Phylogeny and Metadata Network Database Pipeline for Epidemiologic Surveillance

Garrick Stott[a], Leke Lyu[a], Gabriella Veytsel[a], Jacky Kuo[b], Ryan Lewis[b], Armand Brown[c], Kayo Fujimoto[b], and Justin Bahl[a]

a. Institute of Bioinformatics, Department of Infectious Diseases, Department of Epidemiology and Biostatistics, Center for the Ecology of Infectious Diseases, University of Georgia, Athens, GA
b. Department of Health Promotion Behavioral Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX
c. Bureau of Epidemiology, Houston Health Department, Houston, TX

Abstract

The ongoing SARS-CoV-2 pandemic has highlighted the difficulties in integrating disparate data sources for epidemiologic surveillance given an increasing volume of data and recurring analyses. To address this challenge, we have developed a Nextflow workflow with builds a graph database to integrate phylogenetic trees, sample metadata, and community surveillance data for phylodynamic inference. As an example use case, we used 8,568 samples collected over the course of the SARS-CoV-2 pandemic in Houston, TX with contact tracing available. We split the data into one-week intervals, simulating the influx of data over the course of an ongoing outbreak. We then generated maximum likelihood trees from the partitioned datasets and inferred a potential transmission network using a forest of minimum spanning trees built on the patristic distances between samples.

Objective

To build a portable pipeline for phylogeny generation and enable analyses across multiple phylogenies over time.

Methods

Data

We downloaded 8,568 samples from GISAID collected over the course of the SARS-CoV-2 pandemic (May 2020-February 2022) in Houston, TX with some degree of contact tracing available through the Houston Department of Health. From these 8,568 samples, 167 individuals were known to be exposed at a specific venue, with 2 individuals exposed at two different venues, and no two samples are known to have been connected to each other through these venues. In addition, to help determine the best patristic distance threshold, we built a larger tree on all samples from Texas over the Delta wave of the pandemic (March-October 2021) to analyze at what size communities merge with one another.

Building Phylogenetic Trees

We used MAFFT to align these sequences to the Wuhan/Hu-1/2019 reference strain. To generate our phylogenetic trees for downstream analysis, we used both IQ-Tree 2 and FastTree in our pipeline but leveraged IQ-Tree 2 for the example analysis discussed here. We used a GTR model with 1000 ultrafast bootstrap replicates to generate each of the 76 trees, rooting to the Wuhan/Hu-1/2019 reference strain.

Phylogeny Storage

Each tree can be stored in the graph using a Tree Alignment Graph (TAG), a method developed as part of the Open Tree of Life Project, demonstrated in Figure 1. We then calculate the patristic distances between each leaf using the ape package in R. Users also have the option to generate these distances from phylogenies loaded into the graph. The full graph schema is in Figure 2.

