Analysis of SARS-CoV-2 temporal molecular networks using global and local topological characteristics

Fiona Senchyna[1] and Rahul Singh[1]

[1]Department of Computer Science, San Francisco State University, 1600 Holloway Ave., San Francisco, CA 94132, USA
fsenchyna@mail.sfsu.edu, rahul@sfsu.edu

Abstract. The global COVID-19 pandemic continues to have a devastating impact on human population health. In an effort to fully characterize the virus, a significant volume of SARS-CoV-2 genomes have been collected from infected individuals and sequenced. Comprehensive application of this molecular data toward epidemiological analysis in large parts has employed methods arising from phylogenetics. While undeniably valuable, phylogenetic methods have their limitations. For instance, due to their rooted structure, outgroup samples are often needed to contextualize genetic relationships inferred by branching. In this paper we describe an alternative: global and local topological characterization of neighborhood graphs relating viral genomes collected from samples in longitudinal studies. The applicability of our approach is demonstrated by constructing and analyzing such graphs using two distinct datasets from Israel and France, respectively.

Keywords: SARS-CoV-2, Graph topology, Network theory, Computational epidemiology

1 Introduction

The rapid dissemination of Coronavirus-Diesase-19 (COVID-19) since its first recorded outbreak in December 2019 has led to a worldwide pandemic with devastating consequences. According to the World Health Organization (WHO), currently over 300 million cases have been confirmed globally, including over 5 million deaths [1]. Consequently, the global research community has taken tremendous efforts to study the etiological agent, SARS-CoV-2, of COVID-19. The unprecedented volume of whole genomes sequenced and made publicly available has led to rapid advancements in areas such as drug development, diagnostics, and understanding of the pathogenicity and epidemiology of the virus [2–6]. To date, the GISAID (Global Initiative on Sharing Avian Influenza Data) database, a popular publicly available repository for SARS-CoV-2 sequence data, contains over 11 million genomes [7].

In molecular epidemiology, genomes sampled from infected individuals are related to one another based on sequence similarity, typically within a phylogenetic

framework. Inferences are then made regarding the spread and prevalence of a virus within a population. Although originally intended to determine the relatedness of different taxa, phylogenetics has been co-opted and modified for analysis of pathogen transmission [8–10]. Phylodynamic analysis, which studies the interaction and influence of epidemiological, immunological, and evolutionary processes on viral evolution and genetic variation, similarly infers viral population levels over time based on a phylogenetic tree [11]. Regarding SARS-CoV-2, phylogenetic and phylodynamic studies have been used to estimate the source and date of origin of infection, the temporal reproductive number, geographical spread, and the role of super spreaders [5, 6, 9].

Although demonstrably valuable, phylogenetics has its limitations when analyzing disease spread. Phylogenetics methods, particularly those employing Bayesian models, have many parameters that can be challenging to estimate (*e.g.*, the substitution model, molecular clock, and priors). Often, the phylogenetic tree has to be recomputed if a new sequence is added [12] – a significant overhead in large epidemiological settings. Additionally, the constraint of a tree structure limits the topology of the patterns that can be hypothesized and studies. Indeed, the topology of infection spread generally does not conform to the constraints of a single source and predefined branching tree structure [13]. By contrast, a different view of the information arises when it is modeled as a network (graph). Network representations of data have allowed for increased understanding of several biological phenomena (*e.g.*, gene and protein functions, human neural networks, and epidemiological contact tracing) [13–17]. Network properties derived from such representations can be divided into "global" and "local". Global network properties include degree distribution, diameter, path length, and centrality and characterize the connectivity of the entire dataset. Local properties on the other hand, characterize a network in terms of the connectivity of its node to nodes in a local neighborhood. Small, induced subgraphs, called *graphlets*, of the larger network, are one such local topological property. In practice, graphlets are typically defined to consist of graphs containing 3 to 5 nodes. This yields 29 unique graphlet structures whose presence can be used to characterize the local structure of a network. The *relative graphlet frequency* (RGF) distance between two networks can be used as a network comparison measure by comparing the frequency of all 29 *graphlets* in both networks [14]. Local connectivity can also be investigated with the graph Laplacian, which partitions the graph based on an optimal cut, and can reveal communities of nodes within the graph [18].

Molecular genetic networks have been computed for Human Immunodeficiency virus (HIV) and Hepatitis C virus (HCV) for contact tracing purposes [16, 19–21]. In these networks, viral genetic samples taken from infected individuals are represented as nodes. Edges are added between two nodes if the genetic distance between the pair of samples is below a certain distance threshold. HIV and HCV are blood-borne viruses, and their transmission is often associated with high-risk behaviors [22, 23]. In contrast, SARS-CoV-2 is a highly transmissible airborne virus. The resulting large volume of un-sampled hosts makes it virtually impossible to accurately perform contact tracing from sampled sequences alone [23]. However, analysis of changes in topological properties of a SARS-CoV-2 genetic network can provide insight about

the accumulation (or lack thereof) in variation of the virus within a population over time. Here, temporal genetic networks were built for two datasets separately based on a genetic distance threshold of $2x10^{-4}$. Global and local properties of the graphs were analyzed to characterize the dataset and relate to underlying biological changes, including the use of graph cuts to identify emerging viral subtypes within the datasets.

