




# The Challenge of Branch-Aware Data Manifold Exploration

D.M. Bot<sup>1</sup>  J. Peeters<sup>1</sup>  and J. Aerts<sup>1,2</sup> 

<sup>1</sup>Data Science Institute (DSI), Hasselt University, Belgium

<sup>2</sup>Leuven Statistisch Centrum (LStat), KU Leuven, Belgium

## Abstract

Branches within clusters can represent meaningful subgroups that should be explored. In general, automatically detecting branching structures within clusters requires analysing the distances between data points and a centrality metric, resulting in a complex two-dimensional hierarchy. This poster describes abstractions for this data and formulates requirements for a visualisation, building towards a comprehensive branch-aware cluster exploration interface.

## CCS Concepts

• *Computing methodologies* → *Cluster analysis; Dimensionality reduction and manifold learning;*

## 1. Introduction

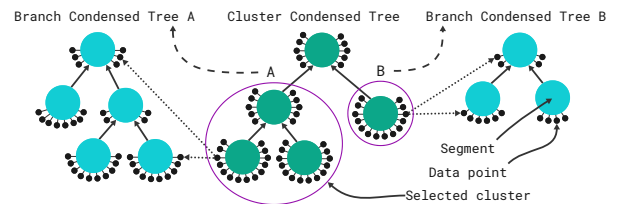
Detecting subgroups in unfamiliar data is an essential data exploration step. Commonly, clustering algorithms detect groups of similar data points based on distance. While branches in the data's manifold can represent meaningful subgroups (see, f.i. [RM79, PZH\*19]), clustering algorithms generally cannot detect them. Intuitively, this can be explained by the observation that within clusters, there is a path between all points that travels only through other data points that 'lie close together'. Instead, detecting branches within clusters requires using a centrality metric. The main idea is to filter out a cluster's central core, breaking the path between branches and allowing them to be detected as clusters. We refer to [Car14] for a detailed formal discussion.

Several approaches for combining distance and centrality information exist. For this poster, we focus on a method (under development) that decouples both dimensions by detecting clusters using data point distances, describing the connectivity within the clusters as networks, and performing a filtration over the centrality to detect branches within clusters.

In this poster, we describe a data abstraction for the resulting hierarchies; formulate the questions that should be answered by a visualisation summarising data this way; and discuss the benefits and shortcomings of our preliminary designs, building towards a unified exploration interface.

## 2. Data Abstraction

The primary data structure to describe is a *condensed tree* as used in the HDBSCAN\* clustering algorithm [MH17, CMZS15]. HDBSCAN\* clusters points by their approximate local density using a single linkage dendrogram as the basis of the algorithm. The condensed tree simplifies this dendrogram by pruning it with a mini-



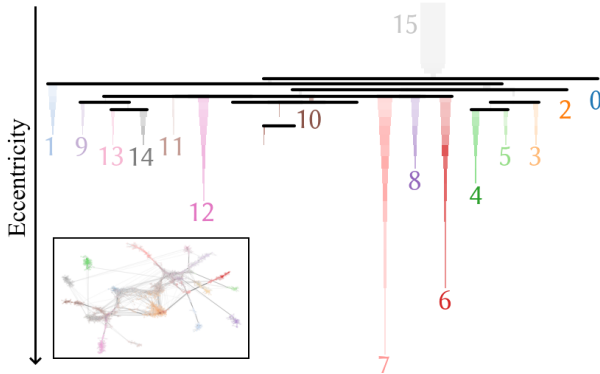
**Figure 1:** Schematic condensed trees. Coloured circles indicate segments, and black points represent data points. Points in the sub-tree of selected segments (A, B) also occur in the branch condensed tree; dotted arrows indicate possible data point matches.

imum cluster size  $n$ . Conceptually, the condensed tree can be seen as a directed tree structure with two types of nodes: segments (i.e. internal nodes) and points (i.e. leaf nodes). Edges occur either between segments, indicating at which distance they merge, or between a point and a segment, specifying at which distance the point joins the segment. The tree has two more interesting properties: (1) each segment is an ancestor of at least  $n$  points, and (2) the edges between segments form a binary tree.

HDBSCAN\* selects segments from the condensed tree based on their *stability* to be the final detected clusters. The entire sub-tree below the chosen segments is considered to belong to a single cluster. Our branch detection approach then constructs another condensed tree describing the branching hierarchy for each selected cluster. In these trees, the edges provide the eccentricity value (i.e.,  $1 / \text{centrality}$ ) at which points and segments merge (see Figure 1).

## 3. Task Requirements

A visualisation of the condensed trees should be able to answer several questions and considerations.



**Figure 2:** A branch condensed tree design. Selected segments are labelled and given a hue. Relative lightness encodes the average density along the branches. 2D projections (bottom-left) can be coloured similarly to provide additional context.

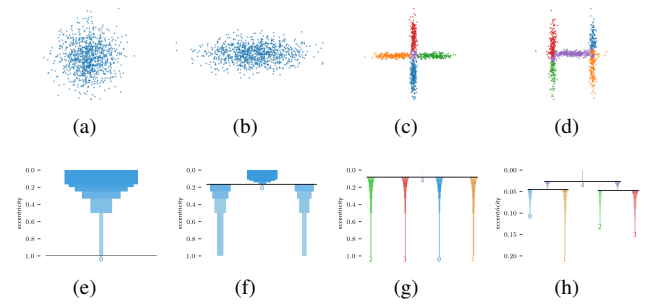
Firstly, a visualisation should show which segments were selected and communicate why they were chosen by showing their *stability*. A segment’s *stability* is defined as the sum of distance (or centrality) ranges for which points are part of the segment. This metric combines three values: (1) the distance (or centrality) range for which the segment exists, i.e. the segment’s persistence, (2) the number of children in the segment, and (3) how long each child is part of the segment.

