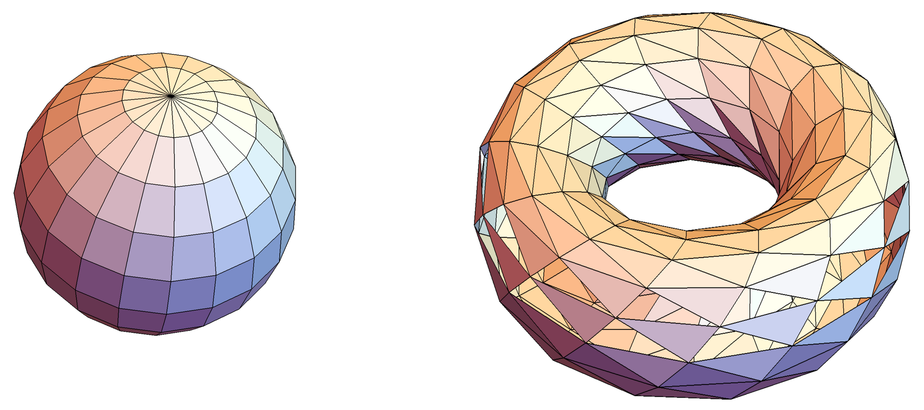
Topological Data Analysis of Attributed Networks using Diffusion Frèchet Functions with Ego Networks

Warren Keil, Mehmet Aktas
Department of Mathematics and Statistics, University of Central Oklahoma

Introduction

- Topological data analysis (TDA) has been a very active area of research in the past couple of decades. Its approach to data analysis is very different compared to most other methods.
- Network analysis on attributed networks, such as social media networks, has also been a very popular and exciting research topic in recent years.
- This research project aims to try to use the tools of TDA, along with recent results in multi-scale modeling and social network analysis to find new ways to study attributed networks.

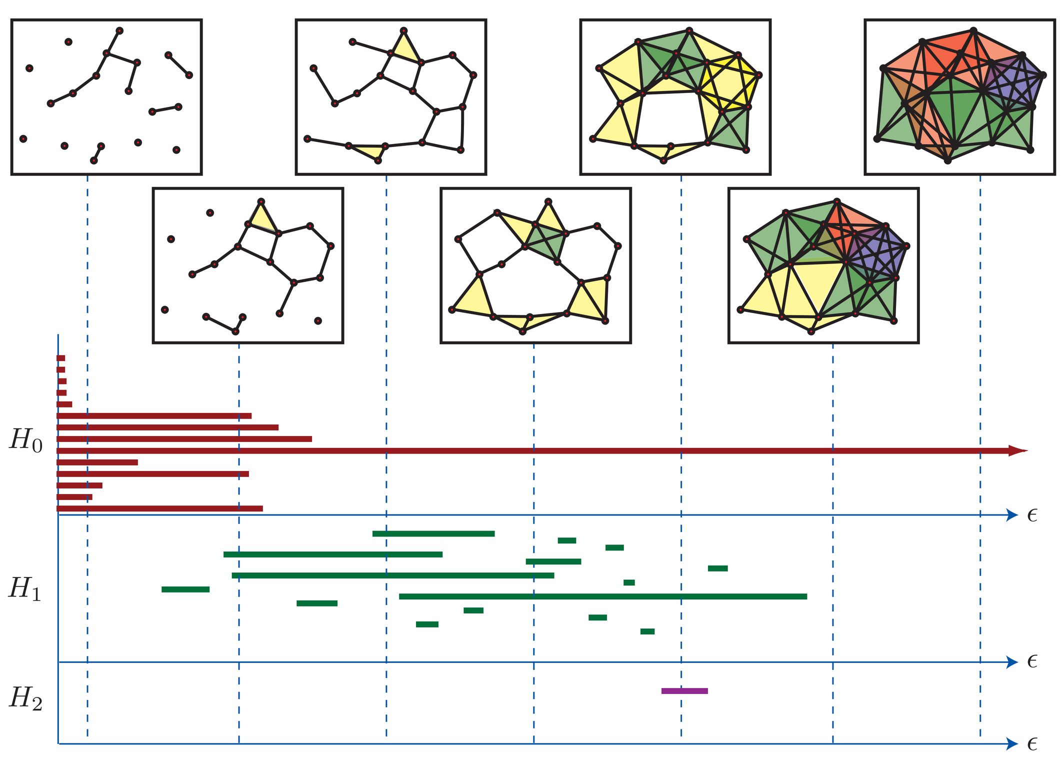


Objectives

1. Apply techniques from topological data analysis to attributed network data
2. Extend recent studies of Ego networks on social networks to Amazon product data
3. Utilize the diffusion Frèchet function to map the network data to a metric space
4. Employ clustering techniques to analyze results

