Topological Data Analysis

2022–2023

Lecture 14

Mapper

22 December 2022

Mapper

G. Singh, F. Mémoli, G. Carlsson, *Topological methods for the analysis of high dimensional data sets and 3D object recognition*, Eurographics Symposium on Point-Based Graphics (2007)

Mapper is a combination of

- dimensionality reduction,
- clustering,
- and graph theory.

It is a powerful visualization method for Topological Data Analysis that does not rely on persistent homology.

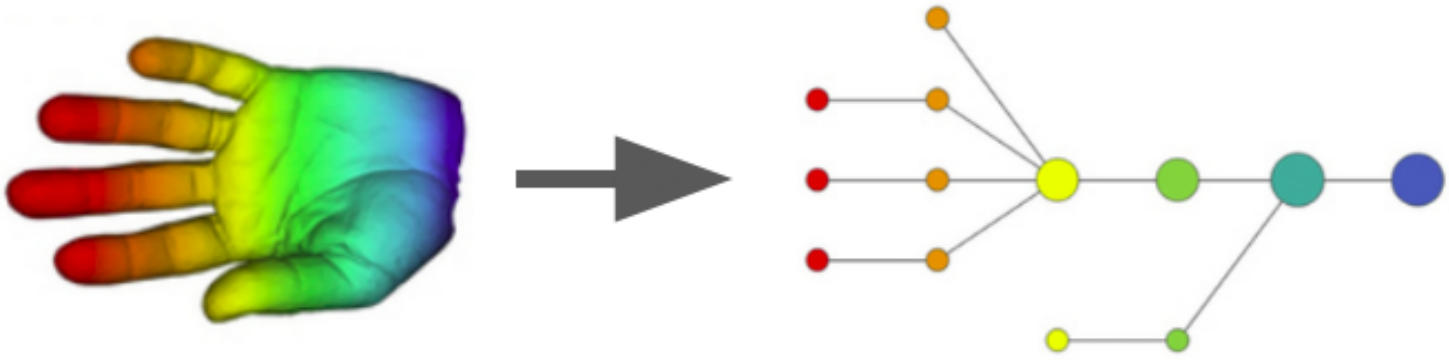
Mapper

Input:

- A data set X,
- a parameter space Z (often a subset of $\mathbb{R}$),
- a function $f : X \to Z$, called a **filter function,**
- and a clustering algorithm, e.g., single linkage.

Output:

- A coloured graph.



Mapper

Let $\mathcal{U} = \{U_i\}_{i \in I}$ be a collection of open sets in a topological space. The **nerve** of $\mathcal{U}$ is the simplicial complex $\boldsymbol{N(\mathcal{U})}$ whose vertex set is I and where $\{i_0, \ldots, i_k\}$ is a k-simplex if and only if

