Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Discovering spatial-temporal patterns via complex networks in investigating COVID-19 pandemic in the United States

Yue Pan a, Limao Zhang b,*, Juliette Unwin c, Miroslaw J. Skibniewski d,e,f

a Shanghai Key Laboratory for Digital Maintenance of Buildings and Infrastructure, Department of Civil Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, 200240, China
b School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore
c MRC Centre for Global Infectious Disease Analysis, United Kingdom
d Department of Civil and Environmental Engineering, University of Maryland, College Park, MD 20742-3021, USA
e Chaoyang University of Technology, 413310 Taichung, Taiwan
f Polish Academy of Sciences Institute of Theoretical and Applied Informatics, 44-100 Gliwice, Poland

A novel approach combining time series analysis and complex network theory is proposed to deeply explore characteristics of the COVID-19 pandemic in some parts of the United States (US). It merges as a new way to provide a systematic view and complementary information of COVID-19 progression in the US, enabling evidence-based responses towards pandemic intervention and prevention. To begin with, the Principal Component Analysis (PCA) varimax is adopted to fuse observed time-series data about the pandemic evolution in each state across the US. Then, relationships between the pandemic progress of two individual states are measured by different synchrony metrics, which can then be mapped into networks under unique topological characteristics. Lastly, the hidden knowledge in the established networks can be revealed from different perspectives by network structure measurement, community detection, and online random forest, which helps to inform data-driven decisions for battling the pandemic. It has been found that states gathered in the same community by diffusion entropy reducer (DER) are prone to be geographically close and share a similar pattern and tendency of COVID-19 evolution. Social factors regarding the political party, Gross Domestic Product (GDP), and population density are possible to be significantly associated with the two detected communities within a constructed network. Moreover, the cluster-specific predictor based on online random forest and sliding window is proven useful in dynamically capturing and predicting the epidemiological trends for each community, which can reach the highest.

1. Introduction

The novel and infectious coronavirus disease called COVID-19 was firstly detected in Wuhan, China in late December 2019, and then continued to expand to more than 210 countries and territories across the world. From the records at the Johns Hopkins University tracker (Jhuomcr, 2020), COVID-19 has resulted in a global total of more than 217 million confirmed cases and 4.51 million deaths at the end of August 2021. Due to great transmission ability from person to person and high spread speed, it is known that COVID-19 has developed to be a global pandemic, causing enormous challenges to human society in terms of public health, economic recession, political disruption, and others (Lak et al., 2021; Maiti et al., 2021). In this regard, a great concern is to fully understand the patterns of COVID-19 evolution dynamics in certain regions, which can hold promise to offer control and actions to combat the pandemic. Such a data-driven smart investigation paves a new way to mitigate the adverse impact from COVID-19 spreading, which is believed to be well fitted in sustainable societies (Rahman et al., 2020; Shorfuzzaman et al., 2021).

So far, the United States (US) has become the country with the highest number of reported infections. To be specific, the US reported the first case on Jan 20, 2020, and the epidemic has spread rapidly throughout all 50 states by March 17, 2020 (IHME COVID-19 Forecasting Team, 2020; Unwin et al., 2020). It has been found that thousands of daily new cases and deaths are reported at the national level, and the risk of infection varies from state to state. Moreover, a recent

---

* Corresponding author.
E-mail addresses: panyue001@sjtu.edu.cn (Y. Pan), limao.zhang@ntu.edu.sg (L. Zhang), h.unwin@imperial.ac.uk (J. Unwin).
model estimates that three times more Americans will be potentially contagious than the actual reported number of infections due to the limited testing capability and restrictive case definition (Noh & Danuser, 2021). The failure of the authorities and leaders to take adequate steps and preventions in time is considered to be one of the leading causes that put the country in high threat of COVID-19, which unfortunately trigger an economic recession and cause grave losses of lives (Woolf et al., 2021). In other words, the US is hit extremely hard by the pandemic, which therefore deserves more of our attention.

Due to the advantages of artificial intelligence (AI), such as automation, efficiency, digitalization, flexibility, and others (Pan & Zhang, 2021a), AI can provide novel solutions in fighting against the disease, which could even be a key factor to the success of the health care system (Chimmula & Zhang, 2020; Pan & Zhang, 2021b). In this regard, many researchers have attempted to treat various AI algorithms as a real-time forecasting tool, which can well fit the given data and then estimate the future trend of COVID-19 at the national scale or world level (Chew et al., 2021a). Accordingly, some preventive measures can be generated at early stages to slow down the spread of the virus. For example, Chew et al. (2021a) developed a hybrid deep learning algorithm under the deep neural network and natural language processing feature extraction method to forecast the growth of COVID-19 cases globally using Twitter data. Pan et al. (2021) built a random forest model with the rolling time-window to learn data about environmental and social factors for predicting the confirmed cases and death cases in four Asian countries, including Japan, South Korea, Pakistan, and Nepal. Andelić et al. (2021) estimated the epidemiology curve for the COVID-19 pandemic in the US by the genetic programming algorithm. Chimmula and Zhang (2020) implemented the long-shot-term memory (LSTM) to represent the spatio-temporal components and predict the transmission rate of COVID-19 in Canada. Although the growth rate of confirmed cases and deaths could be abstractly described and accurately estimated, a clear limitation is that the complex relationships of COVID-19 development between pairs of states are still unexplored. Since the virus spread is likely to demonstrate geographic trends and continue developing, some spatial-temporal patterns are bound to embed in the publicly available datasets about ongoing situations of all 50 states. Herein, the spatial-temporal characteristics of confirmed cases and deaths provide knowledge about the distribution and evolution of COVID-19 in each state of the US, which deserved in-depth investigation. Therefore, its scientific contribution is to develop a novel network analysis framework to take full advantage of the US-related data collected by a COVID-19 tracking project for developing sustainable societies, which can not only reveal the virus transmission characteristics regarding spatial distribution, epidemiological conditions, and social factors, but also make an accurate prediction about the dynamic development trend of COVID-19. That is to say, the unique innovation is to synthetically implement techniques of information fusion, network exploration, community detection, knowledge discovery, machine learning prediction, which can eventually provide a systematic viewpoint to examine dynamic behaviors of COVID-19 in states of the US, explore their inherent connections, and predict the pandemic spread. Through information fusion techniques, observational multiple time series about the accumulated number of positive cases, deaths, and total test results in states of the US can be deeply investigated. The advantage of such data analysis lies in the effective extraction of temporal-spatial patterns of COVID-19 progression across the US and high prediction accuracy about future pandemic evolution, which can potentially, support data-driven suggestions towards pandemic intervention and containment. In this regard, there are three research questions to be addressed: (1) How to effectively quantify synchrony between time series data describing each state, which is of great importance to providing implicit information to connect states within the network logically; (2) How to deeply explore the established networks from views of the topological analysis, community detection, and hidden knowledge discovery to gain actionable insights into the pandemic progress, which helps to enhance the understanding of COVID-19 behavior across different states; (3) How to deploy the idea of online learning and sliding time window to make convincing predictions about the growing trend of positive cases and deaths in each detected community, which can serve as an adaptable decision-making tool to put forward countermeasures in advance.

The remaining part of the paper is organized as follows. In Section 2, critical methods involving in the developed complex network-enabled approach are introduced in detail. In Section 3, the proposed approach has been applied in the COVID-19 evolution data about the US to verify its effectiveness. In Section 4, the great performance of the online random forest in predicting the spread of COVID-19 in the future is demonstrated. In Section 5, the main conclusions are summed up.

2. Data and methods

This paper intends to fuse time series data about the daily positive cases, deaths, and total test results in states of the US together to capture underlying mechanisms of complex systems, another superiority of such a hybrid method lies in its ability to tackle correlation between time series data and forecast future development, since complementary features from time-series data can be mapped into the topological structure (Zhou et al., 2018). Since COVID-19 records are actually time-series data, they can be reasonably transformed into the complex network as a visibility graph (Demetzis et al., 2020). Herein, the network representation can be used to describe the trend of the epidemic in the US based on notions of nodes and edges, where nodes stand for states and edges reflect the relations between states. The significant benefit of complex network analysis is to shed light on the inner workings of the pandemic transmission between states, resulting in the important patterns, properties, and dynamics of COVID-19 evolution across states of the US. To sum up, the motivation of this paper is to comprehensively uncover hidden knowledge underlying information about the pandemic evolution towards the sustainable development goal using various intelligent methods, including time-series analysis, complex network theory, and online random forests. The difference between our work and other existing studies is that this paper goes beyond the simple task of tracking and forecasting the COVID-19 pandemic, which can further bridge the gap between intelligent computing and COVID-19 evolution investigating. Therefore, its scientific contribution is to develop a novel network analysis framework to take full advantage of the US-related data collected by a COVID-19 tracking project for developing sustainable societies, which can not only reveal the virus transmission characteristics regarding spatial distribution, epidemiological conditions, and social factors, but also make an accurate prediction about the dynamic development trend of COVID-19. That is to say, the unique innovation is to synthetically implement techniques of information fusion, network exploration, community detection, knowledge discovery, machine learning prediction, which can eventually provide a systematic viewpoint to examine dynamic behaviors of COVID-19 in states of the US, explore their inherent connections, and predict the pandemic spread. Through information fusion techniques, observational multiple time series about the accumulated number of positive cases, deaths, and total test results in states of the US can be deeply investigated. The advantage of such data analysis lies in the effective extraction of temporal-spatial patterns of COVID-19 progression across the US and high prediction accuracy about future pandemic evolution, which can potentially, support data-driven suggestions towards pandemic intervention and containment. In this regard, there are three research questions to be addressed: (1) How to effectively quantify synchrony between time series data describing each state, which is of great importance to providing implicit information to connect states within the network logically; (2) How to deeply explore the established networks from views of the topological analysis, community detection, and hidden knowledge discovery to gain actionable insights into the pandemic progress, which helps to enhance the understanding of COVID-19 behavior across different states; (3) How to deploy the idea of online learning and sliding time window to make convincing predictions about the growing trend of positive cases and deaths in each detected community, which can serve as an adaptable decision-making tool to put forward countermeasures in advance.
the spatial-temporal pattern. The whole process has been summarized in Fig. 1 with two critical parts, namely data analysis preparation and data analysis implementation. In brief, this work begins with the creation of network maps using multiple time series associated with the epidemic progression that has been preprocessed into one sequence by Principal Component Analysis (PCA) varimax. Three different networks based on the mechanism of Pearson correlation, TLCC, and DTW are generated, where nodes represent the individual state, and the topology is constructed by quantifying synchrony between time series data. Therefore, these networks are feasible to graphically describe the quantifiable relationships and similarity of virus dynamics between states, which deserve a range of complex network analyzes, including topological property measurement at the network level and node level, community detection by DER algorithm, knowledge discovery related to evolution process and social factors, and cluster-specific prediction using the online random forest. All these methods are implemented by Python 3.7. The practical value lies in a better understanding of the dynamic behavior of COVID-19 from a new point, which can help policymakers to inform more reliable strategies for effectively limiting or even blocking the disease spread. Some important methods have been introduced below.

