The Challenge of Branch-Aware Data Manifold Exploration

Daniël M. Bot[1], Jannes Peeters[1], Jan Aerts[1,2]

[1]Data Science Institute (DSI), Hasselt University, Belgium.

[2]Leuven Statistisch Centrum (LStat), KU Leuven, Belgium.

Introduction

While branches in the data's manifold can represent meaningful subgroups (see, f.i., [RM79, PZH*19]), clustering algorithms generally cannot detect them. Instead, detecting branches within clusters requires using a centrality metric [Car14].

Our branch detection method 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.

Data Abstraction

The primary data structure to describe is a *condensed tree* as used in the HDBSCAN* clustering algorithm [MH17, CMZS15]. Conceptually, the condensed tree can be seen as a directed tree structure with two types of nodes: segments and points.

HDBSCAN* selects segments from the condensed tree based on their *stability* to be the final detected clusters. Our branch detection approach then constructs another condensed tree describing the branching hierarchy for each selected cluster. Figure 1 schematically shows this construction.

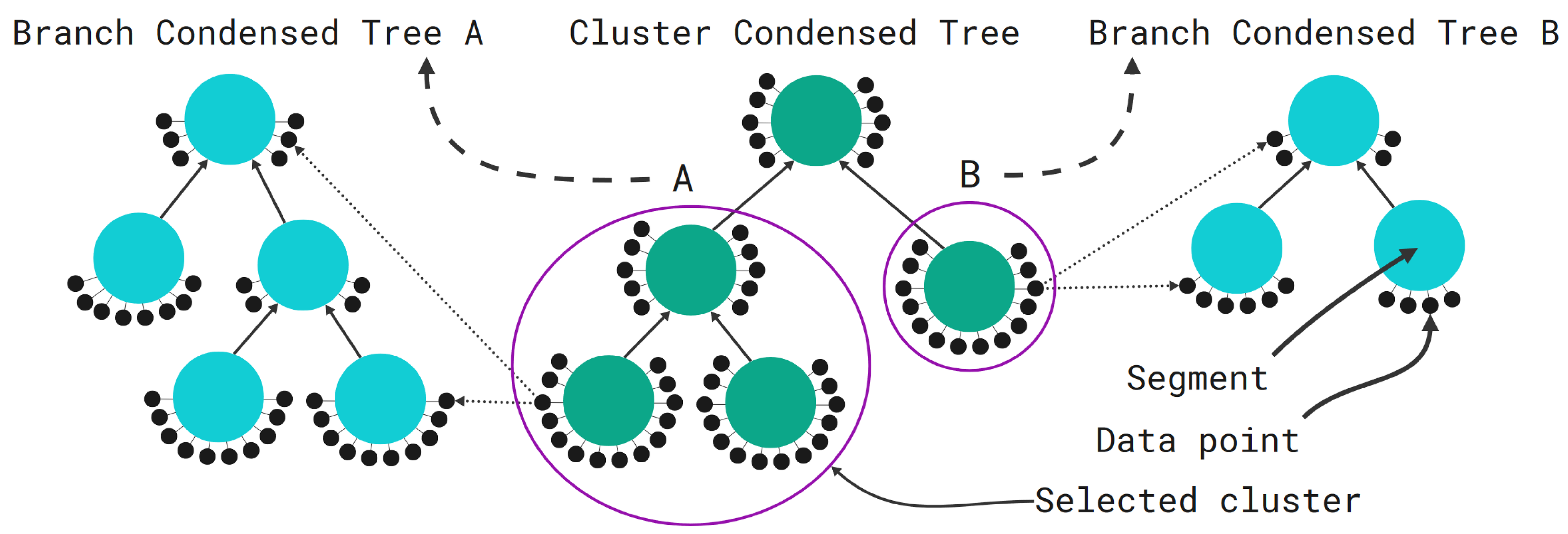

Fig 1. Schematic condensed trees. Coloured circles indicate segments, and black points represent data points. Points in the sub-tree of selected segments (A, B) also occur in the branch condensed tree; dotted arrows indicate possible data point matches.

