Cluster graphs

Rick Jardine

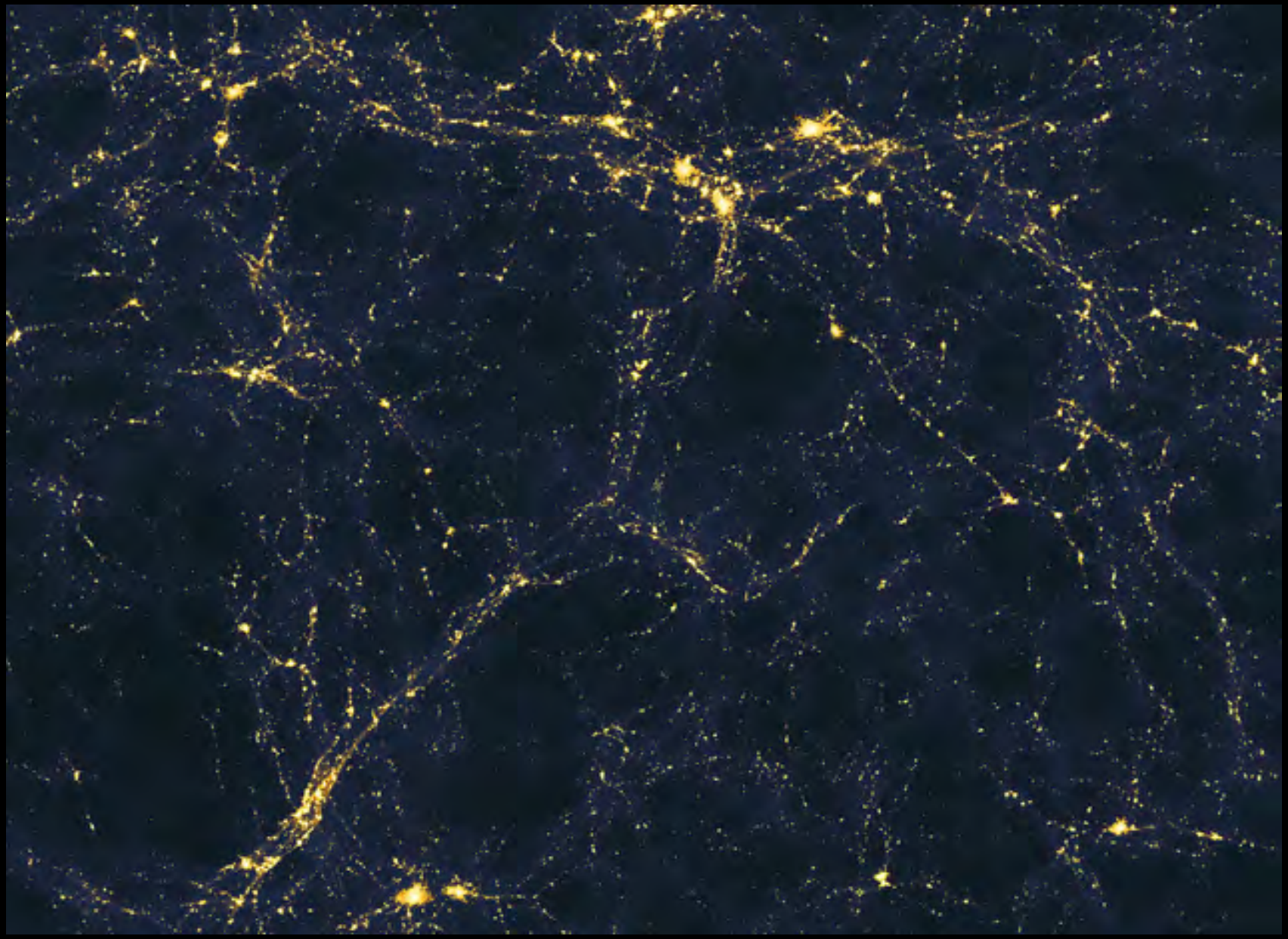
School of Mathematical and Statistical Sciences
Western University

