Barycenter Analysis in the UIUC Mathematics Department

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

December 14, 2023

Abstract

The Barycenter Team conducted a work that involves examining phylogenetic trees obtained from data supplied by a previous group in the UIUC Mathematics Department. We employ agglomerative clustering to produce phylogenetic trees. The goal is to determine the median or barycenter of these trees, which reflect references, MSC categories, and journals. The approach we use involves iterative operations utilizing the "pathtrees.py" script to accurately compute the barycenter. This study provides a comprehensive overview of the tree production process, including the mathematical techniques for determining the barycenter. Additionally, it explores potential avenues for future enhancements in our research.

1 Data Preprocessing

1.1 Similarity Matrices

First we parse the Journals Data, References Data, and MSC Data to create the corresponding similarity matrices.

1.1.1 Similarity Matrix for Journals

For any two professors i and j, the similarity matrix S for journals is defined as:

$$S[i, j] = \text{Number of unique journals in common between prof } i \text{ and prof } j$$

For example, if professor A published papers in journals X and Y, while professor B published in journals X and Z, then the entry $S[A, B]$ of this matrix would be 1.

1.1.2 Similarity Matrix for MSC

The MSC similarity matrix S considers both main and side codes used by the professors:

$$S[i, j] = 2 \cdot (\text{Number of main codes in common}) + 1 \cdot (\text{Number of side codes in common})$$

Note: Only the first two digits of each MSC code are considered.

For example, if professor A's codes are [X, Y, (Z), (W)] and professor B's are [Y, (Z), T, (U)], then the entry $S[A, B]$ would be $2 + 1 = 3$ as they share the main code Y and the side code (Z).

1.1.3 Similarity Matrix for References

For references, the similarity matrix S is given by:

$$S[i, j] = \text{Number of times prof } i \text{ and prof } j \text{ have cited the same reference}$$

For instance, if professor A referred to papers x, y, z in her papers, and professor B referred to papers w, x, y in his papers, then the entry $S[A, B]$ of this matrix would be 2.

1.2 Distance Matrices

The entries in a distance matrix are inversely related to similarity; a larger entry signifies a greater distance or lesser similarity between entities. We construct a distance matrix from a similarity matrix through the following transformation:

Let S be the similarity matrix where $S[i, j]$ is the similarity between professors i and j. The corresponding distance matrix D is computed as:

$$D[i, j] = \begin{cases} 0 & \text{if } i = j, \\ \frac{1}{(S[i,j])^r} & \text{if } i \neq j \end{cases}$$

where r is a parameter that adjusts the rate of conversion from similarity to distance. The transformation is applied element-wise to the non-diagonal entries of S, and the diagonal is set to zero, reflecting that the distance of every professor to themselves is zero.

Heatmap for Distance Matrices:

