PhyloView: A System to Visualize the Ecology of Infectious Diseases using Phylogenetic Data

Minh Tri Le
George Mason University, USA
mle35@gmu.edu

David Attaway
ESRI, USA
DAttaway@esri.com

Taylor Anderson
George Mason University, USA
tander6@gmu.edu

Hamdi Kavak
George Mason University, USA
hkavak@gmu.edu

Amira Roess
George Mason University, USA
aroess@gmu.edu

Andreas Züfle
George Mason University, USA
azufle@gmu.edu

2022 23rd IEEE International Conference on Mobile Data Management (MDM) | 978-1-6654-5176-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/MDM55031.2022.00051

***Abstract*—Since the onset of the COVID-19 pandemic, millions of coronavirus sequences have been rapidly deposited in publicly available repositories. The sequences have been used primarily to monitor the evolution and transmission of the virus. In addition, the data can be combined with spatiotemporal information and mapped over space and time to understand transmission dynamics further. For example, the first COVID-19 cases in Australia were genetically related to the dominant strain in Wuhan, China, and spread via international travel. These data are currently available through the Global Initiative on Sharing Avian Influenza Data (GISAID) yet generally remains an untapped resource for data scientists to analyze such multi-dimensional data. Therefore, in this study, we demonstrate a system named Phyloview, a highly interactive visual environment that can be used to examine the spatiotemporal evolution of COVID-19 (from-to) over time using the case study of Louisiana, USA. PhyloView (powered by ArcGIS Insights) facilitates the visualization and exploration of the different dimensions of the phylogenetic data and can be layered with other types of spatiotemporal data for further investigation. Our system has the potential to be shared as a model to be used by health officials that can access relevant data through GISAID, visualize, and analyze it. Such data is essential for a better understanding, predicting, and responding to infectious diseases.**

***Index Terms*—COVID-19, phylogeny, genomic sequence, spatiotemporal data, data science**

I. Introduction

More than 8.7 million SARS-CoV-2 genome sequences from 241 countries and territories have been shared publicly as of late February 2022 and made available to research communities through a repository called the Global Initiative on Sharing Avian Influenza Data (GISAID) [22], [29]. Originally an initiative to foster the collaborative sharing of influenza virus data, GISAID is now a trusted and widely-used platform for the rapid sharing of data from all influenza viruses and SARS-CoV-2, including genetic sequences as well as related clinical, epidemiological, and geographical data [10].

A phylogenetic tree is a hierarchical visualization of the pathway through evolutionary time from a common ancestor to its genetic decedents [1]. Figure 1 shows an example of a phylogenetic tree for GISAID data obtained for Louisiana, USA [12], [13]. The location of the sequenced cases is generalized to different geographic scales, but is most detailed for cases that fall within Louisiana, in which case the location of the sequenced case is identified at the parish level (similar to a US county). An ancestor or descendent node that falls outside Louisiana may be generalized to the state, country, or continent level. The root of this tree corresponds to the first observed case of COVID-19 in Wuhan, China, and is the only node in this data set corresponding to the city level.

Fig. 1. A phylogenetic tree to visualize evolutionary relationships. Each node corresponds to an observed and sequenced case. Edges correspond to inferred most recent common ancestor of a node. Nodes contain spatial and temporal information which PhyloView leverages for visual analysis.

Since the phylogenetic data of the virus contain both spatial and temporal data, they can be mapped over both space and time. Figure 2 shows the phylogenetic tree data as described above, but transposed into geographic space. Leveraging such phylogenetic information, which includes not only information about "where" but also "from-where", "to-where", and "when" enables much greater insights into understanding the spread and evolution of viruses, which may improve our preparedness for future diseases.

While phylogenetic data are currently available through GISAID, it generally remains an untapped source for data scientists who can analyze such multi-dimensional data. Therefore, the goals of this demonstration paper are to 1) show the potential of combining phylogenetic epidemiology with spatiotemporal data science to understand the spread and

2375-0324/22/$31.00 ©2022 IEEE
DOI 10.1109/MDM55031.2022.00051

evolution of diseases, 2) raise awareness of the availability and dimensions of phylogenetic data to the data science and engineering community, 3) provide a system, namely PhyloView (powered by Esri ArcGIS Insights [6]) that has the potential to support local health departments and researchers across the world for rapid analyses of COVID-19 transmission in space and time, and 4) help establish new research directions that combine phylogeny and data science to find hotspots, to identify key nodes in phylogenetic disease networks, and to leverage this understanding to prepare for future epidemics and pandemics.

After surveying the related work on data-driven understanding of disease spread in Section II, we formally define spatiotemporal phylogenetic data as used by PhyloView in Section III. Section IV describes how spatiotemporal phylogenetic data can be obtained for analysis through public data repositories. In Section V, we describe various visualization approaches used for understanding different aspects of the genome data. Our proposed PhyloView system which leverages and connect these visualization approaches is described in Section VI. In Section VII we present applications of PhyloView while Section VIII concludes the paper.

II. Related Work

Since the onset of the COVID-19 pandemic, the spatiotemporal data science community has made tremendous advances to improve our understanding of the spread of the COVID-19 virus [34], [35], [31]. Solutions have been proposed to map the change of mobility in response to the pandemic [9], to detect clusters of cases [17], to improve contact-tracing solutions [23], [32], to predict new cases over space and time [2], and to develop data-driven simulations of disease spread [26], [33]. The works commonly use fine-grained human mobility data and COVID-19 incident data. Phylogenetic data offer an opportunity to enrich COVID-19 incidence information with flow information by leveraging empirical observations that directly capture the evolution and spread of the disease. Yet, to the best of our knowledge, no existing work in data mining and data engineering communities uses any phylogenetic data.

