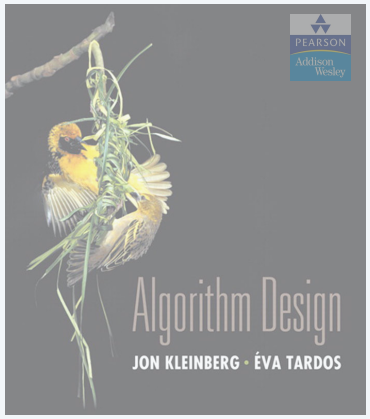


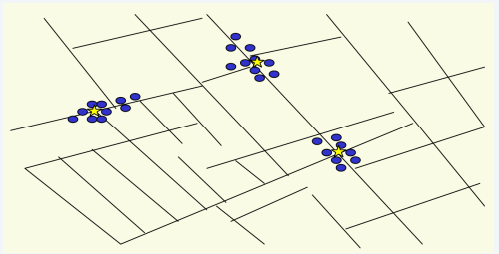
SECTION 4.7

4. GREEDY ALGORITHMS II

- *Dijkstra's algorithm*
- *minimum spanning trees*
- *Prim, Kruskal, Boruvka*
- *single-link clustering*
- *min-cost arborescences*

Clustering

Goal. Given a set U of n objects labeled $p_1, \ldots, p_n$, partition into clusters so that objects in different clusters are far apart.



outbreak of cholera deaths in London in 1850s (Nina Mishra)

Applications.

- Routing in mobile ad-hoc networks.
- Document categorization for web search.
- Similarity searching in medical image databases
- Cluster celestial objects into stars, quasars, galaxies.
- ...

Clustering of maximum spacing

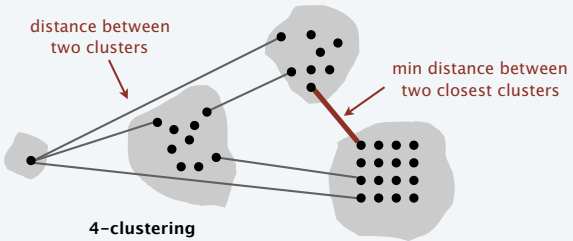
k-clustering. Divide objects into k non-empty groups.

Distance function. Numeric value specifying "closeness" of two objects.

- $d(p_i, p_j) = 0$ iff $p_i = p_j$ [identity of indiscernibles]
- $d(p_i, p_j) \geq 0$ [non-negativity]
- $d(p_i, p_j) = d(p_j, p_i)$ [symmetry]

Spacing. Min distance between any pair of points in different clusters.

Goal. Given an integer k, find a k-clustering of maximum spacing.

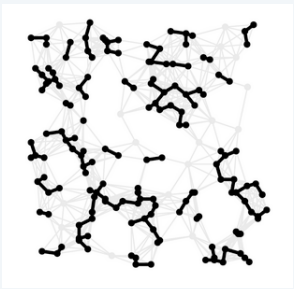


4-clustering

Greedy clustering algorithm

"Well-known" algorithm in science literature for single-linkage k-clustering:

- Form a graph on the node set U, corresponding to n clusters.
- Find the closest pair of objects such that each object is in a different cluster, and add an edge between them.
- Repeat $n - k$ times (until there are exactly k clusters).



Key observation. This procedure is precisely Kruskal's algorithm (except we stop when there are k connected components).

Alternative. Find an MST and delete the $k - 1$ longest edges.

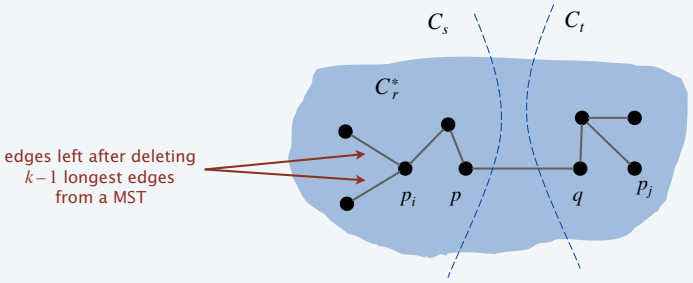
Greedy clustering algorithm: analysis

Theorem. Let C^* denote the clustering $C_1^*, \ldots, C_k^*$ formed by deleting the $k-1$ longest edges of an MST. Then, C^* is a k-clustering of max spacing.

Pf.

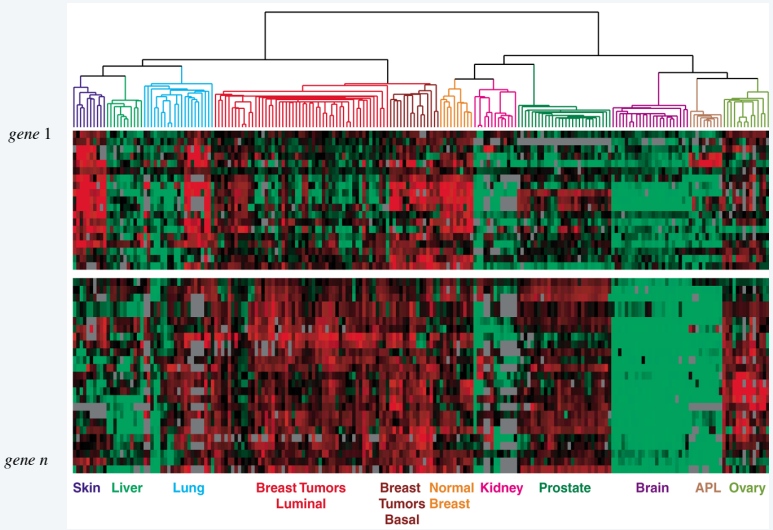
- Let C denote any other clustering $C_1, \ldots, C_k$.
- Let p_i and p_j be in the same cluster in C^*, say C_r^*, but different clusters in C, say C_s and C_t.
- Some edge (p, q) on $p_i - p_j$ path in C_r^* spans two different clusters in C.
- Spacing of C^* = length d^* of the $(k-1)^{\text{st}}$ longest edge in MST.
- Edge (p, q) has length $\le d^*$ since it was added by Kruskal.
- Spacing of C is $\le d^*$ since p and q are in different clusters. ▪

this is the edge Kruskal would have added next had we not stopped it



Dendrogram of cancers in human

Tumors in similar tissues cluster together.



Reference: Botstein & Brown group

gene expressed
gene not expressed

Minimum spanning trees: quiz 5

Which MST algorithm should you use for single-link clustering?

number of objects n can be very large

A. Kruskal (stop when there are k components).

B. Prim (delete $k-1$ longest edges).

C. Either A or B.

D. Neither A nor B.