Navigating the COVID-19 Data Landscape: Automated Hypothesis Generation using Topological Data Analysis

Methun Kamruzzaman (hkz8wk@virginia.edu)
University of Virginia
Matthew Bielskas
University of Virginia
Bala Krishnamoorthy
Washington State University Vancouver
Achla Marathe
University of Virginia
Anil Vullikanti
University of Virginia
Ananth Kalyanaraman
Washington State University

Research Article

Keywords: Covid-19, Topological Data Analysis, mapper, narrative, epoch, path, are.

Posted Date: May 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-470082/v1

Navigating the COVID-19 Data Landscape: Automated Hypothesis Generation using Topological Data Analysis

Methun Kamruzzaman[2*], Matthew Bielskas[2], Bala Krishnamoorthy[3], Achla Marathe[2,4], Anil Vullikanti[2,5], Ananth Kalyanaraman[1]

1 School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA
2 Biocomplexity Institute, University of Virginia, Charlottesville, VA
3 Department of Mathematics and Statistics, Washington State University, Vancouver, WA
4 Department of Public Health Sciences, University of Virginia, Charlottesville, VA
5 Department of Computer Science, University of Virginia, Charlottesville, VA

* corresponding author; hkz8wk@virginia.edu

Abstract

Background: There have been significant spatio-temporal variations in how the COVID-19 pandemic has unfolded in different parts of the country. While a huge amount of data related to the COVID-19 pandemic has become available including data on disease outcomes, different kinds of behaviors and diverse interventions, along with a number of co-variates, analyzing this data to characterize the spatio-temporal variations and gleaning actionable insights and hypotheses on different factors driving the pandemic remains a big challenge. The **objective** of this study is to identify in an unsupervised fashion the key spatio-temporal patterns, anomalies, and associated factors in the spread of COVID-19 in different regions that can be used in the development of models and the planning of public health policies.
Methods: We present a topological data analysis (TDA) framework for exploring COVID-19 data that supports two types of analytical functions: i) discover *disease epochs* in the trajectories that reveal spatio-temporal events of interest; and ii) modeland better elucidate *interactions* between variables of different disease outcomes (e.g., number of cases, hospitalizations, deaths) and intervention mechanisms (e.g., social distancing, contract tracing).
Results: Our TDA framework reveals several insights in an automated manner by identifying co-evolving and divergent cohorts of states with respect to various disease outcomes (e.g., number of new cases) and measures of behavior or interventions (e.g. social distancing, COVID exposure, hospital bed utilization). Our framework also identifies the branching points at which the different cohorts start evolving separately.
Conclusions: The illustrative case studies show that our TDA-based analytical framework can help navigate the epidemic data landscape in an automated and guided manner, and can provide insights to formulate hypotheses and devise sound, data-aided public health policies.

Keywords: Covid-19, Topological Data Analysis, mapper, narrative, epoch, path, flare.

1 Background

COVID-19 has become one of the deadliest pandemics in human history. Spread of such a highly infectious disease in large populations is a complex dynamical process with a large number of factors at play, including the social contact structure in the population, individual variability in susceptibility and response, a wide variety of individual behaviors (e.g., use of masks and other PPEs, compliance with social distancing and level of mobility), and various interventions and response strategies implemented by governments at local, state, and national levels (e.g., closure of schools, and travel restrictions). But unlike past pandemics, a huge amount of data on COVID-19 has become available in short time [23–30] including: (1) time series of disease outcomes, e.g., number of cases, hospitalizations, deaths, and (2) behaviors, e.g., social distancing, PPE usage.

These variables have complex interactions, which have resulted in significant spatio-temporal variations in the epidemic profiles across the states within US, and across countries. For instance, we find that different states exhibit different outcomes even when similar interventions are implemented, e.g., New York peaks in late Spring vs. California peaks later in Summer. Finding such spatio-temporal patterns could be highly insightful for policy planners, as they try to incorporate the lessons in real-time, from previous incidences elsewhere or consider local factors (e.g., demographics). However, this requires (a) discovery of such spatio-temporal patterns from the data, and (b) narratives that would help explain patterns in a manner that is consistent with raw data observations and scientific models. These requirements motivate our work.

Contributions: We present a principled approach to analyze COVID-19 and other pandemic data. Specifically, our approach utilizes the spatio-temporal depths of disease variables to extract two types of insights.

i) Our approach can be used to discover **disease epochs**, which are parts of the trajectory that show specific patterns of behavior—co-evolution or divergence—in an unsupervised and automated fashion. *Co-evolution* refers to different regions that exhibit similar disease outcomes around the same time interval. *Divergence* refers to branching events where different regions start to diverge in their behavior.

ii) Our approach can be used to elucidate the **interaction** between key variables over space and time, e.g., the number of cases affected by interventions such as social distancing and testing, in different regions of the country and at different time intervals; or correlations between incidence rate and region-wise threshold in testing and tracing.

These disease epochs and interactions can be used to formulate narratives for policy makers in real-time, and guide more targeted interventions as the pandemic progresses. Fig. 1 shows an example output generated by our approach.

Our approach is based on topological data analysis (TDA). Topology is the branch of mathematics that deals with structure of spaces. TDA is an emerging area with demonstrated ability to glean insightful structural information in the context of many applications [15]. However, application of TDA to epidemics data is not a straightforward exercise. It requires significant domain-specific customization and adaptation.

In this work, we adapt the `Hyppo-X` platform [11], an open source implementation for TDA previously applied in the contexts of plant phenomics [9, 12] and antibiotic use in

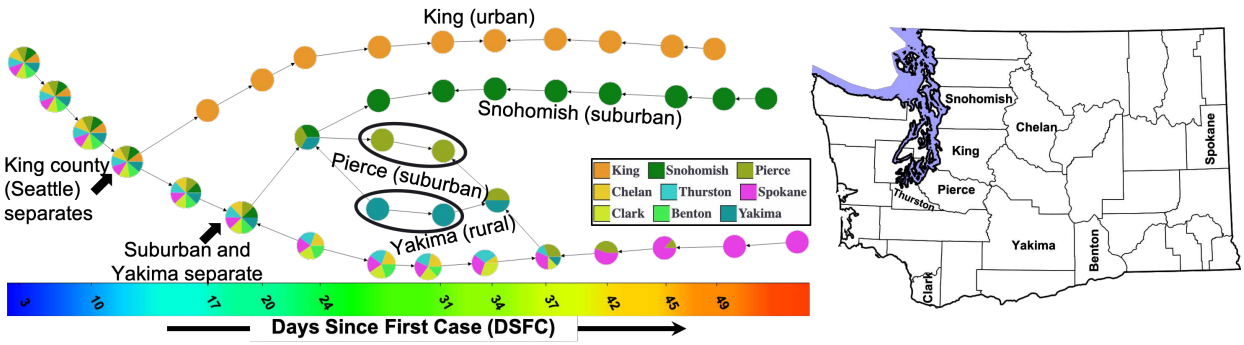


Figure 1: A topological object created by our approach for nine counties of Washington state (right), during their first 50 days of COVID-19. The data consists of time-series of DSFC for each county. Nodes represent clusters of counties (shown as pie-charts) that show similar growth rate in COVID-19 cases at different time intervals (horizontal bar). Edges connect clusters that intersect in points, and edge direction points to the direction of increase in the case rate. Multiple subpaths corresponding to epochs of co-evolving counties and branching events are visible.

hospitals [16] to COVID-19 data. Through illustrative case studies we show that specific features such as paths and *flares* (defined later) in the TDA object can be used to derive new insights. However, these features have to be tailored to epidemic data, as there can be non-monotonicity along paths.

Insights from our approach make it amenable for narrative generation and hypothesis formulation. The data-driven hypotheses generated by TDA provide more concrete and credible questions to be investigated by researchers and public health experts, and would likely lead to more credible explanations. Fig. 1 shows an example of a TDA analysis, which could be used for structured knowledge representation—this can be combined with automated narrative generation using templates [13].

1.1 Related Work

Infectious disease modeling and limitations: Mathematical modeling methods, such as metapopulation models, statistical models and agent-based models, play a significant role in pandemic response [1, 3, 6, 7, 17, 20, 22]. Metapopulation and statistical models are easier to set up than agent-based models (ABM), but ABMs are more adaptable to study micro level behaviors. However, all such models have to be calibrated to disease outcome data [21]. The significant changes in spatio-temporal dynamics due to exogenous factors (e.g., new travel restrictions) make calibration, and hence modeling very challenging. This motivates principled methods, such as ours, that use *raw incidence data* to find key spatio-temporal epidemic patterns.

TDA and Applications TDA has several key advantages over traditional data analysis and dimensionality reduction techniques such as PCA, multi-dimensional scaling, and cluster analysis, owing to its robustness, visualization capabilities, and coordinate-free properties

[15]. Since initial work on the `Mapper` framework by Singh et al. [19], TDA has gained wider use with a demonstrated ability to glean insightful structural information in the context of many applications such as cancer [14], microbial ecology [8], wildfires [2], agricultural phenomics [12], and more. To the best of our knowledge, our effort represents one of the first to adopt TDA for epidemics.

2 Methods

`Hyppo-X` [12] is a specific implementation of the `Mapper` TDA framework [19], along with functions to mine various structural features [9,10]. We adapt the tool for epidemic use cases by: (1) identifying new structural features that capture the complexity of epidemic outcomes, and (2) modifying the parameter choices and implementations of the original framework.

Our methods construct a highly compact representation of high dimensional COVID-19 spatio-temporal data that reveals interesting epochs in disease progression and interplays between significant variables, and aids in the generation of hypotheses and explanations via visual exploration.

