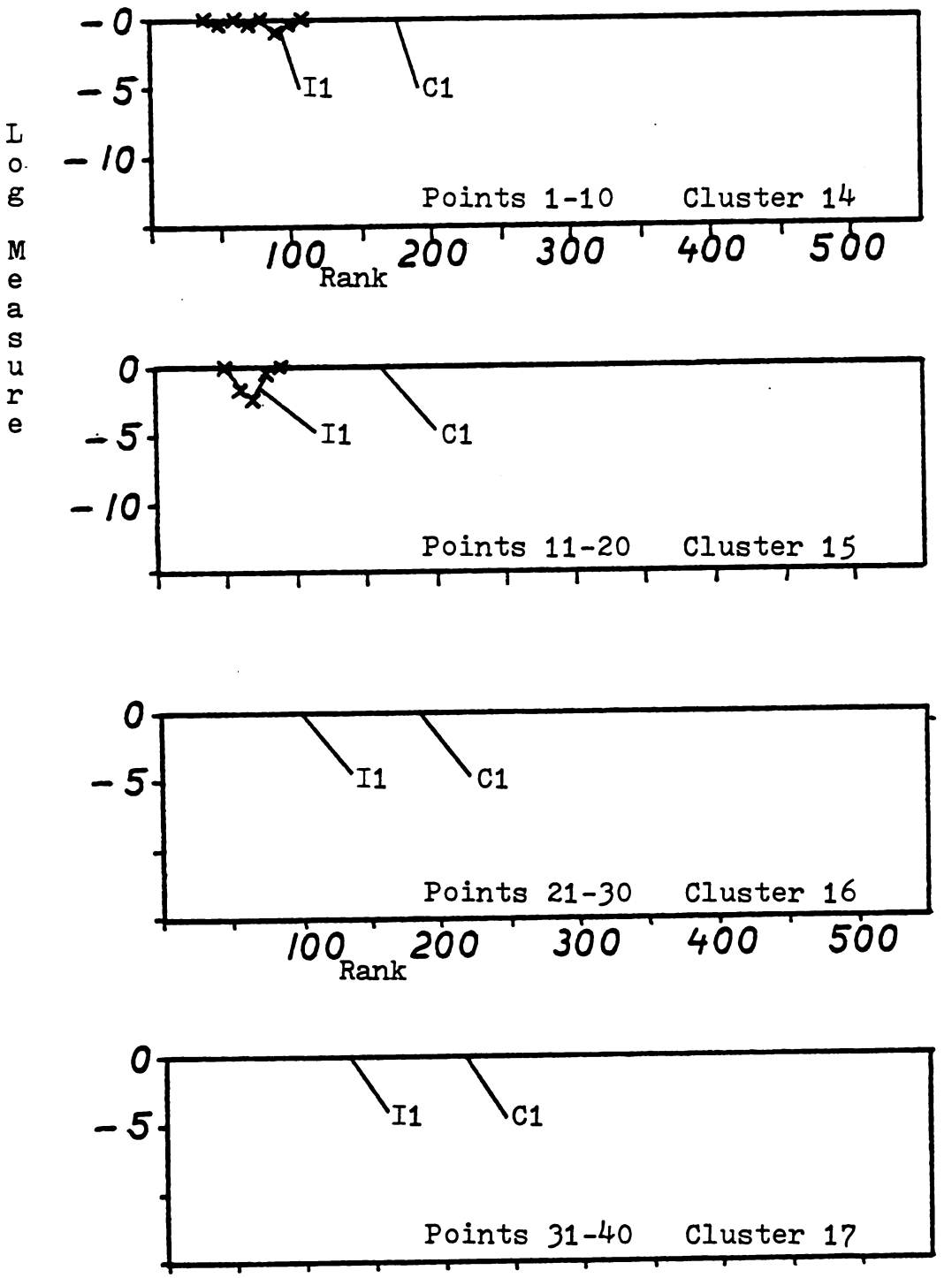


Figure 4.6a. Probability Profiles of Clusters 14-17

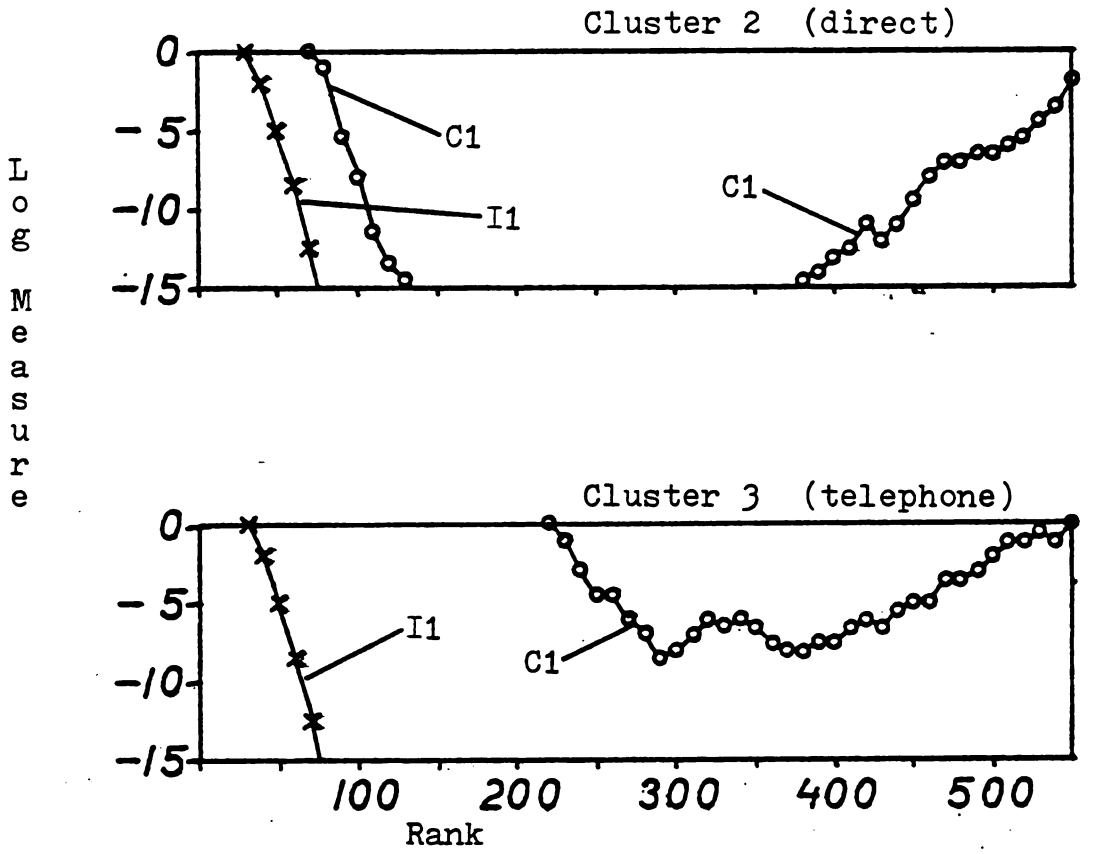


Figure 4.6b. Probability Profiles of Clusters 2,3

better for cluster 2 (direct recording), less than 10**-15 for a wide range ranks, but is also very good for cluster 3 (telephone), with a low value below 10**-8. We conclude that the data may be thought of as organized into two clusters, (2 and 3), while the four clusters, (14 - 17), are not valid.

Profiles were obtained for all the 45 potential clusters listed in Table 4.2. Table 4.2 also lists the lowest value observed for the compactness measure C1 and for the isolation measure I1. Table 4.3 lists the clusters with more than four nodes which appear as single link clusters, as complete link clusters and as k-clusters for more than one k value. The clusters which have a compactness measure C1 (or C2) less than 10**-5, and those which have an isolation measure I1 (or I2) less than 10**-5 are marked.