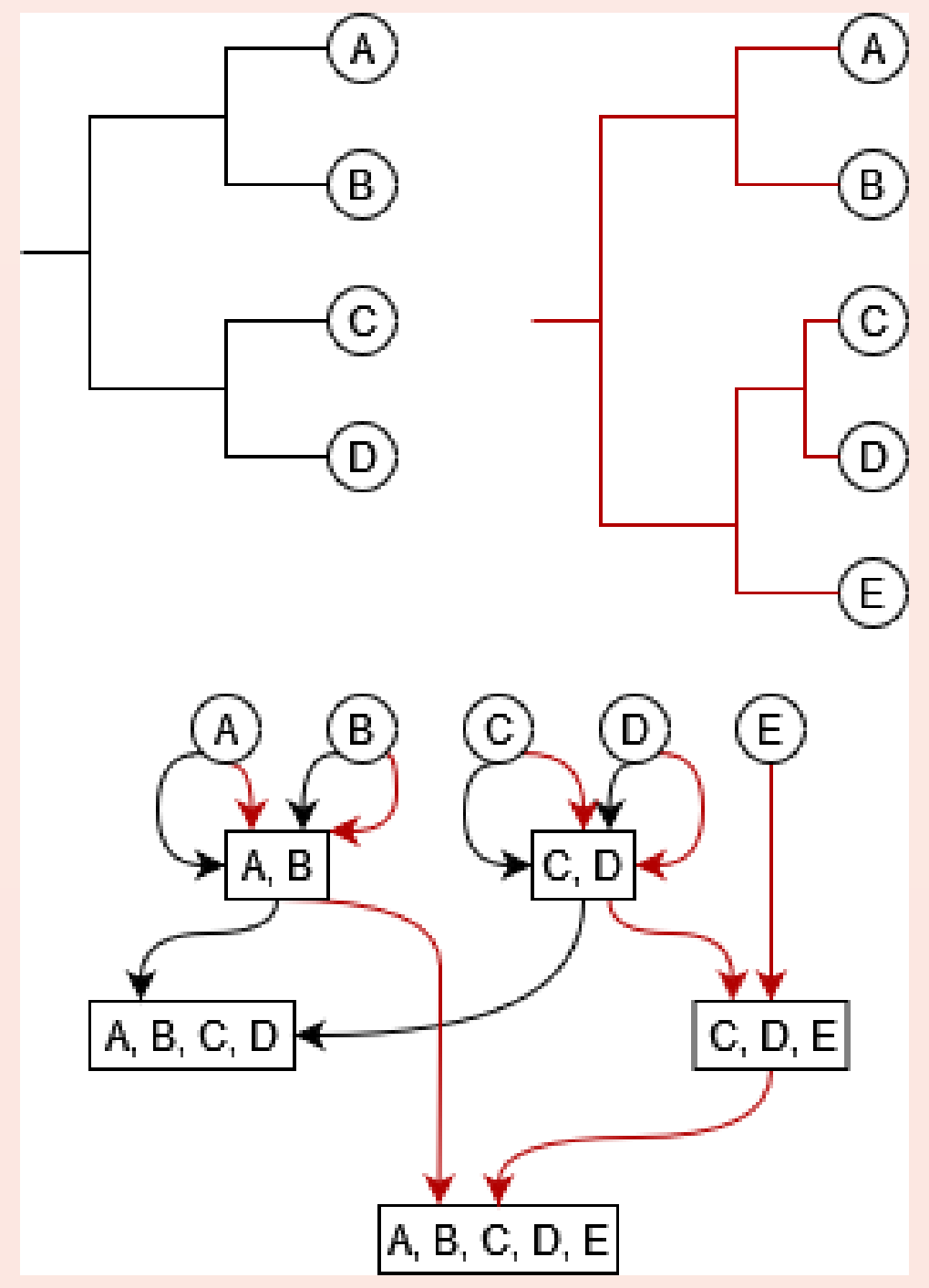
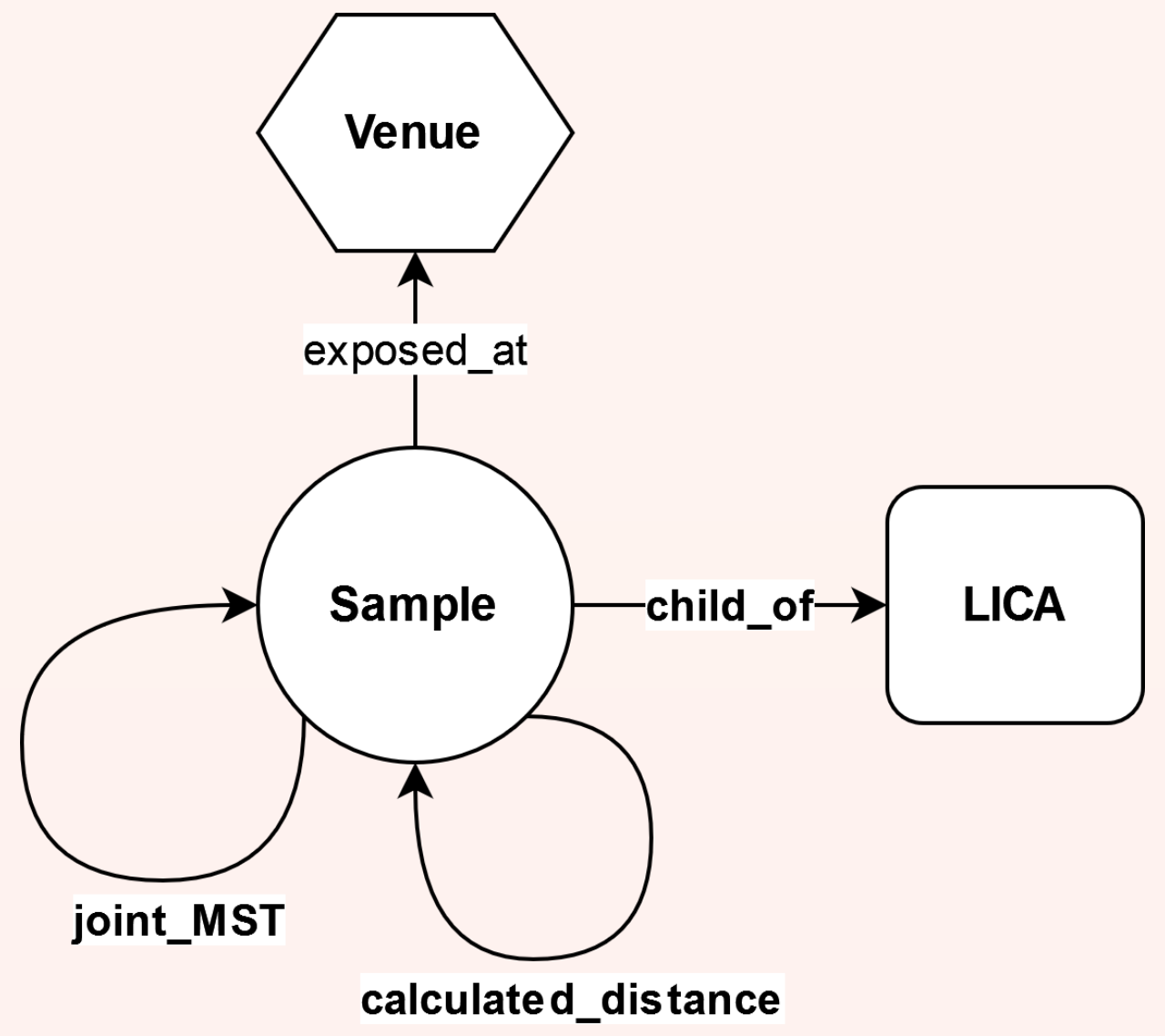


