Canad. Math. Bull. Vol. **63** (3), 2020 pp. 562–576
http://dx.doi.org/10.4153/S000843951900064X
© Canadian Mathematical Society 2019

Stable Components and Layers

J. F. Jardine

Abstract. Component graphs $\Gamma_0(F)$ are defined for arrays of sets F, and, in particular, for arrays of path components for Vietoris–Rips complexes and Lesnick complexes. The path components of $\Gamma_0(F)$ are the *stable components* of the array F. The stable components for the system of Lesnick complexes $\{L_{s,k}(X)\}$ for a finite data set X decompose into *layers*, which are themselves path components of a graph. Combinatorial scoring functions are defined for layers and stable components.

Introduction

Astronomers say that a cluster is a "group of stars or galaxies forming a relatively close association".



Clusters are distinguished by relative density; they are concentrated collections of objects, surrounded by voids.

In the late 1700s, the brother and sister team William and Caroline Herschel found and classified "stellar over-densities" by counting stars in grids of regions of space. The method used today follows the same principle, although it is done with sophisticated imaging equipment coupled with computer analysis that filters out light artifacts. Relative to big picture items such as the cosmic microwave background, clusters are anomalies — they are small, dense collections of stellar objects.

The same sort of big picture/little picture dichotomy is present for other large data sets. In financial data, a dense, relatively small collection of rapid low-value transactions could point to an instance of money laundering, while a large scale analysis can detect sectoral or global market fluctuations.

Colloquially, clusters are collections of data points in relative close proximity within some space that is defined by a finite list of parameters. We will assume, more precisely, that a data set (or a data cloud) X is a finite collection of points inside real

Received by the editors May 16, 2019; revised October 20, 2019.
Published online on Cambridge Core October 23, 2019.
This research was partially supported by the Natural Sciences and Engineering Research Council of Canada.
AMS subject classification: **55U10**, 68R10, 62H30.
Keywords: cluster, graph, stable component, layer.

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press
CrossMark

vector space $\mathbb{R}^n$. The motivating idea behind clustering is to find regions in the ambient space that contain dense populations of elements of the data set X. This is usually done in data analysis by applying partitioning methods to the data set X, or by studying hierarchies of such partitions.

The definitional structure and methods of this paper represent a departure from traditional clustering, although partitions and hierarchies (really, trees) of partitions are both used. The central objects of study are partition elements (or path components, or clusters) that persist through changes of distance parameters, or density parameters, or both — these objects are called *stable components*. The idea is to assign a more precise meaning to the colloquial version of clusters, such as clusters of stars, which are isolated groupings of objects in a data set.

To this end, we recall some basic constructions of topological data analysis in the first section of this paper. Specifically, the data set X determines an ascending sequence of simplicial complexes $V_s(X)$, the Vietoris–Rips complexes, in which the simplices consist of sets of data points having mutual distance bounded by a nonnegative real number s. These complexes, in turn, are filtered by Lesnick subcomplexes $L_{s,k}(X)$, where the filtration is determined by a density parameter k.

Each of these complexes has a functorial set of path components $\pi_0 V_s(X)$ (resp. $\pi_0 L_{s,k}(X)$), which partition subsets of the data set X. The resulting diagrams of partitions have associated hierarchies $\Gamma(V_*(X))$ and $\Gamma(L_{*,*}(X))$, containing *component graphs* $\Gamma_0(\pi_0 V_*(X))$ and $\Gamma_0(L_{*,*}(X))$, respectively. The component graphs consist of path components that do not change through some variation in the defining parameters s and/or k. The component graphs themselves have path components, and these are the *stable components* for the data set X, in various incarnations.

Stable components further decompose into *layers,* for filtrations derived from the Lesnick filtration $L_{s,k}(X)$, in which the underlying sets of vertices may not be constant. The layers of the filtration $L_{*,*}(X)$ are defined by computing path components of the *layer subgraph* $\Gamma_0'(L_{s,k}(X))$ of the component graph. The edges of the layer subgraph are defined by path components that have the same *size* (or cardinality) through changes of defining parameters.

Layers are defined and discussed in the second section of this paper. Every stable component is a disjoint union of its constituent layers, while the layers form graphs that are relatively easy to visualize geometrically as unions of squares.

Naive calculations of stable components and layers produce large collections of small objects. In particular, the individual elements of X are stable components for the Vietoris–Rips filtration. It is typical to interpret small stable components or layers as noise, and remove them from the output of a particular algorithm. This can be done with a "scoring" technique, for which noise objects can be interpreted as stable components having low scores. That said, one could be most interested in stable components having relatively low scores, such as in algorithms that detect money laundering or smaller scale star clusters.

Scoring appears as an analytic device in statistical approaches to clustering (see [5], for example). An alternative combinatorial method of scoring is presented in the third section of this paper. Basically, the score $\sigma(P)$ of a stable component or the score $\sigma(L)$ of a layer L is the sum of the cardinalities of the path components that appear in its list of vertices. This number is most effectively calculated by making yet another graph

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

out of the vertices of the Lesnick complexes $L_{s,k}(X)$ and computing the cardinality of the set of vertices of a pullback of a stable component or layer within this new graph. This method of scoring is additive, so that the score of a stable component is the sum of the scores of its constituent layers.

A simple theoretical problem is presented in the final section of this paper. We start with a data set $X \subset \mathbb{R}^n$ and add a point y that is very close to one of the elements of X to form a new data set $Y = X \sqcup \{y\}$, and then we want to compare the stable components for X and Y.

Suitably interpreted, the inclusion $i : X \subset Y$ induces maps of Vietoris–Rips complexes $V_s(X) \to V_s(Y)$ that are natural with respect to the distance parameter s, with a corresponding map of hierarchies

$$\Gamma(X) := \Gamma(V_*(X)) \xrightarrow{i_*} \Gamma(V_*(Y)) =: \Gamma(Y).$$

The language of layers is used to compare stable components with respect to this map. In this context, a layer of X can be viewed as an edge

$$\big(s, [x]_s\big) \longrightarrow \big(t, [x]_t\big)$$

of $\Gamma(X)$, such that $[x]_s = [x]_t$ as subsets of X and a maximality condition on the length $t - s$ is satisfied. Here, $[x]_s$ is the path component of x in $V_s(X)$.

