Lecture: Nonlinear dimensionality reduction: Diffusion Maps

Foundations of Data Science: Algorithms and Mathematical Foundations

Mihai Cucuringu
mihai.cucuringu@stats.ox.ac.uk

CDT in Mathematics of Random System
University of Oxford

September 30, 2019

Outlook

- inferring meaningful spatial and structural information from incomplete data sets of pairwise interactions between nodes in a network
- the way people interact in many aspects of everyday life often reflect surprisingly well geopolitical boundaries
- inhomogeneity of connections in networks leads to natural divisions, and identifying such divisions can provide valuable insight into how interactions in a network are influenced by its topology
- finding network communities (groups of tightly connected nodes) has been extensively studied in recent years

Real-world network:

- a county-to-county migration network constructed from 1995-2000 US Census data

Diffusion maps

- diffusion maps were introduced in S. Lafon's Ph.D. Thesis in 2004 as a dimensionality reduction tool
- connected data analysis and clustering techniques based on eigenvectors of similarity matrices with the geometric structure of non-linear manifolds
- in recent years, diffusion maps have gained a lot of popularity
- often called Laplacian eigenmaps, these manifold learning techniques identify significant variables that live in a lower dimensional space, while preserving the local proximity between data points

Laplacian Eigenmaps

- consider a set of N points $V = \{x_1, x_2, \ldots, x_N\}$ in an n-dimensional space $\mathbb{R}^n$
- each point (typically) characterizes an image (or an audio stream, text string, etc.)
- if two images x_i and x_j are similar, then $||x_i - x_j||$ is small
- a popular measure of similarity between points in $\mathbb{R}^n$ is defined using the Gaussian kernel

$$w_{ij} = e^{-||x_i - x_j||^2/\epsilon}$$

  so that the closer x_i is from x_j, the larger w_{ij}
- the matrix $W = (w_{ij})_{1 \leq i,j \leq N}$ is symmetric and has positive coefficients
- to normalize W, we define the diagonal matrix D, with $D_{ii} = \sum_{j=1}^{N} w_{ij}$ and define L by

$$L = D^{-1} W$$

  such that every row of L sums to 1.

Laplacian Eigenmaps

- define the symmetric matrix $S = D^{-1/2} W D^{-1/2}$
- note S is similar to L, since one can write

$$S = D^{1/2} D^{-1} \; W D^{-1/2} = D^{1/2} L D^{-1/2}$$

- as a symmetric matrix, S has an orthogonal basis of eigenvectors $v_0, v_1, \ldots, v_{N-1}$, and N real eigenvalues $1 = \lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{N-1}$
- If we decompose S as

$$S = V \Lambda V^T$$

with

$$V V^T = V^T V = I$$

$$\Lambda = Diag(\lambda_0, \lambda_1, \ldots, \lambda_{N-1})$$

then L becomes

$$L = \Psi \Lambda \Phi^T$$

where $\Psi = D^{-1/2} V$ and $\Phi = D^{1/2} V$.

Laplacian Eigenmaps

- L is a row-stochastic matrix, $\lambda_0 = 1$ and $\psi_0 = (1, 1, \ldots, 1)^T$, we disregard this trivial pair
- interpret L as a random walk matrix on a weighted graph $G = (V, E, W)$, where the set of nodes consists of the points x_i, and there is an edge between nodes i and j if and ony if $w_{ij} > 0$
- L_{ij} denotes the transition probability from point x_i to x_j in one step time $\Delta t = \epsilon$

$$Pr\{x(t+\epsilon) = x_j | x(t) = x_i\} = L_{ij}.$$

$$w_{ij} = e^{-||x_i - x_j||^2/\epsilon}$$

- ϵ is the squared radius of the neighborhood used to infer local geometric and density information
- w_{ij} is $O(1)$ when x_i and x_j are in a ball of radius $\sqrt{\epsilon}$, but it is exponentially small for points that are more than $\sqrt{\epsilon}$ apart
- ϵ represents the discrete time step at which the random walk jumps from one point to another

Laplacian Eigenmaps

- Interpreting the eigenvectors as functions over our data set, the *diffusion map* (also called *Laplacian eigenmap)* maps points from the original space to the first k eigenvectors, $\mathcal{L} : V \mapsto \mathbb{R}^k$, is defined as

$$\mathcal{L}_t(x_j) = (\lambda_1^t \psi_1(j), \lambda_2^t \psi_2(j), \ldots, \lambda_k^t \psi_k(j)) \tag{1}$$

- using the left and right eigenvectors of L

$$L_{ij} = \sum_{r=0}^{N-1} \lambda_r \phi_r(i) \psi_r(j)$$

- note that $L_{ij}^t = \sum_{r=0}^{N-1} \lambda_r^t \phi_r(i) \psi_r(j)$
- the probability distribution of a random walk landing at location x_j after exactly t steps, starting at x_i

$$L_{ij}^t = Pr\{x(t) = x_j | x(0) = x_i\}$$

- given the random walk interpretation, quantify the similarity between two points according to the evolution of their probability distributions

$$D_t^2(i,j) = \sum_{k=1}^{N} (L_{ik}^t - L_{jk}^t)^2 \frac{1}{d_k},$$

where the weight $\frac{1}{d_k}$ takes into account the empirical local density of the points by giving larger weight to the vertices of lower degree. $D_t(i,j)$ is the *diffusion distance* at time t.