2.1 Data analysis preparation

2.1.1 Data collection

The data repository used in this study is derived from a website of a COVID Tracking Project (https://covidtracking.com/data/download). Daily data on the COVID-19 pandemic for the US and individual states are updated every day on this website. Three kinds of information are extracted to describe the dynamic behaviors of COVID-19 state-wise, including the cumulative count of positive cases, deaths, and test results. As a more detailed data definition, positive cases include the confirmed plus probable COVID-19 case diagnosis. Deaths represent the fatalities from all positive cases. Total test results report the number of the viral Polymerase Chain Reaction (PCR) testing by state. Since the growth rate, referring to the percentage of a variable over a period of time, is beneficial in evaluating the changes in circumstances and making a prediction about the future performance, it is necessary to convert the three obtained time series into the form of growth rate by Eq. (1) (Chew et al., 2021a).

\[ G_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \times \frac{1}{p} \]  

(1)

where \( G_t \) is the daily growth rate of positive cases or deaths or test results, \( Y_t \) is the cumulative number of positive cases or deaths or test results at the current day \( t \), \( Y_{t-1} \) is the cumulative number of positive cases or deaths or test results at the previous day \( t-1 \), and \( p \) donates the size of the population in each state. The population of the investigated US states is downloaded from the website https://worldpopulationreview.com/states. For calculating the growth rate, the difference between two consecutive days is divided by the value of the beginning value, and then the result is normalized by the size of the population in this state. In the end, there are three streams of data on behalf of the growth rate for a state, which basically manifest how much the COVID-19 condition of a state changes over time. Each time series contains 343 data points during a certain period of Mar 7, 2020, and Feb 12, 2021. Therefore, for all 50 states in the US, a total of 343 x 50 sets of observations are

---

Fig. 1. Methodology framework of the proposed network-enabled analysis.
aggregated in the prepared dataset for further exploration.

2.1.2. PCA varimax calculation

Since the growth of positive cases, deaths, and test results in an individual state is dependent, the COVID-19 development can be considered to be on-orthogonal. Therefore, the principal component analysis (PCA) (Kaiser, 1958; Richman, 1986) is firstly adopted to deal with the observations about daily data on the COVID-19 pandemic for the US. To be more specific, PCA is a statistical procedure for extracting features and preserving variability as much as possible, resulting in low-dimensional representations of a large dataset. An important application of PCA is data fusion (Shahdoosti and Ghassemian, 2016). To improve the interpretability of the resulting PCA components, a rotation algorithm called the Varimax criterion is deployed, aiming to emphasize a small number of important variables through maximizing the value of $V$ defined in Eq. (2) (Hannachi et al., 2007). It benefits in maximizing both the sum of variances of the loadings within the factors and the differences in low and high loadings on a particular factor, allowing for generating more localized spatial patterns related to components and correlating the time series of weights (Nowack et al., 2020). As an explanation of loadings, it represents correlations between variables and factors. For example, the varimax-rotated PCA is particularly useful to understand the large dataset and complex system in atmosphere science, which can catch the regional characteristic of atmospheric teleconnections to make it closer to the physical condition.

$$V = \frac{1}{p} \sum_{j=1}^{m} \left( \left( \sum_{i=1}^{n} x_{ij}^2 \right)^{\frac{3}{2}} - \frac{1}{p} \left( \sum_{i=1}^{n} y_{ij} \right)^2 \right)$$

(2)

where $\bar{y}_j$ is the scaled loading, the size of a square rotation matrix is $m \times m$, and the matrix size of loadings is $p \times m$.

2.1.3. Complex network establishment

After PCA varimax, multiple time series collected by the COVID-19 Tracking Project are transformed into a symbol sequence to well reserve the main features of the original data, which can serve as the rich resource for topological mapping. Therefore, the next question that arises is how to properly bridge the time series and network diagrams. Herein, a data-driven complex network is expected to be built to graphically examine the intrinsic interaction of pandemic evolution within states of the US, where individual states and their dynamic relations of long-period time series are mapped into nodes and edges. Notably, a key step in network construction is to calculate mutual dependency between pairs of time series describing two states. As the determination of node connectivity, the following three approaches are utilized to measure synchrony between time series from a different point of view.

(1) Pearson correlation coefficient can be easily estimated by Eq. (3) (Benesty et al., 2009). Its absolute value closer to 1 means the two states are more correlated with each other. An interesting thing to note is that it may take several days until the increasing cases and mortality of a certain state affect another one, indicating the COVID-19 influence will not occur concurrently.

$$\rho_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

(3)

Dynamic time warping (DTW) is another flexible way to handle the time series that are not in sync, and thus two time-dependent sequences can be wrapped in a nonlinear fashion to determine the optimal match (Müller, 2007). Instead of correlation, the Euclidean distance is calculated to represent the point-to-point relationship between two given time series $X$ and $Y$. The smallest distance is stored as the cost for aligning the point $x_i$ to $y_j$, which can be utilized to form a time warping path standing for the point-to-point alignment. The desired warping path is identified by the minimal total costs of aligning its points, whose corresponding DTW distance is defined as Eq. (4) (Li, Chen, & Wu, 2010).

$$\text{DTW}(X,Y) = \min_{W} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_i - y_j)^2$$

(4)

where $X$ and $Y$ are two time series in the length of $m$ and $n$, $E(X)$ and $\sigma(X)$ are the average value and standard deviation of the sequence $X$, and $E(Y)$ and $\sigma(Y)$ are the average value and standard deviation of the sequence $Y$, respectively.

2.2. Data analysis implementation

2.2.1. Network structure measurement

When the complex network has been rationally developed, it is necessary to gain an overview of its topological characteristics (Hong et al., 2019). For this purpose, several measures are employed to qualitatively assess networks at the network or node level. More specifically, under the goal of studying the network as a whole, four network-level metrics are numerically exploring the hidden relationship in times series by defining in Eq. (2) (Hannachi et al., 2007). The average value and standard deviation of the sequence $X$ and $\sigma(Y)$, respectively.

- **Network diameter**: The maximum number of possible edges.

- **Network density**: The longest value of all the calculated shortest paths between two nodes.

- **Average clustering coefficient**: The average probability that two neighbors of a node are linked to each other.

- **Average path length**: The average shortest path between two nodes, which is averaged over the total number of pairs.
indicators of interest are used, including the network diameter, network density, average clustering coefficient, and average path length, as outlined in Table 1. Besides, the adopted node-level metrics chiefly consider four types of centrality in Table 2, including the degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. These four metrics aim to measure the importance of nodes based on the topological structure. For a clearer explanation, the higher the value of centrality, the more important the node is. The recognized important node will take a more central position and exert more influence within the established network, which can be regarded as the leader to catch more attention.

Note for Table 1: m and n are the total numbers of edges and nodes in the graph, respectively, $d_{ij}$ is the shortest distance between the node i and j, $k_i$ is the degree of the node $v_i$, and $e_i$ is the number of edges between neighbors of $v_i$.

Note for Table 2: n is the total number of nodes, $d_i$ is the shortest distance between the node i and j, $\sigma_0(v)$ is the total number of shortest paths from node i to node j, $\sigma_1(v)$ is the number of these paths through node v, $A$ is the adjacent matrix, and $x_j$ is the eigenvector centrality for node j.

2.2.2. Diffusion entropy reducer (DER)

Since the complex network is not modeled randomly, it can be assumed that there is a great level of the organization embedded in it. To further reveal the intrinsic structure of a network, a pivotal tool called community detection (also known as node clustering) is of great worth. It relies on information encoded in the network topology to partition the entire network into multiple communities (also known as clusters) (Pan et al., 2020). Generally speaking, nodes within a community tend to be more densely connected with stronger relationships or more common properties than external ones, while the connection among inter-communities is relatively weak. Among a variety of developed community detection methods, a novel algorithm named the diffusion entropy reducer (DER) proposed in 2015 (Kozdoba & Mannor, 2015) is adopted due to its promising performance in detecting meaningful node clusters on several standard graph benchmarks. The core idea of DER is to integrate the concept of short random walks into the k-means clustering, which is briefly presented below.