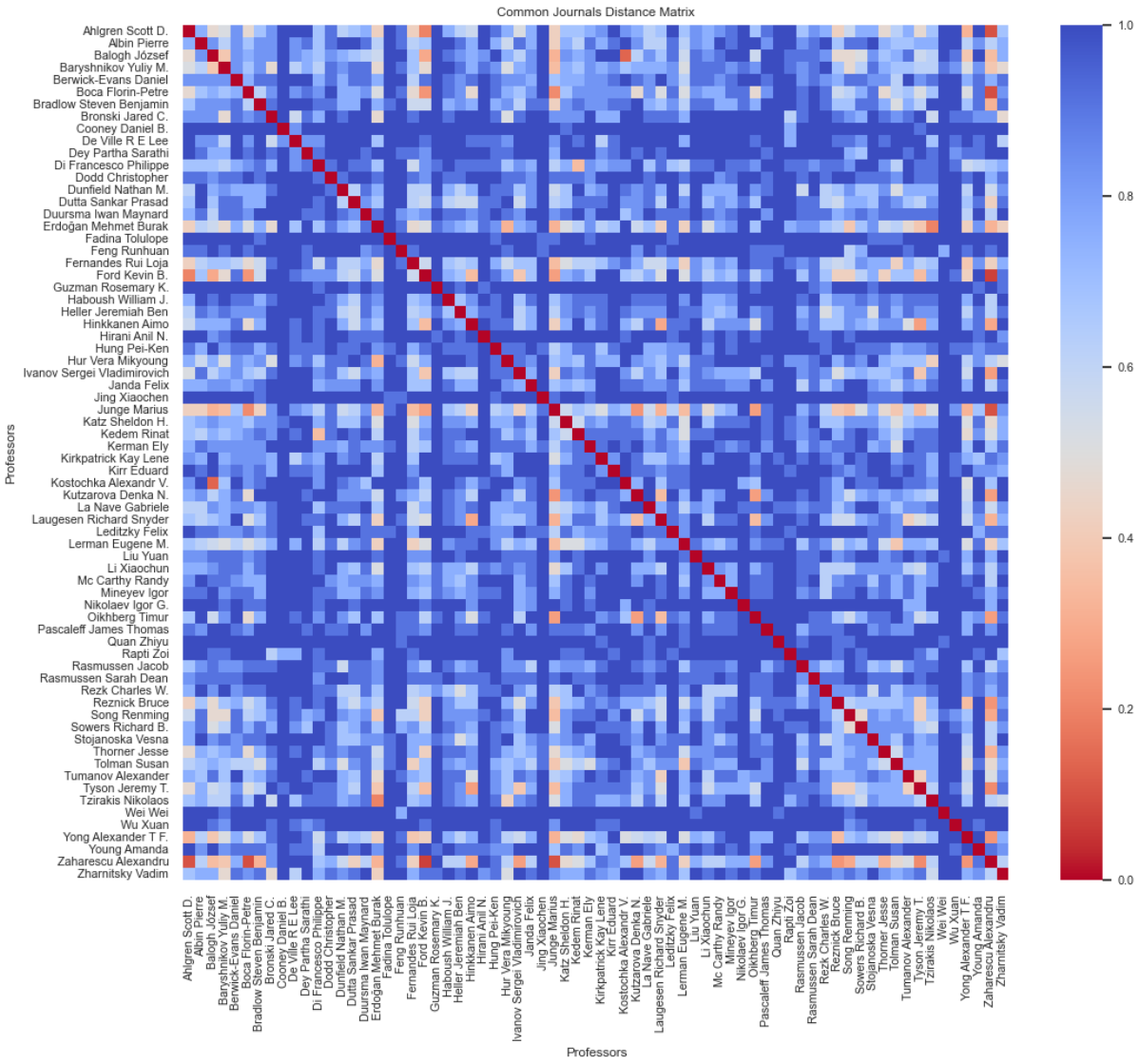

Figure 1: Heatmap for Journal Distance Matrix

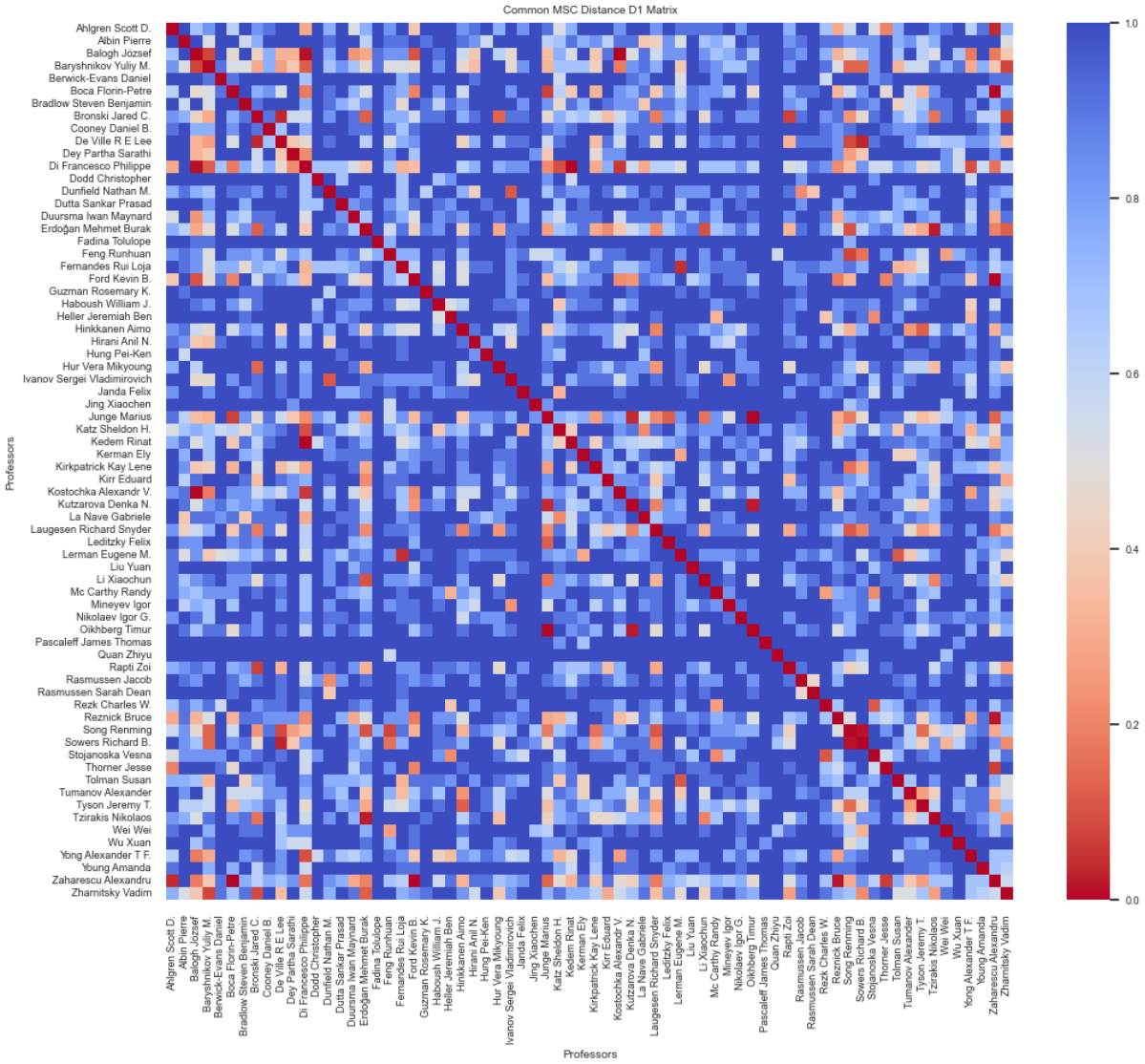

Figure 2: Heatmap for MSC Distance Matrix

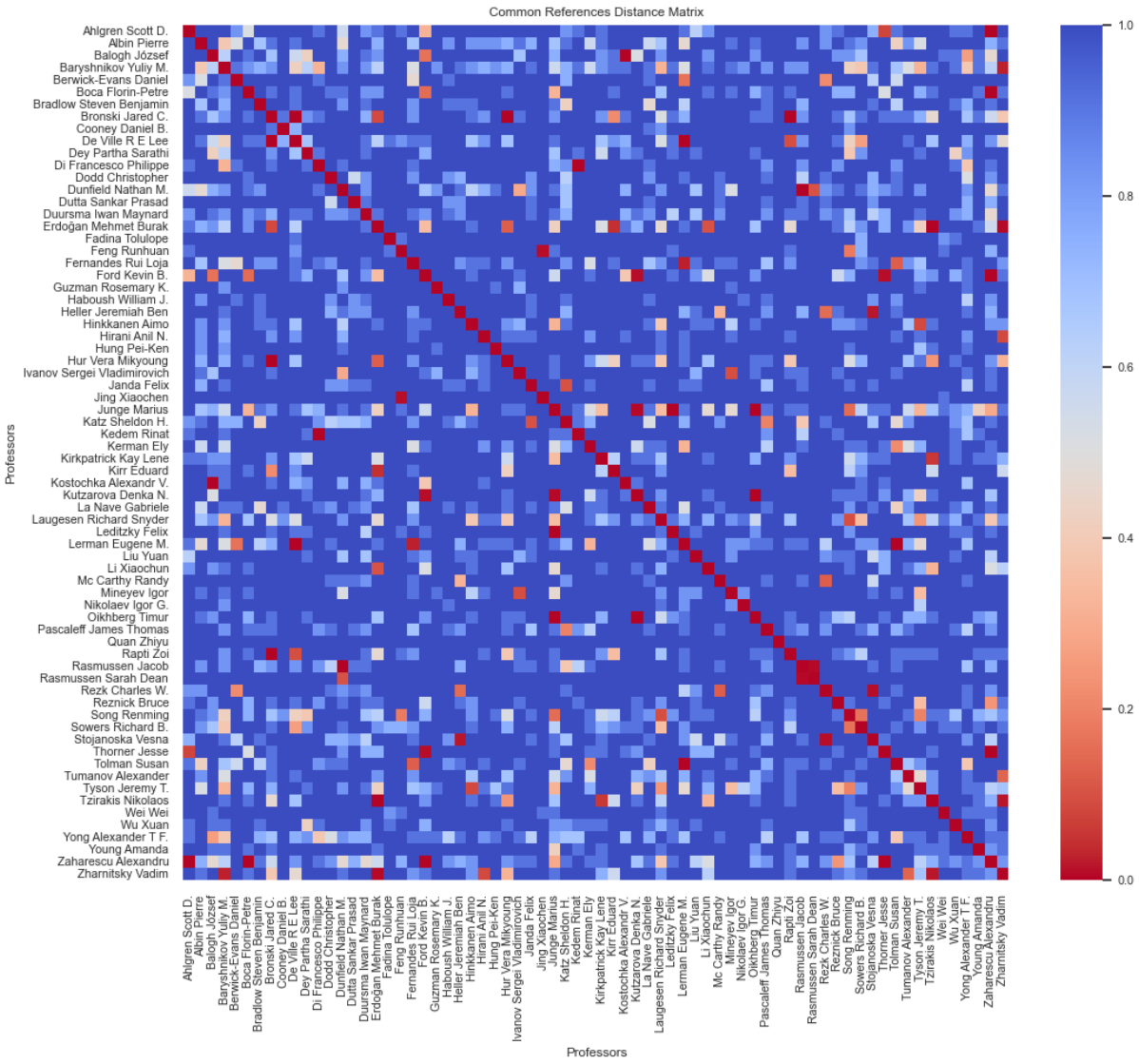

Figure 3: Heatmap for Reference Distance Matrix

2 Methodology

The general process involves the following steps:

1. Download data from the MathSciNet website and save it in JSON format files using the following sub-steps:

    a. Save netid and user password to the `credential.py` file. Inspect ids for buttons on the login pages, use the `.click()` function and `.sleep()` from the webdriver package to perform auto sign-in.

    b. Inspect the id of the search bar, search for publications of math professors using the `.click()` and `.sleep()` functions.

    c. Click on each publication to download the information of references, MSC Code, and Journal for the corresponding paper.

d. Save the information for each professor in a JSON file named `[professor]_papers.json` (e.g., `TolmanSusan_papers.json`) in the following format:

```
{
"PaperA": {"Title", "PaperID", "Author", "
    Journal_Name", "Publication_Year", "
    References", "Codes"},
"PaperB": {...},
...
}
```

where "Codes" refers to the MSC category code.

2. Extract raw data on MSC codes, References, and Journals for each professor in the department:

   a. Extract the "Journal_Name" index for Journals, "References" index for References, and "Codes" for MSC Codes from the `[professor]_papers.json` files.

3. Compare the data between each professor to quantify their similarity, and then store the data into a similarity matrix or distance matrix.

