CS 6170: Computational Topology, Spring 2019 Lecture 13

Topological Data Analysis for Data Scientists

Dr. Bei Wang

School of Computing
Scientific Computing and Imaging Institute (SCI)
University of Utah
www.sci.utah.edu/~beiwang
beiwang@sci.utah.edu

Feb 19, 2019

Mapper Algorithm

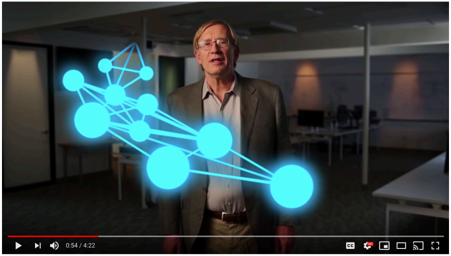
A Comprehensive Review

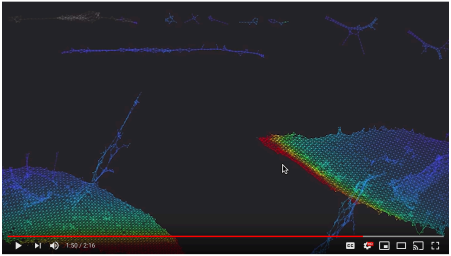
Singh et al. (2007); Lum et al. (2013)

History of mapper algorithm

- Singh et al. (2007)
- At the core of several data analysis startups
- Ayasdi: topological data analysis, machine learning and visualization `https://www.ayasdi.com/`
- Alpine Data: (topological) data analysis at scale, `http://alpinedata.com/`
- Quantopo, LLC. `https://www.kdnuggets.com/2018/01/topological-data-analysis.html`

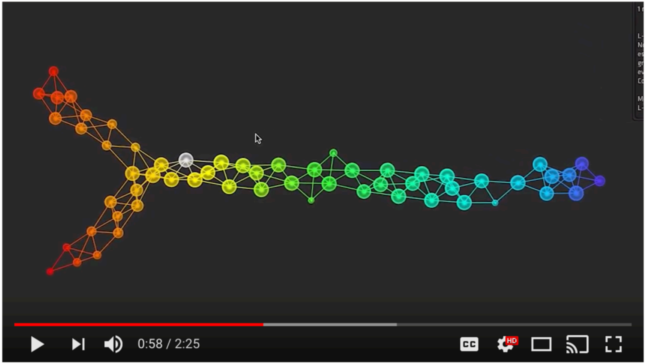
Ayasdi: TDA and Fraud Detection





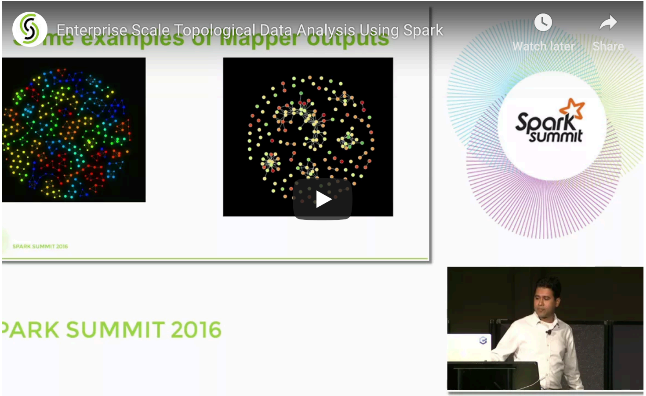
https://www.youtube.com/watch?v=XfWibrh6stw
https://www.youtube.com/watch?v=L8o4an5nh4E

Ayasdi: Patient Stratification



https://www.youtube.com/watch?v=FmfIJ3-UuaI

Alpine Data (Acquired by Tibco in November 2017)