To begin with, a finite undirected network $G = (V, E)$ represented by the adjacency matrix $A = (a_{ij})$ will be divided into subsets in the equal size of $n/k$, where $a_{ij} \geq 0$ indicates the weight of an edge, $n=|V|$ is the number of total nodes, and $k$ is the number of communities. Then, a dynamic procedure called random walk is performed to extract valuable information from even sparse networks based on a transition matrix $T = D^{-1}A$, where $D_{ii} = d_i = \sum_j a_{ij}$ is an $n \times n$ diagonal matrix, and $d_i$ is the degree of a node. In other words, an L-length random walk starting from a node can be created under the station measure $x(i) = \frac{d_i}{\sum_i d_i}$, and then such sampling process will repeat N times to generate sequences $\{x^1, x^2, \ldots, x^N\}$. For a sequence $x$ beginning at the node i after t steps, its distribution can be denoted by $w_i^t$ as the ith row of the matrix $T^t$. Also, the distribution about the average empirical measures of the sequence $x$ can be set as Eq. (5), while the distribution of the random walk from $x_i$ ($\pi_t$ means the restriction of the measure $\pi$ to the subset $CCV$) can be expressed by Eq. (6). Afterward, the k-means-like mechanism is conducted on results of random walk to automatically detect disjoint and separable subsets under the rule of making $P_k = \{i \in V|c = \text{argmax}_{l} \mu_l\}$ unchanged and $\mu_l = \mu_{P_k}$, where $D(w_i, \mu_j)$ is a distance function calculated by Eq. (7). To quantitatively measure the performance of the DER algorithm, eight metrics from Table 2

| Metrics                  | Equation                              | Description                                                                 |
|--------------------------|---------------------------------------|------------------------------------------------------------------------------|
| Degree centrality         | $C_d(n_i) = \frac{\sum_{j=1}^{n_i} a_{ij}}{n_i}$ | To count the total amount of edges a node owns                                |
| Closeness centrality      | $C_c(n_i) = \frac{n-1}{\sum_{j=1}^{n_i} d_j}$ | To calculate the reciprocal of the sum of the shortest distance between an individual node to others |
| Betweenness centrality    | $C_b(n_i) = \frac{1}{n-1} \sum_{j=1}^{n_i} \frac{a_{ij}[v]}{\sigma_1(v)}$ | To estimate the times a node acting as a bridge in the shortest path between two other nodes |
| Eigenvector centrality    | $C_e(n_i) = \frac{1}{2} \sum_{j=1}^{n_i} A ij x_j$ | To measure the degree of a node along with the degree of its neighbors        |

Table 3

| Metric                              | Equation                              | Description                                                                 |
|-------------------------------------|---------------------------------------|------------------------------------------------------------------------------|
| Newman-Girvan modularity            | $Q = \frac{1}{2m} \sum_{\pi} (A_{ij} - \frac{k_i k_j}{2m}) \delta(\pi_i, \pi_j)$ | The subtraction of the fraction of edges within communities from the expected fraction of such edges by a random network |
| x modularity                        | $Z = \frac{\sum_{C \in C} m_{CC} - \sum_{C \in C} \left( \frac{D_{CC}}{2m} \right)^2}{\sqrt{\sum_{C \in C} \left( \frac{D_{CC}}{2m} \right)} \left( 1 - \sum_{C \in C} \left( \frac{D_{CC}}{2m} \right)^2 \right)}$ | The solution of the resolution limit of the original modularity               |
| Average internal degree             | $\bar{p} = \frac{2m}{n-1}$            | The mean of an internal degree of communities                                |
| Scaled density                      | $\sigma = 1 - \frac{\sum a_{ii}}{\frac{\sum a_{ii}}{n}}$ | The multiplication of the original density by the size of a community         |
| Average transitivity                | $\epsilon = 1 - \frac{\sum a_{ii}}{\frac{\sum a_{ii}}{n}} \frac{3}{t}$ | The average clustering coefficient of nodes                                  |
| Average embeddedness of nodes in a community (Orman et al., 2011) | $f_{oam} = \frac{1}{n_1} \sum_{u \in E} |\{u \in C \cup \{u,v\} : (u,v) \notin C \}| > d_u$ | A proportion of the internal degree to the total degree of the node          |
| Fraction over median degree         | $f_{omd} = \frac{1}{n_1} \sum_{u \in E} |\{u \in C \cup \{u,v\} : (u,v) \notin C \}| > d_u$ | The fraction of the community nodes that have an internal degree              |
| Average out degree fraction         | $f_{odf} = \frac{1}{n_1} \sum_{u \in E} |\{u \in C \cup \{u,v\} : (u,v) \notin C \}| > d_u$ | Larger than the median value of the degree                                  |
| Average ratio of nodes in a community that points outside this community | $f_{oafo} = \frac{1}{n_1} \sum_{u \in E} |\{u \in C \cup \{u,v\} : (u,v) \notin C \}| > d_u$ | The average ratio of nodes in a community that points out side this community |
Table 3 are utilized.

\[ w_i = \frac{1}{L} \left( w_i^1 + w_i^2 + \ldots + w_i^L \right) \]  

(5)

\[ \mu_i = \frac{1}{\sum_{c \in C} \sum_{I \in \delta}(c \setminus i)} \sum_{I \in \delta}(c \setminus i) \]  

(6)

\[ D(w, \mu) = \sum_{i \in C} w_i \log \mu_i \]  

(7)

where the denominator of Eq. (6) is the full degree of the subset \( C \).

Note: \( m \) is the number of edges, \( k_i \) is the degree of a node \( n_i \), \( \delta(\cdot) \) is the Kronecker delta, \( C \) is the detected communities, \( n_i \) is the number of communities, and \( d_m \) is the median value of internal degree \( d(u) \).

2.2.3. Online random forests

To make the utmost of each discovered community within the complex network, an important task is to understand its nature of COVID-19 development and predict its future trend on a daily basis. For the purpose of intelligent prediction, a machine learning tool named random forests is applied, which are typically regarded as ensembles of decision trees that are combined by bagging. The random forest has gained great popularity mainly due to its merits of easy interpretation, high computational efficiency, and competitive prediction performance. It has also been successful in making the spatio-temporal estimation of COVID-19 daily cases worldwide (Yesilkaya, 2020). In general, the traditional random forests are implemented offline, which need to access the whole dataset before model training. However, a fact remains that data will arrive in sequence as the event goes, meaning that it is impossible to know the entire data beforehand in most cases. Besides, the off-line model is hard to capture and adapt to new conditions, since the change of data distribution over time is not taken into account. The divided training and testing sets are unable to correct the early decisions. That is to say, the off-line trained random forests could be inapplicable to explore time series data in some senses, possibly leading to a relatively great prediction bias (Zhong et al., 2020). To resolve this situation, an online learning mechanism can be considered to combine into the traditional random forests. The superiority of online learning is that the model can automatically adapt and update new information bases on the streaming data. Moreover, the developed online random forest is much easier to train than the popular LSTM that requires memory-bandwidth-bound computation and is prone to overfitting.

It is known that the records about the pandemic evolution in the US are complex, changeable, and non-linear in nature. To make this kind of time series suitable for regression supervised learning, a sliding window in a certain size of the lag is necessarily added into a univariate time series that is created by PCA varimax. That is to say, the historical data from the prior time steps can be accumulated in a sliding window, which is then used for forecasting the next time step. Herein, the consecutive establishment of random forest models can be realized by learning incrementally inputs from both the previous data and new observations associated with the COVID-19 dynamic behavior. The idea of online learning assists in improving the performance of random forests over time, and thus the online random forest is expected to exhibit an enhanced predictive ability and adaptivity in this research. Based upon the prepared dataset, the detailed execution of the developed online random forest aims to make predictions about successive 200 days from Jul 28, 2020 to Feb 12, 2021. To conceptualize this, the process is also depicted in Fig. 2.

In the beginning, only information about the first 143 days is available. When the width of the window is set as 14 days, the first data sample can be generated with features from Day 1–14 and a label from Day 15. Then, the sliding window moves right by a day to create the second data sample using features from Day 2–15 and a label from Day 16. In the same way, a total of 129 data samples can be produced, which can act as a training set for building up the 1st random forest model. Later, the fitted model is tested on the last 14 days of the existing data (Day 130–143) to predict the next day (Day 144). When data about Day 144 arrives, an additional data sample can be incorporated into the training set. Therefore, the 2nd random forest model is constructed through learning from an updated training set with 130 samples, which are utilized for predicting the condition on Day 145. That is to say, the training set will be constantly updated based on a continuous stream of data. Once the renewed training dataset is prepared, a new random forest model will be trained to fit its set of parameters. Such process of model construction will repeat until no new data enters, which herein can make the model better adapt to data variation changing in the spread and severity of the pandemic. In this example, the last step is to train the 200th random forest model by 328 samples covering data from Day 1 to Day 342, and then the model is tested on Day 329–342 for forecasting Day 343. Finally, the predictive performance of the proposed model is evaluated by three common metrics defined in Eqs. (8)–(10), including the correlation coefficient \( R^2 \), root mean squared error (RMSE), and mean absolute error (MAE).

\[ R^2 = 1 - \frac{\sum_{i=1}^N (y_{i,a} - y_{a,i})^2}{\sum_{i=1}^N (y_{i,a} - \bar{y}_{a})^2} \]  

(8)
\[ RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_{p,i} - y_{a,i})^2} \]  
\[ MAE = \frac{1}{N} \sum_{i=1}^{N} |y_{p,i} - y_{a,i}| \]  

where \( N \) donates the number of predicted data, \( y_{p,i} \) and \( y_{a,i} \) are the predictive value and actual value, respectively, and \( \frac{1}{N} \sum_{i=1}^{N} \) represents the mean of actual value.