Fig. 1. Tree Alignment Graph. We leverage Tree Alignment Graphs (TAG) to store the phylogenetic trees in our database. These networks allow for efficient storage of phylogenies in the graph and provide a means to extract both the original trees as well as subsets or combinations of the trees.

Fig. 2. Graph Database Schema. Schema allows for multiple separate, but related, networks to be overlayed and queried together, for example, TAGs for phylogenetic trees, venue affiliation networks, and known contact networks.



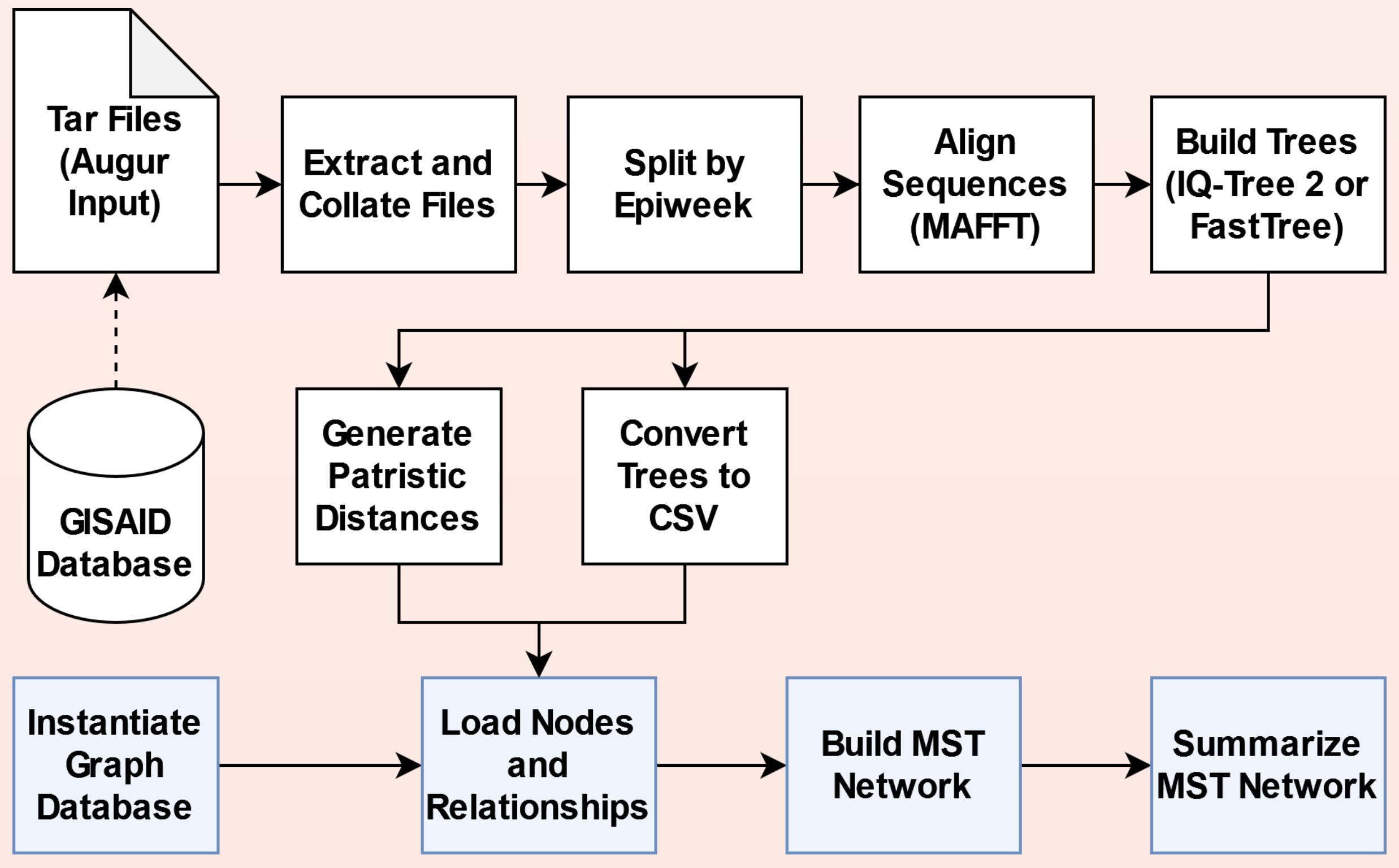
Venue
exposed_at
Sample
child_of
LICA
joint_MST
calculated_distance