The map $\Gamma(X) \to \Gamma(Y)$ breaks up layers of X having both $y \notin [x]_s$ and $y \in [x]_t$ in $V_s(Y)$ and $V_t(Y)$, respectively. Otherwise, layers of X are mapped to "partial" layers of Y. These partial layers expand to layers of Y, having a size that can be approximated. The situation is summarized in Proposition 4.7 below.

More precise answers to the question of how layers and their scores vary through the hierarchy comparison $i_* : \Gamma(X) \to \Gamma(Y)$ would be available in specific examples.

The hierarchy $\Gamma(Y)$ is a refinement of $\Gamma(X)$, and the map i_* is a deformation retract in a very strong sense by the homotopy interleaving methods of [1]. Statements of this form give coarse information about layers, while the precise layer structure of a data set depends strongly on the distances between its points. Adding a single point to a data set X can introduce many new distances between points of the resulting data set Y.

1 Clusters, Graphs, and Stable Components

Clustering is a long-standing enterprise of data analysis, which consists of many definitions and methods. Basically, the standard techniques amount to either partitioning a data set or constructing hierarchies of partitions.

For example, K-means clustering starts with a set of points in a data set. This set of points partitions the data set into regions of nearest neighbours, or Voronoi cells. The algorithm proceeds by finding centres of the cells and then finding nearest neighbour sets for this set of centres to produce a new set of regions. The procedure stops when the set of centres stabilizes. The Voronoi cells determined by the resulting set of stable points partition the data set, and the partitions are expected to contain dense regions of data points near their centres.

Hierarchical clustering algorithms, such as single linkage clustering, require fewer assumptions. The inductive step in single linkage clustering assumes the existence of

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

a partition

$$X = P_1 \sqcup \cdots \sqcup P_k \subset \mathbb{R}^n$$

of the data set X. Find subsets P_i and P_j that are closest together in the ambient space $\mathbb{R}^n$, and then form a coarser partition by taking the union of $Q = P_i \cup P_j$ while keeping the other partition subsets fixed. There is a corresponding function that relates the first partition to the new one. The algorithm typically starts with the discrete partition $X = \bigsqcup_{x \in X} \{x\}$, and the last possible step in the resulting hierarchy of partitions would be the singleton partition consisting of X alone. The algorithm is typically stopped when it reaches a partition of X with sufficiently many points in some partition members, where the phrase "sufficiently many" is open to interpretation. The diagram of partitions and relations between them forms a dendogram, which is a type of tree.

The methods of topological data analysis produce multiple variations of these general themes.

Suppose that $X \subset \mathbb{R}^n$ is a finite data set and choose a non-negative real number s.

The *Vietoris–Rips* complex $V_s(X)$ is a simplicial complex with simplices consisting of sets $\{x_0, \ldots, x_k\}$ of points of X such that $d(x_i, x_j) \leq s$ for all i, j. The vertices of $V_s(X)$ are the elements of X, and the set of path components $\pi_0 V_s(X)$ of $V_s(X)$ defines a partition of the data set X.

Explicitly, elements $x, y \in X$ are in the same path component of $V_s(X)$ if there is a series of "short hops" (length $\leq s$) from x to y through elements of X. We define an equivalence relation on X in this way, and the equivalence classes are the path components of $V_s(X)$. Calculation of the set of path components of $V_s(X)$ can be done with an algorithm.

If $s \leq t$, then there is an induced inclusion of simplicial complexes

$$V_s(X) \subset V_t(X),$$

and a corresponding function $\pi_0 V_s(X) \to \pi_0 V_t(X)$ that relates partitionings given by the respective sets of path components. Because X is finite, there are only finitely many numbers

$$0 = s_0 < s_1 < \cdots < s_p,$$

which I call *phase change* numbers, that can occur as distances between elements of X. In the corresponding string of inclusions

$$X = V_{s_0}(X) \subset V_{s_1}(X) \subset \cdots \subset V_{s_p}(X),$$

the complex $V_{s_0}(X)$ is the discrete set X, while $V_{s_p}(X)$ is a big simplex, which I write as $\Delta^X =: \Delta^N$, where $N + 1$ is the number of elements of X. The space $V_{s_p}(X) = \Delta^X$ is contractible.

The contractibility of $V_{s_p}(X)$ implies that its set of path components $\pi_0 V_{s_p}(X)$ is a singleton set, and I express this by writing $\pi_0 V_{s_p}(X) = *$.

Remark 1.1 One way to produce the complexes $V_{s_i}(X)$, altogether, is to first find all distances s_i between data points, and to determine the maximum distance between elements for all subsets $\sigma = \{x_0, x_1, \ldots, x_p\}$ of X. This maximum distance for a subset σ is one of the s_i, and so σ is a simplex of $V_{s_i}(X)$. This construction is simple enough, but note the exponential complexity [7].

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

The corresponding picture

$$X = \pi_0 V_{s_0}(X) \to \pi_0 V_{s_1}(X) \to \cdots \to \pi_0 V_{s_p}(X) = * \tag{1.1}$$

of surjective functions between partitions defines a tree by a method specified below (Remark 1.4), and as such defines a *hierarchical clustering*.

The Vietoris–Rips complex $V_s(X)$ can be filtered by density. Suppose that k is a non-negative number. Then the complex $V_s(X)$ has a "full" subcomplex $L_{s,k}(X)$, which I call a *Lesnick complex*, whose simplices consist of vertices having at least k neighbours relative to the parameter s. See also [5].

The Lesnick complexes $L_{s,k}(X)$, $k \geq 0$ filter the Vietoris–Rips complex $V_s(X)$, and changing either the distance parameter s or the density parameter k defines an array of inclusions

$$\begin{array}{ccc} L_{s,k}(X) & \longrightarrow & L_{t,k}(X) \\ \uparrow & & \uparrow \\ L_{s,k+1}(X) & \to & L_{t,k+1}(X) \end{array}$$

of simplicial complexes.

