Novel Graph Topology Learning for Spatio-Temporal Analysis of COVID-19 Spread

Baoling Shan, Xin Yuan, *Member, IEEE*, Wei Ni, *Senior Member, IEEE*, Xin Wang, *Fellow, IEEE*, Ren Ping Liu, *Senior Member, IEEE*, and Eryk Dutkiewicz, *Senior Member, IEEE*

***Abstract*—This article presents a new graph-learning technique to accurately infer the graph structure of COVID-19 data, helping to reveal the correlation of pandemic dynamics among different countries and identify influential countries for pandemic response analysis. The new technique estimates the graph Laplacian of the COVID-19 data by first deriving analytically its precise eigenvectors, also known as graph Fourier transform (GFT) basis. Given the eigenvectors, the eigenvalues of the graph Laplacian are readily estimated using convex optimization. With the graph Laplacian, we analyze the confirmed cases of different COVID-19 variants among European countries based on centrality measures and identify a different set of the most influential and representative countries from the current techniques. The accuracy of the new method is validated by repurposing part of COVID-19 data to be the test data and gauging the capability of the method to recover missing test data, showing 33.3% better in root mean squared error (RMSE) and 11.11% better in correlation of determination than existing techniques. The set of identified influential countries by the method is anticipated to be meaningful and contribute to the study of COVID-19 spread.**

***Index Terms*—Graph learning, graph Laplacian, COVID-19.**

I. Introduction

GLOBAL health, economic, and social challenges are escalating due to the Coronavirus disease 2019 (COVID-19) pandemic. As of April 2022, Europe had 192.09 million confirmed cases and over two million deaths.[1] The SARS-CoV-2 virus has undergone numerous genetic changes since its discovery [1]. While some of these changes do not affect the virus's behavior, others may affect how easily it is transmitted. Changes beneficial to the virus tend to spread more quickly, which means that variants harboring them gradually replace other circulating variants [2]. In November 2020, SARS-CoV-2 Alpha was first detected in the United Kingdom, which was estimated to be 50% more transmissible than the original strain. From July 2021 to October 2021, SARS-CoV-2 Delta prevailed in Europe. The SARS-CoV-2 Omicron variant took over from the SARS-CoV-2 Delta variant in Europe in November 2021. Earlier studies demonstrated that Omicron can, to a degree, evade the protective effects of antibodies induced by vaccinations or natural infections. Large portions of the European population are susceptible to infection, leading to sharp increases in COVID-19 cases and unprecedented community spread.

Comprehending the spatio-temporal characteristics of the virus spread is the key to controlling the spread of the pandemic. Studies show that the global spread of the COVID-19 pandemic did not process uniformly [3], [4]. An outbreak's size and condition are influenced by the characteristics of virus spread [5]. Unfortunately, it is difficult to implement evidence-based policies for COVID-19 due to a lack of adequate evidence in policy-making and research [6]. While it is possible to estimate the growth rates of confirmed cases and deaths [7], the relationships between pairs of countries are still unknown as far as the COVID-19 development is concerned. Datasets about ongoing situations in different countries are likely to show spatial-temporal patterns since virus spread tends to follow geographic trends. A spatial-temporal analysis of confirmed COVID-19 cases may also shed light on its evolution. The record of pandemic evolution in Europe is known to be complex, variable, and non-linear. Consequently, it is essential to uncover hidden information about SARS-CoV-2 as new virus variants emerge.

One way to understand the spreading dynamic of the pandemic is to generate and analyze COVID-19 pandemic diffusion graph topologies with the graph-theoretic metrics [8], [9], [10], [11]. In addition to illustrating spatial and temporal connections between places, spatio-temporal maps can potentially indicate changes in pandemic risks [12].

Existing studies have examined the spread of epidemics as a complex system by assessing the degree of correlation or synchronization between time-series data. A deeper understanding of the transmission dynamics of the new variants of SARS-CoV-2 requires new methods beyond assessing correlation or synchronization. There is a need to explore the underlying local structures in the data and reveal the relationships between different countries to understand the spatio-temporal spread of the virus.

Manuscript received 7 September 2022; revised 16 March 2023; accepted 11 April 2023. Date of publication 21 April 2023; date of current version 6 June 2023. *(Corresponding author: Wei Ni.)*

Baoling Shan, Ren Ping Liu, and Eryk Dutkiewicz are with the Global Big Data Technologies Center, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: baoling.Shan@student.uts.edu.au; renping.liu@uts.edu.au; Eryk.Dutkiewicz@uts.edu.au).

Xin Yuan and Wei Ni are with the Data61, CSIRO, Marsfield, NSW 2122, Australia (e-mail: Xin.Yuan@data61.csiro.au; wei.ni@data61.csiro.au).

Xin Wang is with the Department of Communication Science and Engineering, Fudan University, Shanghai 200433, China (e-mail: xwang11@fudan.edu.cn).

Digital Object Identifier 10.1109/JBHI.2023.3267789

[1]https://coronavirus.jhu.edu/data

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This paper aims to uncover the hidden knowledge that underpins the evolution of the pandemic, examine the underlying relationship among countries, and understand the spreading pattern of SARS-CoV-2 variants, e.g., by taking Europe as an example. A new effective graph learning algorithm is proposed to estimate the graph Laplacian of the COVID-19 data, where we first obtain the closed-form expression for the eigenvectors of the graph Laplacian, also known as graph Fourier transform (GFT) basis. Given the eigenvectors, we transform the estimation of the graph Laplacian to a readily solvable, convex problem of estimating its eigenvalues. With the graph Laplacian estimated, we perform an in-depth spatio-temporal analysis of COVID-19 data and shed insights into the COVID-19 spread in Europe.

The main contributions of this paper are listed below.

1) We establish the closed-form expression for the eigenvectors of the graph Laplacian by revealing the intrinsic dependence between the frequency-domain representation of the data and the eigenvectors (that transform the original graph data to the frequency domain).
2) With the closed-form eigenvectors, we estimate the eigenvalues of the graph Laplacian efficiently using convex optimization techniques and then recover the graph Laplacian underlying the COVID-19 data.
3) We analyze the COVID-19 spread based on the inferred graph during different periods at a network level and at a node level. The most influential or representative European countries in COVID-19 spread are identified in different periods.

By applying the new algorithm, we analyze the evolving number of confirmed COVID-19 cases, reveal the spatio-temporal patterns of different SARS-CoV-2 variants among the European countries in different spread periods, and identify a different set of the most influential European countries from the existing techniques. These influential countries deserve attention and can potentially provide insights to help policymakers inform reliable strategies to manage the virus spread.