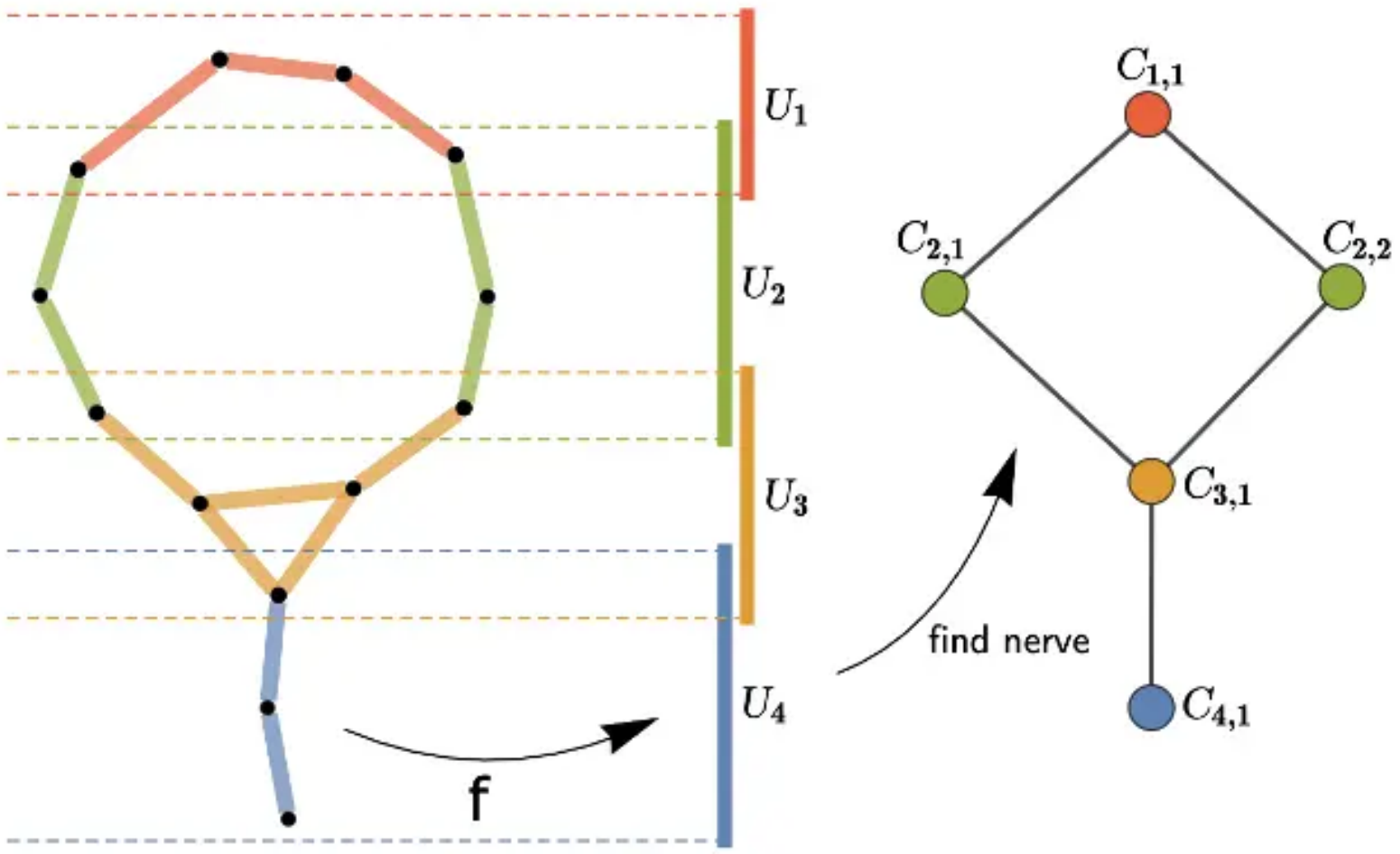
$$U_{i_0} \cap \cdots \cap U_{i_k} \neq \emptyset.$$

Example: If $X = \{x_i\}_{i \in I}$ is a point cloud in $\mathbb{R}^n$, then $\mathcal{B} = \{B_{t/2}(x_i)\}_{i \in I}$ is an open cover of X for each $t > 0$, and the nerve $N(\mathcal{B})$ is the (open ball) Čech complex $C_t(X)$.

Given a continuous map $f : X \to Z$ between topological spaces,

- if $\mathcal{U} = \{U_i\}_{i \in I}$ is an open cover of the image $f(X)$,
- then $f^*\mathcal{U} = \{C_\alpha(f^{-1}(U_i))\}_{i,\alpha}$ is an open cover of X, where $i \in I$ and C_α denotes connected components.

