# Finding the Math Department's Deep Structure

## Illinois Geometry Lab

Faculty Mentor: Yuliy Baryshnikov
Project Leaders: Haoyuan Li, Anji Dong, Ning Jiang
IGL Scholars: Prajeet Basu, Zifan Dong, Ziqi Xu, Adam Wawrowski, Qingyu Yi

University of Illinois at Urbana-Champaign


Midterm Presentation
October 15, 2023

# Project Goals

A team did this project related to database creation, data cleaning, initial proximity measures, and some clustering efforts 2 years ago.

Based on their work, our primary objective is to explore the convergence of research interests among mathematics professors to form clusters, representing distinct areas of mathematics.

- Barycenters in the Space of (Phylogenetic) Trees:
  - Generate distance matrices and phylogenetic trees
  - Evaluate the resulting trees and implied clusterings for consistency with the existing department structure.
  - Find plausible barycenters to detect the intrinsic research clusters.
- Spectral Methods:
  - Generate feature matrices and perform dimensionality reduction using SVD and PCA
  - Assess cluster quality to describe departmental strength, and explore alternative clustering methods like Spectral Clustering.

# Proximity measures

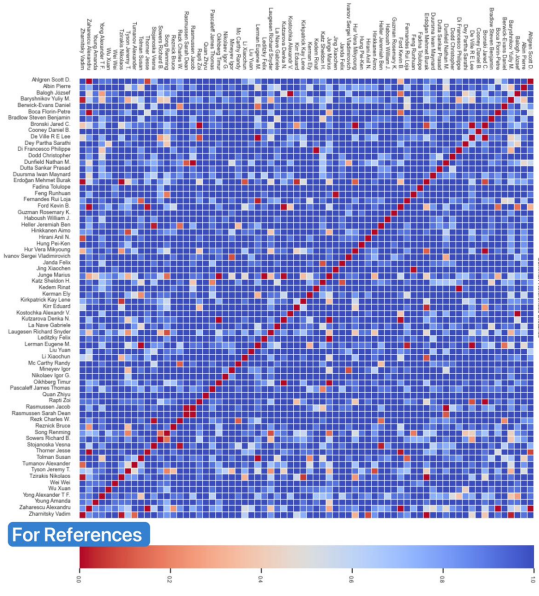
1. Raw Data: Sort out all the data and generate raw data.
2. Similarity matrix: Parse the raw data to create the corresponding similarity matrices for journals and references.
   - For References S[i, j] = Number of times Prof i and Prof j use the same reference.
   - For Journals S[i, j] = Sum of all papers written in the common journals of Prof i and Prof j.
3. Distance Matrix: Define the distance matrix $distance = 1/(r^n)$. n is the entries in similarity matrix and r is a parameter.

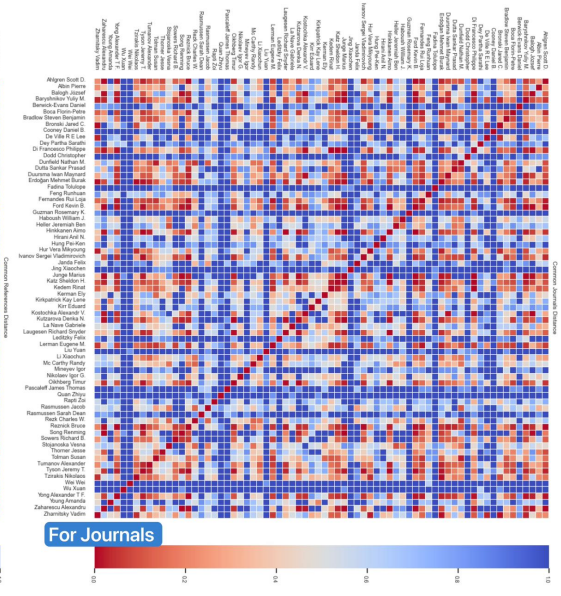
| For Journals | Ahlgren Scott D. | Albin Pierre | Balogh József | Baryshnikov Yuliy M. | Berwick-Evans Daniel |
|---|---|---|---|---|---|
| Ahlgren Scott D. | 56 | 14 | 22 | 30 | 2 |
| Albin Pierre | 14 | 26 | 5 | 20 | 2 |
| Balogh József | 22 | 5 | 177 | 26 | 0 |
| Baryshnikov Yuliy M. | 30 | 20 | 26 | 75 | 0 |
| Berwick-Evans Daniel | 2 | 2 | 0 | 0 | 8 |

