Overlapping Hierarchical Clustering (OHC)

Ian Jeantet, Zoltan Miklos, David Gross-Amblard

Univ Rennes, CNRS, IRISA

ian.jeantet@irisa.fr

IDA 2020, April 27 – 29

Motivations

This work is part of the **EPIQUE ANR Program**[1] that aims to build automatic tools to study the evolution of science from large-scale datasets. Particular constraints of our use case:

- The goal is to build a **hierarchy** over a set of key words.
- The clusters need to be **overlapping** as a key word can belong to several topics.
- Key words are embedded[2] and a topic should be a **dense** area of the embedding space.

[1] ANR-16-CE38-0002 - [2] Mikolov et al., 2013

Motivations

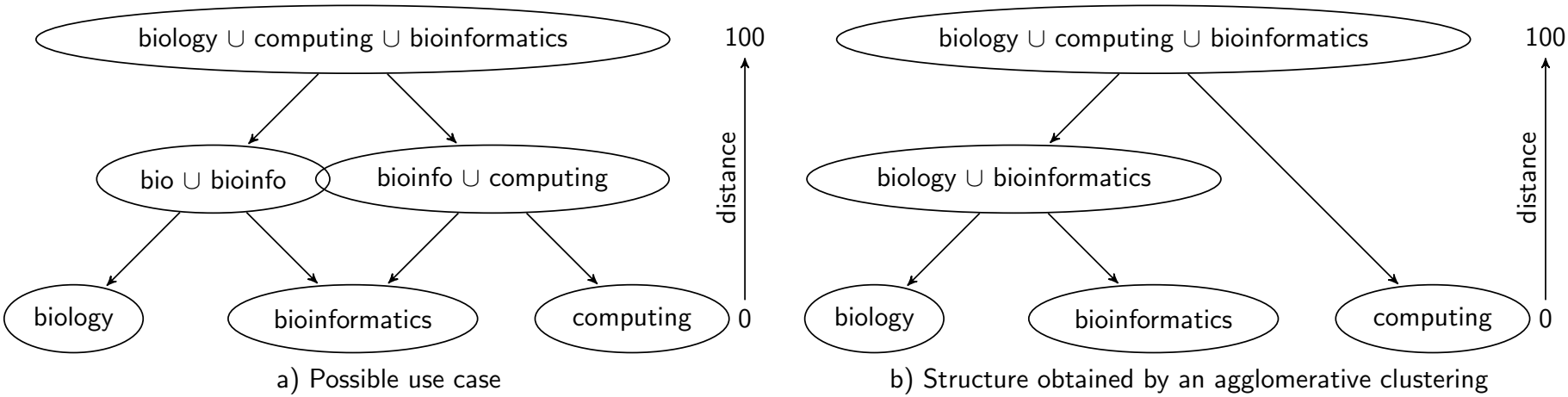
Reminder on classical agglomerative clustering:

- Produce easy-to-understand tree structures.
- Optimal and well studied algorithms (SLINK[1], HDBSCAN[2], SOHC[3], Pyramidal clustering[4]).
- Many variants based on different metrics and linkage criteria.

[1]Sibson et al., 1973 - [2]Campello et al., 2015 - [3]Chen et al., 2008 - [4]Diday, 1984

Motivations

Problems with existing algorithms on an example:



a) Possible use case

b) Structure obtained by an agglomerative clustering

Overlapping hierarchical clustering framework

Principle of the algorithm:

- Initialisation:
  - Start with singleton clusters and the 0-neighbourhood graph.
- Main loop:
  - Add links by increasing order in the graph,
  - Look for impacted clusters,
  - Use a density-based merging threshold λ to decide whether to grow the clusters or not.
- Ending:
  - Stop when there is only one cluster left.

Expressiveness

- Let's compare different structures on this example.
- We can clearly identify the 2 clusters circled in red.
- The point F may force the 2 clusters to merge too soon. We would prefer to have an overlap on this point instead.