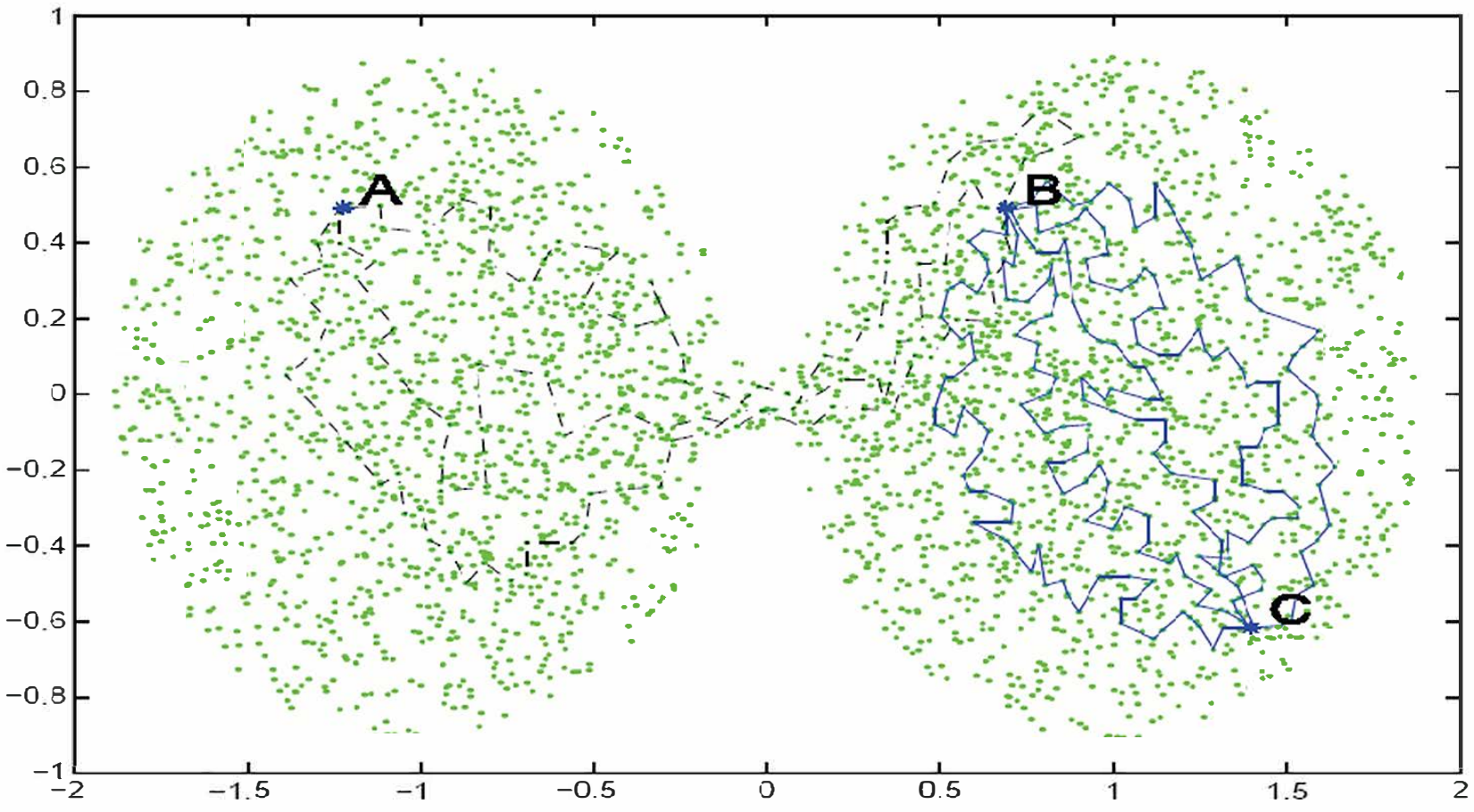
Diffusion Maps

- a matter of choice to tune the parameter t corresponding to the number of time steps of the random walk (used $t = 1$)
- using different values of t corresponds to rescaling the axis
- the Euclidean distance between two points in the diffusion map space introduced in (1) is given by

$$||\mathcal{L}(x_i) - \mathcal{L}(x_j)||^2 = \sum_{r=1}^{N-1} \left(\lambda_r^t \psi_r(i) - \lambda_r^t \psi_r(j)\right)^2 . \qquad (2)$$

- Nadler et al. (2005) have shown that the expression (2) equals the diffusion distance $D_t^2(i, j)$, when $k = N - 1$ (when using $N - 1$ eigenvectors)
- for ease of visualization, use the top $k = 2$ eigenvectors for the projections

Diffusion distance vs Euclidean distance



Limitations of the Euclidean distance

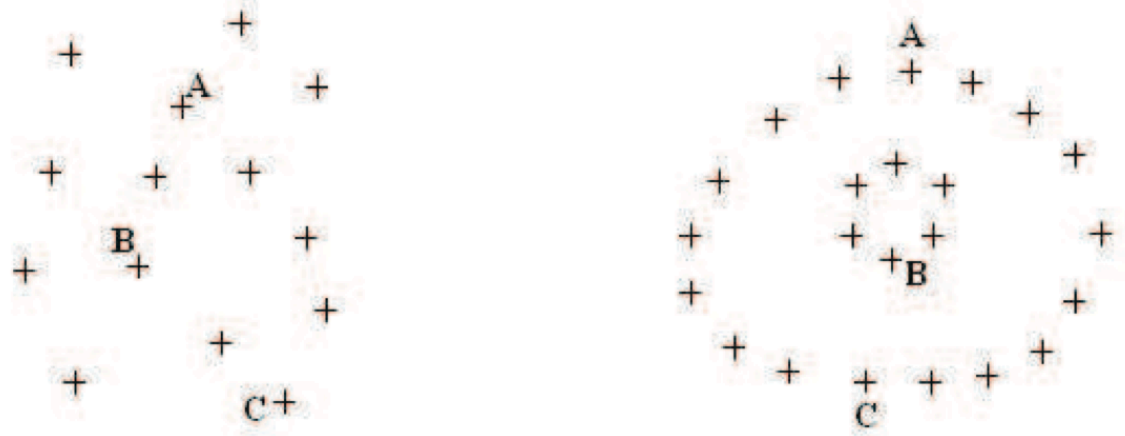


Figure: Euclidean distance may not be relevant to properly understand the distance (or similarity) between two points.

Is C is more similar to point B or to point A?

- (left) the natural answer is: C is more similar to B.
- (right) less obvious given the other observed data points... C should be more similar to A. Need a new metric for which C and A are closer than C and B given the geometry of the observed data.

Eigenvector colourings

- denote by $\mathcal{C}_k$ the colouring of the N data points given by the eigenvector ψ_k
- colour of point $x_i \in V$ is given by the i-th entry in ψ_k, i.e.

$$\mathcal{C}_k(x_i) = \psi_k(i), \text{ for all } k = 0, \ldots, N-1 \text{ and } i = 1, \ldots, N.$$

- $\mathcal{C}_k$: *eigenvector colouring* of order k
- do not confuse with the "graph colouring" terminology
- colourbar: red denotes high values and blue denotes low values, in the eigenvector entries
- in practice, only the first k eigenvectors are used in the diffusion map introduced in (1), with $k << N-1$ chosen such that $\lambda_1^t \geq \lambda_2^t \ldots \geq \lambda_k^t > \delta$ but $\lambda_{k+1}^t < \delta$, where δ is a chosen tolerance
- show how one can extract relevant information from eigenvectors of much lower order

Eigenvector localization

- The phenomenon of *eigenvector localization* occurs when most of the components of an eigenvector are zero or close to zero, and almost all the mass is localized on a relatively small subset of nodes.

- On the contrary, *delocalized eigenvectors* have most of their components small and of roughly the same magnitude.