5 rows × 70 columns

| For References | Ahlgren Scott D. | Albin Pierre | Balogh József | Baryshnikov Yuliy M. | Berwick-Evans Daniel |
|---|---|---|---|---|---|
| Ahlgren Scott D. | 475 | 0 | 0 | 0 | 2 |
| Albin Pierre | 0 | 601 | 0 | 10 | 7 |
| Balogh József | 0 | 0 | 1699 | 5 | 0 |
| Baryshnikov Yuliy M. | 0 | 10 | 5 | 607 | 1 |
| Berwick-Evans Daniel | 2 | 7 | 0 | 1 | 164 |

5 rows × 70 columns



# Hierarchical clusterings and their aggregation

- Phylogenetic trees: We use the simple linkage method to construct phylogenetic trees based on a distance matrix.
- Newick Format trees: We utilize hierarchical clustering to generate Newick format trees for identifying barycenters.
- Next steps:
  - Find barycenter in the space of the phylogenetic trees.
    The method is based on the paper "Geodesics to Characterize the Phylogenetic Landscape," where we employ Sturm's algorithm and the Pathtrees package in Python.

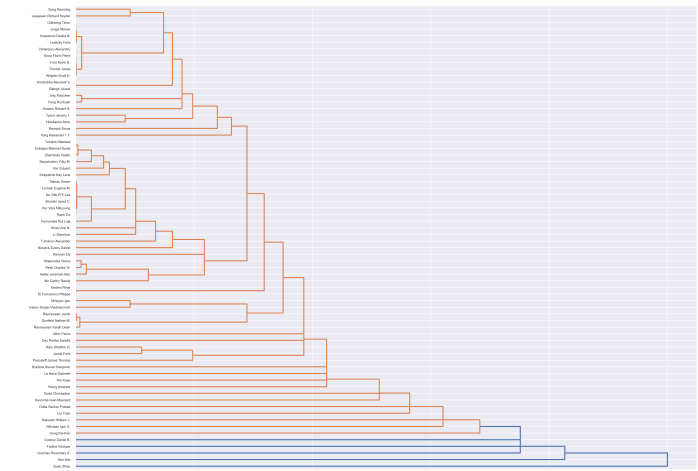


Figure: For References

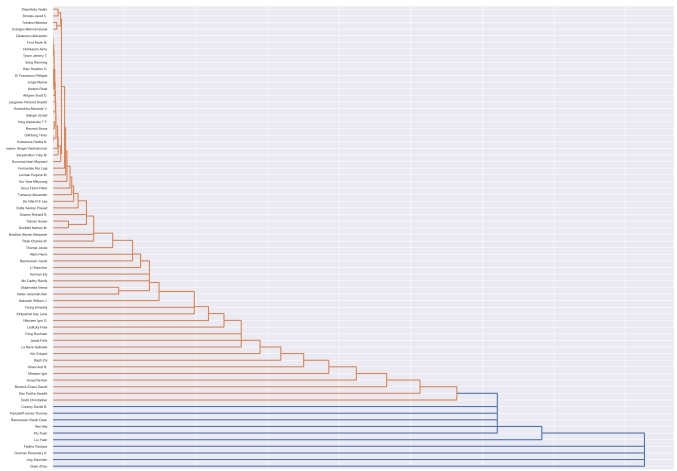
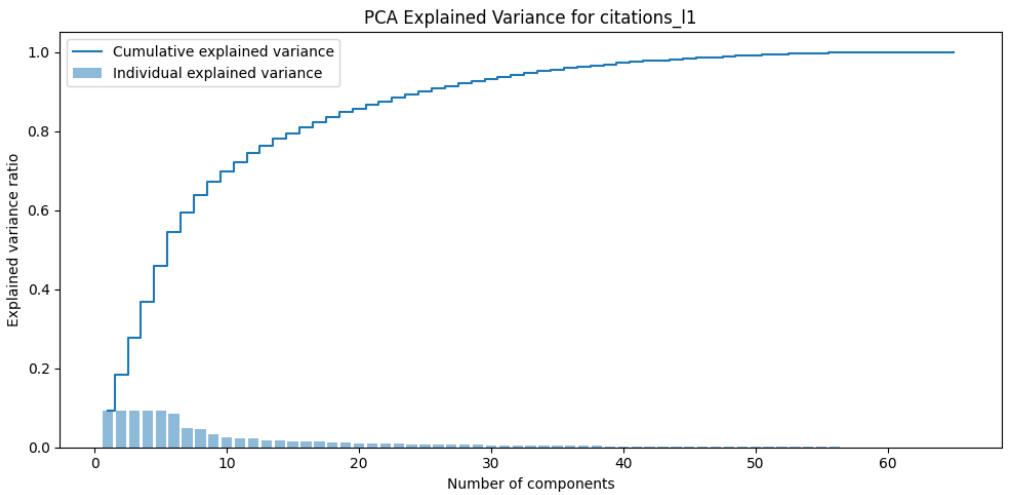