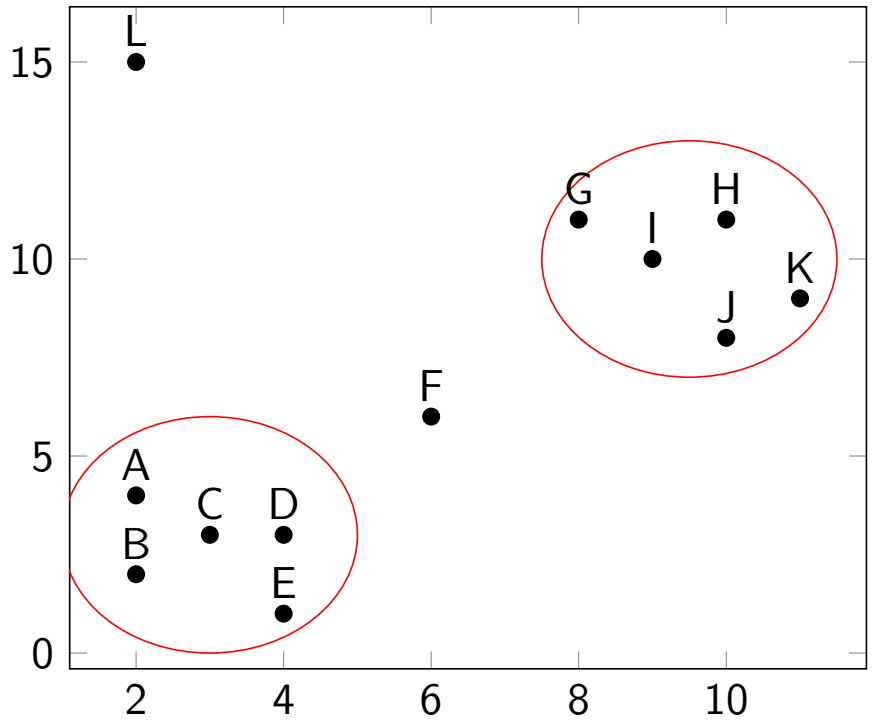


Figure: A practical example

Chaining effect problem

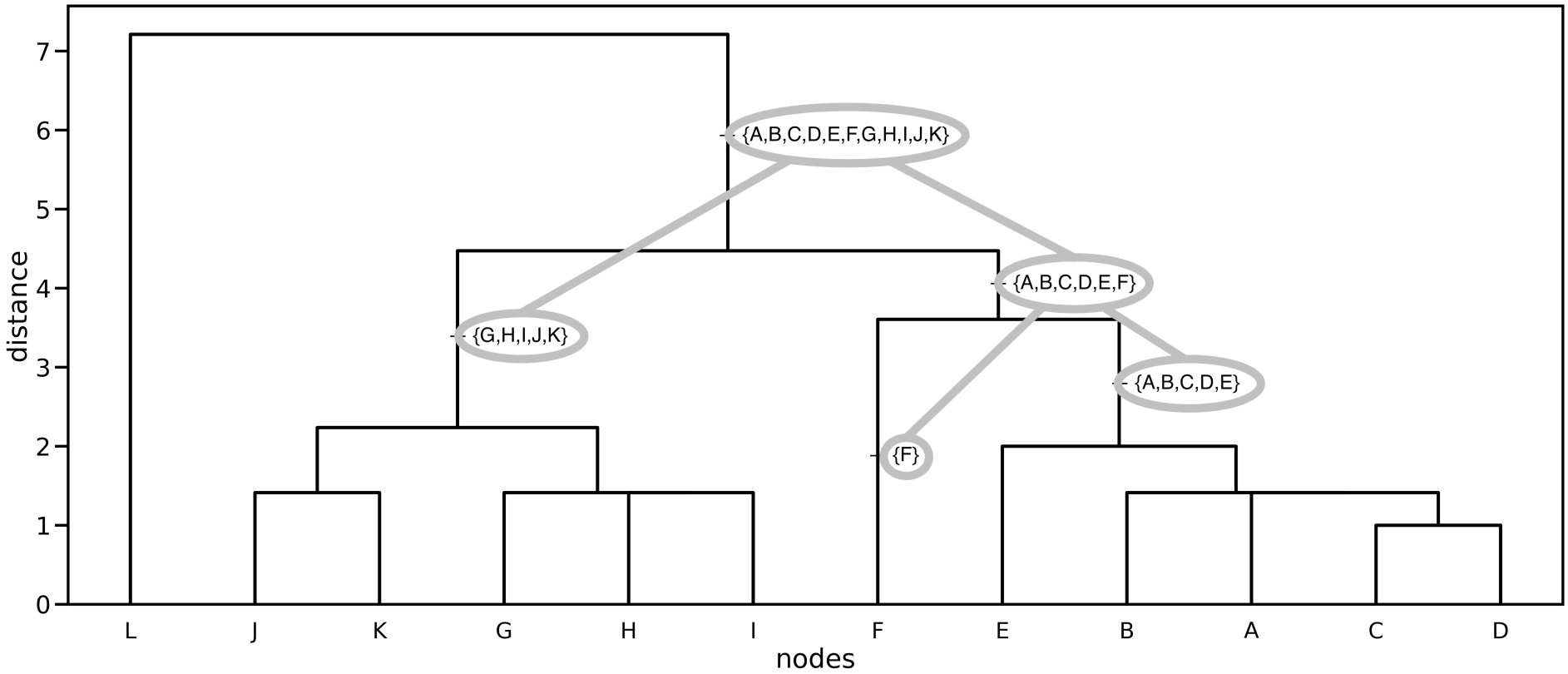


Figure: *SLINK* dendrogram obtained from the practical example.

Overlapping solution

- Level 4: 2 red clusters + 2 isolated points.
- Level 6: F in integrated to $\{A, B, C, D, E\}$.
- Level 7: F is shared with $\{G, H, I, J, K\}$.

