Validity of Clusters Formed by Graph-Theoretic Cluster Methods*

WILLIAM H. E. DAY
Department of Computer Science, Southern Methodist University, Dallas, Texas 75275

Received 19 April 1976; revised 25 May 1977

ABSTRACT

Let P be a finite set of objects to be classified, and let G be a loopless labeled undirected graph, with point set P, in which two points are adjacent if the objects they represent are similar with respect to a specified criterion. A cluster method is simply a particular function ρ that maps G into a set $\rho(G)$ of subsets of P, the clusters associated with G. In this paper I describe graph-theoretic criteria to measure the validity or relative significance of clusters formed by such methods. Using several of these criteria, I classify a sequence of flat cluster methods in which the notions of cluster are based on various graph-theoretic concepts of internal coherence. Basic issues of cluster validity are illustrated by examples using Jardine and Sibson's B_k flat cluster methods.

1. INTRODUCTION AND SUMMARY

In this paper, P is a finite set of objects to be classified, and G is a loopless labeled undirected graph, with point set P, in which two points are joined by a line if the objects they represent are similar with respect to a specified criterion. Line lengths and relative positions of points have no significance in such graphs; thus the graphs in Fig. 1 (a) and (b) are identical. The set of all loopless labeled undirected graphs with point set P is denoted by $\mathcal{G}(P)$.

A graph-theoretic cluster method operates on each graph G in $\mathcal{G}(P)$ to obtain a simple clustering of the objects in P. This clustering is a set of subsets of P such that each object in P is an element of at least one subset in the clustering, and no subset in the clustering contains any other subset in the clustering. The subsets in the clustering obtained from G are called

*This research was supported in part by grants GJ-40487 and DCR75-10930 of the National Science Foundation.

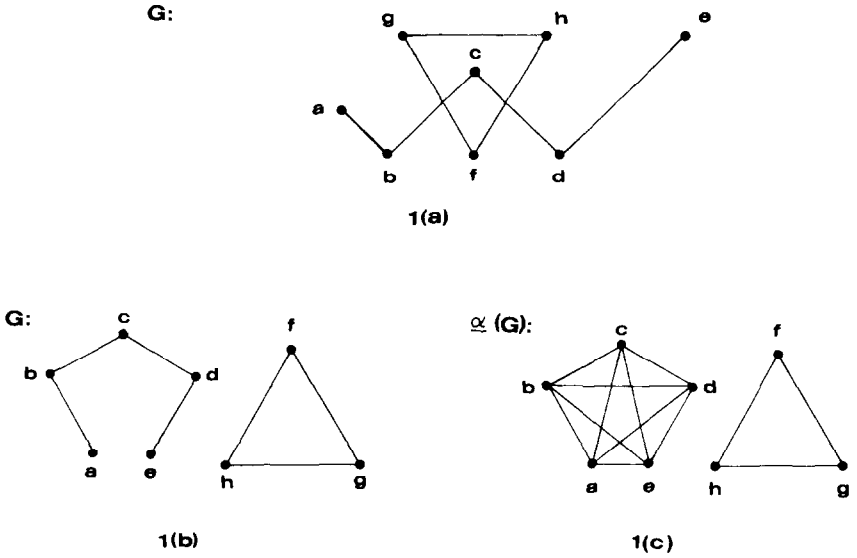


FIG. 1. Loopless, labeled, undirected graphs.

clusters and usually are characterized in G by properties of coherence and isolation.

Graph-theoretic cluster methods may be used to obtain numerically stratified clusterings from numerical measures of the dissimilarity of pairs of objects in P [13]. However, the operation of the cluster methods described here does not require this more complex data environment; to achieve clarity, we focus on the problem of obtaining simple, as opposed to stratified, clusterings.

A basic example of a graph-theoretic cluster method depends on the concept of a clique, or maximal complete subgraph, of a graph G. The *clique cluster method* is a function $\omega: \mathcal{G}(P) \to \mathcal{P}(\mathcal{P}(P))$, where $\mathcal{P}(X)$ denotes the set of all subsets of the set X, and where for every graph G in $\mathcal{G}(P)$, $\omega(G) = \{Q\colon Q \text{ is the point set of a clique in } G\}$. When ω operates on the graph G in Fig. 1 (b), the simple clustering $\omega(G) = \{\{a,b\},\{b,c\},\{c,d\},\{d,e\},\{f,gh\}\}$ is obtained.

The clique cluster method exhibits a number of characteristic properties. Its clusters are dense and compact, since each is a maximal set such that every two points in the set are adjacent in the corresponding graph. The clusters are neither strung out nor straggly. The clusters may be poorly separated, since two clusters containing m and $n \geqslant m$ points may overlap in as many as $m-1$ points. The clique cluster method actually preserves the information represented by the graph, rather than simplifying it, since the

graph G may be reconstructed from the clustering $\omega(G)$. Finally, no efficient implementation exists for the method, since Moon and Moser [18] have shown that the maximum number of cliques in any graph on p points increases exponentially with p.

The functional representation of graph-theoretic cluster methods may be extended by separating their operation into sequential phases of graph transformation and cluster enumeration. First the initial graph G is transformed (presumably with controlled information loss) into another graph $\underline{G}$ in such a way that the cliques of $\underline{G}$ represent the clusters to be associated with G. Then the ω function is applied to $\underline{G}$ to enumerate the simple clustering.

Single-linkage clustering [22] (the dendritic method of Florek et al. [6]) is a typical example of a graph-theoretic cluster method that may be described in this way. Let a graph be called partitioned if each of its components is complete; then the single-linkage graph transformation phase is represented by a function $\underline{\alpha}: \mathcal{G}(P) \rightarrow \mathcal{G}(P)$ that maps each graph G into the least partitioned graph containing G as a subgraph. It is easy to establish that in fact $\underline{\alpha}(G) = \bigwedge \{H \in \mathcal{G}(P) : H \text{ is partitioned}, G \leqslant H\}$ where $\bigwedge \mathcal{H}$ is the graph intersection of the graphs in the set $\mathcal{H}$, and where the notation $G \leqslant H$ indicates that G is a subgraph of H. The *single-linkage cluster method* is simply a function $\alpha: \mathcal{G}(P) \rightarrow \mathcal{P}(\mathcal{P}(P))$ defined as the function composition of $\underline{\alpha}$ and ω such that $\alpha(G) = \omega \cdot \underline{\alpha}(G) = \omega(\underline{\alpha}(G))$ for every G in $\mathcal{G}(P)$. When $\underline{\alpha}$ operates on the graph G in Fig. 1 (b), the partitioned graph in Fig. 1 (c) is obtained; when α operates on this same G, the simple clustering $\alpha(G) = \{\{a,b,c,d,e\}, \{f,g,h\}\}$ is obtained.