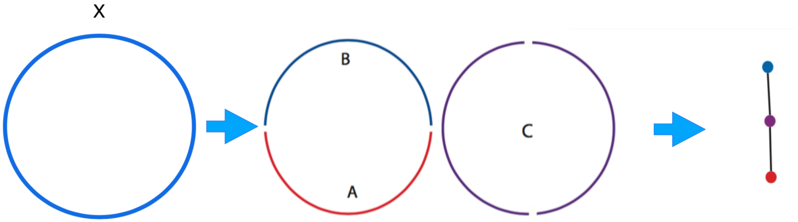
Enterprise Scale Topological Data Analysis Using Spark
https://databricks.com/session/enterprise-scale-topological-data-analysis-using-spark

Mapper algorithm and visualization of HD data

- Singh et al. (2007)
- Qualitative understanding of HD point cloud data through visualization
- Combining DR with graph visualization
- Desirable properties of visualization for HD data:
  - Insensitive to metric (approximation to similarity): robust to small changes to the metric
  - Understanding sensitivity to parameter changes: provide useful summary of behavior under all choices of parameters
  - Exploratory, multi-scale representations: at various levels of resolution, comparison.

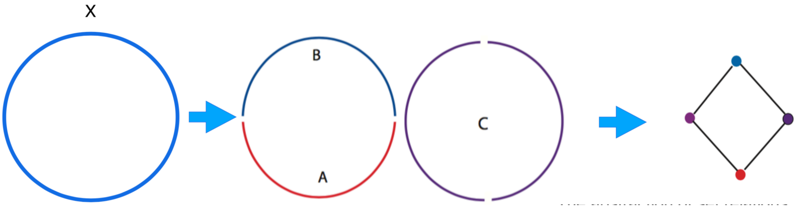
Mapper: Motivation and High-Level Intuitions

Covering a circle by sets



Carlsson (2009)

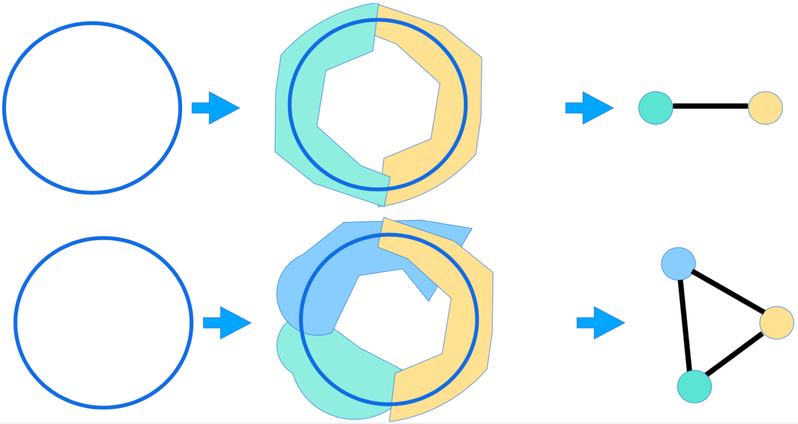
- X: a unit circle
- A covering $\mathcal{U}$ of X is given by the 3 sets $A = \{(x, y) \mid y < 0\}$, $B = \{(x, y) \mid y > 0\}$ and $C = \{(x, y) \mid y \neq \pm 1\}$.
- Obtain an abstraction of set relations based on overlaps of *sets*.



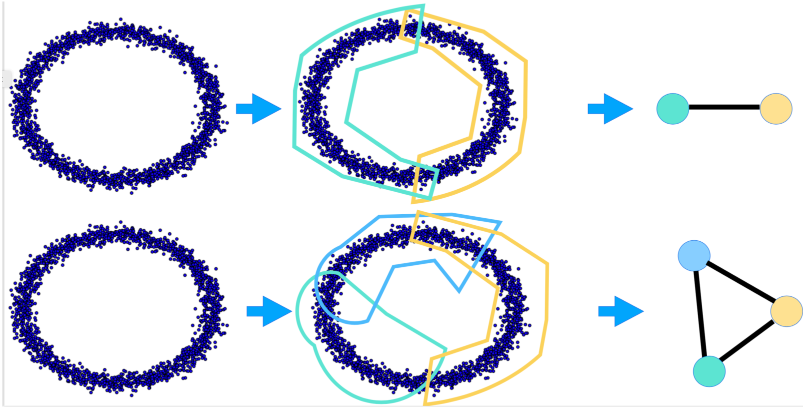
Carlsson (2009)

- Obtain an abstraction of set relations based on overlaps among the *connected components of sets*.
- Roughly speaking, this is the concept of the *nerve*.