2 Data and Methods

2.1 Data and Preprocessing

Analysis of SARS-CoV-2 molecular evolution within a population was performed on two distinct datasets described previously [5, 6]. Samples in each dataset originated from France and Israel, respectively, and were collected during the first wave of the pandemic in the early months of 2020. For each sequence, the collection date was known. SARS-CoV-2 genomes were downloaded from the GISAID database (https://www.gisaid.org). Accession numbers for the Israel dataset (IDS) (n=212) are EPI_ISL_447258 - EPI_ISL_447469; and for French dataset (FDS) (n=186) are EPI_ISL_414624-7,29-38, EPI_ISL_415649-54, EPI_ISL_416493-502,504-506, 508-513, EPI_ISL_416745-52, 54, 56-58, EPI_ISL_417333-4, 36-40, EPI_ISL_418218-40, EPI_ISL_418412-31, EPI_ISL_419168-88, EPI_ISL_420038-64, EPI_ISL_420604-25, and EPI_ISL_421500-1. A reference sequence (originating from the first recorded outbreak in Wuhan, China) was downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/, accession number MN908947). Separately, the genomes for each dataset were aligned to the reference sequence with MAFFT [24]. Non-coding regions were removed from all genomes according to the reference sequence annotation. Additionally, samples were removed from further downstream analysis if the coding region contained more than 1% ambiguous nucleotides or gaps. Insertions and deletions (indels) were ignored due to lack of clarity between indels and ambiguous nucleotides. The final size consisted of 171 and 181 samples for IDS and FDS, respectively. After removal of non-coding regions, sequences had a nucleotide length of 29,132.

2.2 Construction of Temporal Networks

Each dataset was represented as a set of temporally evolving networks,

$$G(t) = (V(t), E(t)) \tag{1}$$

where $t = \{t_1, \ldots t_n\}$ are the set of time points corresponding to the sample collection times. Consequently, $\Delta G = G(t_i) - G(t_{i-1})$ defines the incremental change in the network between the times t_i and t_j and $V(t) = \{v_1, \ldots, v_n\}$ represents the samples collected on or before t. The edge, $e_{ij} \in E(t)$ connects v_i and $v_j \in V(t)$ if the genetic distance between v_i and v_j is below a threshold and indicates the respective samples to be genetically close within the time spanned by t. The reader may note that no constraints are placed on the specific time-points at which the data is gathered.

Pairwise genetic distances are calculated as the number of sites where the two sequences differed divided by the total length of the sequence (hamming distance). For each pairwise comparison, positions with ambiguous nucleotides or gaps are ignored. A genetic distance threshold of $2x10^{-4}$ was empirically chosen to connect sample nodes in the network. This threshold ensured that the majority of pairwise distances were below the cutoff value in both datasets: 53.8% (7,813/14,535) for IDS and 70.8% (11,532/16,290) for FDS (Figure 1). Nevertheless, the threshold was high enough to prevent formation of highly connected networks that lacked any meaningful topology. IDS produced 27 graphs from March 17th to April 22nd, 2020. FDS produced 26 graphs dated from February 26th to March 24th.

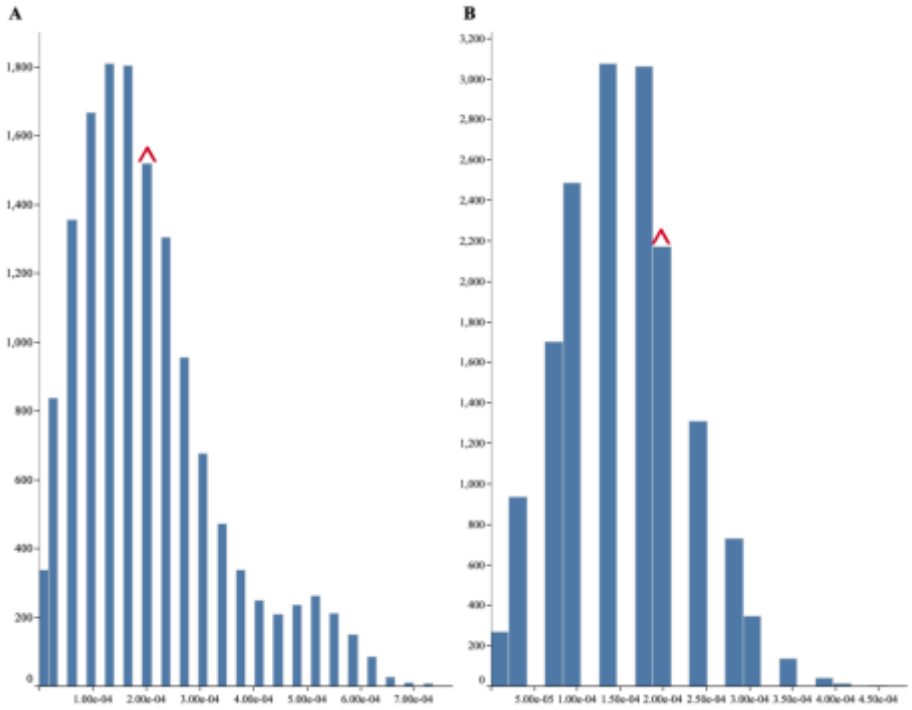

Fig. 1. Genetic distance distribution for (A) IDS and (B) FDS. Genetic distances are on the x-axis and number of pairwise comparisons are on the y-axis. ^ indicates the bin containing the cut-off threshold in both plots.

2.3 Global Network Analysis

To assess how nucleotide changes in the viral population are reflected in a network, the following global properties were calculated: degree distribution, average clustering coefficient, average path length, and diameter. Centrality measures including degree, closeness, and betweenness centrality were also calculated to identify those nodes most central to the network and relate to the nucleotide constitution of the respective datasets.

