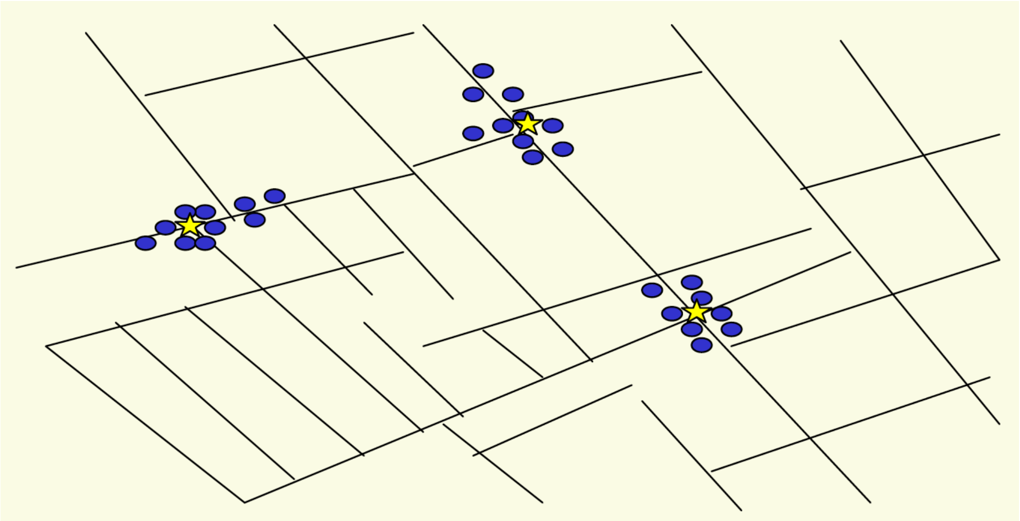
# 4.7 Clustering



Outbreak of cholera deaths in London in 1850s, analyzed by Dr. John Snow.
(Nina Mishra, HP Labs)

- k-Clustering with max-spacing problem
- Greedy solution (inc. proof)

# Clustering

**Clustering.** Given a set U of n objects labeled $p_1, \ldots, p_n$, classify into coherent groups.

↑ photos, documents, micro-organisms

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

↑ number of corresponding pixels whose intensities differ by some threshold

**Fundamental problem.** Divide into clusters so that points in different clusters are far apart.

- Routing in mobile ad hoc networks.
- Identify patterns in gene expression.
- Document categorization for web search.
- Similarity searching in medical image databases
- Skycat: cluster $10^9$ sky objects into stars, quasars, galaxies.

TUDelft

# Clustering of Maximum Spacing

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

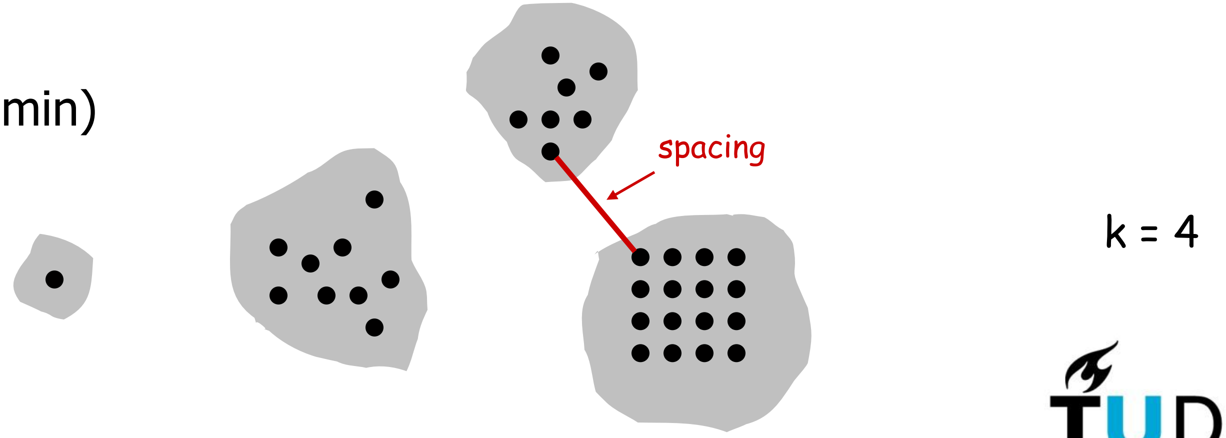
Distance function. Assume it satisfies several natural properties.