Extensive numerical tests are carried out to assess the graph learning accuracy of the new algorithm. Compared with the latest techniques, the new algorithm has the minimum root mean squared estimation error (RMSE) and the maximum correlation of determination (R^2) with at least 33.3% and 11.11% improvements, respectively, thereby corroborating the results of our COVID-19 analysis.

Following is an overview of the remainder of this paper. Section II summarizes the existing techniques. Section III presents the materials, system model and problem statement, respectively. In Section IV, the new closed-form expression is analytically established for the GFT basis, and accordingly, the new algorithm is developed to learn the graph Laplacian. In Section V, the proposed algorithm is experimentally validated in terms of accuracy and performed to analyze European COVID-19 data, with conclusions provided in Section VI.

Symbols and Notations: Matrices and vectors are represented by boldface uppercase and lowercase letters, respectively. $(\cdot)^T$ denotes transpose. $\|\cdot\|_F$ denotes the Frobenius norm. $\mathrm{eigen}[\cdot]$ yields the eigenvectors. Used notations are collated in Table I.

TABLE I
NOTATION AND DEFINITION

Notation	Definition
P	Number of days during a COVID-19 period studied
N	Number of countries considered
$\mathbf{X}$	COVID-19 data during a period studied
$\mathcal{G}$	The underlying graph of the COVID-19 data during a period
$\mathcal{V}$	Set of vertices corresponding to the considered countries in graph $\mathcal{G}$
$\mathcal{E}$	Set of edges in graph $\mathcal{G}$
$\mathbf{W}, \mathbf{L}$	Weighted adjacency and Laplacian matrices of graph $\mathcal{G}$
$\mathbf{U}, \mathbf{\Lambda}$	Orthonormal eigenvector and eigenvalue matrices of $\mathbf{L}$
$\mathbf{u}_i$	The i-th column of $\mathbf{U}$, i.e., the i-th eigenvector of $\mathbf{L}$
$\mathbf{S}$	Frequency-domain representation of the COVID-19 data during a period
$\mathcal{K}$	Index set for the K largest entries of $\{\|\mathbf{u}_i^T\mathbf{X}\|\}_{i=1}^N$
$\mathcal{K}^C$	Complementary set of $\mathcal{K}$
$\mathbf{U}_\mathcal{K}$ (or $\mathbf{U}_{\mathcal{K}^c}$)	Eigenvector matrix collating the columns of $\mathbf{U}$ indexed by $\mathcal{K}$ (or $\mathcal{K}^c$)

II. Related Work

Since the pandemic outbreak, researchers from various fields have extensively investigated the spread of the disease. The complex network theory based on a pair-wise configuration has been widely used to model the topological relationship of the COVID-19 data from a global perspective [4], [13], [14]. Azad et al. [13] conducted a social network analysis to trace the COVID-19 spread in India based on the travel history of infected patients and revealed that international travel played a key role in the pandemic outbreak in a country. Jo et al. [14] developed an infected network using the contact tracing information of confirmed cases, and found that governmental measures had a strong impact on the COVID-19 spread network in Seoul. Through modeling tourism mobility as a complex network, Tsiotas et al. [4] created a multidimensional framework to understand the COVID-19 spread across countries. Chu et al. [15] constructed an air travel network structure to visualize the connectedness and evolution of the pandemic. Travel subnetworks were formed by aggregating airport data at the national level and adding it to a matrix capturing the flight recurrences between countries. Using a similar conceptualization, they also developed a pandemic space approach [16] that uses the historical correlation of confirmed cases to locate the connections between different countries. By integrating Bayesian parameter inference with a Watts-Strogatz small-world network epidemiological model, Syga et al. [17] inferred a time-varying COVID-19 transmission network in Germany. It was shown that government interventions reduced random contacts and transmission probabilities.

A number of approaches have been developed to infer the pandemic's time-dependent transmission network, compared to previous works on network-based models. For instance, the correlation coefficients were exploited to capture the linear/nonlinear and symmetric pairwise matrix between different regions [11], [12], [18], [19]. So et al. [11] constructed dynamic pandemic networks over time for 164 countries to predict and assess the pandemic risk using network statistics. The connections in the networks were established based on the correlation

of changes in the number of confirmed cases between the two countries. Pan et al. [12] used the Pearson correlation coefficient, time-lagged cross-correlation, and dynamic time wrapping to examine interactions in the evolution of pandemics across the different states of the US. McMahon et al. [18] examined the spatial correlations of new active cases across different states in the US and assessed their magnitude over time. Their results showed stronger correlations between urban areas compared to rural areas, revealing that the pandemic spread was largely driven by travel between cities. Using spatio-temporal correlation, Aral et al. [19] identified distinct spatial clusters and spatial associations among COVID-19 cases in Turkey, revealing that spatial analysis helped explain the spread of the disease.

Alguliyev et al. [20] created a conceptual graph model by taking into account various epidemiological traits of COVID-19, such as social distance, the period of contact with an infected individual, and demographic characteristics based on location, thereby enabling a visual representation of virus propagation. This helps determine undetected cases of infection. Ieracitano et al. [21] adopted a deep learning technique based on fuzzy logic to create a classification system for the early detection of COVID-19 cases using portable chest X-ray (CXR) images. Absar et al. [22] developed a computer-assisted system for the automatic classification of COVID-19 CXR images using Support Vector Machine (SVM) to enable fast diagnosis of COVID-19.

Pearson correlation directly captures the pairwise relationship between two regions and inherently generates a transport network with a non-random topology, but ignores the causal relationship implied by underlying network structures [23]. Graph learning techniques have been adopted to effectively infer graphs from observed data [24], [25], [26], [27]. Dong et al. [24] explored the graph Laplacian using *a-priori* structural information to minimize variations of smooth signals. Kalofolias et al. [25] utilized primal-dual optimization to construct graph inference as learning the weighted adjacency matrix. Saboksayr et al. [26] further generalized the method of [25] to support more general multi-class smooth data observation. The graph Laplacian was learned using block-coordinate descent (BCD) in [27]. This algorithm decomposed the original Laplacian estimation problem into subproblems and then solved them alternately at each iteration.