Lesnick says that the array $\{L_{s,k}(X)\}$ is the degree Rips filtration of the Vietoris–Rips system $\{V_s(X)\}$, [4].

Example 1.2 The partitioning algorithm DBSCAN* ("Density based spatial clustering of applicatons with noise") amounts to a calculation of $\pi_0 L_{s,k}(X)$, for fixed tunable distance parameter s and density parameter k.

Example 1.3 The "hierarchical" version HDBSCAN* [2,5] of DBSCAN* is the production (and interpretation) of a tree that is associated with the string of functions

$$\pi_0 L_{s_0,k}(X) \longrightarrow \pi_0 L_{s_1,k}(X) \longrightarrow \cdots \longrightarrow \pi_0 L_{s_p,k}(X) = *. \tag{1.2}$$

Here, $L_{s_i,k}(X)$ could be empty for small s_i, and we are assuming that k is bounded above by the cardinality of X. There is one tunable parameter in this case, namely the density k.

Starting with the data set $X \subset \mathbb{R}^n$ as above, we continue to examine the string of functions

$$X = \pi_0 V_{s_0}(X) \longrightarrow \pi_0 V_{s_1}(X) \longrightarrow \cdots \longrightarrow \pi_0 V_{s_p}(X) = *$$

as in (1.1), but with a different interpretation.

These functions together determine a graph $\Gamma(X)$, whose vertices are pairs $(s_i, [x])$, where $[x]$ (sometimes $[x] = [x]_i$ or $[x] = [x]_{s_i}$, for precision) is a path component of a data point $x \in X$ in the complex $V_{s_i}(X)$. The edges of the graph have the form $(s_i, [x]) \to (s_{i+1}, [x])$. The path component represented by x in $V_{s_i}(X)$ maps to the path component represented by x in $V_{s_{i+1}}(X)$ under the function $\pi_0 V_{s_i}(X) \to \pi_0 V_{s_{i+1}}(X)$, as is standard.

The graph $\Gamma(X)$ is the *hierarchy graph* for $\pi_0 V_*(X)$. This graph is a tree and is the hierarchical clustering arising from the complexes $V_s(X)$, but we go further.

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

The vertex $(s_i, [x])$ is a *branch point* if, equivalently,

(a) there are distinct path components $[u], [v]$ of $V_{s_{i-1}}(X)$ such that $[u] = [v] = [x]$ in $\pi_0 V_{s_i}(X)$, or
(b) the inclusion $[x]_{i-1} \subset [x]_i$ of path components is not surjective.

Remove all edges that terminate in branch points from the hierarchy graph $\Gamma(X)$, to produce a subgraph $\Gamma_0(X)$, called the *component graph*. The path components of the graph $\Gamma_0(X)$ are the *stable components* of the data set X.

Remark 1.4 These definitions can be generalized to arbitrary strings of functions

$$F\colon\ F_0 \xrightarrow{\alpha} F_1 \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} F_p.$$

The *hierarchy graph* $\Gamma(F)$ has vertices (i, x) with $x \in F_i$, and has edges

$$(i, x) \to (i+1, \alpha(x)).$$

I say that the vertex (i, x) is a *branch point* of $\Gamma(F)$ if there are distinct elements $y, z \in F_{i-1}$ such that $\alpha(y) = \alpha(z) = x$. Remove all edges terminating in branch points from the graph $\Gamma(F)$ to form the *component graph* $\Gamma_0(F)$, and then the *stable components* of F are the path components of the graph $\Gamma_0(F)$.

At this level of generality, the hierarchy graph $\Gamma(F)$ is a disjoint union of trees, one for each element of the set F_p.

Example 1.5 The hierarchy graph $\Gamma(\pi_0 L_{*,k}(X))$ that is associated with the string of functions (1.2) is the tree of the HDBSCAN* algorithm.

The stable components for the string (1.2), meaning the path components of the component graph $\Gamma_0(\pi_0 L_{*.k}(X))$, are said to be clusters in [5, Sec.2.3].

Applying the path component construction π_0 to the array of simplicial complexes

$$\begin{array}{ccc} L_{s_i,k}(X) & \longrightarrow & L_{s_{i+1},k}(X) \\ \uparrow & & \uparrow \\ L_{s_i,k+1}(X) & \longrightarrow & L_{s_{i+1},k+1}(X) \end{array}$$

produces an array of functions

$$\begin{array}{ccc} \pi_0 L_{s_i,k}(X) & \longrightarrow & \pi_0 L_{s_{i+1},k}(X) \\ \uparrow & & \uparrow \\ \pi_0 L_{s_i,k+1}(X) & \longrightarrow & \pi_0 L_{s_{i+1},k+1}(X). \end{array}$$

This array is finite, and it is a special case of a finite array of functions F having the form

$$\begin{array}{ccccc} F_{0,0} & \xrightarrow{\alpha} & F_{1,0} & \longrightarrow & \cdots \\ {\scriptstyle\beta}\uparrow & & \uparrow{\scriptstyle\beta} & & \\ F_{0,1} & \xrightarrow[\alpha]{} & F_{1,1} & \longrightarrow & \cdots \\ \uparrow & & \uparrow & & \\ \vdots & & \vdots, & & \end{array}$$

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

where $F_{i,k} = \pi_0 L_{s_i,k}(X)$.

There is again a "hierarchy" graph $\Gamma(F)$ with vertices $((m,k),x)$ with $x \in F_{m,k}$, and edges

$$\big((m,k),x\big) \longrightarrow \big((m+1,k),\alpha(x)\big) \quad \text{and} \quad \big((m,k+1),y\big) \longrightarrow \big((m,k),\beta(y)\big),$$

called *horizontal* and *vertical* edges, respectively.