- $d(p_i, p_j) = 0$ iff $p_i = p_j$ (identity of indiscernibles)
- $d(p_i, p_j) \geq 0$ (nonnegativity)
- $d(p_i, p_j) = d(p_j, p_i)$ (symmetry)

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

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

Q. How? (1 min)



# Greedy Clustering Algorithm

Single-link k-clustering algorithm.

- Form graph (without edges) on vertex set, corresponding to n clusters.
- Find *closest pair of objects* from different clusters
- Add edge between them.
- Repeat n-k times until there are exactly k clusters left.

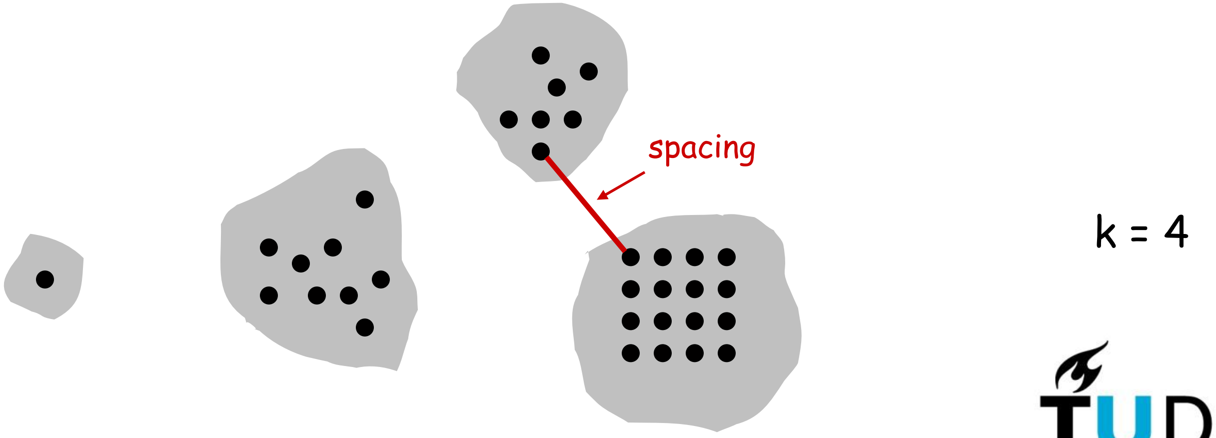
Key observation. This procedure is precisely Kruskal's algorithm (except we stop when there are k connected components).

Remark. Equivalent to finding an MST and deleting the k-1 most expensive edges (i.e. Reverse-Delete).
Each cluster has then a MST.

# Greedy Clustering Algorithm: Analysis

**Theorem.** Let C* denote the clustering $C^*_1, \ldots, C^*_k$ formed by deleting the k-1 most expensive edges of a MST by Kruskal. C* is a k-clustering of max spacing.

**Pf.** (standard optimality proof: any other cluster has smaller spacing)



# Greedy Clustering Algorithm: Analysis

Theorem. Let C* denote the clustering $C^*_1, \ldots, C^*_k$ formed by deleting the k-1 most expensive edges of a MST by Kruskal. C* is a k-clustering of max spacing.

Pf. (standard optimality proof: any other cluster has smaller spacing)