At the same time, studies in epidemiology and biology have analyzed the evolution of COVID-19 and transmission patterns [4], [18], [16]. These studies focused mainly on the genealogical aspects of the virus, the evolution of the virus towards different strains, the effect of mutations on the human immune response, and the epidemiologic relationships between cases. Other work has focused on understanding the evolution of COVID-19, in space and time, using phylogenetic data [7], [28], but to our knowledge no studies provide the opportunity to explore multiple dimensions (strain, phylogeny, spatial-temporal patterns) of the data concurrently.

III. Spatiotemporal Phylogenetic Data

In the same way that DNA paternity testing can determine the (likely) biological parents of a human, a phylogenetic analysis of the genetic sequence of the RNA (ribonucleic acid) of a virus like SARS-CoV-2 can determine the (likely) most recent ancestor of the virus [3]. The reason we cannot directly infer

Fig. 2. Geographic representation of COVID-19 phylogeny.

the direct "biological parent" of a virus is that only a small fraction of observed viral cases are sequenced. For example, from January 2020 to November 2021, 4,963,718 sequences have been shared via GISAID. Of these, the USA has shared 1,526,198 sequences, the highest number of sequences shared globally. Yet, this effort captures only 3.31% of the reported cases in the US [11]. An overview of the coverage of COVID-19 genome sequencing in different area across the globe is found in [24].

Thus, when a case of COVID-19 of a person Z is sequenced, the direct parent of this case (from the person Y that exposed Z) may not have been sequenced and thus, may not be included in the database. The same may apply for the person X that exposed Y, and so on. The most recent ancestor refers to the youngest ancestor of Z, which has been sequenced (and thus, is contained in the database).

In addition to purely phylogenetic data based on RNA sequencing, public databases also provide location information of an observed case and time information. We formally define a spatiotemporal phylogenetic database as follows.

Definition 1 (Spatiotemporal Phylogenetic Database): Let $\mathcal{C} = \{c_1, ..., c_N\}$ be a collection of N cases and let $c_i = (t_i, s_i)$ where $t_i \in \mathcal{T}$ is a time-stamp from a time domain $\mathcal{T}$ (such as daily dates since Jan. 1st 2020) and $s_i \in ([-90, 90], [-180, 180])$ is a latitude/longitute pair encoding a geolocation. Further, let $\mathcal{P} : \mathcal{C} \mapsto \mathcal{C} \cup \perp$ denote a functional relation that maps each case to their most recent ancestor ($\perp$ in case of the first case, which has no ancestor).

IV. Phylogenetic Data Availability

The GISAID provides free and open access to genomic data of influenza viruses and the coronavirus responsible for the COVID-19 pandemic [30]. As of late February 2022, more than 8.7 million sequenced cases of COVID-19 have been shared, including spatiotemporal and phylogenetic information as described in Definition 1. These data are shared via two separate files: The genomic epidemiology database contains spatiotemporal information for each case, including date, geocoordinates, age, and gender of the reported case. This dataset also includes the sequenced genome of each case. The full phylogenetic tree is shared in the Audacity platform [5] within the GISAID platform, which provides a single text file that can be parsed into a tree representation. This file also includes a table that matches each of the nodes of this tree to a unique

identifier that can be joined with the genomic epidemiology database to link spatiotemporal information. GISAID directly allows users to select spatial regions and time intervals. This enables, for example, local health departments to obtain data for their region, or to support contact tracers with possible origins of an observed case during a specific time of interest.

The goal of this work is to allow users and health officials to rapidly visualize and explore spatiotemporal phylogenetic data obtained from GISAID. Therefore, PhyloView not only sheds light on this data for the spatiotemporal data science and data engineering community but also provides a tool so that researchers and public health partners can rapidly visualize, explore, and analyze the data and gain insight towards its use for better understanding the evolution and spread of the virus.

V. Visual Analysis of Infectious Disease Data

What makes the analysis of phylogenetic data challenging is the multitude of different data types combining spatial, temporal, variant, and genomic information. This section describes various visualization approaches used for better understanding different aspects of the data. Specifically, Section V-A describes how to visualize phylogenetic relationships using a phylogenetic tree. Then, Section V-B presents using a spatiotemporal disease ecology map to understand the spread of a disease in space and time. Section V-C proposes a geographic diseases spread network to understand the spread between regions and Section V-D describes the use of disease strain networks to understand the evolution of a virus. Once these different aspects of visualizing infectious diseases data are described, we will introduce our proposed system PhyloView in Section VI to connect these different visualization approaches for a more holistic view of the spread and evolution of infectious diseases across space and time.

A. Phylogenetic Tree Visualization

We note that the most recent ancestor relationship is acyclic, since for any $(c_i, c_j) \in \mathcal{P}$, denoting that c_j is the most recent ancestor of c_i, it must hold that $t_j < t_i$, that is, the case c_j must have occurred before t_i in time. Since the relation $\mathcal{P}$ is acyclic and maps each case to at most one ancestor (except for a root for which no ancestor is sequenced), the graph is defined by $\mathcal{P}$ is a forest. In the case where all cases have a single common ancestor, that is the case where the first case has been sequenced, the graph defined by $\mathcal{P}$ is a tree rooted in the first case. We call this tree defined by the most recent sequenced ancestor relationship $\mathcal{P}$ a *phylogenetic tree*.