The vertex $((m,k),x)$ is a *horizontal branch point* if there are two elements y, $z \in F_{m-1,k}$ such that $\alpha(y) = \alpha(z) = x$. Similarly, $((m,k),x)$ is a *vertical branch point* if there are elements $u, v \in F_{m,k+1}$ such that $\beta(u) = \beta(v) = x$.

Remove all horizontal edges terminating in horizontal branch points and remove all vertical edges terminating in vertical branch points from the graph $\Gamma(F)$, to form the *component graph* $\Gamma_0(F)$. Observe that the graphs $\Gamma(F)$ and $\Gamma_0(F)$ have the same vertices. The path components $\pi_0\Gamma_0(F)$ are the *stable components* of the array F.

Example 1.6 The set of vertices for the complex $L_{s,k}(X)$ is empty for $k \geq m$, for some m. Holding s fixed and letting k vary gives a string of functions

$$\varnothing = \pi_0 L_{s,m}(X) \longrightarrow \pi_0 L_{s,m-1}(X) \longrightarrow \cdots \longrightarrow \pi_0 L_{s,0}(X) = \pi_0 V_s(X).$$

The corresponding graph $\Gamma(\pi_0 L_{s,*}(X))$ is a disjoint union of trees, which is the analogue of the cluster tree of a density function [6] for the present context. The cluster tree is also a type of hierarchy graph; see also [3].

2 Layers

Suppose again that $X \subset \mathbb{R}^n$ is a finite data set, and consider the ascending sequence of Vietoris–Rips complexes

$$X = V_{s_0}(X) \subset V_{s_1}(X) \subset \cdots \subset V_{s_p}(X) \tag{2.1}$$

that is associated with the phase change numbers $0 = s_0 < s_1 < s_2 < \cdots < s_p$. Form the associated string of functions

$$X = \pi_0 V_{s_0}(X) \longrightarrow \pi_0 V_{s_2}(X) \longrightarrow \cdots \longrightarrow \pi_0 V_{s_p}(X) = *$$

between path components.

Given an edge $(s_i, [x]_{s_i}) \to (s_{i+1}, [x]_{s_{i+1}})$ in the associated hierarchy graph $\Gamma(F) = \Gamma(X)$, the vertex $(s_{i+1}, [x]_{s_{i+1}})$ is not a branch point if and only if the cardinalities of the path components $[x]_{s_i} \in \pi_0 V_{s_i}(X)$ and $[x]_{s_{i+1}} \in \pi_0 V_{s_{i+1}}(X)$ coincide. This means that $[x]_{s_i} = [x]_{s_{i+1}}$ as subsets of X.

It follows that a stable component for the sequence $F = \{\pi_0 V_{s_j}(X)\}$ is a string of edges

$$P\colon \big(i, [x]_i\big) \longrightarrow (i+1, [x]_{i+1}) \longrightarrow \cdots \longrightarrow \big(i+n, [x]_{i+n}\big)$$

in $\Gamma(F)$ of maximal length such that all components $[x]_j \in \pi_0 V_{s_j}(X)$ coincide as subsets of the data set X.

Each stable component P as above contains a unique branch point, namely the vertex $(i, [x]_i)$ at the beginning of the string. It follows that stable components of X can be identified with the branch points of the hierarchy $\Gamma(F) = \Gamma(X)$.

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

Fatten up the sequence (2.1) of Vietoris–Rips complexes to the array $\{L_{s_i,k}(X)\}$ of Lesnick complexes, write $F_{i,k} = \pi_0 L_{s_i,k}(X)$, and form the graph $\Gamma(F)$ as in the last section.

Form a subgraph $\Gamma'_0(F)$ of $\Gamma(F)$ by saying that

$$(2.2) \qquad \big((i,k),[x]\big) \longrightarrow \big((i+1,k),[x]\big) \quad \text{or} \quad \big((i,k),[z]\big) \longrightarrow \big((i,k-1),[z]\big)$$

is an edge of $\Gamma'_0(F)$ if and only if the path components $[x] \in \pi_0 L_{i,k}(X)$ and $[x] \in \pi_0 L_{i+1,k}(X)$ (resp. $[z] \in \pi_0 L_{i,k}(X)$ and $[z] \in \pi_0 L_{i,k-1}(X)$) coincide as subsets of X.

Remark 2.1 If $[z] = [z]_{i,k}$ is path component of $L_{i,k}(X)$, then z represents a path component $[z]_{i,k-1}$ of $L_{i,k-1}(X)$, and there is an inclusion $[z]_{i,k} \subset [z]_{i,k-1}$ of subsets of X. If the element $((i,k-1),[z]) = ((i,k-1),[z]_{i,k-1})$ is a vertical branch point, then $[z]_{i,k-1}$ contains a path component from $L_{i,k}(X)$ in addition to $[z]_{i,k}$, and so $[z]_{i,k} \neq [z]_{i,k-1}$ as subsets of X.

Similar considerations hold in the horizontal case: if $[x]_{i,k} = [x]_{i+1,k}$ as subsets of X, then $((i+1,k),[x]_{i+1,k})$ cannot be a horizontal branch point.

The graph $\Gamma'_0(F)$ is the *layer graph* for F. The edges of $\Gamma'_0(F)$ are edges of the graph $\Gamma(F)$ for which the size of path component subsets is preserved.

If either of the edges in (2.2) is in the layer subgraph $\Gamma'_0(F)$, then the target in each case cannot be a horizontal (resp. vertical) branch point, because of the preservation of size of path components. It follows that the layer graph $\Gamma'_0(F)$ is a subgraph of the component graph $\Gamma_0(F)$.

Suppose that $((s,k),[x])$ is a vertex of $\Gamma'_0(F)$, and suppose given a path

$$P\colon ((t_0,k_0),[y_0]) \longrightarrow \cdots \longrightarrow \big((t_n,k_n),[y_n]\big)\big) = \big((s,k),[x]\big)$$

in $\Gamma'_0(F)$ that terminates at $((s,k),[x])$. Then $x \in X$ is in all path components $[y_i] \in \pi_0 L_{t_i,k_i}(X)$, since these subsets of X are constant through the path, so that we can rewrite the path P as

$$P\colon \big((t_0,k_0),[x]\big) \longrightarrow \cdots \longrightarrow \big((t_n,k_n),[x]\big) = \big((s,k),[x]\big).$$

In general, the path component L of a vertex $((s,k),[x])$ in $\Gamma'_0(F)$ consists of vertices of the form $((t,r),[x])$.