The single-linkage cluster method exhibits several characteristic properties. Its clusters are nonoverlapping and so are well separated. Efficient implementations of the method exist [9,21]. A single-linkage cluster may exhibit a low degree of internal coherence, as illustrated, for example, by the cluster $\{a,b,c,d,e\}$ in the graph in Fig. 1 (b). This same cluster is highly elongated or strung out, and thus exhibits the phenomenon known as chaining. In some applications, the existence of intermediate objects in a chain is adequate justification for classifying objects together in a cluster. Otherwise, the occurrence of chaining in single-linkage clusters makes the use of this cluster method inappropriate to the application.

Attempts to solve the chaining problem are of two types. In the first type, information concerning the relative similarities of pairs of objects is used to prevent the formation of highly elongated clusters. The *complete linkage method* introduced by Sørensen [24] and the *average linkage methods* of Sokal and Michener [23] are examples of this approach. These methods suffer from function discontinuities [13], so that trivial errors in the original similarity measurements may cause nontrivial discrepancies in the clusterings obtained.

Cluster methods of the second type replace highly elongated clusters by sequences of smaller, but overlapping, clusters. In the graph of Fig. 1 (b), for instance, it may be acceptable to recognize a sequence $\{a,b\}$, $\{b,c\}$, $\{c,d\}$, $\{d,e\}$ of overlapping clusters in place of the single linkage cluster $\{a,b,c,d,e\}$. Jardine and Sibson's *flat cluster methods* [13,20] are examples of methods using this approach; they result from a rigorous formal development of classificatory systems that addresses the chaining problem by permitting a restricted amount of cluster overlap.

Nevertheless the presence of cluster overlap in such methods brings concomitant problems in measuring the validity or relative significance of clusters, and these issues are addressed in subsequent sections. Flat cluster methods are formally introduced in Sec. 2, and many of the subsequent results are stated in terms of them. A desirable consistency property is proposed for graph-theoretic cluster methods: that cluster formation in each graph component should occur independently of the cluster formation in other graph components. A refinement of this idea leads in Sec. 3 to the concept of authentic cluster; nonauthentic, or specious, clusters are illustrated using the B_k flat cluster methods. These concepts of cluster validity are extended to cluster methods simply by calling a method authentic if each of its clusters is authentic. Several families of flat cluster methods based on various notions of connectivity are described in Sec. 4, and a characterization is obtained of the authentic methods among them. Since even among authentic clusters there is considerable variation in apparent internal coherence, two measures of cluster cohesion are described in Sec. 5 for flat cluster methods. Intuitively, the cohesion index measures how close a cluster is to being completely connected; while the attenuation index measures how close a cluster is to being unidentifiable.

2. FLAT CLUSTER METHODS

To assist in the investigation of flat cluster methods, Jardine and Sibson established a one-to-one correspondence [13,20] between each flat cluster method ρ: $\mathcal{G}(P)\to\mathcal{P}(\mathcal{P}(P))$ and a subset $\underline{\underline{\rho}}$ of $\mathcal{G}(P)$, called an indicator family of graphs, satisfying four graph-theoretic conditions. Frequently it is convenient to describe a flat cluster method in terms of its corresponding indicator family. To see how this is done, let $K(P)$ denote the complete graph having P as its point set and having every unordered pair of distinct points in P as a line. Furthermore, if G is a graph and if ϕ is a permutation of P, then the "relabeled" graph ϕG has P as its point set and $\{\{\phi(u),\phi(v)\}:\{u,v\}$ *is a line of* $G\}$ as its line set. An *indicator family on* P [13,20] is then a subset $\underline{\underline{\rho}}=\underline{\underline{\rho}}nP)$ of $\mathcal{G}(P)$ satisfying the following conditions:

(I1) If $\mathcal{F}\subseteq\underline{\underline{\rho}}$, then $\wedge\mathcal{F}\in\underline{\underline{\rho}}$ (closed under arbitrary intersection).

(I2) If $G \in \underline{\underline{\rho}}$ and ϕ is a permutation of P, then $\phi G \in \underline{\underline{\rho}}$ (closed under arbitrary relabeling).

(I3) $K(P) \in \underline{\underline{\rho}}$ non-empty.

(I4) $\underline{\underline{\rho}} \neq \{K(P)\}$ (nontrivial).

Given an indicator family $\underline{\underline{\rho}}$ on P, let $\rho : \mathcal{G}(P) \to \mathcal{G}(P)$ be the function such that $\rho(G) = \bigwedge \{H \in \underline{\underline{\rho}} : G \leqslant H\}$ for every G in $\mathcal{G}(P)$; then the function $\rho = \omega \cdot \underline{\rho} : \mathcal{G}(P) \to \mathcal{P}(\mathcal{P}(P))$ is the flat cluster method on P that corresponds to $\underline{\underline{\rho}}$.

It is easy to establish by indicator-family construction that there is one flat cluster method on two points and there are three flat cluster methods on three points. Table 1 gives a complete enumeration of the 33 flat cluster methods on four points; each indicator family appearing there is characterized by a maximum set of pairwise nonisomorphic graphs contained in the family. A preliminary investigation of mine indicates that there are 3208 indicator families on five points.

Four indicator families, denoted by $\underline{\underline{\tau}}$, $\underline{\underline{\sigma}}$, $\underline{\underline{\alpha}}$, and $\underline{\underline{\omega}}$, are easily defined on any finite set P having at least two points. Let $\overline{K}(P)$ denote the graph with point set P and having no lines. Then $\underline{\underline{\tau}}(P) = \{\overline{K}(P), K(P)\}$ is trivially an indicator family on P whose corresponding flat cluster method $\tau = \omega \cdot \underline{\tau}$ operates on graphs as follows: if $G \cong \overline{K}(P)$, then $\tau(G) = \{\{u\} : u \in P\}$ so that none of the points are clustered together; but if $G \neq \overline{K}(P)$, then $\tau(G) = \{P\}$ so that all of the points are clustered together. Let $\underline{\underline{\alpha}}(P)$ denote the set of all partitioned graphs in $\mathcal{G}(P)$; then $\underline{\underline{\alpha}}$ is the indicator family on P whose corresponding flat cluster method $\alpha = \omega \cdot \underline{\alpha}$ is the single-linkage cluster method. The subset $\underline{\underline{\sigma}}(P)$ of $\underline{\underline{\alpha}}(P)$, in which every graph has at most one nontrivial component, is an indicator family on P. Finally $\underline{\underline{\omega}}(P) = \mathcal{G}(P)$ is the indicator family on P whose corresponding flat cluster method is the clique cluster method ω.