To illustrate the relationship $\mathcal{P}$, Figure 1 shows an example phylogenetic tree for COVID-19 cases sequenced with a geographic focus on Louisiana. In this case, the phylogenetic information yields a single tree that is rooted in the first case observed in Wuhan, China, as their most recent common ancestor. We note that for a specific region or time interval, the phylogenetic information may yield a forest with multiple roots, rather than a single tree, as some cases may not have a single common most recent ancestor.

B. Spatiotemporal Disease Ecology Map

The most recent sequenced ancestor relationship $\mathcal{P}$ allows us to understand the phylogenetic ancestor of an observed case. Phylogenetic databases such as the Global Initiative on Sharing Avian Influenza Data (GISAID) [22], [29] additionally provide the location and time of observed cases. Such spatio-temporal information has been used to understand the location of cases and clusters among them as done in previous studies [27], [8]. But by combining spatiotemporal information with phylogenetic data, we gain information to understand for each observed (and sequenced) case not only where it occurred, but also where it came from. Drawing the cases on a map and connecting them by their phylogenetic relationships $\mathcal{P}$ yields a spatiotemporal diseases ecology map. Figure 2 shows such a spatiotemporal diseases ecology map for the Louisiana data for which the phylogenetic tree is shown in Figure 1. Note that while Figure 2 only shows cases in Louisiana, their phylogenetic ancestry links to cases outside of Louisiana and across the world. Figure 2 lacks the details to gain a deep understanding of how a disease may have spread from/to a specific Parish in Louisiana. What is needed is a system that allows exploring such data by allowing to Zoom in and select specific regions, highlight incoming/outgoing edges from the region, and that allows to "replay" the occurrence of these edges over time to gain an understanding of both spatial and temporal ecology of the diseases. Our proposed system to provide this functionality is described in Section III.

C. Geographic Disease Spread Networks

A different approach for visualizing both the location of cases and the phylogenetic relationship $\mathcal{P}$ between locations is using a geographic disease spread network as shown in Figure 3. In such a network, the nodes of a phylogenetic tree (see Figure 1) are aggregated by region. Thus, a node of a geographic disease spread network corresponds to entire region and the visualized size of the node corresponds to the count of the number of cases within the region. Correspondingly, an edge between two regions corresponds to the number of phylogenetic ancestry relationships between them. It is noted that the resulting network is no longer a tree, as there may now exist cycles in this network. For example, there might be a case in Region A that is the most recent ancestor of a case in Region B and this case may be the most recent ancestor of a (later) case in Region A. We also note that some of the areas from disconnected components appear to contradict the assumption of a fully connected tree $\mathcal{P}$. This is due to some of cases in the GISAID dataset having missing data for their location. These cases are omitted in the geographic disease spread network and the children of these cases (in the phylogenetic tree defined by relation $\mathcal{P}$) have no ancestor in the geographic disease spread network. While interesting, the geographic disease spread network of Figure 3 does not allow us to gain any deeper understanding of the spatiotemporal ecology of disease due to the overwhelming number of nodes and edges. To gain a better understanding, we need a system that allows us to zoom into sub-networks

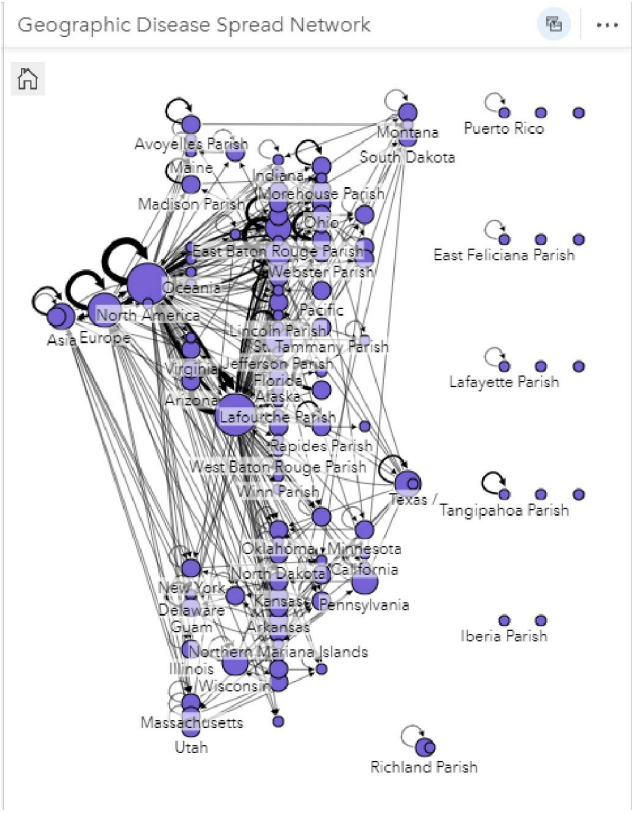

Fig. 3. The Geographic Spread Network for the Louisiana data. Nodes of the network correspond to spatial regions. Links between regions correspond to the most recent sequenced ancestor relationship $\mathcal{P}$. The size of a node corresponds to the number of cases observed in a region and the size of an edge corresponds to the number of observed ancestor relationships between two regions.