Infectious disease data have several distinctive traits: i) most variables of interest are continuous and time-varying; ii) variables may be interdependent (e.g., outcome of an intervention, or intervention to an outcome); iii) response of one variable to another could have a time delay (or lag); and iv) spatial connectivity, in addition to local policies, may affect a variable's evolution.

Input: A high dimensional point cloud where each "point" represents a distinct [location, time] combination (see Section 3.1). The location can be a county, state, or a country; time is the date of observation as recorded in raw incidence data (or line lists). The dimensions represent various disease attributes recorded for each point. These include variables of disease outcome, e.g., number of new cases, hospitalizations, deaths, exposure, etc.; and variables that capture intervention or capacities, e.g., social distancing index, number of tests, contact tracing, and PPE availability.

Navigation: The output is a compact visual representation of the high-dimensional point cloud, which can help elucidate (a) the interplay between different variables of choice; and (b) reveal interesting periods (or *epochs*) in disease progression. We focus on two types of epochs.

i) Epochs of *co-evolution* refer to intervals in time where points from different geographic locations show similar disease trajectories, e.g., growth trends in the number of cases, or case-positivity rates. This information can be used by public health officials to understand the effects of policies, and to coordinate them across regions.

ii) Epochs of *divergence* refer to intervals in time where a branching event in disease trajectories is observed, in which various geographical locations show divergent behavior. Domain experts can query branches in such epochs, identify attributes that vary substantially across the branched paths, and use them in explanations.

A third type of insight we want to uncover is time-agnostic: are there different geographical locations that show similar "behavior" across two different time periods? For instance,

while forming the intervention strategy for location X, knowing how another location Y, which had a similar set of attributes, responded to an intervention strategy could be helpful.

Modeling interplay between variables and topological object construction. We capture the interplay between different subsets of variables using the notion of *filter* and *performance* variables of a `Mapper`. A *filter* f is a variable used to bin the points into a predetermined set of overlapping fixed-size intervals. For instance, with time as a filter, a 100-day pandemic data set can be broken down into 20 bins of size 6 days each, assuming a 20% overlap. More generally, the filter function could be high dimensional where multiple variables are combined to determine the binning [15, 19]. In this work, we consider multiple variables as individual filter functions. A *performance* variable, on the other hand, is used to separate points that fall within each bin into disjoint clusters. Each cluster represents a set of points that share the same range of values in the filter space and exhibit similar performance.

For the epidemic use case, we select variables that are typically treated as potentially "causal" by a domain expert as filters (e.g., social distancing, time), while performance variables capture typical disease outcomes (e.g., new cases, hospitalizations). Note that we can use multiple filters and multiple performance variables, and we can use this feature to study the interplays between multiple sets of variables.

Topological object example: Fig. 1 shows an example topological object built using time measured in Days Since First Case (DSFC) as the filter, and rate of change of new cases as the performance variable. Given a high dimensional point cloud X, a choice of filters $F = \{f_1, \ldots, f_k\}$, and a pairwise distance function g defined over a (set of) performance variable(s), we build a topological object [12] using the following major steps:

1. For each of the filters $f_i \in F$, create $n_i > 0$ buckets that evenly subdivide the range of values of f_i and assign a centroid c_{ij} for each bucket ($j \in [1, n_i]$). In Figure 1, we chose the set of filters $F = \{\text{DSFC}\}$ and we created $n_1 = 15$ buckets for this filter.
2. With reference to its centroid, expand each bucket along each dimension (for k filters it will be k dimensions) by a percentage α_j for $j \in [1, k]$. This expansion creates regions of overlap between adjacent buckets. In Figure 1, each of the original buckets (except the terminal buckets) were expanded by 25% along both directions (left and right). The left most bucket expanded to the right direction and the right most bucket expanded to the left direction by the same percentage.
3. Identify all points that belong to each expanded bucket, i.e., points whose f_i values fall within the corresponding range for f_i specified by the bucket. Apply a distance-based clustering method, e.g., DBSCAN [5], on the performance variable(s) g to separate points within each k-dimensional bin into clusters. The DBSCAN parameters used in Figure 1 are: a) density=2 and b) radius=10. With these choices of clustering parameters, we clustered all points in each bin based on their *rate of change of number of cases* values.
4. Construct a directed graph $G(V, E)$ where every node is a cluster and an edge connects any two nodes whose clusters intersect. Since clusters from a single bin do not share points, edges connect clusters from adjacent (or nearby) bins that both include points from the overlapping regions of their bins. Edges are directed toward nodes with higher cluster

mean of (a chosen) performance variable. In the topological object shown in Figure 1, the numbers of bins n_i and overlap percentages α_j were chosen such that each edge connects a node representing a cluster from a bin to another node representing a cluster from an adjacent bin that shares at least one point with the first cluster.

We employ ideas from persistent homology [4] to choose the number of bins and overlap % so that the final topological object is *stable*. For instance, with the number of bins for a filter f_i fixed, we vary the % overlap across a range of values, and construct the topological objects for each overlap %. We identify a large range of overlap values across which the number of components in the object does not change, and choose the midpoint of this range as the nominal overlap %.

Extraction of Epochs: Exploration of the topological object begins with the domain expert driving the navigation and extraction process, as guided by some of the automated feature extraction capabilities in the tool.

Co-evolution: Co-evolving epochs are obtained by detecting simple, i.e., unbranched, paths in the trajectories that consist of points from two or more geographical locations. This process is enabled by selecting the pie-chart view mode for the topological object, where each cluster is shown as a pie-chart. For instance in Fig. 1, the bottom-most segment from DSFC interval [20,38] consists of a combination of five WA State counties, showing similar number of new cases observed within this period of time.

Divergence: Divergent epochs are identified by parts of the trajectories with branches. Such a feature is called a *flare* [10], and consists of a "stem" leading up to a branching node, with two or more directed paths emerging out of it. Each of these directed paths suggests a monotonic increase in the performance variable along the clusters of that path. However, in the case of epidemic data, we study two sub-types of flares:

a) *Forks* are flares where the branching paths do *not* merge later, suggesting a permanent or long-term monotonic separation between the different paths.
b) *Loops* are flares where two branching paths merge later, suggesting a short-lived monotonicity along one of the paths before it reverses course to re-merge with the other path, suggesting an "ebb and flow" of the disease trajectory.

When there is such ebb and flow, one of the merging paths typically represents a *spine* that consists of a trail of clusters where a majority of the points lie. Techniques from subgraph mining [18] can also be used for extraction of epochs, though they have to be extended to incorporate the additional information associated with nodes.

In Fig. 1, we observe all these cases of flares—with the King and Snohomish counties in Washington State, representing forked paths, while the more rural/suburban counties such as Pierce and Yakima looping back to the spine represented by the other five counties.

Time-agnostic features: Time is typically selected as one of the filters to capture the temporal nature of the dynamic phenomenon. However, its inclusion as a filter is not required. The domain expert may want to observe epochs from across the time scale—for instance, States observing a certain degree of intervention, e.g., social distancing, that also

exhibit a similar level of case-positivity rate. A careful selection of filters can allow the domain expert to detect structural features in a time-agnostic fashion.

Spatial features: Our approach also allows a domain expert to uncover spatial relationships in disease characteristics. For instance, it has been well understood that disease rates at neighboring counties can influence the disease rates locally. Such spatial coherence can be captured in our framework by including spatial attributes (e.g., lat/long) as filters.

3 Results

3.1 Experimental Setup

Data sets: We use data from the University of Maryland COVID-19 Impact Analysis Platform (IAP)[a], which has data for all of U.S. (both state- and county-level) available for a duration of 252 days of the pandemic since January 2020. Each "point" is a [state,day] combination, and we have $N = 9,119$ points. Each daily entry has more than 40 variables. For our analysis, we selected a subset based on a preliminary Pearson correlation analysis and domain expert's advice:

i) *Social Distancing Index (SDI)*: An integer in $[0, 100]$, with 0 indicating no social distancing and 100 indicating total compliance. We took a 7-day moving average using daily SDI, since we observed a weekly pattern;
ii) *New cases*: Daily number of confirmed COVID-19 cases per $1,000$ people. We used a 7-day moving average for this variable as well.
iii) *COVID Exposure*: Daily number exposed to the virus per $1,000$ people. We took a 14-day moving average following the expected incubation period[b];
iv) *Tests done:* Daily number of tests per $1,000$ people; and
v) *%Hospital bed utilization*: Daily percentage of hospital beds occupied by COVID-19 patients.

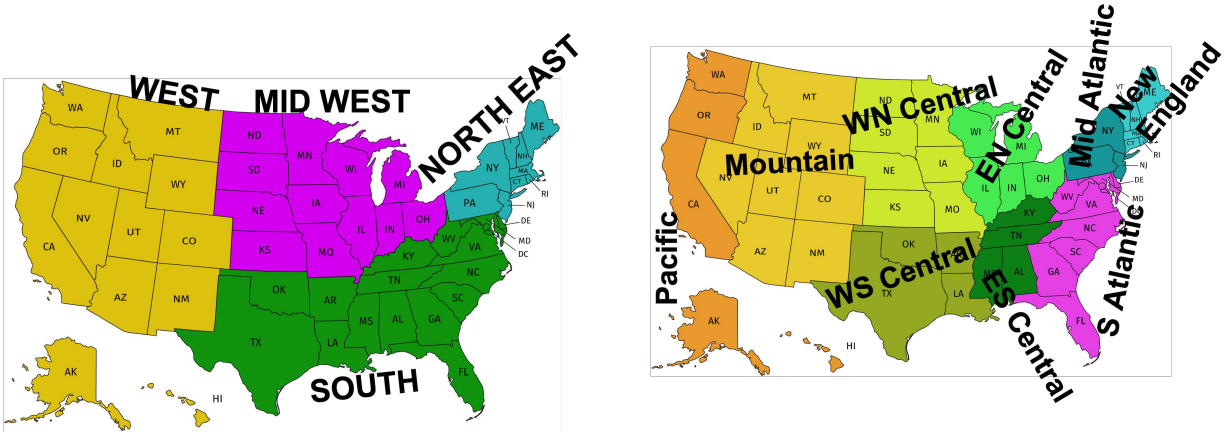
In addition, we use two derived variables: *Days Since First Case (DSFC) and case-positivity rate*. DSFC is the number of days that have elapsed since the day of the first case observed in a state. Even though data is available from 1/1/2020, states recorded their first case at different dates. We mark those days as day 0 for each state. Results using case-positivity rate are omitted due to space limitations.