3. Complex network development

The goal of this research is to study the characteristics of COVID-19 evolution in each state of the US and reveal their potential relationships using graph theory and time series analysis. It is believed that the complex network theory contributes to a supplemental understanding of the pandemic across the US. Before data mining, it is necessary to prepare a dataset containing time series in terms of three features during March 7, 2020–February 12, 2021, including the daily growth rate of COVID-19 positive cases, deaths, and test result in 50 individual states. This dataset can well capture the dynamic behavior of COVID-19 in every state. Fig. 3 presents a framework incorporating three key steps to generate network maps. It begins with preprocessing features by PCA varimax, which is responsible for projecting three features into a one-dimensional subspace. The PCA varimax acts as an effective tool to reduce the data dimensionality for simplifying the complex problem. Herein, two principle components are retained and rotated. Then, the synchrony between time series representing different individual states is measured, which provides rich evidence to reasonably build the complex network to describe the relationships of pandemic evolution among states. To make it clear, step 1 in Fig. 3 takes the state California (CA) as an example. Via PCA varimax, the three time series about CA in Fig. 3(a) can be converted into a univariate one in Fig. 3(b). This newly-generated time series can retain most of the information in the original data whilst dropping the less informative eigepairs, and thus the main pattern about the pandemic progress in CA can be extracted and drive the data analysis to be faster and easier. That is to say, PCA varimax helps to reduce the dimension of multivariate data and extract their main modes of variability about the COVID-19 progression for further exploration.

Data about the remaining 49 states was processed in the same way. As a result, a total of 50 time series are generated in the first step to represent the original data space corresponding to pandemic evolution in 50 states of the US. Since these time series are valuable information for inferring and assessing the change of pandemic situation, a meaningful task is to feed them into the second step for measuring synchrony between time series.

It is worth noting that the spatial-temporal correlation is a significant characteristic of COVID-19 development in different states of the US, indicating that time and location are bound to affect the severity and evolution of the pandemic. To capture such hidden knowledge, step 2 calculates three measures of synchrony named Pearson, time-lagged cross correlation (TLCC), and dynamic time warping (DTW), which are effective in discovering quantifiable relationships from time series of states. To illustrate, two states CA and AL (Alabama) are taken as an example, as shown in Fig. 3(c,d). More specifically, Fig. 3(c) returns the result of the Pearson and TLCC methods, and Fig. 3(d) provides the result of DTW approaches. The value of the Pearson correlation between CA and AL is 0.752, which can be considered that conditions in CA and AL are highly correlated. The peak correlation from TLCC is 0.841 at the time lag of three days, revealing a possible causality that the effect from AL will take three days to propagate to CA. Through dynamically comparing two sequences from CA and AL that vary in time, DTW is calculated to be 110.48, aiming to determine whether the phase in one state matches another even if they are not in sync. The computing process for Pearson, TLCC, and DTW will repeat in all pairs of states, which can ultimately generate three similarity matrices in a size of 50 × 50. Since it is unnecessary to compare each state to itself, and thus value in the diagonal line of the matrix is 0. Due to the symmetry, the part below the main diagonal is visualized. It is observed that the deeper blue in Fig. 3(e,f) and the lighter blue in Fig. 3(g) means the higher similarity or stronger correlation between the pandemic evolution of each pair of states.

The third step is for topological mapping. Since all states in the US can be regarded as a whole system, it is reasonable to take their inherent interactions into account for building a complex network, resulting in a systematic evaluation of the overall condition of COVID-19 dynamics in the US. Based on the resulted similarity matrix, a fully connected network can be constructed owning 50 nodes and 2500 links. This kind of original network yields a complex topology structure, which is bound to embed some redundant information to raise the difficulty in network investigation. Meanwhile, there is no need to explore the original network, since of particular interest in this research is the states with the top strongest relationships. For this concern, 500 links with the largest value of Pearson or TLCC and the lowest value of DTW are preserved herein, accounting for 20% of all possible links within a network. The link removal strategy can make the network less dense, thereby simplifying the problem and decreasing the computational complexity. Accordingly, the similarity matrices in Fig. 3(e–g) can be easily converted into three adjacency matrices that are visualized in a square matrix in Fig. 3(h–j) and allow for network representation. That is to say, the rationality for building an undirected graph heavily relies on the adjacency matrix, which provides evidence about the interactions in elements to shape the complex system. When the value in an adjacency matrix is larger than 0, two nodes have to be connected. On the contrary, the value of 0 indicates the absence of connections.

Finally, three different complex networks can be obtained, as displayed in Fig. 3(k–m), where nodes are representative of individual states of the US and the weight of an edge quantifying the relationship between the time series of two states is highlighted by color. The deep exploration of network topology can potentially reveal valuable patterns of pandemic evolution in the US.

4. Result analysis

4.1. Characteristics of established networks

The main role of the three established state-level networks is to model the dynamic behavior of a complex system. All networks are composed of multiple states with non-linear interactions. Since they are built based on different formation mechanisms named Pearson, TLCC, and DTW, they are bound to have their own properties. If all the 50 nodes within the network are fully connected, the network will contain 2500 links to unnecessarily increase the complexity of network analysis. For simplicity, we can only focus on the most important links. Herein, the criterion is set as follows: two individual states with the top 20% highest value of Pearson and TLCC or the top 20% smallest DTW need to be connected. As a result, there are only 500 links remaining in each network, while the number of connected states is 45, 44, and 40 in the network of Pearson, TLCC, and DTW, respectively. From the view of a network, Table 4 summarizes the topography of the entire network that is measured by a variety of network-level metrics. It can be seen that the DTW network owns the largest graph density, where around 64.1% of the total potential connections in a network can be actually observed. Inversely, the Pearson network has the largest network diameter of 5, indicating this network is more decentralized than others. As for the average clustering coefficient, its value in all the three networks exceeds 0.8, implying that there is a comparatively large tendency for states in the built networks to group together. In other words, the three targeted networks from Pearson, TLCC, and DTW can be considered as the high-clustering network containing some subgroups of states with a closer
Fig. 3. Proposed framework for network representation using COVID-19 data about states of the US. (a) and (b) The PCA varimax for feature preprocessing in the state CA. (c) and (d) Measurement of synchrony between their time series data obtained from PCA varimax for the state AL and CA. (e–g) The correlation or distance matrix for all pairs of states. (h–j) The adjacent matrixes based on methods of Pearson, TLCC, and DTW with the top 20% links representing the higher similarity between time series. (k–m) The complex network based on three adjacent matrixes.
relationship or higher similarity, and thus, the further exploration about community detection is meaningful. Besides, the average path length of the three networks is below the value of 2, which also verifies that the three undirected and weighted networks are highly clustered. Another thing to note is that these networks can well satisfy two properties: one is an average shortest path length representing a high level of global reachability; the other is the high clustering indicating great local connectivity. Therefore, it can be assumed that all established networks have a small-world property (Golbeck, 2013).

From the view of nodes, it is necessary to measure the prominence of a state to clearly reveal the intricate organization of a complex network. Since no general definition of centrality has developed, four centrality metrics that are defined by varied concepts, including the degree centrality, closeness centrality, eigenvector centrality, and betweenness centrality are calculated herein to quantify the importance of states in each network. Therefore, the top 10 states can be determined by different measurements of network centrality, which are more important than the other states. That is to say, variation trends of time series from other states are more likely to be consistent with these identified states that own comparatively a high value of centrality. It means that these states ranked high tend to contain common patterns of the COVID-19 spread in the US, which deserve more attention. These key states have been visualized in maps of Fig. 4, where the intensity of color and the size of the circle are scaled by the value of centrality. Through investigating these recognized significant states, it paves a convenient way for us to capture and estimate the rough trend and direction of the dynamical progress of COVID-19 in the US. Based on the centrality measurement, OH (Ohio) and UT (Utah) exhibit the maximum centrality in the Pearson network. IN (Indiana) and OH are recognized as the most important in the TLCC network. The top states under the highest centrality in the DTW network are MD (Maryland), OH, and UT. It can be therefore assumed that OH and UT are the representative of states in the east and west, respectively.