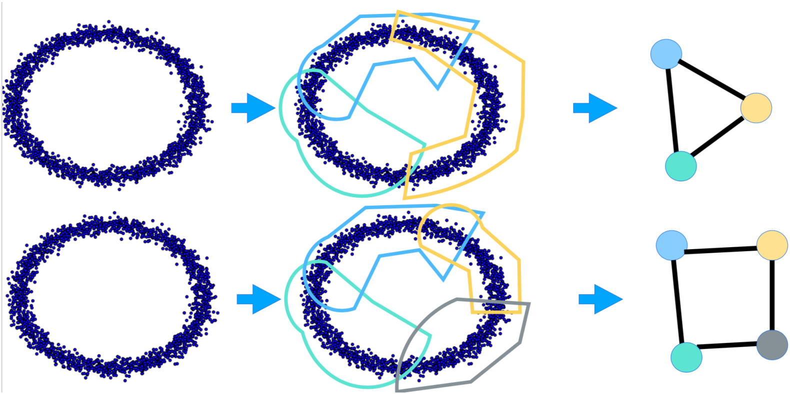
Covering a circle by sets: manifold setting



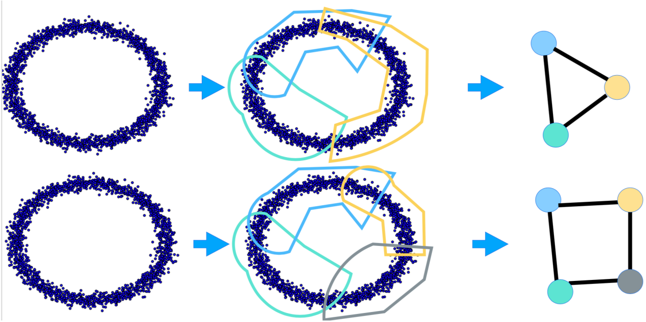
Covering a circle by sets: point cloud setting



Covering a circle by sets: change of scale



Covering of a point cloud at the right scale



- Given point cloud data and a covering...
- Taking the nerve of the covering can sometimes capture the shape of the data at the *right* scale(s)

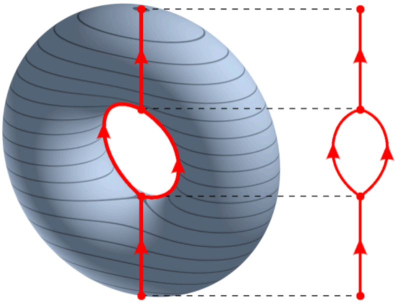
Mapper algorithm: advantages in HD data analysis

- Qualitative analysis, simplification and visualization of HD data sets and functions on these data sets.
- *Data summarization/skeletonization*: Extracting simple descriptions of HD data sets in the form of simplicial complexes or graphs
- *Function-induced clustering*: partial clustering of the data guided by a set of functions defined on the data.
- *Flexibility*: any clustering algorithm may be used with Mapper.
- *Exploratory and multi-scale*: explore parameters at all scales if possible.

Mapper: The Mathematical Formulation

Reeb Graph

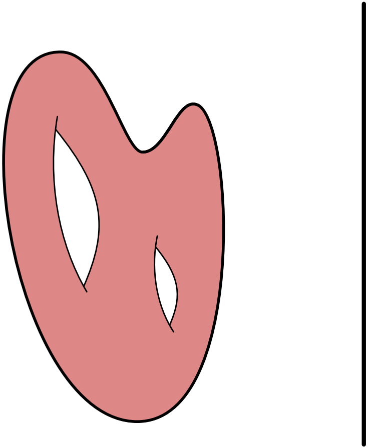
- Let $f : \mathbb{X} \to \mathbb{R}^d$ be a generic, continuous mapping
- Two points $x, y \in \mathbb{X}$ are *equivalent*, demoted by $x \sim y$, if $f(x) = f(y)$ and x and y belong to the same path-connected component of the pre-image of f, $f^{-1}(f(x)) = f^{-1}(f(y))$.
- The *Reeb graph*, $\mathcal{R}(X, f) = \mathbb{X}/\sim$, is the quotient space contained by identifying equivalent points together with the quotient topology inherited from $\mathbb{X}$.