Secondly, designs should communicate the shape of the detected clusters. The branch-condensed trees encode this information. They often contain three segments for clusters without branches: one root and two leaves. These leaves represent the outsides of the cluster growing inward, and their stability corresponds to the cluster elongation. Clusters with branches will have more leaves in their branch-condensed trees. The order in which the leaves merge represents the cluster’s shape. For example, an X-shaped cluster will have four leaves connecting close to the maximum centrality. In contrast, for an H-shaped cluster, the four leaves first join into two separate segments before those merge close to the maximum centrality. See Figure 3 for an illustration of these examples.

Finally, a visualisation should communicate the density profile over the cluster shapes. The existence and position of density maxima are important for interpreting the detected clusters and branches, as they express the variability and likelihood of similar observations.

#### 4. Preliminary Design and Example

Figure 2 shows a preliminary branch condensed tree design summarising *C. elegans*’ cell development data through gene expressions [PZH\*19]. The design adapts [MHA17]’s condensed tree plot. Segments are visualised as a hierarchically laid-out binary tree. The (logarithmic) vertical axis encodes the centrality, so the heights represent persistence. Segment widths encode the number of data points in the segment at each distance value. In contrast to [MHA17]’s design, only the direct children of a segment are counted, rather than all children in a segment’s sub-tree. The area, therefore, no longer encodes *stability*, and the tree resembles a *Reeb*



**Figure 3:** Branch condensed tree captures the shape of clusters. 2D point clouds coloured by the detected branch subgroups (top row) with their corresponding branch-condensed trees (bottom row).

graph. A benefit of this approach is that one does not have to look at the change in width to see where the points lie within the shape, effectively prioritising shape interpretation over stability comparison. Interestingly, the figure reveals that the density appears highest within the branches. This could support the interpretation that the branches represent distinct developmental end-states, which could be expected to occur more often than in-between states.

For data with a single cluster, simply showing the branch condensed tree effectively summarises the data’s shape. Unlike dimensionality reduction plots, the tree does not rely on 2D coordinates that tend to (over)emphasise longer distances. However, the tree requires more interpretation to understand the shape it represents.

How to scale this design to multiple clusters still needs to be determined. Simply showing numerous branch-condensed trees would not communicate why clusters were selected. In addition, the colour coding is limited in number by distinct hues. Furthermore, the current design must still be adapted to show how many local density maxima occur at a particular centrality along a branch.

One potential alternative design that remains to be explored is based on [GSWD18]’s Bubble Tree Map. Their visualisation uses the area of circles to encode a quantity of interest and the borders to encode the uncertainty of that quantity. A direct application to our problem is challenging, as two intersecting hierarchies have to be encoded. However, visualising the selected clusters, branches, and density maxima should be feasible while communicating their stability. Another alternative could be a Mapper-like [SMC07] summary graph, encoding the selected branches as nodes. However, encoding the hierarchies to communicate why segments were selected would be non-trivial.

#### 5. Conclusions

This poster presented the challenge of visualising nested hierarchies for a branch-based data exploration method. One preliminary design was shown, and its benefits and limitations were briefly discussed.

#### 6. Acknowledgements

This work was supported in part by Hasselt University BOF grants ADMIRE [BOF21GP17] and [BOF21DOC19].

## References

- [Car14] CARLSSON G.: Topological pattern recognition for point cloud data. *Acta Numer.* 23, 2014 (may 2014), 289–368. doi:10.1017/S0962492914000051. 1
- [CMZS15] CAMPELLO R. J. G. B., MOULAVI D., ZIMEK A., SANDER J.: Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. Data* 10, 1 (jul 2015), 1–51. doi:10.1145/2733381. 1
- [GSWD18] GORTLER J., SCHULZ C., WEISKOPF D., DEUSSEN O.: Bubble Treemaps for Uncertainty Visualization. *IEEE Trans. Vis. Comput. Graph.* 24, 1 (jan 2018), 719–728. doi:10.1109/TVCG.2017.2743959. 2
- [MH17] MCINNES L., HEALY J.: Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (2017), pp. 33–42. doi:10.1109/ICDMW.2017.12. 1
- [MHA17] MCINNES L., HEALY J., ASTELS S.: hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2, 11 (2017), 205. doi:10.21105/JOSS.00205. 2
- [PZH\*19] PACKER J. S., ZHU Q., HUYNH C., SIVARAMAKRISHNAN P., PRESTON E., DUECK H., STEFANIK D., TAN K., TRAPNELL C., KIM J., WATERSTON R. H., MURRAY J. I.: A lineage-resolved molecular atlas of *C. Elegans* embryogenesis at single-cell resolution. *Science* (80-. ). 365, 6459 (2019). doi:10.1126/science.aax1971. 1, 2
- [RM79] REAVEN G. M., MILLER R. G.: An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia* 16, 1 (jan 1979), 17–24. doi:10.1007/BF00423145. 1
- [SMC07] SINGH G., MÉMOLI F., CARLSSON G.: Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *PGB@ Eurographics* 2 (sep 2007). 2