Department's Deep Structure

IGL Scholars: Dev Patel, Peiyan Wu, Yizhi Zhang, Zheyu Zhang
Project Mentor: Sujeet Bhalerao, Faculty Advisor: Prof. Yuliy Baryshnikov

Introduction

The goal of this project was to make a program that accepts as input a list of mathematicians, say from a department at any institution, and detects intrinsic research clusters (i.e. groups of people working in close areas, publishing in the same journals etc.) within the department. This classification is based on data scraped from the online database MathSciNet using the Python programming language. This data allows us to define proximity measures (that is, a notion of how close faculty members are), using which we can apply hierarchical clustering algorithms to gain insight into research structure of the math department. Finally, the project aimed to visualize results of the clustering using phylogenetic trees (dendrograms).

The current program is based on Python and the user must install a Python environment on their device. Next, the user must create a text file that contains the list of names of faculty members from the department to be clustered into research areas. This text file is to be used as input for the program.

Clustering

The main clustering method we used on analyzing our data is agglomerative hierarchical clustering. Roughly, this is a method for clustering in which each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. The precise algorithm for merging clusters depends on the linkage criterion, which determines distances between sets of observations. Given a linkage criteria $D(a, b)$ that defines the distance between any pair of clusters (A, B), the clustering algorithm is described below:

Algorithm:

1. Initialize all data points as clusters of one element.
2. Compute distance $D(A, B)$ between all clusters.
3. Find the shortest distance r between any pair of clusters.
4. For each pair of cluster (A, B) such that $D(A, B) = r$, merge them into a single cluster. $(A, B) \to A'$
5. Save the current state labeled with r.
6. Repeat step 2-5 until every data point is merged into a single cluster.

This algorithm produces a tree structure of clusters, also called dendrogram. The linkage criteria we used is unweighted average linkage (UPGMA):

Given metric d. For clusters A and B,

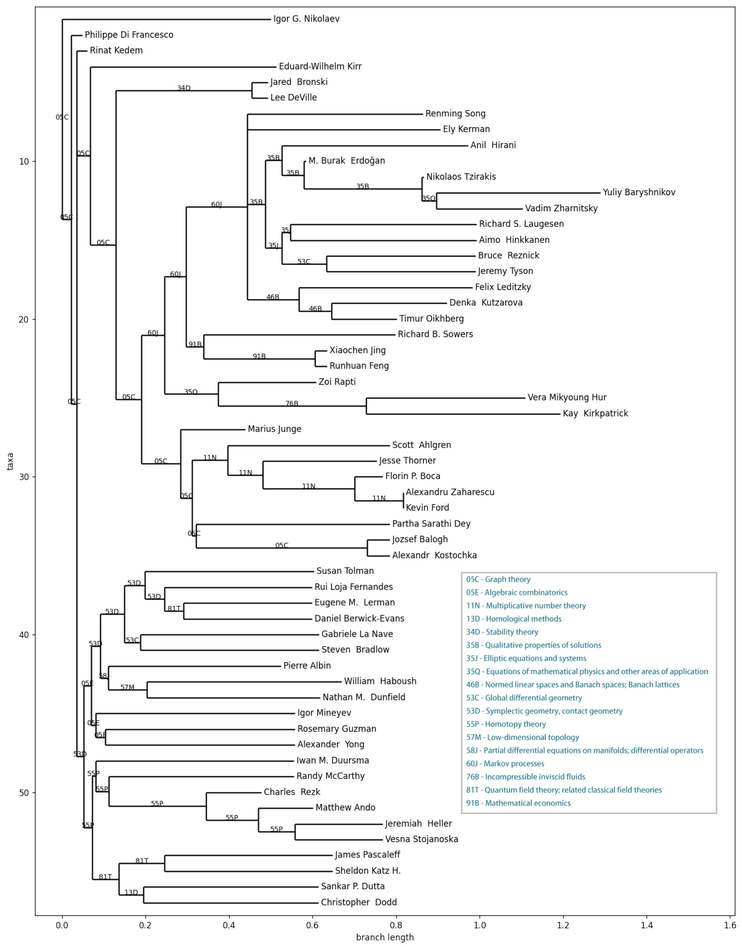
$$D(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

After experimenting with many sources of data, we determined that citations and references data are most suitable for our task.
Since different areas of study can have very different research output, we also normalized the data we collected to prevent biases in clustering: Let X denote number of joint citation between faculty members A, B,

$$\hat{X}_{A,B} = \frac{X_{A,B}}{\sqrt{A_{\text{total}} \cdot B_{\text{total}}}}$$

Result

The below result is the consensus dendrogram of clustering results from normalized joint citation and common references data. The branches are also labeled with the "majority" MathScinet classification code of the corresponding sub-cluster. The label for the classification code is found by using the extracted data to find the classification code under which the maximum number of papers are classified.

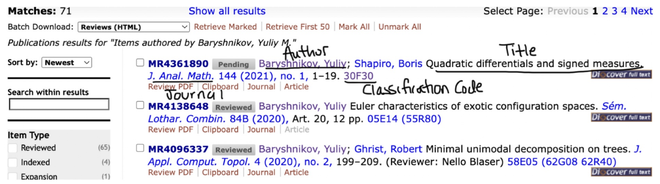


Each leaf branch corresponding to a faculty in math department. The clustering shows how close they work together. If two faculties are in the same leaf cluster, then they most likely work in the same area, and people in close clusters are likely to be working in related areas.

Data Collection

Web Scraping

We used python packages Selenium and BeautifulSoup to extract data from MathSciNet. An example of a search result on MathSciNet for publications from a faculty member is given below.



- We extracted faculty names, references, citations, MathScinet classification codes, and journal names for each publication of each faculty member in the department.
- Next, we generated similarity matrices (number of common journals, citations, references, etc). For each dataset and corresponding similarity matrix created, the $(i, j)^{th}$ entry is listed in the table below.
- Finally, we also created matrices for SVD-based clustering method with respect to classification codes, references, and citations.

Joint Citation: Number of times each pair of faculties are cited in the same paper.
Joint Publication: Number of paper each pair of faculties co-authored.
Common Reference: The number of papers a pair of scholars both referred to in their publications.
Common Journals: Number of papers each pair of faculties published in same journals.
Directed Citation: Number of times one faculty cited another faculty in his/her publications.
Reference Matrix: Number of times any author has reference(s) for one specific paper.
Citation Matrix: Number of times any author has citation(s) for one specific paper.
Classification Matrix: Number of papers author has under specific classification.

Future Directions

- Integrate the data collection and cluster analysis programs into a single package.
- Integrate zbMath, formerly Zentralblatt MATH, which is another online database, as additional data source.
- Extend this program to other fields such as Physics and Computer Science Research.
- Explore SVD-based clustering methods.
- Obtain finer classifications and visualize them more clearly.

References

[1] Gunnar Carlsson Facundo Memoli. Characterization, Stability and Convergence of Hierarchical Clustering Methods. *Journal of Machine Learning Research*, 11:1425-1470, (2010)
[2] Jon. Kleinberg. An Impossibility Theorem for Clustering. *Adv Neural Inform Process Syst (NIPS)*, 15, (2003)
[3] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25.11, 1422â1423, (2009)

Support for this project was provided by the Illinois Geometry Lab and the Department of Mathematics at the University of Illinois at Urbana-Champaign.