For our spatial studies, we used data from all 50 states of the U.S. and District of Columbia (D.C.). Given the large number of states, and to keep the visual objects tractable, we use two ways of grouping the states into a) regions and b) divisions, as shown in Fig. 2a.

3.2 Case Studies

We present a set of concrete case studies using the modified implementation of `Hyppo-X` on COVID-19 data sets. The goal is to evaluate the ability of TDA to produce data-driven insights and hypotheses on COVID-19 data, using automated knowledge representation and subsequent navigation. A vast number of experiments can be performed with TDA, using

[a] `https://data.covid.umd.edu/`
[b] `https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html`



(a) (b)

Figure 2: U.S. map by: a) regions and b) divisions.

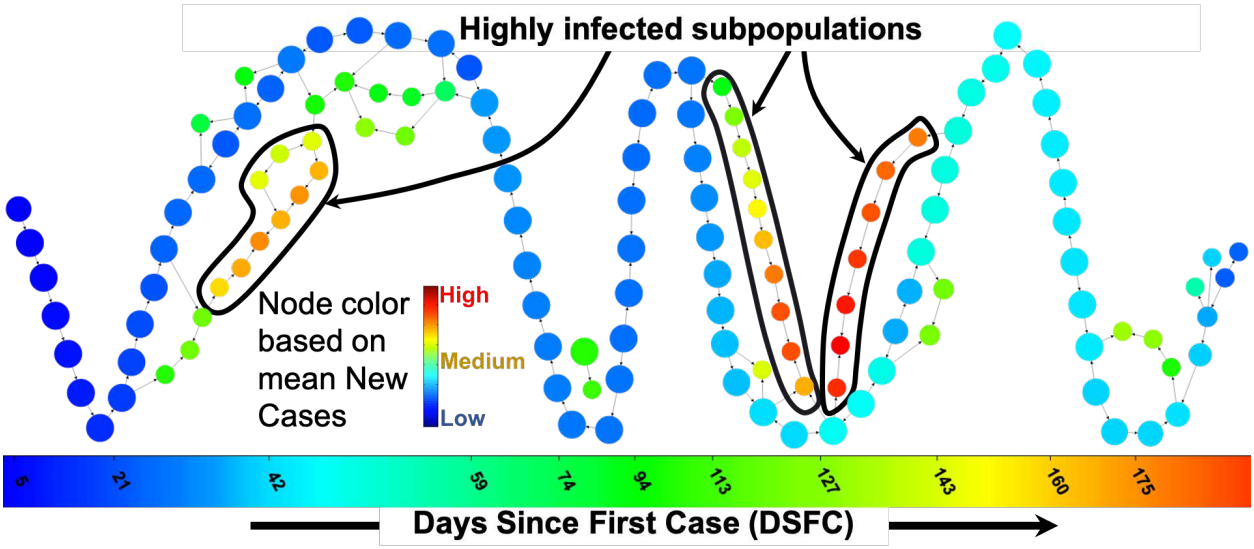
different combinations of variables. We select a handful of experiments for our case studies, keeping in mind the natural choices for a domain expert. For each case study, we formulate a question that can provide one or more testable hypotheses to public health experts. For example, how do different points in space and time co-evolve or diverge in behavior? From this information we aim to learn more focused sub-population level variations and similarities in behavior, as opposed to population level effects.

Case Study** 1**: Decoding of co-evolving and divergent groups of states with time. What subset of states co-evolved and what subset of states diverged from one another, as the pandemic spread through the U.S.? To answer this question, we ran `Hyppo-X` using DSFC as the filter and new cases as the clustering attribute. Fig. 3 shows the resulting topological object. The key observations are as follows:

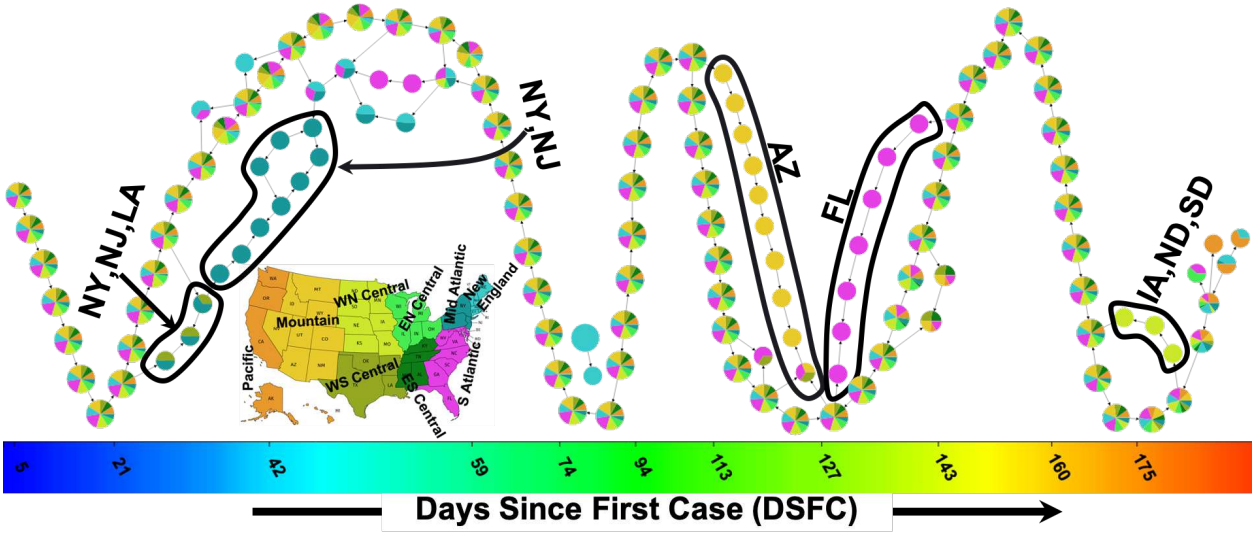
1. Two clear peaks are visible around DSFC epoch: [29,46] and DSFC epoch: [114,136].
2. The peaks can be attributed to two different cohorts of states. The first peak is owing to the co-evolving group mainly consisting of NY and NJ; whereas the second peak is owing to AZ and FL.
3. In all cases of flares, we see loops (i.e., the branched paths merge back to the spine). This demonstrates the ebb-and-flow characteristic expected in epidemic curves.

Case Study** 2**: How does social distancing relate to new cases? Intuitively, social distancing index (SDI) is expected to influence (as an intervention) and be influenced (high SDI due to high cases) by new cases. To capture the interaction between these two variables, we used SDI as the filter and new cases as the clustering attribute. The resulting topological object is shown in Fig. 4. The key observations are as follows:

1. Clusters with a higher number of cases appear more toward the higher end of SDI spectrum. While this may look counter intuitive, a higher number of cases increases the desire to comply with social distancing measures. The North Eastern states were some of the first states to get hit as they started to institute stricter protocols.