Topological Data Analysis

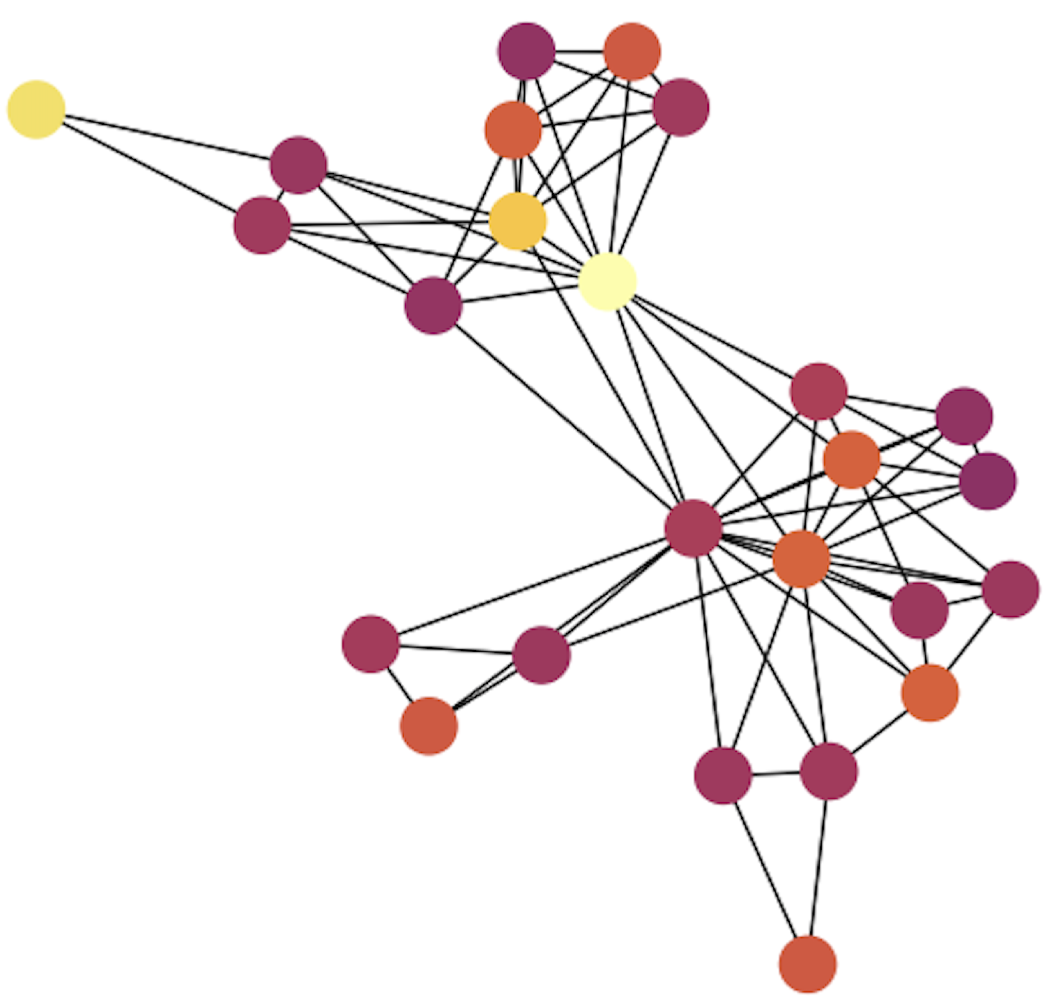
- Starts with the assumption that shape has meaning
- Topology is concerned with properties of shapes that are invariant under continuous deformations
- A common method of topological data analysis is persistent homology
- Map the data points to some metric space.
- Embed a simplicial complex over these points for all values of a parameter t



Source: Ghrist, R. Bulletin of the American Mathematical Society 45.1 (2008): 61-75.