Perhaps the best known flat cluster methods are the B_k, or *fine k-clustering*, methods introduced by Jardine and Sibson [11, 12, 13]. The corresponding indicator families $\underline{\underline{B}}_k$ are (somewhat awkwardly) described in graph-theoretic terms as follows. Let $G - x$ denote the graph having all of the points of G and having every line of G except x. If Q is a nonempty subset of P and if G is a graph in $\mathcal{G}(P)$, then the subgraph of G induced by Q is denoted by $\langle G|Q\rangle$; the spanning subgraph of G induced by Q always has the points of G as its points and is denoted by $\langle G|Q\rangle_s$. Then, for every positive integer k,

$$\underline{\underline{B}}_k(P) = \{G \in \mathcal{G}(P) : \text{if } |Q| \geqslant k+2 \text{ and } \langle K(P)|Q\rangle_s - x \leqslant G, \text{ then } x \text{ is also a line in } G\}.$$

It is easy to verify that $\underline{\underline{B}}_1(P) = \underline{\underline{\alpha}}(P)$ so that B_1 is in fact the single-linkage

TABLE 1
Flat Cluster Methods on Four Points

FLAT CLUSTER METHOD	INDICATOR FAMILY											PROPERTIES C = Consistent A = Authentic
$\rho_1 = \tau$	x										x	A
ρ_2	x	x									x	A
ρ_3	x	x	x								x	A
ρ_4	x			x							x	
ρ_5	x	x		x							x	A
ρ_6	x	x	x	x							x	A
$\rho_7 = \sigma$	x	x			x						x	A
ρ_8	x	x	x		x						x	A
$\rho_9 = \alpha$	x	x		x	x						x	C A
ρ_{10}	x	x	x	x	x						x	C A
ρ_{11}	x	x	x	x		x					x	A
ρ_{12}	x	x	x	x	x	x					x	C A
ρ_{13}	x	x					x				x	
ρ_{14}	x	x	x				x				x	A
ρ_{15}	x	x		x			x				x	
ρ_{16}	x	x	x	x			x				x	A
ρ_{17}	x	x	x		x		x				x	A
ρ_{18}	x	x	x	x	x		x				x	C A
ρ_{19}	x	x	x	x		x	x				x	A
ρ_{20}	x	x	x	x	x	x	x				x	C A
ρ_{21}	x	x	x	x	x	x	x	x			x	C A
ρ_{22}	x			x					x		x	
ρ_{23}	x	x		x					x		x	
ρ_{24}	x	x	x	x					x		x	
ρ_{25}	x	x	x	x	x				x		x	C
ρ_{26}	x	x	x	x		x			x		x	A
ρ_{27}	x	x	x	x			x		x		x	
ρ_{28}	x	x	x	x	x	x			x		x	C A
ρ_{29}	x	x	x	x	x		x		x		x	C
ρ_{30}	x	x	x	x		x	x		x		x	A
ρ_{31}	x	x	x	x	x	x	x		x		x	C A
ρ_{32}	x	x	x	x	x	x	x	x	x		x	C A
$\rho_{33} = \omega$	x	x	x	x	x	x	x	x	x	x	x	C A

cluster method; similarly, if $p=|P|$, then $\underline{\underline{B}}_{p-1}(P)=\underline{\underline{\omega}}(P)$ so that B_{p-1} is the clique cluster method. Since $\underline{\underline{B}}_k(P)\subseteq\underline{\underline{B}}_{k+1}(P)$ for every $1\leqslant k\leqslant p-2$, the B_k methods form a parameterized sequence between these extremes. The parameter k represents a restriction on the amount of cluster overlap, namely that distinct clusters in any B_k clustering may overlap in at most $k-1$ points. When B_2 operates on the graph G in Fig. 1 (b), the simple clustering $B_2(G)=\{\{a,b\},\{b,c\},\{c,d\},\{d,e\},\{f,g,h\}\}$ is obtained; the straggly cluster $\{a,b,c,d,e\}$ of the single-linkage method is again replaced by smaller overlapping clusters.

The B_k methods exhibit a desirable consistency property: that cluster formation in each component of a graph occurs independently of cluster formation in other components of the graph. Specifically, a cluster method ρ: $\mathcal{G}(P)\to\mathcal{P}(\mathcal{P}(P))$ is called *consistent*, if, for every cluster $Q\in\rho(G)$, there exists a component $\langle G|R\rangle$ of G containing Q such that $Q\in\rho(\langle G|R\rangle_s)$. The consistency of the B_2 method is illustrated in Fig. 2 by the graph G which has components $\langle G|Q_1\rangle$ and $\langle G|Q_2\rangle$; then

$$B_2(G)=\{\underline{\{a,b,c,d\}},\ \underline{\{e,f\}},\ \underline{\{f,g\}},\ \underline{\{g,h\}},\ \underline{\{e,h\}}\}$$

while $B_2(\langle G|Q_1\rangle_s)=\{\underline{\{a,b,c,d\}},\ \{e\},\{f\},\{g\},\{h\}\}$

and $B_2(\langle G|Q_2\rangle_s)=\{\{a\},\{b\},\{c\},\{d\},\underline{\{e,f\}},\underline{\{f,g\}},\underline{\{g,h\}},\underline{\{e,h\}}\}$.

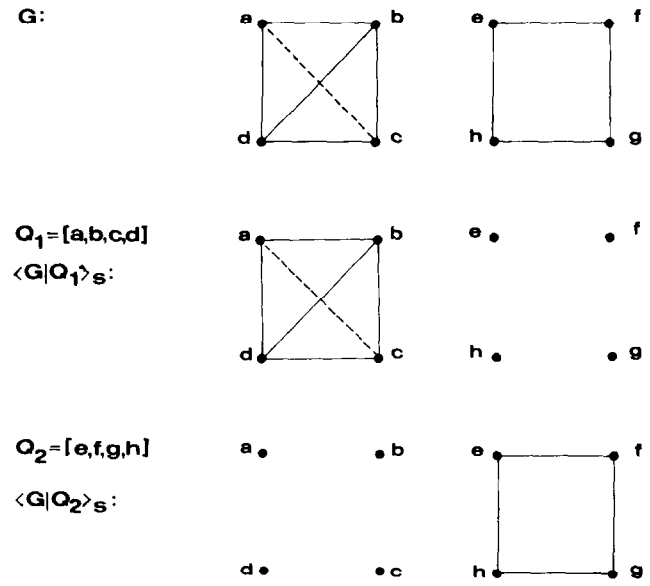


FIG. 2. Consistency in the B_2 method. [Dashed lines are in $\underline{B}_2(H)$ but are not in H.]

The consistent flat cluster methods have a useful characterization in terms of their indicator families. Let $\bigvee\mathfrak{F}$ denote the graph union of the graphs in the set $\mathfrak{F}$, and let two graphs be called line-independent if their sets of endpoints are disjoint. Then Day [2] has shown that a flat cluster method $\rho: \mathcal{G}(P) \to \mathcal{P}(\mathcal{P}(P))$ is consistent if ond only if the indicator family $\underline{\underline{\rho}}(P)$ satisfies the following conditions.