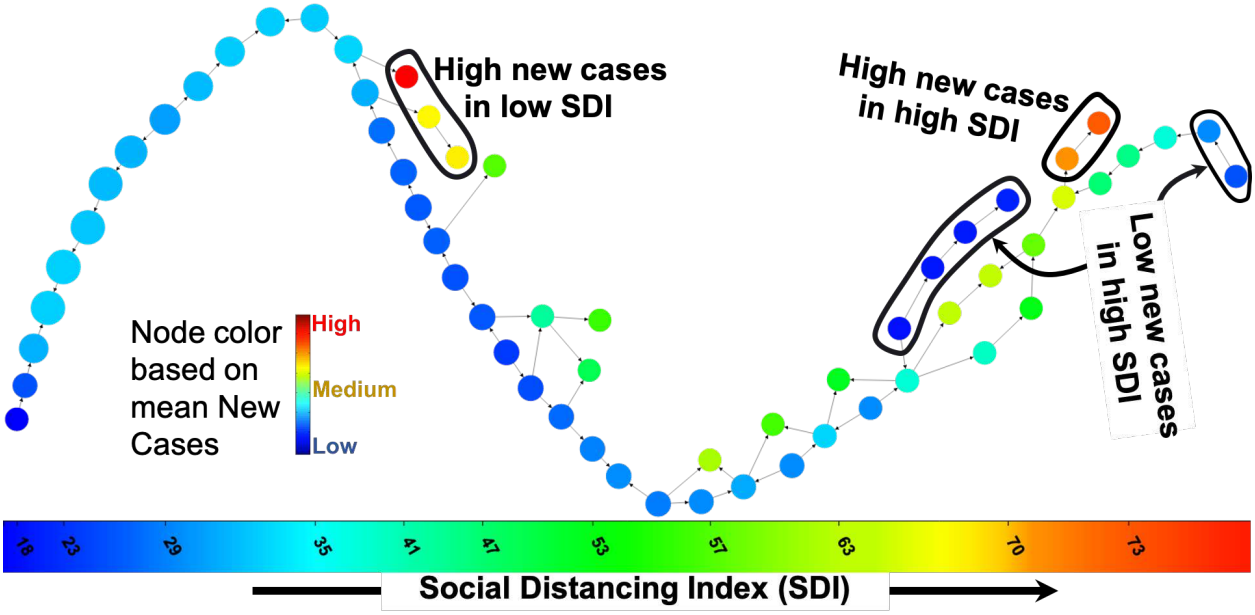


(a) Nodes (clusters) colored by new cases

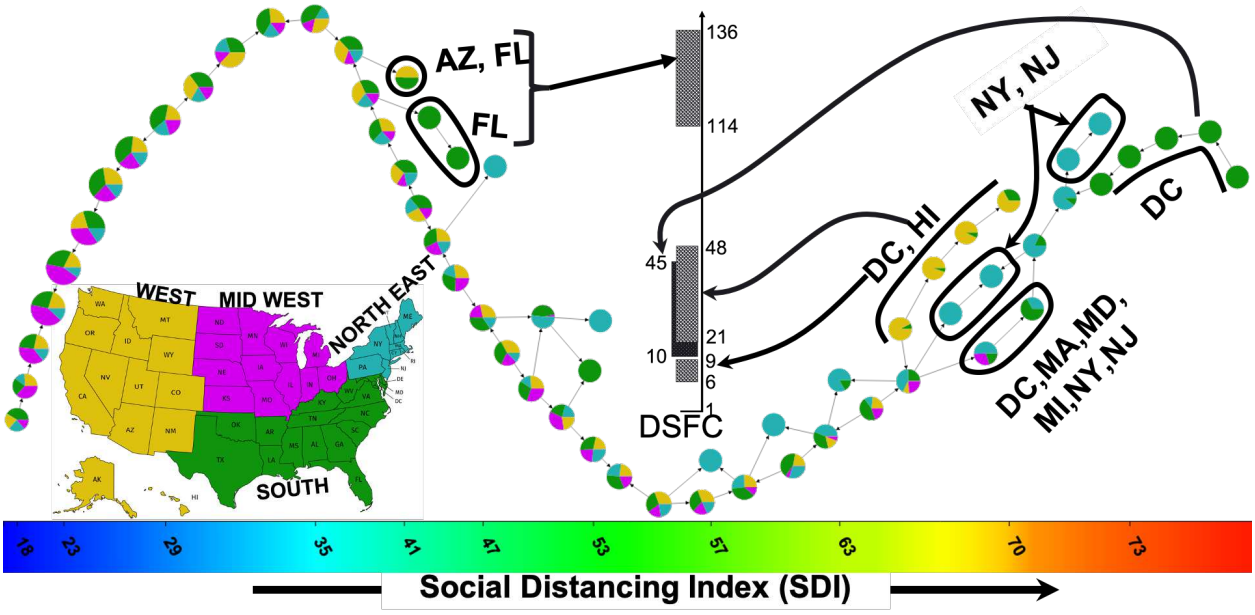


(b) Nodes (clusters) shown as a pie-chart of U.S. divisions

Figure 3: Topological object with DSFC as filter and new cases as clustering attribute. `Hyppo-X` parameters: # bins: 80; overlap: 35%; and DBSCAN (density: 2, radius: 0.032) for clustering. The horizontal color bar shows the DSFC scale. There is *no* Y-axis—the object is simply placed to match various intervals of the time scale. Each node is a cluster of points ([state,DSFC]) that had similar number of cases in the same time interval, and each edge (directed) connects two intersecting clusters.



(a) Nodes (clusters) colored by new cases



(b) Nodes (clusters) shown as a pie-chart of U.S. regions

Figure 4: Topological object with Social Distancing Index (SDI) as filter and new cases for clustering. `Hyppo-X` parameters: # bins: 40; overlap: 35%; and DBSCAN (density: 2, radius: 0.035) for clustering. Horizontal color bar shows the SDI scale, and each node is a cluster of [state,SDI] observations that can come from different points in time.

2. On the same end of the spectrum, we also note states such as DC and HI that also have a high SDI but very low cases. This is because those states started implementing shutdowns proactively (following other states). We confirm this by observing that the epochs corresponding to DC and HI cohort branch is [6,48] (early in the season).
3. In contrast, AZ and FL on the left end of the SDI spectrum exhibit a large number of new cases and low SDI. In fact, by examining Fig. 3b alongside, we confirm that AZ and FL form the second wave of peaks (DSFC epoch [114,136]). This shows that the states AZ and FL were mostly responding to the larger number of cases with higher social distancing, instead of being proactive.

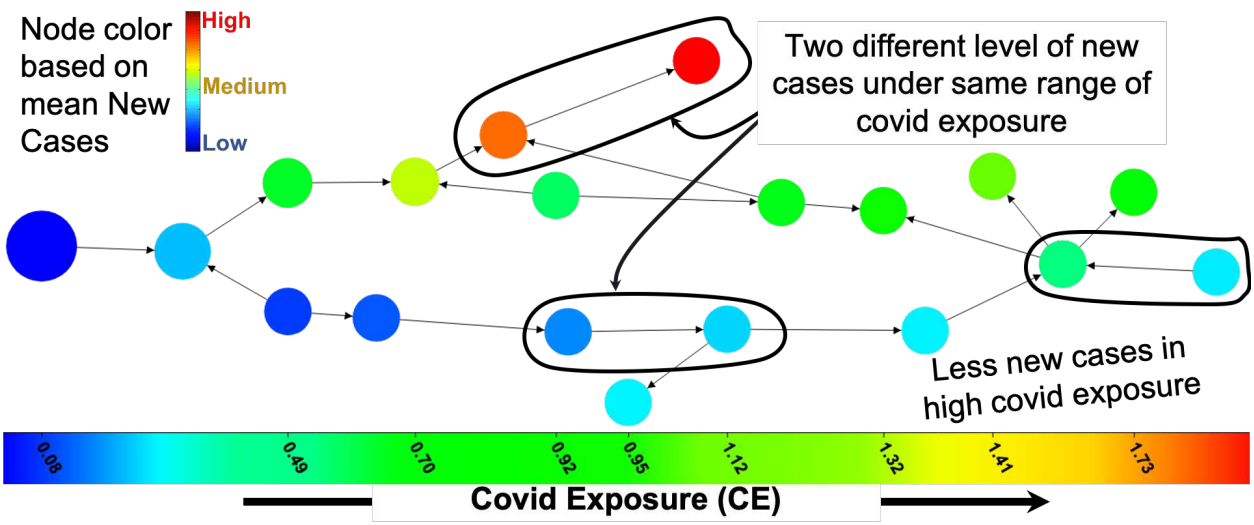
This case study provides two insights: a) visualizing TDA objects generated from two (or more) different filter studies could be more revealing/explanatory; and b) TDA-based observations can potentially help in formulating hypotheses that are more action driven, e.g., could AZ and FL have recorded smaller number of cases had they adopted SDI early on? While it will be too simplistic to associate one variable as causal, this data-driven study certainly points toward an important role for SDI in controlling spread. Use of multiple filters (using other intervention measures) could reveal a more complete and robust picture.

We note that one of the factors *not* captured in this study is the potential role of *lag* between these two variables. Social distancing, as with other interventions, is expected to take some time to exert an influence in the new cases. We conducted a correlation study taking into account a wide range of lag values. We found that different states have different lags at which these two variables are most highly correlated. For instance, lags with the highest correlations for CA and FL states were 18 and 29 days, respectively. Hence we did not explore the effect of lags further. But in general, one could use our TDA framework on data modified with varying lag values to parse out its effects.

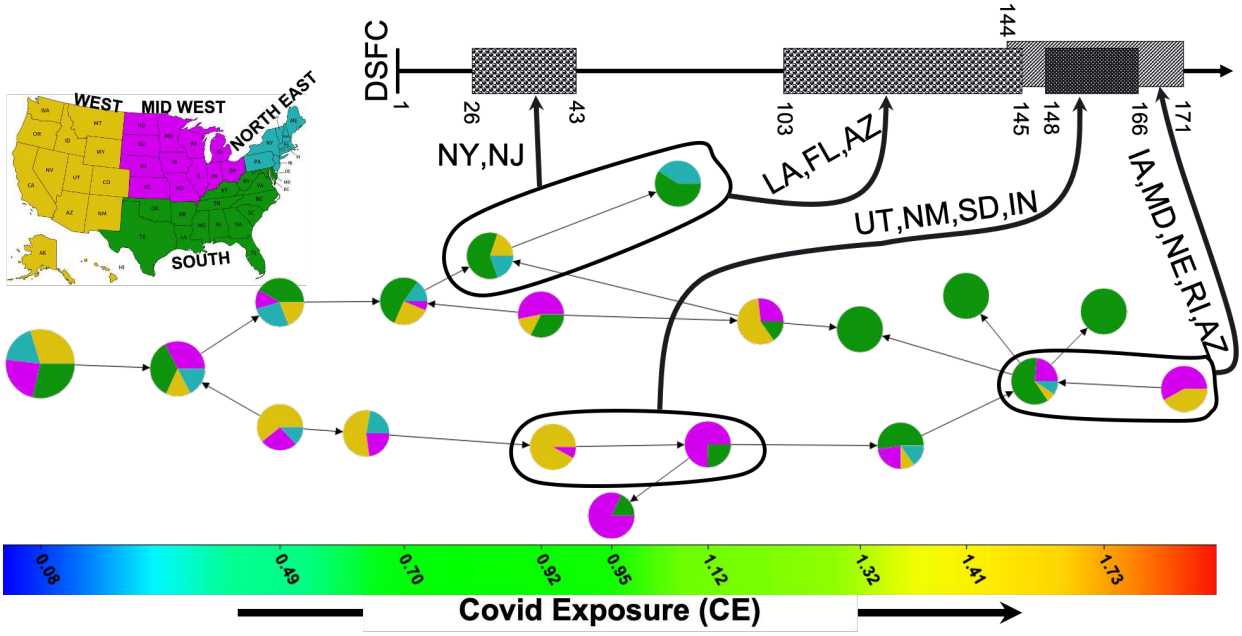
Case Study 3: How does COVID exposure correlate with new cases? COVID exposure represents the estimated number of people potentially exposed to the virus per 1,000 people. It is computed mainly using contact tracing, and hence is a good indicator of the state's tracing capability. Fig. 5 shows the topological object obtained using Covid exposure as the filter and new cases as the clustering attribute. The key observations are as follows:

1. While high exposure is expected to correlate with high number of cases in general, clusters with large numbers of new cases appear through a wide range of exposure.
2. The mid-range exposure values show two divergent subgroups: NY and NJ occupy the top branch with higher cases and UT and NM (with SD, IN) occupy the lower branch with low cases. Upon a closer examination, we found the DSFC epochs associated with both these groups to be significantly different, i.e., NY and NJ: [26,43] whereas UT and NM: [148,166].

At the start of the pandemic, states that were impacted early such as NY and NJ were not set up to do contact tracing, and hence had low correlation with exposure. On the other hand, states such as UT and NM had the advantage of time to be able to put in resources for better tracing and tracking, which resulted in higher correlation with new cases.



(a) Nodes (clusters) colored by new cases



(b) Nodes (clusters) shown as a pie-chart of U.S. regions

Figure 5: Topological object with COVID Exposure (CE) as filter and new cases for clustering. `Hyppo-X` parameters: # bins: 10; overlap: 10%; and DBSCAN (density: 2, radius: 0.01) for clustering. Horizontal color bar shows the CE scale, and each node is a cluster of points [state,CE]. Part (b) shows each node as a DOY (day of year) bar corresponding to active epochs.

Case Study 4: How does hospital bed utilization rate correlate with new cases? The (estimated) percentage of hospital beds occupied by COVID-19 patients (%HBU) points to the capacity of a state to handle the pandemic. We can expect a larger hospitalization rate when new cases increase. This is confirmed by the topological object shown in the inset in Fig. 6. However, there are a few other observations:

1. We observed more branching (in the form of loops) toward the higher end of the %HBU spectrum, suggesting a diversity in the number of new cases.
2. Furthermore, we observed outlier states (short flares growing from the spine): IA, SD for low %HBU, and MS and (AZ, FL) for mid-range %HBU. These states peaked late summer, while their hospitalizations were lower.

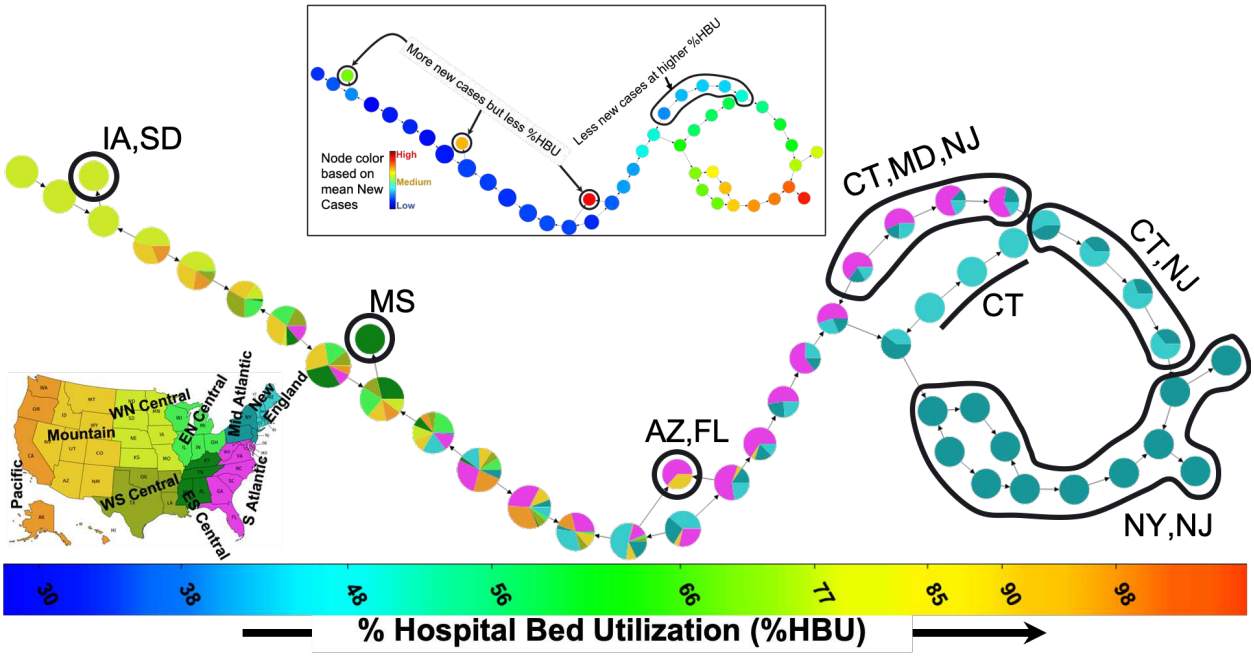


Figure 6: Topological object with %Hospital Bed Utilization (%HBU) as filter and new cases for clustering. `Hyppo-X` parameters: # bins: 30; overlap: 45%; DBSCAN (density: 2, radius: 0.03).

4 Discussion

Our results on complex spatio-temporal datasets show that the TDA framework and epochs—co-evolving and divergent—provide useful insights into epidemic outcomes and efficacy of intervention. As with any unsupervised learning method, specific results depend on a multitude of parameters. But observations that are robust across many settings, e.g., co-evolution epochs with AZ and FL across multiple filters, suggest underlying connections. With more follow up analyses, such observations can be translated into testable hypotheses.

Our tool is ready to be used by public health researchers and professionals: (1) the modified implementation of `Hyppo-X` is available open source [11], along with an interactive display that has been used by domain experts to navigate the entire object, examine values

in the clusters, and change visualizations easily; and (2) we have set up the tool with publicly available COVID-19 datasets, and the analyses can be done easily as the data gets revised.

Our analysis also suggests effective values for the parameters, though a user can easily explore variations. The `Hyppo-X` framework already supports identification of different kinds of features such as paths and flares. Testable hypotheses can be easily derived from these features.

Our TDA framework is not intended to be, and not designed to function as, a predictive model for any Covid-19 measure. It forms a foundational step that can reveal insights from Covid-19 which can then be used as a guide to devise interventions or policies. In the longer term, our methods can be extended in a number of ways. We can couple our analysis with other statistical and analytical methods, providing a way to test many of these hypotheses, e.g., through statistical and simulation based models. We should also be able to incorporate more sophisticated subgraph mining techniques on the TDA objects to identify other kinds of features, which might become visible only at certain spatio-temporal scales; but this could become challenging with a large number of filters. Epochs and other related features provide structured knowledge representations, which can be coupled with narrative generation techniques from NLP, e.g., [13], and assist agencies such as the CDC in providing weekly summaries of the pandemic evolution.

Application of TDA methodology is novel in the public health domain. There are no other tools that can provide this kind of spatio-temporal analysis and visualization of the data. TDA enables analysis of co-evolution of patterns, feedback between variables, splitting and reconnecting of co-occurring events, among other things, in an automated way. This data-driven methodology allows researchers to generate hypotheses that are supported by the data and not biased by the prior beliefs of the user. The illustrative case studies demonstrate the utility of TDA methodology in the context of COVID-19 pandemic. It provides another tool to researchers and public health professionals who are feeling inundated with the data as cases and deaths rise and the urgency to understand and treat this disease becomes more dire by the day.

5 Conclusions

We present a TDA-based analytical framework for exploring the structure of COVID-19 data that identifies key spatio-temporal patterns, anomalies, and associated factors in the spread of COVID-19 in different regions, all in an unsupervised manner. We present several illustrative case studies which show that our framework can help navigate the epidemic data landscape in an automated and guided manner, and can provide insights to formulate hypotheses and devise sound, data-aided public health policies.

6 List of Abbreviations

The following abbreviations have been introduced and used in the paper.

TDA	Topological Data Analysis
DSFC	Days Since First Case
IAP	Impact Analysis Platform
SDI	Social Distancing Index
CE	Covid Exposure
DOY	Day of Year
HBU	Hospital Beds Utilized

7 Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The datasets analyzed in the current study are taken from the University of Maryland COVID-19 Impact Analysis Platform (IAP), available at `https://data.covid.umd.edu/`. The computational analysis was carried out using our TDA platform `Hyppo-X` [11] available at `https://xperthut.github.io/HYPPO-X`.

Competing interests: The authors declare that they have no competing interests.

Funding: National Science Foundation (NSF) grant DBI-1661348.

Authors' contributions: All authors contributed equally.

Acknowledgements: Not applicable.

References

1. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco J J, and Vespignani A Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences 2009; (106):21484–21489.

2. Bendick R, and Hoylman Z H Topological data analysis reveals parameters with prognostic skill for extreme wildfire size. Environmental Research Letters 2020.