Task Requirements

1. Show which segments were selected and communicate their *stability*.
2. Communicate the shape of the detected clusters. Figure 2 demonstrates how branch-condensed trees show clusters' shapes.
3. 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.

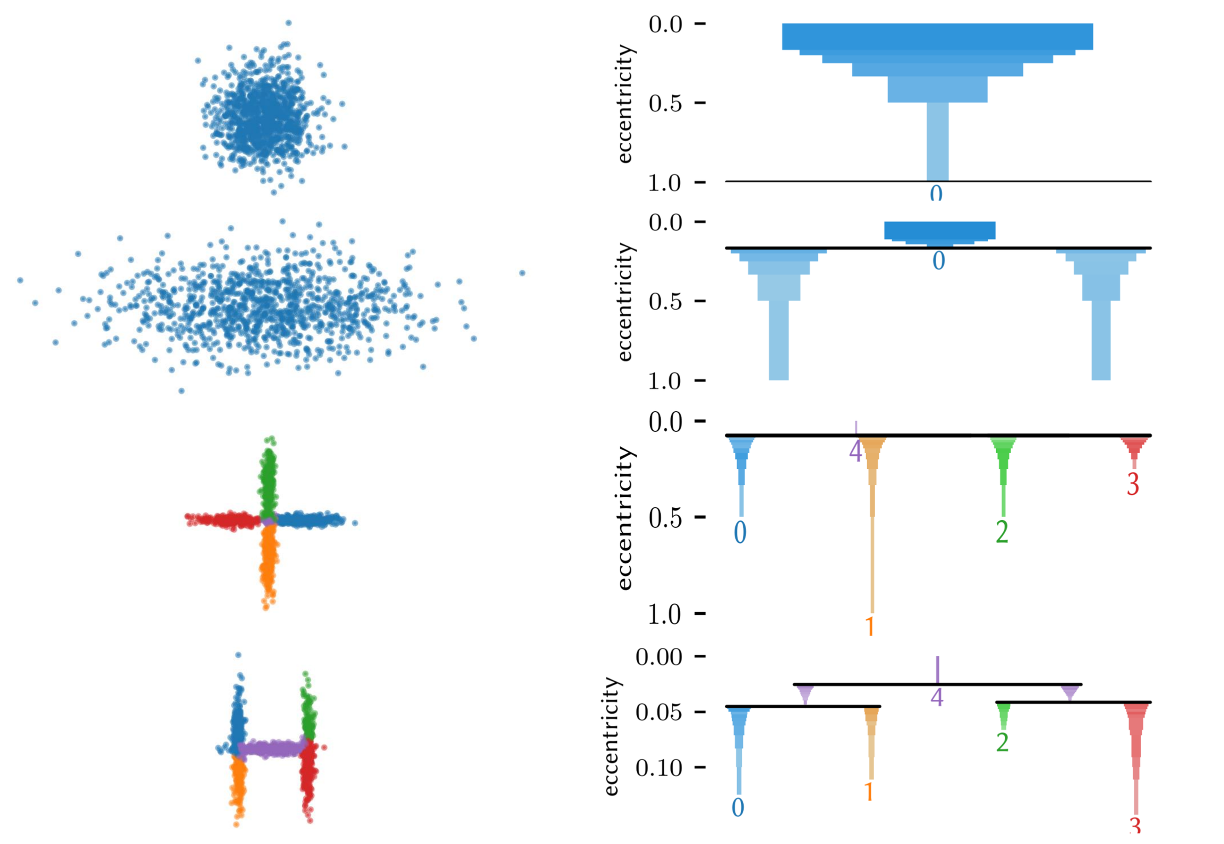

Fig 2. Branch condensed tree captures the shape of clusters. 2D point clouds coloured by the detected branch subgroups (left) with their corresponding branch-condensed trees (right).