4. Generate the initial phylogenetic trees for each distance matrix using hierarchical clustering, along with their corresponding heat maps.

5. Use the `pathtrees.py` file in the pathtrees repository to iteratively generate the barycenter of the trees:

   a. Initialize with `numpathtrees = 3` and the starting trees MSC and REF. The output is the first pathtree.

   b. Take the output from the previous step, increase `numpathtrees` to 4, and include the Journal tree to generate two pathtrees.

   c. For $N > 0$ iterations:

      i. Input `numpathtrees = 3 + N` and the last output tree, alternating with the trees MSC, REF, and Journal depending on the iteration stage.

      ii. The final output after N iterations will be the barycenter of the trees.

   This iterative process can be repeated to find increasingly precise barycenters.

6. Aggregate the iterations into a single file to transform into visualized trees, with the final iteration representing the current barycenter, using tree plotting code.

The methodology starts with the generation of similarity and distance matrices from the provided data, followed by the creation of phylogenetic trees using agglomerative clustering techniques. The trees are then converted into Newick format for further analysis. The method uses the Java package GTP (Owen and Provan., 2011) to generate the geodesic between pairs of trees, pathtrees for generating intermediate trees on the shortest path between two arbitrary trees in the Python package Pathtrees and several other standard Python modules. Using iteration to find reasonable median, the barycenter, of three trees.

2.1 Phylogenetic tree and Newick Format tree

A phylogenetic tree is a diagrammatic representation that showcases the evolutionary relationships among various biological species or entities based upon similarities and differences in their physical or genetic characteristics. The structure of a phylogenetic tree is a branching system where each branch point, or node, represents the common ancestor of the species diverging from that point. The length of the branches can reflect the genetic changes or chronological time, depending on the type of phylogenetic tree constructed. These trees are constructed through the analysis of morphological or genetic data, where algorithms and statistical models determine the likelihood of evolutionary paths.

The Newick format is a way of representing graph-theoretical trees with a specific syntax in a text form, primarily used to denote phylogenetic trees. In the Newick format, trees are represented as a nested set of parentheses with branch lengths specified by numbers. Each set of parentheses corresponds to a node, and the enclosed leaf nodes and subnodes represent descendants of that node. Leaves in the tree are typically represented by the name of the data item—like the species or genes they represent—while internal nodes may be named or left blank if the focus is on the tree's structure. For example, a simple tree with three descendants might be represented as ((A:0.1,B:0.2):0.3,C:0.4);, where A, B, and C are the leaf nodes, and the numbers represent branch lengths.

2.2 Billera-Holmes-Vogtmann (BHV) treespace and shortest paths

The Billera-Holmes-Vogtmann (BHV) treespace is a concept from the mathematical field of geometric phylogenetics, which integrates geometry with evolutionary biology. In mathematics, the BHV treespace is defined as a space where each point represents a phylogenetic tree, a tree that depicts the evolutionary relationships among a set of species or taxa.

The Geodesic Treepath Problem (GTP) algorithm is a polynomial algorithm to compute the geodesic distance between two phylogenetic trees, which was introduced by Billera et al. (2001).

In BHV treespace, there exists a unique shortest path, or geodesic, that joins two phylogenetic trees T1 and T2. This geodesic may be calculated using the Geodesic Treepath Problem (GTP) technique (Owen and Provan., 2011).

2.3 Iteration and Barycenter

We offer a method to build intermediate trees on the shortest path between two arbitrary trees, called pathtrees, based on the Billera-Holmes-Vogtmann (BHV) distance between pairs of trees. These pathtrees give a structured way to investigate intermediate neighboorhoods between trees of interest in the BHV tree space. We implemented our algorithm in the Python package PATHTREES to process these trees. This tool facilitates an iterative refinement process, enabling us to approach the true barycenter with each iteration. The process is documented comprehensively, with iterations aggregated into a single file that is then transformed into visualized tree structures. These visualizations are crucial for interpreting the progressive accuracy of our barycenter calculations.