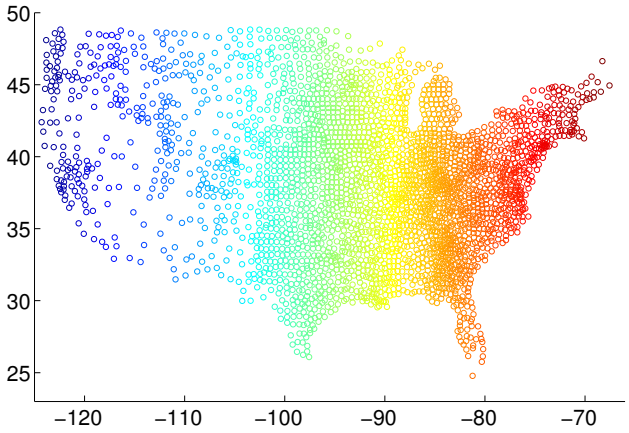
2000 US Census data set

- reports the number of people that migrated from every county to every other county within US during 1995-2000
- $M = (M_{ij})_{1 \le i,j \le N}$ the total number of people that migrated between county i and county j (so $M_{ij} = M_{ji}$)
- $N = 3107$ denotes the number of counties in mainland US
- let P_i denote the population of county i
- different similarity measures

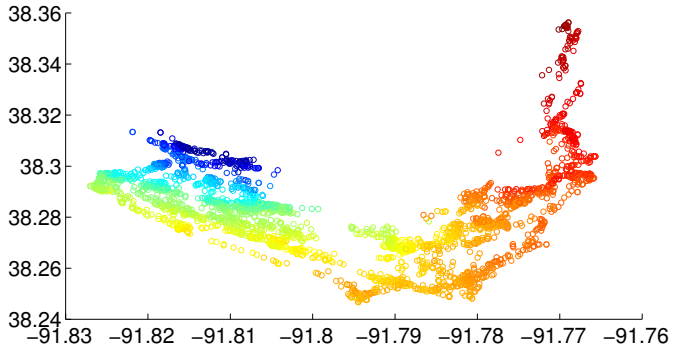
$$W_{ij}^{(1)} = \frac{M_{ij}^2}{P_i P_j}; \qquad W_{ij}^{(2)} = \frac{M_{ij}}{P_i + P_j}; \qquad W_{ij}^{(3)} = 5500 \frac{M_{ij}}{P_i P_j}$$

- Midwest gets placed closer to the west coast, but further from the east coast
- colourings based on latitude reveal the north-south separation
- $W^{(1)}$ does a better job at separating the east and west coasts, while $W^{(2)}$ highlights best the separation between north and south

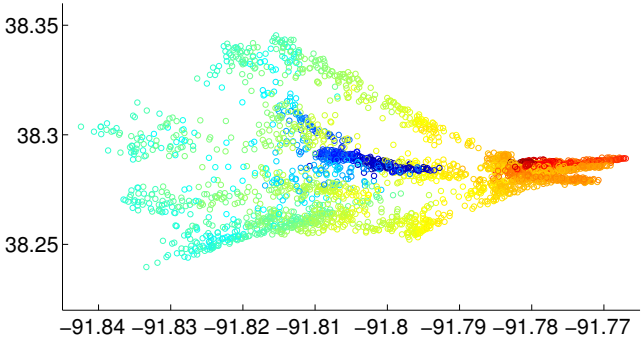
Colored by longitude



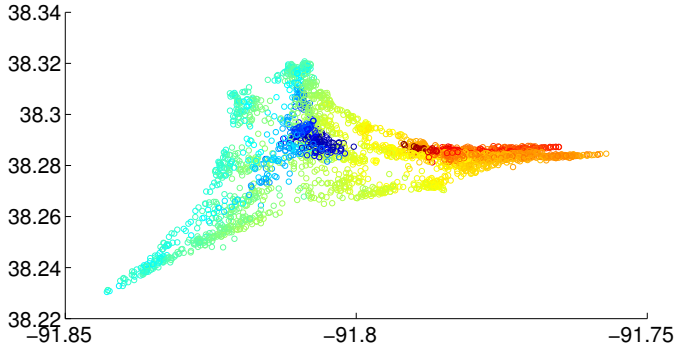
(a) Map of USA, coloured by longitude



(b) $W_{ij}^{(1)} = \frac{M_{ij}^2}{P_i P_j}$



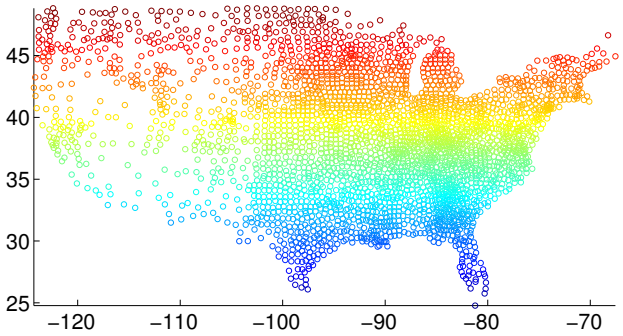
(c) $W_{ij}^{(2)} = \frac{M_{ij}}{P_i + P_j}$



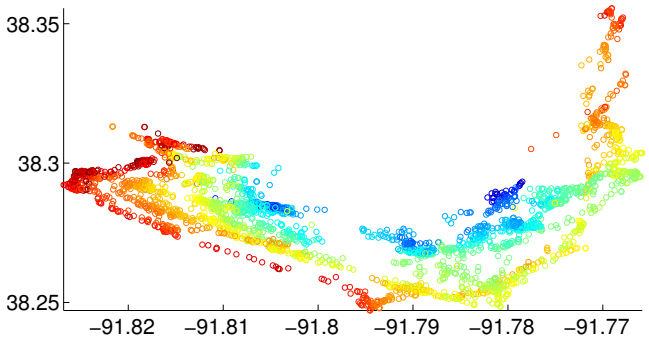
(d) $W_{ij}^{(3)} = \frac{M_{ij}}{P_i P_j}$

Figure: Diffusion map reconstructions from the top two eigenvectors, for various similarities, with nodes colored by longitude

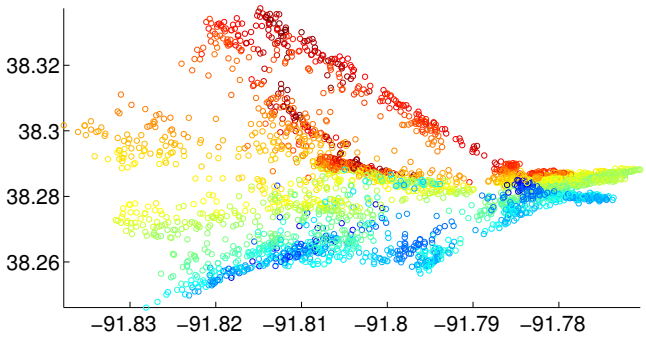
Colored by latitude



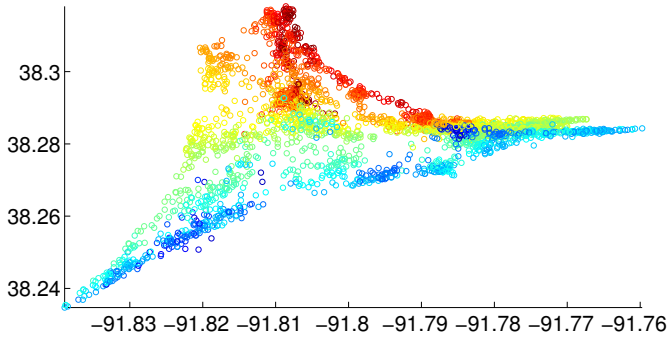
(a) Map of USA, coloured by latitude



(b) $W_{ij}^{(1)} = \frac{M_{ij}^2}{P_i P_j}$



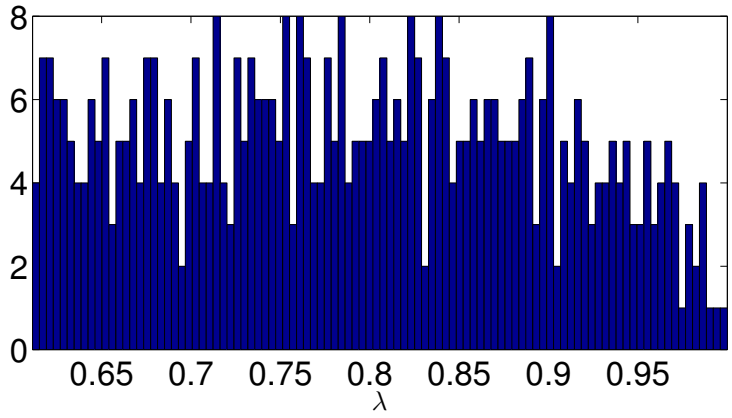
(c) $W_{ij}^{(2)} = \frac{M_{ij}}{P_i + P_j}$