With three exceptions, all the clusters with 10 or more points are both compact and isolated at the 10**-5 level. Clusters 9 and 18 are isolated, and compact among clusters of their isolation. Cluster 11 is compact, and isolated among clusters of its compactness. Only three clusters fail to be either compact or isolated by the best case measures. Of these, the most interesting is cluster 30, identified in Figure 4.1, which appears as a single link and complete link cluster and as a k-cluster for k = 2 and 3. This cluster is somewhat compact, with measure

Table 4.3 Validity of Hierarchical Clusters

Single Link Clusters

Cluster Number	Size	Validity I1	C1	I2	C2	Cluster Number	Size	Validity I1	C1	I2	C2
2	20	x	x			21	9		x	x	
3	20	x	x			24	5			x	x
4	19	x	x			25	5		x		
6	16	x	x			26	5	x			x
8	15	x	x			28	5	x			x
13	11	x	x			30	5			x	x
19	10	x	x			31	5				x

Complete Link Clusters

Cluster Number	Size	Validity I1	C1	I2	C2	Cluster Number	Size	Validity I1	C1	I2	C2
1	24	x	x			24	5			x	x
6	16	x	x			25	5		x		
9	14	x			x	26	5	x			x
12	11	x	x			28	5	x			x
18	10	x			x	30	5			x	x
20	10	x	x			31	5				x
22	7		x								

k-Clusters, for more than one k of k = 1, 2, 3, 4 and 7

Cluster Number	Size	Validity I1	C1	I2	C2	Cluster Number	Size	Validity I1	C1	I2	C2
2	20	x	x			10	14	x	x		
3	20	x	x			11	12		x	x	
4	19	x	x			19	10	x	x		
5	18	x	x			25	5		x		
6	16	x	x			28	5	x			x
7	15	x	x			30	5			x	x
8	15	x	x								

x indicates the measure has a minimum value less than 10 ** (-5).

C1 below 10**-4, but the isolation measure I1 never falls below 10**-3. The probability of a single link cluster lifetime as good as the observed value of 36 with a formation rank of 24 is .000078, which indicates that the cluster is isolated. Also, the isolation measure I2 is below 10**-5. If we are willing to accept a cluster with C1 below 10**-4 as compact, then we also conclude that cluster 30 is isolated among clusters with the same compactness index.

Clusters 22 and 25, identified in Figure 4.1, are compact, but are not isolated at the 10**-5 level, even among clusters of their compactness. Cluster 25 is a single link cluster with lifetime l = 22 and Prob (L >= l) = .0038 under the random graph null hypothesis. From the probability profiles we conclude that this low value is due to the compactness of the cluster, rather than its isolation. It is interesting to compare cluster 25 with another single link cluster, number 21, also identified in Figure 4.1. The lifetime of cluster 21, l = 7 with Prob (L >= l) = .060, is much shorter than the lifetime of cluster 25, yet the validity measures C1 and I1 are both lower for cluster 21 than for cluster 25. The probability profiles provide information on the interaction of the cluster with other points in the data set at many different ranks, and thus may provide a

much different view of the cluster than a test which observes only a limited range of ranks.

4.6. Conclusions

The probability profiles show that almost all the potential clusters identified by the three clustering methods used here are significantly compact and isolated. This data set is well described by clusters. It is far from being a random data set under the random graph null hypothesis.

Aside from statements regarding the birth and life times of the clusters, which are useful only for comparisons among clusters, the only tools available in the literature for testing the validity of clusters derived from proximity matrices are the single link lifetime test of Ling [LIN73a] and the complete link extraneous edges test of Baker and Hubert [BAK76]. The extraneous edges test cannot be applied to a 40-node data set since the required Monte Carlo runs for 40-node data have not been published and are time consuming to obtain. The single link cluster lifetime test is inferior to the use of cluster profiles in two ways. First, cluster profiles yield unique information about the compactness and isolation of a cluster, while the lifetime test combines the two requirements into a single test. Second, cluster profiles may be applied to any subset, including

k-clusters and complete link clusters, while the lifetime test is only applicable to single link clusters.

Another conclusion must be considered and cannot be easily dismissed. If the intrinisic dimensionality of the data set is two, as suggested by the results of the MDSCAL program, then the results in Chapter 3 lead to the conclusion that probability bounds developed under the random graph null hypothesis are not appropriate and will lead to gross underestimates of the probability that a validity index will have a value as good as the observed value. Even if the intrinsic dimensionality is near 38, the distributions developed under the random graph null hypothesis may not give a true picture of the validity of the clusters. If the intrinsic dimensionality is two, the distributions used in the cluster profiles must be determined by simulation under an appropriate two dimensional null hypothesis. Since no other test is available, this simulation is necessary if validity tests are to be applied to clusters created by methods other than single link or complete link clustering. We must join with Ling and Matula in warning of the danger inherent in applying tests based on the random graph null hypothesis to situations where the assumptions of the hypothesis may be violated.