(I5) If $Q \subseteq P$ then $\langle K(P)|Q\rangle_s \in \underline{\underline{\rho}}$.

(I6) If the graphs in $\mathfrak{F} \subseteq \underline{\underline{\rho}}$ are pairwise line-independent, then $\bigvee\mathfrak{F} \in \underline{\underline{\rho}}$.

With this result, it is easy to see that only 12 of the 33 flat cluster methods in Table I are consistent.

3. AUTHENTIC AND SPECIOUS CLUSTERS

The B_2 method exhibits a second desirable consistency property: in every graph, formation of each cluster occurs independently of other cluster formation even within the same component of the graph. Specifically, given cluster method ρ: $\mathcal{G}(P) \to \mathcal{P}(\mathcal{P}(P))$ and graph G in $\mathcal{G}(P)$, a cluster Q in $\rho(G)$ is called *authentic* in G if Q also is a cluster in $\rho(\langle G|Q\rangle_s)$; otherwise Q is called *specious* in G. Clusters formed by the B_k methods may be specious when $k \geqslant 3$. Consider, for example, the clusters $Q_1 = \{a,b,c,d,e\}$ and $Q_2 = \{d,e,f,g\}$ formed when the B_3 method operates on the graph G in Fig. 3: Q_1 is authentic in G since $B_3(\langle G|Q_1\rangle_s) = \{Q_1, \{f\}, \{g\}\}$; but Q_2 is specious since $B_3(\langle G|Q_2\rangle_s) = \{\{a\}, \{b\}, \{c\}, \{d,f,g\}, \{e,f,g\}\}$. The points Q_2 seem inadequately linked in G, since the formation of Q_2 as a cluster depends on lines in G with endpoints external to Q_2. Put another way, the B_3 method allows clusters to overlap so that they are inadequately separated from each other; in the example, Q_2 is inadequately separated from Q_1.

Specious clusters may exhibit other unsound features: they may be very large; a great distance may separate a specious cluster from a graph feature affecting its formation; the points of a specious cluster in a graph may be very poorly connected in the graph. The B_3 method illustrates these features in the following examples.

Specious clusters may be very large. When the B_3 method operates on the graph G in Fig. 4 (a), it obtains the specious cluster $\{1,\ldots,100\}$ and the authentic cluster $Q = \{99,\ldots,103\}$. The formation of the specious cluster depends on the presence in G of each of the lines in $\langle G|Q\rangle$; for example, when the line $x = \{102,103\}$ is removed from G, then the clustering $B_3(G - x)$ contains 198 authentic clusters, each with three points. Generally, if $k \geqslant 3$ and $p = |P| \geqslant k+3$, then a graph G exists with the property that $B_k(G)$ contains a specious cluster on $p-k$ points.

A great distance may separate a specious cluster from a line affecting its formation. In this context, the distance between two points in a connected

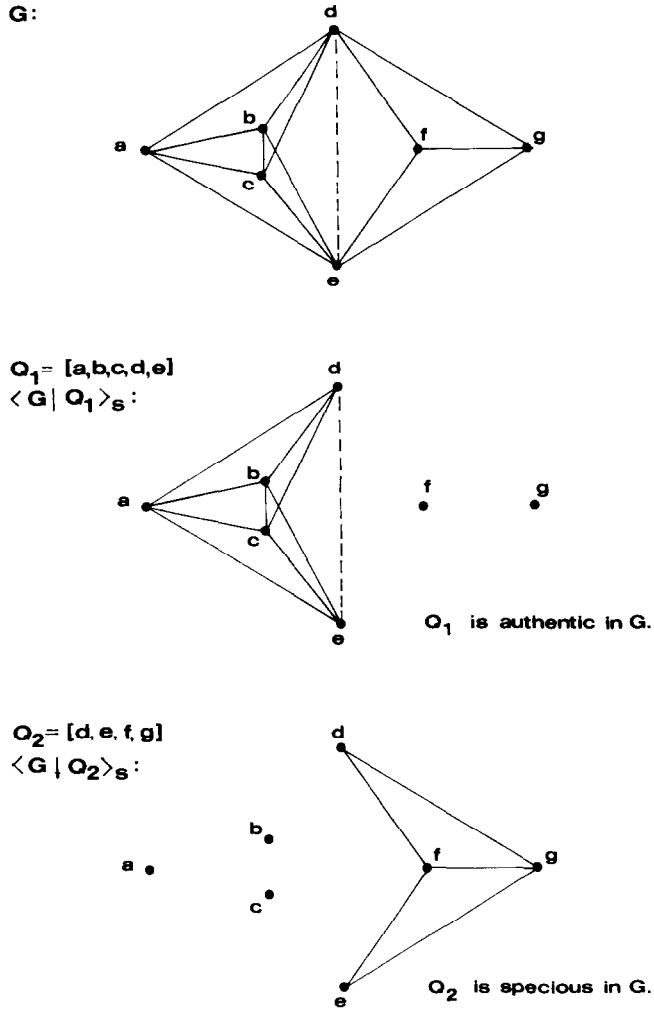
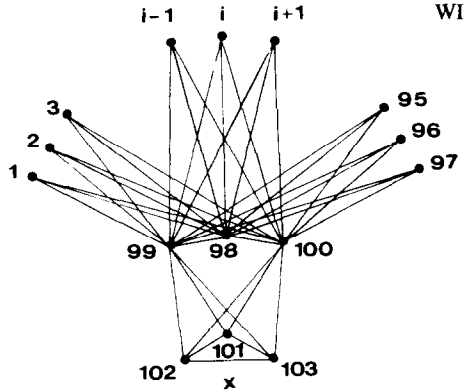
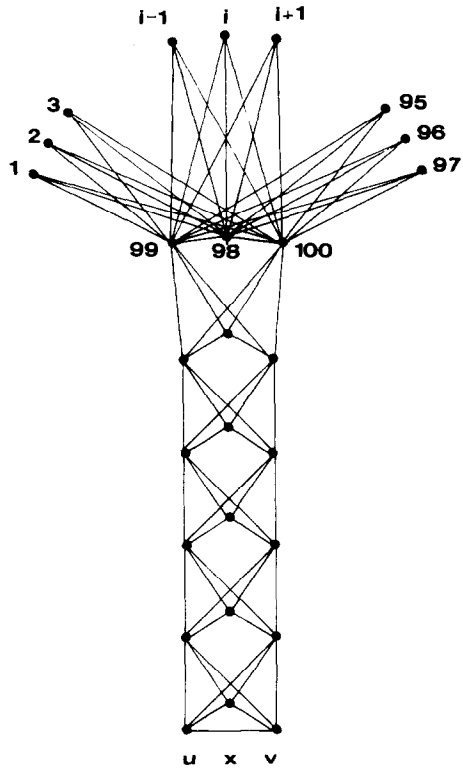


FIG. 3. Clusters formed by the B_3 method. [Dashed lines are in $B_3(H)$ but are not in H.]

graph is the length of a shortest path joining them, and the diameter of a connected graph is the greatest distance between any pair of points. The specious cluster $Q=\{1,\ldots,100\}$ forms when the B_3 method operates on the graph G in Fig. 4 (b); its formation depends on the line $x=\{u,v\}$ since $B_3(G-x)$ has clusters on three points only. The endpoints of x are a distance of 5 from the nearest points of Q while, by comparison, the diameter of G is 6. Generally, if $k \geqslant 3$, $d \geqslant 1$, and $p=|P| \geqslant kd+3$, then a



(a)



(b)

FIG. 4. Large specious clusters formed by the B_3 method.

graph G exists with diameter $d+1$ and with the property that $B_k(G)$ contains a specious cluster on $p-kd$ points that is a distance d in G from a line affecting its formation.