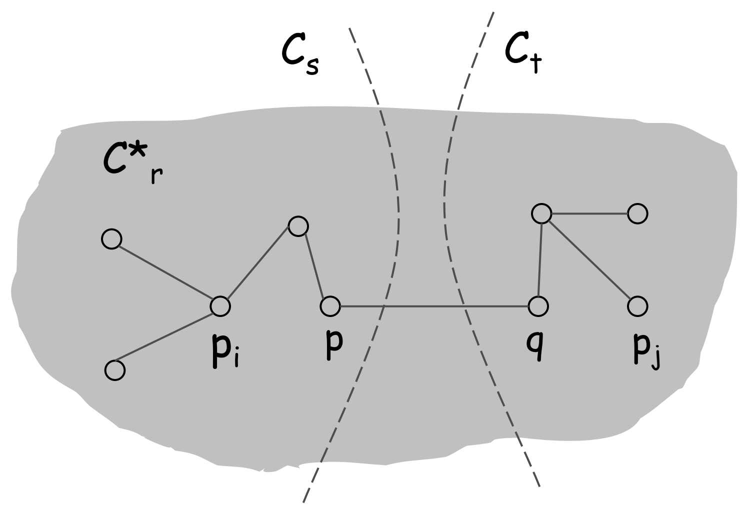
- The spacing of C* is the length d* of the $(k-1)^{st}$ most expensive edge.
- Let C denote some other clustering $C_1, \ldots, C_k$.
- Q. How do we know that spacing is less than (or equal to) d*?
- …
- Spacing of C is $\leq$ d*
- Nothing assumed of C, so holds for all C. ▪

# Greedy Clustering Algorithm: Analysis

**Theorem.** Let $C^*$ denote the clustering $C^*_1, \ldots, C^*_k$ formed by deleting the k-1 most expensive edges of a MST by Kruskal. $C^*$ is a k-clustering of max spacing.

**Pf.** (standard optimality proof: any other cluster has smaller spacing)

- The spacing of $C^*$ is the length $d^*$ of the $(k-1)^{st}$ most expensive edge.
- Let C denote some other clustering $C_1, \ldots, C_k$.
- Let $p_i$, $p_j$ be in the same cluster in $C^*$, say $C^*_r$, but different clusters in C, say $C_s$ and $C_t$.
- ...
- So spacing of C is $\leq d^*$ since p and q are in different clusters.
- Nothing assumed of C, so holds for all C. ▪



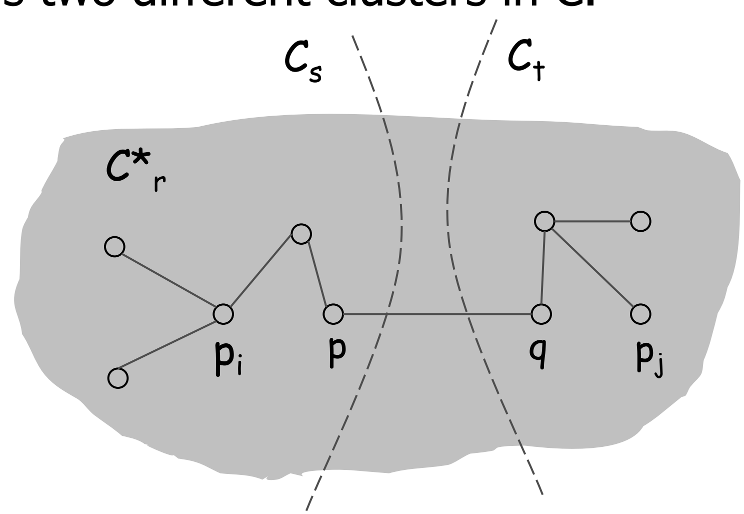

This proof can be found on page 160-161.

# Greedy Clustering Algorithm: Analysis

Theorem. Let $C^*$ denote the clustering $C^*_1, \ldots, C^*_k$ formed by deleting the k-1 most expensive edges of a MST by Kruskal. $C^*$ is a k-clustering of max spacing.

Pf. (standard optimality proof: any other cluster has smaller spacing)

- The spacing of $C^*$ is the length $d^*$ of the $(k-1)^{st}$ most expensive edge.
- Let C denote some other clustering $C_1, \ldots, C_k$.
- Let $p_i$, $p_j$ be in the same cluster in $C^*$, say $C^*_r$, but different clusters in C, say $C_s$ and $C_t$. In same cluster in $C^*$, so $p_i$-$p_j$ path in MST in $C^*_r$.
- Some edge (p, q) on $p_i$-$p_j$ path in $C^*_r$ spans two different clusters in C.
- All edges on $p_i$-$p_j$ path have length $\leq d^*$ since Kruskal chose them.
- So spacing of C is $\leq d^*$ since p and q are in different clusters.
- Nothing assumed of C, so holds for all C. ▪




This proof can be found on page 160-161.