Ego Networks

- Type of network analysis that looks at subgraphs
- Very useful in analysis of attributed networks
- Can assist in detecting substructures in networks
- Provides a mean of using the attributes of the node to assign weights to the edges



Diffusion Frèchet Functions

- Multi-scale function useful in a variety of analysis
- Takes classical Frèchet function and replaces Euclidean distance with diffusion distance from the heat (diffusion) equation
- Able to detect multi-modal data and other patterns that traditional functions sometimes miss
- Ideal in preserving the shape of the data
- Proven stable with respect to the Wasserstein distance

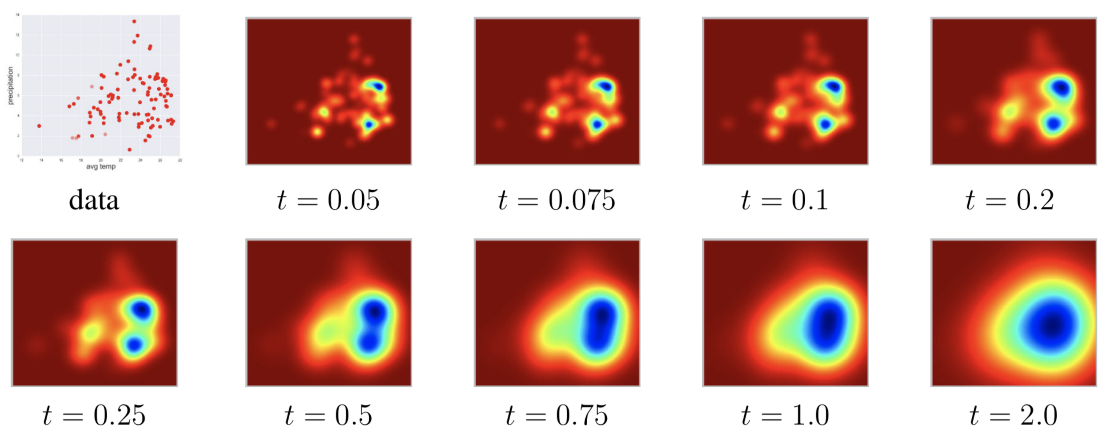
Diffusion Frèchet Function on Euclidean Space [2,3]

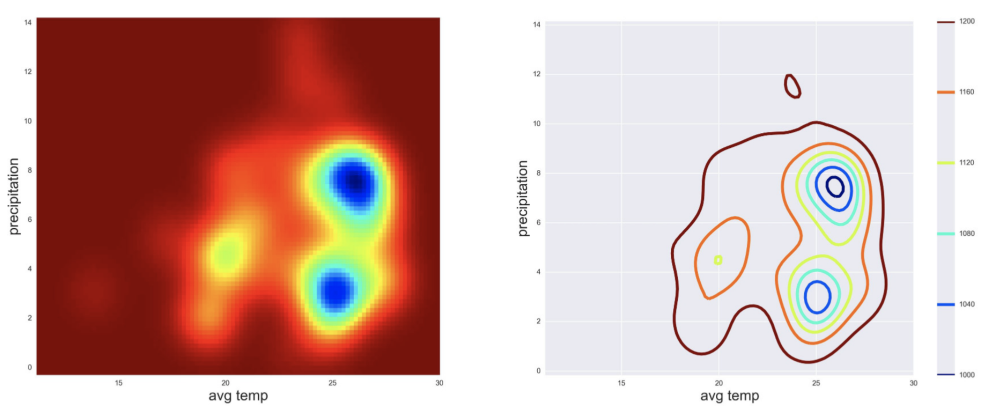
- Let $k_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the fundamental solution of the Heat equation.
- $k_t(x, y)$ can be interpreted as the temperature at time t at point y when the heat source was at point x at time 0.
- If $k_t(x, y)$ is large, it means that heat diffuses fast from x to y.
- Associate each point x in $\mathbb{R}^d$ with the function $k_t(x, \cdot) = k_{t,x}$ in the space of square-integrable functions $L^2(\mathbb{R}^d)$
- If heat diffuses in a similar way from points $x, y \in \mathbb{R}^d$ to any other point $z \in \mathbb{R}^d$, the functions $k_{t,x}$ and $k_{t,y}$ will be close in $L^2(\mathbb{R}^d)$.
- The diffusion distance $dt : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, for $t > 0$, is

$$d_t(x, y) := \|k_{t,x} - k_{t,y}\|_2.$$

- The diffusion Frèchet function, is

$$V_{\alpha,t}(x) := \int_{\mathbb{R}^d} d_t^2(x, y)\alpha(dy)$$





Diffusion Frèchet Function on weighted networks [2,3]

To define a heat diffusion process on networks, we must define an analog of the Laplacian.