For a vertex $((t,r),[x])$ and a path P as above, if (t,r) satisfies $t_0 \le t \le t_n = s$ and $k = k_n \le r \le k_0$, then $x \in L_{t_0,k_0}(X)$ so that $x \in L_{t,r}(X)$, and x represents a path component $[x]$ for the three complexes

$$L_{t_0,k_0}(X) \subset L_{t,r}(X) \subset L_{t_n,k_n}(X).$$

It follows that the three path components represented by x coincide, so that $((t,r),[x])$ is in the path component of $((s,k),[x])$.

It follows that every path P in the layer graph $\Gamma'(F)$ generates a *square* $\mathrm{sq}(P)$ of vertices $((t,r),[x])$ with $t_0 \le t \le s$ and $k \le r \le k_0$, and this square lies in the path component of $((s,k),[x])$.

Suppose now that L is a path component in $\Gamma'_0(F)$ of a vertex $((t,r),[y])$. Find the elements $((s_i,k_i),[y])$, $i = 1,\dots,n$, of the component L that are maximal in s_i and minimal in k_i. Find all paths $Q_1,\dots,Q_n$ in L that terminate in one of the $((s_i,k_i),[y])$, and are maximal in the sense that they cannot be extended to longer

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

paths. Then the path component L in $\Gamma_0'(F)$ is a union of the squares associated with these maximal paths, in the sense that

$$L = \bigcup_{i=1}^{n} \mathrm{sq}(Q_i). \tag{2.3}$$

We have a graph inclusion $\Gamma_0'(F) \subset \Gamma_0(F)$, and both graphs have the same vertices. It follows that each path component (stable component) P of the component graph $\Gamma_0(F)$ is a disjoint union of components of the graph $\Gamma_0'(F)$, meaning that $P = \bigsqcup_j L_j$, where the subsets L_j are path components of $\Gamma_0'(F)$ that are contained in P. In other words, each stable component is a disjoint union of layers. At the same time, we know from (2.3) that each layer L_j is a union of squares.

These observations, taken together, give a geometric picture of the stable components for the Lesnick filtration $\{L_{s,k}(X)\}$ of a data set X.

Example 2.2 The distinction between stable components and layers potentially appears in all subdiagrams of the array $\{L_{s_i,k}(X)\}$ for which the complexes involved do not share a common set of vertices.

Such is the case for the diagram of complexes

$$L_{s_0,k}(X) \subset L_{s_1,k}(X) \subset \cdots \subset L_{s_p,k}(X),$$

that produces the HDBSCAN* algorithm of Example 1.3. The stable components (or clusters of [5]) break up into disjoint unions of layers in this case, but this decomposition remains to be interpreted.

3 Scoring

Continue with the data set $X \subset \mathbb{R}^n$; suppose again that the distinct lengths between elements of the finite set X have the form

$$0 = s_0 < s_1 < \cdots < s_p$$

and form the array of Lesnick complexes $\{L_{s_i,k}(X)\}$. These complexes are subcomplexes of the largest Vietoris–Rips complex $V_{s_p}(X)$, which can be identified with a large simplex Δ^N, where X has $N+1$ elements.

Write $K_{s_i,k}$ for the set of vertices of the complex $L_{s_i,k}(X)$. Then there is an array of inclusions of vertices $\{K_{s_i,k}\}$ and a collection of surjective functions

$$p\colon K_{s_i,k} \longrightarrow \pi_0 L_{s_i,k}(X). \tag{3.1}$$

These functions p are natural in i and k, and together define a map of arrays. All sets $K_{s_i,k}$ are subsets of the data set X. The sets $K_{s_i,0}$ coincide with X.

The function p of (3.1) is defined by $p(x) = [x]$, where $[x]$ is the path component represented by the vertex x. The subset $p^{-1}([x])$ of $K_{s_i,k}$ is the set of members of the path component $[x]$ of the complex $L_{s_i,k}(X)$.

The array of sets $K_{s_i,k}$ defines a graph $\Gamma(K_{*,*})$ as before. Vertices are pairs $((s_i,k),y)$ with $y \in K_{s_i,k}$, and there are horizontal and vertical edges having the respective forms

$$\big((s_i,k),y\big) \longrightarrow \big((s_{i+1},k),y\big) \quad \text{and} \quad \big((s_i,k),y\big) \longrightarrow \big((s_i,k-1),y\big).$$

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

The functions p of (3.1) define a graph homomorphism

$$p\colon \Gamma(K_{*,*}) \longrightarrow \Gamma\big(\pi_0 L_{*,*}(X)\big),$$

which is defined on vertices by $p((s_i,k),y) = ((s_i,k),[y])$.

Given $x \in X$, there is a subgraph $\Gamma_x(K) \subset \Gamma(K_{*,*})$ that has vertices $((s_j,k),x)$ with $x \in K_{s_j,k}$. Each subgraph $\Gamma_x(K)$ is connected, and the subgraphs $\Gamma_x(K)$ are the connected components of $\Gamma(K_{*,*})$. It follows that there is a graph isomorphism

$$\bigsqcup_{x\in K_0} \Gamma_x(K) \xrightarrow{\cong} \Gamma(K_{*,*}).$$

Let P be a stable component for the diagram $\pi_0 L_{*,*}(X)$, and form the pullback diagram

(3.2)
$$\begin{array}{ccc} P \cap \Gamma(K_{*,*}) & \longrightarrow & \Gamma(K_{*,*}) \\ \downarrow & & \downarrow p \\ P & \xrightarrow[i]{} & \Gamma(\pi_0 L_{*,*}(X)), \end{array}$$

where i is the composite inclusion $P \subset \Gamma_0(\pi_0 L_{*,*}(X)) \subset \Gamma(\pi_0 L_{*,*}(X))$ of graphs.

