Classify then Summarize or Summarize then Classify

Melvin F. Janowitz

DIMACS, Rutgers University
Piscataway, NJ 08854

Workshop Honoring Edwin Diday
held on September 4, 2007

What is Cluster Analysis?

- Software package?
- Collection of Computer Algorithms?
- Type of multivariate statistical analysis?
- Branch of discrete mathematics?
- Should not be recognized as a separate discipline.

Want this to be a discipline. So need a mathematical model.
Two views:

1. Input data has a true hierarchical structure.
   - Data you are given has possible errors.
   - Clustering estimates the true structure.
2. Cluster analysis suggests possible internal structure for data. The suggestions may or may not be valid.

Background From Jardine and Sibson

Book: Mathematical Taxonomy by N. Jardine and R. Sibson, Wiley, New York, 1971. This got me started.

Underlying finite set to be classified: E

$\Sigma(E)$ the reflexive symmetric relations on E.

Dissimilarity coefficient (DC) $d : E \times E \to \Re_0^+$

- $d(a, b) = d(b, a)$
- $d(a, a) = 0$

d is an ultrametric if also

- $d(a, b) \leq \max\{d(a, c), d(b, c)\}$ for all $a, b, c \in E$.

Numerically stratified clustering (NSC) $Td : \Re_0^+ \to \Sigma(E)$ a residual mapping (lattice theoretic idea) in that

- There is an h such that $Td(h) = E \times E$.
- $Td(\bigwedge h_i) = \bigcap Td(h_i)$.

NSCs and DCs are in one-one correspondence.

In the book a cluster method is viewed as a transformation of a DC to an ultrametric. Careful (but limited) mathematical model.

Connection with Symbolic Data Analysis

• If every object in E has a specified collection of attributes, it is straightforward to compute a DC.

• What if there is some variation or uncertainty involving the values of the attributes within an object?

• One could take a mean, or a median, or some other statistic summarizing the data belonging to each object.

This involves a summary before one classifies anything.
Might be wise to defer any summary as long as possible.

One might view the attributes as taking values in an interval. or one might view them as belonging to a distribution. This places us into the framework of symbolic data analysis, but it also puts us into a discipline called Percentile Clustering (Janowitz and Schweizer, Math. Social Sciences **18**, pp. 135-186). Included here would be dissimilarities taking values in a confidence interval. In these cases, we need to be able to have DCs taking values in a poset with smallest member 0.

What is a dissimilarity coefficient?

Mapping d from ordered pairs of objects to some partially ordered set (Often the non-negative reals).

Higher values of $d(x, y)$ make (x, y) more dissimilar (less similar). So $d(x, y)$ measures the dissimilarity.

- Another view:
- $d(x, y)$ represents the levels at which (x, y) is a candidate for clustering.
- Basic property: If (x, y) is a candidate for clustering at level h, and $h \leq k$, then (x, y) is a candidate at level k. k provides a less strict criterion.
- Cluster method: At each level h, decide which cluster candidates actually get clustered. View clusters as possible classifications.

Clustering Based on a Poset

L is poset with smallest element 0 where dissimilarities measured.

$\mathcal{F}(L) =$ order filters of L ordered by

$F \leq G \iff G \subseteq F$.

$F \neq \emptyset$, $x \in F, x \leq y$ implies $y \in F$.

Principal filter: $F_h = \{y \in L : y \geq h\}$.

$\mathcal{F}(L)$ is a complete distributive lattice.

DC: $D : E \times E \to \mathcal{F}(L)$ such that

$D(a, b) = D(b, a)$

$D(a, a) \leq D(a, b)$. Might want $D(a, a) = F_0$.

Can take $D(a, b)$ to be principal filters. Ultrametric if also

$D(a, b) \leq D(a, c) \vee D(b, c)$ for all $a, b, c \in E$.

$SD : L \to \Sigma(E)$ (Symmetric relations on E).

SD gives cluster candidates at level h.