- Let $v_1, \ldots, v_n$ be the nodes of a weighted network K and W be the $n \times n$ weighted adjacency matrix.
- The graph Laplacian is the matrix Δ defined by $\Delta = D - W$ where D is the diagonal matrix with diagonal entries $d_{ii} = \sum_{k=1}^{n} w_{ik}$.
- The heat kernel can be expressed as $k_t(i, j) = \sum_{k=1}^{n} e^{-\lambda_k t}\phi(i)\phi(j)$ where $0 \leq \lambda_1 \leq \ldots \leq \lambda_n$ are the eigenvalues of Δ with orthonormal eigenvectors $\phi_1, \ldots, \phi_n$.
- The diffusion distance between v_i and v_j is

$$d_t^2(i, j) = \sum_{k=1}^{n} e^{-2\lambda_k t}(\phi_k(i) - \phi_k(j))^2$$

- The diffusion Frèchet function for weighted networks is

$$F_{\xi,t}(i) = \sum_{j=1}^{n} d_t^2(i, j)\xi_j$$

Methodology

1. For each item in Amazon dataset, compute N-Ego network

2. Calculate diffusion Fréchet distance of each node of Ego networks

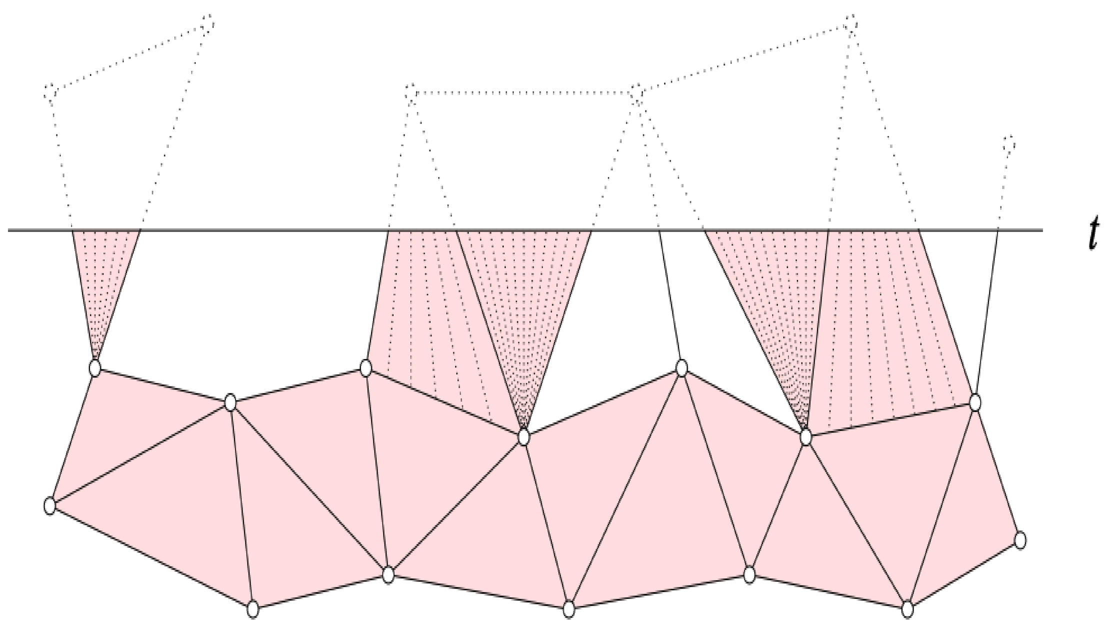
3. Compute lower star filtration using diffusion distance as height

