# COMP229: Introduction to Data Science
# Lecture 25: hierarchical clustering

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

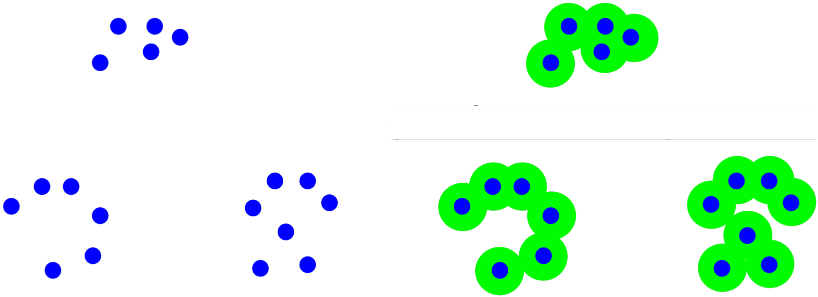
# The clustering problem

Given a cloud $C$ of points (a finite set with only pairwise distances), split the cloud $C$ into subsets called *clusters* in such a way that the points within the same cluster are more similar to each other than to points from all others cluster.

There are 1000s of clustering algorithms, because the concepts of a single cluster and a similarity within a cluster can be defined in many ways.

# Single-linkage clustering

**Definition 25.1.** The *single-linkage* clustering of a cloud $C$ with a distance threshold $\alpha > 0$ puts points $p, q$ into the same cluster if $d(p, q) \leq \alpha$.



Does it mean that any points within the same cluster are at most $\alpha$ away from each other?

# The potential weaknesses

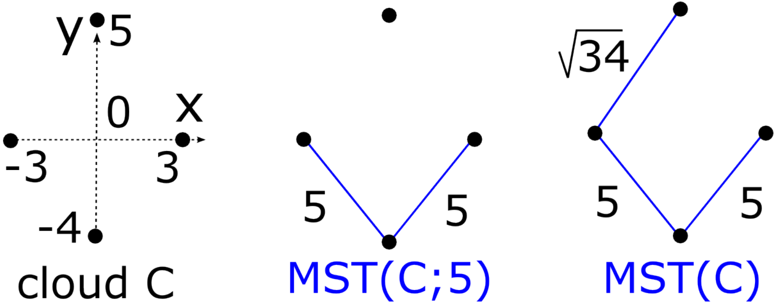
A threshold $\alpha$ can be hard to choose, e.g. for high-dimensional (not easily visualisable) data.

$C$ might have two big subsets connected by a chain of short links that are shorter than $\alpha$, which will lead to one cluster instead of two or more.

Other algorithms, e.g. the complete-linkage clustering, may resolve this difficulty by considering sizes of clusters or a density of points.

# A Minimum Spanning Tree MST($C$)

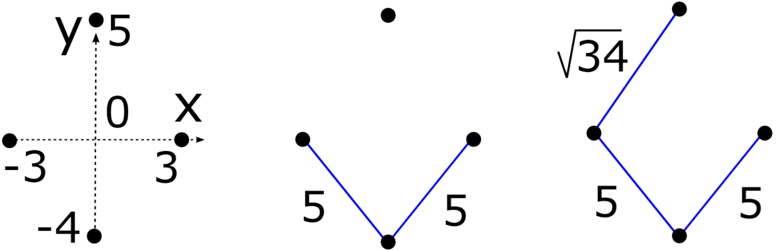
**Definition 25.2.** A *Minimum Spanning Tree* of a cloud $C$ is a tree (connected graph) that has the vertex set $C$ and the minimum total length.



If all pairwise distances between points of $C$ are different, then MST($C$) is unique.

# MST and single-linkage clustering

**Claim 25.3.** For any $\alpha > 0$, let $\mathrm{MST}(C;\alpha)$ be a Minimum Spanning Tree $\mathrm{MST}(C)$ without edges of length $> \alpha$. Then connected components of $\mathrm{MST}(C;\alpha)$ are *single-linkage clusters* of $C$.



cloud C

MST(C;5)

MST(C)

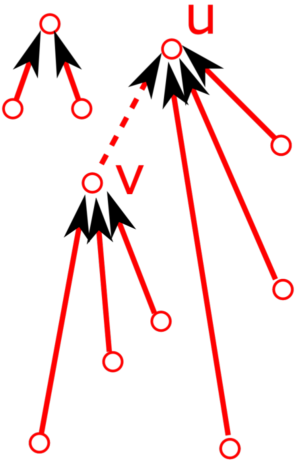
# MST($G$) of a connected graph $G$

**Definition 25.4.** Let $G$ be a graph whose edges have lengths (weights). MST($G$) is a tree that has all the vertices of $G$ and a minimum total length.

**Claim 25.5.** If a graph $G$ has $m$ edges, then MST($G$) can be computed in time $O(m \log m)$.

*Outline.* All edges are sorted in the increasing order in time $O(m \log m)$. We start from $n$ isolated vertices and add one edge at a time. Every vertex has a parent link (initially to itself) to another vertex in the same connected component.

# The union–find structure

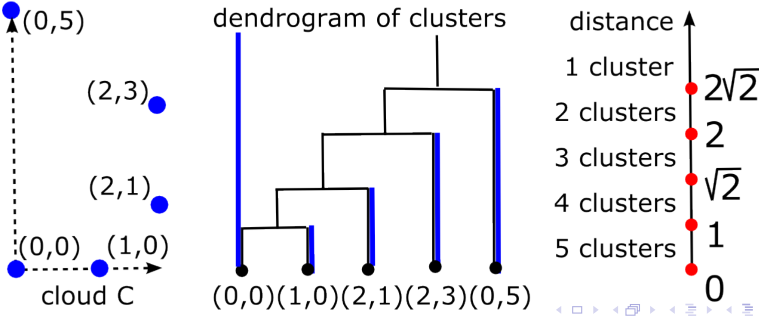


When an edge between $u, v$ is added, we check if $u, v$ are in the same component by comparing their roots (highest parents) obtained by going up along parent links. If $u, v$ are in different components, their roots are joined by a new parent link. Continue until a forest becomes one big tree.

Join a shorter tree to a higher one. The maximum height is $O(\log n)$, this is the time is per edge. □

# The dendrogram of clustering

**Definition 25.6.** In the single-linkage clustering, when clusters (or trees in a union-find structure) merge (become connected), the *dendrogram* is an abstract diagram visualising these mergers below.



# Hierarchical clustering

**Definition 25.7**. A *hierarchical* clustering builds a hierarchy of clusters such that any cluster at a lower level is a subset of a clusters at a higher level, hence can be described by a dendrogram.

The typical strategies are *agglomerative* (merging smaller clusters into larger ones) or *divisive* (subdividing large clusters into smaller ones).

Other hierarchical clustering algorithms are complete-linkage, mean-linkage, centroid-linkage.

# Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2–3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

**Question**. For a cloud of $n$ points $C \subset \mathbb{R}^2$, what's the complexity to find single-linkage clusters?

# Answer to the quiz and summary

**Answer**. The brute-force algorithm through a complete graph on any cloud $C$ needs $O(n^2 \log n)$ time. We'll discuss $O(n \log n)$ time for $C \subset \mathbb{R}^2$.

- *Single-linkage* clusters for a threshold $\alpha$ are equivalence classes of the relation generated by the distance condition $d(p, q) \leq \alpha$.
- A *Minimum Spanning Tree* of a cloud $C$ has the vertices $C$ and the minimum total length.
- MST($G$) of a graph $G$ with $m$ edges is found in time $O(m \log m)$ by a union-find structure.