The top 10 important states identified by centrality have a slight

| Characteristics | Pearson | TLCC | DTW |
|-----------------|---------|------|-----|
| Number of nodes | 45      | 44   | 40  |
| Number of links | 500     | 500  | 500 |
| Graph Density   | 0.505   | 0.529| 0.641|
| Network diameter | 5       | 4    | 3   |
| Average clustering coefficient | 0.814 | 0.809 | 0.866 |
| Average path length | 1.635 | 1.542 | 1.371 |

Fig. 4. Maps of the top 10 important states in each network determined by centrality. (a,b) Results of the Pearson network. (c,d) Results of the TLCC network. (e,f) Results of the DTW network. (Note: dc, cc, ed, and bc are the abbreviation of degree centrality, closeness centrality, eigenvector centrality, and betweenness centrality.)
difference in the three established networks. A collection of recognized states in three networks along with their relationships are visualized in the Venn diagram of Fig. 5(a). According to the overlapping of three circles on behalf of a commonality, seven common states are considered to play the most important roles in all three networks. Besides, results from the Pearson network and the TLCC network have a higher level of commonality, while there are more recognized states in the DTW network standing alone outside the overlapping zone. For quantitatively understanding the similarities and differences between pair-wise networks, the Jaccard index is calculated by dividing the size of the intersection by the size of the union of the important states. As a result, the Jaccard index of the Pearson and TCLL network is calculated to be 0.706, which is 0.414 and 0.373 higher than the pair of the Pearson and DTW networks and the pair of TLCC and DTW networks, respectively. In the intersection of three circles, a total of seven states are shared in common, including CO (Colorado), MD, NV (Nevada), OH, PA (Pennsylvania), UT, and WV (West Virginia), which play key roles in all the three networks. It can be assumed that other 43 states are prone to share a similar tendency as the pandemic evolution in these seven detected states that are reasonably considered as the representatives to explore the path of COVID-19 in the US, and thus more attention needs to be paid to them.

Their corresponding conditions in terms of three indicators, including the accumulated positive cases, deaths, and total test results at the end of Feb 12, 2021, have been provided in Fig. 5(b). The three neighboring states called PA, OH, and MD locating in the east of the US conduct more tests and suffer from more positive cases and deaths. Reserve happens in the four neighboring western states named CO, NV, UT, and WV (West Virginia), which play key roles in all the three networks. It can be assumed that other 43 states are prone to share a similar tendency as the pandemic evolution in these seven detected states that are reasonably considered as the representatives to explore the path of COVID-19 in the US, and thus more attention needs to be paid to them.

According to the overlapping of three circles on behalf of a commonality, seven common states are considered to play the most important roles in all three networks. Besides, results from the Pearson network and the TLCC network have a higher level of commonality, while there are more recognized states in the DTW network standing alone outside the overlapping zone. For quantitatively understanding the similarities and differences between pair-wise networks, the Jaccard index is calculated by dividing the size of the intersection by the size of the union of the important states. As a result, the Jaccard index of the Pearson and TCLL network is calculated to be 0.706, which is 0.414 and 0.373 higher than the pair of the Pearson and DTW networks and the pair of TLCC and DTW networks, respectively. In the intersection of three circles, a total of seven states are shared in common, including CO (Colorado), MD, NV (Nevada), OH, PA (Pennsylvania), UT, and WV (West Virginia), which play key roles in all the three networks. It can be assumed that other 43 states are prone to share a similar tendency as the pandemic evolution in these seven detected states that are reasonably considered as the representatives to explore the path of COVID-19 in the US, and thus more attention needs to be paid to them.

4.2. Community detection for state clustering

Due to the large value of the average clustering coefficient, subgroups of states in a high topological dense will exist within all the three established networks. An important topic of interest to fully understand the network structure is called community detection. For this purpose, 10 state-of-the-art algorithms have been tried to decompose an entire complex network into several meaningful groups for the discovery of internal network organization. Regarding all alternative algorithms, the grid-search strategy is implemented on parameter space, aiming to search the optimal parameters automatically to maximize the value of modularity in each sub-group. The performance of these 10 methods has been compared based on the evaluation metrics about the quality of extracted communities. Evaluation of the quality of extracted communities has been listed in Table 5. According to the comparison of two modularity scores, it can be found that Diffusion Entropy Reducer (DER) turns out to be the superior one resulting in the highest Newman-Girvan modularity of 0.129 (Pearson network), 0.119 (TLCC network), 0.054 (DTW network) and the highest z modularity of 0.260 (Pearson network), 0.239 (TLCC network), 0.108 (DTW network). Since DER can create the best partitions under the highest degree of compactness, it is adopted as the preferable technique to shed light on the structure of the three built networks.

Since the whole network will not be broken up randomly, some hidden information that cannot be obtained via simple observation is certain to be incorporated in the detected communities. Fig. 6 presents the clustering results derived from DER. States grouped within each community are clearly visualized in the complex network, and their geographical distributions are provided on the map. The node size in Fig. 6(a,c,e) is proportional to the node’s degree. The null community in Fig. 6(b,d,f) represents the isolated nodes, meaning that these states do not contain the established network. Evidently, two disjoint parts that
are distinguished in blue and orange color are made up of each network. States grouped into the same community are better connected internally and more sparsely linked to states in the rest of the network. Therefore, a cohesive cluster is made up of a set of states owning a similar pattern and tendency of time series about the COVID-19 evolution. Especially in the Pearson network and TLCC network, the distribution of states for forming a community is comparatively regular. It can be seen that states in central and northern America tend to be grouped into a community, while states in the north, south, and east of the US constitute another community. Although the DTW network has no such distinguishing pattern regarding the spatial distribution, a general regularly still exists that a community is more likely to consist of states that are very close geographically. It also implies that the neighboring states will share more sequential features in common.

In short, the value of community detection in this research should be highlighted from two perspectives. For one thing, the DER algorithm can separate different color points apart to return a clear community division. Consequently, groups of states that are unable to be directly observed can be identified relying on network topology. For another, convincing results from community detection have shown their flexibility in capturing the hidden spatial relations about the pandemic progress in different states, thereby implying that geographical location could be an important difference between two communities. In a sense, these meaningful communities can further validate the great reasonability and effectiveness of the developed process for network construction and partition. Besides, each detected community has its own structural features, which have been measured by a series of metrics and summarized in Table 6.

### 4.3. Knowledge discovery from detected communities

Apart from network structure and spatial distribution, another point to be aware of is that there is a lot of knowledge associated with the practical value embedded in the two communities. It is believed that each community locating in a specific space can demonstrate some typical characteristics, such as the pandemic behavior, the corresponding epidemiological and operational conditions, and social factors, and thus deeper investigations are required. From the point of pandemic behavior, three indicators including the number of positive cases, deaths, and total test results are taken into account. Their corresponding increase rates on a weekly basis have been provided in Fig. 7, where the mean variations of all states in communities 1 and 2 are differentiated by orange and blue. The light orange and blue stand for a 95% confidence interval of the target time series, and the central line is the mean. The characteristics of communities derived from different networks are varied. Observably, apparent hysteresis exists in Fig. 7(a,b,d,e) reflecting the Pearson network and TLCC network. The peak at the increased rate of positive cases and death in community 1 lags around 10 weeks against the time series of community 2. Especially during week 30–40, conditions of states grouped into community 2 in Pearson and TLCC were worse off, which caused a steeper growth curve about positive cases and deaths and reached up to the maximum at a faster speed. Then, the severity of COVID-19 in community 2 would take several days to extend to states in community 1 and exert negative impacts on them. This offers one possible explanation about the 10-week delay of the peak in community 1. In the DTW network, although communities 1 and 2 exhibit very similar times series about the increase of the positive cases and deaths (Fig. 7(g,h)), their different features root in the total test results. From Fig. 7(i), it seems clear that community 1 in the DTW network is prone to aggregate states experiencing higher growth in tests.

From the point of epidemiological and operational conditions at the state level, the accumulated data on the last day of the targeted period (Feb 12, 2021) is taken to evaluate and infer the COVID-19 scenarios between two communities. As shown in Fig. 8, knowledge embedded in communities from three established networks is diverse which could supplement each other to facilitate better track and more complete comprehension of the pandemic evolution across the US. For the Pearson network and TLCC network, it can be found from Fig. 8(a,b) that corresponding states in community 2 are more likely to have larger value of accumulated positive cases/population, but they averagely achieve a lower value of accumulated deaths/population than community 1. To be more specific, the mean of accumulated positive results/population about communities 1 and 2 is 1.115 (IQR: [0.939, 1.204]) and 0.720 (IQR: [0.615, 0.843]), respectively.

Another possible reason lies in the testing level. States in community 1 generally shows a lower infection rate while community 2 could probably demonstrate a higher survival rate. Therefore, an important characteristic behind these two networks to be noted is that community 1 generally shows a lower infection rate while community 2 could probably demonstrate a higher survival rate. Another possible reason lies in the testing level. States in community 1 may pick up more severe cases, leading to high mortality. Distinctively, the significant difference in DTW network that can be directly perceived can be discovered from Fig. 8(c), where the average value of accumulated total test results/population about communities 1 and 2 is 1.115 (IQR: [0.939, 1.204]) and 0.720 (IQR: [0.615, 0.843]), respectively. That is to say, if states are more active in conducting PCR tests to check infection, they are more likely to gather together.