Mapper

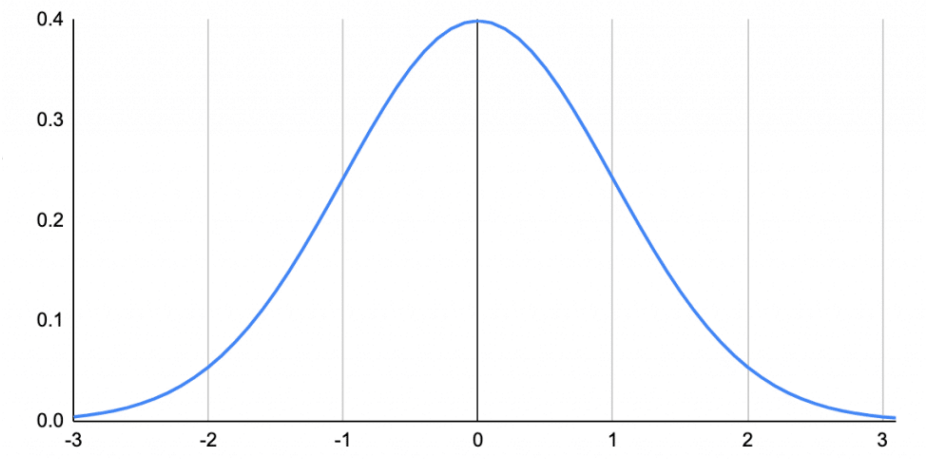


Mapper

Example: Take $X = \mathbb{R}$ and $f : X \to \mathbb{R}$ the density function for a standard Gaussian distribution. Choose the open cover

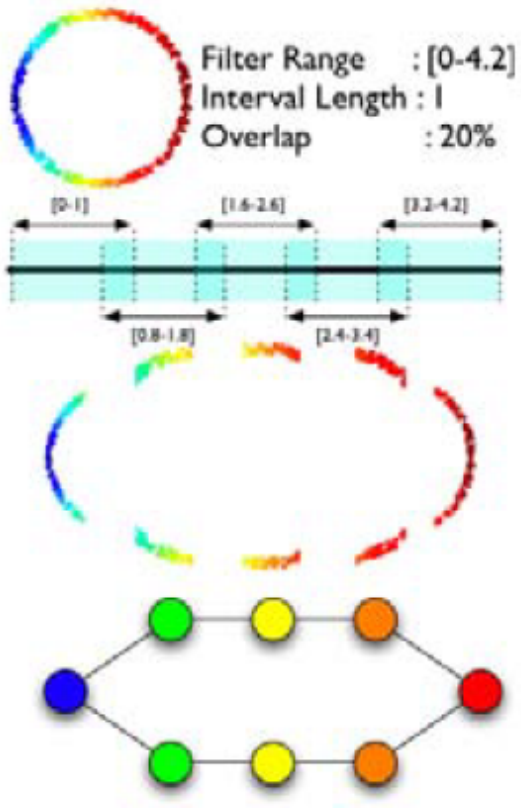
$$U_1 = [0, 0.15), \; U_2 = (0.1, 0.25), \; U_3 = (0.2, 0.35), \; U_4 = (0.3, 0.5).$$

Then the nerve of the cover $f^*\mathcal{U}$ is as follows:

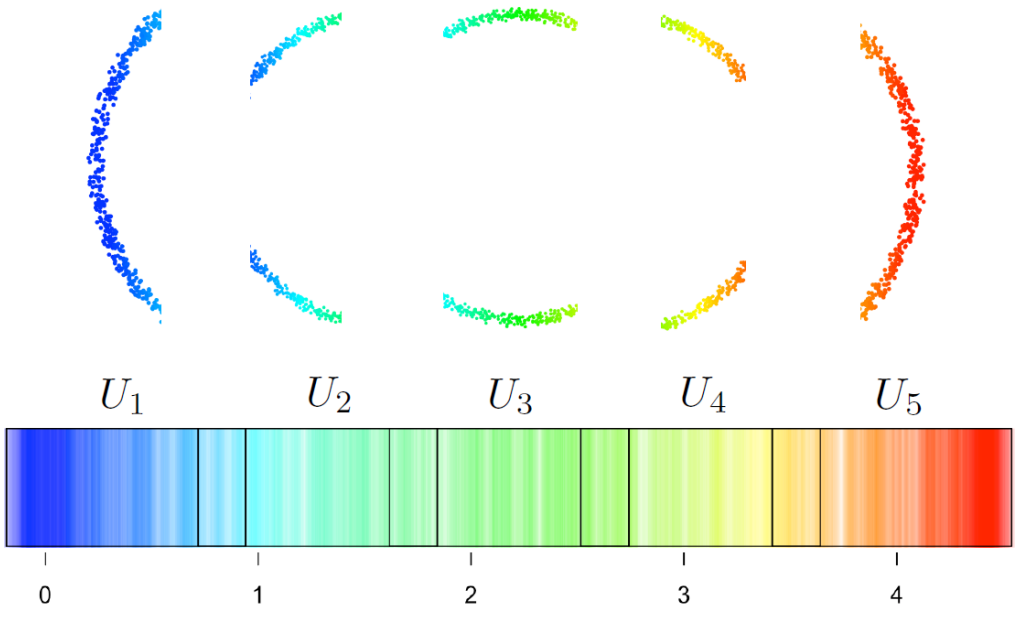


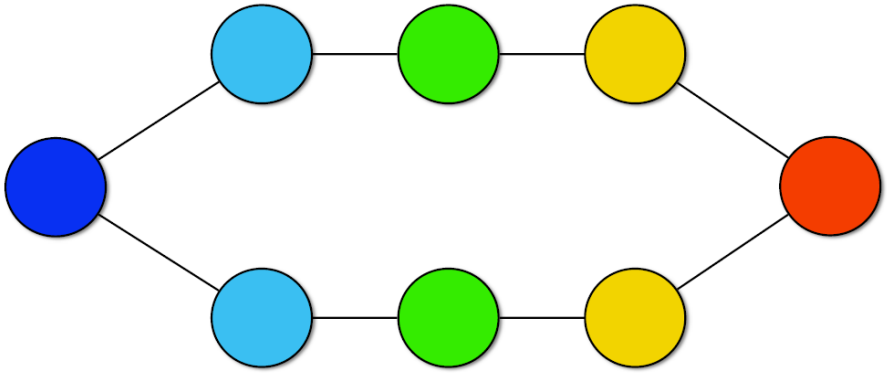
Mapper

Example: Let X be a random sample of points in a circle, and choose $f(x) = \|x - p\|$ where p is the left-most point in X. Cover the range of f with 5 intervals. The resulting graph is:



Mapper





Mapper

Example: The Miller–Reaven diabetes study (1985)

Six variables were measured in a sample of 145 patients, yielding a 6-dimensional data set. In the original study, a 3-dimensional image of the data was obtained:

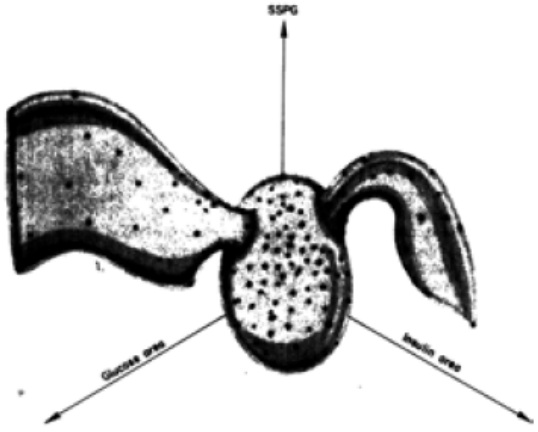


Figure 4: *Refer to Section 5.1 for details. Three dimensional projection of the diabetes data obtained using projection and pursuit.*

