Finding the Math Department's Deep Structure

Zifan Dong, Adam Wawrowski, Ziqi Xu, Prajeet Basu, Qingyu Yi, Anji Dong, Ning Jiang, Haoyuan Li, Yuliy Baryshnikov

Introduction

This project explores the dynamic nature of our department's strengths by identifying research clusters based on shared areas and collaborative patterns. It aims to assess the significance of these clusters within their respective fields, focusing on strengths and growth. Building on a previous Spring '22 IGL project, we extract and aggregate these clusters using existing data.

Barycenter Team Goal

Use the phylogenetic trees generated and find barycenters in the space of (phylogenetic) trees

Spectral Method Team Goal

Use dimension-reducing methods to cluster the faculty of the UIUC Math Department based on identifiers of research.

Data

In Spring 2022, a previous IGL group worked on this project to gather the data we used in this project. Their repository collects data about publications by math faculty from [MathSciNet](https://mathscinet.ams.org/mathscinet/index.html).

We used this data to construct two types of matrices for hierarchical clustering based on four identifiers: Mathematics Subject Classification (MSC), References, Citations, and Journals.

Count Matrix

In this matrix, we have the members of the faculty as rows and the list of identifiers for the research papers as columns. Cell i, j represents the number of times that a paper from faculty i identifies with j.

Similarity Matrix

In this matrix, we have the members of faculty on the rows and columns, where cell i, j represents a metric of similarity based on an identifier of their research between faculty i and j.

We derive the Similarity matrices from the Count Matrices. We used different identifiers of the published research to cluster the faculty based on different metrics. This allows for comparing results across identifiers.

Methods and Results

Barycenter Team:

We then generate the initial phylogenetic trees for each distance matrix using hierarchial clustering, along with their corresponding heat maps.