The points of a specious cluster in a graph may be very poorly connected in the graph. When the B_3 method operates on the graph G in Fig. 5, it obtains the specious cluster $Q=\{a,b,c\}$ and three authentic clusters on five points. The set Q in G is called independent since no two of its points are joined by a line. Specious clusters that are independent sets may be arbitrarily large. Generally, if $k \geqslant 3$, $m \geqslant 3$, and $p=|P|=km(m-1)/2+m$, then a graph G exists with the property that $B_k(G)$ contains $m(m-1)/2$ authentic clusters on $k+2$ points and one specious cluster on m points that is also an independent set.

The concept of cluster authenticity may be used to identify a class of cluster methods with a strict but gratifying property. A cluster method ρ: $\mathcal{G}(P) \to \mathcal{P}(\mathcal{P}(P))$ is called *authentic* if, for each graph G in $\mathcal{G}(P)$, every cluster in $\rho(G)$ is authentic; otherwise ρ is called *specious*. The α and ω cluster methods are always authentic, and it is easy to verify that 24 of the 33 flat cluster methods on four points are authentic. Consistency and authenticity in cluster methods are unrelated concepts in the sense that each may occur independently of the other. For example, in the flat cluster methods on four points: α is authentic and consistent; σ is authentic but inconsistent; ρ_{25} is consistent but specious; and ρ_4 is neither authentic nor consistent.

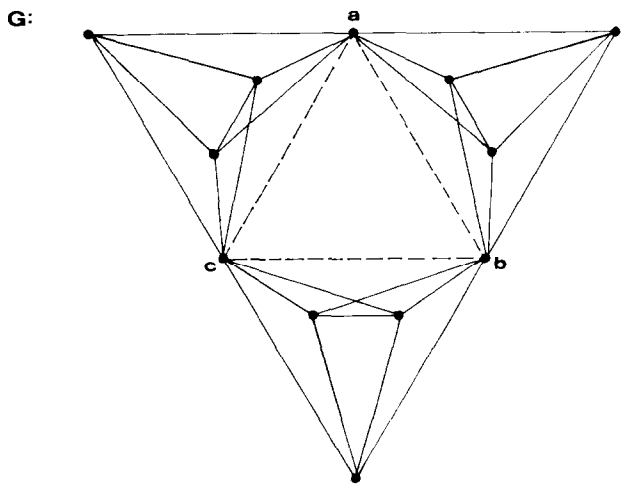


FIG. 5. A specious cluster formed by the B_3 method. [Dashed lines are in $B_3(G)$ but are not in G.]

Since the presence of specious clusters raises complex questions concerning cluster validity, the user of a cluster method may wish to establish whether the method is authentic. In a preliminary investigation of this problem, Day [3] obtained a graph-theoretic characterization of authentic flat cluster methods.

4. FLAT CLUSTER METHODS BASED ON CONNECTIVITY

In the last decade there has been a recurrent interest in relating the notion of cluster to graph-theoretic concepts of connectivity [4, 10, 15, 17]. Usually such cluster definitions identify as clusters the point sets of maximal subgraphs characterized by some measure of internal coherence or of isolation. In the single-linkage cluster method, for example, a cluster is the point set of a component, or maximal connected subgraph, of a graph. The cluster exhibits at least a minimal degree of internal coherence, since every two distinct points in the cluster are joined in the component by at least one path, and it is nicely isolated from other clusters, since points in distinct clusters are never joined by a path in the graph. As another example, a cluster identified by the clique cluster method is the point set of a clique, or maximal complete subgraph, of the graph. The cluster exhibits a high degree of internal coherence, since every two distinct points in the cluster are endpoints of a line in the graph, but it may be rather poorly isolated from other clusters, since a clique on n points may overlap another clique in as many as $n-1$ points. Other proposed cluster definitions have been based on concepts of minimum degree [14, 19, 25, 26], line connectivity [15, 16], and point connectivity [15], where a graph has minimum degree k if each point is adjacent to at least k other points; a graph is k-line-connected if every two distinct points are joined by at least k line-disjoint paths; and a graph is k-point-connected if every two distinct points are joined by at least k point-disjoint paths.

These cluster definitions all share a common structure: first a property ρ is selected that is applicable to subgraphs of a graph; then a cluster is identified as the point set of *ρ-component* or maximal subgraph exhibiting property ρ. Table 2 lists the properties mentioned here and gives also corresponding ρ-component names.

Each property in Table 2 may be used to construct a corresponding flat cluster method. To see this, for each property ρ in Table 2 let $\underline{\underline{\rho}}(P)$ denote the following subset of $\mathcal{G}(P)$:

$$\underline{\underline{\rho}}(P) = \{ G \in \mathcal{G}(P) : \text{every } \rho\text{-component in } G \text{ is complete} \}. \qquad (1)$$

THEOREM 1

Let P be a finite set of $p \geqslant 2$ points. If k is a positive integer, then $\underline{\underline{\alpha}}(P)$,

$\underline{\underline{\delta}}_k(P)$, $\underline{\underline{\lambda}}_k(P)$, $\underline{\underline{\kappa}}_k(P)$, *and* $\underline{\underline{\omega}}(P)$ *are consistent indicator families such that*

$$\underline{\underline{\alpha}}(P) \subseteq \underline{\underline{\delta}}_k(P) \subseteq \underline{\underline{\lambda}}_k(P) \subseteq K_k(P) \subseteq \underline{\underline{B}}_k(P) \subseteq \underline{\underline{\omega}}(P).$$

Furthermore,

$$\underline{\underline{\alpha}}(P) = \underline{\underline{\delta}}_1(P) = \underline{\underline{\lambda}}_1(P) = \underline{\underline{\kappa}}_1(P) = \underline{\underline{B}}_1(P)$$

and

$$\underline{\underline{\delta}}_{p-1}(P) = \underline{\underline{\lambda}}_{p-1}(P) = \underline{\underline{\kappa}}_{p-1}(P) = \underline{\underline{B}}_{p-1}(P) = \underline{\underline{\omega}}(P).$$