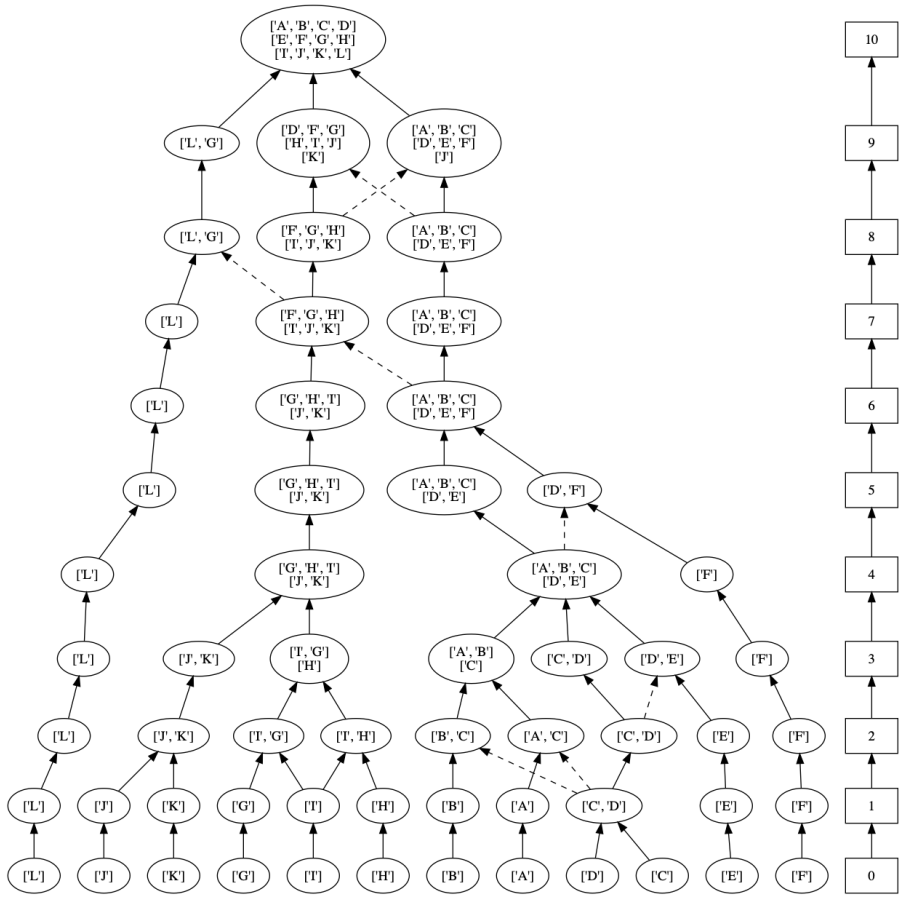


Figure: OHC quasi-dendrogram obtained with $\lambda = 0.2$.

How to compare quasi-dendrograms?

- Idea: match levels that have the same number of clusters and compare these clusters in the sense of the Jaccard similarity.
- Problem: quasi-dendrograms can have several levels with the same number of clusters.