$SD(h) = \{(a, b) : h \in D(a, b)\}$.

$h \leq k$ implies $SD(h) \subseteq SD(k)$.

If L has a largest member 1, then $SD(1) = E \times E$.

$h \in D(a, b) \iff (a, b) \in SD(h)$.

Detour into theory

If we take DCs as mappings $d\colon E \times E \to L$ where L is poset with 0, single linkage clustering not valid. Single linkage operation:

For $h \in L$, let $R_h = \{(a,b) : d(a,b) \leq h\}$. Output has at h the transitive relation $E_h = \gamma(R_h)$ generated by R_h. For this to work $\gamma(R_h \cap R_k)$ must equal $\gamma(R_h) \cap \gamma(R_k)$.

Not true unless L is a chain.

One solution: For L a chain, the map $Td : L \to \Sigma(E)$ has the property that the preimage of every principal filter is a principal filter. Can relax this to assume that each pre-image is the finite union of principal filters.

(Applications of the theory of partially ordered sets to cluster analysis, Banach Center Publications **9**, 1982, pp. 305-319.)

Another solution (Present view): Assume d takes values in $\mathcal{F}(L)$.

An example involving numerical data

	Water	Protein	Fat	Lactose	Ash
1. Bison	86.9	4.8	1.7	5.7	0.9
2. Buffalo	82.1	5.9	7.9	4.7	0.78
3. Camel	87.7	3.5	3.4	4.8	0.71
4. Cat	81.6	10.1	6.3	4.4	0.75
5. Deer	65.9	10.4	19.7	2.6	1.4
6. Dog	76.3	9.3	9.5	3	1.2
7. Dolphin	44.9	10.6	34.9	0.9	0.53
8. Donkey	90.3	1.7	1.4	6.2	0.4

Composition of Mammal Milk (Clustan)

Original data has 25 mammal species. Just wanted short example. Used DC taking values in $\Re_0^{+5}$ where $\Re_0^+$ denotes non-negative reals. Here is construction. Used squared Euclidean distance on each attribute to construct five separate DCs, then represent them as columns in a single dissimilarity matrix having 28 rows and 5 columns and denoted P. Use vector ordering inherited from $\Re_0^{+5}$.

A Version of Complete Linkage Clustering

We illustrate complete linkage clustering with the data at hand. The clusters implied by the minimal members $M(P)$ of P are all formed. We then remove them from P to form P_1. members of P_1. If k is such a level, we look at the clusters implied by the members of $M(P)$ strictly under k. These are all formed. To get the clusters at level k, we merge any clusters for which all links have been made (including any links at level k), and continue the process. We illustrate this numerically. Here is the list of edges of P.

level	edge	level	edge	level	edge	level	edge
1.	12	8.	23	15.	35	22.	48
2.	13	9.	24	16.	36	23.	56
3.	14	10.	25	17.	37	24.	57
4.	15	11.	26	18.	38	25.	58
5.	16	12.	27	19.	45	26.	67
6.	17	13.	28	20.	46	27.	68
7.	18	14.	34	21.	47	28.	78

Sample Calculations

The minimal members of $M(P)$ (height 0) are levels $\{1, 2, 7, 8, 9, 11, 19, 20, 21, 23, 24\}$. The next layer (height 1) of levels is $\{3, 5, 12, 14, 18, 26\}$.

Let's examine the clusters for these levels.

level 3 lies only above level 9. Thus we must cluster 24. level 3 has as a cluster candidate 14. In single linkage clustering we would have the cluster 124. But complete linkage would not merge 14 with 24 since there is no link between 1 and 2. Thus at level 3, 24 is the only non-singleton cluster.