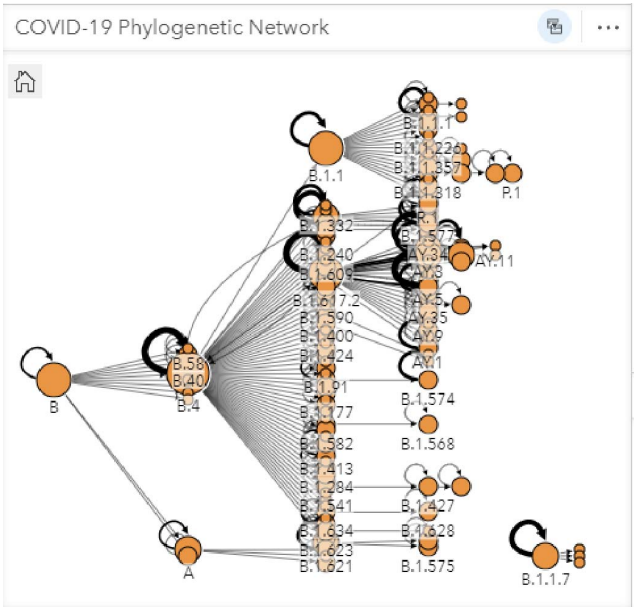

Fig. 4. The Phylogenetic Network for the Louisiana data. Nodes correspond to strains of SARS-CoV-2 variants and mutations (clades) of interest. Links between clades correspond to the most recent sequenced ancestor relationship $\mathcal{P}$. The node size corresponds to the number of cases observed for a clade, and the size of an edge corresponds to the number of observed ancestor relationships between two clades.

of interest, select specific nodes, and filter adjacent nodes. For example, the local health department of a specific parish in Louisiana may want to Zoom into the local subnetwork around the node corresponding to the parish. Our proposed system PhyloView, described in Section III provides this functionality.

D. Disease Strain Evolution Networks

In addition to investigating the spatiotemporal ecology of a disease in a region, it may also be of interest for health officials to understand the different strains of various that emerged (or may emerge) in a region. Figure 4 shows the phylogenetic network for the Louisiana dataset used in previous examples, in which cases are grouped by their Nextstrain clade [14]. Again, we observe disconnected components in the phylogenetic network due to missing information of the clade on some of the cases observed in the GISAID dataset. Interesting in Figure 4 is the Nextstrain clade B.1.1.7, which is, during the time interval depicted, newly emerging. This is the strain that originated in Britain, has since spread across the globe to over 90 countries, and is widely known as the Alpha Variant. Such visualization is interesting for health officials but also for biologists and epidemiologists to understand the spread of different strains of an infectious virus on a global level but also at a local level to protect vulnerable communities. What is needed is a system that allows us to select data for specific communities, visualize the phylogenetic network of the selected community, and select time frames to visually analyze the emergence of strains to improve our understanding of infectious diseases ecology.

VI. PhyloView: System Description

PhyloView allows users to explore any set of spatiotemporal phylogenetic data visually by fusing location analytics with open data science and business intelligence workflows. It enables health departments to visualize data of interest and to understand infectious disease ecology. This visualization empowers analysts of all skill levels to directly connect data, perform advanced analytics, and dissect results for helping health officials to understand lessons learned from the COVID-19 pandemic and better prepare for future pandemics.

In the following Section VI-A, we describe the functionality of PhyloView using data from the state of Louisiana in the United States, including mapping of disease spread, visualization of phylogenetic information, and cross-connection between visualizations to support a better understanding of disease ecology. This demonstration is publicly available at https://insights.arcgis.com/#/view/499be6d89e6e4f0c8289d244dea8a05a. Then, Section VI-C uses a different dataset that, instead of focusing on a single area, visualizes the spread of COVID-19 across the entire globe. Section VI-D describes the underlying data model and