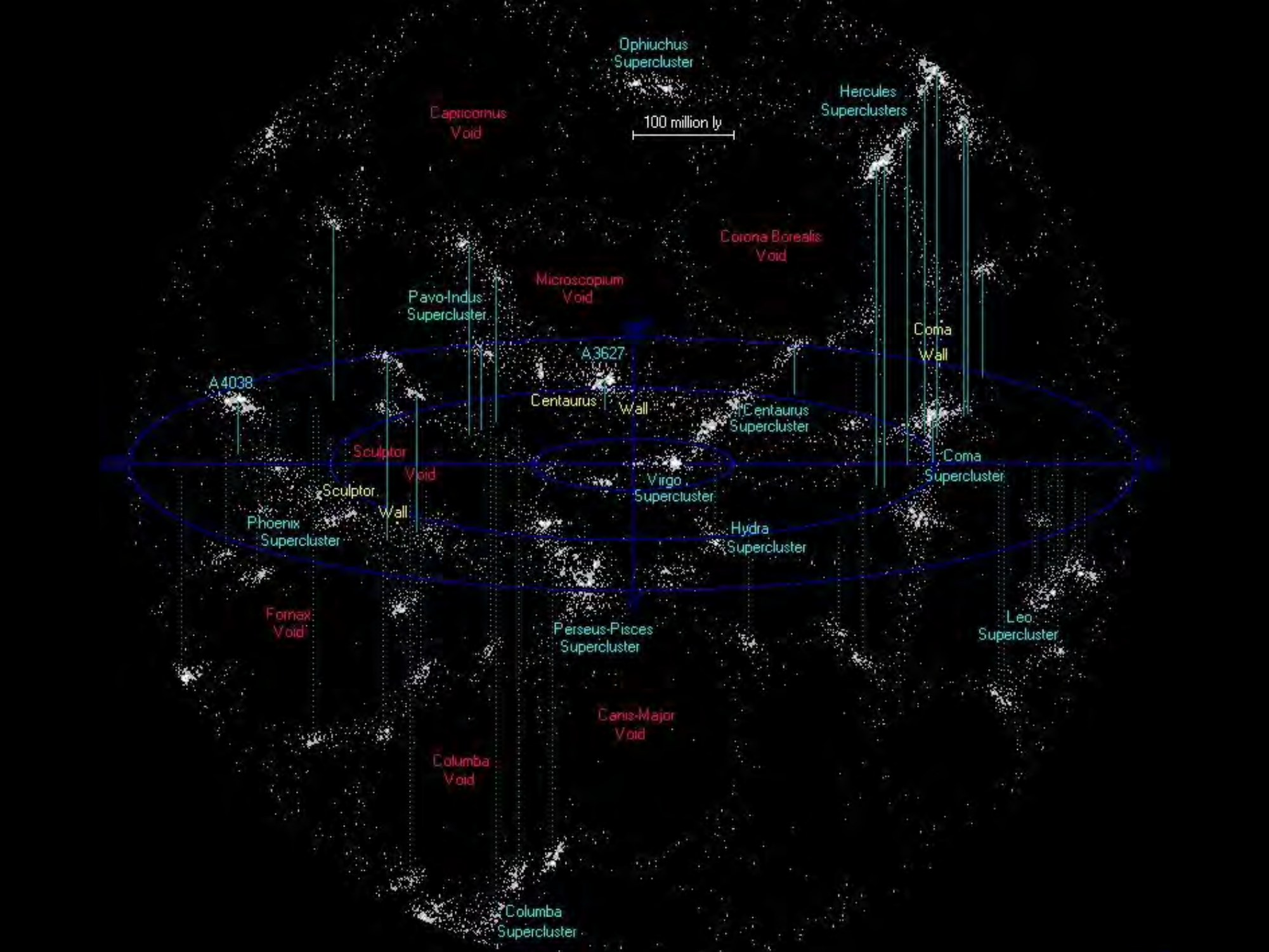
October 26, 2017

Foundational Mathematics for Machine Learning

Tutte Institute
Canadian Security Establishment (CSE)
Ottawa

May 23 – June 9, 2016




Ophiuchus Supercluster
Hercules Superclusters
Capricornus Void
100 million ly
Corona Borealis Void
Microscopium Void
Pavo-Indus Supercluster
Coma Wall
A3627
A4038
Centaurus Wall
Centaurus Supercluster
Sculptor Void
Coma Supercluster
Virgo Supercluster
Sculptor Wall
Phoenix Supercluster
Hydra Supercluster
Fornax Void
Perseus-Pisces Supercluster
Leo Supercluster
Canis-Major Void
Columba Void
Columba Supercluster

Topological Data Analysis

A **data cloud** is a finite set of points $X \subset \mathbb{R}^N$. (a metric space)

Basic idea: analyze regions of the data cloud X, by density.

Rips complex: $s > 0$: $V_s(X)$ has simplices $\{x_0, \ldots, x_n\}$ st $d(x_i, x_j) < s$ for all i, j.

- If $s < t$, then $V_s(X) < V_t(X)$
- $V_s(X)$ is discrete for s small, contractible for s big.
- There are only finitely many isomorphism types $V_i(X) = V_{s_i}(X)$.