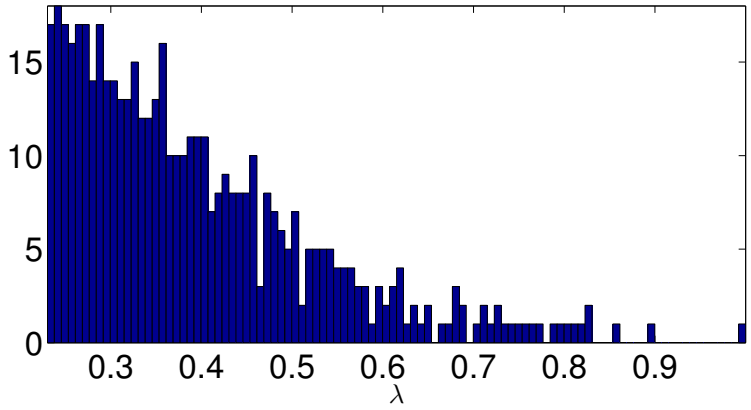
(d) $W_{ij}^{(3)} = 5500 \frac{M_{ij}}{P_i P_j}$

Figure: Diffusion map reconstructions from the top two eigenvectors, for various similarities, with nodes colored by longitude

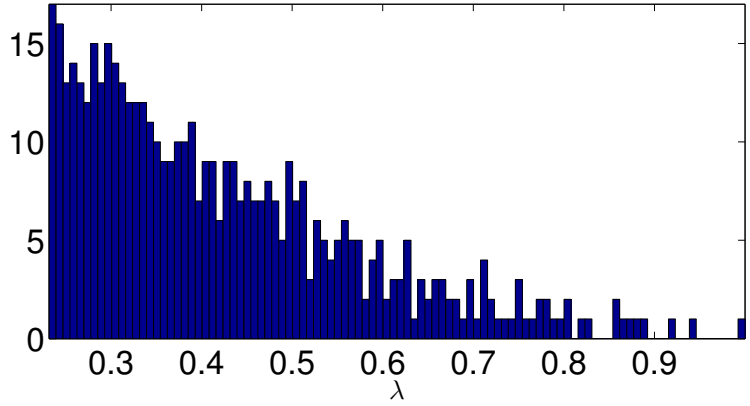
Spectrum



(a) $L = D^{-1}W^{(1)}$



(b) $L = D^{-1}W^{(2)}$



(c) $L = D^{-1}W^{(3)}$

Figure: Histogram of the top 500 eigenvalues of matrix L for different similarities.

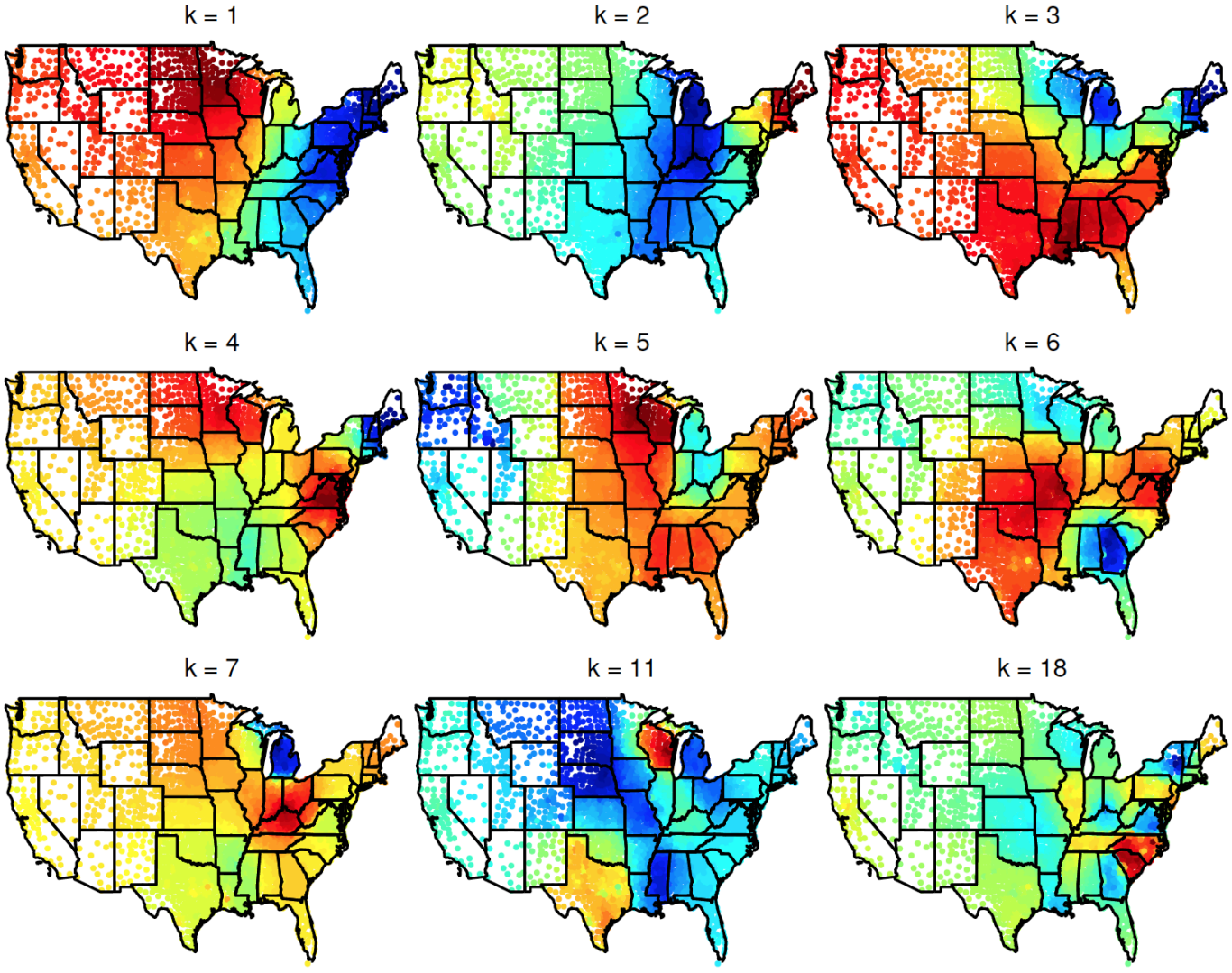


Figure: Eigenvector colourings for the similarity matrix $W_{ij} = \frac{M_{ij}^2}{P_i P_j}$.