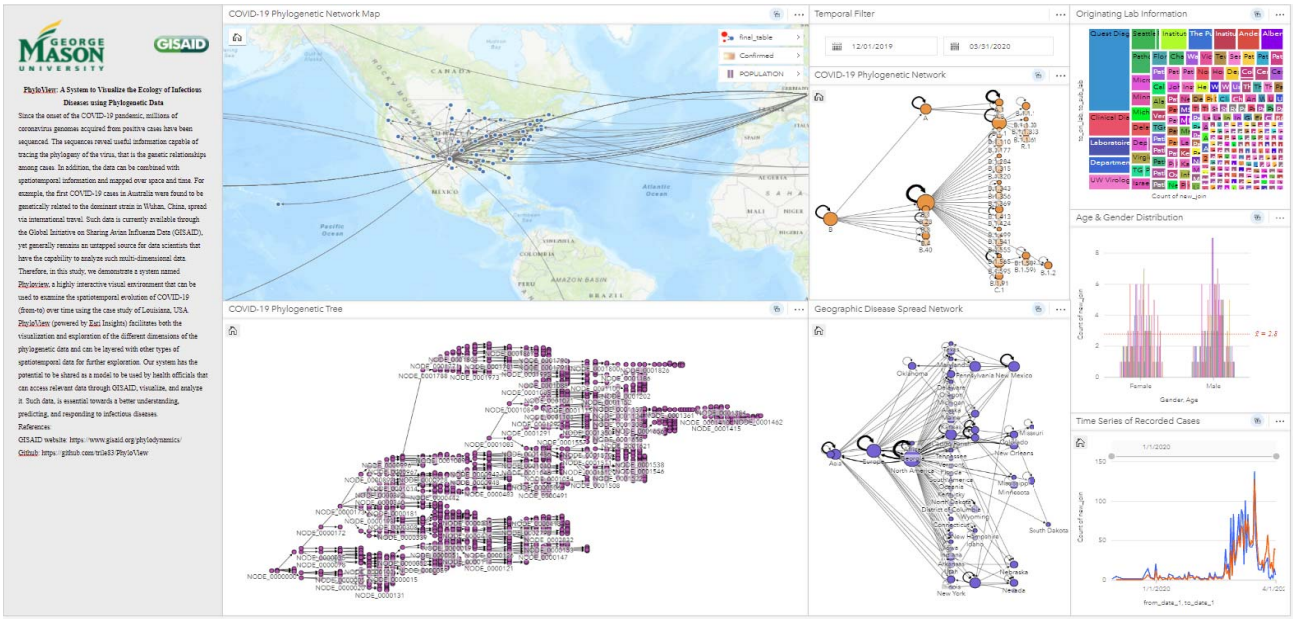

Fig. 5. An overview of Phyloview. a) Top-Left: A map showing the geographic spread of the virus using directed links that connect observed cases to the location of their most recent ancestors; b) Top-Middle: Control that filters data based on time and a network visualization of SARS-CoV-2 strains; c) Top-Right: Attribution of the sources/laboratories/publications that contributed each of the selected cases and below, the age and gender distribution of the cases; d) Bottom-Left: The phylogenetic tree of all currently selected cases; e) Bottom-Middle: A network visualization of virus evolution by geography where nodes represent locations and edges (and their weight) correspond to the number of phylogenetic connections between two locations; f) Bottom-right: Time series of the number of cases observed during the selected time interval.

how datasets were preprocessed and linked. We provide a link to our code repository to reproduce our demonstration using the Louisiana data and replicate our analysis with data for any spatial region, time interval, and infectious disease. In addition, to help the interested reader understand the functionality of PhyloView, we have prepared a short tutorial video to help quickly describe the functionality of PhyloView. This video can be found at https://youtu.be/S-8ibOLtxTY.

A. PhyloView Graphical User Interface

Figure 5 shows a screenshot of PhyloView using 1700 cases from Louisiana and their phylogenetic information. We note that for the purpose of this demonstration, our data were not obtained directly from the GISAID database, since the GISAID license agreement does not allow to re-share any data. Instead, to make our demonstration reproducible, we obtained data from GISAID's website on Phylodynamics of Pandemic Coronavirus in Louisiana, USA [12], [13]. This dataset, which contains 4,447 genomes collected between March 2020 and February 2022, as well as instructions on how to obtain it, are found in our GitHub repository [20].

Panel 1 (Temporal Filter) of PhyloView allows the user to select the time interval of the loaded data. By default, all data are shown in PhyloView, but some tasks may require showing only cases before or after a certain event.

Panel 2 (COVID-19 Phylogenetic Network Map) shows a map of all cases and their phylogenetic relationship, denoted by directed arrows to understand the spread of selected cases across the world visually. This panel allows users to select individual cases or phylogenetic links or select locations regionally by drawing a box.

Panel 3 (COVID-19 Phylogenetic Tree) visualizes the phylogenetic tree (or forest) for the selected data. PhyloView also shows summary statistics of other attributes of the selected cases.

Panel 4 (Originating Lab Information) references the laboratories that have sequenced the currently selected cases using a treemap. Clicking on each of the shown labs provides detailed information about the lab and corresponding publications.

Panel 5 (Age & Gender Distribution) depicts socio-demographic information of the selected cases, including age and gender information.

Panel 6 (Geographic Disease Spread Network) provides a network where nodes represent locations and links represent the genetic relationship between cases observed in those locations. This graph can be used to understand and visualize how the virus travels among places to understand which places may have high importance (for example, in terms of PageRank [25]) in the infection network.

Panel 7 (COVID-19 Phylogenetic Network) describes the relationship between different strains of COVID-19 as the virus evolves over space and time. And finally, Panel 8 (Time Series of Recorded Cases) measures the magnitude of cases over time where blue measures the number of "from" cases and orange represents the number of "to" cases.

All panels allow for the dynamic selection of data reflected in other panels, referred to as cross-filtering. For example, the user may select a specific case on the map, and both map and the phylogenetic tree will be updated to show only this case and all its phylogenetic descendants. Similarly, a case (and thus, a subtree) can be selected on the phylogenetic tree, and the map will automatically be updated to show the ecology of this specific case. This allows to investigate the spread that originated from a particular case rapidly.

B. Usecase: Louisiana, USA

To illustrate an application of PhyloView, imagine a local health department of a Louisiana county (called a parish) such as Caddo Parish that wants to understand the local ecology of COVID-19, i.e., to understand where cases in Caddo Parish originated from and where cases spread within and from Caddo Parish. Using PhyloView, Health officials can load GISAID data as shown in Figure 5 and select Caddo Parish (and neighboring parishes) on the map panel by drawing a box. This will automatically update all other panels to show only cases in the selected spatial region. The phylogenetic tree immediately shows subtrees to understand which cases have caused major spread. Non-selected nodes and edges are greyed out in the phylogenetic tree, allowing users to browse the path (towards the root in Wuhan, China) to understand the original trajectory of each observed case in Caddo Parish. The geographic disease spread network also provides a concise visualization of all places where cases entered to exit Caddo Parish.

C. Usecase: World-Wide Diseases Ecology

We also want to demonstrate PhyloView in the context of a global dataset. For this purpose, we use the "nextregions" dataset provided by GISAID [10] which is part of their EpiCoV database [1].