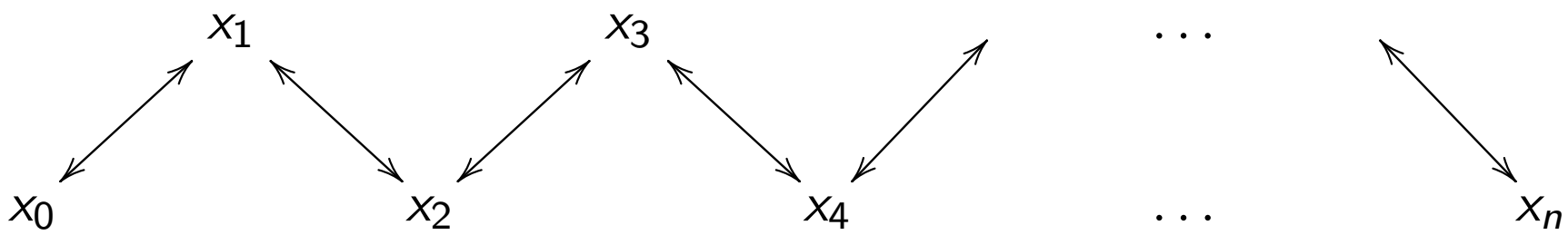
Have an sequence of complexes ("filtration", "dynamical system")

$$V_1(X) \subset V_2(X) \subset \cdots \subset V_k(X) \subset \ldots$$

What we care about is points and 1-simplices of $V_s(X)$: pairs of points (x, y) such that $d(x, y) < s$.

Path components

Say that points x, y are in the same **path component** of $V_s(X)$ (write $x \sim_s y$) if there is a string of segments (1-simplices)



in X with $x = x_0$, $y = x_n$, and $d(x_i, x_{i+1}) < s$ for all i.

Each pair (x_i, x_{i+1}) defines a 1-simplex of $V_s(X)$. The picture defines a polygonal path of 1-simplices of $V_s(X)$ between x and y.

x is related to y in $V_s(X)$ if there is a series of short hops (of length $< s$) through points of X.

$\pi_0 V_s(X) =$ the set of equivalence classes under $\sim_s$, is the set of **path components** of $V_s(X)$.

Varying the parameter s

If $s < t$ and $x \sim_s y$, then $x \sim_t y$.

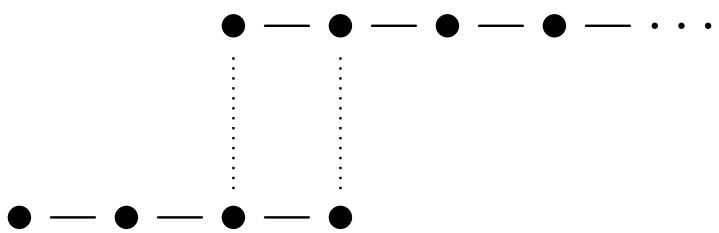
Hops of length $< s$ are of length $< t$.