Minimum Spanning Trees

Finally, as our example use case, we built 2000 minimum weight spanning trees across each of the patristic distance networks, filtering on patristic distances less than 6 months. While pairwise genetic distance is often used instead due to its low computational cost, patristic distances can use more of the information available in the sequence alignment, allowing for easier differentiation between rapidly and slowly evolving sites. After generating a network of minimum weight spanning trees, we summarize them into joint edges to build our inferred transmission network.

Nextflow Pipeline

To make this workflow and modular for future development, we built a Nextflow pipeline that will generate a graph database from a directory of .tar files from GISAID, .fasta sequence files, or a collection of Newick or Nexus formatted tree files. In addition, one can opt to generate trees either using FastTree or IQ-Tree 2 and skip storage of phylogenetic trees if they only plan to use the patristic distance network for their analyses. A diagram showing the complete workflow is shown below in Figure 3.



Tar Files (Augur Input)
Extract and Collate Files
Split by Epiweek
Align Sequences (MAFFT)
Build Trees (IQ-Tree 2 or FastTree)
GISAID Database
Generate Patristic Distances
Convert Trees to CSV
Instantiate Graph Database
Load Nodes and Relationships
Build MST Network
Summarize MST Network

Fig. 3. Nextflow Pipeline Design. Our Nextflow pipeline takes in a set of .tar files in the format used by GISAID, merges these files, and separates them (cumulatively) by CDC Epi Week before running them through the alignment and tree-building process. In blue are steps completed within the Neo4j graph database.

Results

We began our analysis with a slightly larger dataset covering all of Texas in order to observe how the number of communities changes with our patristic distance threshold. We observe a peak at around 21 days patristic distance, after which point new clusters stop forming and existing clusters begin to merge, as seen in Figure 4. We use this threshold in subsequent visualizations to highlight edges which are more likely to have originated from the same cluster.