There is a graph isomorphism

$$\bigsqcup_{x\in X} P \cap \Gamma_x(K) \xrightarrow{\cong} P \cap \Gamma(K_{*,*}),$$

while the set of vertices $(P \cap \Gamma(K_{*,*})_0$ of the graph $P \cap \Gamma(K_{*,*})$ is a disjoint union of the sets $p^{-1}((s_i,k),[y])) \subset K(s_i,k)$ with $((s_i,k),[y]) \in P$.

It is standard to write $|F|$ for the number of elements in a finite set F. It follows that there is an identity

(3.3)
$$\sum_{x\in X} |(P \cap \Gamma_x(K))_0| = \sum_{((s_i,k),[y])\in P} |[y]|,$$

where $(P \cap \Gamma_x(K))_0$ is the set of vertices of the graph $P \cap \Gamma_x(K)$. The number

$$\zeta(x,P) = |(P \cap \Gamma_x(K))_0|$$

is the number of vertices in P that are represented by x. It is a combinatorial *stability measure* of x with respect to P. This description of $\zeta(x,P)$ is adapted from the stability measure for x of [5], which is an analytic invariant.

The combinatorial persistence *score* $\sigma(P)$ of a stable component P is the sum of all stability measures $\zeta(x,P)$:

$$\sigma(P) = \sum_{x\in K_0} \zeta(x,P),$$

so that

$$\sigma(P) = \sum_{x\in X} |(P \cap \Gamma_x(K))_0| = \sum_{((s_i,k),[y])\in P} |[y]|,$$

on account of the identity (3.3).

Analogs of diagram (3.2) lead to similar analyses for all subobjects of the graph $\Gamma(\pi_0 L_{*,*}(X))$, including

(a) layers for $\Gamma(\pi_0 L_{*,*}(X))$,

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

(b) stable components and layers for the graph $\Gamma(\pi_0 L_{*,k}(X))$ (HDBSCAN* case), and
(c) stable components (which are also layers) for the graph $\pi_0 L_{*,0}(X) = \pi_0 K_*(X)$.

If the subobject $L \subset \Gamma(\pi_0 L_{*,*}(X))$ is a layer, then we have a pullback diagram

$$\begin{array}{ccc} L \cap \Gamma(K_{*,*}) & \longrightarrow & \Gamma(K_{*,*}) \\ \downarrow & & \downarrow p \\ L & \xrightarrow[i]{} & \Gamma(\pi_0 L_{*,*}(X)), \end{array}$$

where i is the inclusion of the layer L. Then

$$L \cap \Gamma(K_{*,*}) = \coprod_{x \in X} L \cap \Gamma_x(K).$$

In this case, either $L \cap \Gamma(K_{*,*}) = \varnothing$, or the intersection has a vertex $((s_i, k), x)$ such that x represents every vertex of L. It follows that

$$|(L \cap \Gamma(K_{*,*})_0| = |L_0| \cdot |[x]|,$$

where x is any choice of representative for a vertex $((s_i, k), [x])$ of L, and L_0 is the set of vertices of the layer L. This is consistent with counting cardinalities of the fibres $p^{-1}((s_i, k), [x])$, as dictated by the right-hand side of equation (3.3).

The score $\sigma(L)$ of a layer L then has the rather simple form

$$\sigma(L) = \sum_{x \in X} |(L \cap \Gamma_x(K))_0| = \sum_{((s_i,k),[y]) \in L} |[y]| = |L_0| \cdot |[x]|.$$

Note finally that since a stable component P is a disjoint union

$$P = L_1 \sqcup \cdots \sqcup L_k$$

and scoring for P and L amounts to counting fibres for the graph map p, then there is a relation

$$\sigma(P) = \sum_{i=1}^{k} \sigma(L_i),$$

which relates the score of P to the scores of its constituent layers.

4 Adding a Point

Suppose that $Y = X \sqcup \{y\}$ and that $d(y, x_0) < r$ for some $x_0 \in X$. The number r should be tiny; this will be made precise later.

In this section, we compare stable components for the hierarchies $\Gamma(X) = \Gamma(V_*(X))$ and $\Gamma(Y) = \Gamma(V_*(Y))$ arising from the respective Vietoris–Rips complexes.

The comparison requires a more versatile approach: we expand the systems $V_*(X)$ and $V_*(Y)$ (that are indexed by phase change numbers in previous sections) to functors

$$s \longmapsto V_s(X) \quad \text{and} \quad s \longmapsto V_s(Y),$$

respectively, for numbers s in an interval $[0, R]$, where R is an upper bound for the phase change numbers of both X and Y.

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

We are then entitled to simplicial set maps $i : V_s(X) \subset V_s(Y)$ that are natural in $s \in [0, R]$, which are induced by the inclusion $X \subset Y$. The interesting behaviour of both of these systems, in isolation, occurs at the respective phase shift numbers, but a robust comparison mechanism requires all parameters.

The functor $s \mapsto V_s(X)$ defines a set-valued functor $s \mapsto \pi_0 V_s(X)$, and an expanded hierarchy $\Gamma(X)$ can be defined by analogy with what we have above: $\Gamma(X)$ is a graph (or a category) with vertices $(s, [x]_s)$ with $s \in [0, R]$ and $[x]_s \in \pi_0 V_s(X)$, and has edges $(s, [x]_s) \to (t, [x]_t)$ with $s \leq t$. We will sometimes write $[x]_{s,X}$ for $[x]_s$ where comparisons of data sets are involved.

An inclusion of data sets $X \subset Y$ then induces a morphism of graphs

$$i\colon \Gamma(X) \longrightarrow \Gamma(Y).$$

For such comparisons, it is convenient to phrase the discussion in terms of layers. The layers of $\Gamma(X)$ can be identified with edges $L\colon (s, [x]_s) \to (t, [x]_t)$ of $\Gamma(X)$ such that

(a) s and t are phase shift numbers of X;
(b) $[x]_s = [x]_t$ as subsets of X;
(c) the *length* $t - s$ of the edge L is maximal with respect to the first two conditions.