In [28], Sardellitti et al. uncovered a block sparse representation of signals without assuming any diffusion process over the graph. They associated a graph with band-limited data observations and used alternative optimization (AO) to learn an orthonormal sparsifying transform. With obtained transform, the problem of graph estimation was converted into a convex one, and the graph Laplacian was recovered using convex optimization methods. Humbert et al. [29] considered band-limited graph data with properties of smoothness, as well as global sparse frequency representation. Like [28], AO, barrier methods, and manifold optimization were iterated to learn graph Laplacian approximately in [29].

III. Materials and System Model

The analysis is based on the open-access dataset of the daily counts of confirmed COVID-19 cases reported officially by different countries, territories, and regions, and published by the WHO.[2] The daily data on the COVID-19 pandemic for European countries are updated every day. We collect the data from January 2020 to April 2022, and divide this period into four based on the statistics from the WHO, as shown in Fig. 1. The first period is the early stage of the pandemic outbreak, between March 2020 and October 2020, when the original strain of the virus dominated the spread. The second period is from November 2020 to May 2021, when the Alpha variant was dominant. The third stage is from June 2021 to October 2021, when the Delta variant broke out. The fourth stage is from November 2021 to April 2022, when the Omicron variant rapidly replaced the Delta and became the dominating variant in most European countries.

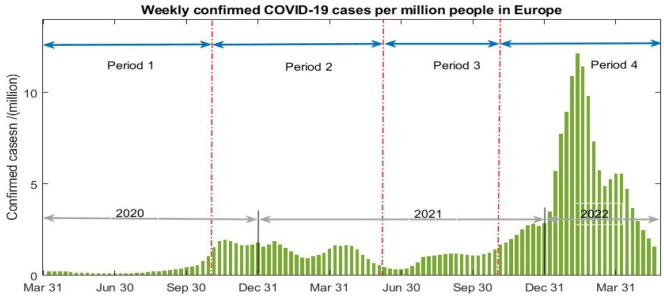


Fig. 1. Weekly confirmed COVID-19 cases per million people.

For each period, we analyze the SARS-CoV-2 time series data of the 44 European countries published by the WHO, by extracting a graph with 44 nodes from the data. The graph is denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathcal{V} = 1, 2, \ldots, N$ is the set of N vertices, with $N = 44$ being the number of countries. The set of edges, denoted by $\mathcal{E}$, is a subset of $\mathcal{V} \times \mathcal{V}$. The weighted adjacency matrix of the graph $\mathcal{G}$, denoted by $\mathbf{W} \in \mathbb{R}^{N \times N}$, indicates the extent to which two countries are correlated with respect to COVID-19 spread. $W_{ij} = W_{ji} \neq 0$ for $\forall (i, j) \in \mathcal{E}$. Each vertex in the graph is associated with a European country and corresponds to the time series recording daily confirmed cases per million people in the country.

For each of the periods, we use $\mathbf{x}_p \in \mathbb{R}^{N \times 1}, \forall p \in \{1, \ldots, P\}$ to denote the COVID-19 records of the N countries on the p-th day of the period, where P is the number of days in the period. The COVID-19 data of the European countries during the period are arranged in an $N \times P$ matrix, denoted by $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_P] \in \mathbb{R}^{N \times P}$. Here, $\mathbf{X}$ is band-limited and its frequency-domain representation has finite bandwidth; in other words, the virus spreads across countries, rather than breaks out simultaneously in all countries.

To derive information about the underlying topology of $\mathcal{G}$, we need to estimate the graph Laplacian $\mathbf{L}$. According to the definition in [27], graph Laplacian is a positive semi-definite matrix with positive diagonal elements and non-positive off-diagonal elements, which can be rewritten as:

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}\text{diag}(\boldsymbol{\lambda})\mathbf{U}^T. \quad (1)$$

Here, $\boldsymbol{\lambda} = \text{diag}(\lambda_1, \ldots, \lambda_N)$ is a diagonal matrix consisting of the non-negative eigenvalues of the Laplacian, and $\mathbf{U} =$

[2][Online]. Available: https://covid19.who.int/WHO-COVID-19-global-data.csv

$[\mathbf{u}_1, \ldots, \mathbf{u}_N]$ is an orthonormal matrix comprising the corresponding eigenvectors. According to [30], the GFT is the projection of $\mathbf{X}$ on $\mathbf{U}$, i.e., the subspace spanned by the eigenvectors of $\mathbf{L}$. The GFT of the COVID-19 data $\mathbf{x}_p, \forall p \in \{1, \ldots, P\}$ on p-th day, denoted by $\mathbf{s}_p$, is given by

$$\mathbf{s}_p = \mathbf{U}^T \mathbf{x}_p. \tag{2}$$

Let $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_p] \in \mathbb{R}^{N \times P}$. Then,

$$\mathbf{X} = \mathbf{U}\mathbf{S}. \tag{3}$$

With the sparsity of $\mathbf{s}_p, \forall p$, we define $\mathbf{S}$ as a K-block sparse matrix with rows consisting of multiple all-zero vectors. Here, K indicates the frequency-domain bandwidth of the COVID-19 data $\mathbf{X}$. K is obtained empirically in prior or enumerated to find its proper value [31]. The set $\mathcal{B}_K$ contains all K-block sparse matrices, defined as $\mathcal{B}_K \triangleq \{\mathbf{S} \in \mathbb{R}^{N \times P}, \mathbf{S}(i,:) = \mathbf{0}, \forall i \notin \mathcal{K} \subseteq \mathcal{V},\ K = |\mathcal{K}|\}$. Here, $\mathbf{S}(i,:)$ denotes the i-th row of $\mathbf{S}$, and the set $\mathcal{K} \subseteq \mathcal{V}$ collates the indexes to the K most significant frequency components of the $\mathbf{X}$.

IV. Proposed Graph Inference for COVID-19 Spread Analysis

COVID-19 data analysis plays an important role in identifying the most influential countries or regions in the spread of the virus and understanding how the virus spreads among countries. In this section, we propose a new graph learning technique, which accurately and efficiently extracts the underlying graph topological information of the COVID-19 data, reveals the fine-grained similarity (or correlation) between different countries in the virus spread process, and helps identify the most influential countries that present strong representativeness. More specifically, the technique extracts the graph Laplacian matrix $\mathbf{L}$ of the COVID-19 data in each period by first deriving the eigenvectors $\mathbf{U}$ of the matrix and then solving the eigenvalues $\mathbf{\Lambda}$ efficiently using convex optimization techniques. By applying the graph extracted and centrality measures, we identify the influential countries that can play a key role in the study of the COVID-19 spread.