Figure 6 shows the spread of COVID-19 on a global scale utilizing available dataset on the GISAID website. The global COVID-19 data ranges from December 7, 2019 to December 31, 2021. This dataset includes a total of 5800 recorded cases across the world. Similar to the Louisiana dataset, this dataset provides the detailed location of each case (aggregated to city, county, or state level). However, a major difference in this dataset is that it does not provide a unique identifier of the most recent ancestor of a case. Instead, it only provides the continent where the most recent ancestor of a case was observed. Thus, a case observed in New York, USA having a most recent ancestor case in Italy, would correctly show New York, USA as the location, but would show Europe as the location of the most recent ancestor. Thus, this dataset yields much less information, but allows us to visualize the global spread on a high level.

This data allows to study, using the map of Figure 6, how COVID-19 spread across the planet having the sources of each case aggregated to continental level. We observe a fully connected network across the continents having each continent having phylogenetic links to places across the globe. This shows that the virus spread from all continents and, regardless where it originated, all continents have contributed to the pandemic spread of COVID-19 except for Australia and South America having a much lower number of outgoing edges. This is also evident from the Geographic Disease Spread Network (bottom right of Figure 6), showing three major hubs corresponding to the United States, Europe, and Asia. We encourage the interested reader to explore each of the panels by zooming in and out of spatial regions on the map, or into specific subnetworks of the geographic disease spread network using an instance of PhyloView that we have made available at https://insights.arcgis.com/index.html#/view/3e44d3af2a534efda911009420051e32.

[1]Upon logging into GISAID, navigating to the EpiCoV tab, then clicking Downloads, selecting "nextregions" and choosing the global version of this dataset.

D. Backend Model

PhyloView is powered by an analytical model that joins, selects, aggregates, and projects the GISAID data, including phylogenetic links between cases, spatiotemporal case information, socio-demographic information (age, gender) of cases, and information about the submitting laboratories and publications of each case. The model can be viewed by selecting the icon in the upper right-hand corner of the browser referred to as "Analysis View." The model that powers PhyloView is publicly available with a subscription to Esri ArcGIS Insights. Interested researchers and practitioners can search the repository using the keyword "PhyloView", download a copy of the PhyloView model, and "plug in" their own data obtained from GISAID. Users without access to Esri ArcGIS Insights can visualize and explore data in Phyloview as is but can not plug in their own data.

VII. Applications of PhyloView

Our goal is that PhyloView is used by the scientific community and stakeholders in public health to not only better understand the ecology of existing and future infectious diseases, but to identify the potential for the use of phylogenetic data in studies beyond biology and epidemiology. We hope that this platform will foster interdisciplinary collaborations between public health, biology and epidemiology, and data scientists. Here, we discuss some practical applications for phylogenetic data and PhyloView for local health departments and epidemiologists.

A. A System for Local Health Departments

The United States alone has more than 3,000 local health departments across the country. These city, county, metropolitan, district, and tribal departments work every day to protect and promote health and well-being for all people in their communities. For example, the Loudon County Health Department in Virginia identifies conditions that potentially affect the health of the public, provides communication on health-related matters to mitigate impacts, and works to control the spread of diseases [21]. We envision that PhyloView can be

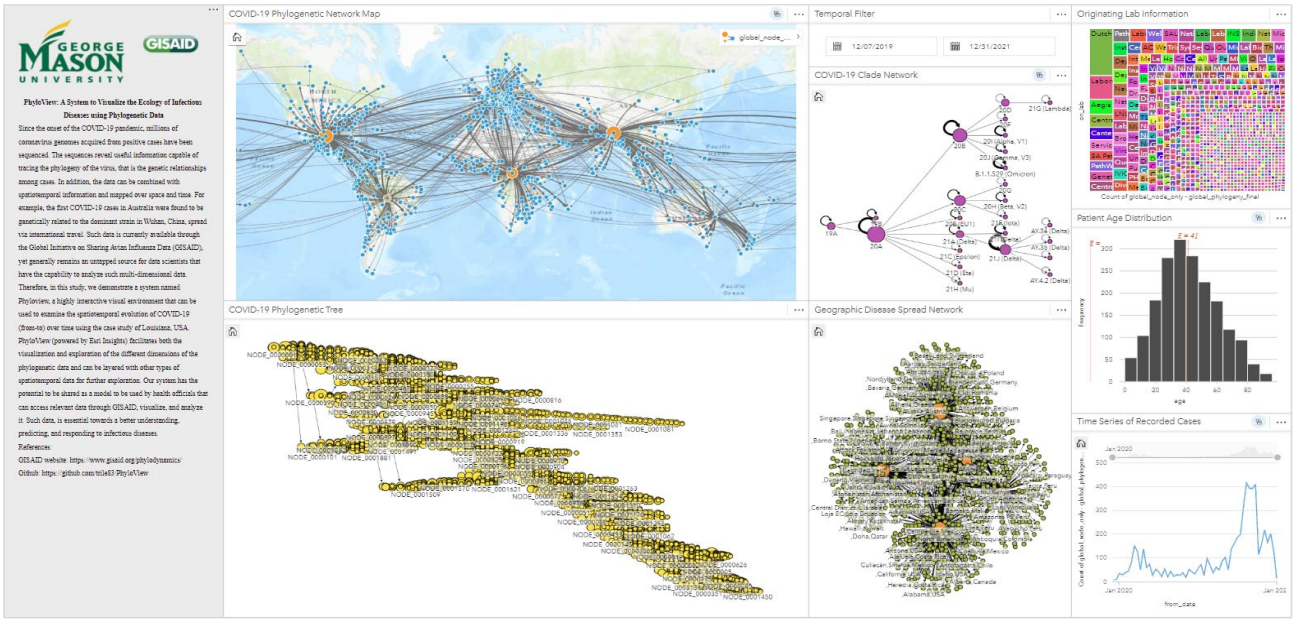

Fig. 6. A screenshot of using PhyloView to visually analyze the global spread of COVID-19 on a global dataset. Top-Left: Visualization of the spatio-temporal ecology of COVID-19 across the globe. For each case, the origin locations are aggregated to continent level for visualization. Bottom left: The phylogenetic tree of all cases included in this data. Top-Right: Phylogenetic network of the clades of cases observed in this dataset. Bottom-right: Geographic disease spread network visualizaing the spread between continents and countries/cities.