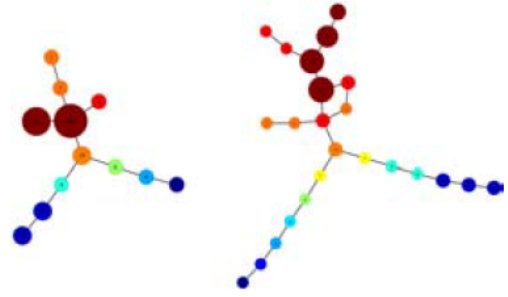
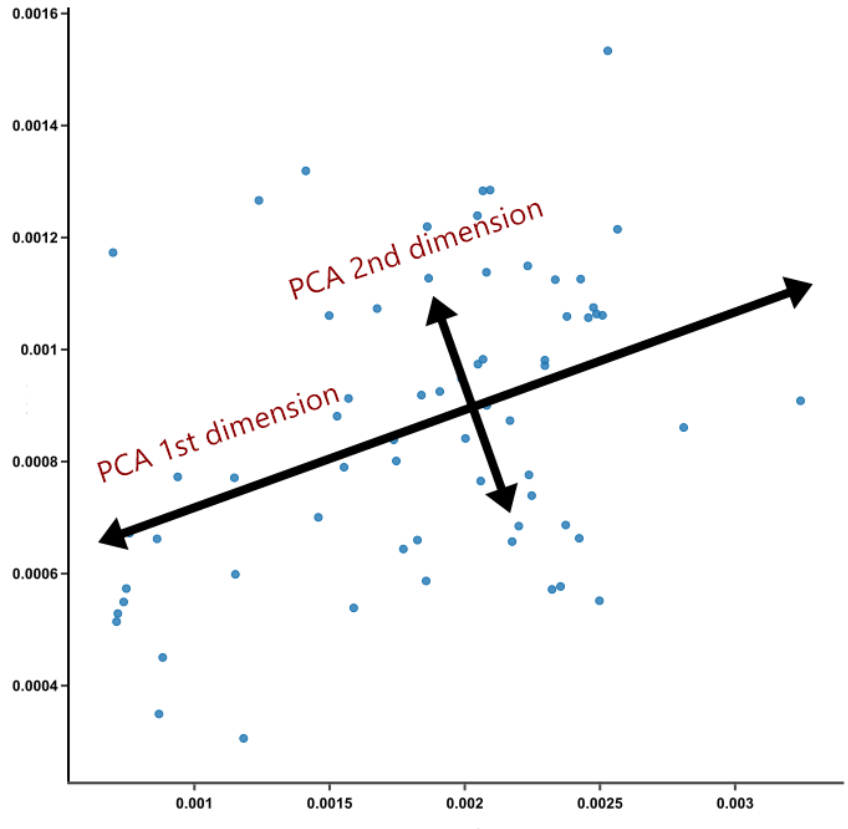


Figure 5: *Refer to Section 5.1 for details. On the left is a "low-resolution"* `Mapper` *output which was computed using 3 intervals in the range of the filter with a 50% overlap. On the right is a "high-resolution"* `Mapper` *output computed using 4 intervals in the range of the filter with a 50% overlap. The colors encode the density values, with red indicative of high density, and blue of low. The size of the node and the number in it indicate the size of the cluster. The low density ends reflect type I and type II diabetes. The flares occurring in Figure 4 occur here as flares with blue ends.*

Mapper

Choice of a filter function: use dimensionality reduction methods such as **principal component analysis** (PCA).