A. Graph Topology Extraction

First, we estimate the graph Laplacian $\mathbf{L}$ and hence, the graph topology $\mathcal{G}$ substantiating the COVID-19 data $\mathbf{X}$. By taking into account the band-limitedness of $\mathbf{X}$, we formulate the problem as

$$\min_{\mathbf{L},\mathbf{U} \in \mathbb{R}^{N \times N}, \mathbf{S} \in \mathbb{R}^{N \times P}} \|\mathbf{X} - \mathbf{U}\mathbf{S}\|_F^2 + f(\mathbf{L}, \mathbf{X}) \tag{4a}$$

$$\text{s.t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_N, \tag{4b}$$

$$\mathbf{S} \in \mathcal{B}_K, \tag{4c}$$

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T,\ \mathbf{L} \in \mathbb{L},\ \text{tr}(\mathbf{L}) = N, \tag{4d}$$

$$\mathbf{u}_1 = \frac{1}{\sqrt{N}}\mathbf{1}. \tag{4e}$$

The objective (4a) is composed of two terms. The first accounts for data fidelity through a quadratic loss penalizing any discrepancy between $\mathbf{U}\mathbf{S}$ and $\mathbf{X}$. The second term provides a regularization function. According to [24] and [28], we set

$$f(\mathbf{L}, \mathbf{X}) = \text{tr}(\mathbf{X}^T\mathbf{L}\mathbf{X}) + \alpha \left\|\text{vec}(\mathbf{L})\right\|_1.$$

Constraint (4b) guarantees that the matrix $\mathbf{U}$ is unitary, satisfying the decomposition in (1); constraint (4c) enforces that the GFT coefficient matrix $\mathbf{S}$ is K-block sparse; constraint (4d) ensures that $\mathbf{L}$ complies with the requirement of a legitimate graph Laplacian, and $\mathbb{L}$ contains all legitimate candidates for $\mathbf{L}$[27]:

$$\mathbb{L} = \{\mathbf{L} \succeq \mathbf{0} | \mathbf{L}\mathbf{1} = \mathbf{0}, L_{ij} = L_{ji} \leq 0, \forall i \neq j\}. \tag{5}$$

According to $\mathbf{L}\mathbf{1} = \mathbf{0}$ in (5), we conclude that 0 is an eigenvalue of $\mathbf{L}$ and corresponds to the eigenvector $\mathbf{u}_1 = \frac{1}{\sqrt{N}}\mathbf{1}$, i.e., the first column of $\mathbf{U}$; see (4e).

Remark 1: To address the non-convexity of (4) caused by the non-convex orthonormality constraint in (4b) and the sparsity constraint in (4c), we decouple and solve (4) in two phases. Given the COVID-19 data $\mathbf{X}$, we first estimate the GFT basis $\mathbf{U}$ by minimizing $\|\mathbf{X} - \mathbf{U}\mathbf{S}\|_F^2$ subject to $\mathbf{U}^T\mathbf{U} = \mathbf{I}_N$, $\mathbf{S} \in \mathcal{B}_K$ and $\mathbf{u}_1 = \frac{1}{\sqrt{N}}\mathbf{1}$. In the second step, we estimate the eigenvalues $\mathbf{\Lambda}$ by minimizing the regularizer $\text{tr}(\mathbf{X}^T\mathbf{L}\mathbf{X}) + \alpha\|\text{vec}(\mathbf{L})\|_1$ with the obtained $\mathbf{U}$.

1) Extraction of Eigenvectors: Starting with the GFT basis, $\mathbf{U}$, provides a way to identify the intrinsic structure in the COVID-19 data that are related to the underlying pandemic network, even without the *a-priori* information of graph Laplacian $\mathbf{L}$. To estimate $\mathbf{U}$ from $\mathbf{X}$ complies with the definition of the GFT, i.e., $\mathbf{X} = \mathbf{U}\,\mathbf{S}$.

By utilizing the orthonormality property of $\mathbf{U}$ in (4b), we have $\|\mathbf{X} - \mathbf{U}\mathbf{S}\|_F^2 = \|\mathbf{U}^T\mathbf{X} - \mathbf{S}\|_F^2$. We start with the first part of problem (4), as given by [28, eq. 8]

$$\min_{\mathbf{U} \in \mathbb{R}^{N \times N}, \mathbf{S} \in \mathbb{R}^{N \times P}} \left\|\mathbf{U}^T\mathbf{X} - \mathbf{S}\right\|_F^2, \text{ s.t. } (4b),\ (4c),\ (4e). \tag{6}$$

Despite the convex objective function, problem (6) is non-convex due to the orthonormality in (4b) and the sparsity in (4c). Since both $\mathbf{U}$ and $\mathbf{S}$ are unknown, we reorganize (6) as

$$\min_{\mathbf{U} \in \mathbb{R}^{N \times N}} \min_{\mathbf{S} \in \mathcal{B}_K} \sum_{i=1}^{N} \left\|\mathbf{u}_i^T\mathbf{X} - \mathbf{S}(i,:)\right\|_2^2, \text{ s.t. } (4b),\ (4e), \tag{7}$$

which can be rewritten as

$$\min_{\mathbf{U} \in \mathbb{R}^{N \times N}} \left(\min_{\mathbf{S} \in \mathcal{B}_K} \sum_{i \in \mathcal{K}} \left\|\mathbf{u}_i^T\mathbf{X} - \mathbf{S}(i,:)\right\|_2^2 + \sum_{i \notin \mathcal{K}} \left\|\mathbf{u}_i^T\mathbf{X}\right\|_2^2 \right)$$
$$\text{s.t.} \quad (4b),\ (4e). \tag{8}$$

By closely analyzing the objective function of (8), it can be noticed that the optimal $\mathcal{K}$ comprises the indices of the K largest entries of $\{\|\mathbf{u}_i^T\mathbf{X}\|\}_i^N$, and satisfies

$$\mathbf{S}(i,:) = \begin{cases} \mathbf{u}_i^T\mathbf{X}, & \text{if } i \in \mathcal{K}; \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{9}$$

Therefore, the objective of (8) is reduced to only include the $(N - K)$ smallest entries of $\{||\mathbf{u}_i^T\mathbf{X}||\}_i^N$, after optimizing $\mathbf{S}$ to suppress $\sum_{i \in \mathcal{K}} \|\mathbf{u}_i^T\mathbf{X} - \mathbf{S}(i,:)\|_2^2$ using (9). To minimize this

objective with respect to $\mathbf{S}$, we aim to seek the optimal $\mathbf{U}$, represented as $\mathbf{U}^*$, in (6).