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

Require: $N > 0$

$MSC \leftarrow$ MSC Tree
$REF \leftarrow$ Reference Tree
$JOUR \leftarrow$ Journal Tree
$numpathtrees \leftarrow 3$
$OutPut0 \leftarrow$ internalpathtrees(MSC, REF, $numpathtrees$)
$numpathtrees \leftarrow 4$
$outPut1 \leftarrow$ internalpathtrees($OutPut0$, $JOUR$, $numpathtrees$)
while $N > 0$ **do**
    $numpathtrees \leftarrow 3 + N$
    $outPutN \leftarrow$ internalpathtrees(MSC, REF, $JOUR$, $numpathtrees$)
    $N \leftarrow N - 1$
end while

3 Results and Future work

After completing 17 iterations, our algorithm has yielded an initial barycenter of the phylogenetic trees. This barycenter is a significant stride towards decoding the intricate academic focus areas within the department. Future enhancements may involve the integration of additional perspectives, coupled with an increased number of iterations, to achieve a closer convergence to the true barycenter. Such advancements could unveil deeper structural insights and connections within the UIUC Mathematics Department.

The 3 original phylogenetic trees for journal, MSC and reference.

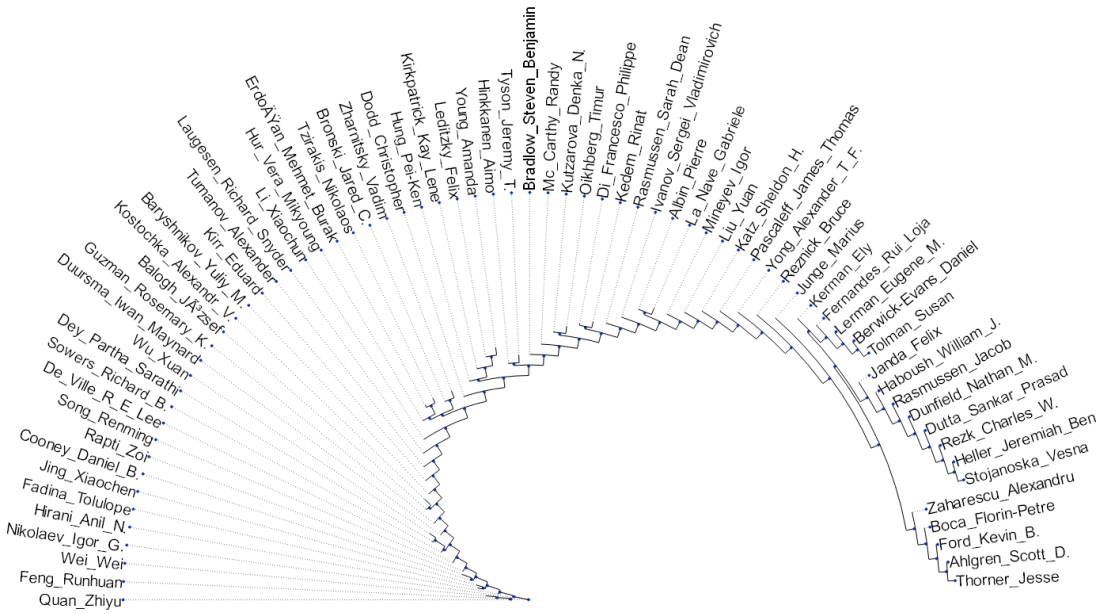

Figure 4: Journal phylogenetic trees

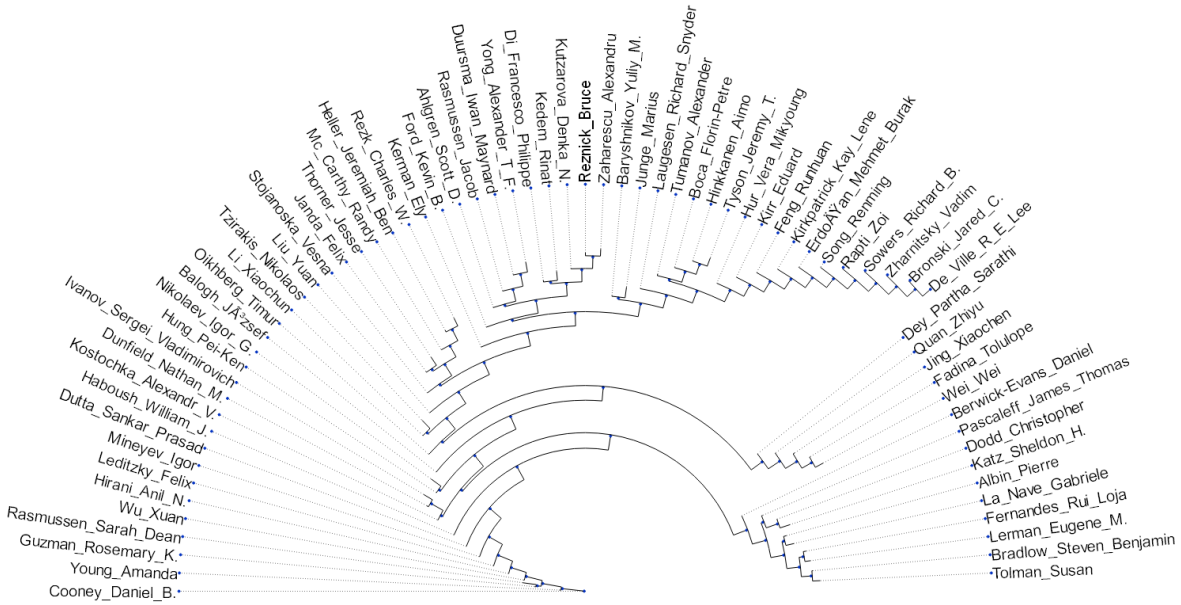

Figure 5: MSC phylogenetic trees

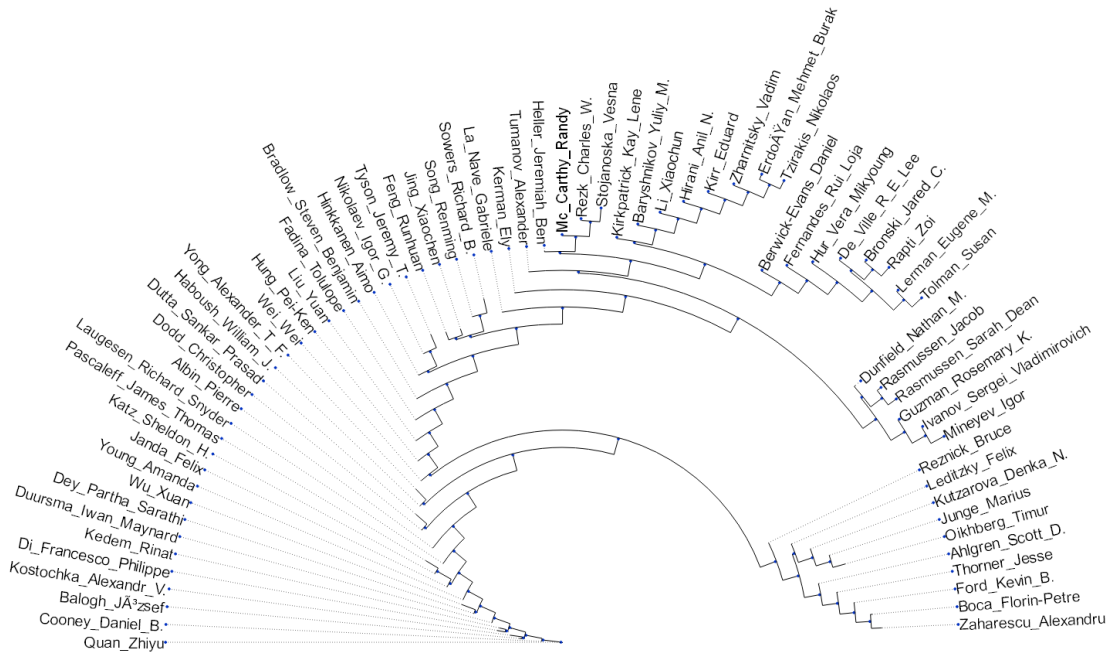

Figure 6: Reference phylogenetic trees

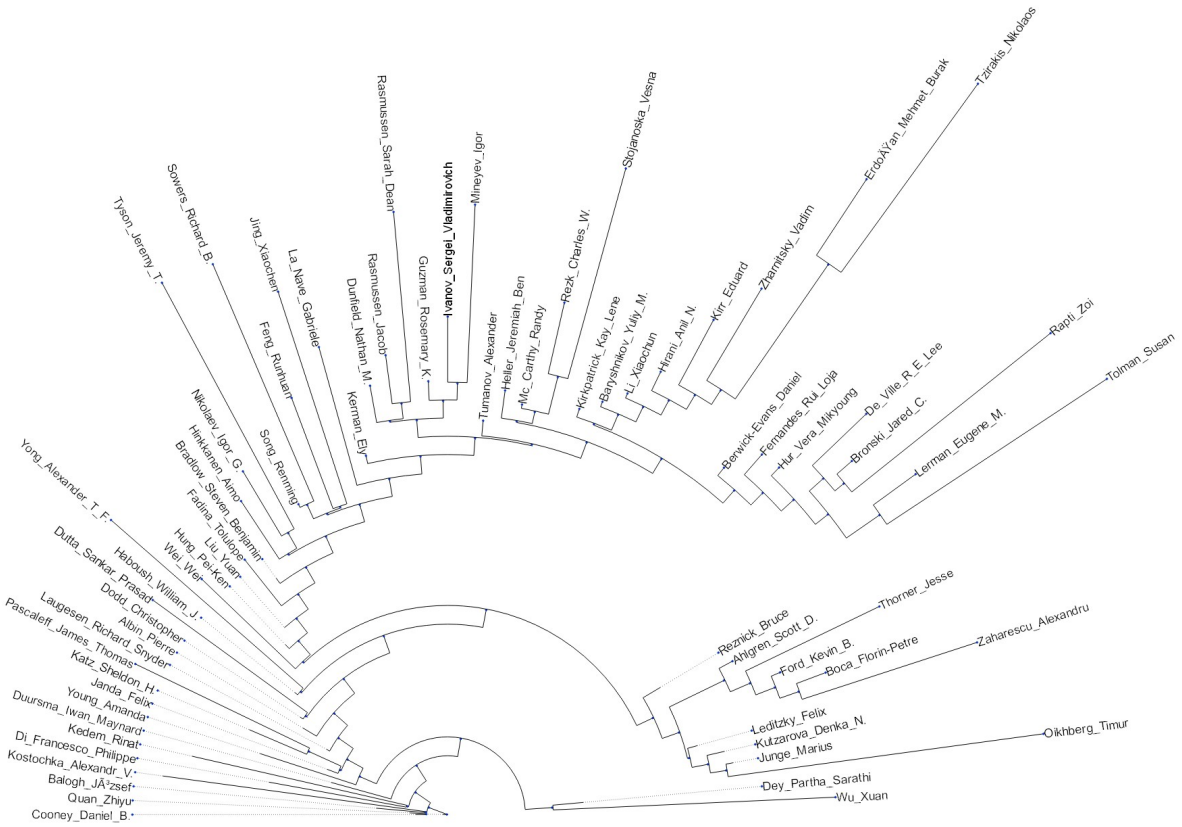

Figure 7: final barycenter after 17 iterations

References

Billera, L., Holmes, S., and Vogtmann., K. (2001). Geometry of the space of phylogenetic treesn. *Advances in Applied Mathematics*, 27:733–767.

Owen, M. and Provan., J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans Comput Biol Bioinf*, 8(1):2–13.