The vertex $(s, [x]_s)$ at the initial vertex of a layer must be a branch point, and the full collection of layers of $\Gamma(X)$ can be identified with its set $Br(X)$ of branch points. The layers of $\Gamma(X)$ are the path components (the stable components) of the component graph $\Gamma_0(X)$ of the first section.

A *partial layer* for X is an edge $(s, [x]_s) \to (t, [x]_t)$ of $\Gamma(X)$ such that $[x]_s = [x]_t$. In other words, a partial layer is an edge that satisfies only condition (b) above.

Returning to the situation $X \subset Y = X \sqcup \{y\}$ of interest, write

$$0 = s_0 < s_1 < \cdots < s_k$$

for the list of phase change numbers (*i.e.*, distances between points) for X.

We suppose that y is close to an element $x_0 \in X$, in the sense that $d(y, x_0) < r$, where

$$r < s_{i+1} - s_i \tag{4.1}$$

for all $i \geq 0$. The inequality (4.1) is the meaning of the requirement that the number r should be tiny.

Lemma 4.1 *Suppose that $x \in X$ has the property that y is not a member of the path component $[x]_{s,Y}$ of $V_s(Y)$. Then the inclusion $[x]_{s,X} \subset [x]_{s,Y}$ is a bijection.*

Proof Suppose that $x' \in [x]_{s,Y}$. Then $y \notin [x]_{s,Y}$, so there is a path

$$x = y_0 \longleftrightarrow y_1 \longleftrightarrow \cdots \longleftrightarrow y_r = x'$$

such that all $y_i \in X$. It follows that $x' \in [x]_{s,X}$. ∎

Corollary 4.2 *Suppose that $s \leq t$ and that $y \notin [x]_{t,Y}$ in $V_t(Y)$. If $(s, [x]_{s,X}) \to (t, [x]_{t,X})$ is a partial layer for X, then $(s, [x]_{s,Y}) \to (t, [x]_{t,Y})$ is a partial layer for Y.*

Note that if $y \notin [x]_{t,Y}$, then $y \notin [x]_{s,Y}$, because $[x]_{s,Y} \subset [x]_{t,Y}$ for $s \leq t$.

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

Proof of Corollary 4.2 There is a commutative diagram of inclusions

$$\begin{array}{ccc} [x]_{s,X} & \xrightarrow{=} & [x]_{t,X} \\ {\scriptstyle =}\downarrow & & \downarrow{\scriptstyle =} \\ [x]_{s,Y} & \longrightarrow & [x]_{t,Y} \end{array}$$

in which the vertical maps are bijections by Lemma 4.1. The claim follows. ∎

Remark There could be layers $(s,[x]_{s,X}) \to (t,[x]_{t,X})$ of X such that $y \notin [x]_{s,Y}$ and $y \in [x]_{t,Y}$. In that case, the corresponding edge $(s,[x]_{s,Y}) \to (t,[x]_{t,Y})$ cannot be a partial layer of Y, because the set $[x]_{t,Y}$ is strictly larger than $[x]_{s,Y}$.

In general, if L is a simplicial complex and V is a set of vertices of L, the *full subcomplex* of L on the set V consists of those simplices σ of L having all of their vertices in the set V.

Write $V_s(X)(y)$ for the full subcomplex of $V_s(X)$, on the set of those vertices x such that $[x]_{s,Y} = [y]_{s,Y}$ in $V_s(Y)$.

Lemma 4.3 *Suppose that $s_i \leq s < s_{i+1} - r$ for all $i \geq 1$. Then we have the following.*

(i) *The space $V_s(X)(y)$ is connected.*
(ii) *The function $\pi_0 V_s(X) \to \pi_0 V_s(Y)$ is a bijection.*

Proof (i) Suppose that x is a vertex of $V_s(X)(y)$. Then there is a sequence of 1-simplices

$$x = y_0 \longleftrightarrow y_1 \longleftrightarrow \cdots \longleftrightarrow y_n \longleftrightarrow y$$

in $V_s(Y)$ with $y_i \in X$. Then $d(y,x_0) < r$, so $d(y_n,x_0) < s + r$ and $[x] = [x_0]$ in $V_{s+r}(X)$. But $V_s(X) = V_{s+r}(X)$, since $s_i \leq s < s+r < s_{i+1}$, and so $[x] = [x_0]$ in $V_s(X)$. This is true for all vertices $x \in V_s(X)(y)$, so that $V_s(X)(y)$ is connected.

(ii) Suppose that x_1, x_2 are vertices of $V_s(X)$ such that $[y] \neq [x_1]$ and $[y] \neq [x_2]$ in $V_s(Y)$. Then it follows from Lemma 4.1 that $[x_1]_{s,X} = [x_2]_{s,X}$ in $V_s(Y)$ if and only if $[x_1]_{s,Y} = [x_2]_{s,Y}$ in $V_s(X)$.

The function $i_*\colon \pi_0 V_s(X) \to \pi_0 V_s(Y)$ is surjective, since $d(y,x_0) < r < s_1 \leq s$, so that $[y]_{s,Y} = [x_0]_{s,Y}$. The previous paragraphs imply that i_* is injective. ∎

Corollary 4.4 *Under the assumptions for Lemma 4.3, the map $i_*\colon \pi_0 V_{s_i}(X) \to \pi_0 V_{s_i}(Y)$ is a bijection at all phase change numbers s_i for X such that $i \geq 1$.*

Corollary 4.5 *Suppose that the edge $(s,[x]_{s,X}) \to (t,[x]_{t,X})$ is a partial layer of X. Suppose that $y \in [x]_{s,Y}$ and that s and t are phase change numbers for X, with $s_1 \leq s \leq t$.*

Then the edge $(s,[x]_{s,Y}) \to (t,[x]_{t,Y})$ of $\Gamma(Y)$ is a partial layer of Y.