2.4 Local Topological Analysis

To ensure that small inconsequential connected components do not impact comparisons of networks, only the largest subnetwork, $s_i(G(t))$, for $G = (G_1, \ldots, G_n)$ was used for local connectivity analysis. This was done as the largest subnetwork contained over 98% of total samples in the majority of graphs. Henceforth, the major subnetworks will be referred to as $G_1, \ldots, G_n$. That is, for notational simplicity, we are simply using the notation G_i to denote the largest subnetwork of G_i. We employed the notion of graphlets to conduct the local connectivity analysis using *Graph Crunch 2* [14]. To identify topological changes between sampling periods, the *relative graphlet frequency* distance (RGF distance) was calculated between consecutive graphs, $G(t_i)$ and $G(t_{i+1})$. The RGF distance (D) is defined as follows [25]:

$$D(\Delta, \Omega) = \sum_{i=1}^{29} |F_i(\Delta) - F_i(\Omega)| \tag{2}$$

where Δ and Ω represent $G(t_i)$ and $G(t_{i+1})$, and $i \in \{1, \ldots, 29\}$ in (2) are the number of distinct *graphlets*. Further,

$$F_i(\Phi) = {-\log(N_i(\Phi))} \Big/ {\sum_{i=1}^{29} N_i(\Phi)} \tag{3}$$

The function F_i divides the log of the frequency of a graphlet by the sum of the frequencies of all graphlets to ensure the difference in node size between Δ and Ω is accounted for (3).

2.5 Quantification of Nucleotide Variation

The consensus sequences for each $G = (G_1, \ldots, G_n)$ were calculated. Single nucleotide variants (SNVs) were identified for each sample by pairwise comparison to the reference sequence. For any single SNV, if its frequency within the population was above 0.1 it was deemed a mutation of interest. To quantify change over time, nucleotide diversity of G_t and the difference in diversity between G_t and samples added at G_{t+1} (Δ_{t+1}) were calculated. Nucleotide diversity was characterized according to the definition by Nei and Li [26].

2.6 Spectral Network Partitioning

The connected components in each of $G_1, \ldots, G_n$ were split using spectral partitioning. Let the Laplacian matrix of a network G, $L(G)$, be defined as follows:

$$L(G) = D(G) - A(G) \tag{4}$$

Where $A(G)$ is the adjacency matrix and $D(G)$ is diagonal matrix of the graph (4). Partitioning of the graph into connected components was accomplished through eigendecomposition of $L(G)$. Nodes (samples) are split based on whether their values

in the eigenvector associated with the second smallest eigenvalue is above or below a defined threshold. Five thresholds were tested for the initial splitting of the graph (-0.008, -0.0075, -0.007, -0.005, and 0). The quality of a partition was quantified using the normalized cut value induced by that partition. A threshold of -0.007 was chosen as it consistently gave the lowest normalized cut value for all graphs in both datasets. The connected components were recursively partitioned into connected components of at least two nodes to investigate further groupings of samples when applicable.

2.7 Phylogenetic Analysis.

As a comparison to the spectral partitioning of the graph into clusters, phylogenetic analysis was performed on the genome sequences. A maximum likelihood phylogenetic tree was constructed using RAxML (Randomized Accelerated Maximum Likelihood) v1.0.0 [27] with a GTR substitution model and 100 bootstrap replicates.

3 Results and Analysis

3.1 Genetic Characterization of the Viral Population

IDS. The initial consensus sequence had 4 SNVs compared to the reference sequence. These were C3037T, C14408T, A23403G, and G25563T. On March 21st, the nucleotide at position 1059 in the consensus changed from C to T. The proportion of samples containing a C at this position decreased from 60% (6/10) to 42.86% (6/14) (-28.57%). There were 11 mutations of interest that were not part of the consensus sequence. These included C2416T (overall frequency, 0.11), C11916T (0.17), C18998T (0.15), G28881A (0.11), G28882A (0.11), and G28883C (0.12). The nucleotide diversity between consecutive graphs remained consistent (median, 1.99×10^{-4}, interquartile range (IQR), 1.83×10^{-4} - 2.06×10^{-4}). Similarly, the absolute difference in diversity between samples within G_t and Δ_{t+1} was also very small (median, 5.15×10^{-5}, IQR, 2.41×10^{-5} - 7.21×10^{-5}).

FDS. Compared to the reference sequence**,** the consensus sequence had 4 SNVs (C3037T, C14408T, A23403G, and G25563T). The nucleotide at position 1059 of the consensus changed from T to C on March 4th, the opposite of IDS. The frequency of C in this position increased from 26.67% (4/15) to 51.61% (16/31) (+93.51%). This nucleotide is in the open reading frame (ORF) 1a region of the SARS-CoV-2 genome and encodes the gene *Nsp2*. The reference sequence contains C and the mutation to T is non-synonymous. However, the full functionality of *Nsp2* has yet to be fully understood and so the effect of this mutation on viral fitness is unknown [28, 29].Two SNVs, C2416T (0.13) and C15324T (0.34), were deemed of interest. Like IDS, there was little variation in nucleotide diversity between consecutive graphs (median, 1.28×10^{-4}, IQR, 1.22×10^{-4} -1.47×10^{-4}). The absolute difference in diversity between G_t and Δ_{t+1} was also minimal (median, 2.46×10^{-5}, IQR, 1.27×10^{-5} - 4.51×10^{-5}).