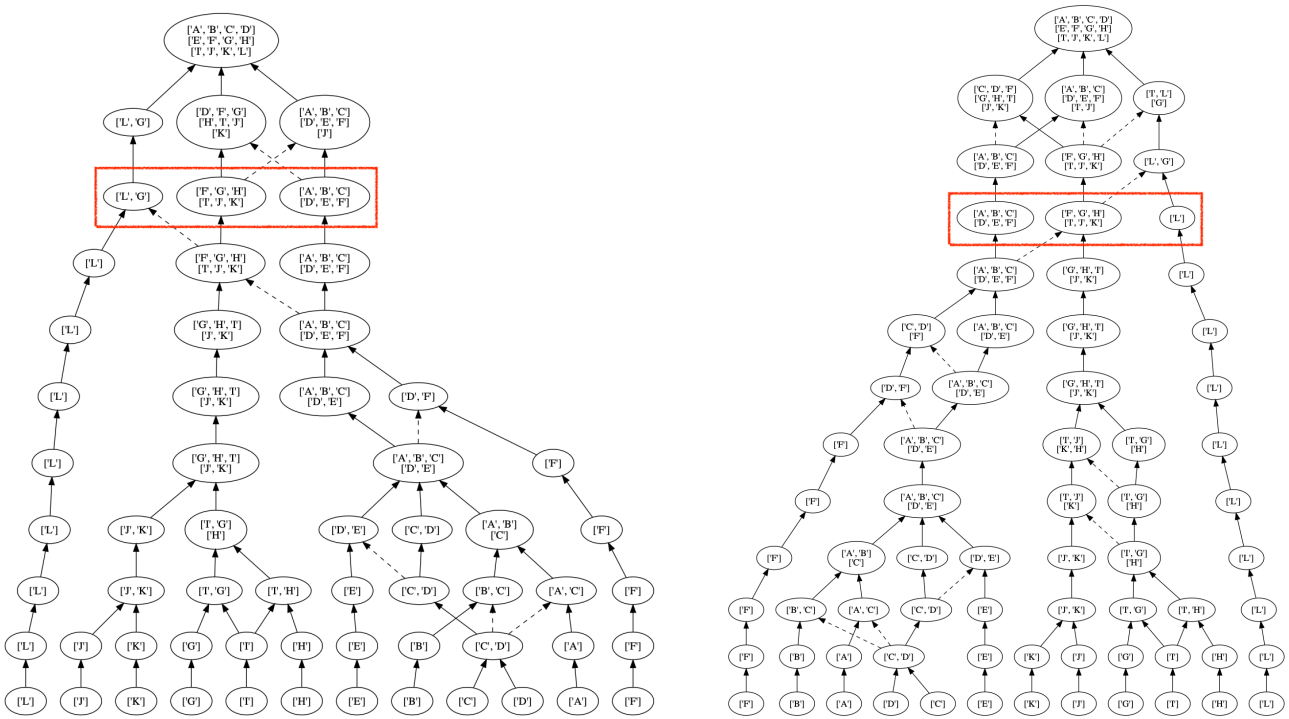


Table: Similarity between 2 quasi-dendrograms.

Similarity

Proposal: try to find the best match between these levels but by keeping the structural constraint.