Similar reasoning shows that at level 5, we have the cluster 12346, while complete linkage only has 1234. At level 12, we have

SL: 12347, 56 and CL: 1234, 56.

level	Single	Complete
14	234	24
18	138	13
26	123, 4567	123, 456

The nontrivial clusters

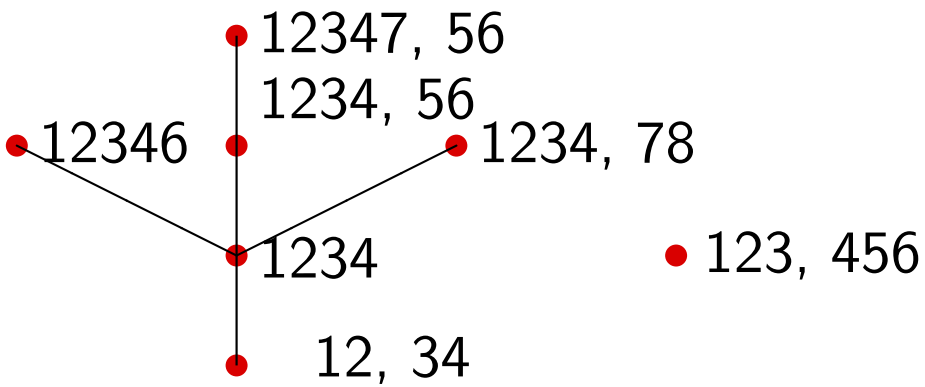

Figure: The nontrivial clusters (ignoring levels at which they occur)

Remember: Clusters just suggest structure!