There is a function of equivalence classes (path components)

$$\pi_0 V_s(X) \to \pi_0 V_t(X),$$

which is induced by the inclusion $V_s(X) \subset V_t(X)$.

Picture:



Cluster

We get a family of maps between path component sets

$$\pi_0 V_1(X) \to \pi_0 V_2(X) \to \cdots \to \pi_0 V_k(X) \to \dots$$

A "cluster" is a path component in some $V_i(X)$ that does not vary with i, "for a while".

How to express that? Suppose given functions

$$F(1) \xrightarrow{\alpha} F(2) \xrightarrow{\alpha} \dots \xrightarrow{\alpha} F(k) \xrightarrow{\alpha} \dots$$

For $p < q$, $x \in F(p)$, $y \in F(q)$, say that $x \sim y$ if $\alpha^{q-p}(x) = y$ and $(\alpha^{q-k})^{-1}(y) = \{\alpha^{k-p}(x)\}$ for all $p \leq k \leq q$:

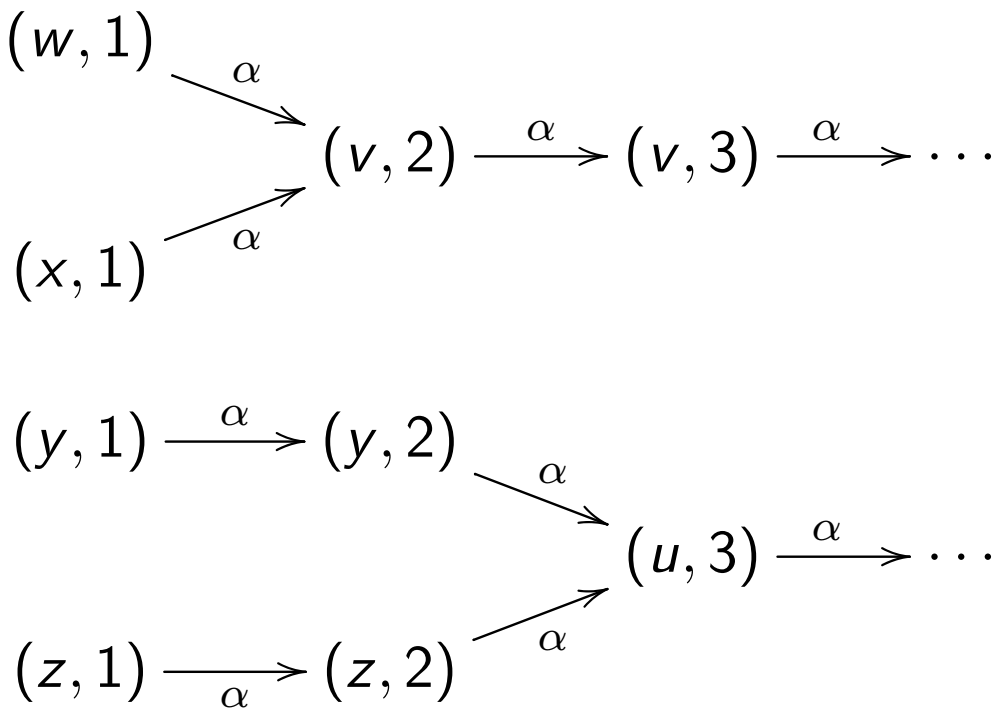
$$F(p) \to F(k) \to F(q)$$

Clusters are equivalence classes in $\cup_p F(p)$.

The graph $\Gamma(F)$

Sets and functions: $F: \ F(1) \xrightarrow{\alpha} F(2) \xrightarrow{\alpha} \ldots \xrightarrow{\alpha} F(k) \xrightarrow{\alpha} \ldots$

Graph $\Gamma(F)$: vertices (x, i), $x \in F(i)$, edges $(x, i) \to (\alpha(x), i+1)$.