used by such local health departments to trace the evolution of the spread of disease (including influenza and Covid-19), to understand the effect of ports of entry to local communities, to identify unknown incoming or outgoing pathways of spread that could be a public health risk, and to identify clusters of infections in space and time.

B. A System for Epidemiologists

PhyloView can be used by biologists and epidemiologists as a tool to better understand the transmission dynamics and pathways of infectious diseases. The challenge is to concurrently visualize all dimensions of disease spread including spatial, temporal, and phylogenetic information. For that purpose, PhyloView connects all panels such that when selecting a spatial region, a temporal interval, or a phylogenetic subtree in one panel automatically updates all other panels to allow for the examination of the relationships between different data dimensions. Users are encouraged to draw a spatial box on the map of PhyloView to select a spatial region, select a temporal subset, or select a node in the phylogenetic tree to select the corresponding subtree. While solutions exist to visualize each of the dimensions individually, to the best of our knowledge, PhyloView is the first tool to allow visual analysis of all three dimensions at the same time. With its ability to load additional layers of data, Phyloview would allow for an analysis of the associations between characteristics of a location (transportation and mobility networks, socio-demographic characteristics) and the emergence of new strains. Identifying transmission pathways that facilitate the spreading and evolution of a disease may be useful in future outbreak events. Although we describe these use cases separately, the findings of this analysis would support public health decision making.

C. A System for Data Scientists

Traditional analysis of infectious diseases data focuses on the location and time of infected individuals. Thus, the traditional focus is on the "to" field of phylogenetic data records. However, less attention is paid to the "from" fields which are enabled through phylogenetic analysis, thus answers not only the question of where infectious diseases have spread to, but also where it has come from. Phylogenetic data, such as used by PhyloView in Figure 5 enables such research as it enriches observed cases of an infectious disease with information of where the case originated from. A possible challenge for mobility data science could be to identify nodes that exhibit a significantly large number of outgoing nodes. Such "hubs" or "influencers" could then be investigated further for possible causes of such increased spread. Similar research has been done to understand the most influential nodes on the web [25], in communication networks [15], and in social networks [19]. For geographic disease spread networks, new algorithms may be needed to properly model population density and geographic autocorrelation. Beyond finding "most influencial" nodes in a geographic disease spread network, another challenge is to find frequent pathways which may span multiple geographical nodes to understand global disease ecology and to take preventive actions to mitigate the spread of future diseases.

VIII. Conclusions

PhyloView is a system that allows to visually analyze phylogenetic information of disease data across space and time. PhyloView allows understanding not only where and when cases of infectious diseases such as COVID-19 were observed, but also where each of the cases emerged from to improve our understanding of infectious disease ecology. This understanding is facilitated by multiple tools for visual analytics, including a) phylogenetic tree view, b) a spatiotemporal disease ecology map, c) a geographic diseases spread network, and d) a phylogenetic network. PhyloView connects each of these tools, such that selecting cases based on their phylogeny is visualized on the map and vice versa.

A shortcoming of PhyloView is data availability. While sequenced cases of infectious diseases such as influenza and COVID-19 are shared by GISAID [10], only a small fraction (~3.3% in the United States) of observed cases of infectious diseases are sequenced at all [11]. Users of PhyloView should remain cautious of the resulting uncertainty in the data and acknowledge that important pathways of disease ecology may be missing in the data.

We hope that PhyloView will help local health departments rapidly understand the spread of future infectious diseases in their community to support mitigative decision-making. In addition, we hope that PhyloView may be used by biologists and epidemiologists as a research tool and we hope that PhyloView will inspire the spatiotemporal data science community to investigate phylogenetic data to improve our understanding of the spread of the COVID-19 virus to improve our preparedness for future pandemics. On the flip side, we hope that data-driven research using phylogenetic data will help justify more cases to be sequenced by the epidemiology community, thus creating a synergy between our two disciplines.

Acknowledgements

This research has been supported by National Science Foundation grants DEB-2109647 and DEB-2030685.

References

[1] D. Baum et al. Reading a phylogenetic tree: the meaning of monophyletic groups. *Nature Education*, 1(1):190, 2008.

[2] G. Bobashev, I. Segovia-Dominguez, Y. R. Gel, J. Rineer, S. Rhea, and H. Sui. Geospatial forecasting of covid-19 spread and risk of reaching hospital capacity. *SIGSPATIAL Special*, 12(2):25–32, 2020.

[3] L. L. Cavalli-Sforza and A. W. Edwards. Phylogenetic analysis. models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1):233, 1967.

[4] T. Chookajorn. Evolving covid-19 conundrum and its impact. *Proceedings of the National Academy of Sciences*, 2020.

[5] S. Elbe and G. Buckland-Merrett. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Global challenges*, 1(1):33–46, 2017.