https://en.wikipedia.org/wiki/Reeb_graph

Reeb Graph in Shape Analysis

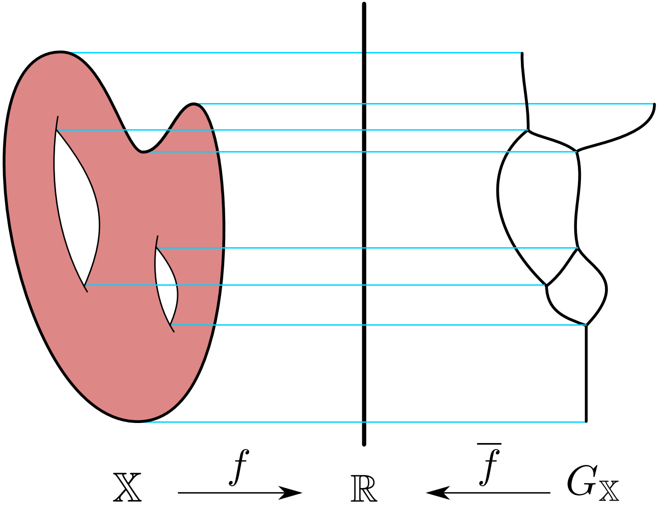
Reeb Graph



$$\mathbb{X} \xrightarrow{f} \mathbb{R}$$

- Input: $(\mathbb{X}, f)$
- $f : \mathbb{X} \to \mathbb{R}$

Reeb Graph



- Output: $(G_{\mathbb{X}}, \bar{f})$
- $G_{\mathbb{X}} := \mathcal{R}(\mathbb{X}, f)$, $\bar{f} : G_{\mathbb{X}} \to \mathbb{R}$

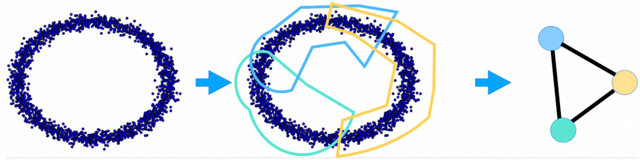
Cover and Nerve

- An *open cover* of a topological space $\mathbb{X}$ is a colletion $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ of open sets for some indexing set A such that

$$\bigcup_{\alpha \in A} U_\alpha = \mathbb{X}.$$

- Given a cover $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ of $\mathbb{X}$, let $\mathrm{Nrv}(\mathcal{U})$ demote the simplicial complex that corresponds to the *nerve* of the cover $\mathcal{U}$,

$$\mathrm{Nrv}(\mathcal{U}) = \{\sigma \in A \mid \cap_{\alpha \in \sigma} U_\alpha \neq \emptyset\}.$$



Mapper

- Given $f : \mathbb{X} \to \mathbb{R}^d$.
- Fix a cover $\mathcal{U} = \{U_\alpha\}$ of $f(\mathbb{X})$.
- The collection $f^{-1}(\mathcal{U}) = \{f^{-1}(U_\alpha)\}$ is a cover of $\mathbb{X}$.
- Let $f^*(\mathcal{U})$ be the cover which splits the sets of $f^{-1}(\mathcal{U})$ into path-connected components.
- Then Mapper is the nerve of this cover.