Proof. It is straightforward to establish that $\underline{\underline{\alpha}}(P)$, $\underline{\underline{\delta}}_k(P)$, $\underline{\underline{\lambda}}_k(P)$, $\underline{\underline{\kappa}}_k(P)$, and $\underline{\underline{\omega}}(P)$ each satisfy conditions (I1) through (I6) and so are consistent indicator families. Now $\underline{\underline{\alpha}}(P) \subseteq \underline{\underline{\delta}}_k(P)$, since every k-cluster is connected; $\underline{\underline{\delta}}_k(P) \subseteq \underline{\underline{\lambda}}_k(P)$, since every k-line-component is connected with minimum degree k; $\underline{\underline{\lambda}}_k(P) \subseteq \underline{\underline{\kappa}}_k(P)$, since every k-component is k-line-connected; $\underline{\underline{\kappa}}_k(P) \subseteq \underline{\underline{B}}_k(P)$, since two complete subgraphs on $k+1$ points are subgraphs of the same k-component if they overlap in k points; and $\underline{\underline{B}}_k(P) \subseteq \mathcal{G}(P) = \underline{\underline{\omega}}(P)$. Trivially, $\underline{\underline{B}}_1(P) \subseteq \underline{\underline{\alpha}}(P)$ and $\underline{\underline{\omega}}(P) \subseteq \underline{\underline{\delta}}_{p-1}(P)$.

The properties in Table 2 are examples of a special property type called *normal* property in [1]. In that paper Day established that if P is a finite set on at least two points and if ρ is a normal property, then $\underline{\underline{\rho}}(P)$ as defined by (1) is an indicator family on P.

TABLE 2
Concepts of Connectedness in Graphs

Property ρ	ρ-Component name	Indicator family
Connected	Component	$\underline{\underline{\alpha}}(P)$
Connected with minimum degree k	k-cluster [14]	$\underline{\underline{\delta}}_k(P)$
k-line-connected	k-line-component	$\underline{\underline{\lambda}}_k(P)$
k-point-connected	k-component	$\underline{\underline{\kappa}}_k(P)$
Complete	Clique	$\underline{\underline{\omega}}(P)$

It would be gratifying to find that the consistent flat cluster methods δ_k, λ_k, and κ_k are also authentic for arbitrary k. The next theorem establishes that in general this is not the case, although nontrivial authentic flat cluster methods of these types exist when $k=2$ and $|P| \geqslant 5$.

THEOREM 2

Let P be a finite set of $p \geqslant 2$ points. If k is a positive integer and if $\rho \in \{\delta_k, \lambda_k, \kappa_k\}$, then the flat cluster method ρ: $\mathcal{G}(P) \to \mathcal{P}(\mathcal{P}(P))$ is authentic if and only if $k \leqslant 2$ or $p \leqslant k+2$.

Proof. Let $k \leqslant 2$, and suppose that $Q \in \rho(G)$. If $\langle G|Q\rangle$ is complete or has property ρ, then $Q \in \rho(\langle G|Q\rangle_s)$. If $\langle G|Q\rangle$ is not complete and does not have property ρ, then there exists a clique $\langle \underline{\rho}(G)|R\rangle$ distinct from $\langle \underline{\rho}(G)|Q\rangle$ such that $|Q \cap R| \geqslant 2$ and $|R| \geqslant k+2$; but this is impossible, since $\langle \underline{\rho}(G)|Q \cup R\rangle$ also has property ρ. Hence ρ is an authentic method.

Let $p \leqslant k+2$ and suppose that $Q \in \rho(G)$. If $\langle G|Q\rangle$ is not complete, then G contains a noncomplete ρ-component, so that $k+2 \leqslant |Q| \leqslant p$; but then $Q = P$ and $G = \langle G|Q\rangle_s$, so that $Q \in \rho(\langle G|Q\rangle_s)$. If $\langle G|Q\rangle$ is complete, then $Q \in \rho(\langle G|Q\rangle_s)$. Hence ρ is an authentic method.

Let $k > 2$ and $p > k+2$, and suppose that $\{s,t,u,v_1,\ldots,v_k\} \subseteq P$. Let $G \in \mathcal{G}(P)$ be the graph in which $\langle G|\{v_1,\ldots,v_k\}\rangle$ is complete and both s and t are adjacent to every point in $\{u,v_1,\ldots,v_k\}$; then $\{s,t,u\}$ is a specious cluster in G, so that ρ is a specious method.

In like manner one can prove that the B_k method is authentic if and only if $k \leqslant 2$ or $|P| \leqslant k+2$.

It is disappointing to find all of these flat cluster methods specious when $k > 2$ and $|P| > k+2$. For the δ_k and λ_k methods, however, a simple criterion concerning cluster cardinality differentiates authentic and specious clusters.

THEOREM 3

Let $\rho : \mathcal{G}(P) \to \mathcal{P}(\mathcal{P}(P))$ be a flat cluster method where $\rho \in \{\delta_k, \lambda_k\}$ for positive integer k. If $G \in \mathcal{G}(P)$ and $Q \in \rho(G)$, then Q is authentic in G if and only if $|Q| > k$ or $\langle G|Q\rangle$ is complete.

Proof. Suppose that $|Q| > k$. If $\langle G|Q\rangle$ is complete or has property ρ, then $Q \in \rho(\langle G|Q\rangle_s)$. If $\langle G|Q\rangle$ is not complete and does not have property ρ, then there exists a clique $\langle \underline{\rho}(G)|R\rangle$ distinct from $\langle \underline{\rho}(G)|Q\rangle$ such that $|Q \cap R| \geqslant 2$ and $|R| \geqslant k+2$; but this is impossible, since $\langle \underline{\rho}(G)|Q \cup R\rangle$ also has property ρ. Hence Q is authentic in G.

Suppose that $\langle G|Q\rangle$ is complete; then $Q \in \rho(\langle G|Q\rangle_s)$, so that Q is authentic in G.

Suppose that $|Q| \leqslant k$ and $\langle G|Q\rangle$ is not complete. Since no subgraph of $\langle G|Q\rangle$ has property ρ, $\langle G|Q\rangle_s \in \underline{\underline{\rho}}(P)$, so that $Q \notin \rho(\langle G|Q\rangle_s)$. Hence Q is specious in G.