1.	Bison	5.	Deer
2.	Buffalo	6.	Dog
3.	Camel	7.	Dolphin
4.	Cat	8.	Donkey

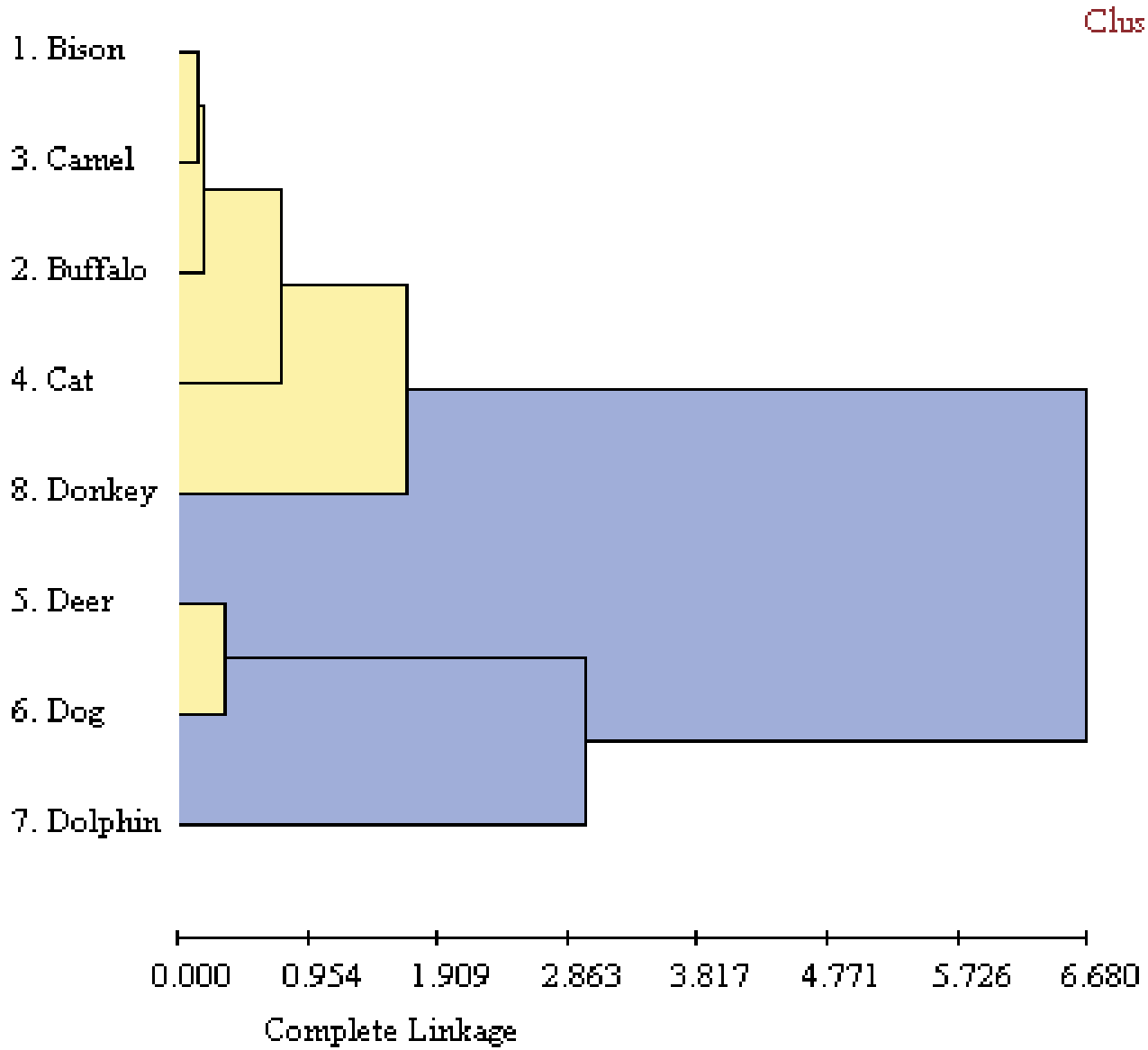

Figure: Complete linkage using standard clustering

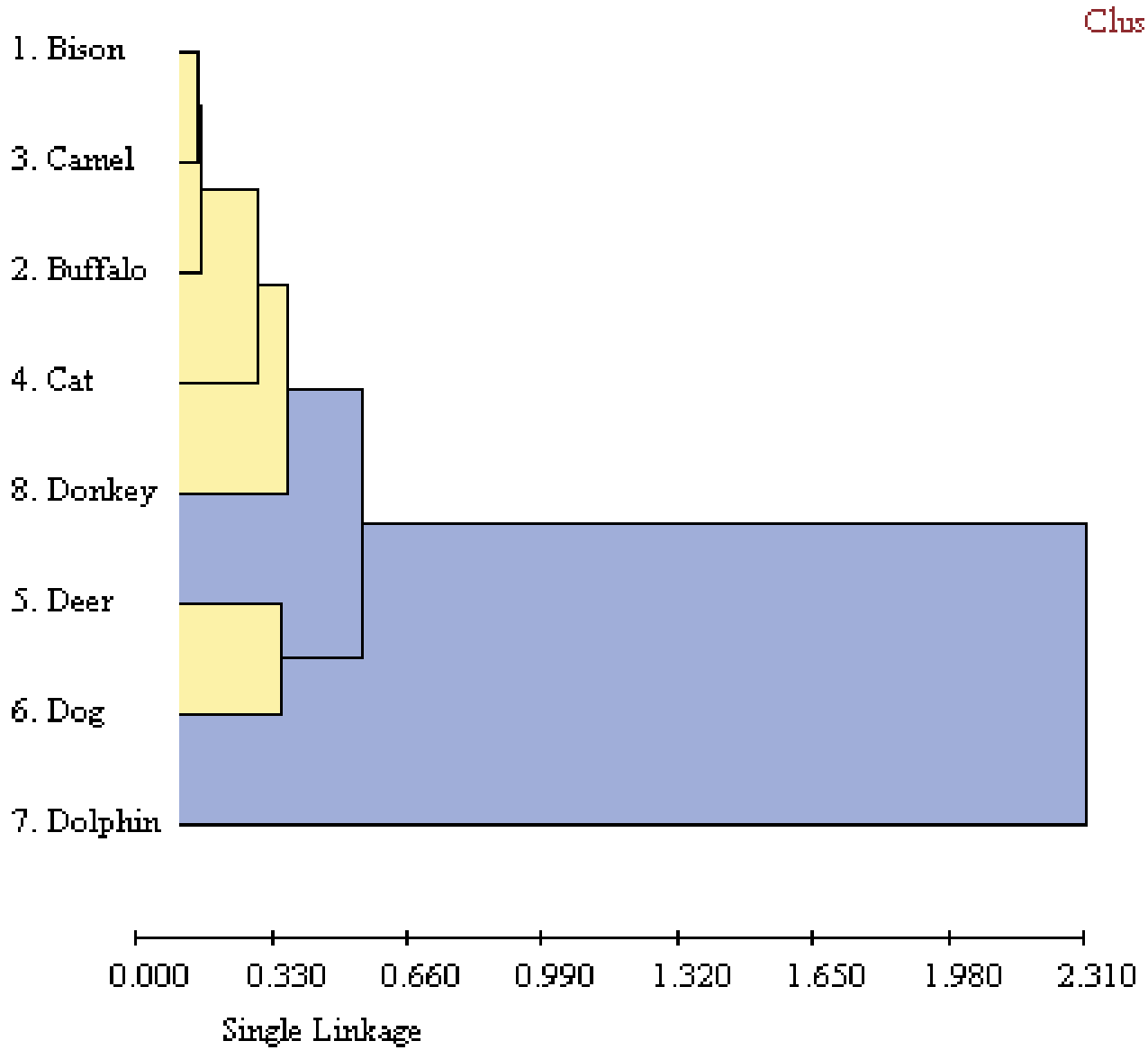

Figure: Single linkage using standard clustering

Vietnam casualties

Here is a second example. It relates to US and South Vietnamese combat deaths during the Viet Nam war over a period of 6 years. The data was taken from Hartigan, *Clustering Algorithms*, (Wiley, New York, 1975), p. 175. Will not repeat data here. Example was discussed in Janowitz and Schweizer, Ordinal and Percentile Clustering, Math. Social Sciences, **18** (1989), 135-186.

Description: 72 monthly totals, one for US and one for South Viet Nam. Period was from January, 1966 to December, 1971. Will label the data with the letters $a, b, c, d, e, f, g, h, i, j, k, l$ in chronological order.

Description of technique: We used Squared Euclidean distance to create a 72 by 72 dissimilarity matrix $\hat{d}$. The dissimilarity we are after is then based on the 12 groups of data (6 per group). Thus to get the dissimilarity $D(a, b)$, we use the distribution formed by the 36 entries $\{\hat{d}(i, j) : 1 \leq i \leq 6, 7 \leq j \leq 12\}$.

We take the 30th, 50th and 70th percentiles of this distribution. Thus the DC D is a 66 by 3 array of numbers. We order this array with the vector space ordering
$x \leq y \Longleftrightarrow$ row for $x \leq$ row for y. We label the groups with the letters from a to l, and these represent in chronological order the months

JAN-JUN, 1966, JUL-DEC, 1966, . . . , JUL-DEC, 1971.

This poset has one minimal member: the triple for ab, and two maximal members, the triples for be, and el. Let's see how the complete linkage algorithms works.

We begin by clustering ab at the level (340.3, 451.7, 1087.6). This level is not only minimal, it is the smallest member of D. There is only one entry at height 2, and it corresponds to cd. Thus at that level, we have the clusters ab and cd.

Height 3: the pairs bc, hj, jl.

For bc: $bc > cd > ab$, so we have the clusters ab, cd.

ab and cd do not merge to form abcd because no links at bd, ad, ac.

For hj: hj $> cd > ab$, so clusters ab, cd, hj are formed.

For jl: jl $> cd > ab$ so clusters ab, cd, jl.

Groups e, i and k seem to not cluster with other entries. The clusters that want to form involve ab, cd, fg, and hjl.

The next slide gives a graphic view of the data, and following that a slide that looks at the various clusterings.