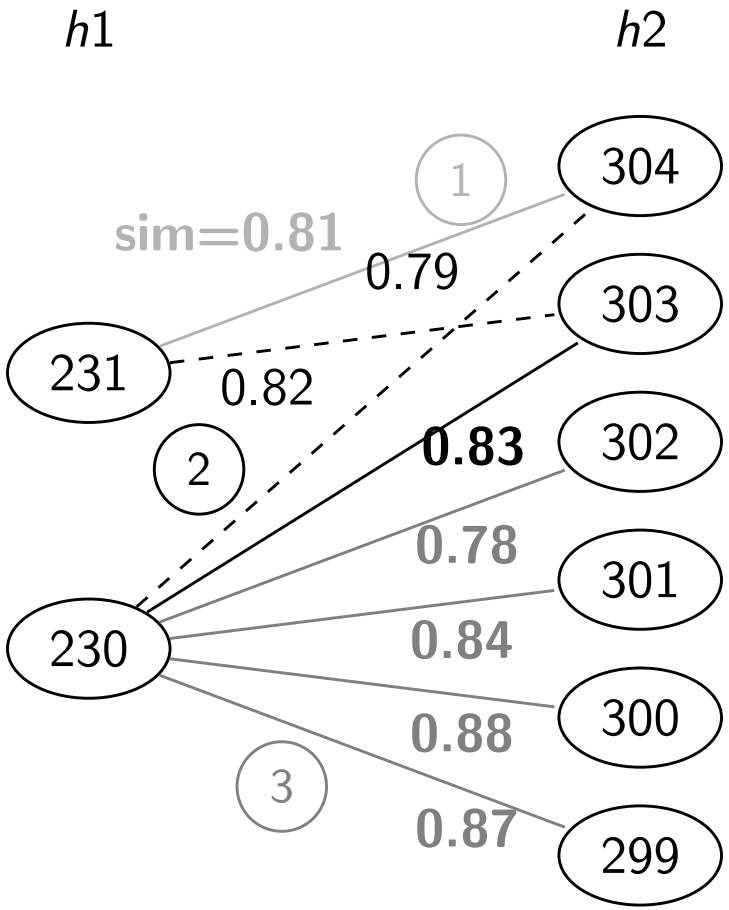


Figure: Similarity example on levels of the same size of the hierarchies $h1$ and $h2$.

Recap

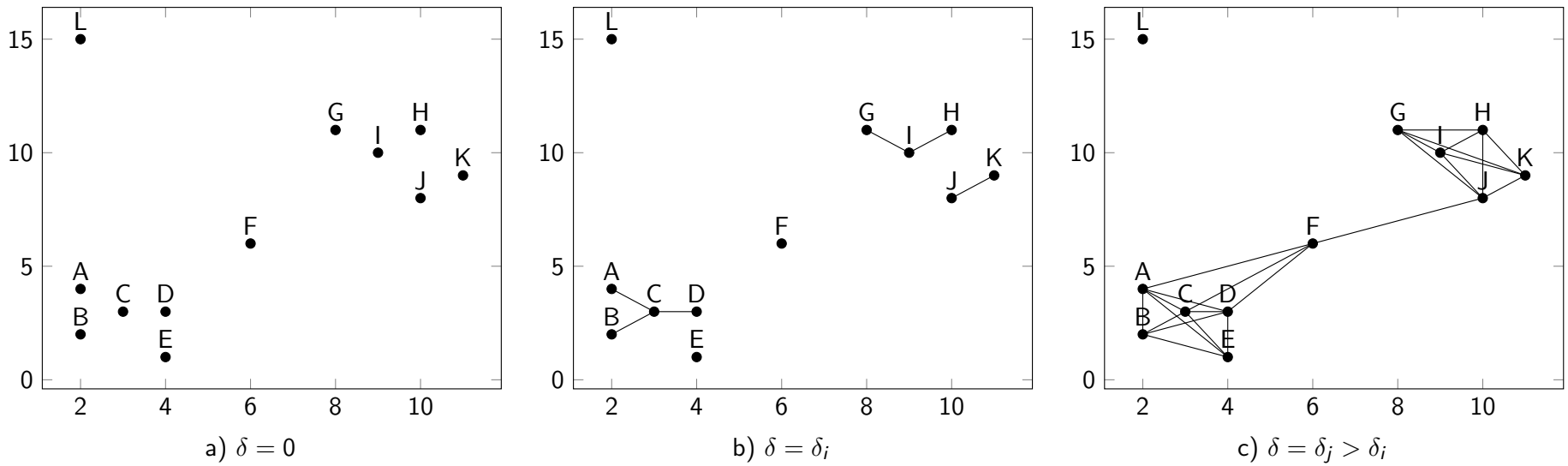
Our contributions:

- We propose an agglomerative clustering algorithm that allows and builds overlaps during the process,
- We proved and measured the conservativity of our algorithm with a particular tuning,
- We introduce a similarity measure to compare classical and quasi-dendrograms.

Appendix

δ-neighbourhood graph

A δ-neighbourhood graph is a graph containing all the edges between 2 of its vertices if their distance is smaller than δ.



a) $\delta = 0$ b) $\delta = \delta_i$ c) $\delta = \delta_j > \delta_i$

Table: A practical example.

Cluster growth

How are the clusters growing?