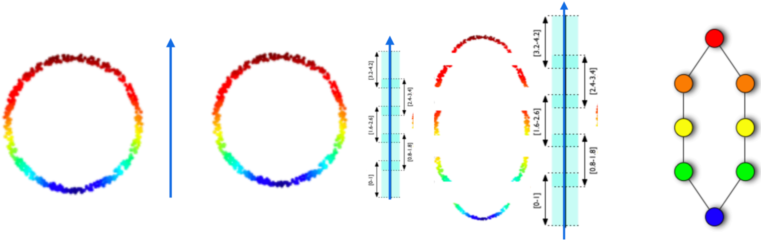
$$M(\mathcal{U}, f) := \mathrm{Nrv}(f^*(\mathcal{U})).$$

Mapper: The Main Algorithm and Variations

Input, output, implementation

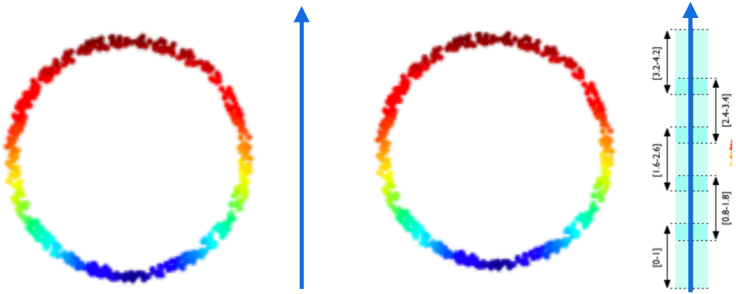
- Input:
  - Point cloud data X
  - Distance metric on the point cloud d_X
  - Functions on the point cloud: filter functions or lens $f : X \to \mathbb{R}$
- Output:
  - The mapper graph G_X: a summary of the data as a graph or a simplicial complex based on function-induced clustering,
  - Interface with (interactive) visualization, statistics and ML
- Parameters:
  - Parameters for the chosen clustering algorithms
  - Filter functions $f_1, f_2, \cdots$, etc.
  - Number of intervals m
  - Amount of interval overlap p
  - Color functions, etc.

Mapper algorithm by example



Singh et al. (2007)

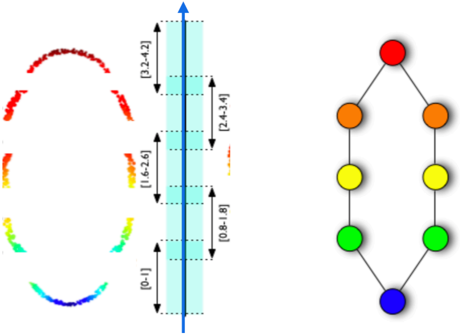
Mapper algorithm by example



Singh et al. (2007)

1. Input: a point cloud with a filter function, e.g., a height function. Assume there is a distance (metric) defined between any two points in the point cloud.
2. Cover the range of the function with intervals: using # of intervals, and amount % of overlap as parameters. E.g., # of intervals $= 5$, overlap $= 25\%$.

Mapper algorithm by example



Singh et al. (2007)

3. Look at the points in the domain that falls into each interval, and apply clustering (e.g., DBSCAN) to these points. E.g., following the inverse map.
4. Obtaining the nerve of all clusters (a covering) in the domain. E.g., here it is a graph representation that summarizes the data. Such a graph can interface with ML and interactive visualization.

Clustering inside the mapper algorithm

- Almost any clustering algorithm can be used
- Assume there is a notion of distance (metric) between a pair of points in the data domain (distance can be computed or provided)
- Clustering is equivalent to a notion of connected component in the point cloud setting
- Commonly used clustering algorithms:
  - Density-based spatial clustering of applications with noise (DBSCAN)
  - Single-linkage clustering
  - K-means, etc.
- Desirable properties:
  - Not restricted to Euclidean distance; can take distance matrix input
  - Do not require specifying the number of clusters beforehand

Parameters for the covering

- Number of intervals: m
    - Increasing m will increase the $\#$ of clusters we observe
    - May create more empty clusters (small number of points per cluster)
    - May capture finer features of the data
    - If density varies, pick up clusters with high density
- Percentage of overlap: p
    - Increasing p will increase the connectivities among the clusters
    - Sometimes robust in dealing with noise

Filter functions

- A filter function can be given a prior, e.g. car purchasing price
- It can also be derived from the properties of the point cloud itself
  - Density estimation