A Graphic View of the Data

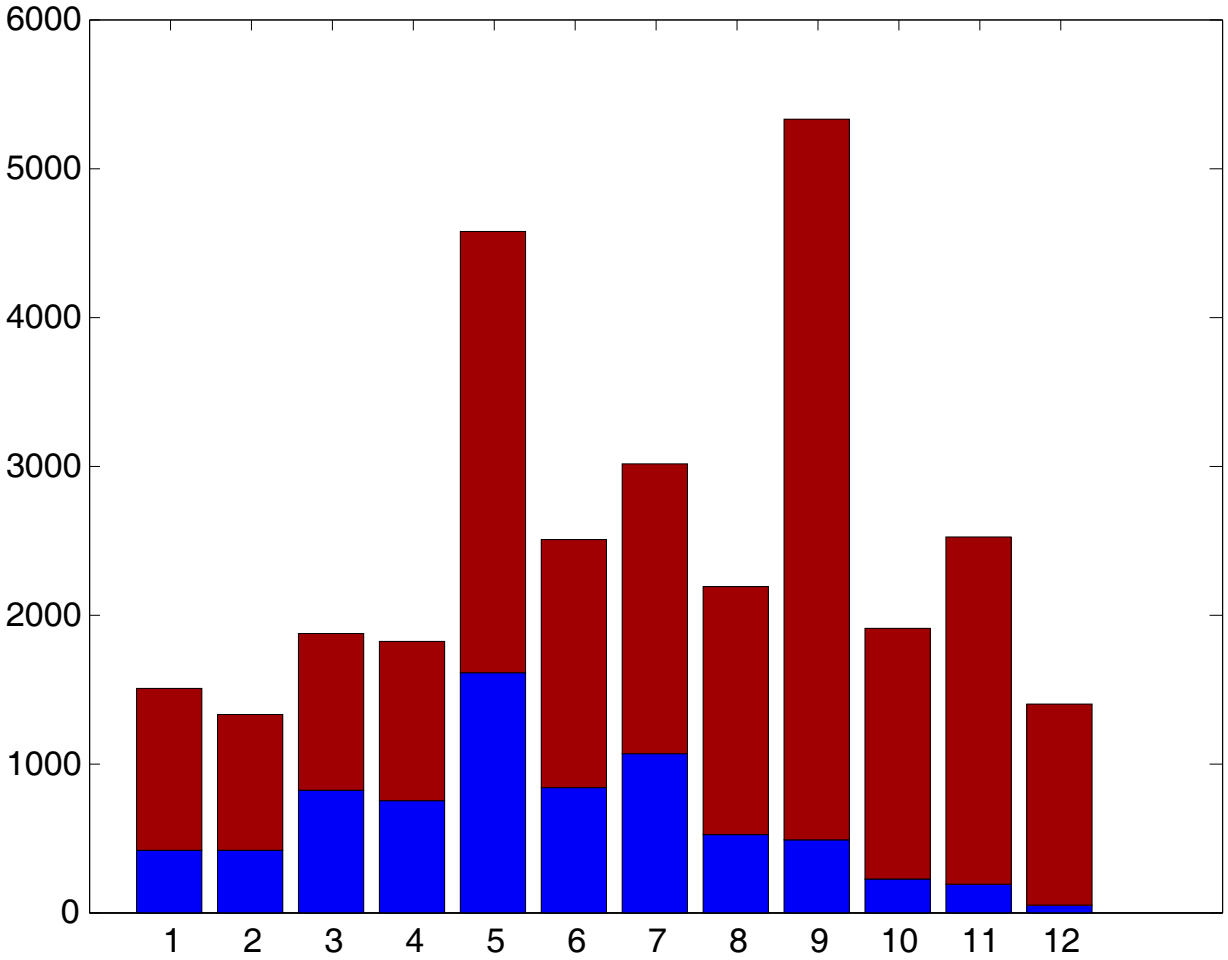

Figure: Vietnam Casualty Data

Note: Blue is US casualties, red is SVN.