In addition, another way to easily uncover the significant difference between communities is the linear regression analysis. In Fig. 8(d,f), linear regression analysis is performed to fit a linear equation for the observed data from two pairs of features: one is accumulated positive

### Table 5

Clustering evaluation and comparison from different community detection algorithms.

| Algorithm                      | Year | Pearson Newman-Girvan modularity | Z modularity | Pearson Newman-Girvan modularity | Z modularity | TLCC Newman-Girvan modularity | Z modularity | DTW Newman-Girvan modularity | Z modularity |
|--------------------------------|------|-----------------------------------|--------------|-----------------------------------|--------------|--------------------------------|--------------|--------------------------------|--------------|
| Girvan Newman (Girvan & Newman, 2002) | 2002 | 0.084                            | 0.040        | 0.000                            | 0.000        | 0.000                          | 0.000        | 0.107                          | -0.001       |
| Greedy modularity (Clauset, Newman, & Moore, 2004) | 2004 | 0.122                            | 0.255        | 0.111                            | 0.228        | 0.059                          | 0.059        | 0.120                          | -0.001       |
| Walktrap (Pons & Latapy, 2005) | 2005 | 0.078                            | 0.173        | 0.114                            | 0.228        | 0.023                          | 0.023        | 0.046                          | -0.002       |
| Eigenvector (Newman, 2006)      | 2006 | 0.122                            | 0.255        | 0.118                            | 0.236        | 0.054                          | 0.054        | 0.111                          | -0.002       |
| Louvain (Blondel et al., 2008)  | 2008 | 0.122                            | 0.255        | 0.116                            | 0.232        | 0.052                          | 0.052        | 0.107                          | -0.002       |
| Belief (Zhang & Moore, 2014)    | 2014 | 0.013                            | 0.034        | 0.008                            | 0.021        | -0.032                         | -0.032       | 0.032                          | -0.002       |
| Significance communities (Traag et al., 2013) | 2014 | 0.066                            | 0.135        | 0.040                            | 0.081        | -0.016                         | -0.016       | 0.032                          | -0.002       |
| Diffusion Entropy Reducer (DER) (Kozdoba & Mannor, 2015) | 2015 | 0.129                            | 0.260        | 0.119                            | 0.239        | 0.054                          | 0.054        | 0.108                          | -0.002       |
| Fluid (Parés et al., 2017)      | 2017 | 0.098                            | 0.209        | 0.075                            | 0.167        | 0.054                          | 0.054        | 0.107                          | -0.002       |
| Leiden (Traag et al., 2019)     | 2018 | 0.123                            | 0.256        | 0.116                            | 0.232        | 0.052                          | 0.052        | 0.104                          | -0.002       |
When the $p$-value is below the significant level of 0.05, the observed data provide enough evidence to reject the null hypothesis that the independent variable has no correlations with the dependent variable. In other words, the positive-death models can better fit data about a certain community, which are more convincing than the positive-test models. That implies that modeling the relationship between accumulated positive cases/population and accumulated deaths/population could promise practical value. To further assess the COVID-19 risk in different communities, the fitted lines can be visually compared by the slope coefficients. It is evident that an obvious distinction between two communities in the Pearson network and TLCC network lies in the relationship of accumulated positive cases/population and accumulated deaths/population, where community 1 with a larger slope indicates that an increase of positive cases is associated with a greater increase in

![Fig. 6. Community detection results based on the DER algorithm. (a,b) Results of the Pearson network with 26 states in community 1 and 19 states in community 2. (c,d) Results of the TLCC network with 25 states in community 1 and 19 states in community 2. (e,f) Results of the DTW network with 20 states in community 1 and 20 states in community 2.](image)

![Table 6](table)

| Metrics                   | Pearson Mean ($\pm$ std) | TLCC Mean ($\pm$ std) | DTW Mean ($\pm$ std) |
|---------------------------|--------------------------|-----------------------|----------------------|
| Newman-Girvan modularity  | 0.129 ($\pm$ 0.054)      | 0.119 ($\pm$ 0.054)   | 0.054 ($\pm$ 0.054)  |
| Z modularity              | 0.260 ($\pm$ 0.108)      | 0.239 ($\pm$ 0.108)   | 0.108 ($\pm$ 0.108)  |
| Average internal degree   | 14.269 ($\pm$ 4.269)     | 14.429 ($\pm$ 4.851)  | 13.850 ($\pm$ 0.550) |
| Scaled density            | 1.284 ($\pm$ 0.184)      | 1.263 ($\pm$ 0.257)   | 1.137 ($\pm$ 0.045)  |
| Average transitivity      | 0.790 ($\pm$ 0.067)      | 0.817 ($\pm$ 0.098)   | 0.882 ($\pm$ 0.002)  |
| Average embeddedness      | 0.732 ($\pm$ 0.032)      | 0.702 ($\pm$ 0.047)   | 0.595 ($\pm$ 0.033)  |
| Fraction over median degree | 0.422 ($\pm$ 0.001) | 0.404 ($\pm$ 0.036) | 0.450 ($\pm$ 0.000) |
| Average out degree fraction | 7.470 ($\pm$ 1.162) | 7.781 ($\pm$ 1.061) | 11.150 ($\pm$ 0.000) |
deaths. It also validates that the Pearson network and TLCC networks are prone to group states that are at a comparatively high risk of mortality into the same cluster. In the DTW network, the relationship between accumulated positive cases/population and accumulated total test results/population in states of community 1 is significantly different from community 2.

From the point of social factors, the two detected communities can potentially coordinate with the political party, GDP, and population density of their corresponding states. The relationships between communities and three social factors are depicted in the chord graphs in Fig. 9. According to the size of the arc that is proportional to the number of states, an interesting finding is that community 1 in the Pearson network and TLCC network tend to support the Democratic Party of the United States, whose GDP and population density are likely to keep above the median. The social condition of community 2 is the opposite. However, these links in the DTW network are much weaker than the other two networks. The social characteristics of two communities in three networks are detailed in Table 7. It is observed that the p-value for the three social factors in the Pearson network is all smaller than 0.05. That is to say, the Pearson network deserves deeper explorations than...
the other two networks, since its social factors can significantly differ between communities. More specifically, 17 of the 26 states belonging to community 1 support the Democratic Party, while only 3.158% of states in community 2 support the Republican Party. The median GDP in community 1 (268.32, IQR: [90.983, 543.633]) is more than twice that in community 2 (171.84, IQR: [52.090, 332.610]). Meanwhile, community 1 (134.2, IQR: [58.195, 260.988]) tends to triple in median of population density from community 2 (56.640, IQR: [11.090, 167.010]). Similarly, the difference between GDP and population density within two communities can also be detected in the TLCC network (p-value < 0.05). That is to say, there are some associations between COVID-19 conditions and social factors in the discovered communities from both the Pearson network and TLCC network, which therefore paves a new way to understand the nature of COVID-19 across the US.

Note: The unit of GDP is in billion chained US dollars. The unit of population density is the number of residuals per square mile. IQR is the interquartile range. The p-value for the categorical variable named party is calculated by the Fisher’s exact test. The p-value for the continuous variable named GDP and population density is calculated by the Wilcoxon rank-sum test. The p-value < 0.05 has been highlighted.

4.4. Trend prediction

Notably, it is not a straightforward task to accurately estimate the future trend of the COVID-19 evolution over time and space. The difficulty mainly comes from two aspects. For one thing, the pandemic could keep fluctuating day by day and vary state by state. For another, the implementation of non-pharmaceutical interventions along with their influence is hard to predict and can dynamically change. Based upon meaningful clusters retrieved from community detection, it is believed that states belonging to the same community could share comparatively a high similarity in time series about the pandemic spread. For simplicity, states organized into a community can be treated as a whole to understand the nature of COVID-19 spread and make cluster-specific time series predictions at the community level. Herein, the online random forest based on a sliding window can be employed to continuously learn daily records and automatically update the predictive model. Due to its powerful ability to capture the dynamic of virus behavior, it is bound to improve the quality of predictions over time. Through forecasting the daily increase of positive cases and deaths in each community, it is possible to alert the local government in advance and formulate
In this case, the prediction begins at the 144th day (Jul 28, 2020), and the forecasting process lasts 200 consecutive days. A detailed description of establishing the online random forest has been provided in the methodology. Time series describing communities derived from Pearson, TLCC, and DTW networks are collected on a daily basis, which will be well fitted by their own cluster-specific forecasted models to predict the unknown future. Notably, the use of prior time steps to predict the next step is a key of the sliding window method, aiming to ensure the overall model performance in predicting COVID-19 trends in terms of daily increase of positive cases and deaths. In this regard, an essential skill is to select the appropriate size of a rolling time window to catch the real-time changes caused by both the pandemic dynamics and control measures. Herein, the number of previous days for prediction is set as 7 days, 14 days, 21 days, and 28 days to conduct a series of comparative experiments. The corresponding results have been listed in Appendix Tables 1–3, where the best window size has been highlighted by bold font. It is evident that a proper window size can drive the model to gain a significant decrease in RMSE and MAE. Take the increase of deaths in community 1 from the Pearson network as an example. Under the rolling time-window size of 21 days, its RMSE and MAE can reach the minimum, which can be reduced by about $7.5 \times 10^{-6}$ and $1.07 \times 10^{-8}$ compared to the 7-days window.

Based on these determined optimal window sizes, Fig. 10 visualizes the promising prediction results in terms of the daily increase of positive cases and deaths against their observed value, and Table 8 measures the prediction performance. Time series about the last 200 days of the studying period are predicted on a daily basis. The light orange and deep orange lines are the ground truth and prediction of community 1, respectively. The light blue and deep blue lines are the ground truth and prediction of community 2, respectively. The scatter plot inside evaluates the prediction performance in an intuitive manner, which displays the degree of deviation for the estimated value from the actual value. It is observed that curves in Fig. 10 representing the predictive value seem to be highly consistent with the trend and value of the ground truth. Such a high level of agreement between the estimated and measured values has occurred in all communities from three networks during the targeted period. Besides, scatter plots as the subplots of Fig. 10 show that points that compare the actual and predicted value are likely to concentrate on $y=x$, which also validates the reliability of the proposed prediction model. As listed in Table 8, the optimal window width for

![Fig. 10](image-url)
forecasting the death rate is larger than 14 days in all discovered communities. That is to say, historical conditions from the previous two weeks or more even sooner are taken into account, which may exert great impacts on the current mortality. In contrast, the estimation on the case increase tends to rely on a smaller time window involving fewer historical data within 14 days. It can be assumed that the hysteretic nature in mortality changes will be more significant. Moreover, all communities can reach low RMSE and MAE and high R2, numerically revealing the success and generalization of the online random forest in complex networks under different mechanisms. In particular, the task for forecasting the growth of positive cases can achieve the average R2 value higher than 0.874, which is likely to be more precise than the prediction of death increase with the average R2 below 0.765. Besides, the prediction accuracy of different networks can be ranked from the best to the worst: Pearson > TLCC > DTW (an increase of positive cases), and DTW > TLCC > Pearson (an increase of deaths). Therefore, communities embedding in the Pearson and TLCC networks are more suitable to predict the daily increase of future cases, while communities from the DTW network are more likely to exhibit a temporal trend in mortality to better estimate how the sequence of deaths will keep on in the near future. 