3. Chinazzi M, Davis J T, Ajelli M, Gioannini C, Litvinova M, Merler S, Pastore y Piontti A, Mu K, Rossi L, Sun K, Viboud C, Xiong X, Yu H, Halloran M E, Longini I M, and Vespignani A The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. Science 2020; 368(6489):395–400. https://science.sciencemag.org/content/368/6489/395.

4. Edelsbrunner H, Letscher D, and Zomorodian A Topological persistence and simplification. Discrete and Computational Geometry 2002; (28):511–533.

5. Ester M, Kriegel H.-P, Sander J, and Xu X A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96) 1996:226–231.

6. Eubank S, Guclu H, Kumar V A, Marathe M V, Srinivasan A, Toroczkai Z, and Wang N Modelling disease outbreaks in realistic urban social networks. Nature 2004; 429(6988):180–184.

7. Ferguson N, Laydon D, Nedjati Gilani G, Imai N, Ainslie K, Baguelin M, Bhatia S, Boonyasiri A, Cucunuba Perez Z, Cuomo-Dannenburg G, et al Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. Imperial College Technical Reports 2020.

8. Guzmán-Sáenz A, Haiminen N, Basu S, and Parida L Signal enrichment with strain-level resolution in metagenomes using topological data analysis. BMC genomics 2019; 20(2):194.

9. Kalyanaraman A, Kamruzzaman M, and Krishnamoorthy B Interesting Paths in the Mapper Complex. Journal of Computational Geometry 2019; 10(1):500–531. https://arxiv.org/abs/1712.10197.

10. Kamruzzaman M, Kalyanaraman A, and Krishnamoorthy B Detecting divergent subpopulations in phenomics data using interesting flares. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB'18) 2018; 155–164.

11. Kamruzzaman M, Kalyanaraman A, and Krishnamoorthy B HYPPO-X: A software library for visual analytics on complex high dimensional data. `https://xperthut.github.io/HYPPO-X` 2019.

12. Kamruzzaman M, Kalyanaraman A, Krishnamoorthy B, Hey S, and Schnable P S Hyppo-X: Toward a scalable exploratory framework for complex high-dimensional phenomics data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2019. https://arxiv.org/abs/1707.04362.

13. Kybartas B and Bidarra R A survey on story generation techniques for authoring computational narratives. IEEE Transactions on Computational Intelligence and AI in Games 2017; 9(3):239–253.

14. Lockwood S and Krishnamoorthy B Topological features in cancer gene expression data. In Proceedings of the Pacific Symposium on Biocomputing (PSB); 2015; (20):108–119. https://arxiv.org/abs/1410.3198.

15. Lum P Y, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J G, and Carlsson G Extracting insights from the shape of complex data using topology. Scientific Reports 2013; 3(1236):1.

16. Madhobi K F, Kamruzzaman M, Kalyanaraman A, Lofgren E, Moehring, R and Krishnamoorthy B A visual analytics framework for analysis of patient trajectories. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics 2019; 15–24.

17. Marathe M and Vullikanti A Computational epidemiology. Communications of the ACM 2013; 56(7):88–96.

18. Reza T, Ripeanu M, Tripoul N, Sanders G and Pearce R Prunejuice: Pruning trillion-edge graphs to a precise pattern-matching solution. In Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC'18) 2018.

19. Singh G, Memoli F, and Carlsson G Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In Proceedings of the Symposium on Point Based Graphics (SPG'07) 2007; 91–100.

20. Venkatramanan S, Chen J, Fadikar A, Gupta S, Higdon D, Lewis B, Marathe M, Mortveit H, and Vullikanti A Optimizing spatial allocation of seasonal influenza vaccine under temporal constraints. PLoS Computational Biology 2019.

21. Zhang L, Ghader S, Pack M, Xiong C, Darzi A, Yang M, Sun Q, Kabiri A, and Hu S An interactive COVID-19 mobility impact and social distancing analysis platform 05 2020.

22. Zhang Q, Sun K, Chinazzi M, Pastore y Piontti A, Dean N E, Rojas D P, Merler S, Mistry D, Poletti P, Rossi L, Bray M, Halloran M E, Longini I M, and Vespignani A Spread of zika virus in the Americas. PNAS 2017; 114(22):E4334–E4343.

23. University of Maryland, COVID-19 Impact Analysis Platform. `https://data.covid.umd.edu/` 2020.

24. See how your community is moving around differently due to COVID-19. `https://www.google.com/covid19/mobility/` 2020.

25. Coronavirus (Covid-19) Data in the United States. `https://github.com/nytimes/covid-19-data` 2020.

26. Coronavirus (Covid-19) Data in AWS S3 bucket as data lake. `https://go.aws/covid-19-data-lake` 2020.

27. University of Virginia, COVID-19 Surveillance Dashboard. `http://ncov.bii.virginia.edu/dashboard/` 2020.

28. MIDAS 2019 Novel Coronavirus Repository. https://github.com/midas-network/COVID-19 2020.

29. New and Improved COVID Symptom Survey Tracks Testing and Mask-Wearing. https://delphi.cmu.edu/blog/2020/10/12/new-and-improved-covid-symptom-survey-tracks-testing-and-mask-wearing/ 2020.

30. Household Pulse Survey Public Use File (PUF). https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html 2020.

31. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/past-reports/09042020.html 2020.

Figures

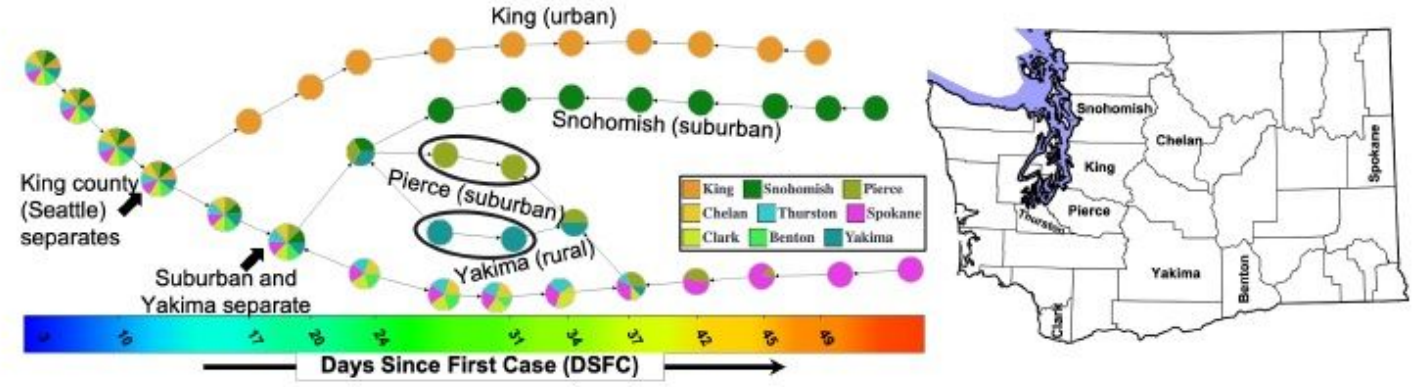
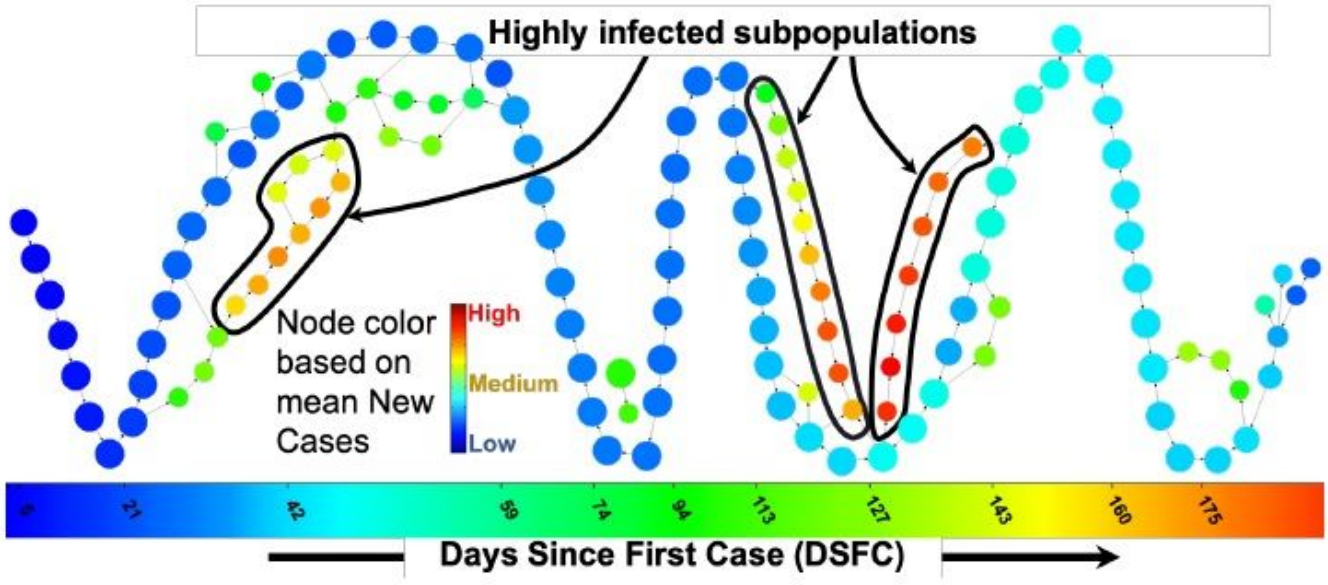