A **branch point** is a vertex (x, i) with more than one incoming edge $(y, i-1) \to (x, i)$.

The cluster graph

Remove all edges of $\Gamma(F)$ terminating in branch points to construct subgraph $\Gamma_0(F) \subset \Gamma(F)$

$\Gamma_0(F)$ is the **cluster graph** for F.

Graphs have path components, and the **clusters** are the path components of $\Gamma_0(F)$, ie. elements of $\pi_0\Gamma_0(F)$.

Alternatively: A cluster of F is a path

$$(x_0, i) \to (x_1, i+1) \to \cdots \to (x_p, i+p)$$

of max length in $\Gamma(F)$ st no $(x_j, i+j)$ is a branch point for $j > 0$.

NB: (x_0, i) is a branch point, or x_0 has no preimage in $F(i-1)$.

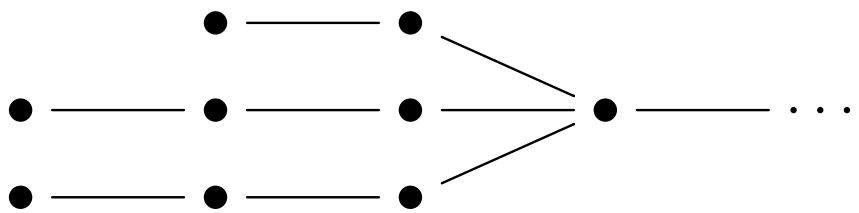
Example: A cluster of $\{\pi_0 V_i(X)\}$ starts with a path component $[x] \in \pi_0(V_i(X))$ which was strictly smaller in $V_{i-1}(X)$ (branch point) and has a fixed size through the maps

$$V_i(X) \to V_{i+1}(X) \to \cdots \to V_{i+p}(X)$$

for some maximal p.

Noise

The isolated groups of bright spots define "small" clusters. They join other clusters at some parameter value, which could be large.



The small clusters are "noise", up to some interpretation.

Two ways to address this:

1) Every element of $x_s \in \pi_0 V_s(X)$ has a cardinality $|x_s|$. Score each cluster

$$P: \ (x_s, s) \to (x_{s+1}, s+1) \to \cdots \to (x_{s+p}, s+p)$$

by setting $\sigma(P) = |x_s| \cdot p$. Compare scores of clusters.

2) Throw away the path components of small size during the computation process.

Comments

1) The **score**

$$\sigma(P) = |x_s| \cdot p = \sum_{(x_i,j)\in P} |x_i|$$

is the sum of the cardinalities $|x_i|$ of all path components appearing in the cluster P.

2) Clusters with big voids around them have higher scores than clusters of same size surrounded by smaller voids.

3) Scoring is relatively expensive. It can only be done after all other calculations.

4) Throwing away small path components (eg isolated stars, small groups) is brutal but computationally effective — can be done before constructing the cluster graph.

Higher dimensional persistence

The Rips complex has subcomplexes ("Lesnick complexes")

$$\cdots \subset L_{s,k+1}(X) \subset L_{s,k}(X) \subset \ldots L_{s,0}(X) = V_s(X)$$

defined by valence of vertices, and natural in s.

$x \in L_{s,k}(X)$ if it is a member of at least k edges ... another type of density measure.

Have a rectangular array of inclusions of complexes

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

all with potentially different vertices.