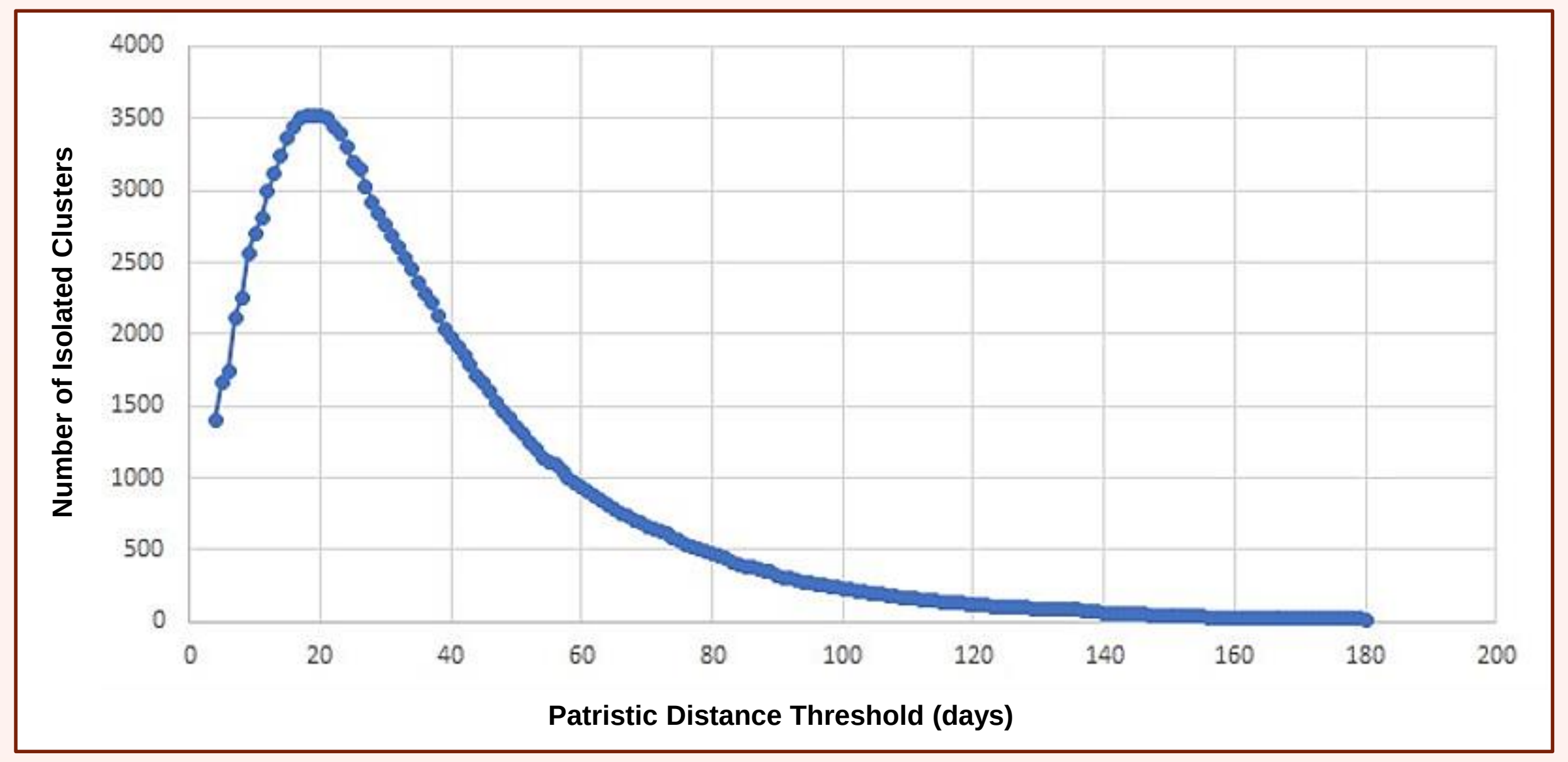


Fig. 4. Number of Isolated Clusters by Patristic Distance Threshold. This plot shows the number of weak components (i.e. isolated clusters) at various thresholds of patristic distance. Cluster count is maximized around 21 days.