Proof The spaces $V_s(X)(y)$ and $V_t(X)(y)$ are connected by Lemma 4.3. It follows that

$$\begin{aligned} [x]_{s,Y} &= \{y\} \sqcup \big(V_s(X)(y)\big)_0 = \{y\} \sqcup [x]_{s,X} \\ &= \{y\} \sqcup [x]_{t,X} = \{y\} \sqcup \big(V_t(Y)(t)\big)_0 = [x]_{t,Y} \end{aligned}$$

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

in Y. ∎

Corollary 4.6 *Suppose that $(s, [x]_{s,Y}) \to (t, [x]_{t,Y})$ is a partial layer of Y such that $y \in [x]_{s,Y}$ and $x \neq y$. Suppose that s, t are phase change numbers for X with $s_1 \leq s \leq t$.*

Then the edge $(s, [x]_{s,X}) \to (t, [x]_{t,X})$ is a partial layer of X.

Proof The function $V_s(X)(y)_0 \to V_t(X)(y)_0$ is a bijection, since $[x]_{s,Y} \to [x]_{t,Y}$ is a bijection, and the spaces $V_s(X)(y)$ and $V_t(X)(y)$ are path connected by Lemma 4.3. It follows that the function $[x]_{s,X} \to [x]_{t,X}$ is a bijection. ∎

Suppose that $L\colon (s, [x]_{s,X}) \to (t, [x]_{t,X})$ is a layer of X such that either $y \in [x]_{s,Y}$ or $y \notin [x]_{t,Y}$ and $s_1 \leq s$. In both cases, the edge $(s, [x]_{s,Y}) \to (t, [x]_{t,Y})$ is a partial layer for Y, by Corollaries 4.5 and 4.2, respectively.

Suppose that s' and t' are phase change numbers of X such that

$$s_1 \leq s' \leq s \leq t \leq t'$$

and such that $(s', [x]_{s',Y}) \to (t', [x]_{t',Y})$ is a partial layer of Y. Then we have the following.

(a) If $y \in [x]_{s,Y}$, then $y \in [x]_{s',Y}$. The maps $[x]_{s',X} \to [x]_{s,X}$ and $[x]_{t,X} \to [x]_{t',X}$ are bijections on account of the connectedness assertion of Lemma 4.3, and it follows that $(s', [x]_{s',X}) \to (t', [x]_{t',X})$ is a partial layer of X. But L is a layer, so that $s' = s$ and $t = t'$ by the maximality condition.

(b) Suppose that $y \notin [x]_{t,Y}$. Then $y \notin [x]_{t',Y}$ since $[x]_{t,Y} = [x]_{t',Y}$. Then the edge $(s', [x]_{s',X}) \to (t', [x]_{t',X})$ is a partial layer of X, since the diagram of functions

$$\begin{array}{ccc} [x]_{s',X} & \longrightarrow & [x]_{t',X} \\ {\scriptstyle =}\downarrow & & \downarrow{\scriptstyle =} \\ [x]_{s',Y} & \xrightarrow[=]{} & [x]_{t',Y} \end{array}$$

commutes, as in the proof of Corollary 4.2. The maximality condition for the layer L again implies that $s' = s$ and $t = t'$.

We have proved the following result.

Proposition 4.7 *Suppose that $L\colon (s_i, [x]_{s_i,X}) \to (s_j, [x]_{s_j,X})$ is a layer of X with $s_1 \leq s_i$. Suppose that either $y \in [x]_{s_i,Y}$ or $y \notin [x]_{s_j,Y}$.*

Then the edge $L\colon (s_i, [x]_{s_i,Y}) \to (s_j, [x]_{s_j,Y})$ is a partial layer for Y, and the layer $(s, [x]_{s,Y}) \to (t, [x]_{t,Y})$ containing L must have $s_{i-1} < s \leq s_i$ and $s_j \leq t < s_{j+1}$.

The numbers s and t in the statement of Proposition 4.7 are phase change numbers for Y. One can show that the phase change numbers for Y lie in the open intervals $(s_i - r, s_i + r)$ centred at the phase change numbers s_i for X. This is a constraint on the position of the numbers s and t above, but the differences $s_i - s_{i-1}$ and $s_{j+1} - s_j$ could be large, and there is no theoretical information, for example, on whether s is close to s_i or close to s_{i-1}.

Acknowledgments I would like to thank the referee for raising the question that is the subject of the final section of this paper. This question is quite natural, and the

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press

collection of ideas behind the solution that is presented here has proved to be quite provocative.

References

[1] A. J. Blumberg and M. Lesnick, *Universality of the homotopy interleaving distance.* 2017. arxiv:1705.01690.

[2] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, *Hierarchical density estimates for data clustering, visualization, and outlier detection.*. ACM Transactions on Knowledge Discovery from Data (TKDD) 10(2016), 1–51. https://doi.org/10.1145/2733381

[3] G. Carlsson and F. Memoli, *Multiparameter hierarchical clustering methods.* In: *Classification as a tool for research*. Stud. Classification Data Anal. Knowledge Organ, Springer, Berlin, 2010, pp. 63–70. https://doi.org/10.1007/978-3-642-10745-0_6

[4] M. Lesnick and M. Wright, *RIVET: visualization and analysis of two-dimensional persistent homology.* 2019. http://rivet.online.

[5] L. McInnes and J. Healy, *2017 IEEE International Conference on Data Mining Workshops. (ICDM Workshops 2017, New Orleans, LA, USA (November 18–21))*, IEEE, 2017, pp. 33–42. https://doi.org/10.1109/ICDMW.2017.12

[6] W. Stuetzle, *Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample*. J. Classification 20(2003), 25–47. https://doi.org/10.1007/s00357-003-0004-6

[7] A. Zomorodian, *Fast construction of the Vietoris-Rips complex*. Computers & Graphics 34(2010), 263–271. https://doi.org/10.1016/j.cag.2010.03.007

Department of Mathematics, University of Western Ontario, London, Ontario
e-mail : jardine@uwo.ca

https://doi.org/10.4153/S000843951900064X Published online by Cambridge University Press