3.2 Changes in Global Properties

IDS. We found that most global network properties experienced little change. The diameter gradually increased over time, shifting from 2 to 3 then growing to 5 as new samples were added. The median clustering coefficient was 0.88 (IQR, 0.86 – 0.89) and the median average path length was 1.58 (IQR, 1.55-1.61). There was some change in the most central nodes according to degree and closeness centrality. However, the most central nodes were not vastly different in their genomic constitution. They differed from the consensus sequence by 0-3 nucleotides (EPI_ISL_447408, EPI_ISL_447310, EPI_ISL_447305, EPI_ISL_447277, EPI_ISL_447284). The joining of a smaller subnetwork to the major subnetwork on March 30th did not influence degree or closeness centrality, but the node with the highest betweenness centrality did shift from EPI_ISL_447277 to EPI_ISL_447447. EPI_ISL_447277 had 3 nucleotide differences from the consensus sequence, while EPI_ISL_447447 had 6. This change is most likely due to the difference in the measures of centrality. Both the degree and closeness centrality measure the relation of a node to all other nodes in the graph, by node degree or length of shortest paths, respectively. Whereas betweenness centrality measures the impact of a node on the shortest paths between all other pairs of nodes in the graph. A divergence in a small proportion of the samples from the consensus sequence would, therefore, naturally have a larger effect on betweenness than closeness or degree. There were no significant findings in the changes in the degree distribution.

FDS. there was little change in the FDS diameter (1-4), average clustering coefficient (median, 0.91, IQR, 0.89-0.92) and average path length (median, 1.17, IQR, 1.15-1.28) over time. Initially, between February 26th and March 2nd, the graph was fully connected and so all samples were equally central. As the graph progressed through time, the nodes with the highest degree, closeness, and betweenness centrality overlapped substantially. These samples differed by 0-3 mutations from the consensus sequence. By the end of the period, the nodes with the highest betweenness centrality were identical to the consensus sequence except at position 25563 (EPI_ISL_414631, EPI_ISL_416494, EPI_ISL_420047). This mutation was present in the population at a frequency of 0.42. Four additional nodes shared the highest degree and closeness centrality. These nodes were an exact match to the consensus sequence (EPI_ISL_418219, EPI_ISL_417336, EPI_ISL_418425, EPI_ISL_419168). Again, the degree distribution did not provide significant insight.

3.3 Association between RGF Distance and Genetic Variation

The RGF distance between consecutive IDS graphs ranged from 0 to 5.57. Most distances were below 1 (77%, 20/26). However, there were two periods where the RGF distance sharply increased. The first was between March 20th and March 25th, reaching a maximum distance of 5.57, and the second between April 19th and 20th, with a distance of 2.01 (figure 2). For FDS, the RGF distance between consecutive graphs ranged from 0.02 to 8.58. Similar to IDS, most distances were below 1 (88%, 22/25).

Again, there were two peaks where large distances were recorded: that being 8.58 between March 2nd and 4th, and 1.98 between March 21st and 22nd (Figure 2). Graphs pertaining to the first peak in RGF distances for IDS and FDS are illustrated in Figures 3 and 4, respectively.

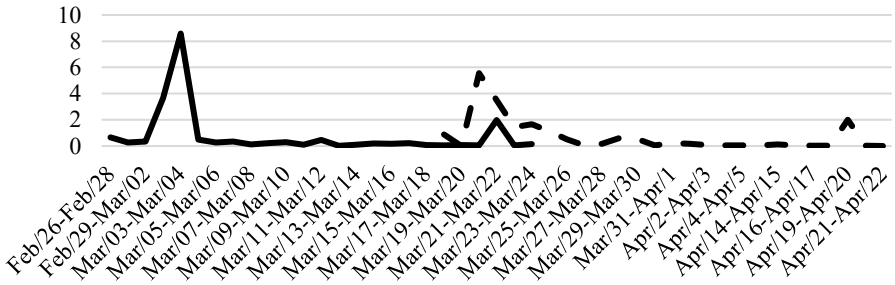

Fig. 2. RGF distance (y-axis) between consecutive graphs (x-axis) in FDS (solid) and IDS (dashed).

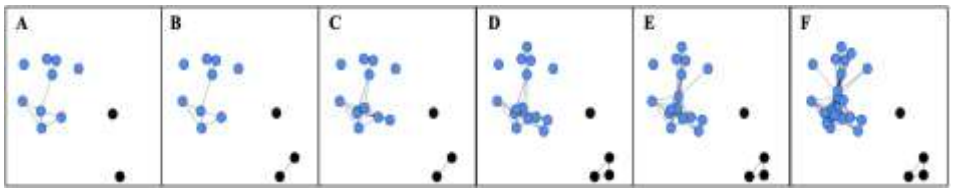

Fig. 3. Change in IDS graphs across time, including March (A) 19th, (B) 20th, (C) 21st, (D) 22nd, (E) 23rd, and (F) 24th. Nodes within the largest subnetwork are in blue.

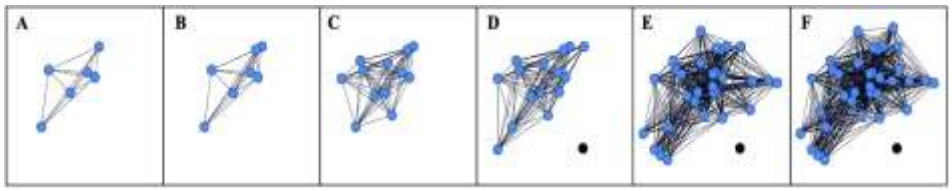

Fig. 4. Progression of the FDS connected component across February (A) 28th and (B) 29th, and March (C) 2nd, (D) 3rd, (E) 4th, and (F) 5th. Nodes of largest subnetwork are in blue.