- Let's start with this ($\delta_i = 2.23$)-neighbourhood graph.
- Let's assume we have 2 clusters, $C_{i,1} = \{A, B, C, D, E\}$ and $C_{i,2} = \{F\}$.
- The density of $C_{i,1}$ is $9/10 = 0.9$.
- The density of $C_{i,2}$ is 1.

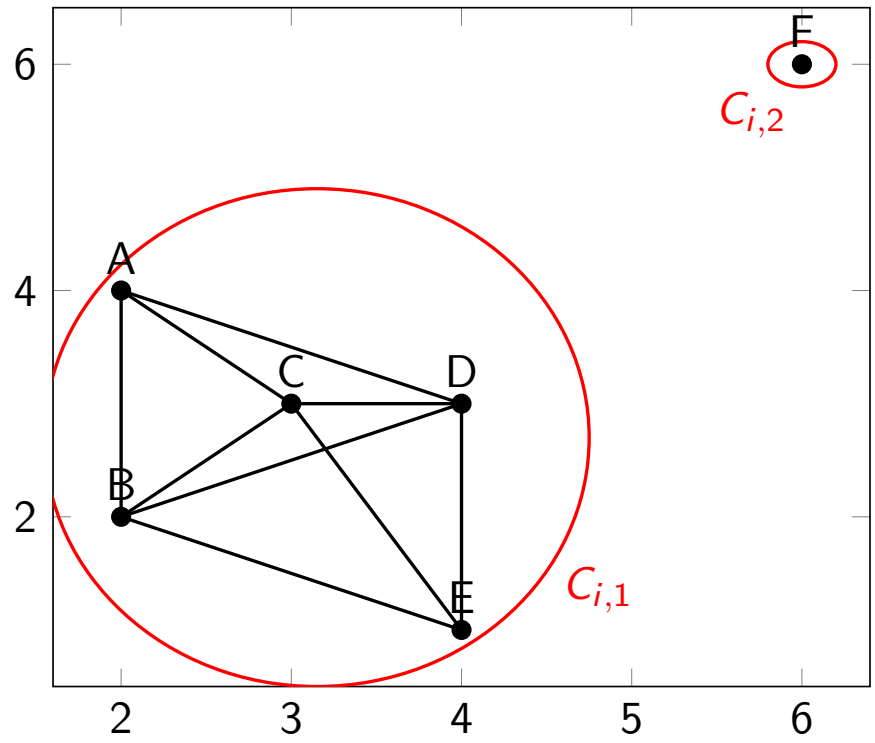


Figure: Starting with the δ_i-neighbourhood graph.

Cluster growth

- The density of $C_{i,1}$ is now 1.
- The density of a new cluster formed by adding F to $C_{i,1}$ is $11/15 \simeq 0.73$.
- Case 1: $\lambda < 0.27 = 1 - 0.73$ then
  - F stays out of the cluster and $C_{j,1} = C_{i,1}$
  - However $C_{j,2} = \{D\} \cup C_{i,2}$ as the density of $\{D\} \cup \{F\}$ is 1.
- Case 2: $\lambda \geq 0.27$ then the new cluster C_j is now $\{A, B, C, D, E, F\}$.

