Cluster Analysis

Range of techniques aimed at sorting data into clusters, with objects in a cluster more similar to each other than to those in other clusters.

"cluster" is not a well-defined term. There are many (elementary) approaches to clustering.

We'll focus only on "hierarchical clustering". This is a range of techniques which requires a notion of distance $d(x,y)$ between objects x, y to be clustered. The results of these techniques can be represented as dendrograms / phylogentic trees.

Example Distances between objects

	h	m	r	c	w
h	0	11	10	14	22
m	11	0	3	13	21
r	10	3	0	12	20
c	14	13	12	0	16
w	22	21	20	16	0

$V = \{h, m, r, c, w\}$

Table defines a metric on C.

$G(V, t)$ = graph with vertex set V, and one undirected edge $\{x, y\}$ for all $x, y \in V$ with $d(x, y) \leq t$.

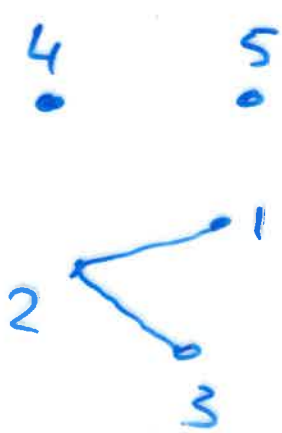
choose $t \geq 0$

for $t' > t$ we have an inclusion of graphs

$$G(V, t) \hookrightarrow G(V, t').$$

for instance,

$G(V, 10)$



See Computer

$\pi_0(G(V,10)) = \{x, y, z\}$

Say,

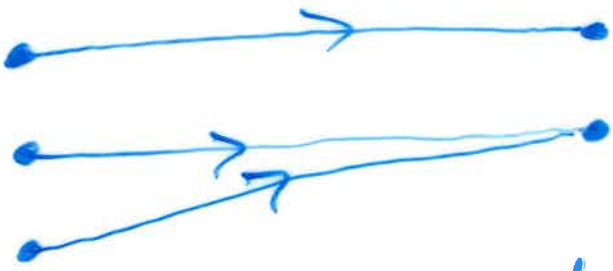
thus three connected components.

The construction π_0 induced a set-theoretic function

$$\pi_0(G(V,t)) \longrightarrow \pi_0(G(V,t'))$$

for any $t' > t$. For instance

$$\pi_0(G(V,10)) \longrightarrow \pi_0(G(V,14))$$



three components

two components.

The functions

$$\pi_0(G(V,t)) \to \pi_0(G(V,t'))$$

for $0 \le t < t' < \infty$ can be represented as a dendrogram.

See computer

The leaves of the dendrogram represent the objects to be clustered. Two objects x, y are considered to be in different clusters if the path from leaf x to leaf y is "long."

A <u>barcode</u> is obtained from a dendrogram by removing those edges that represent the merging of clusters.

see computer

See pdf slides