Extracting Geography From Datasets in Social Sciences

Yuke Li
yuke.li@yale.edu

Tianhao Wu
Yale University, tianhao.wu@yale.edu

Nicholas Marshall
Yale University, nicholas.marshall@yale.edu

Stefan Steinerberger
Yale University, stefan.steinerberger@yale.edu

Yale
DAY OF DATA 2016

Extracting Geography From Datasets in Social Sciences

Yuke Li[1], Tianhao Wu[2], Nicholas F. Marshall[3] and Stefan Steinerberger[3]

[1]Department of Political Science, [2]Department of Statistics, [3]Department of Mathematics

The Actual Geography of Human Interactions

Motivated by the fact that many forms of interactions stretch beyond the traditional geographical confines in the age of globalization, we define aresearch question of the actual geography of human interactions. By applying the methods of dimensionality reduction, we study the innate geometry of several datasets in social sciences and, in doing so, extract the actual geography of the underlying interactions. The significance of this practice is thatthe local-global boundaries for every possible form of human interactions, for which data is available, can then be redefined.

Dimensionality Reduction for Network

The agents interactions can be represented by a directed and weighted graph, where the set of nodes represents the collection of agents, the set of edges represents the connections among the agents and, as the interactions are usually quantifiable and the edges are weighted.

In order to extract the actual geography, we reduce the number of the dimensions in the data. Dimensionality reduction relates to a collection of methods that aims to convert a set of data with a huge number of dimensions into data with fewer dimensions, ensuring that the converted data restores the main information concisely.

We proceed by mapping n entities to a complete, weighted graph

$$\phi_1 : \{\text{collection of entities}\} \to \text{complete, weighted graph on } n \text{ vertices}$$

that encodes the given information. The crucial part is the construction of a map

$$\phi_2 : \text{complete, weighted graph on } n \text{ vertices} \to \mathbb{R}^2$$

that preserves as much information as possible of the data of the "interactions" among those regions. The composition is therefore

$$\phi_2 \circ \phi_1 : \{\text{collection of entities}\} \to \mathbb{R}^2$$

is the desired mapping.

Principal Component Analysis and Diffusion Embedding

The main tools of dimensionality reduction we use are Principal Component Analysis (PCA) developed by Hotelling (1933, 1936) and Diffusion Embedding introduced by Coifman & Lafon (2006).

1 PCA

Suppose that a set of n regions is given. The data of their interactions is denoted by a $n \times n$ matrix $\mathbf{T}$:

$$\mathbf{T}(i,j) = \text{Amount of Interactions to } j \text{ from } i \quad \text{for } i,j = 1,\dots,n,$$

where $\mathbf{T}(i,i) = 0$ by convention. In order to adjust for the different sizes of the regions, we will normalize $\mathbf{T}$ by dividing the rows of $\mathbf{T}$ by the rows sums. Explicitly, we define a data matrix $\mathbf{X}$ as $\mathbf{X}(i,j) = \frac{\mathbf{T}(i,j)}{\sum_{k=1}^{n}\mathbf{T}(i,k)}$. We consider the columns of $\mathbf{X}$ as samples from an n-dimensional space. By construction, the dimensions have been normalized to facilitate comparison.

Next, we use PCA to reduce the dimensions of the data from n to two. The PCA algorithm consists of two steps. First, we subtract the mean from each from of $\mathbf{X}$ and define $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ denotes a n-dimension column vector of all 1s, and $\mathbf{1}^T$ denotes the transpose of $\mathbf{1}$. Next, we compute the Singular Value Decomposition (SVD) of $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

where $\mathbf{U}$ is an orthogonal matrix of left singular vectors, $\mathbf{S}$ is a diagonal matrix of the corresponding singular values, and $\mathbf{V}$ is an orthogonal matrix of the corresponding right singular vectors. The right singular vectors $\mathbf{V}$ are the eigenvectors of the covariance matrix:

$$\mathbf{Y}\mathbf{Y}^T = (\mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\mathbf{1}^T)(\mathbf{X} - \frac{1}{n}\mathbf{X}\mathbf{1}\mathbf{1}^T)^T$$

We define an embedding ϕ mapping the regions into $\mathbb{R}^2$ by

$$i \overset{\phi}{\mapsto} (\mathbf{V}(i,1), \mathbf{V}(i,2)) \in \mathbb{R}^2$$

for $i = 1,\dots,n$.