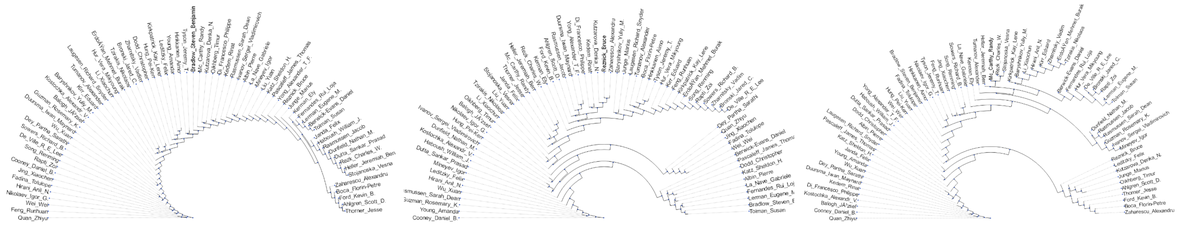

Figure 1: Journal - MSC - Reference

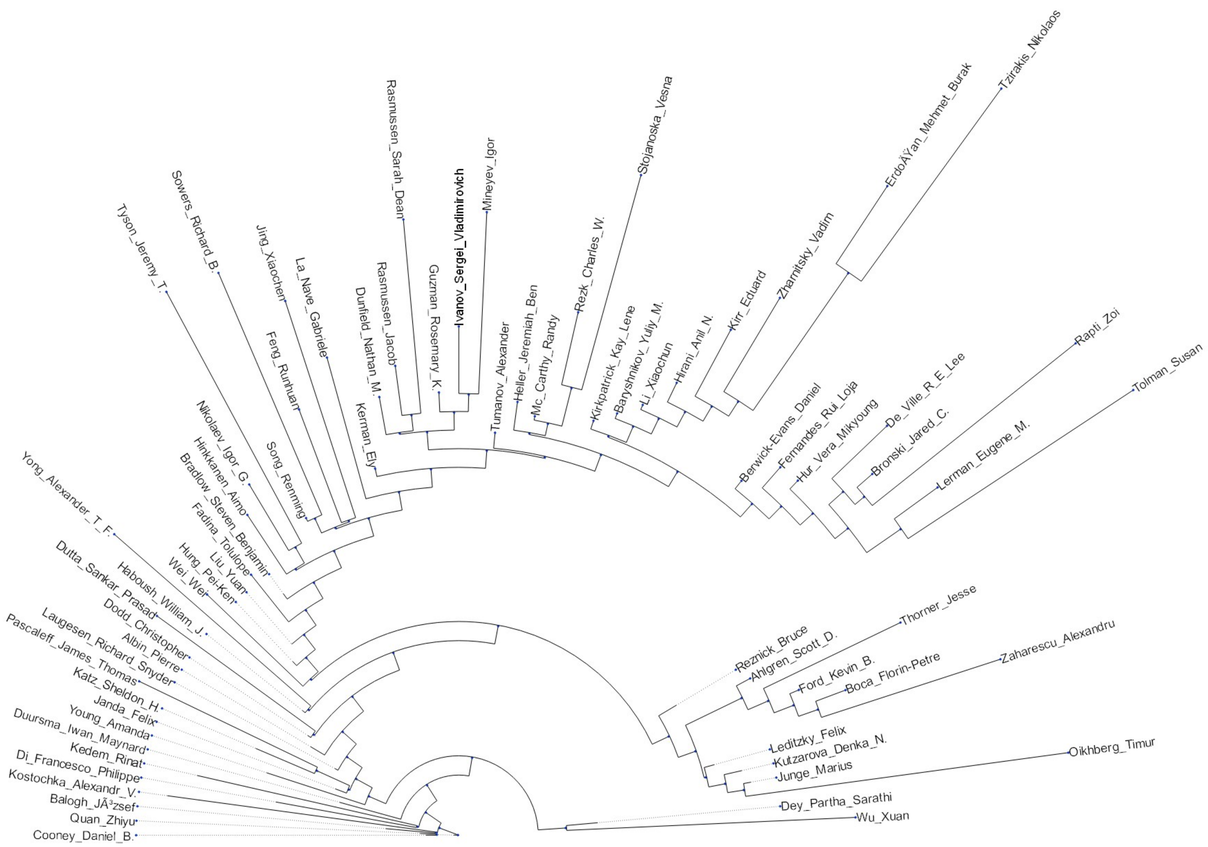

Figure 2: Barycenter 17

We take the three trees and use the "pathtrees.py" file in the pathtrees repository to generate the barycenter through an iterative process: (note numpathtree = of endpoint trees + pathtrees) In general, this process can be repeated for as many iterations as needed to keep finding more and more precise barycenters. We aggregate our iterations into a single file which can be transformed into visualized trees, with the final iteration being the current barycenter, using our tree plotting code. The above is the barycenter after the algorithm is repeated 17 times.

Algorithm 1 Generates the Barycenter between MSC, References, and Journals

Require: $N > 0$
- $MSC \leftarrow MSCTree$
- $REF \leftarrow ReferenceTree$
- $JOUR \leftarrow JournalTree$
- $numpathtrees = 3$
- $OutPut0 \leftarrow internalpathtrees(MSC, REF, numpathtrees)$
- $numpathtrees = 4$
- $outPut1 \leftarrow internalpathtrees(OutPut0, JOUR, numpathtrees)$
- **while** $N > 0$ **do**
  - $numpathtrees = 3 + N$
  - $outPutN \leftarrow internalpathtrees(MSC, REF, JOUR, numpathtrees)$
  - $N \leftarrow N - 1$

Spectral Team:

After preprocessing, we performed PCA on the matrices. Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used in data analysis. It transforms the original features of a dataset into a new set of uncorrelated variables called principal components. By retaining only a subset of these components, PCA helps capture the most significant information in the data while reducing its dimensionality. After performing PCA, we conducted the following two clustering methods on the data:

Affinity Propagation is a clustering algorithm that identifies exemplars within a dataset, representing data points that best summarize the entire set. It relies on the concept of message passing between data points to iteratively determine the most representative exemplars and assign other points to them, effectively partitioning the data into clusters.

K-means is a clustering algorithm that partitions a dataset into a predetermined number of clusters. It iteratively assigns data points to clusters based on the mean of the features within each cluster and updates cluster centroids until convergence.

Using these two methods, we found a mixed result in the quality of the clusters, which will need to be improved upon.

Future Directions

Barycenter Team:

In the future, in addition to journals, references, and MSC, more perspectives can be used to generate the Barycenter of the trees to find the deeper structure of the UIUC Mathematics Department. Meanwhile, the iterative process can be repeated more and more times to make the results converge to the barycenter infinitely, thus increasing the accuracy of the results.

Spectral Team:

Although we were successful in generating clusters of the faculty in terms of References, Citations, Journals, and MSC, only the clusters for the MSC identifier were logically sound. As a result, the next step would be to work on identifying the problems with clustering based on References, Citations, and Journals, and fixing these problems.