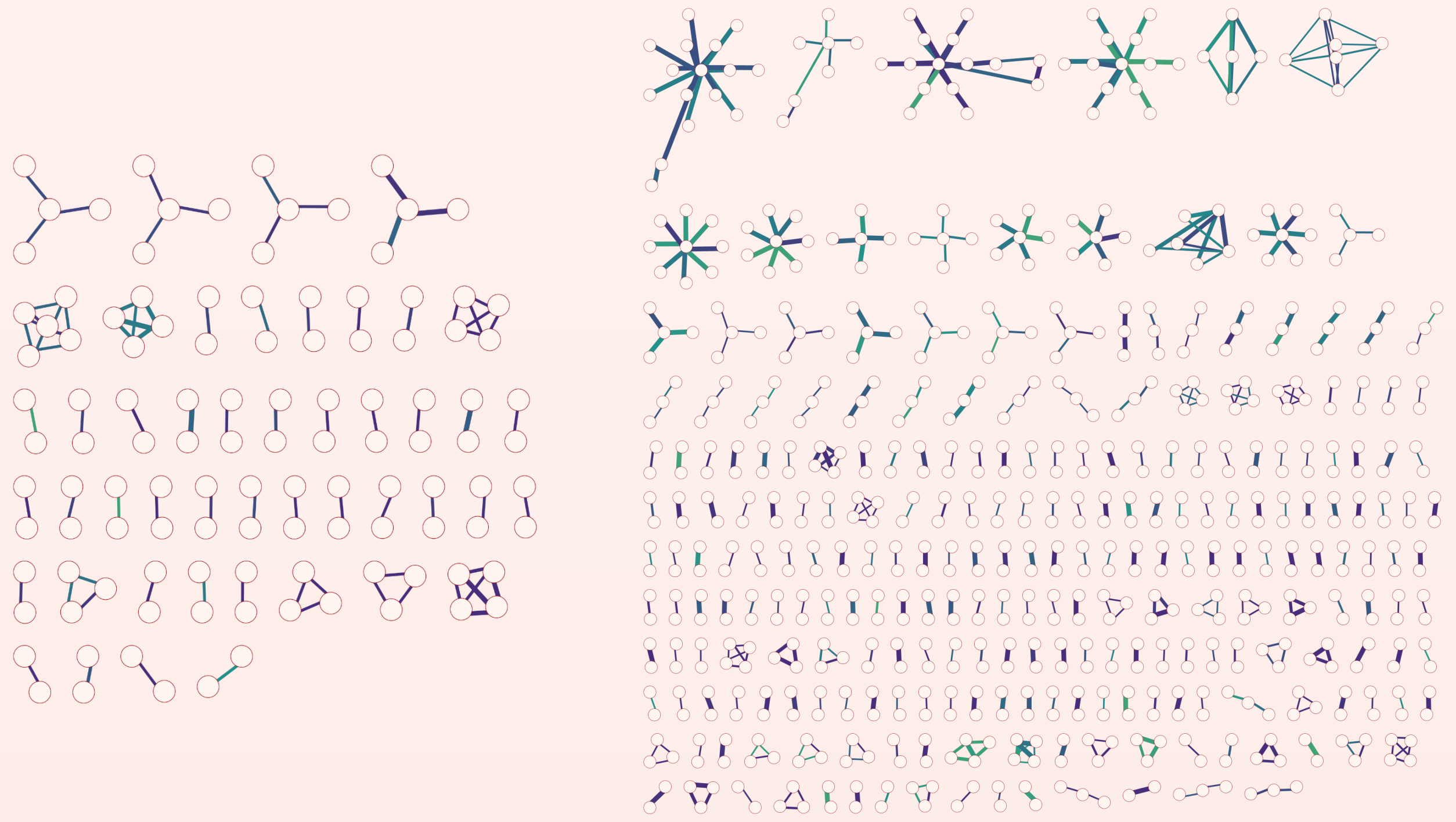


Fig. 5. Joint MST Network (21-day Threshold) for Epi Weeks 202101 (left) and 202201 (right). Edge thickness indicates percentage of MSTs concording with that edge, edge color indicates patristic distance, purple for within 10 days, green for 21 days, and yellow for greater than 21 days.