4. Store persistent homology information in barcodes

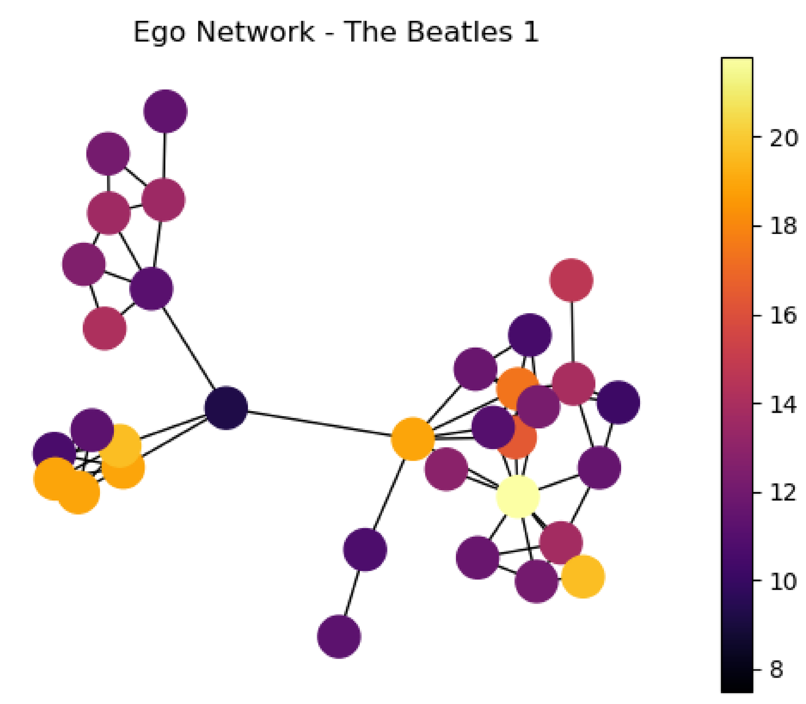
5. Compute Wasserstein distance between each barcode. Store in dissimilarity matrix

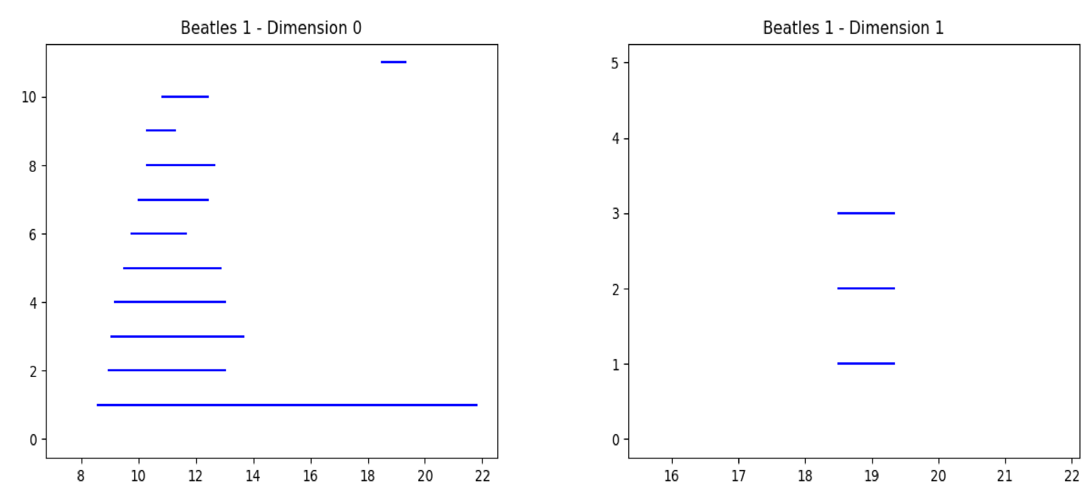
6. Perform hierarchical clustering using Wasserstein distances