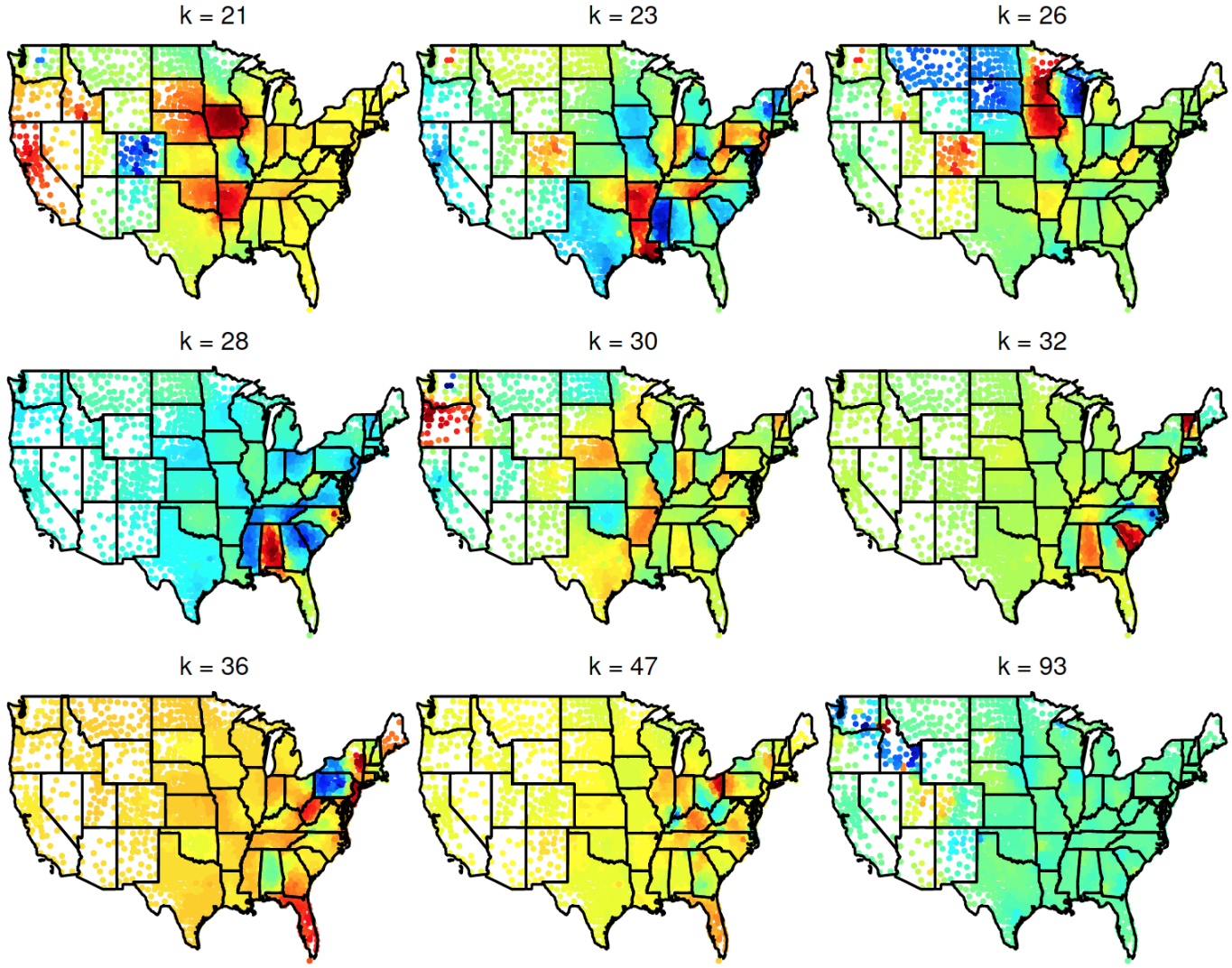
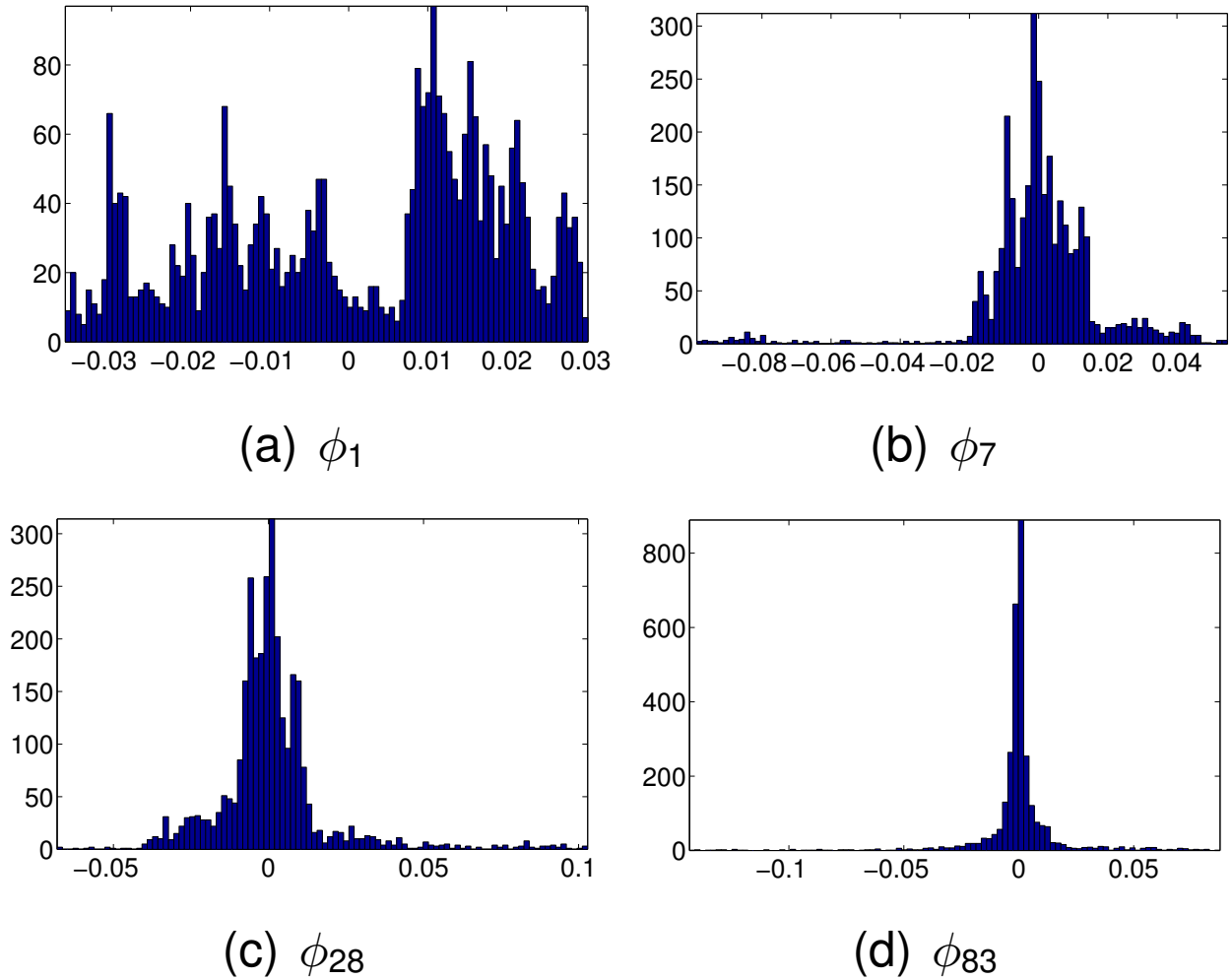


Figure: Further eigenvector colourings for the similarity matrix $W_{ij} = \frac{M_{ij}^2}{P_i P_j}$.