Principal Component Analysis and Diffusion Embedding

2 Diffusion Embedding

Let $\mathbf{T}$ denote the $n \times n$ matrix and define a symmetric affinity kernel by $\mathbf{K} = \frac{1}{2}\left(\mathbf{T} + \mathbf{T}^T\right)$. Starting with the symmetric affinity kernel $\mathbf{K}$, we will proceed with the Laplace-Beltrami normalization diffusion maps algorithm. First, we define a diagonal matrix $\mathbf{D}$ whose entries of the sum of the rows of $\mathbf{K}$, i.e. $\mathbf{D} = \text{diag}\,(\mathbf{A1})$, where $\mathbf{1}$ denotes an n-dimensional column vector of 1s. Using $\mathbf{D}$, we will normalize $\mathbf{K}$ defining

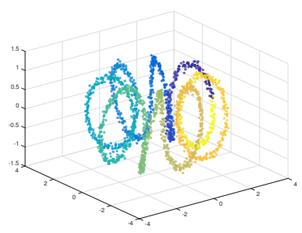
$$\mathbf{W} = \mathbf{D}^{-1}\mathbf{K}\mathbf{D}^{-1}$$

Finally, we will define the Markov matrix

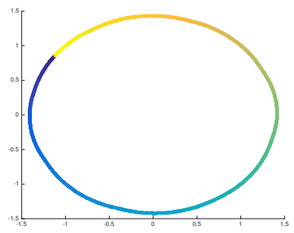
$$\mathbf{P} = \mathbf{W}\mathbf{Q}^{-1}$$

where $\mathbf{Q} = \text{diag}(\mathbf{W1})$. The eigenvalues and eigenvectors of $\mathbf{P}$ are used to define the diffusion map. Specifically, we decompose $\mathbf{P} = \mathbf{\Psi}\mathbf{\Lambda}\mathbf{\Psi}^{-1}$ and define a embedding φ from the n countries to $\mathbb{R}^2$ by

$$i \overset{\varphi}{\mapsto} (\mathbf{\Lambda}(2,2)\mathbf{\Psi}(i,2), \mathbf{\Lambda}(3,3)\mathbf{\Psi}(i,3))$$