Interestingly, change in the consensus sequence (described in section 3.1) corresponded with the highest RGF distance for both IDS and FDS. While the connection between RGF distance and a change in the consensus sequence is intriguing, no significant genetic variation could be found in either dataset to explain the second peak. From the data studied here, the RGF distance may simply reflect the gradual accumulation of variation. Further longitudinal data with significant heterogeneity in variants is needed to study how local topological changes can be characterized using measures such as the RGF distance.

3.4 Laplacian Network Partitioning versus Phylogenetic Analysis

Initially, as new samples were added, the partitioning of IDS differed from graph to graph. The median normalized cut value between this time was 0.88 (IQR, 0.58-1.02). From March 30th through the remainder of the studied period, the partition in the graph remained the same. The partition consisted of two connected components with 164 and 5 samples, respectively, and the median normalized cut value decreased to 0.09 (IQR, 0.09-0.09). Samples in the latter component (EPI_ISL_447324, EPI_ISL_447322, EPI_ISL_447319, EPI_ISL_447271, and EPI_ISL_447265) shared 7 SNVs that were uncommon in the entire dataset. Those were G11083T (overall frequency, 0.06), C14805T (0.04), T17247C (0.04), C17676T (0.03), G26144T (0.04), G26660T (0.03), and C29627T (0.04). C17676T and G26660T were present only in this component and not in any sample in the larger component. None of the samples contained the 11 SNVs of interest. The samples that connected the small and large partitions were EPI_ISL_447271 and EPI_ISL_447447, respectively. EPI_ISL_447447 also had the highest betweenness centrality in the graph, as described in 3.2. The normalized cut value for the French dataset remained relatively consistent throughout the studied period (median=1.18; IQR, 1-1.23). Additionally, until the last 3 days, the partitioning changed between graphs of different time periods. During these last three days, the cut, or the number of edges required to be removed to partition the graph, was 7.71% (638/8273), 6.72% (740/11011), and 6.6% (761/11532), respectively. The smaller partition consisted of 9 nodes (EPI_ISL_416746, EPI_ISL_414631, EPI_ISL_416494, EPI_ISL_418235, EPI_ISL_418426, EPI_ISL_418428, EPI_ISL_420047, EPI_ISL_420043, EPI_ISL_419178). 6 out of 9 samples shared mutations in the nucleocapsid phosphoprotein at positions G28881A, G28882A, G28883C. These mutations were not found in any sample in the larger connected component. In both IDS and FDS, the larger connected components had the C3037T, C14408T, A23403G mutations in over 95% and 100% of the samples, respectively. These mutations have been acknowledged as marker mutations of major clades and transmission clusters by GISAID and others [7, 30, 31], while the shared SNVS in the smaller IDS and FDS connected components have also been identified as markers of SARS-CoV-2 subtypes [32]. Partitioning of the last tracked graph, G_n, was compared to a phylogenetic tree representation of the data. As can be seen in figure 5, the importance of samples grouped together by spectral partitioning are less obvious when illustrated as a clade grouping on a phylogenetic tree and are even not necessarily within the same clade (figure 5B).

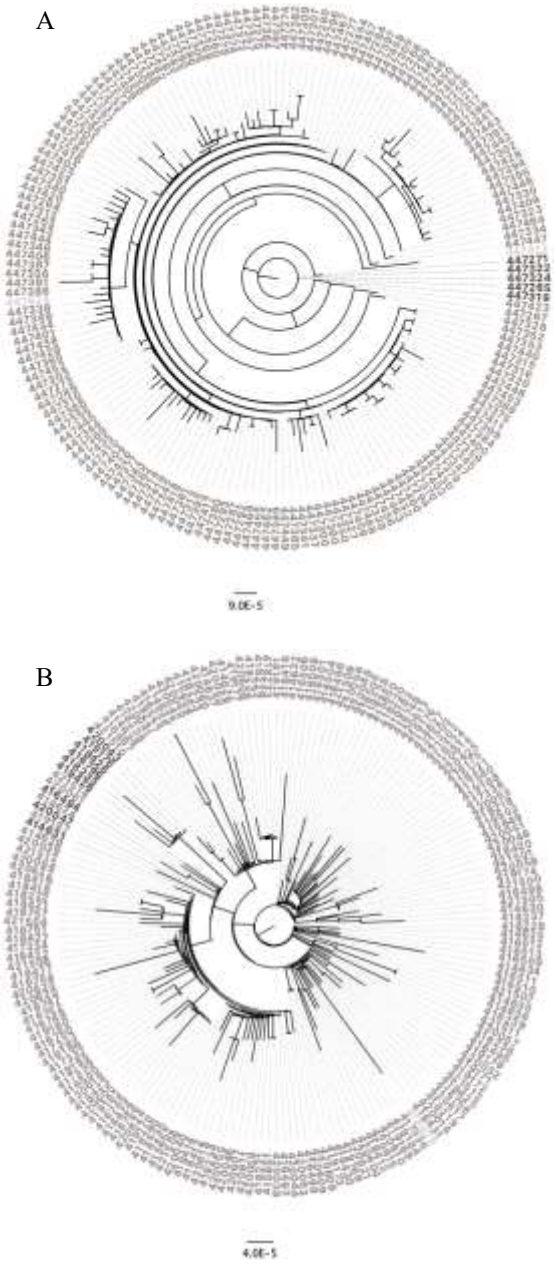

Fig. 5. Phylogenetic tree for (A) IDS and (B) FDS. Colors indicate samples grouped together by spectral partitioning (darkest and second darkest grey, respectively) and samples not part of the major subnetwork (light grey). Numberings refer to the GISAID sample accession numbers.