A display of the clusters

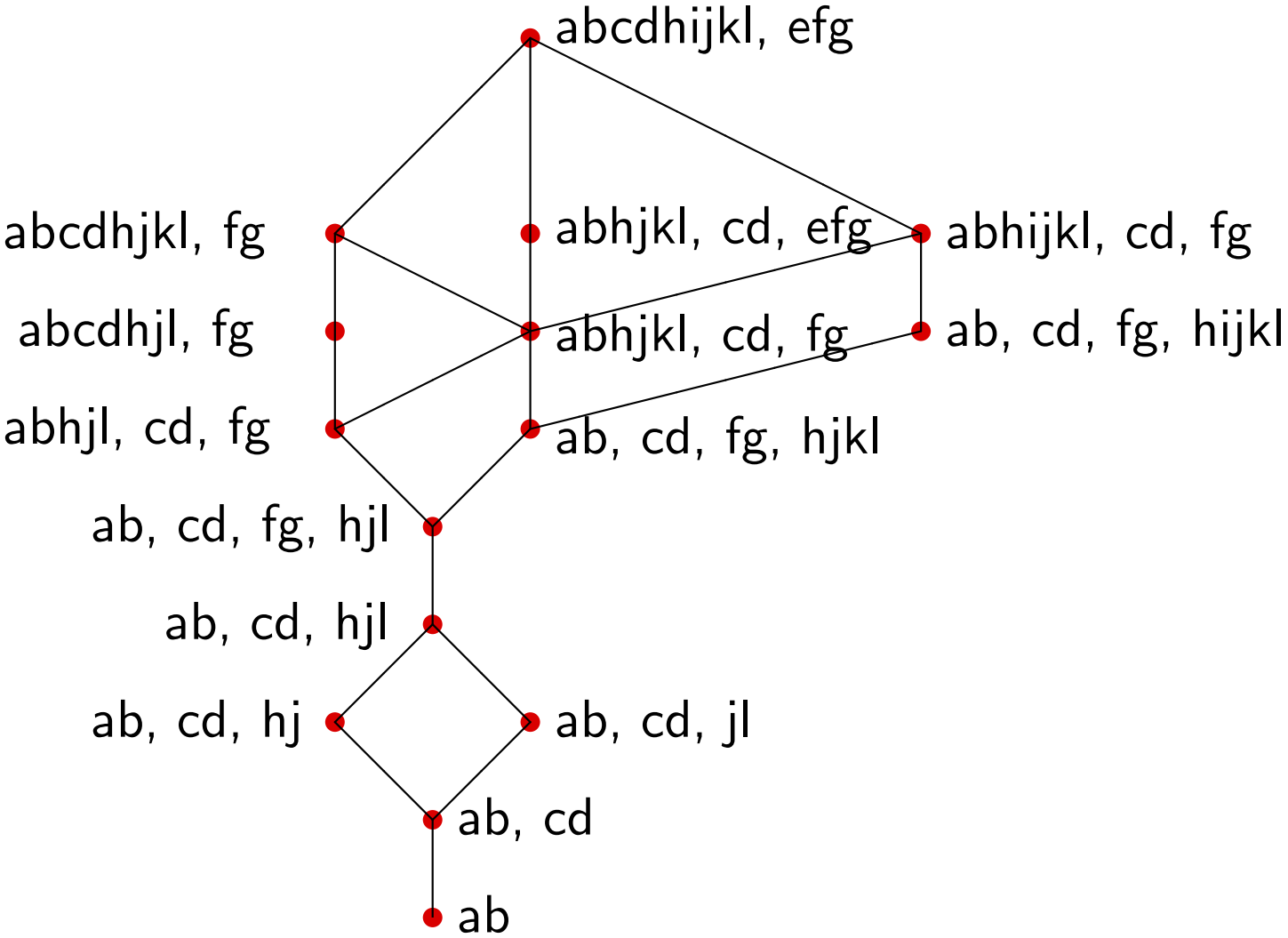

Figure: The Nontrivial Clusters

