Minimum Spanning Trees

Algorithms: Design and Analysis, Part II

Correctness of Greedy Clustering

Correctness Claim

Theorem: Single-link clustering finds the max-spacing k-clustering.

Proof: Let $C_1, \ldots, C_k =$ greedy clustering with spacing S.

Let $\hat{C}_1, \ldots, \hat{C}_k =$ arbitrary other clustering.

Need to show: Spacing of $\hat{C}_1, \ldots, \hat{C}_k$ is $\leq S$.

Correctness Proof

Case 1: $\hat{C}_i$'s are the same as the C_i's [maybe after renaming] $\Rightarrow$ has the same spacing S.

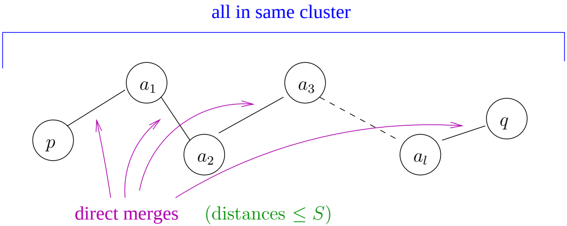
Case 2: Otherwise, can find a point pair p, q such that
(A) p, q in the same greedy cluster C_i
(B) p, q in different clusters $\hat{C}_i, \hat{C}_j$

Property of greedy algorithm: If two points x, y "directly merged at some point", then $d(x, y) \leq S$. [Distance between merged point pairs only goes up.]

Easy case: If p, q directly merged at some point, $S \geq d(p, q) \geq$ spacing of $\hat{C}_1, \ldots, \hat{C}_k$.

Correctness Proof (con'd)

Tricky case: p, q "indirectly merged" through multiple direct merges.



Let $p, a_1, \ldots, a_l, q$ be the path of direct greedy merges connecting p & q.

Key point: Since $p \in \hat{C}_i$ and $q \notin \hat{C}_i$, $\exists$ consecutive pair a_j, a_{j+1} with $a_j \in \hat{C}_i, a_{j+1} \notin \hat{C}_i \Rightarrow S \geq d(a_j, a_{j+1}) \geq$ Spacing of $\hat{C}_1, \ldots, \hat{C}_k$ QED!

since a_j, a_{j+1} directly merged

since a_j, a_{j+1} separated