Figure 1

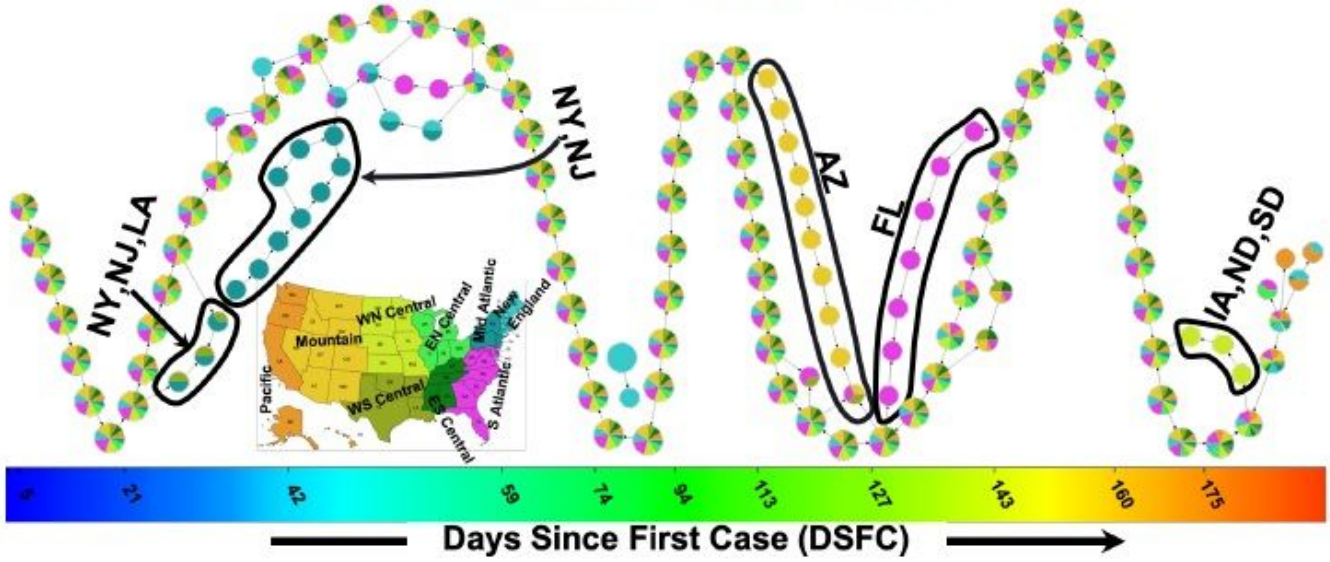
A topological object created by our approach for nine counties ofWashington state (right), during their rst 50 days of COVID-19. The data consists of time-series of DSFC for each county. Nodes represent clusters of counties (shown as pie-charts) that show similar growth rate in COVID-19 cases at different time intervals (horizontal bar). Edges connect clusters that intersect in points, and edge direction points to the direction of increase in the case rate. Multiple subpaths corresponding to epochs of co-evolving counties and branching events are visible.

Figure 2

U.S. map by: a) regions and b) divisions.



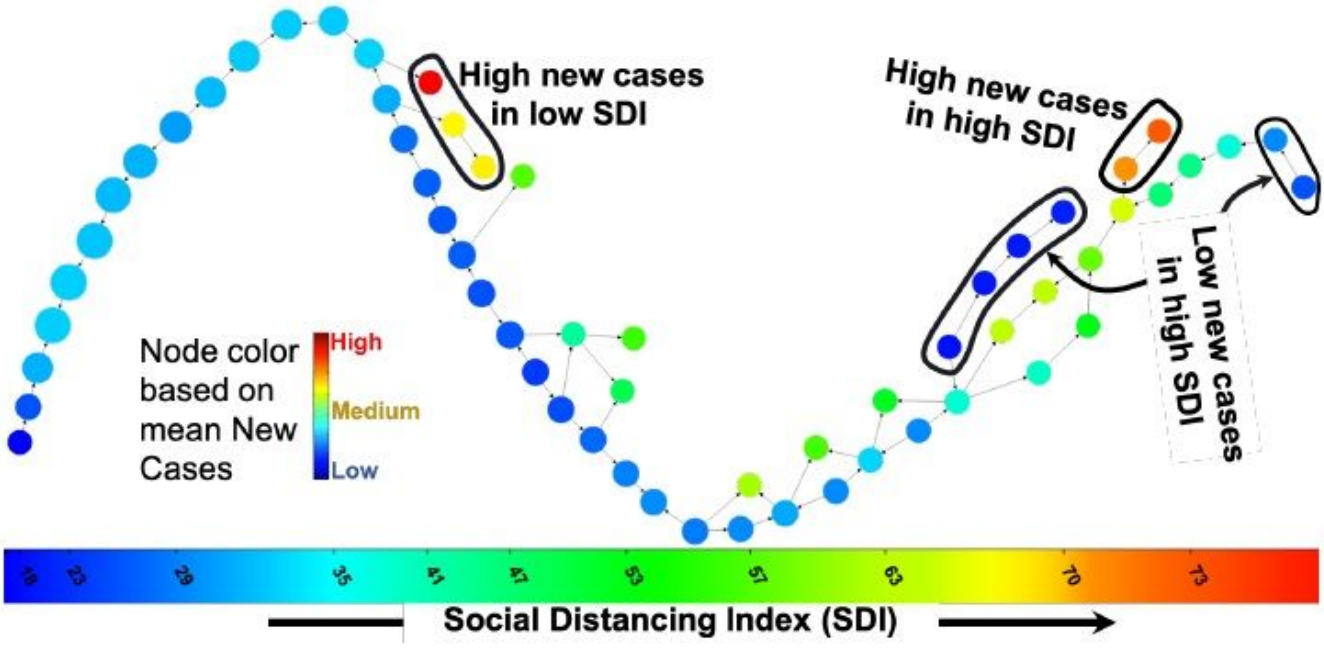
(a) Nodes (clusters) colored by new cases



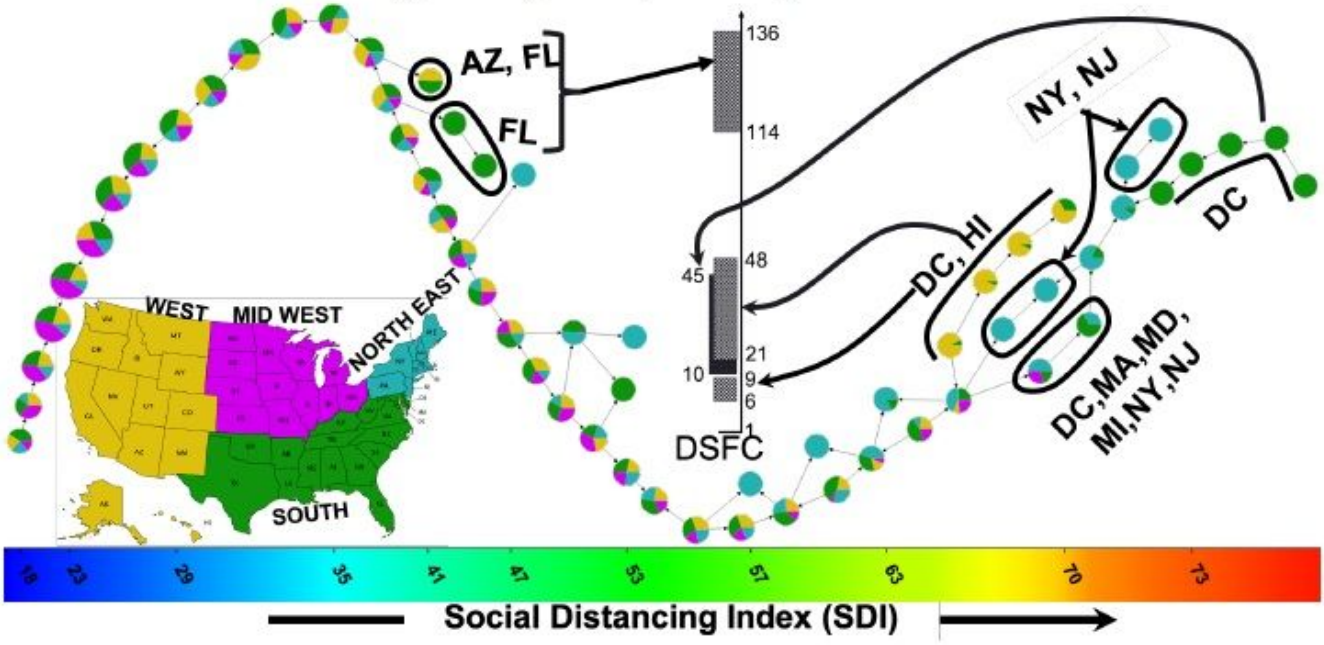
(b) Nodes (clusters) shown as a pie-chart of U.S. divisions

Figure 3

Topological object with DSFC as lter and new cases as clustering attribute. Hyppo-X parameters: # bins: 80; overlap: 35%; and DBSCAN (density: 2, radius: 0:032) for clustering. The horizontal color bar shows the DSFC scale. There is no Y-axis|the object is simply placed to match various intervals of the time scale. Each node is a cluster of points ([state,DSFC]) that had similar number of cases in the same time interval, and each edge (directed) connects two intersecting clusters.