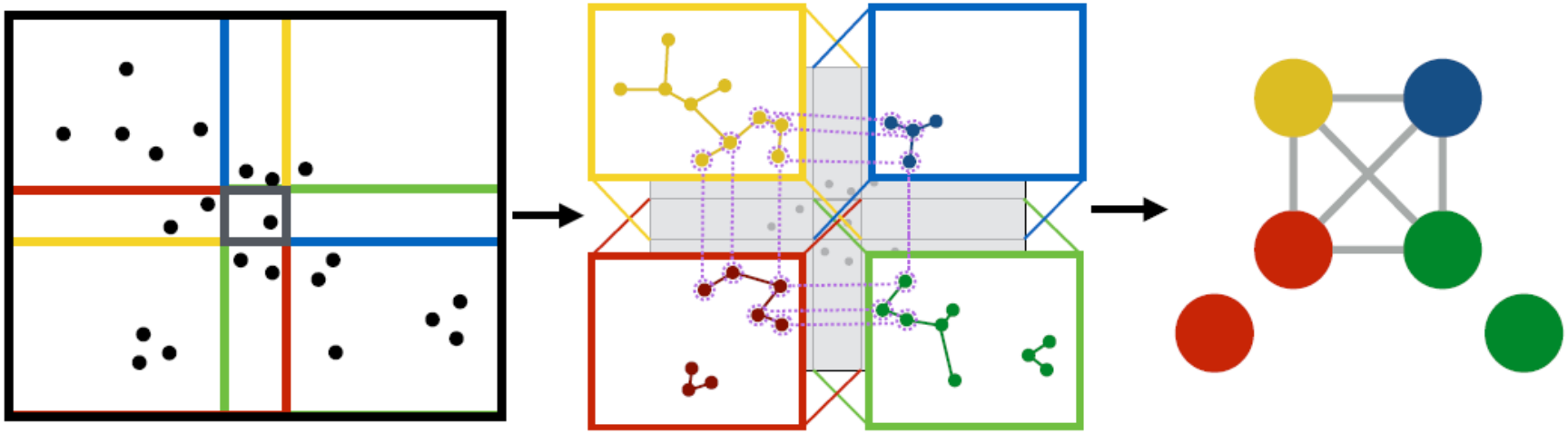
Mapper

An example where the parameter space *Z* is 2-dimensional:

$f : X \to Z$

$X_\alpha = f^{-1}U_\alpha$

$f^*(\mathcal{U}) \to M(\mathcal{U}, f)$



$\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$

$C(X_\alpha) \to f^*(\mathcal{U})$

(1-skeleton example)

