Graph Mining
SD212
6. Hierarchical clustering

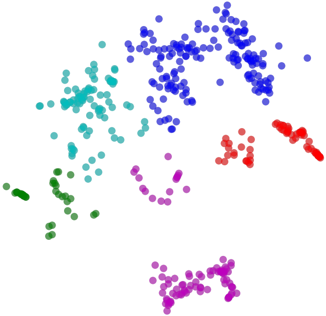
Thomas Bonald

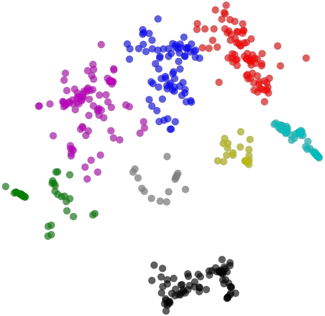
2018 – 2019

TELECOM ParisTech

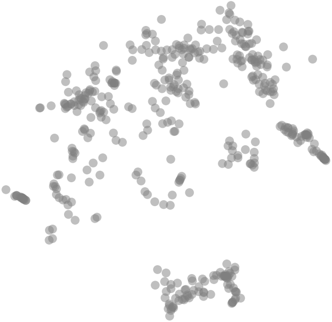
Motivation

- What is a **good** clustering?
- Which **resolution**?

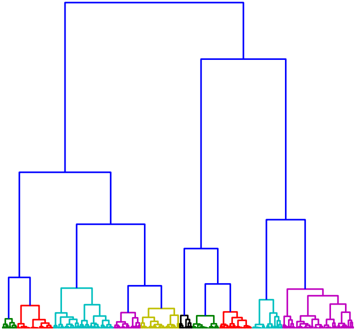




Hierarchical clustering



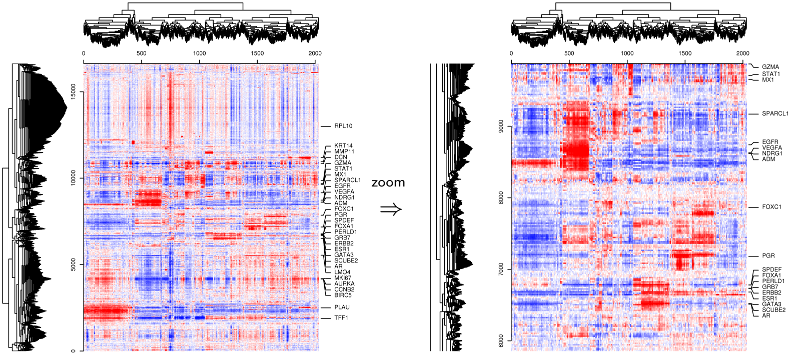
Data



Binary tree

Example in biology

2,035 tumors, 16,634 non-redundant genes



Wirapati 2009

Hierarchical clustering: vector data

Divisive algorithms

- e.g., through successive k-means

Agglomerative algorithms

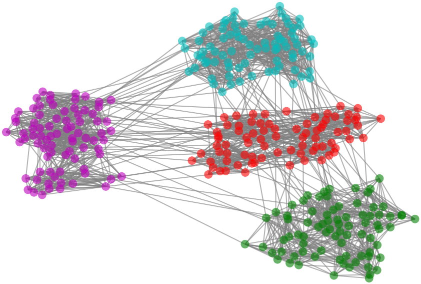
- Successive merges of the closest clusters $a, b \subset \{1, \ldots, n\}$

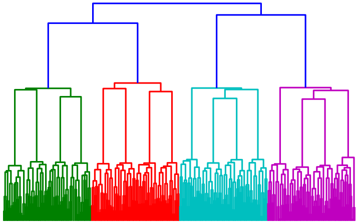
Linkage	$d(a, b)$
Single	$\min_{i\in a, j\in b} \|\|x_i - x_j\|\|$
Complete	$\max_{i\in a, j\in b} \|\|x_i - x_j\|\|$
Average	$\frac{1}{\|a\|\|b\|} \sum_{i\in a, j\in b} \|\|x_i - x_j\|\|$
Ward	$\frac{\|a\|\|b\|}{\|a\|+\|b\|} \|\|g_a - g_b\|\|^2$

Lance & Williams 1967

- Local search by the **nearest-neighbor chain**
  Murtagh 1983

Hierarchical clustering: graph data

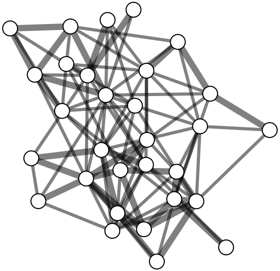




Framework

Consider a weighted, undirected, connected graph $G = (V, E)$

- n nodes, m edges
- A, weighted adjacency matrix
- $w_i = \sum_j A_{ij}$, weight of node i
- $v = \sum_i w_i = \sum_{i,j} A_{ij}$,
  volume of the graph