Example: C. Elegans Single Cell

Figure 3A shows a preliminary branch condensed tree design summarising *C. elegans*' cell development data [PZH*19]. The design adapts [MHA17]'s condensed tree plot, showing segments as a hierarchically laid-out binary tree. 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, effectively prioritising *shape* interpretation over *stability* comparison.

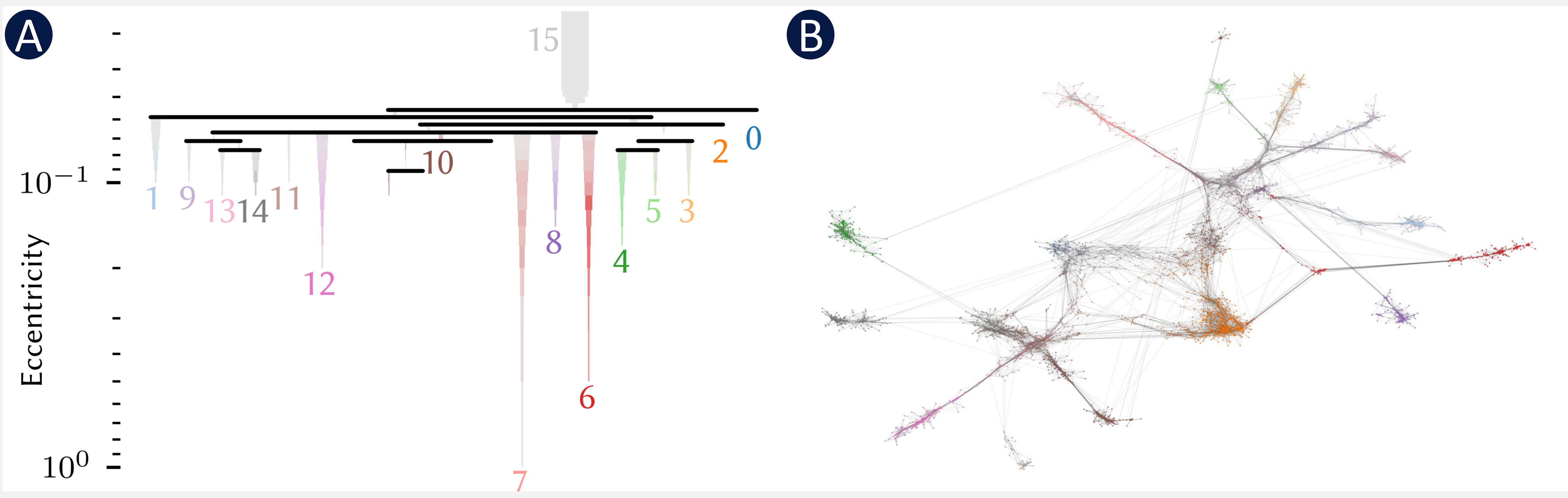

Fig 3. A preliminary branch condensed tree design (a) with corresponding densMAP [NBC21] projection (b). Selected branches are labelled and given a hue. Relative lightness encodes the average density along the branches. Points in the 2D projection are coloured by their membership to each branch to provide additional context.

Potential Alternatives

- An adapted Bubble Tree Map [GSWD18].
- A Mapper-like [SMC07] summary graph.
- A Rivet-like [LW15] exploration view.

Future Work:

- How to scale to multiple clusters? Simply showing numerous branch-condensed trees would not communicate why clusters were selected.
- How to scale to more sub-groups? The colour coding is limited in number by distinct hues.
- How to communicate how many local density maxima occur at a particular centrality along a branch?

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.
[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.
[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.
[LW15] LESNICK M., WRIGHT M.: Interactive Visualization of 2-D Persistence Modules, *arxiv,* dec 2015.
[MH17] MCINNES L., HEALY J.: Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (2017), pp. 33–42.
[MHA17] MCINNES L., HEALY J., ASTELS S.: hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software 2*, 11 (2017), 205.
[NBC21] NARAYAN A., BERGER B., CHO H.: Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nat. Biotechnol. 39*, 6 (jun 2021), 765–774.
[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).
[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.
[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).