Mapper

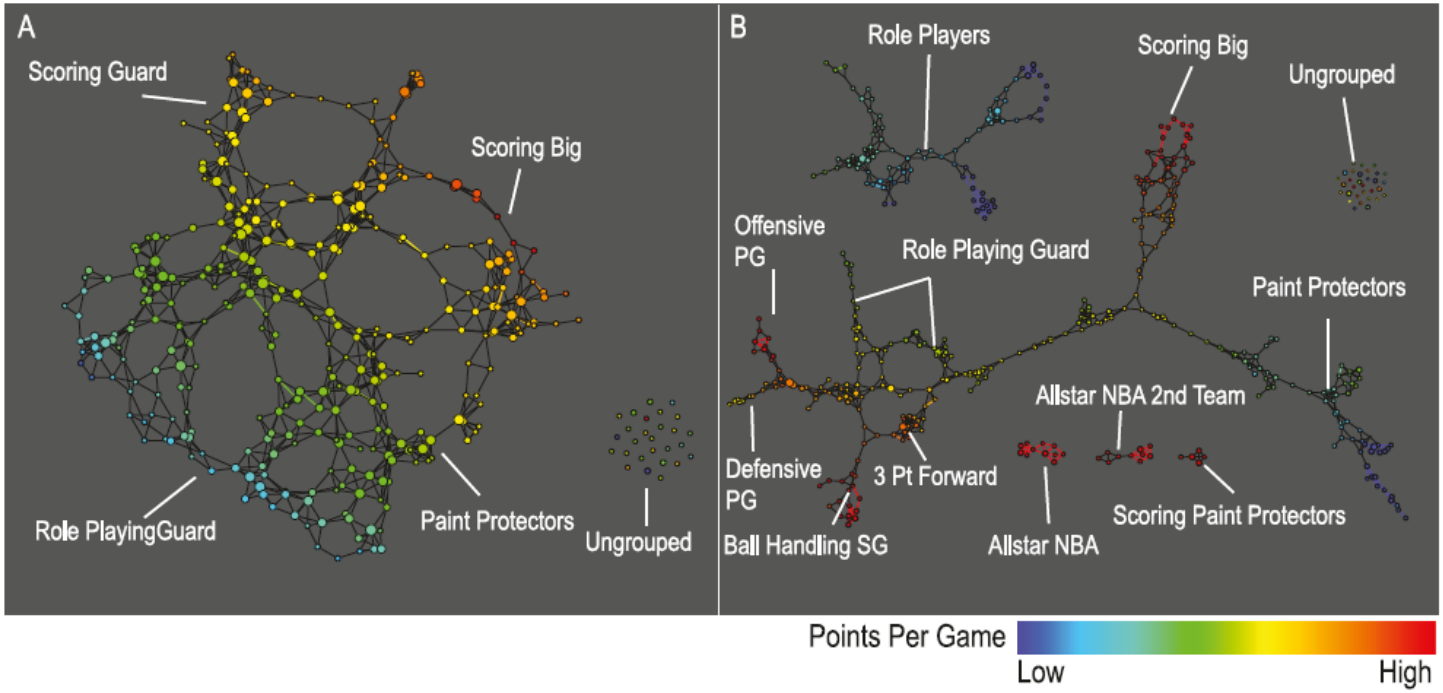
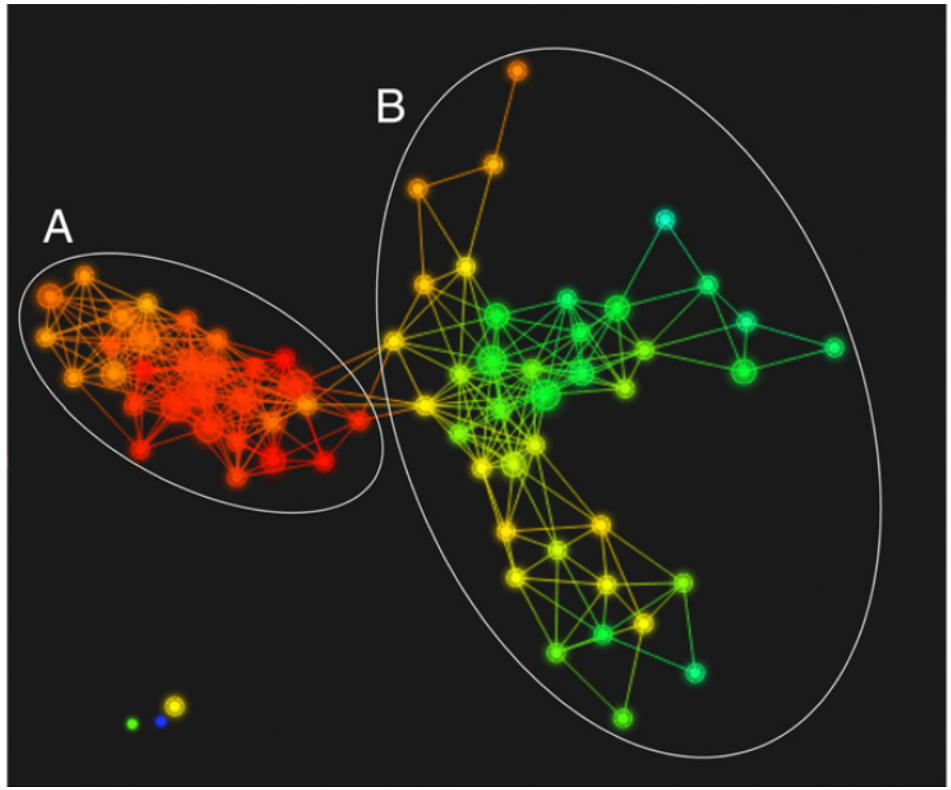


Figure 5 | The following map of players was constructed using the principal and secondary SVD filters at two different resolutions. A) Low resolution map at 20 intervals for each filter B) High resolution map at 30 intervals for each filter. The overlap is such at that each interval overlaps with half of the adjacent intervals, the graphs are colored by points per game, and a variance normalized Euclidean distance metric is applied. Metric: Variance Normalized Euclidean; Lens: Principal SVD Value (Resolution 20, Gain 2.0x, Equalized) and Secondary SVD Value (Resolution 20, Gain 2.0x, Equalized). Color: red: high values, blue: low values.