Substitute (9) into the objective of (8). Then, problem (6) can be written as

$$\begin{aligned}\mathbf{U}^* &= \arg\min_{\mathbf{U}} \sum_{i\notin\mathcal{K}} \left\|\mathbf{u}_i^T\mathbf{X}\right\|_2^2 = \arg\min_{\mathbf{U}} \left\|\mathbf{U}_{\mathcal{K}^c}^T\mathbf{X}\right\|_F^2 \\ &= \arg\max_{\mathbf{U}} \left\|\mathbf{U}_{\mathcal{K}}^T\mathbf{X}\right\|_F^2, \end{aligned} \tag{10}$$

where $\mathcal{K}^c$ denotes the complementary set of $\mathcal{K}$, i.e., $\mathcal{K}^c = \mathcal{V}\setminus\mathcal{K}$; and the matrices $\mathbf{U}_{\mathcal{K}}$ and $\mathbf{U}_{\mathcal{K}^c}$ collate the column-vectors of $\mathbf{U}$ with indexes collected in $\mathcal{K}$ and $\mathcal{K}^c$, respectively.

Despite the non-convexity of (10), the goal of (10) is to identify the K-dimensional subspace in which the COVID-19 data $\mathbf{X}$ has the largest orthogonal projection; i.e.,

$$\arg\max_{\mathbf{U}} \left\|\mathbf{U}_{\mathcal{K}}^T\mathbf{X}\right\|_F^2 = \arg\max_{\mathbf{U}} \ \mathrm{tr}\left(\mathbf{P}_{\mathbf{U}_{\mathcal{K}}}\mathbf{X}\mathbf{X}^T\right), \tag{11}$$

where $\mathbf{P}_{\mathbf{U}_{\mathcal{K}}} = \mathbf{U}_{\mathcal{K}}\mathbf{U}_{\mathcal{K}}^T$ is the orthogonal projector onto the subspace spanned by $\mathbf{U}_{\mathcal{K}}$.

Using (11), we reformulate the problem (6) as

$$\mathbf{U}^* = \arg\max_{\mathbf{U}} \mathrm{tr}\left(\mathbf{P}_{\mathbf{U}_{\mathcal{K}}}\mathbf{X}\mathbf{X}^T\right), \ \text{s.t.} \ (4e). \tag{12}$$

Theorem 1: By examining the two cases of $\mathbf{u}_1 \notin \mathbf{U}_{\mathcal{K}}$ and $\mathbf{u}_1 \in \mathbf{U}_{\mathcal{K}}$, the optimal solution to problem (6), denoted by $\mathbf{U}^* = [\mathbf{U}_{\mathcal{K}}^*, \mathbf{U}_{\mathcal{K}^c}^*]$, can be obtained as

$$\mathbf{U}^* = \mathrm{eigen}\left[\left(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T\right)\mathbf{X}\mathbf{X}^T\left(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T\right)^T\right]. \tag{13}$$

Proof: Please refer to Appendix A.

2) Extraction of Eigenvalues: Given the $\mathcal{K}$-band-limited COVID-19 data with the optimal $\mathbf{U}^*$ gained from (13), the graph Laplacian $\mathbf{L}$ is written as

$$\mathbf{L} = [\mathbf{U}_{\mathcal{K}}, \mathbf{U}_{\mathcal{K}^c}] \begin{bmatrix} \mathbf{\Lambda}_{\mathcal{K}} & \\ & \mathbf{\Lambda}_{\mathcal{K}^c} \end{bmatrix} [\mathbf{U}_{\mathcal{K}}, \mathbf{U}_{\mathcal{K}^c}]^T. \tag{14}$$

where $\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_{\mathcal{K}} & \\ & \mathbf{\Lambda}_{\mathcal{K}^c} \end{bmatrix}$. By plugging (14), $\mathrm{tr}(\mathbf{X}^T\mathbf{L}\mathbf{X})$ is written as

$$\begin{aligned}\mathrm{tr}(\mathbf{X}^T\mathbf{L}\mathbf{X}) &= \mathrm{tr}(\mathbf{X}^T(\mathbf{U}_{\mathcal{K}}\mathbf{\Lambda}_{\mathcal{K}}\mathbf{U}_{\mathcal{K}}^T)\mathbf{X} + \mathbf{X}^T(\mathbf{U}_{\mathcal{K}^c}\mathbf{\Lambda}_{\mathcal{K}^c}\mathbf{U}_{\mathcal{K}^c}^T)\mathbf{X}) \\ &= \mathrm{tr}(\mathbf{S}_{\mathcal{K}}^T\mathbf{\Lambda}_{\mathcal{K}}\mathbf{S}_{\mathcal{K}}). \end{aligned} \tag{15}$$

Problem (4) becomes

$$\begin{aligned}\min_{\mathbf{\Lambda}_{\mathcal{K}},\mathbf{\Lambda}_{\mathcal{K}^c},\mathbf{L}} \ & \mathrm{tr}(\mathbf{S}_{\mathcal{K}}^T\mathbf{\Lambda}_{\mathcal{K}}\mathbf{S}_{\mathcal{K}}) + \alpha\left\|\mathrm{vec}(\mathbf{L})\right\|_1 \\ \text{s.t.} \ & \mathbf{L} = [\mathbf{U}_{\mathcal{K}}, \mathbf{U}_{\mathcal{K}^c}] \begin{bmatrix} \mathbf{\Lambda}_{\mathcal{K}} & \\ & \mathbf{\Lambda}_{\mathcal{K}^c} \end{bmatrix} [\mathbf{U}_{\mathcal{K}}, \mathbf{U}_{\mathcal{K}^c}]^T, \\ & \mathbf{\Lambda}_{\mathcal{K}} \succeq \mathbf{0}, \mathbf{\Lambda}_{\mathcal{K}^c} \succeq \mathbf{0}, \\ & \mathbf{L}\mathbf{1} = \mathbf{0}, \\ & \mathrm{tr}(\mathbf{L}) = N, \\ & L_{ij} = L_{ji} \le 0, \forall i \ne j. \end{aligned} \tag{16}$$

Since its objective and constraints are convex or affine, problem (16) is convex and can be solved by CVX toolboxes. With $\mathbf{U}$ and $\mathbf{\Lambda}$ obtained, we can obtain the graph Laplacian $\mathbf{L}$ underlying the European COVID-19 data using (1).