Figure: For Journals

# Dimensionality reduction

- Dataset: Our data frame indicates the number of times a particular professor (in a given row) has cited a particular paper (in the corresponding column)
- Most of the columns have just 1 faculty linked to it. In other words, the vectors spanning the columns are nearly orthogonal.
- SVD: decomposes the original data matrix into three matrices: $\mathbf{U}$, $\Sigma$, and $\mathbf{V}^T$.
- The first few columns of $\mathbf{V}$ (corresponding to the largest singular values) represent the most significant directions of data variance.



# Scatterplots

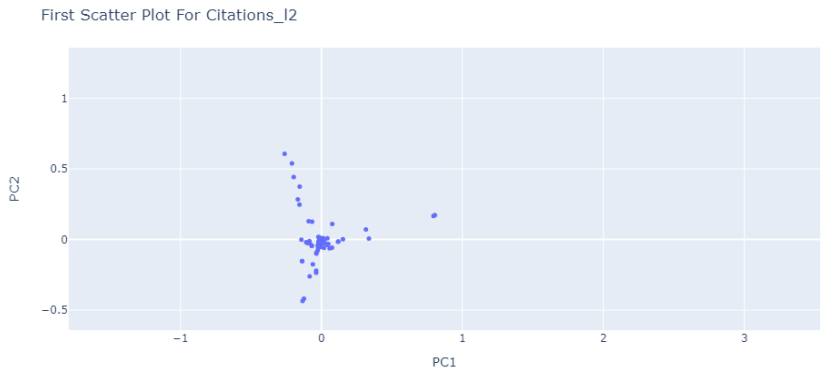



Figure: First Scatter Plot with Citations Data

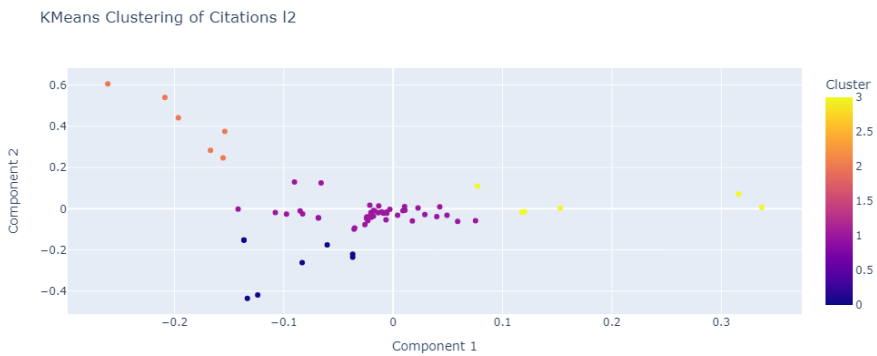



Figure: KMeans Clustering after removing the outliers

# Thank you

Thank you!