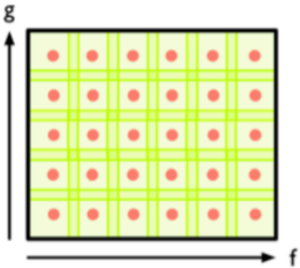
$$f_\epsilon(x) = C_\epsilon \sum_{y \in X} \exp \frac{-d(x,y)^2}{\epsilon}$$

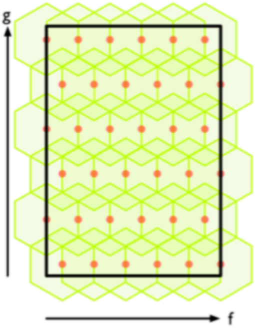
  - Eccentricity

$$E_p(x) = \left(\frac{\sum_{y \in X} d(x,y)^p}{N} \right)^{1/p}$$

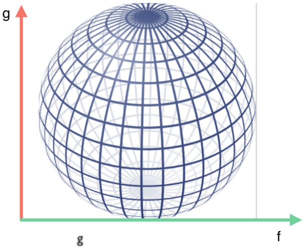
  - Distance to a point in the data
  - Graph laplacians

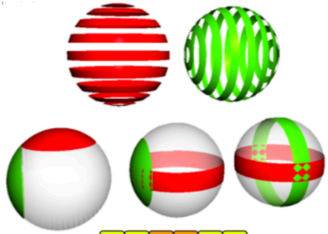
1D vs 2D Mapper

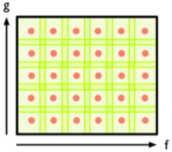




- 1 vs. 2 filter function(s)
- 1D intervals vs. 2D intervals.
- The covering of the domain of the function is no longer by intervals Instead, by rectangles or other geometric shapes, etc.





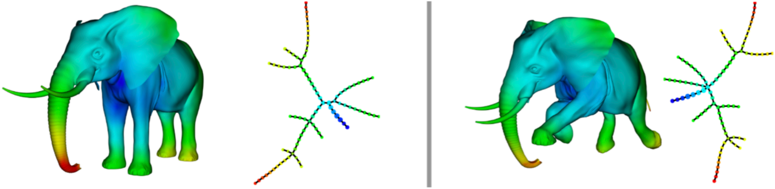


0	0	1	1	0	0
0	1	1	1	1	0
1	1	2	2	1	1
1	1	2	2	1	1
0	1	1	1	1	0
0	0	1	1	0	0

- Lower right: Count the number of connected components per 2D interval (square in the range).

Mapper: Applications

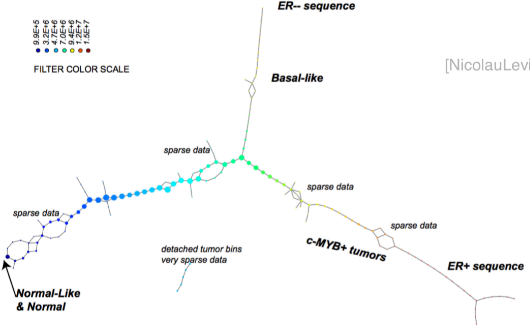
Shape skeletonization & classification



Singh et al. (2007)

- See Kepler Mapper demo examples: cat, lion, horse...

Breast cancer dataset A

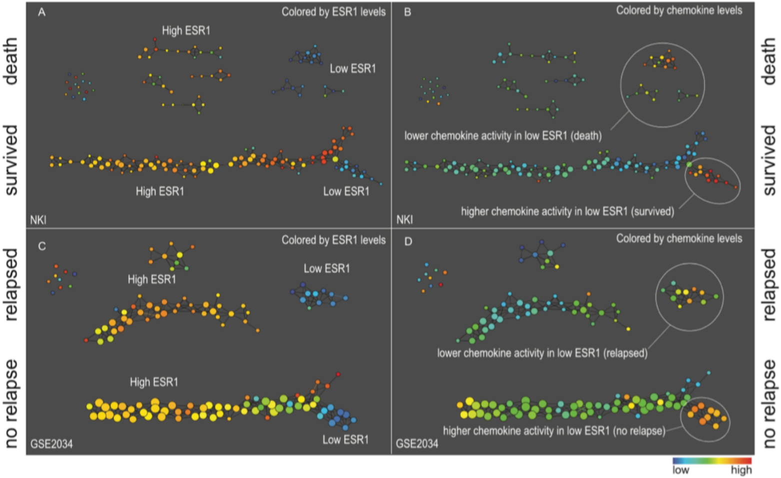


[NicolauLevineCarlsson2011]