B. Influential Country Identification

Next, given the graph topology $\mathbf{L}$ underlying the COVID-19 data and indicating the propagation of the virus, we proceed to estimate the spread pattern of the four variants among the European countries. As shown in Table II, three node-level metrics, including degree centrality [32], closeness centrality [33], and betweenness centrality [34], are used to measure the influence of individual countries in the COVID-19 spread, where d_{ij} represents the shortest distance between nodes i and j in the extracted graph, σ_{ij} is the total number of shortest paths between nodes i and j, and $\sigma_{ij}(v)$ denotes the number of these paths through node v.

- Degree centrality measures the number of connections a node has, helping identify the most connected nodes to the rest of the pandemic networks [32].
- Closeness centrality measures the inverse of the sum of the distances between a node and all other nodes in the network, which helps to identify nodes that are central and easily reachable within the network [33].
- Betweenness centrality measures the importance of a node in maintaining the shortest paths between other nodes in the network, helping to identify nodes that play a critical role in connecting different parts of the network [34].

The higher centrality a country has, the more influential it is and the more attention it deserves. In other words, the countries ranked high in terms of the centrality measures are likely to present the important COVID-19 spread patterns.

Many other existing methods, such as node embeddings [35], DeepWalk [36], spectral clustering [37], and influence maximization [38], aimed to efficiently find influential nodes in large-scale graphs, e.g., social networks with thousands or even millions of nodes, often still based on the above classical centrality measures. Nevertheless, the graph considered consists of only $N = 44$ vertices (for 44 European countries). Computational complexity is less of a concern.

We also take two network-level metrics in Table II, i.e., average path length [39] and global efficiency [40], to explore the spread of the pandemic.

- Average path length measures the average number of hops required to get from one node to another node in the network [39]. A short average path length indicates a highly connected network, contributing to the fast spread of the pandemic [41].
- Global efficiency measures the average inverse shortest path length between all pairs of nodes, indicating how quickly the virus can spread [40]. A high global efficiency indicates a dense and well-connected network with fast virus propagation, while a low global efficiency indicates a fragmented and poorly connected network deterring the virus propagation.

V. Method Assessment and Results

In this section, we first experimentally validate the superiority of the proposed technique to existing approaches in graph learning accuracy of the COVID-19 data. Then, we use the technique to conduct an in-depth analysis of COVID-19 data, and shed different insights into pandemic spread from existing

TABLE II
TOPOLOGICAL CHARACTERISTICS OF THE LEARNED COMPLEX NETWORKS

Metric	Formula	Description
Degree centrality	$C_d(n_i) = \frac{\sum_1^j e_{ij}}{N-1}$	The number of edges directed towards node i.
Closeness centrality	$C_c(n_i) = \frac{N-1}{\sum_{i \neq j} d_{ij}}$	The average length of the shortest paths from node i to the rest of the nodes.
Betweenness centrality	$C_b(n_i) = \frac{\sum_{i,j \neq v} \frac{\sigma_{ij}(v)}{\sigma_{ij}}}{(N-1)(N-2)}$	The number of times that a node serves as an intermediate relay along the shortest paths.
Average path length	$\frac{\sum_{i \neq j} d_{ij}}{N(N-1)}$	The average length of all the shortest paths in a graph.
Global efficiency	$\frac{N(N-1)}{\sum_{i \neq j} d_{ij}}$	The efficiency of information exchange between all node pairs.

techniques. The analysis is based on the open-access WHO dataset of the daily counts of confirmed COVID-19 cases in the 44 European countries.

Apart from the proposed algorithm, we evaluate the state-of-the-art solutions: Saboksayr's algorithm [26], Sardellitti's Total Variation (TV) algorithm [28], Sardellitti's Estimated-Signal-Aid (ESA) algorithm [28], and Humbert's algorithm [29].

- *Saboksyr' algorithm [25]:* This is a scalable and time-efficient primal-dual algorithm that learns the topological structures of time series represented by the weighted adjacency matrices of graphs. However, this method has no explicit generative model for the observations. In other words, the model's accuracy may not be adequate for numerous real-world datasets that exhibit localized behaviors or exhibit piecewise smoothness.
- *Sardellitti's TV graph learning algorithm [28]:* The approach involves a two-step scheme: (a) learning the orthonormal sparsifying transform from data using AO, and (b) recovering the Laplacian from the sparsifying transform using convex optimization. The algorithm is reasonably computationally efficient by exploiting convex optimization techniques. However, the effectiveness of the overall process is compromised due to the AO-based approximation in the first step, which penalizes the fidelity of the orthogonal sparsifying transform.
- *Sardellitti's ESA graph learning algorithm [28]:* Different from Sardellitti's TV graph learning algorithm, this algorithm exploits the knowledge of the GFT coefficient matrix of the first step in the second step, where the Laplacian matrix is recovered from the sparsifying transform and the GFT coefficients using convex optimization.
- *Humbert's algorithm [29]:* This is another AO-based algorithm with alternating procedures relying on standard minimization methods, i.e., manifold gradient descent and linear programming. However, only suboptimal solutions can be obtained using the AO method. The computational complexity of the method is also high.

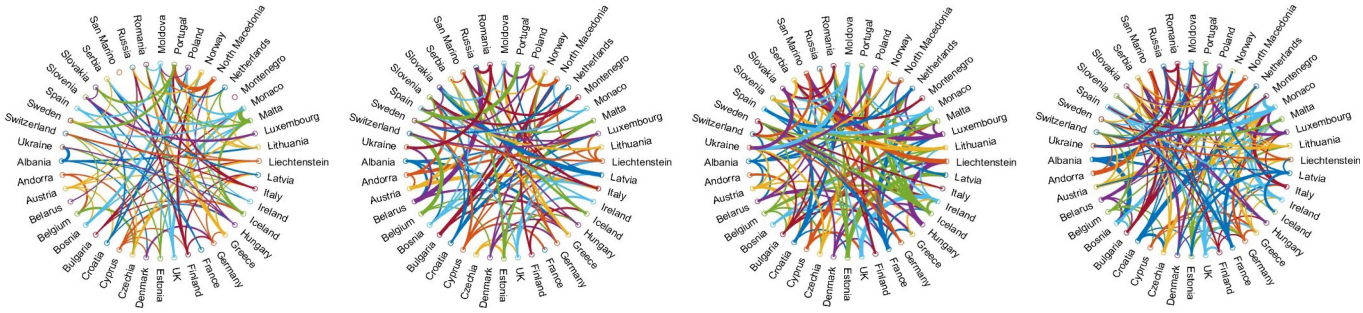
In addition to the above state-of-the-art graph learning techniques, we also compare our proposed algorithm with the state-of-the-art graph neural network (GNN) [42] when assessing the accuracy of the algorithm. The GNN consists of multiple hidden layers with 50 hidden units per layer. In the training stage, the input to the GNN includes the training data and the weighted correlation matrix of the training set. By contrast, the input of the graph learning algorithms is the training set.