(a) ϕ_1 (b) ϕ_7

(c) ϕ_{28} (d) ϕ_{83}

Figure: Histogram of eigenvectors $\phi_1, \phi_7, \phi_{28}, \phi_{83}$ of $L = D^{-1}W^{(1)}$

- ϕ_1 provides a meaningful partitioning that separates the East from the Midwest; entries in $[-0.03, 0.03]$ with few entries of zero magnitude.
- however, eigenvectors ϕ_7, ϕ_{28} and ϕ_{83} are *localized*, i.e. they have their larger entries localized on a specific subregion of the US map (highlighted in blue or red in the eigenvector colorings), while taking small values in magnitude on the rest of the domain.

The graph partitioning problem (GPP)

- Investigate the connection of such geographically cohesive coloured subgraphs with the (GPP)
- In general, the GPP seeks to decompose a graph into K disjoint subgraphs (clusters), while minimizing the sum of the weights of the "cut" edges, i.e., edges with endpoints in different clusters
- Given the number of clusters K, the Weighted-Min-Cut problem is an optimization problem that computes a partition $\mathcal{P}_1, \ldots, \mathcal{P}_K$ of the vertex set, by minimizing the weights of the cut edges

$$\text{Weighted Cut}(\mathcal{P}_1, \ldots, \mathcal{P}_k) = \sum_{i=1}^{k} E_w(\mathcal{P}_i, \overline{\mathcal{P}_i}), \tag{3}$$

where $E_w(X, Y) = \sum_{i \in X, j \in Y} W_{ij}$, and $\overline{X}$ denotes the complement of X.

Spectral clustering

- extensive literature survey on spectral clustering algorithms: *Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416*

  `https://arxiv.org/abs/0711.0189`
- & the popular spectral relaxation introduced by Shi and Malik (early 2000s)
- When dividing a graph into two smaller subgraphs, one wishes to minimize the sum of the weights on the edges across two different subgraphs, and simultaneously, maximize the sum of the weights on the edges within the subgraphs.
- Alternatively, one tries to maximize the ratio between the latter quantity and the former, i.e., between the weights of the inside edges and the weights of the outside edges.
- We regard the US states as the clusters, and investigate the possibility that the isolated coloured regions that emerge correspond to local cuts in the weighted graph

Clustering

- denote by S the matrix of size $N \times N$ ($N = 49$ the number of mainland US states) that aggregates the similarities between counties at the level of states
- if state i has k counties with indices $x_1, \ldots, x_k$, and state j has l counties with indices $y_1, \ldots, y_l$, then we consider the $k \times l$ submatrix

$$\tilde{W}_{i,j} = W_{\{x_1,\ldots,x_k\},\{y_1,\ldots,y_l\}} \tag{4}$$

  and denote by S_{ij} the sum of the kl entries in $\tilde{W}_{i,j}$
- heatmap shows the components of the matrix S on a logarithmic scale, where the intensity of entry (i, j) denotes the aggregated similarity between states i and j

Cluster-Cluster Meta Adjacency Matrix

$S :=$

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}
Cluster 2	.	.	.	.	.
Cluster 3	.				
Cluster 4	.				
Cluster 5	S_{51}	S_{52}	S_{53}	S_{54}	S_{55}

- S_{ii} is "inside degree" of state i, $d_i^{in} = S_{ii}$, which measures the internal similarity between the counties of state i
- denote by $d_i^{out} = \sum_{u=1, u \neq i}^{N} S_{i,u}$ (i.e., the sum of the non-diagonal elements in row i) the "outside degree" of node i, which measures the similarity/migration between the counties of state i and all other counties outside of state i
- denote by $d_i^{ratio} = \frac{d_i^{in}}{d_i^{out}}$, the "ratio degree" of node i which straddles the boundary between intra-state and inter-state migration
- a large ratio degree is a good indicative that a state is very well connected internally, and has little connectivity with the outside world, and thus is a good candidate for a cluster.
- the Table ranks the top 15 states within the US in terms of their ratio degree.