4 Conclusions

Here, we add a temporal dimension to a network representation to elucidate a connection between change in topological properties and viral molecular evolution within a population. Temporal dynamic networks, or time-varying graphs, can be represented as multiple networks acting as "snapshots" of the overall network changing in time.

In the two datasets studied by us, tracking of changes in the network reflected the evolution in the underlying viral genetic population. Small changes in global properties such as diameter and average path length were congruent with the low level of change in the overall nucleotide diversity of the population. Spectral partitioning of the graph was able to highlight communities of samples with shared SNVs not obvious from a phylogenetic construction of the data. The implication of graphlet-based analyses local topological analysis is less clear. Although preliminary results presented here found that the largest RGF distance between two temporally adjacent networks coincided with a shift in the consensus sequence of the population, this finding was not consistent in other temporally adjacent networks with a relatively large RGF distance.

Acknowledgements. This research was funded in part by National Science Foundation grant IIS-1817239.

Genome sequences analyzed is this work were submitted and collected be the following laboratories: IDS was submitted by the Stern Lab and collected by Microbiology laboratory, Assuta Ashdod University-Affiliated Hospital (EPI_ISL_447258 - 80); Microbiology Division, Barzilai University Medical Center (EPI_ISL_447281-310); Clinical Virology Laboratory, Soroka Medical Center and the Faculty of Health Sciences, Ben-Gurion University of the Negev (EPI_ISL_447311-30); Clinical Virology Unit, Hadassah Hebrew University Medical Center (EPI_ISL_447331-82, EPI_ISL_447407-16), Clinical Microbiology Laboratory, The Baruch Padeh Medical Center, Poriya (EPI_ISL_447383-406, EPI_ISL_447417-8); and Clinical Microbiology Laboratory, Sheba Medical Center (EPI_ISL_447419-69).

FDS was submitted by two laboratories, namely the National Reference Center for Viruses of Respiratory Infections, Institut Pasteur, Paris. Samples were collected by the Centre Hositalier Universitaire de Rouen Laboratoire de Virologie (EPI_ISL_414624, EPI_ISL_416494); Centre Hospitalier Régional Universitaire de Nantes Laboratoire de Virologie (EPI_ISL_414625); Centre Hospitalier Compiègne Laboratoire de Biologie (EPI_ISL_414627, EPI_ISL_414629-30, EPI_ISL_414634-8, EPI_ISL_415653-4, EPI_ISL_416495-7, EPI_ISL_418218, EPI_ISL_418220-1, EPI_ISL_418223-5, EPI_ISL_418227-8, EPI_ISL_418231, EPI_ISL_418236-9); Hôpital Robert Debré Laboratoire de Virologie (EPI_ISL_414631-2); Centre Hospitalier René Dubois Laboratoire de Microbiologie - Bât A (EPI_ISL_414633); Hôpital Instruction des Armées - BEGIN (EPI_ISL_415650); CH Jean de Navarre Laboratoire de Biologie (EPI_ISL_416493, EPI_ISL_420044, EPI_ISL_420053); Institut Médico legal - Hop R. Poincaré (EPI_ISL_416498); LABM GH nord Essonne (EPI_ISL_416498); Hopital franco britannique - Service des Urgences (EPI_ISL_416501); CHRU Pontchaillou - Laboratoire de Virologie

(EPI_ISL_416502, EPI_ISL_416504-6, EPI_ISL_416508-13); CHU - Hôpital Cavale Blanche - Labo. de Virologie (EPI_ISL_418219); CHRU Bretonneau - Serv. Bacterio-Virol. (EPI_ISL_418222); EHPAD - Résidences les Cèdres (EPI_ISL_418226); Hopital franco britannique - Laboratoire (EPI_ISL_418229); Clinique AVERAY LA BROUSTE, Med. Polyvalente (EPI_ISL_418230); Service des Urgences (EPI_ISL_418232-3); Cabinet médical (EPI_ISL_418235); Sentinelles network (EPI_ISL_420038, EPI_ISL_420045, EPI_ISL_420055, EPI_ISL_421514), L'Air du Temps (EPI_ISL_420039-40); CH Compiègne Laboratoire de Biologie (EPI_ISL_420041, EPI_ISL_420049-50, EPI_ISL_420056-7, EPI_ISL_421500, EPI_ISL_421509-11); Service de Biologie clinique (EPI_ISL_420042, EPI_ISL_421513); CMIP (EPI_ISL_420043, EPI_ISL_420061); Résidence Villa Caroline (EPI_ISL_420046-7); Service de Biologie Médicale - BP 125 (EPI_ISL_420048, EPI_ISL_420058-60, EPI_ISL_420062, EPI_ISL_420064, EPI_ISL_421501, EPI_ISL_421504-6, EPI_ISL_421512), Résidence Eleusis (EPI_ISL_420051); Résidence les Marines (EPI_ISL_420052); Résidence de maintenon (EPI_ISL_420054); Labo BM - Site de Juvisy - Hopital Général (EPI_ISL_420063); Parc des Dames (EPI_ISL_421502-3); Le Château de Seine-Port (EPI_ISL_421507-8), and unknown (EPI_ISL_414626, EPI_ISL_415649, EPI_ISL_415651-2, EPI_ISL_415649).