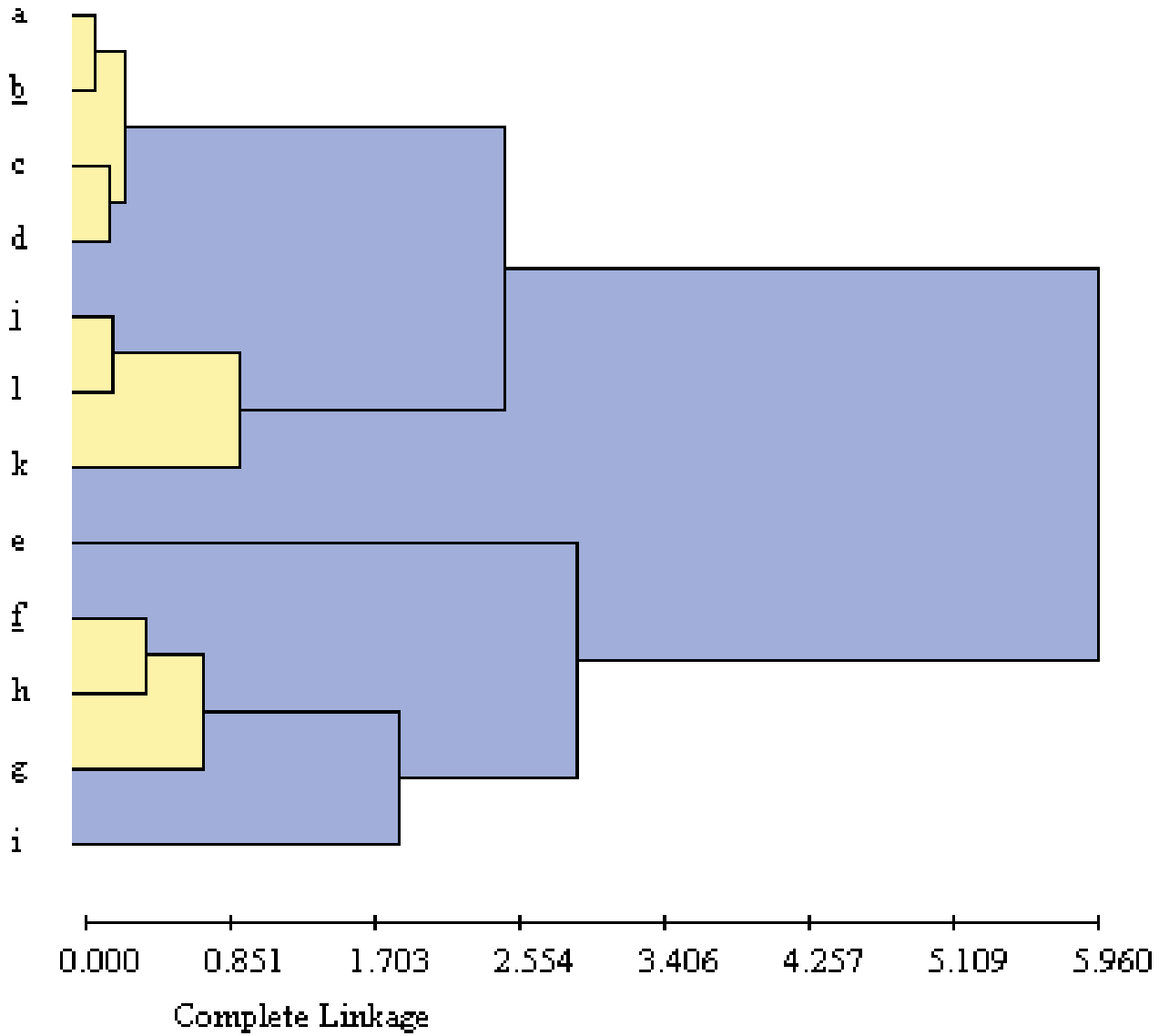

Figure: Clustering Based on Medians of each Group