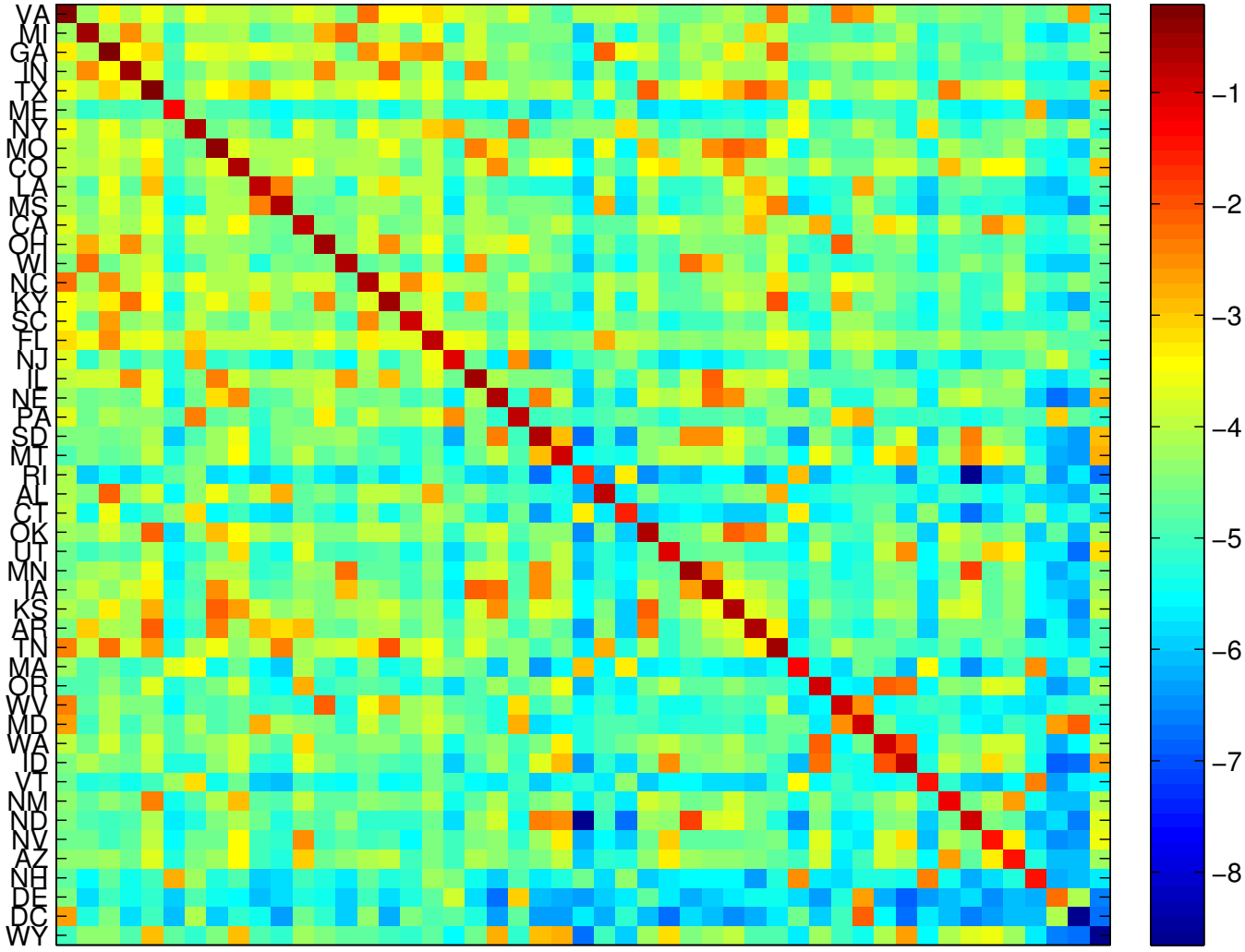


Figure: Heatmap of the inter-state migration flows. Rows (and columns) are sorted by the ratio degrees of the states. The intensity of entry (i, j) denotes, on a logarithmic scale, the similarity between states i and j, i.e., the sum of all entries in the submatrix $\tilde{W}_{i,j}$

rank	state	ratio degree
1.	VA	26.7
2.	MI	20.4
3.	GA	19.9
4.	IN	19.7
5.	TX	19.0
6.	ME	18.9
7.	NY	18.7
8.	MO	18.5
9.	CO	17.1
10.	LA	16.6
11.	MS	16.1
12.	CA	15.7
13.	OH	15.6
14.	WI	14.5

Table: Top 15 states within the US, ordered by ratio degree.