A. Graph Learning-Based Analysis of COVID-19 Data

Fig. 2 provides the pandemic spread networks of 44 European countries over the four different periods obtained by the proposed algorithm, where the parameters of the algorithm are $K = 26$ and $\alpha = 1$ decided in the way delineated at the beginning of Section V-B. The thickness of an edge measures the similarity of the COVID-19 spread between two countries. The virus spreads in the two countries are more likely to be related if the edge is thicker. The density of the edges indicates the extent to which the COVID-19 spread among countries. It is observed in Fig. 2 that the virus spreads are increasingly related among the European countries from Period 1 to Period 4. Not only did the spreads increase between the countries, but the virus spread increasingly widely across more countries.

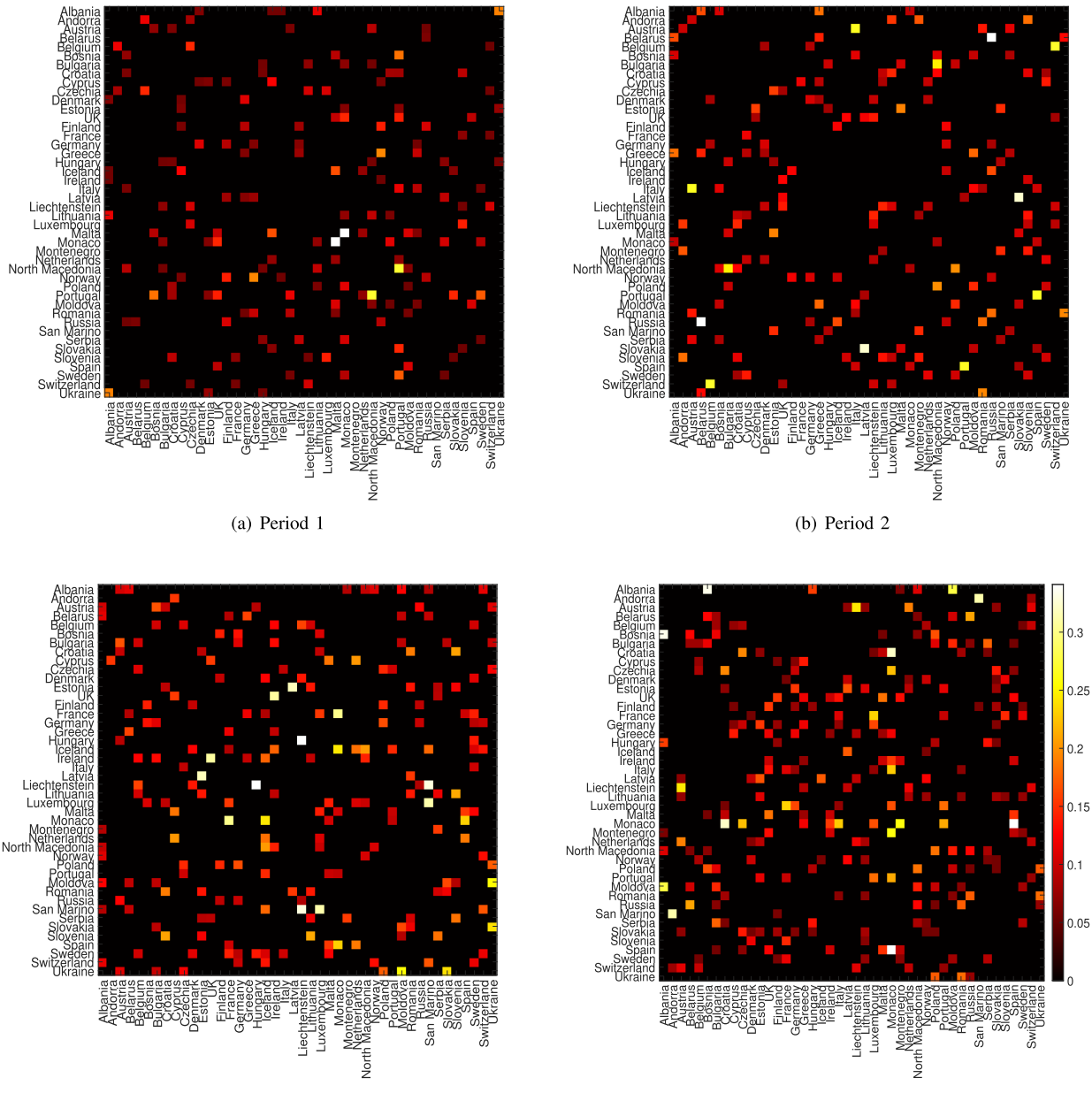
To better illustrate the correlation of the COVID-19 spread between the European countries, Fig. 3 plots the weighted adjacency matrices of the graphs extracted from the COVID-19 data by the proposed algorithm. In the figure, the 44 European countries are sorted alphabetically from Albania to Ukraine along the x- and y-axes. The intensity of the color at each pixel stands for the extent of the correlation between the two countries associated with the pixel. For example, the pixel corresponding to Greece and Norway is lighter than others in Fig. 3(a), indicating that Greece and Norway are highly correlated in Period 1. Likewise, Russia and Belarus are highly correlated in Period 2 in Fig. 3(b). Nevertheless, the number of light-colored pixels increases overall in both Periods 3 and 4 in Figs. 3(c) and 3(d), indicating that the Delta and Omicron variants have higher and stronger propagation characteristics in Europe, which is consistent with the finding made in Fig. 2.

Figs. 4–7 visualize the top 5 countries that are identified to have been the most influential in the process of the COVID-19 virus spread in Europe, using the proposed approach based on the aforementioned three node-level metrics, i.e., degree centrality, closeness centrality, and betweenness centrality. A darker color indicates a country identified by more centrality measures to be among the top 5 most influential countries. For example, Czechia was influential during Period 1 in the sense of all



(a) Period 1 (b) Period 2 (c) Period 3 (d) Period 4

Fig. 2. The learned graph of the COVID-19 spread in the 44 European countries during different periods.