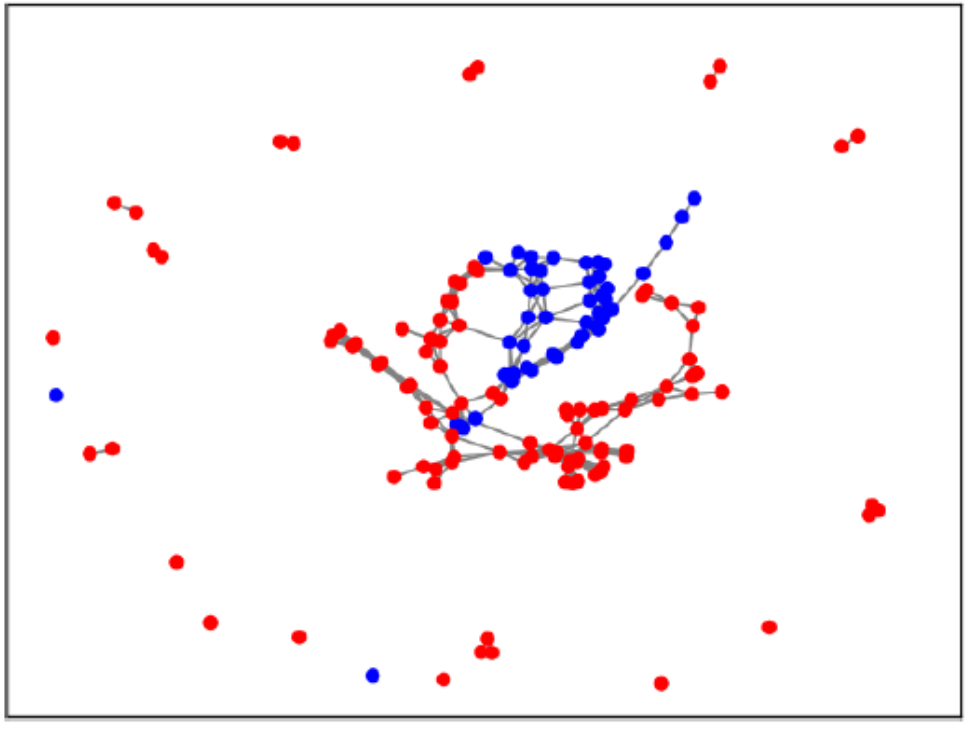
P. Y. Lum et al., *Extracting insights from the shape of complex data using topology*, Scientific Reports 3 (2013), 1236

Mapper



J. L. Bruno et al. (2017), *Longitudinal identification of clinically distinct neurophenotypes in young children with fragile X syndrome*, PNAS 114(40), 10767–10772

Mapper



	Connectivity	Scattering	Homogeneity
Red (malign)	0.94	1.86	0.97
Blue (benign)	0.83	0.49	0.90

Mapper

Kepler Mapper

https://kepler-mapper.scikit-tda.org

Python Mapper

http://danifold.net/mapper/

TDAview (Mapper online)

https://voineagulab.github.io/TDAview/

H. J. van Veen et al. (2019)
Kepler Mapper: A flexible Python implementation of the Mapper algorithm, Journal of Open Source Software 4(42), 1315

K. Walsh, M. A. Voineagu, F. Vafaee, I. Voineagu (2020)
TDAview: an online visualization tool for topological data analysis, Bioinformatics 36, 4805–4809