5. Conclusion

The main thrust of this thesis is a study of the distribution of two measures of cluster validity under two null hypotheses. Section 5.1 summarizes the results and cites the main contributions of the thesis and Section 5.2 suggests areas for future work.

5.1. Summary of Results

The clustering situation under consideration is that in which the information of interest is a matrix of proximities with ordinal scale. This situation often occurs in psychometric and sociometric data. Chapter 1 presents necessary definitions and a review of the literature on cluster validity with emphasis on proximity matrices with ordinal scale. The ordinal information in such a proximity matrix may be represented as a sequence of threshold graphs. This sequence of threshold graphs is used to develop Cluster Profiles, a new tool for the analysis of cluster validity. The probability distributions which are needed for the computation of Cluster Profiles are investigated for the Random Graph Null Hypothsis and for the Uniform Hypercube Null Hypothesis.

The concept of a Cluster Profile, which graphically represents the interaction of a proposed cluster with the environment of points in which it occurs, is developed in

Chapter 2. This concept has not previously been applied to cluster validity studies and allows a more detailed analysis of the compactness and isolation of a cluster than previous cluster testing techniques. The Cluster Profile concept requires indices of cluster validity which may be applied to any subset of points in a threshold graph. Two indices which meet this requirement, the number of internal edges for compactness and the number of linking edges for isolation, are developed in Section 2.1.2.

A contribution of this thesis is the discussion of the classification and utility of various choices for the sample population of clusters in Section 2.2.1. We argue that the best case distribution is the most useful distribution for tests of cluster validity, because it is applicable to any subset of points, and thus to a cluster found by any clustering method, and because it forms an upper bound on the distribution of any sample population which includes one subset from each random graph. Upper bounds on the cumulative distribution function for the number of internal and linking edges using the best case distribution are derived in Sections 2.2.2 and 2.2.3. The derivation of these bounds and the demonstration of their asymptotic behavior in the special case where all internal edges are present, given in Section 2.3, are the main contributions of this thesis. These bounds are used to

calculate two of the validity measures which form the Probability Profile. Two other validity measures, based on fixed validity index distributions, are also developed in Section 2.2.2 and 2.2.3.

Chapter 3 presents a study of the cumulative distributions of the number of internal and linking edges under the Uniform Hypercube Null Hypothesis. The results demonstrate that distributions derived under the Random Graph Null Hypothesis lead to false conclusions with data sets created under the Uniform Hypercube Null Hypothesis. Although several authors have warned against using results based on the Random Graph Null Hypothesis to test the validity of clusters, this is the first explicit demonstration of the extent to which an alternative null hypothesis can affect the distributions. Theoretical distributions under the Uniform Hypercube Null Hypothesis are presented for four points in one dimension, and distributions created by Monte Carlo simulation of the creation of data sets are presented for several medium sized data sets.

An explicit effect of high dimensionality on the distribution of interpoint distances for a set of five points chosen from a uniform distribution in a hypercube is shown in Section 3.2. This effect has not been reported in the literature and sheds light on a subtle issue in intrinsic dimensionality. Two viewpoints of the

meaning of intrinsic dimensionality are discussed and it is shown that the viewpoint which assumes the intrinsic dimensionality to be the number of free parameters which define the allowed positions of the points in space leads to important information on the structure of the underlying population which is suppressed by the alternative viewpoint.

In Chapter 4 Cluster Profiles are used to determine the validity of several potential clusters in a data set consisting of 40 samples of choral speech. We show that for many of the potential clusters, no test of cluster validity in the literature can be applied. The single link cluster lifetime test of Ling [LIN73a] does not apply to subsets of points which are not single link clusters, and the extraneous edges test of Baker and Hubert [BAK76] does not apply to subsets which are not complete link clusters. Furthermore, the distributions required for application of the extraneous edges test are not readily available for 40-point data sets. The Cluster Profiles show that most of the large clusters found by the single link, complete link and k-clustering methods are unusually compact and isolated, and thus we conclude that these clusters are valid, with a proviso concerning intrinsic dimensionality. The Cluster Profiles show that one natural way of organizing this data set into clusters yields valid clusters, while another does not.