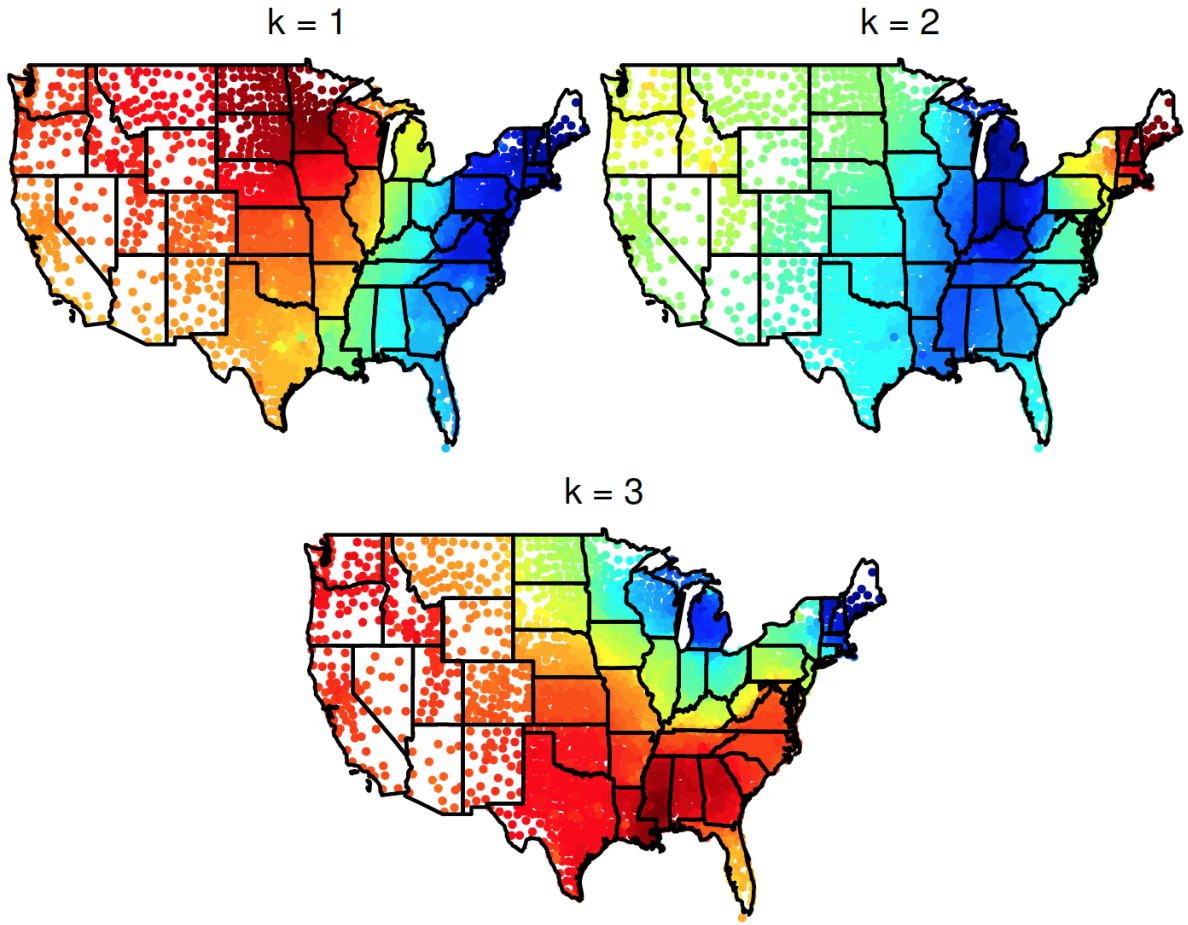
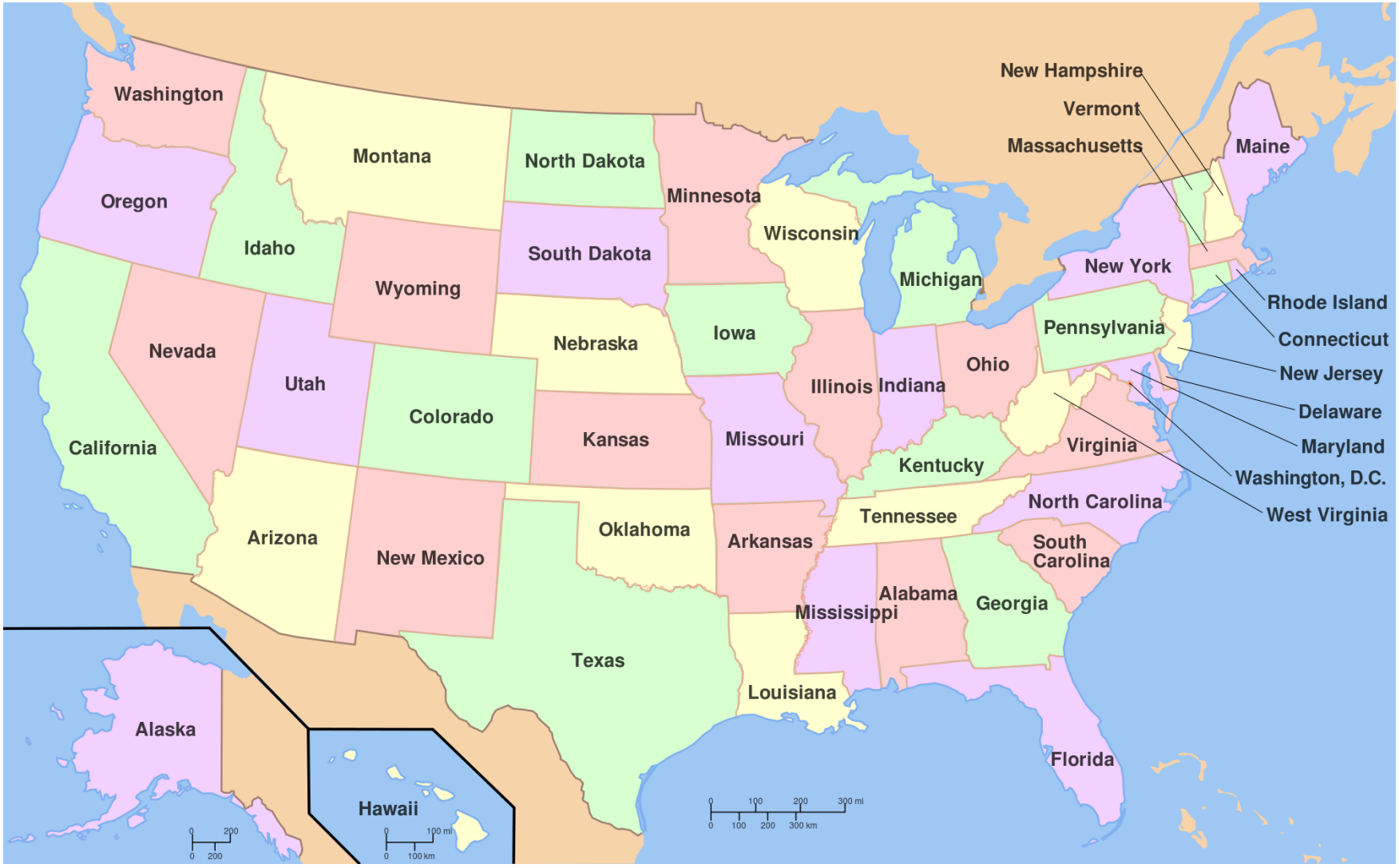
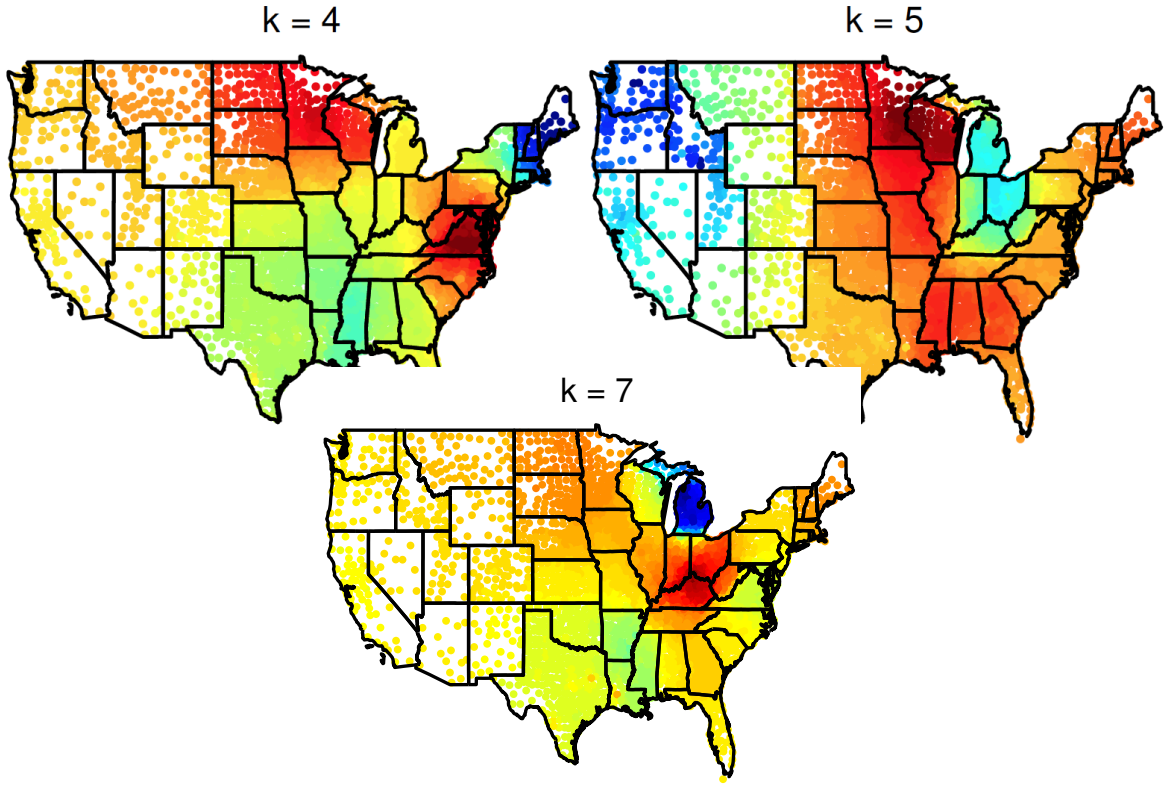


Figure: Top three eigenvectors correspond to global cuts between various coasts within the US. The only state that stands out individually is Michigan (MI) for $k = 3$, which has rank 2.


Washington
Oregon
California
Idaho
Nevada
Montana
Wyoming
Utah
Arizona
Colorado
New Mexico
North Dakota
South Dakota
Nebraska
Kansas
Oklahoma
Texas
Minnesota
Iowa
Missouri
Arkansas
Louisiana
Wisconsin
Illinois
Michigan
Indiana
Ohio
Kentucky
Tennessee
Mississippi
Alabama
Georgia
Florida
South Carolina
North Carolina
Virginia
West Virginia
Pennsylvania
New York
Maryland
Delaware
New Jersey
Connecticut
Rhode Island
Massachusetts
Vermont
New Hampshire
Maine
Alaska
Hawaii
Washington, D.C.

Eigenvector colorings vs Ratio Degree



- $k = 4$: the largest entries correspond to counties in Virginia (VA) which is also ranked 1^{st}
- $k = 5$: Wisconsin (WI) ranked 14
- $k = 6$: the states coloured in dark red and dark blue are Georgia (GA) with rank 3, and Missouri (MO) of rank 8
- $k = 7$: Michigan (MI), of rank 2, stands out as the only dark blue coloured state.