The analog of Theorem 3 is false for both κ_k and B_k methods, since specious clusters, when they exist, may contain as many as $|P|-3$ points. To see this, let $k \geqslant 3$ and $p = |P| \geqslant k+3$, and let P be partitioned into the sets $Q = \{q_1,q_2,q_3\}$, $R = \{r_1,\ldots,r_{k-1}\}$, and $S_i = \{s_i\}$ for $1 \leqslant i \leqslant p-k-2$. Let $H \in \mathcal{G}(P)$ be the graph in which the following subgraphs are complete: $\langle H|Q \cup R\rangle$; $\langle H|R \cup S_1\rangle$; and $\langle H|R \cup S_1 \cup S_i\rangle$ for $2 \leqslant i \leqslant p-k-2$. Finally, let $G = H - x$, where x is the line with r_1 and r_2 as endpoints; then $P \setminus Q$ is a specious cluster in G containing $p-3$ points. No larger specious

cluster is possible, since for κ_k (B_k) two distinct k-components (cliques) of a graph overlap in at most $k-1$ points [8]. The graphs in Figs. 3 and 4(a) illustrate this construction when $k=3$.

Another undesirable feature shared by both κ_k and B_k methods is that a specious cluster in a graph may also be an independent set in which no two points are joined by a line. To see this, let k and m be positive integers not less than 3, let $Q=\{q_1,\ldots,q_m\}$ be a subset of P, and for $1\leqslant i<j\leqslant m$ let S_{ij} be disjoint subsets of k points of $P\setminus Q$. Let $G\in\mathcal{G}(P)$ be the graph in which for each $1\leqslant i<j\leqslant m$, q_i and q_j are both adjacent to every point in S_{ij}, and $\langle G|S_{ij}\rangle$ is complete. Then Q is specious in G and also is an independent set in G. The graph in Fig. 5 illustrates this construction when $k=m=3$.

5. MEASURES OF CLUSTER COHESION

The classification of clusters as specious or authentic is a crude measure of cluster validity, since even among authentic clusters there is considerable variation in their apparent internal cohesion. The single-linkage cluster method identifies five-point clusters in both G_1 and G_{21} of Table 3, yet these graphs are extreme examples of the extent to which connected graphs on five points may be internally connected. In 1966 Estabrook [4] proposed, as a simple measure of cluster cohesion for the single-linkage method, a normalized measure of the number of cluster lines exceeding the minimum number necessary to connect the cluster points. This measure may be restated in a sufficiently general form to apply to the flat cluster methods described in Sec. 4.

Let $\rho:\mathcal{G}(P)\to\mathcal{P}(\mathcal{P}(P))$ be a flat cluster method based on the normal property ρ; let $Q\in\rho(G)$ be an authentic cluster on p points such that $\langle G|Q\rangle$ has q lines; and let $m(p,\rho)$ be the least number of lines in any graph $H\in\mathcal{G}(P)$ such that $Q\in\rho(H)$. Then the *cohesion index* $\chi(Q,\rho(G))$ is defined by

$$\chi(Q,\rho(G))=\frac{q-m(p,\rho)}{p(p-1)/2-m(p,\rho)}$$

when the denominator is nonzero, and is one otherwise. For the single-linkage method, it is well known that $m(p,\alpha)=p-1$, so that the cohesion index is [4]

$$\chi(Q,\alpha(G))=\frac{2(q-p+1)}{(p-1)(p-2)}$$

when $p>2$. The cohesion indices for the remaining flat cluster methods described in Sec. 4 based on the following results.

TABLE 3
Cohesion and Attenuation Indices for Clusters on Five Points

	GRAPH G WITH POINT SET Q																				
	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	G_{11}	G_{12}	G_{13}	G_{14}	G_{15}	G_{16}	G_{17}	G_{18}	G_{19}	G_{20}	G_{21}
No. of lines q	4	4	4	5	5	5	5	5	6	6	6	6	6	7	7	7	7	8	8	9	10
Minimum degree $\delta(G)$	1	1	1	1	1	1	1	2	1	1	2	2	2	1	2	2	2	2	3	3	4
Line connectivity $\lambda(G)$	1	1	1	1	1	1	1	2	1	1	2	2	2	1	2	2	2	2	3	3	4
Point connectivity $\kappa(G)$	1	1	1	1	1	1	1	2	1	1	1	2	2	1	2	2	2	2	3	3	4
$\chi(Q,\alpha(G))$	0	0	0	.17	.17	.17	.17	.17	.33	.33	.33	.33	.33	.5	.5	.5	.5	.67	.67	.83	1.0
$\psi(Q,\alpha(G))$	0	0	0	0	0	0	0	.33	0	0	.33	.33	.33	0	.33	.33	.33	.33	.67	.67	1.0
$\chi(Q,\delta_2(G)) = \chi(Q,\lambda_2(G))$	-	-	-	-	-	-	-	0	-	-	.2	.2	.2	-	.4	.4	.4	.6	.6	.8	1.0
$\psi(Q,\delta_2(G)) = \psi(Q,\lambda_2(G))$	-	-	-	-	-	-	-	0	-	-	0	0	0	-	0	0	0	0	.5	.5	1.0
$\chi(Q,\kappa_2(G))$	-	-	-	-	-	-	-	0	-	-	-	.2	.2	-	.4	.4	.4	.6	.6	.8	1.0
$\psi(Q,\kappa_2(G))$	-	-	-	-	-	-	-	0	-	-	-	0	0	-	0	0	0	0	.5	.5	1.0
$\chi(Q,B_2(G))$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	.33	.33	.67	1.0
$\psi(Q,B_2(G))$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	.5	.5	1.0

THEOREM 4 (Harary [7])

Let k and p be positive integers such that $1<k<p$, and let $\rho_k \in \{\delta_k, \lambda_k, \kappa_k\}$. Then $m(p,\rho_k)=(kp+\Delta)/2$, where $\Delta=0$ (1) when the product kp is even (odd).

THEOREM 5

If k and p are positive integers such that $1 \leqslant k<p$, then $m(p,B_k)=kp-k(k+1)/2$.

Proof. Let Q and G be such that $Q \in B_k(G)$ and $p=|Q|>k$; then $\langle G|Q\rangle$ has minimum degree not less than k and contains a complete subgraph on not less than $k+1$ points, so that $m(p,B_k) \geqslant k(k+1)/2+k(p-k-1)=kp-k(k+1)/2$. Let Q and G be such that $Q=\{v_1,\ldots,v_p\}$ and $\langle G|\{v_1,\ldots,v_k,v_{k+i}\}\rangle$ is complete for every $1 \leqslant i \leqslant p-k$; then $Q \in B_k(G)$ and G has $kp-k(k+1)/2$ lines so that $m(p,B_k) \leqslant kp-k(k+1)/2$.

Thus for $1<k<p-1$ the cohesion indices become

$$\chi(Q,\rho_k(G))=\frac{2q-kp-\Delta}{p(p-1)-kp-\Delta}$$

for $\rho_k \in \{\delta_k,\lambda_k,\kappa_k\}$, and

$$\chi(Q,B_k(G))=\frac{2q-2kp+k(k+1)}{p(p-1)-2kp+k(k+1)}.$$

Table 3 illustrates the use of these indices with clusters on five points.