[6] esri. ArcGIS Insights. https://doc.arcgis.com/en/insights, Accessed 11/10/2021.

[7] P. Forster, L. Forster, C. Renfrew, and M. Forster. Phylogenetic network analysis of sars-cov-2 genomes. *PNAS*, 117(17):9241–9243, 2020.

[8] I. Franch-Pardo, B. M. Napoletano, F. Rosete-Verges, and L. Billa. Spatial analysis and gis in the study of covid-19. a review. *Science of the total environment*, 739:140033, 2020.

[9] S. Gao, J. Rao, Y. Kang, Y. Liang, and J. Kruse. Mapping county-level mobility pattern changes in the united states in response to covid-19. *SIGSpatial Special*, 12(1):16–26, 2020.

[10] GISAID. Enabling rapid and open access to epidemic and pandemic virus data. https://www.gisaid.org/about-us/mission/, 2021.

[11] GISAID. Submission tracker global. https://www.gisaid.org/index.php?id=208, 2021.

[12] GISAID and L. H. Shreveport. Gisaid-louisiana. https://www.gisaid.org/phylodynamics/louisiana-usa/, 2021.

[13] GISAID and L. H. Shreveport. Gisaid-louisiana. https://phylodynamics.pandemicprepardness.org/charon/getDataset?prefix=/SARS-CoV-2/Louisiana, 2021.

[14] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[15] Q. Han. Social influence analysis using mobile phone dataset. In *IEEE MDM*, pages 12–17. IEEE, 2016.

[16] Y. A. Helmy, M. Fawzy, A. Elaswad, A. Sobieh, S. P. Kenney, and A. A. Shehata. The covid-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control. *Journal of clinical medicine*, 9(4):1225, 2020.

[17] A. Hohl, E. Delmelle, and M. Desjardins. Rapid detection of covid-19 clusters in the united states using a prospective space-time scan statistic: An update. *SIGSPATIAL Special*, 12(1):27–33, 2020.

[18] Y. Junejo, M. Ozaslan, M. Safdar, R. A. Khailany, S. Rehman, W. Yousaf, and M. A. Khan. Novel sars-cov-2/covid-19: origin, pathogenesis, genes and genetic variations, immune responses and phylogenetic analysis. *Gene reports*, 20:100752, 2020.

[19] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, pages 137–146, 2003.

[20] M. T. Le, D. Attaway, T. Anderson, H. Kavak, A. Roess, and A. Zufle. Phyloview. https://github.com/trile83/PhyloView, 2021.

[21] Loudon County, Virginia. Loudon County Health Department (https://www.loudoun.gov/111/Health-Department), Accessed, 02/21/2022.

[22] A. Maxmen. One million coronavirus sequences: popular genome site hits mega milestone. *Nature*, 593(7857):21–21, 2021.

[23] M. Mokbel, S. Abbar, and R. Stanojevic. Contact tracing: Beyond the apps. *SIGSPATIAL Special*, 12(2):15–24, 2020.

[24] B. B. Oude Munnink, N. Worp, D. F. Nieuwenhuijse, R. S. Sikkema, B. Haagmans, R. A. Fouchier, and M. Koopmans. The next phase of sars-cov-2 surveillance: real-time molecular epidemiology. *Nature medicine*, 27(9):1518–1524, 2021.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[26] J. Pesavento, A. Chen, R. Yu, J.-S. Kim, H. Kavak, T. Anderson, and A. Züfle. Data-driven mobility models for covid-19 simulation. In *ACM SIGSPATIAL*, pages 29–38, 2020.

[27] R. Pung, C. J. Chiew, B. E. Young, S. Chin, M. I. Chen, H. E. Clapham, A. R. Cook, S. Maurer-Stroh, M. P. Toh, C. Poh, et al. Investigation of three clusters of covid-19 in singapore: implications for surveillance and response measures. *The Lancet*, 395(10229):1039–1046, 2020.

[28] T. Rito, M. B. Richards, M. Pala, M. Correia-Neves, and P. A. Soares. Phylogeography of 27,000 sars-cov-2 genomes: Europe as the major source of the covid-19 pandemic. *Microorganisms*, 8(11):1678, 2020.

[29] A. J. Rodríguez-Morales, G. J. Balbin-Ramon, A. A. Rabaan, R. Sah, K. Dhama, A. Paniz-Mondolfi, P. Pagliano, and S. Esposito. Genomic epidemiology and its importance in the study of the covid-19 pandemic. *genomics*, 1:3, 2020.

[30] Y. Shu and J. McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

[31] Taylor Anderson, J. Yu, and A. Züfle. Proceedings of the 1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19, 2020.

[32] L. Xiong, C. Shahabi, Y. Da, R. Ahuja, V. Hertzberg, L. Waller, X. Jiang, and A. Franklin. React: real-time contact tracing and risk monitoring using privacy-enhanced mobile tracking. *SIGSPATIAL Special*, 12(2):3–14, 2020.

[33] S. Zeighami, C. Shahabi, and J. Krumm. Estimating spread of contact-based contagions in a population through sub-sampling. *Proc. VLDB Endow.*, 14(9):1557–1569, 2021.

[34] A. Züfle. Introduction to this Special Issue: Modeling and Understanding the Spread of COVID-19: (Part I). sigspatial special volume 12 issue 1. pp. 1–2., 2020.

[35] A. Züfle and T. Anderson. Introduction to this Special Issue: Modeling and Understanding the Spread of COVID-19 (Part II).SIGSPATIAL Special Volume 12 Issue 2. pp. 1–2 , 2020.