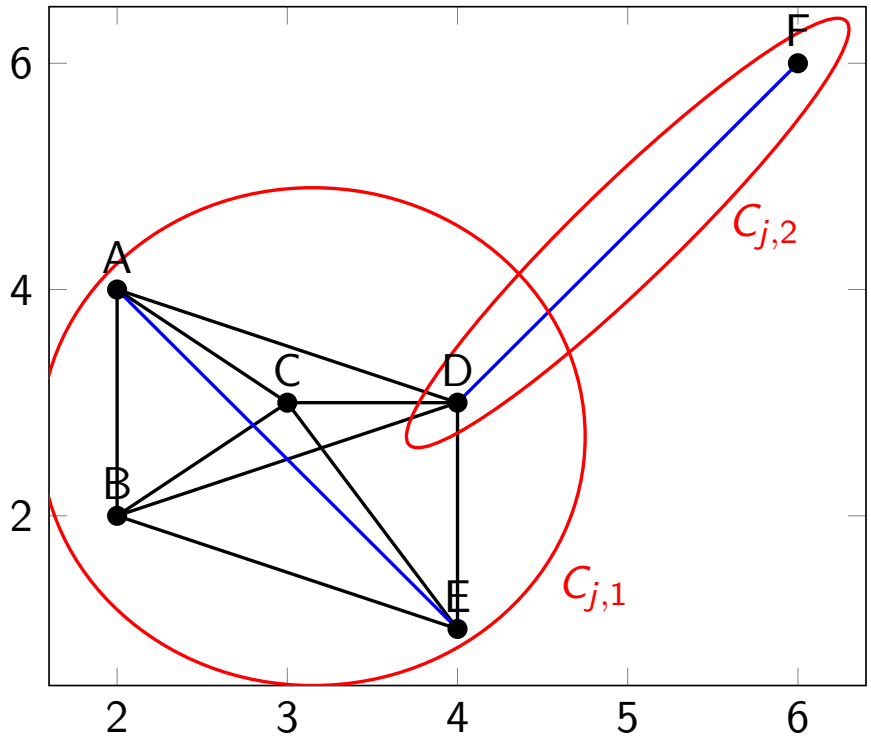


Figure: We increase δ to $\delta_j = 3.6$

Cluster growth

- The density of both $C_{j,1}$ and $C_{j,2}$ is 1.
- The density of a new cluster formed by adding F to $C_{j,1}$ is $12/15 \simeq 0.8$.
- Now with a merging criterion of only 0.2 we form a new cluster.
- Depending on λ, $\{A, B, C, D, E\}$ will persist more or less longer in the hierarchy.