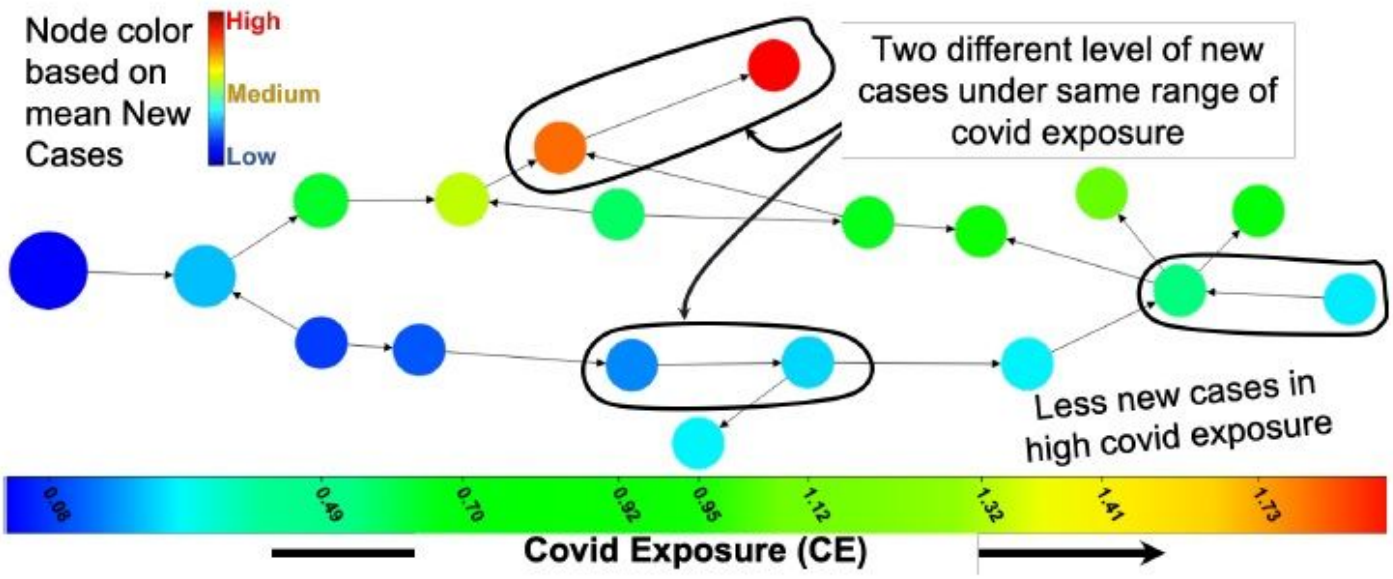
(a) Nodes (clusters) colored by new cases



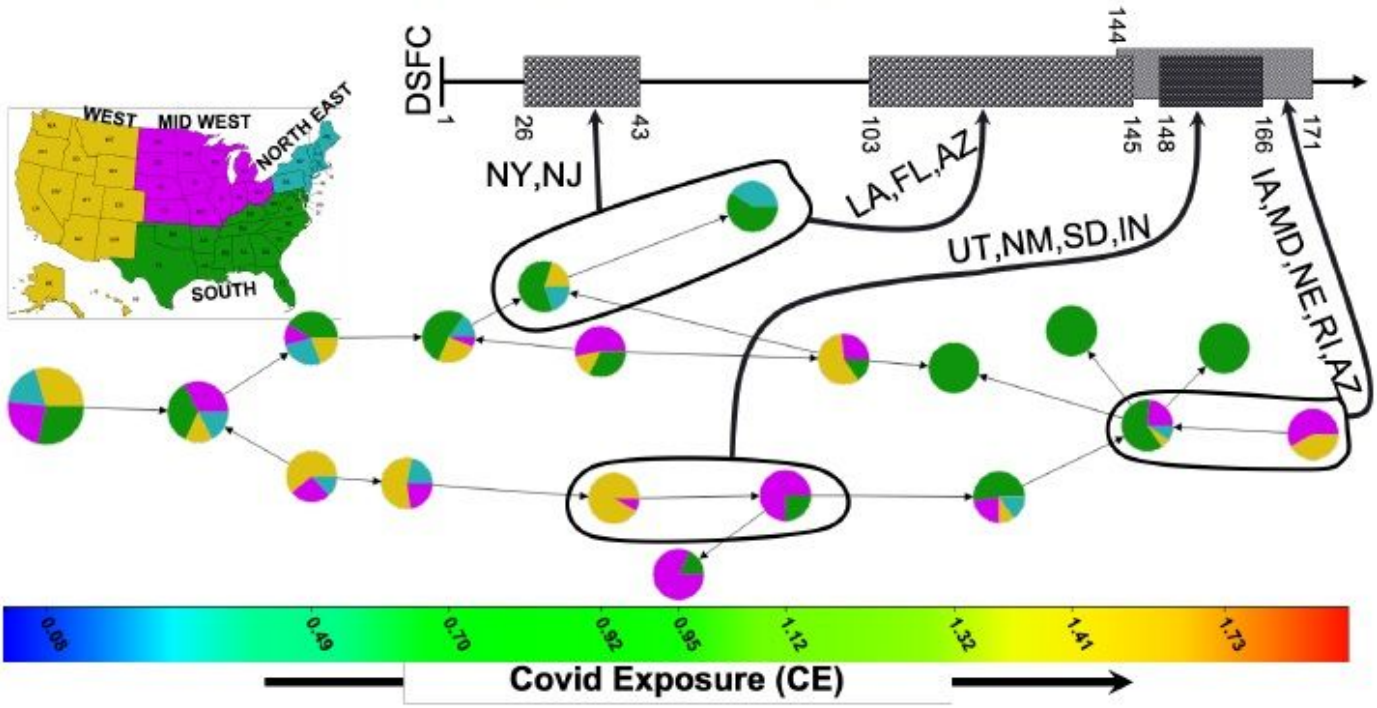
(b) Nodes (clusters) shown as a pie-chart of U.S. regions

Figure 4

Topological object with Social Distancing Index (SDI) as lter and new cases for clustering. Hyppo-X parameters: # bins: 40; overlap: 35%; and DBSCAN (density: 2, radius: 0:035) for clustering. Horizontal color bar shows the SDI scale, and each node is a cluster of [state,SDI] observations that can come from different points in time.



(a) Nodes (clusters) colored by new cases



(b) Nodes (clusters) shown as a pie-chart of U.S. regions

Figure 5

Topological object with COVID Exposure (CE) as lter and new cases for clus- tering. Hyppo-X parameters: # bins: 10; overlap: 10%; and DBSCAN (density: 2, radius: 0:01) for clustering. Horizontal color bar shows the CE scale, and each node is a cluster of points [state,CE]. Part (b) shows each node as a DOY (day of year) bar corresponding to active epochs.

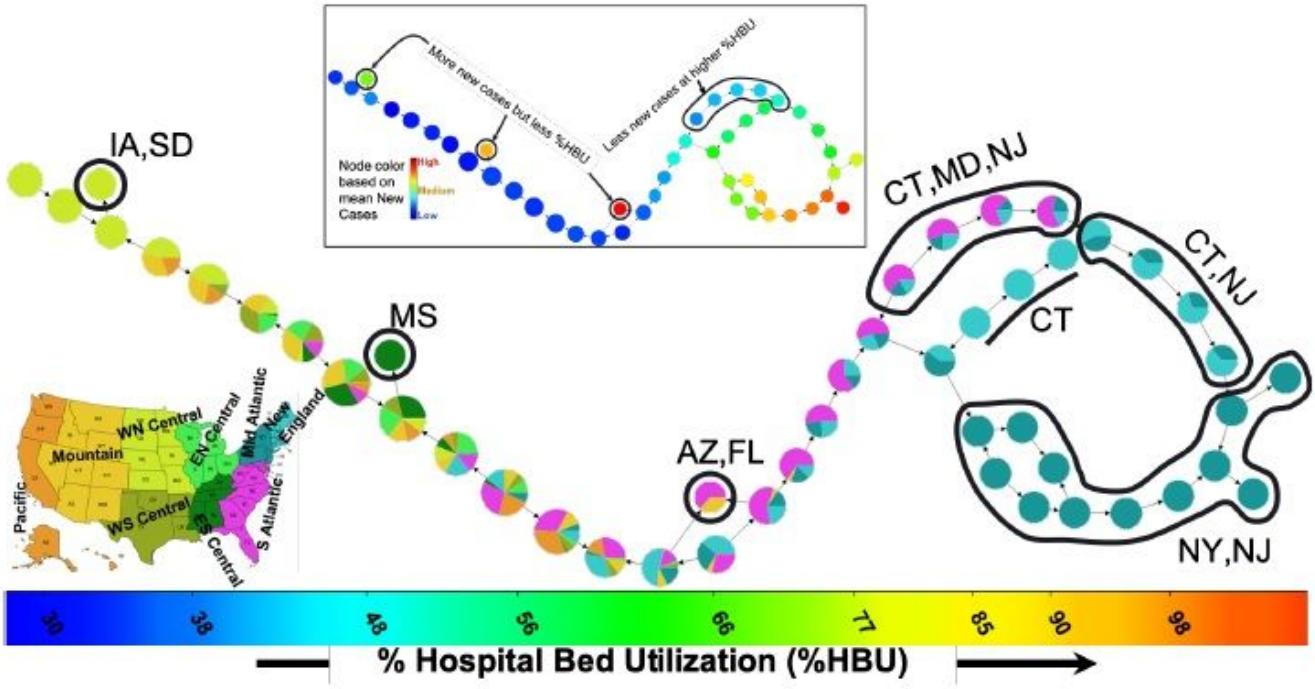


Figure 6

Topological object with %Hospital Bed Utilization (%HBU) as lter and new cases for clustering. Hyppo-X parameters: # bins: 30; overlap: 45%; DBSCAN (density: 2, radius: 0:03).