A note of warning is emphasized in Chapters 3 and 4, namely, tools based on the Random Graph Null Hypothesis, such as Cluster Profiles, which test the validity of clusters, must be applied with caution. If the data are patterns in a space, the distributions derived using the Random Graph Null Hypothesis are not applicable, and the results will overrate the validity of the clusters.

5.2. Areas of Future Work

The best case distributions used in this thesis select the subset of nodes in each random graph which has the optimum validity index. Distributions for the isolation and compactness indices are developed separately. Best case distributions for subsets which satisfy a combination of requirements of compactness and isolation should be developed. Other indices of compactness and isolation and their best case distributions should also be developed.

The validity measures C1 and I1 are upper bounds on the best case distributions of the compactness and isolation indices under the random graph null hypothesis. The development of exact expressions of these distributions is an area for future work. Along the same line, development of an algorithm for determining the optimal compactness and isolation indices for a k-node subset of a graph would improve the simulation used in

this thesis.

The distributions of graph-theory-based tests of cluster validity other than those defined here should be investigated under the Uniform Hypercube Null Hypothesis. The goal of such studies is to find computationally inexpensive tests based on the Random Graph Null Hypothesis, which are robust enough to be applicable in situations where the Uniform Hypercube Null Hypothesis is appropriate. In particular, the distribution of Ling's single link cluster lifetime statistic should be investigated.

In this thesis two versions of a "no clustering" null hypothesis are investigated. The development of tests for rejection of these null hypotheses are an important step in the study of cluster validity. A further step is to investigate alternatives to the null hypotheses. What are the standard hypotheses of "clustering" which are to be accepted when a "no clustering" null hypothesis is rejected? Definition and classification of "clustering" hypotheses is an important area of future work.

The distribution of interpoint distances for a set of points chosen at random from a uniform distribution in a hypercube is studied in this thesis. The interpoint distance distribution for other distributions of the points, such at the multidimensional Gaussian distribution, should be studied. It would be interesting

to know the characterization of distributions for which the ratio of shortest to longest interpoint distances in a set of points drawn from the distribution goes to one as the dimensionality increases.

Finally, the regular changes in the cumulative distribution functions of the validity measures as the dimensionality is varied under the Uniform Hypercube Null Hypothesis suggest that it may be possible to find a functional relationship among these distributions for the various dimensions. Distributions under the Random Graph Null Hypothesis, together with the transformation relating an infinite dimensional distribution to a d-dimensional distribution, would provide an inexpensive test of cluster validity under the Uniform Hypercube Null Hypothesis.

APPENDIX

APPENDIX

An Approximate Best Case Algorithm

We wish to find the optimal value for a validity index over all K-node subsets of the nodes of a graph as required for the simulation of Section 3.4.1, given an adjacency matrix GR(i,j), K, and a parameter TRIES. The following algorithm approximates the optimal value for the number of internal edges, a compactness index. A similar algorithm approximates the optimal value for the number of linking edges, an isolation index.

1. Use the K nodes with highest degrees as the initial subset.
2. Call SEARCH to find a subset with "local" optimal compactness.
3. Set COMP to the compactness index for the subset.
4. Set LOOP to 1.
5. Choose a K-node subset at random.
6. Call SEARCH to find a subset with "local" optimal compactness.
7. Find the compactness index for the subset. If the index is greater than COMP, then set COMP to the new index and GO TO 4.
8. Set LOOP to LOOP + 1. If LOOP <= TRIES then GO TO 5,

else DONE.

Now COMP is a good value for the best case compactness index of the graph. It is a "local" optimal value and is at least as good as TRIES other randomly selected "local" maxima.

The routine SEARCH finds a "local" subset with optimal "local" validity index. A "local" subset is one which can be created by repeatedly exchanging a node in the subset with a node not in the subset such that each exchange improves the validity index as much as possible. The SEARCH algorithm for the compactness index is given below.