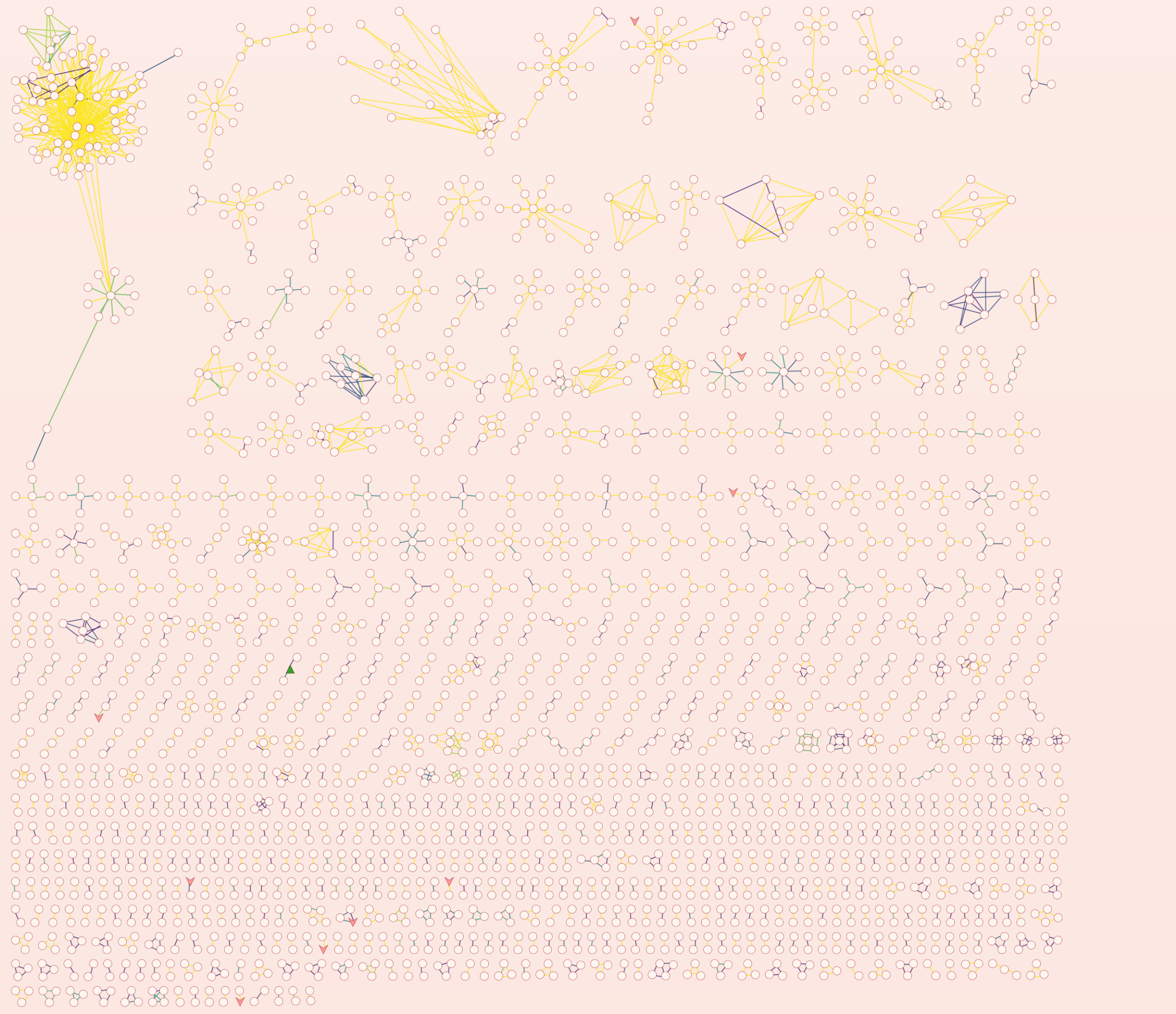


Fig 6. Joint MST Network (Epi Week 202136). Edge thickness indicates percentage of MSTs concording with that edge, edge color indicates patristic distance, purple for within 10 days, green for 21 days, and yellow for greater than 21 days. Red V nodes indicate samples associated with education centers. Small clusters begin to appear around education center samples.