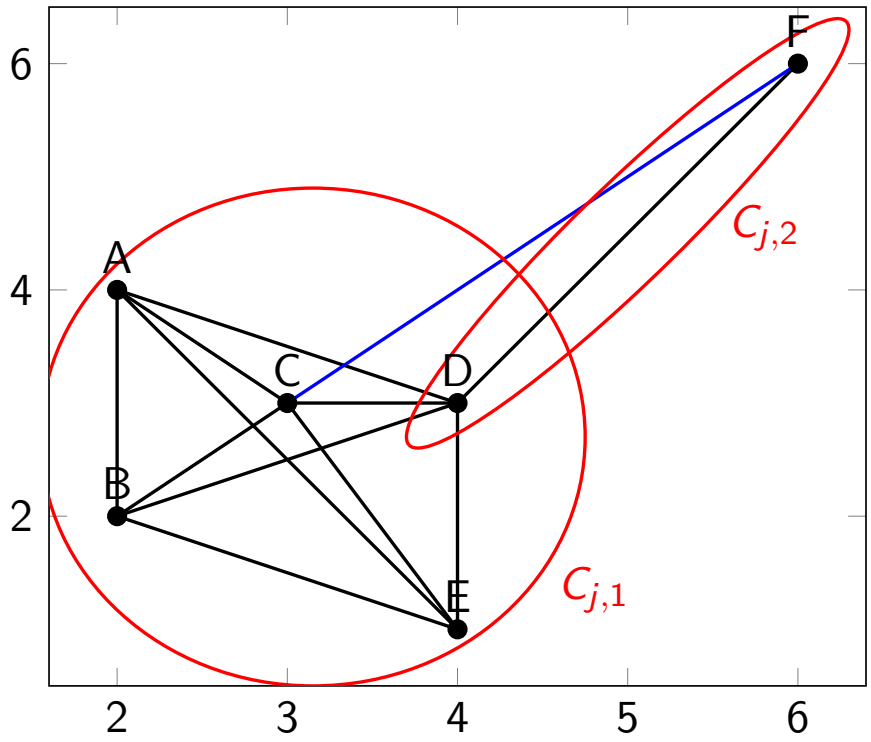
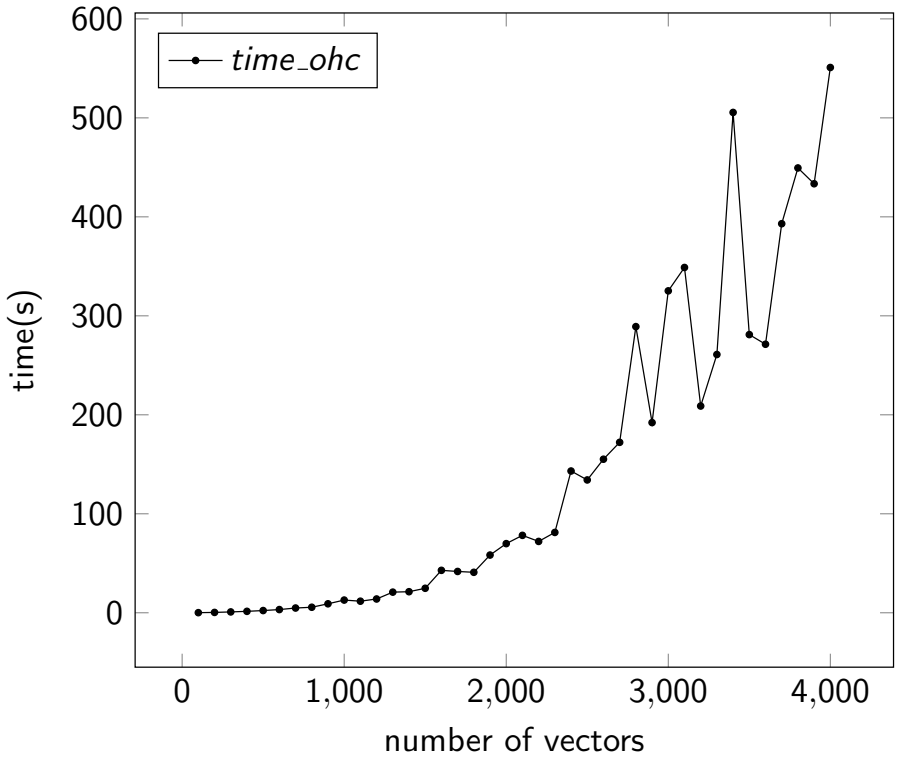
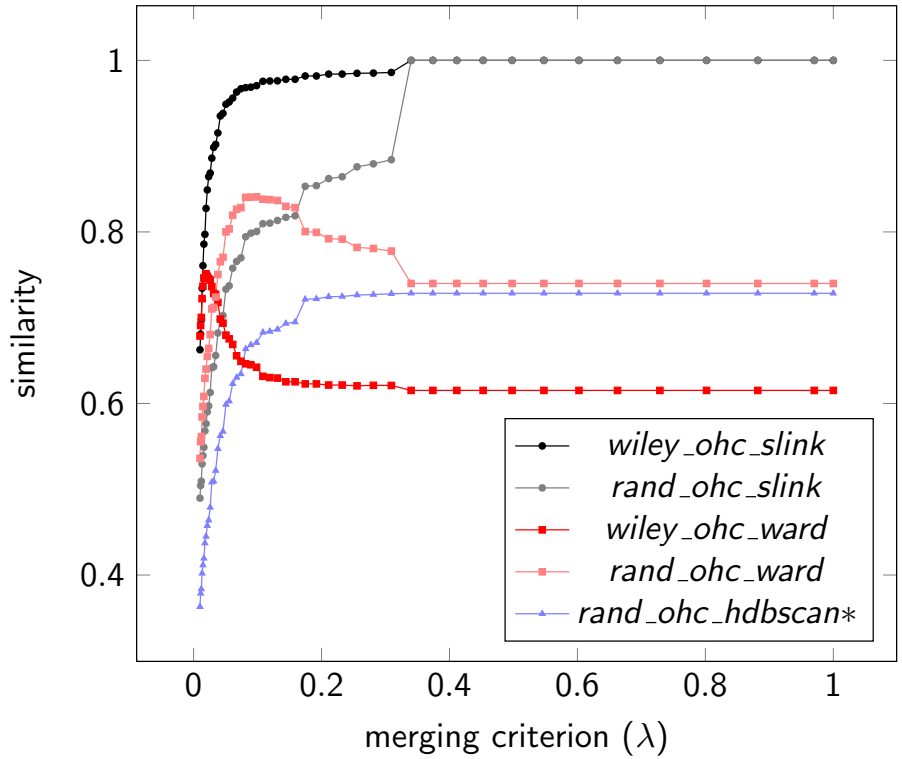


Figure: We increase δ to $\delta_k = 4.2$

Experiments



(a) Computation limitation (up to 4000 points on a regular computer) → Possibility to perform a preprocessing



(b) Similarity between hierarchical structures according to the merging criterion.

Figure: Study of the merging criterion.