1. For each node i, find the change VAL(i) in the number of internal edges of the subset if node i is moved from outside the subset into the subset.
2. Find the highest VAL among nodes outside the subset (BESTOUT) and the lowest VAL among nodes in the subset (WORSTIN). If BESTOUT <= WORSTIN, then DONE.
3. Look for a pair of nodes (IN,OUT) such that VAL(IN) = WORSTIN, VAL(OUT) = BESTOUT and (IN,OUT) is not an edge of the graph. If such a pair is found, then GO TO 5, else if BESTOUT <= WORSTIN + 1, then DONE.
4. Find a pair of nodes (IN,OUT) such that VAL(IN) = WORSTIN, VAL(OUT) = BESTOUT and (IN,OUT) is an edge of the graph.

5. Move node IN out of the subset, move node OUT into the subset and update VAL to reflect the exchange. GO TO 2.

Two versions of random selection of the subset were tried. Weighting nodes with the degree of the node seemed better than weighting nodes uniformly in that more subsets were discovered which improved the indices found using the initial subsets. Results using TRIES = 100 suggest that about 90 percent of the improved indices are found with the first ten random subsets. Since the search time was the determining factor in the time required for each simulation run, TRIES = 10 was used for all the reported runs.

BIBLIOGRAPHY

BIBLIOGRAPHY

ABR64 -- C. T. Abraham, "Evaluation of clusters on the basis of random graph theory." IBM Res. Rept. RC-1177, 1964.

AND73 -- M. R. Anderberg, Cluster Analysis for Applications. New York: Academic Press, 1973.

BAK75 -- Frank B. Baker and Lawrence J. Hubert, "Measuring the power of hierarchical cluster analysis." JASA 70, 31-38 (1975).

BAK76 -- Frank B. Baker and Lawrence J. Hubert, "A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering." JASA 71, 870-878 (1976).

BEN69 -- Robert S. Bennett, "The intrinsic dimensionality of signal collections." IEEE Trans. Inf. Th. IT-15, 517-525 (1969).

BLA77 -- Roger K. Blashfield and Mark S. Aldenderfer, "A consumer report on cluster analysis software." Penn. State, Report for NSF grant DCR #74-20007, 1977.

BRE75 -- Ronald L. Breiger, Scott A. Boorman and Phipps Arabie, "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling." J. of Math. Psych. 12, 328-383 (1975).

BOR75 -- W. J. Borucki, D. H. Card and G. C. Lyle, "A method of using cluster analysis to study statistical dependence in multivariate data." IEEE Trans. Comp. C-24, 1183-1191 (1975).

BRO65 -- K. A. Brownlee, Statistical Theory and Methodology in Science and Engineering, 2nd Ed. New York: Wiley, 1965.

CHE74 -- Chiu Kuan Chen and Harry C. Andrews, "Nonlinear intrinsic dimensionality computations." IEEE Trans. Comp. C-23, 178-184 (1974).

DAY77 -- William H. E. Day, "Validity of clusters formed by graph-theoretic methods." Math. Biosci. 36, 299-317 (1977).

DUB76 -- Richard Dubes and Anil K. Jain, "Clustering techniques: the user's dilemma." Pattern Recognition 8, 247-260 (1976).

DUB77 -- Richard Dubes and Anil K. Jain, "Models and methods in cluster validity." Tech. Rep. TR-77-05, Michigan State University, 1977.

ENG69 -- L. Engleman and J. A. Hartigan, "Percentage points of a test for clusters." JASA 64, 1647-1648 (1969).

ERD59 -- P. Erdos and A. Renyi, "On random graphs 1." Publ. Math. (Dubrecen) 6, 290-297 (1959).

ERD60 -- P. Erdos and A. Renyi, "On the evolution of random graphs." Publ. Math. Inst. Hung. Acad. Sci. 5, 17-61 (1960).

ERD61 -- P. Erdos and A. Renyi, "On the strength of connectedness of a random graph." Act. Math. Hung. 12, 261-267 (1961).