Abstraction

Computing path components gives rectangular array of functions

$$F_{i,k} = \pi_0 L_{s,k}(X): \qquad \begin{array}{ccc} F(s,k) & \xrightarrow{\alpha} & F(s+1,k) \\ {\scriptstyle\beta}\uparrow & & \uparrow{\scriptstyle\beta} \\ F(s,k+1) & \underset{\alpha}{\rightarrow} & F(s+1,k+1) \end{array}$$

There is a (directed) graph $\Gamma(F)$ with vertices $(x,(i,j))$ and edges

$$(x,(i,j)) \to (\alpha(x),(i+1,j)) \text{ and } (x,(i,j)) \to (\beta(x),(i,j+1)).$$

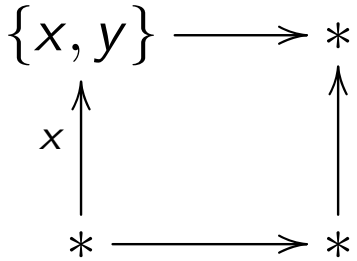
$(x,(i,j))$ is a **horizontal branch point** if there are distinct $(u,(i-1,j))$, $(v,(i-1,j))$ with $\alpha(u) = \alpha(v) = x$. **Vertical branch points** are defined similarly.

Removing edges ending at branch points gives the **cluster graph** $\Gamma_0(F) \subset \Gamma(F)$.

The **clusters** are the path components $\pi_0\Gamma_0(F)$.

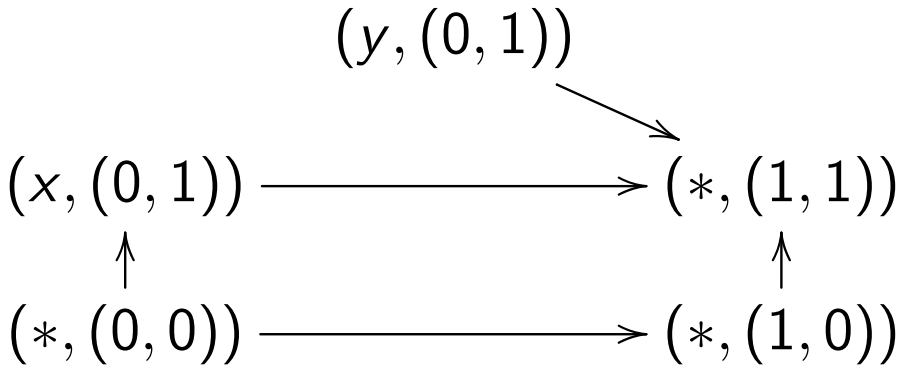
Example

F is the diagram of functions



$*$ is the one point set, and $x : * \to \{x, y\}$ picks out the element x.

Here's $\Gamma(F)$:



$(*, (1, 1))$ is the one (horizontal) branch point.

Example, cont.

$\Gamma_0(F)$ is constructed by removing the edges

$$(x,(0,1)) \to (*,(1,1)) \text{ and } (y,(0,1)) \to (*,(1,1))$$

$\Gamma_0(F)$ is the graph

$$(y,(0,1))$$

$$\begin{array}{ccc} (x,(0,1)) & & (*,(1,1)) \\ \uparrow & & \uparrow \\ (*,(0,0)) & \longrightarrow & (*,(1,0)) \end{array}$$

$\{(y,(0,1))\}$ is a noise object.

Scoring

$L_{s,k}(X)$ of a data cloud X — just an example.

A cluster P is a connected graph consisting of a set of vertices $(x,(s,k))$ with suitable edges.

For each vertex $(x,(s,k))$, the element x is a path component (a set of vertices) in $L_{s,k}(X)$.

The path component x has finite cardinality $|x|$.

The **score** $\sigma(P)$ of the cluster P is defined by

$$\sigma(P) = \sum_{(x,(s,k))\in P} |x|.$$

As before, we deal with noise by throwing away clusters with low scores, or by throwing away points $(x,(s,k))$ with $|x|$ small, or both.

Explanation

Suppose given an ascending sequence of finite sets

$$P: \ P_0 \subset P_1 \subset P_2 \subset \cdots \subset P_n.$$

The score $\sigma(P)$ of the sequence P is given by

$$\sigma(P) = \sum_{i=0}^{n} |P_i|.$$

$$P_1 = P_0 \sqcup (P_1 - P_0)$$
$$P_2 = P_0 \sqcup (P_1 - P_0) \sqcup (P_2 - P_1)$$

Multiplicities: The points of P_0 are counted $n+1$ times, the points of $P_1 - P_0$ are counted n times, ... , the points of $P_n - P_{n-1}$ are counted only once.