Lower Star Filtration

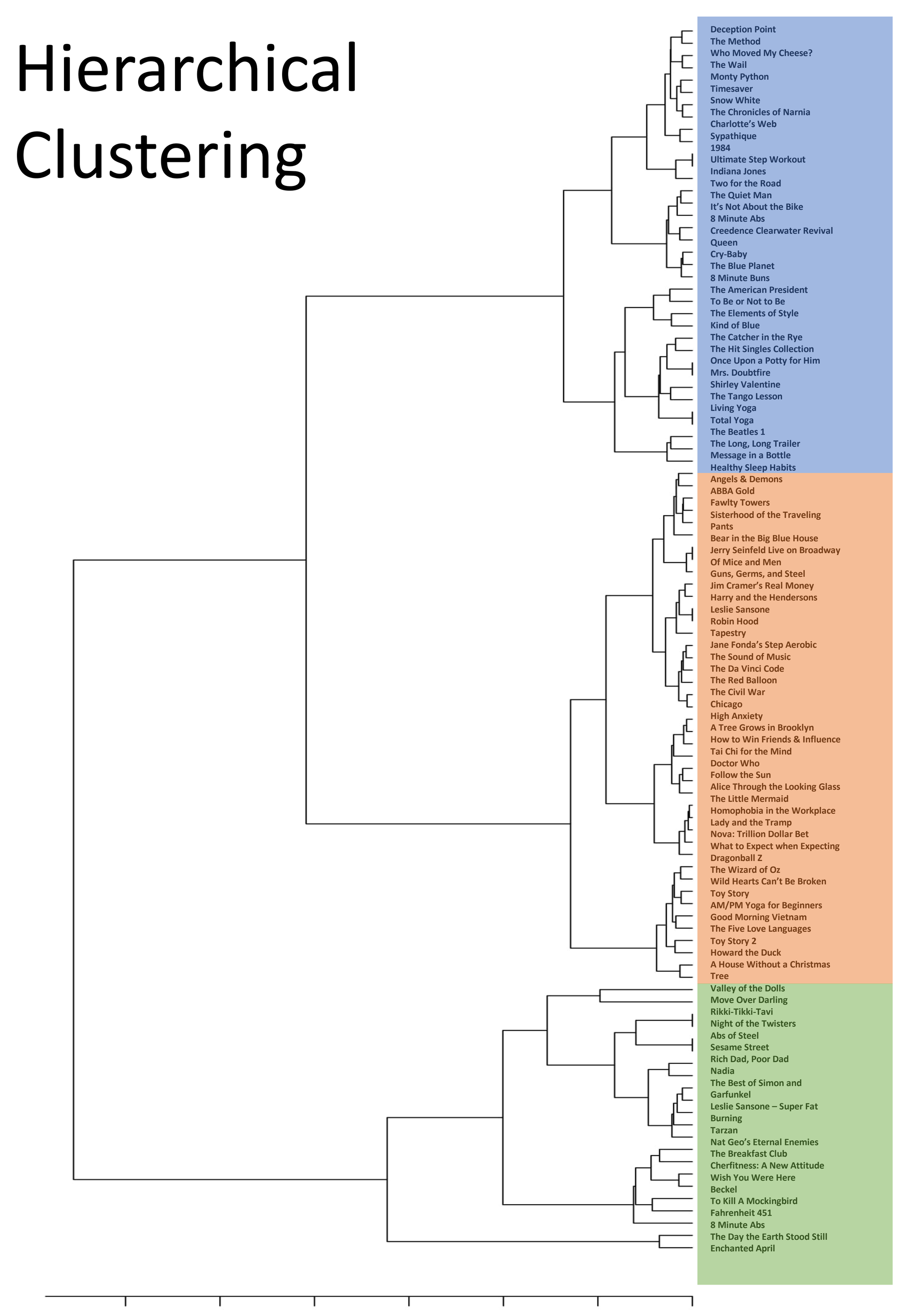


Example: Beatles 1 Album





Clustering



References

[1] Leskovec, Jure, et al. *Learning to Discover Social Circles in Ego Networks.* Stanford Large Network Dataset Collection,snap.stanford.edu/data/. (2015).

[2] Martinez, Diego H. Diaz. *Multiscale Summaries of Probability Measure with Applications to Plant and Microbiome Data*, Dissertation, The Florida State University, (2016).

[3] Martinez, Diego H. Diaz. *Probing the Geometry of Data with Diffusion Frèchet Functions*, Applied and Computational Harmonic Analysis (2018)

[4] Carlsson, Gunnar *Topology and Data*, Bulletin of the American Mathematical Society 46.2255-308, (2009)

[5] Ghrist, Robert. *Barcodes: the persistent topology of data*, Bulletin of the American Mathematical Society 45.1 : 61-75 (2008).