GOW70 -- J. C. Gower and G. J. S. Ross, "Minimum spanning trees and single linkage cluster analysis." Appl. Stat. 18, 54-64 (1970).

HAR67 -- J. A. Hartigan, "Representation of similarity matrices by trees." JASA 62, 1140-1158 (1967).

HUB74a -- Lawrence J. Hubert, "Some applications of graph theory to clustering." Psychometrika 39, 283-309 (1974).

HUB74b -- Lawrence Hubert, "Spanning trees and aspects of clustering." Br. J. Math. Statist. Psychol. 27, 14-28 (1974).

JOH67 -- Stephen C. Johnson, "Hierarchical clustering schemes." Psychometrika 32, 241-254 (1967).

KEL76 -- F. P. Kelly and B. D. Ripley, "A note on Strauss's model for clustering." Biometrika 63, 357-360 (1976).

KRU64 -- J. B. Kruskal, "Nonmetric multidimensional scaling: a numerical method." Psychometrika 29, 115-129 (1964).

LIN72 -- Robert F. Ling, "On the theory and construction of k-clusters." Computer Journal 15, 326-332 (1972).

LIN73a -- Robert F. Ling, "A probability theory of cluster analysis." JASA 68, 159-164 (1973).

LIN73b -- Robert F. Ling, "The expected number of components in random linear graphs." Annals of Probability 1, 876-881 (1973).

LIN75 -- Robert F. Ling, "An exact probability distribution on the connectivity of random graphs." J. Math. Psych. 12, 90-98 (1975).

LIN76 -- Robert F. Ling and George G. Killough, "Probability tables for cluster analysis based on a theory of random graphs." JASA 71, 293-300 (1976).

MAT77 -- David W. Matula, "Graph theoretic techniques for cluster analysis algorithms." In Classification and Clustering. New York: Academic Press, 1977.

MOU70 -- M. D. Mountford, "A test for the difference between clusters." Statistical Ecology 3, 237-251 (1970).

MQU61 -- L. L. McQuitty, "Typal analysis." Ed. and Psych. Meas. 21, 677-697 (1961).

MQU67 -- L. L. McQuitty, "A mutual development of some typological theories and pattern analytic methods." Ed. and Psych. Meas. 27, 21-48 (1967).

MSH77 -- Michael G. McShane and Charles R. Sherman, "Classification of U.S. medical schools." Presented at the Classification Society (NAB) annual meeting, Dartmouth, June 7-9, 1977.

PET79 -- Karl Pettis, Thomas Bailey, Anil K. Jain and Richard Dubes, "An intrinsic dimensionality estimator from near-neighbor information." To be published in IEEE Trans. Pat. Rec. Mach. Int. (1979).

RAP72 -- A. Rapoport and S. Fillenbaum, "An experimental study of semantic structures." In A. K. Romney, R. N. Shepard and S. B. Nerlove (Eds.), Multidimensional Scaling, Vol. II., Pp. 93-131. New York: Seminar Press, 1972.

RID53 -- R. J. Riddell, Jr., and G. E. Uhlenbeck, "On the theory of the virial development of the equation of state of monoatomic gases." J. Chem. Physics 21, 2056-2064 (1953).

RIJ70 -- C. J. van Rijsbergen, "A clustering algorithm." Comp. Jour. 13, 113-115 (1970).

SCH73 -- James V. Schultz and Lawrence J. Hubert, "Data analysis and the connectivity of random graphs." Journal of Mathematical Psychology 10, 421-428 (1973).

SNE77 -- P. H. A. Sneath, "Method for testing the distinctness of clusters." Math. Geo. 9, 123-143 (1977).

STR75 -- David J. Strauss, "A model for clustering." Biometrika 62, 467-475 (1975).

TOS75 -- O. I. Tosi, "The problem of speaker identification and elimination." In Measurement Procedures in Speech, Hearing, and Language (S. Singh, ed.). Baltimore: University Park Press, 1975.

WON77 -- Andrew K. C. Wong and T. S. Liu, "A decision-directed clustering algorithm for discrete data." IEEE Trans. Comp. C-26, 75-82 (1977).

ZAH71 -- C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters." IEEE Trans. Comp. C-20, 68-86 (1971).