The second submitting laboratory was CNR Virus des Infections Respiratoires - France SUD. Samples were collected by CNR Virus des Infections Respiratoires – France SUD (EPI_ISL_416745-6); Institut des Agents Infectieux (IAI) Hospices Civils de Lyon (EPI_ISL_416747-8, EPI_ISL_416750, EPI_ISL_416754, EPI_ISL_416756, EPI_ISL_416758); Centre Hospitalier de Valence (EPI_ISL_416749, EPI_ISL_418414-5, EPI_ISL_418417, EPI_ISL_419168); CHU Gabriel Montpied (EPI_ISL_416751-2); Centre Hospitalier de Bourg en Bresse (EPI_ISL_416757, EPI_ISL_417340, EPI_ISL_418426, EPI_ISL_419183, EPI_ISL_419185-6, EPI_ISL_420620); Institut des Agents Infectieux (IAI), Hospices Civils de Lyon (EPI_ISL_417333-4, EPI_ISL_417336-7, EPI_ISL_417339, EPI_ISL_418420-5, EPI_ISL_418429-31, EPI_ISL_419169-73, EPI_ISL_419177-82, EPI_ISL_419184, EPI_ISL_420604-11, EPI_ISL_420615-6, EPI_ISL_420618-9, EPI_ISL_420621-5); Centre Hospitalier de Macon (EPI_ISL_417338, EPI_ISL_418413, EPI_ISL_419174-6, EPI_ISL_419187-8, EPI_ISL_420612-4); Centre Hospitalier des Vals d'Ardeche (EPI_ISL_418412); GH Les Portes du Sud (EPI_ISL_418416); Centre Hospitalier Saint Joseph Saint Luc (EPI_ISL_418418-9, EPI_ISL_420617); Hopital Privé de l'Est Lyonnais (EPI_ISL_418418-9, EPI_ISL_420617); and Centre Hospitalier Lucien Hussel (EPI_ISL_418428).

5 References

1. World Health Organization, https://covid19.who.int.
2. Shah, V.K., Firmal, P., Alam, A., Ganguly, D., Chattopadhyay, S.: Overview of Immune Response During SARS-CoV-2 Infection: Lessons From the Past. Front Immunol. 11, 1949 (2020). https://doi.org/10.3389/fimmu.2020.01949.

3. Peng, L., Shen, L., Xu, J., Tian, X., Liu, F., Wang, J., Tian, G., Yang, J., Zhou, L.: Prioritizing antiviral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures. Sci Rep. 11, 6248 (2021). https://doi.org/10.1038/s41598-021-83737-5.
4. Naqvi, A.A.T., Fatima, K., Mohammad, T., Fatima, U., Singh, I.K., Singh, A., Atif, S.M., Hariprasad, G., Hasan, G.M., Hassan, M.I.: Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochim Biophys Acta Mol Basis Dis. 1866, 165878 (2020). https://doi.org/10.1016/j.bbadis.2020.165878.
5. Miller, D., Martin, M.A., Harel, N., Tirosh, O., Kustin, T., Meir, M., Sorek, N., Gefen-Halevi, S., Amit, S., Vorontsov, O., Shaag, A., Wolf, D., Peretz, A., Shemer-Avni, Y., Roif-Kaminsky, D., Kopelman, N.M., Huppert, A., Koelle, K., Stern, A.: Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. Nat Commun. 11, 5518 (2020). https://doi.org/10.1038/s41467-020-19248-0.
6. Danesh, G., Elie, B., Michalakis, Y., Sofonea, M.T., Bal, A., Behillil, S., Destras, G., Boutolleau, D., Burrel, S., Marcelin, A.-G., Plantier, J.-C., Thibault, V., Simon-Loriere, E., van der Werf, S., Lina, B., Josset, L., Enouf, V., Alizon, S., the COVID SMIT PSL group: Early phylodynamics analysis of the COVID-19 epidemic in France. Epidemiology (2020). https://doi.org/10.1101/2020.06.03.20119925.
7. Global Initiative on Sharing Avian Influenza Data (GISAID), https://www.gisaid.org.
8. Wymant, C., Hall, M., Ratmann, O., Bonsall, D., Golubchik, T., de Cesare, M., Gall, A., Cornelissen, M., Fraser, C., STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration: PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Molecular Biology and Evolution. 35, 719–733 (2018). https://doi.org/10.1093/molbev/msx304.
9. Sledzieski, S., Zhang, C., Mandoiu, I., Bansal, M.S.: TreeFix-TP: Phylogenetic Error-Correction for Infectious Disease Transmission Network Inference. Pac Symp Biocomput. 26, 119–130 (2021). https://doi.org/10.1142/9789811232701_0012.
10. Didelot, X., Kendall, M., Xu, Y., White, P.J., McCarthy, N.: Genomic Epidemiology Analysis of Infectious Disease Outbreaks Using TransPhylo. Curr Protoc. 1, e60 (2021). https://doi.org/10.1002/cpz1.60.
11. Volz, E.M., Koelle, K., Bedford, T.: Viral Phylodynamics. PLoS Comput Biol. 9, e1002947 (2013). https://doi.org/10.1371/journal.pcbi.1002947.
12. Nascimento, F.F., Reis, M. dos, Yang, Z.: A biologist's guide to Bayesian phylogenetic analysis. Nat Ecol Evol. 1, 1446–1454 (2017). https://doi.org/10.1038/s41559-017-0280-x.
13. Zarrabi, N., Prosperi, M., Belleman, R.G., Colafigli, M., De Luca, A., Sloot, P.M.A.: Combining Epidemiological and Genetic Networks Signifies the Importance of Early Treatment in HIV-1 Transmission. PLoS ONE. 7, e46156 (2012). https://doi.org/10.1371/journal.pone.0046156.