Table 9 and 10 compare the prediction performance of the online random forest and the classical random forest in terms of RMSE and MAE. This comparison experiment validates that the random forest using the online learning technique can outperform its classical version that only takes the previous days’ value, bringing about at least 50% of improvement in lowering the error between the prediction and ground truth.

5. Conclusion

As an effort of promoting sustainable social development, the goal of this research is to propose a novel network-enabled approach for exploring the spatial-temporal spread of COVID-19 among states of the US. This new perspective in the pandemic investigation is motivated by the idea that states across the US constitute a complex social system under high-dimensional data and nonlinear phenomena, which eventually provides insights into the inherent interactions of pandemic evolution among states. The contributions lie in two aspects: (a) From the state of knowledge, the state-level epidemiological situations can be reasonably monitored, evaluated, and estimated through the combination of intelligent data mining methods, including PCA varimax, synchrony calculation, network structure measurement, community detection algorithm named DER, knowledge discovery related to evolution process and social factors, and cluster-specific prediction using the online random forest. (b) From the state of practice, the extracted spatial-temporal patterns from the data-driven complex network can serve as rich scientific evidence to systematically understand the dynamic behavior of COVID-19 from a new point. It can potentially support state-scale prevention and intervention for better controlling the risk of infection and death.

The effectiveness of the developed network-enabled approach has been experimentally verified in the observable data about the number of positive cases, deaths, and total test results in each state of the US during Mar 7, 2020 and Feb 12, 2021. There are some interesting founds to be noted. Firstly, although the three networks derived from different logics have their unique topological characteristics, they recognize seven common states with relatively high centrality as the top important nodes, including CD, MD, NV, OH, PA, UT, and WV. For simplicity, we can focus more on these representative states to study the COVID-19 scenarios in the US. Secondly, since the topological structure of three networks is proven to be highly clustered, the DER-based community detection can gather states owning patterns of time series in great similarity. It is found that two discovered communities within a state can eventually provide insights into the inherent interactions of pandemic evolution among states. The contributions lie in two aspects: (a) From the state of knowledge, the state-level epidemiological situations can be reasonably monitored, evaluated, and estimated through the combination of intelligent data mining methods, including PCA varimax, synchrony calculation, network structure measurement, community detection algorithm named DER, knowledge discovery related to evolution process and social factors, and cluster-specific prediction using the online random forest. (b) From the state of practice, the extracted spatial-temporal patterns from the data-driven complex network can serve as rich scientific evidence to systematically understand the dynamic behavior of COVID-19 from a new point. It can potentially support state-scale prevention and intervention for better controlling the risk of infection and death.

As an effort of promoting sustainable social development, the goal of this research is to propose a novel network-enabled approach for exploring the spatial-temporal spread of COVID-19 among states of the US. This new perspective in the pandemic investigation is motivated by the idea that states across the US constitute a complex social system under high-dimensional data and nonlinear phenomena, which eventually provides insights into the inherent interactions of pandemic evolution among states. The contributions lie in two aspects: (a) From the state of knowledge, the state-level epidemiological situations can be reasonably monitored, evaluated, and estimated through the combination of intelligent data mining methods, including PCA varimax, synchrony calculation, network structure measurement, community detection algorithm named DER, knowledge discovery related to evolution process and social factors, and cluster-specific prediction using the online random forest. (b) From the state of practice, the extracted spatial-temporal patterns from the data-driven complex network can serve as rich scientific evidence to systematically understand the dynamic behavior of COVID-19 from a new point. It can potentially support state-scale prevention and intervention for better controlling the risk of infection and death.

The effectiveness of the developed network-enabled approach has been experimentally verified in the observable data about the number of positive cases, deaths, and total test results in each state of the US during Mar 7, 2020 and Feb 12, 2021. There are some interesting founds to be noted. Firstly, although the three networks derived from different logics have their unique topological characteristics, they recognize seven common states with relatively high centrality as the top important nodes, including CD, MD, NV, OH, PA, UT, and WV. For simplicity, we can focus more on these representative states to study the COVID-19 scenarios in the US. Secondly, since the topological structure of three networks is proven to be highly clustered, the DER-based community detection can gather states owning patterns of time series in great similarity. It is found that two discovered communities within a state can eventually provide insights into the inherent interactions of pandemic evolution among states.
concerning the political party, GDP, and population density of the state are significantly different between two detected communities in the Pearson or TLCC network. Meanwhile, these two communities also show an obvious difference in the rising tendency of positive cases or deaths. As for the DTW network, its communities differ from the operational conditions in light of the total PCR tests. Lastly, the cluster-specific models for time series prediction can keep updating on a daily basis as new data are reported. Online learning facilitates a more precise estimation of case increase and death rate, yielding the average R² value of a network larger than 0.874 and 0.734, respectively. Such preview and predictive information about the near future can be introduced to help policymakers to take early warning, preparation, and response towards the ongoing pandemic, which eventually enhances the resilience of human societies.

6. Future works

There are still some limitations in this study. The future avenues can be performed from the following points. Firstly, the more advanced casual discovery method named PCMCI (Runge et al., 2019) can be conducted to measure the true causal dependency between time series. A time-lagged and directed interdependency network can be reasonably created as a casual interpretation, which helps to further uncover the physical mechanism and infer the causal effects. Secondly, some more useful features, such as the number of vaccinated people, the control measures on the risk of infection, and others, can be taken into account for more reliable analysis and prediction. Thirdly, the current research mainly focuses on discovering hidden knowledge from the accumulated COVID-19 records relying on the complex network theory, which stays at the data mining level. An issue worthy of discussion is that how to make the utmost of the obtained result for containing the virus spread and reducing the infection risk. Thus, it requires more attempts to discuss the implication and insights extracted from the identified spatial-temporal patterns, where multi-objective optimization (Zhang & Lin, 2021) can be conducted to aid health authorities to make data-driven decisions and guide the effective prevention and control suggestions in terms of quarantines, social distance measures, personal protective measures, and vaccination. Also, it is necessary to consult experts and view reports, aiming to verify that whether the results are fully logical and compatible with the current state of the situation and how well the control strategies can actually impede the virus transmission. Fourthly, the application range of the network-enabled analysis can be extended to the country level instead of the state level within the US, which has the potential value in modeling multiple time series of countries into a complex network for enhancing the understanding of COVID-19 dynamic natures around the world.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. All of the sources of funding for the work described in this publication are acknowledged. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property. We declare that they have no competing interests. We confirm that the manuscript has been read and approved by all named authors. We confirm that the order of authors listed in the manuscript has been approved by all named authors.

Acknowledgments

This study is supported by the Ministry of Education Tier 1 Grants, Singapore (No. 04MNP000279C120, No. 04MNP002126C120) and the Start-Up Grant at Nanyang Technological University, Singapore (No. 04NS000423C120).

Appendix

Appendix Table 1. Prediction performance under different sizes of rolling time window for communities in the Pearson network.

| Community | Increase of positive cases |  |  | Increase of deaths |  |  |
|-----------|---------------------------|---|---|-------------------|---|---|
|           | Time window | RMSE | MAE | R²   | Time window | RMSE | MAE | R²   |
| Community 1 | 7 days | 6.80E-05 | 4.62E-05 | 9.01 | 7 days | 1.57E-06 | 9.93E-07 | 0.794 |
| 14 days | 6.73E-05 | 4.54E-05 | 9.02 | 21 days | 1.53E-06 | 9.61E-07 | 0.804 |
| 21 days | 7.04E-05 | 4.73E-05 | 0.89 | 21 days | 1.49E-06 | 9.82E-07 | 0.813 |
| 28 days | 7.27E-05 | 4.84E-05 | 0.88 | 28 days | 1.53E-06 | 1.03E-06 | 0.803 |
| Community 2 | 7 days | 9.90E-05 | 6.62E-05 | 0.86 | 14 days | 2.50E-06 | 1.06E-06 | 0.629 |
| 14 days | 9.70E-05 | 6.74E-05 | 0.87 | 21 days | 2.41E-06 | 1.56E-06 | 0.655 |
| 21 days | 1.00E-04 | 6.84E-05 | 0.86 | 21 days | 2.44E-06 | 1.63E-06 | 0.648 |
| 28 days | 9.82E-05 | 7.04E-05 | 0.87 | 28 days | 2.47E-06 | 1.65E-06 | 0.646 |
References

An, S., Gao, X., Jiang, M., & Sun, X. (2018). Multivariate financial time series in the light of complex network analysis. Physica A: Statistical Mechanics and its Applications, 503, 1241–1255.