Fig. 3. *PAD* analysis of the *NKI* data. The output has three progression arms, because tumors (data points) are ordered by the magnitude of deviation from normal (the *HSM*). Each bin is colored by the mean of the filter map on the points. Blue bins contain tumors whose total deviation from *HSM* is small (normal and *Normal-like* tumors). Red bins contain tumors whose deviation from *HSM* is large. The image of *f* was subdivided into 15 intervals with 80% overlap. All bins are seen (outliers included). Regions of sparse data show branching. Several bins are disconnected from the main graph. The ER^- arm consists mostly of *Basal* tumors. The $c\text{-}MYB^+$ group was chosen within the *ER* arm as the tightest subset, between the two sparse regions.

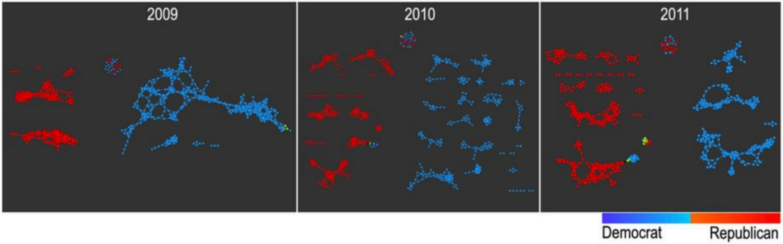
Nicolau et al. (2011)

Breast cancer dataset B



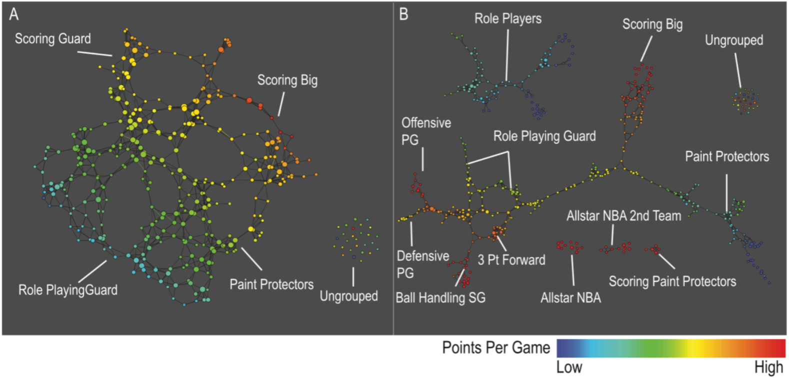
Lum et al. (2013)

Politics



Lum et al. (2013)

Sports



Lum et al. (2013)

Future directions of mapper

Current limitations

- How to choose the stable range of parameters (m, p)
- How to choose the clustering algorithms
- How to choose the filter functions
- Obtain insights with the right color function, etc.

Current and future directions

- Multi-scale mapper, nerves, etc. Dey et al. (2016, 2017)
- Better automatic parameter tuning
- Theoretical understanding of 2D and HD mapper

Implementations of Mapper Algorithms

Open Source Implementations

- Python Mapper: `http://danifold.net/mapper/index.html`
- R implementation: TDAmapper
  `https://cran.r-project.org/web/packages/TDAmapper/index.html`
- Spark Mapper: `https://github.com/log0ymxm/spark-mapper`
- *Kepler-Mapper*: `https://github.com/MLWave/kepler-mapper`

Kepler Mapper Demo (examples folder)

- Circles
- Digits
- Horse
- Breast Cancer

The 3-Torus

n-dimensional torus

- Denote by T^n the n-dimensional torus, which is the topological space:

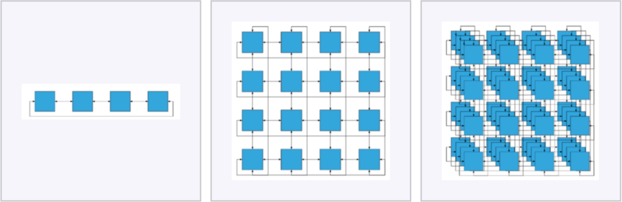
  $$(S^1)^n \cong S^1 \times S^1 \times \dots S^1$$

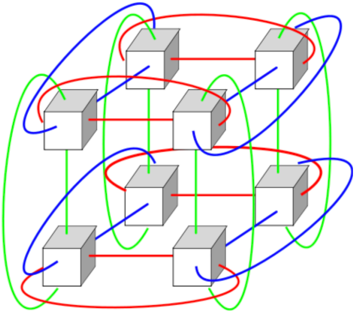
  i.e., the product of S^1, the circle, with itself n times.
- It is equipped with the product topology.
- Betti numbers for 3-dimensional torus (3-torus):

  $$(\beta_0, \beta_1, \beta_2, \beta_3) = (1, 3, 3, 1).$$

https://topospaces.subwiki.org/wiki/Homology_of_torus

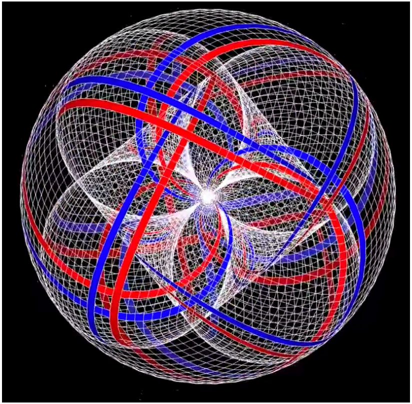
Torus interconnect





A torus interconnect is a switch-less network topology for connecting processing nodes in a parallel computer system.
https://en.wikipedia.org/wiki/Torus_interconnect

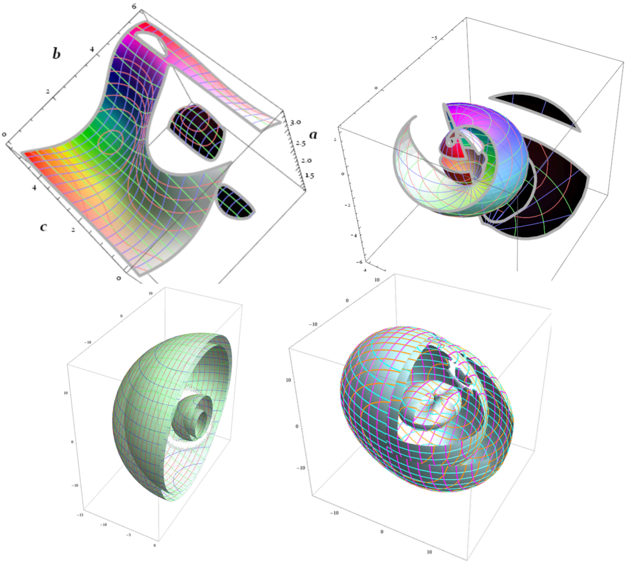
What is the mapper representation of a 3-Torus?



Which filter function to use?

https://www.youtube.com/watch?v=hTlKmGRW2pI

Hint: cross sections of a 3-Torus?



https://mathematica.stackexchange.com/questions/23546/how-can-i-draw-a-3d-cross-section-of-a-3-torus-embedded-in-4d-euclidean-space

References I

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Dey, T. K., Memoli, F., and Wang, Y. (2016). Mutiscale mapper: A framework for topological summarization of data and maps. *Proceedings 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 997–1013.

Dey, T. K., Mémoli, F., and Wang, Y. (2017). Topological analysis of nerves, reeb spaces, mappers, and multiscale mappers. *International Symposium on Computational Geometry (SOCG)*.

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3.

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.

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