14. Kuchaiev, O., Stevanović, A., Hayes, W., Pržulj, N.: GraphCrunch 2: Software tool for network modeling, alignment and clustering. BMC Bioinformatics. 12, 24 (2011). https://doi.org/10.1186/1471-2105-12-24.
15. Hayes, W., Sun, K., Przulj, N.: Graphlet-based measures are suitable for biological network comparison. Bioinformatics. 29, 483–491 (2013). https://doi.org/10.1093/bioinformatics/bts729.
16. Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D., Bunimovich, L., Costenbader, E., Sexton, C., O'Connor, S., Xia, G.-L., Khudyakov, Y.: QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics. 34, 163–170 (2018). https://doi.org/10.1093/bioinformatics/btx402.
17. Vecchio, F., Miraglia, F., Maria Rossini, P.: Connectome: Graph theory application in functional brain network architecture. Clinical Neurophysiology Practice. 2, 206–213 (2017). https://doi.org/10.1016/j.cnp.2017.09.003.
18. Cardoso, D.M., Delorme, C., Rama, P.: Laplacian eigenvectors and eigenvalues and almost equitable partitions. European Journal of Combinatorics. 28, 665–673 (2007). https://doi.org/10.1016/j.ejc.2005.03.006.
19. Kosakovsky Pond, S.L., Weaver, S., Leigh Brown, A.J., Wertheim, J.O.: HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. Mol Biol Evol. 35, 1812–1819 (2018). https://doi.org/10.1093/molbev/msy016.
20. Campo, D.S., Xia, G.-L., Dimitrova, Z., Lin, Y., Forbi, J.C., Ganova-Raeva, L., Punkova, L., Ramachandran, S., Thai, H., Skums, P., Sims, S., Rytsareva, I., Vaughan, G., Roh, H.-J., Purdy, M.A., Sue, A., Khudyakov, Y.: Accurate Genetic Detection of Hepatitis C Virus Transmissions in Outbreak Settings. J Infect Dis. 213, 957–965 (2016). https://doi.org/10.1093/infdis/jiv542.
21. Poon, A.F.Y., Gustafson, R., Daly, P., Zerr, L., Demlow, S.E., Wong, J., Woods, C.K., Hogg, R.S., Krajden, M., Moore, D., Kendall, P., Montaner, J.S.G., Harrigan, P.R.: Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. Lancet HIV. 3, e231-238 (2016). https://doi.org/10.1016/S2352-3018(16)00046-1.
22. Bartenschlager, R., Lohmann, V.: Replication of hepatitis C virus. J Gen Virol. 81, 1631–1648 (2000). https://doi.org/10.1099/0022-1317-81-7-1631.
23. Lorenzo-Redondo, R., Ozer, E.A., Achenbach, C.J., D'Aquila, R.T., Hultquist, J.F.: Molecular epidemiology in the HIV and SARS-CoV-2 pandemics. Curr Opin HIV AIDS. 16, 11–24 (2021). https://doi.org/10.1097/COH.0000000000000660.
24. Katoh, K., Rozewicki, J., Yamada, K.D.: MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 20, 1160–1166 (2019). https://doi.org/10.1093/bib/bbx108.
25. Przulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geometric? Bioinformatics. 20, 3508–3515 (2004). https://doi.org/10.1093/bioinformatics/bth436.

26. Nei, M., Li, W.H.: Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences. 76, 5269–5273 (1979). https://doi.org/10.1073/pnas.76.10.5269.
27. Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A.: RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35, 4453–4455 (2019). https://doi.org/10.1093/bioinformatics/btz305.
28. Ugurel, O.M., Ata, O., Turgut-Balik, D.: An updated analysis of variations in SARS-CoV-2 genome. Turk J Biol. 44, 157–167 (2020). https://doi.org/10.3906/biy-2005-111.
29. Zheng, Y.-X., Wang, L., Kong, W.-S., Chen, H., Wang, X.-N., Meng, Q., Zhang, H.-N., Guo, S.-J., Jiang, H.-W., Tao, S.-C.: Nsp2 has the potential to be a drug target revealed by global identification of SARS-CoV-2 Nsp2-interacting proteins. Acta Biochimica et Biophysica Sinica. 53, 1134–1141 (2021). https://doi.org/10.1093/abbs/gmab088.
30. Yang, X., Dong, N., Chan, E.W.-C., Chen, S.: Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries. Emerging Microbes & Infections. 9, 1287–1299 (2020). https://doi.org/10.1080/22221751.2020.1773745.
31. Bai, Y., Jiang, D., Lon, J.R., Chen, X., Hu, M., Lin, S., Chen, Z., Wang, X., Meng, Y., Du, H.: Comprehensive evolution and molecular characteristics of a large number of SARS-CoV-2 genomes reveal its epidemic trends. International Journal of Infectious Diseases. 100, 164–173 (2020). https://doi.org/10.1016/j.ijid.2020.08.066.
32. Yang, H.-C., Chen, C., Wang, J.-H., Liao, H.-C., Yang, C.-T., Chen, C.-W., Lin, Y.-C., Kao, C.-H., Lu, M.-Y.J., Liao, J.C.: Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. Proc. Natl. Acad. Sci. U.S.A. 117, 30679–30686 (2020). https://doi.org/10.1073/pnas.2007840117.