Ardelt, N., Barense Segota, S., Lorenzini, J., Jurtil, Z., Suuterek, T., Ilagoevici, A., Protic, A., Cabov, T., Filipovit, N., & Cai, Z. (2021). Estimation of covid-19 epidemiology curve of the united states using genetic programming algorithm. International Journal of Environmental Research and Public Health, 18, 959.

Baggio, G., Bassett, D. S., & Pasqualetti, F. (2021). Data-driven control of complex networks. Nature Communications, 12, 1429.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009).extérieur. Journal of Statistical Mechanics: Theory and Experiment, 2008, P10008.

Chew, A. W. Z., Pan, Y., Wang, Y., & Zhang, L. (2021a). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2021, Article 107417.

Chew, A. W. Z., Wang, Y., & Zhang, L. (2021b). Correlating dynamic climate conditions and socioeconomic-governmental factors to spatiotemporal spread of COVID-19 via semantic segmentation deep learning analysis. Sustainable Cities and Society, 75, Article 103231.

Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons and Fractals, 135, Article 109864.

Clauwaert, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical Review E, 70(6), Article 066111.

Demertzis, K., Tsiotas, M., & Magafas, L. (2020). Modeling and forecasting the COVID-19 temporal spread in Greece: An exploratory approach based on complex network defined splits. International Journal of Environmental Research and Public Health, 17, 4693.

Flake, G. W., Lawrence, S., & Giles, C. L. (2000). Efficient identification of web communities. In Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining (pp. 150–160).

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. In Proceedings of the national academy of sciences 99 (pp. 7821–7826).

Goldbeck, J. (2013). Analyzing the social web. Newnes.

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. International Journal of Climatology: A Journal of the Royal Meteorological Society, 27, 1119–1125.

Ho, Y., Tang, M., Wu, Z., Xiao, Z., & Shen, G. Q. (2019). The evolution of patterns within embodied energy flows in the Chinese economy: A multi-regional based complex network approach. Sustainable Cities and Society, 47, Article 101500.

Hu, F., Huang, M., Sun, J., Zhang, X., & Liu, J. (2021). An analysis model of diagnosis and treatment for COVID-19 pandemic based on medical information fusion. Information Fusion, 73, 11–21.

C. Jhuomcr, 2020. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) 2020. https://coronavirus.jhu.edu/map.html.

T.I. HME COVID-19 Forecasting Team. (2020). Modeling COVID-19 scenarios for the United States Nature Medicine, 27, 94–105.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187–200.

Kozloba, M., & Mannor, S. (2015). Community detection via measure space embedding. Advances in Neural Information Processing Systems, 28, 2890–2898.

Lak, A., Sharifi, A., Baie, S., Zali, A., Maher, A., Mostafavi, E., & Khalili, D. (2021). Spatio-temporal patterns of the COVID-19 pandemic, and place-based influential factors at the neighborhood scale in Tehran. Sustainable Cities and Society, 72, Article 103034.

Li, B., Peng, Y., He, H., Wang, M., & Feng, T. (2021). Built environment and early infection of COVID-19 in urban districts: A case study of Hangzhou. Sustainable Cities and Society, 66, Article 102685.

Li, Y., Chen, H., & Wu, Z. (2010). Dynamic time warping distance method for similarity test of multipoint ground motion field. Mathematical Problems in Engineering 2010, Article 749517.

Maiti, A., Zhang, Q., Samnig, S., Pramanik, S., Khakraborti, S., Cerda, A., & Pilla, F. (2021). Exploring spatiotemporal effects of the driving factors on COVID-19 incidences in the contiguous United States. Sustainable Cities and Society, 68, Article 102784.

Miyachi, A., & Kawase, Y. (2016). Z-score-based modularity for community detection in networks. PloS one, 11, Article e0147805.

Müller, Meinard (2007). Dynamic time warping. Information retrieval for music and motion. Berlin, Heidelberg: Springer.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74, Article 036104.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69, Article 026113.

Noh, J., & Danuser, G. (2021). Estimation of the fraction of COVID-19 infected people in US states and countries worldwide. PloS one, 16, Article e0246772.

Nowack, P., Runge, J., Eyring, V., & Haigh, J. D. (2020). Causal networks for climate and related techniques in atmospheric science: A review. Springer.

Pan, Y., & Zhang, L. (2021b). Roles of artificial intelligence in construction engineering and related techniques in atmospheric science: A review. Advances in Neural Information Processing Systems, 34, 265–279.

Pan, Y., Zhang, L., & Skibniewski, M. J. (2020). Clustering of designers based on building information modeling event logs. Computer-Aided Civil and Infrastructure Engineering, 35, 701–718.

Pan, Y., Zhang, L., Yan, Z., Lwin, M. O., & Skibniewski, M. J. (2021). Discovering optimal strategies for mitigating COVID-19 spread using machine learning: Experience from Asia. Sustainable Cities and Society, Article 102524.
Sustainable Cities and Society xxx (xxxx) xxx

Y. Pan et al.

Paréns, F., Gasulla, D. G., Vilalta, A., Moreno, J., Ayguadé, E., Labarta, J., Cortes, U., & Suzumura, T. (2017). Fluid communities: A competitive, scalable and diverse community detection algorithm (pp. 229–240). Springer. International Conference on Complex Networks and their Applications.

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks (pp. 284–293). Springer. International symposium on computer and information sciences.

Qiao, T., Shan, W., Zhang, M., & Liu, C. (2019). How to facilitate knowledge diffusion in well-connected communities. A competitive, scalable and diverse community detection algorithm (pp. 229–240). Springer. International Conference on Complex Networks and their Applications.

Rahman, M. A., Zaman, N., Asyhari, A. T., Al-Turjman, F., Bhuiyan, M. Z. A., Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting fluid communities: A competitive, scalable and diverse community detection algorithm (pp. 229–240). Springer. International Conference on Complex Networks and their Applications.

Richman, M. B. (1986). Rotation of principal components. Journal of Climatology, 6, 293–335.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5, eaau4996.

Shahdoosti, H. R., & Ghassemeni, H. (2016). Combining the spectral PCA and spatial PCA fusion methods by an optimal filter. Information Fusion, 27, 150–160.

Shorffuzzaman, M., Hossain, M. S., & Alhamid, M. F. (2021). Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic. Sustainable Cities and Society, 64, Article 102582.

Traag, V. A., Krings, G., & Van Dooren, P. (2013). Significant scales in community structure. Scientific Reports, 3, 1–10.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports, 9, 1–12.

Tsotras, D., & Magafas, I. (2020). The effect of anti-COVID-19 policies on the evolution of the disease: A complex network analysis of the successful case of Greece. Physics, 2, 325–339.

Unwin, H. J. T., Mishra, S., Bradley, V. C., Gandy, A., Mellan, T. A., Coupland, H., Ish-Horowicz, J., Vollmer, M. A. C., Whittaker, C., Filipp, S. L., Xi, X., Monod, M., Ratmann, O., Hutchinson, M., Valka, F., Zhu, H., Hawryluk, I., Milton, P., Ainslie, K. E. C., Bagarlin, M., Boonyasiri, A., Brazeau, N. F., Cattarino, L., Cucunuba, Z., Cuzzo-Dannenbourg, G., Dorigatti, I., Eales, O. D., Eaton, J. W., van Elsland, S. L., FitzJohn, R. G., Gaythorpe, K. A. M., Green, W., Hinsley, W., Jeffrey, B., Knock, E., Laydon, D. J., Lees, J., Nedjati-Gilani, G., Nouvellet, P., Okell, L., Pag, K. V., Sivertsen, I., Thompson, H. A., Walker, P., Walters, C. E., Watson, O. J., Whittles, L. K., Ghanai, A. C., Ferguson, N. M., Riley, S., Donnelly, C. A., Bhattacharya, S., & Flaxman, S. (2020). State-level tracking of COVID-19 in the United States. Nature Communications, 11, 6189.

Wang, Z., Hill, D. J., Chen, G., & Ding, Z. Y. (2017). Power system cascading risk assessment based on complex network theory. Physica A: Statistical Mechanics and Its Applications, 482, 532–543.

Woolf, S. H., Chapman, D. A., & Lee, J. H. (2021). COVID-19 as the leading cause of death in the United States. Journal of the American Medical Association, 325, 123–124.

Xue, F., Li, W., Chen, K., & Webster, C. J. (2019). BIM reconstruction from 3D point clouds: A semantic registration approach based on multimodal optimization and architectural design knowledge. Advanced Engineering Informatics, 42, 181–213.

Yeşilkömat, C. M. (2020). Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithms. Chaos Solitons and Fractals, 140, Article 110210.

Zhang, L., & Lin, P. (2021). Multi-objective optimization for limiting tunnel-induced damages considering uncertainties. Reliability Engineering and System Safety, 216, 107945.

Zhang, P., & Moore, C. (2014). Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. Proceedings of the National Academy of Sciences, 111, 18144–18149.

Zhao, J., Huang, S., Huang, Q., Wang, H., Leng, G., & Fang, W. (2020). Time-lagged response of vegetation dynamics to climatic and teleconnection factors. Catena, 189, Article 104474.

Zhang, Y., Yang, H., Zhang, Y., & Li, P. (2020). Online random forests regression with memories. Knowledge-Based Systems, 201, Article 106058.

Zhou, C., Kong, T., Jiang, S., Chen, S., Zhou, Y., & Ding, L. (2020). Quantifying the evolution of settlement risk for surrounding environments in underground construction via complex network analysis. Tunnelling and Underground Space Technology, 103, Article 103490.