The single-linkage method identifies as a cluster the point set Q of G_8 in Table 3 even if a line is deleted from G_8 before the method is applied; yet deletion from G_8 of two or more lines always prevents identification of Q as a cluster when the method is applied. This suggests that we might use, as a variant measure of the internal cohesion of a cluster Q in $\rho(G)$, a suitably normalized measure of the minimum number of lines whose deletion from $\langle G|Q\rangle$ prevents identification of Q as a cluster. This measure also may be stated in a form sufficiently general to apply to the flat cluster methods described in Sec. 4.

Let $\rho: \mathcal{G}(P) \to \mathcal{P}(\mathcal{P}(P))$ be a flat cluster method based on the normal property ρ; let $Q \in \rho(G)$ be an authentic cluster on p points such that $\langle G|Q\rangle$ has q lines; and let $n(Q,\rho(G))$ be the minimum number of lines whose deletion from $\langle G|Q\rangle$ prevents identification by ρ of Q as a cluster. Then the *attenuation index* $\psi(Q,\rho(G))$ is defined by

$$\psi(Q,\rho(G))=\frac{n(Q,\rho(G))-1}{\max\{n(Q,\rho(H)): Q \text{ is authentic in } H \in \mathcal{G}(P)\}-1}$$

when the denominator is nonzero, and is one otherwise. The line connectivity $\lambda(G)$ of a graph G is the minimum of lines whose removal results in a disconnected or trivial graph; it follows immediately that

$$\psi(Q,\alpha(G))=\frac{\lambda(\langle G|Q\rangle)-1}{p-2}$$

for $p>2$, while

$$\psi(Q,\lambda_k(G))=\frac{\lambda(\langle G|Q\rangle)-k}{p-k-1}$$

for $p>k+1$. The argument for the δ_k method is slightly more complex, since deletion of lines from $\langle G|Q\rangle$ may prevent the identification of Q as a cluster because the resulting graph either is disconnected or has a point with degree less than k. If $\delta(G)$ denotes the minimum degree among the points of G, then

$$\psi(Q,\delta_k(G))=\frac{\min\{\lambda(\langle G|Q\rangle),\delta(\langle G|Q\rangle)-k+1\}-1}{p-k-1}$$

for $p>k+1$. Table 3 also illustrates the use of these attenuation indices with clusters on five points.

The examples in Table 3 suggest that the attenuation index generally provides a stricter measure of cluster cohesion than the cohesion index. However, since computation of the attenuation index for these cluster methods requires calculation of the line connectivity, its time complexity is proportional to $p^{5/3}q$ [5]; by contrast, computation of the cohesion index has a time complexity proportional to q.

REFERENCES

1 W. H. E. Day, Specification of indicator families by graph-theoretic properties, Tech. Rep. CP 75018, Dept. of Comput. Sci., South. Methodist Univ., Dallas, Texas 75275, Aug. 1975.

2 W. H. E. Day, Flat cluster methods based on concepts of connectedness, Tech. Rep. CS 7606, Dept. of Comput. Sci., South. Methodist Univ., Dallas, Texas 75275, Apr. 1976.

3 W. H. E. Day, Flat cluster methods based on authentic indicator families, Tech. Rep. CS 76011, Dept. of Comput. Sci., South. Methodist Univ., Dallas, Texas 75275, Aug. 1976.

4 G. F. Estabrook, A mathematical model in graph theory for biological classification, *J. Theor. Biol.* **12**, 297–310 (1966).

5 S. Even and R. E. Tarjan, Network flow and testing graph connectivity, *SIAM J. Comput.* **4**, 507–518 (1975).

6 K. Florek, J. Lukaszewicz, J. Perkal, H. Steinhaus, and S. Zubrzycki, Sur la liason et la division des points d'un ensemble fini, *Colloq. Math.* **2**, 282–285 (1951).

7 F. Harary, The maximum connectivity of a graph, *Proc. Natl. Acad. Sci. USA* **48**, 1142–1146 (1962).

8 F. Harary, *Graph Theory,* Addison-Wesley, Reading, Mass., 1969, p. 47.

9 J. Hopcroft and R. Tarjan, Algorithm 447. Efficient algorithms for graph manipulation [H], *Commun. ACM* **16**, 372–378 (1973).

10 L. J. Hubert, Some applications of graph theory to clustering, *Psychometrika* **39**, 283–309 (1974).

11 N. Jardine and R. Sibson, A model for taxonomy, *Math. Biosci.* **2**, 465–482 (1968).

12 N. Jardine and R. Sibson, The construction of hierarchic and nonhierarchic classifications, *Comput. J.* **11**, 177–184 (1968).

13 N. Jardine and R. Sibson, *Mathematical Taxonomy,* Wiley, London, 1971, pp. 54–57, 61–63, 65–69, 92–101.

14 R. F. Ling, On the theory and construction of k-clusters, *Comput. J.* **15**, 326–332 (1972).

15 D. W. Matula, Cluster analysis via graph theoretic techniques, in *Proc. Louisiana Conf. on Combinatorics, Graph Theory, and Computing* (R. C. Mullin, K. B. Reid, and D. P. Roselle, Eds.), Univ. of Manitoba, Winnipeg, 1970, pp. 199–212.

16 D. W. Matula, k-components, clusters and slicings in graphs, *SIAM J. Appl. Math.* **22**, 459–480 (1972).

17 D. W. Matula, Graph theoretic techniques for cluster analysis algorithms, in *Advanced Seminar on Classification and Clustering* (J. Van Ryzin, Ed.), Academic, New York, 1977, pp. 95–129.

18 J. Moon and L. Moser, On cliques in graphs, *Isr. J. Math.* **3**, 23–28 (1965).

19 M. J. Shepherd and A. J. Willmott, Cluster analysis on the Atlas computer, *Comput. J.* **11**, 56–62 (1968).

20 R. Sibson, A model for taxonomy. II, *Math. Biosci.* **6**, 405–430 (1970).

21 R. Sibson, *SLINK*: an optimally efficient algorithm for the single-link cluster method, *Comput. J.* **16**, 30–34 (1973).

22 P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy,* Freeman, San Francisco, Calif., 1973, pp. 216–222.

23 R. R. Sokal and C. D. Michener, A statistical method for evaluating systematic relationships, *Univ. Kansas Sci. Bull.* **38**, 1409–1438 (1958).

24 T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, *Biol. Skr.* **5**, 1–34 (1948).

25 D. Wishart, A generalization of nearest neighbor which reduces chaining effects, in *Numerical Taxonomy* (A. J. Cole, Ed.), Academic, New York, 1969, pp. 282–311.

26 C. T. Zahn, Graph-theoretical methods for detecting and describing Gestalt clusters, *IEEE Trans. Comput.* **C-20**, 68–86 (1971).