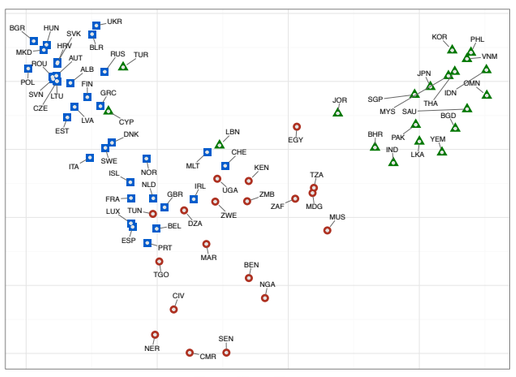


A Noisy Toroidal Helix

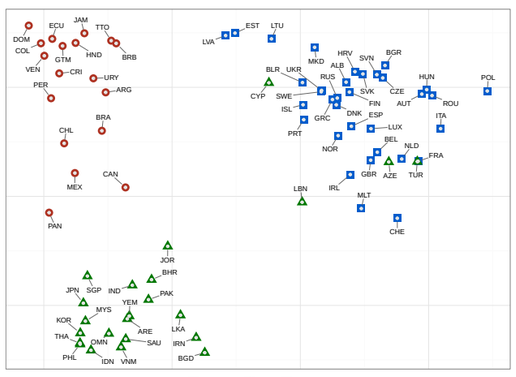


A Perfect Elliptocytosis by Diffusion Embedding

Application on International Trade Flow



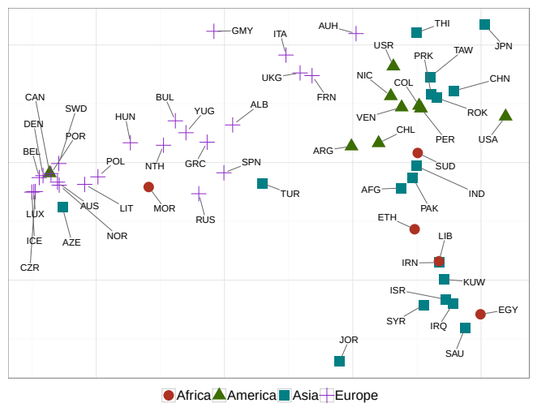
Africa, Asian and Europe Trade in 2009



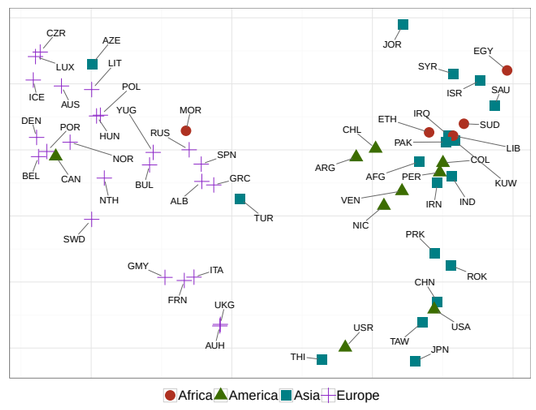
Americas, Asia and Europe Trade in 2009

- Obvious clusters by continent.
- Austria is in a cluster surrounded by the Czech republic, Croatia, Hun- gary, Romania, Slovakia and Slovenia (all were at least partially elements of the Austro-Hungarian empire).
- Close countries: Spain and Portugal (geographic proximity), Belgium (linguistic proximity) and France (both historical and linguistic proximity).
- In the context of trading among America-Asia-European countries, Azerbaijan, Cyprus, Lebanon and Turkey actively trade as if they were European countries.

Applications on Militarized Interstate Disputes



PCA Map on MID

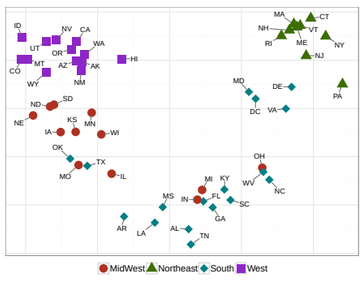


Diffusion Map on MID

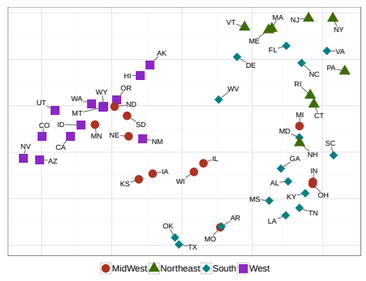
- The countries fall into North African, East Asian, Middle East, European and American clusters.
- Some country dyads have long-term conflicts throughout the hundred years. Therefore, regardless of the method or the normalization, their closeness remains robust. The examples include Germany-France, Germany-Italy, Russia-Turkey, Russia- Poland, Russia-Sweden, Russia-Greece and Russia-Spain Dyads in the European cluster, the China-Japan dyad in the East Asian cluster, Jordan-Syria in the Middle-Eastern cluster, the Chile-Peru-Colombia Colombia-Venezuela, among others.

Mirror the U.S. Map

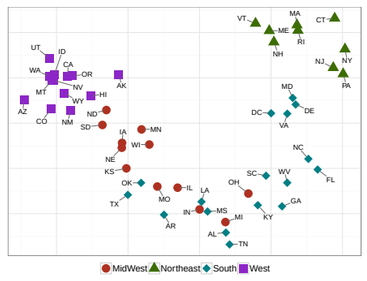
State-to-State flows are also important sources of human interactions, we applied the dimensionality reduction method to three datasets: Commodity Flow, Migration and Air Carrier of the U.S.



Commodity Flow



Air Carriers



Migration

The maps we created mirror the actual U.S. map to an extreme extent, when the differences reveal specifications of the datasets.

References

- Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24 (6), 417441.
- Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28 (3/4), 321377.
- Coifman, R. R., Lafon, S., 2006. Diffusion maps. Applied and Computational Harmonic Analysis 21 (1), 530.