Outline

1. Node sampling
2. Agglomerative algorithm
3. Notion of dendrogram
4. Link with modularity
5. Dendrogram cuts

Node sampling

- The edges of the graph induce a probability distribution on node pairs:

$$p(i,j) = \frac{A_{ij}}{v}$$

- Marginal distribution:

$$p(i) = \sum_{j \in V} p(i,j) = \frac{w_i}{v}$$

- Conditional distribution:

$$p(i|j) = \frac{p(i,j)}{p(j)} = \frac{A_{ij}}{w_j}$$

Outline

1. Node sampling
2. **Agglomerative algorithm**
3. Notion of dendrogram
4. Link with modularity
5. Dendrogram cuts

Agglomerative algorithm

Idea: Merge successively the two "closest" nodes

Linkage

$$\sigma(i,j) = \frac{p(j|i)}{p(j)} = \frac{p(i|j)}{p(i)} = \frac{p(i,j)}{p(i)p(j)} = v\frac{A_{ij}}{w_i w_j}$$

Algorithm

Input: Graph $G = (V, E)$ with $V = \{1, \dots, n\}$

For $t = 1, \dots, n-1$

- $i, j \leftarrow \arg\max_{i,j \in V, i \neq j} \sigma(i,j)$
- merge i, j into node $n + t$
- update σ

Output: List of merges

Merging two nodes

Denote by $i \cup j$ the node resulting from the merge of i and j:

$$p(i \cup j, k) = p(i, k) + p(j, k), \quad \forall k \in V \setminus \{i, j\}$$

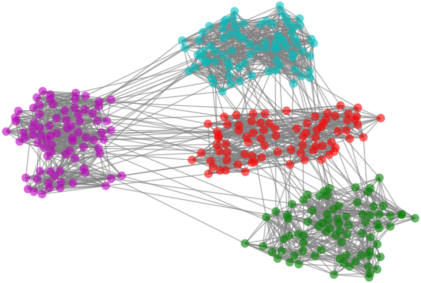
$$p(i \cup j) = p(i) + p(j)$$

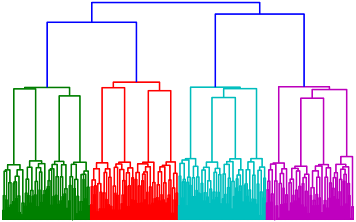
Update formula

$$\forall k \neq i, j, \quad \sigma(i \cup j, k) = \frac{p(i)}{p(i) + p(j)} \sigma(i, k) + \frac{p(j)}{p(i) + p(j)} \sigma(j, k)$$

Outline

Dendrogram





Distance

$$d(i,j) = \frac{1}{\sigma(i,j)} = \frac{w_i w_j}{v A_{ij}}$$

Reducibility

Proposition

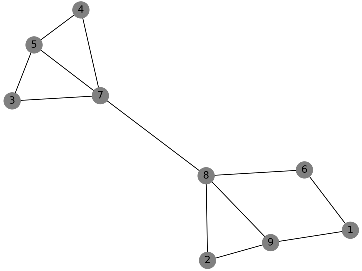
$$d(i \cup j, k) \geq \min(d(i, k), d(j, k))$$

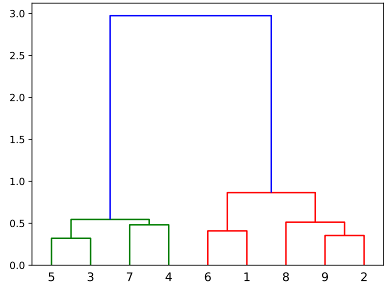
Consequence

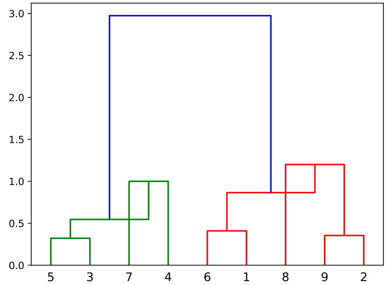
The distances of successive merges is non-decreasing:

$$d_1 \leq d_2 \leq \ldots \leq d_{n-1}$$

Regular vs. non-regular dendrograms







The nearest-neighbor chain

The complexity of the basic algorithm is in $O(n^3)$.
More efficient approach through the **nearest-neighbor chain**:

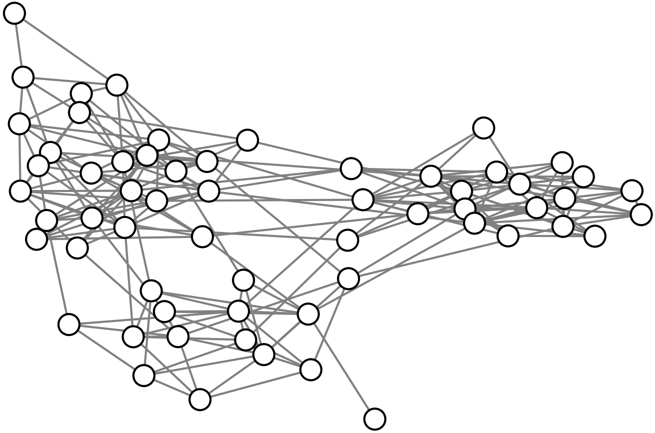
Algorithm

Input: Graph $G = (V, E)$ with $V = \{1, \dots, n\}$
While $|V| > 1$:

- take a node at random
- build the chain of nearest-neighbors
- merge the two last nodes of this chain
- update σ
- restart the chain

Output: List of merges

Example



Loops of the nearest-neighbor chain

Assume you get a loop in the NN chain: $i \to j \to k \to i$

Outline

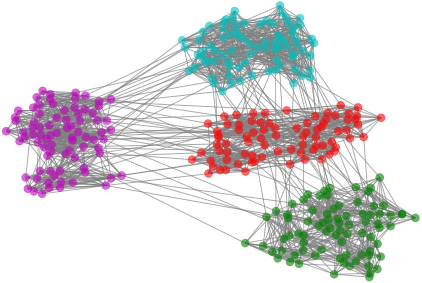
1. Node sampling
2. Agglomerative algorithm
3. Notion of dendrogram
4. **Link with modularity**
5. Dendrogram cuts

Modularity

Modularity at resolution γ:

$$Q_\gamma(C) = \frac{1}{2m} \sum_{i,j \in V} \left(A_{ij} - \gamma \frac{d_i d_j}{2m} \right) \delta_{C(i),C(j)}$$

The fit ($\gamma \to 0$) vs diversity ($\gamma \to +\infty$) trade-off



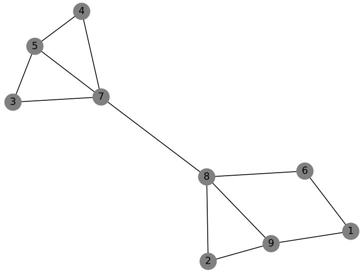
Resolution limit

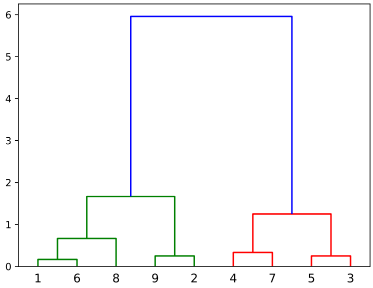
Proposition

The resolution limit, beyond which all clusters have size 1, is:

$$\gamma_1 = \max_{i \neq j} \sigma(i, j).$$

Example

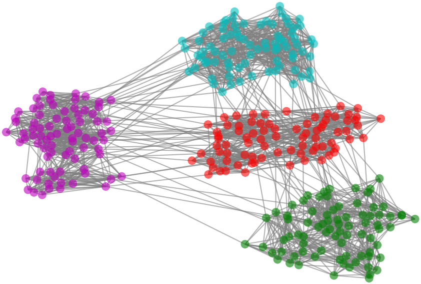
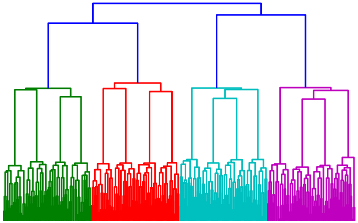




Outline

1. Node sampling
2. Agglomerative algorithm
3. Notion of dendrogram
4. Link with modularity
5. **Dendrogram cuts**

Dendrogram cuts

Summary

Hierarchical clustering is a key technique for the **multi-scale analysis** of graphs:

- **Dendrogram**, a compact representation of the hierarchical structure of a graph
- **Linkage**, based on the sampling ratio:

$$\sigma(i,j) = \frac{p(i,j)}{p(i)p(j)}$$

- The **nearest-neighbor chain**, an efficient algorithm for agglomerative clustering
- Extensions exist for **bipartite** and **directed** graphs