Comments

At least for clusters, we may have the first viable approach to multidimensional persistence.

These ideas apply to arrays of sets of all dimensions. eg. we could vary the data cloud X in persistence applications.

Persistent homology

The Rips complexes $V_s(X)$ have homology groups $H_n(V_s(X))$, $n \geq 0$, (coefficients in a fixed field k), all finite dimensional vector spaces because all complexes are finite.

The inclusions $V_i(X) \subset V_{i+1}(X)$ induce vector space morphisms

$$H_n(V_1(X)) \xrightarrow{t} H_n(V_2(X)) \xrightarrow{t} \dots$$

interpreted as a $k[t]$-module, a **persistence module**.

Standard theorem about finitely generated modules over a principal ideal domain says that a persistence module is a direct sum of finite torsion modules $k[t]/(t^p)$ (shifted).

The decomposition

$$M = \bigoplus_{i \geq 0} H_k(V_i(S))$$

is a fin. gen. graded $k[t]$-module, killed by some minimal power t^r. r is the **exponent** of M.

For homogeneous $z \in M$, the smallest n such that $t^n \cdot z = 0$ is the **period** of z.

There is a hom. $x \in M$ of period $r =$ index of M. Find one.

The quotient $M/(x)$ has exponent $n \leq r$. Find $z \in M/(x)$ of period n, and choose $y \in M$ such that $y \mapsto z$ under $M \to M/(x)$.

By adjusting, can find y of same period as z (a "splitting").

Consequence: $\langle x \rangle \oplus \langle y \rangle \to M$ has trivial kernel.

This is the start of an induction which shows that

$$M \cong \langle x \rangle \oplus \langle y \rangle \oplus \ldots \cong k[t]/(t^r) \oplus k[t]/(t^n) \oplus \cdots \oplus k[t]/(t^m).$$

The adjustment

$y \mapsto z$ under $M \to M/(x)$, and z has exponent n.

$$t^n \cdot y = t^s \cdot a \cdot x,$$

some $a \in k$ (all elements homogeneous).

1) If $s = r$ then the exponent of y is the exponent of z.

2) If $s < r$ then $t^n \cdot y$ has exponent $r - s$, so y has exponent $n + (r - s) \leq r$, so $n \leq s$, and the period of

$$y - t^{s-n} \cdot a \cdot x$$

is n, same as z.

Birth and death

The (finite dimensional, torsion) $k[t]$-module defined by the vector space morphisms

$$H_n(V_1(X)) \xrightarrow{t} H_n(V_2(X)) \xrightarrow{t} \dots$$

is a finite direct sum of modules of the form

$$k \cdot x \underset{\cong}{\xrightarrow{t}} k \cdot (tx) \underset{\cong}{\xrightarrow{t}} \dots \underset{\cong}{\xrightarrow{t}} k \cdot (t^{r-1}x), \ \deg(x) = s$$

(string of isomorphisms of 1-dim. vector spaces), a copy of $k[t]/(t^r)[s]$.

$x \in H_n(V_s(X)$ does not lift to $H_n(V_{s-1}(X))$,
$t^r x = 0 \in H_n(V_{s+r}(X))$.

This is a **persistent homology class**, which is **born** in $H_n(V_s(X))$ and **dies** in $H_n(V_{s+r}(X))$.

A persistent homology class is a cluster in vector spaces.

References


John Healy and Leland McInnes.
Accelerated heirarchical density clustering.
Preprint, arXiv: 1705.07321v2 [stat.ML], 2017.


J.F. Jardine.
Cluster graphs.
Preprint, `http://www.math.uwo.ca/faculty/jardine/preprints/preprints.html`, 2017.


Afra Zomorodian and Gunnar Carlsson.
Computing persistent homology.
Discrete Comput. Geom., 33(2):249–274, 2005.