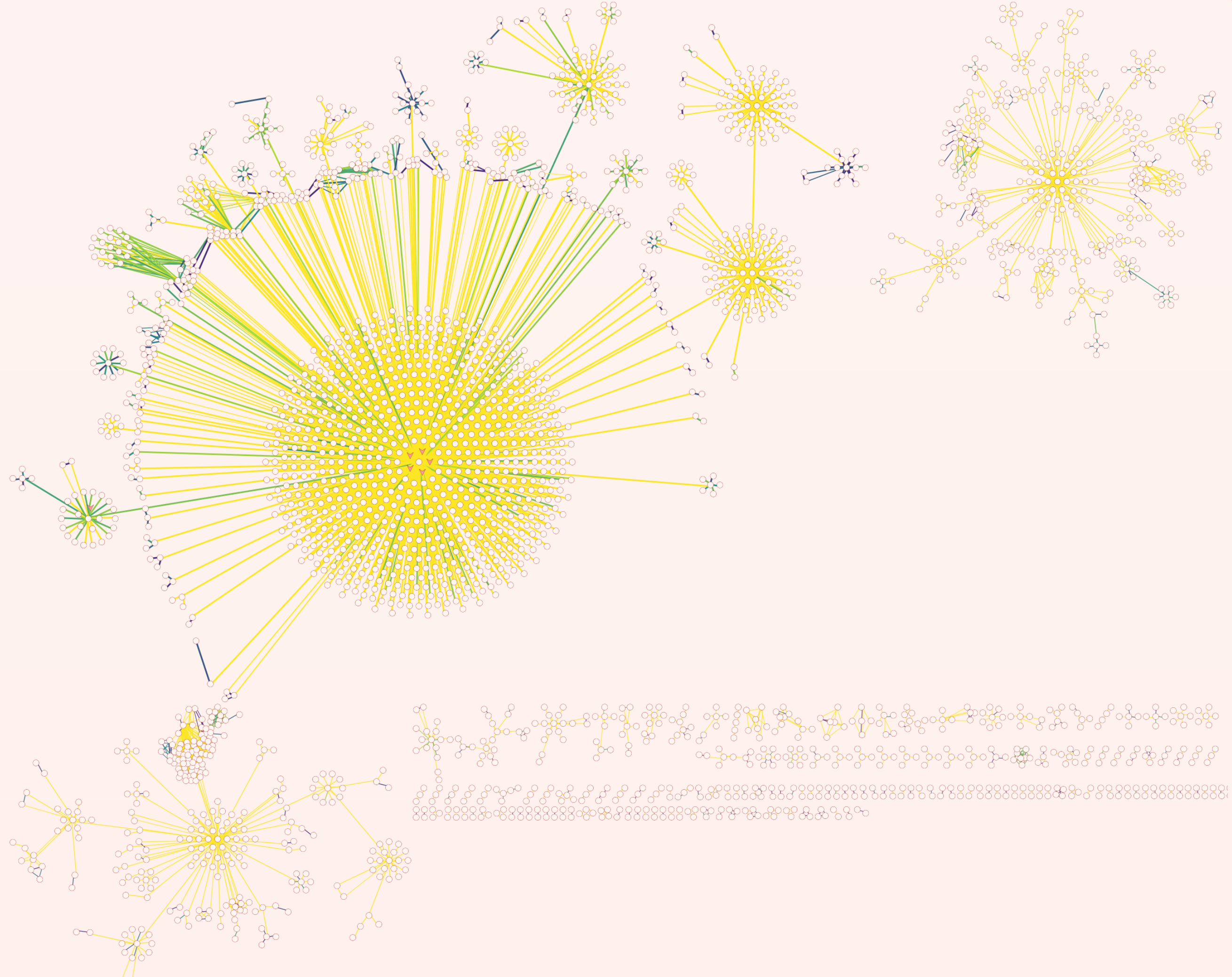


Fig. 7. Joint MST Network (6 Month Threshold) for Epi Week 202201. Edge thickness indicates percentage of MSTs with that edge, edge color indicates patristic distance, purple for within 10 days, green for 21 days, and yellow for greater than 21 days. Red V nodes indicate samples associated with education centers. Education centers are associated with a large hub of smaller clusters.

We then visualized the joint MST networks, focusing on edges with a 21 day or less patristic distance since after this point, it becomes difficult to distinguish significant results from noise. These closest relationships are most clearly shown in Figure 5, comparing networks generated in early 2021 and 2022 respectively. In our last two figures, we highlight an interesting relationship between nodes associated with education centers and larger clusters. Starting in late August, we see networks begin to form around these nodes (Figure 6). They rapidly become significant hubs (albeit with significant distance between clusters) by the start of the following year (Figure 7).

Conclusions

We have developed a Nextflow pipeline which is able to quickly build a Neo4j graph database containing phylogenetic trees and their associated metadata. While this example analysis could be done without necessitating a graph database, we benefit from the scalability and portability of this approach. The graph database offers a plethora of built-in tools useful in public health, such as community detection, centrality measures, and link prediction, enabling future work. In this brief example analysis, we observed a patristic distance threshold of 21 days is most effective for maximizing cluster size in a larger dataset. We also highlight qualitatively the importance of education centers in driving new hubs for the spread of SARS-CoV-2. Future work will focus on quantifying this relationship. In addition, we hope to standardize the formatting of new graph schemas, build native visualization tools, and design new methods for cluster identification and stability measures.

Acknowledgements and Contacts

Data sourced from the GISAID EpiCov Database. This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93021C00018 (NIAID Centers of Excellence for Influenza Research and Response, CEIRR), Centers for Disease Control and Prevention under contract No. 75D30121C10133

Source code on GitHub: https://github.com/glstott/PMeND

Preprint associated with this work: https://doi.org/10.1101/2022.04.19.488067

If you have questions, please contact: Garrick.stott@uga.edu