(a) Period 1 (b) Period 2

(c) Period 3 (d) Period 4

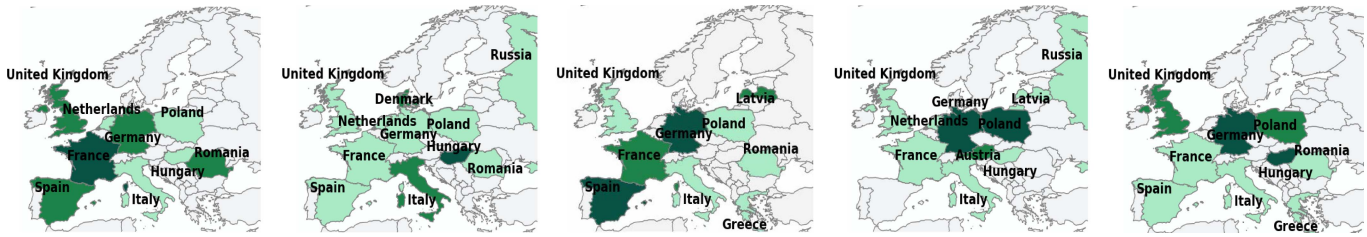
Fig. 3. The weighted matrix of the learned graph of the spread of COVID-19 in 44 European countries during different periods.

(a) Proposed (b) Saboksayr (c) Sar-TV (d) Sar-ESA (e) Humbert

Fig. 4. Influential countries identified during Period 1.

(a) Proposed (b) Saboksayr (c) Sar-TV (d) Sar-ESA (e) Humbert

Fig. 5. Influential countries identified during Period 2.



(a) Proposed (b) Saboksayr (c) Sar-TV (d) Sar-ESA (e) Humbert

Fig. 6. Influential countries identified during Period 3.

(a) Proposed (b) Saboksayr (c) Sar-TV (d) Sar-ESA (e) Humbert

Fig. 7. Influential countries identified during Period 4.

three centrality measures. This makes sense since the different centrality measures are closely related in nature [12].

It is obvious in Figs. 4–7 that the proposed algorithm identifies a different set of the most influential European countries in the COVID-19 spread, compared to the state-of-the-art graph learning methods. Particularly, the proposed algorithm helps identify a small and concentrated set of influential countries in every period of COVID-19 spread; i.e., a country is more likely to be associated with multiple centrality measures. In other words, the influence of a country is more likely to be manifested through multiple measures. Here, the parameters of each method are separately tested and optimized, according to their individual settings.

Fig. 8 quantitatively evaluates how different the top 5 most influential countries are identified by the different algorithms. Specifically, we vectorize the 15 most important countries identified using each of the considered algorithms based on the three centrality measures. The similarity between the 15-element

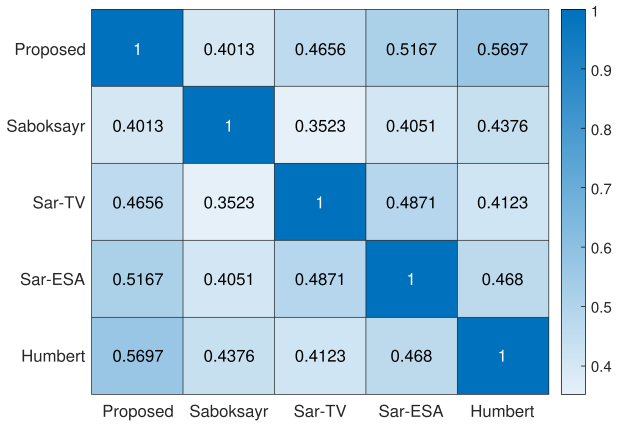


Fig. 8. The correlation of different algorithms.

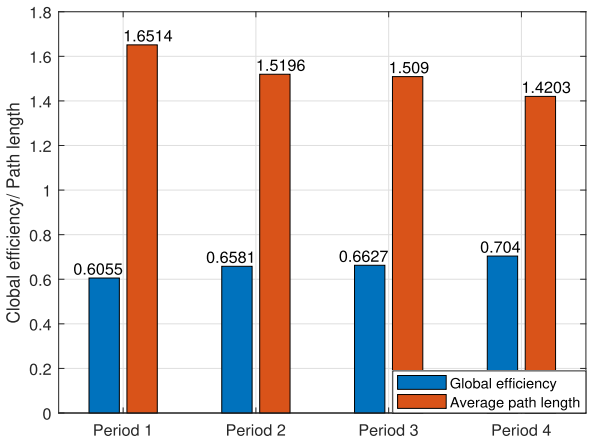


Fig. 9. Average path length and global efficiency corresponding to different periods of COVID-19.

vectors produced by any two of the considered graph learning algorithms, measured by the cosine distance $\frac{\mathbf{V}_1^T\mathbf{V}_2}{|\mathbf{V}_1|\cdot|\mathbf{V}_2|}$, quantifies the similarity between the algorithms, where $\mathbf{V}_1$ and $\mathbf{V}_2$ are the two 15-element column vectors, and $|\cdot|$ stands for the norm.

As shown in Fig. 8, the proposed algorithm yields the highest similarity to Humbert's [29] in terms of their identified important countries (under three centrality measures), followed by Sar-ESA [28], Sar-TV [28], and Saboksayr's [26]. The similarities of the proposed algorithm to the existing algorithms are consistent with the graph learning (and reconstruction) accuracy of the algorithms, as will be shown in Fig. 12. Note that the ground truth regarding the most important countries is unavailable in practice. Given the best graph learning accuracy of the proposed algorithm and the consistent rankings between the accuracies and the similarities of the existing algorithms, it is reasonable to conclude that the countries identified by the proposed algorithm are more accurate and can contribute to more effective study and response to the pandemic.

Fig. 9 plots the average path length and global efficiency of the graph recovered by the proposed graph learning algorithm in the four periods of the COVID-19 pandemic. It is observed that the average path length decreases and the global efficiency increases in the four periods. The Omicron variant (i.e., Period 4) corresponds to the smallest average path length and the largest global efficiency, indicating that the Omicron variant has a higher level of global reachability and infectivity. In contrast, the original strains in the early stage of the pandemic, i.e., Period 1, have higher average path lengths and smaller global efficiencies. This is consistent with the finding in Figs. 2 and 3. The reason can be that during Period 1, the countries responded to the outbreak with stay-at-home or workplace closure, effectively slowing down the increase in confirmed cases.

B. Accuracy Validation of Proposed Graph Learning

Without the ground truth of the graphs underlying the COVID-19 data, we resort to assessing the learning accuracy of the proposed algorithm by obfuscating part of the data and assessing the reconstruction accuracy of this part of data based on the learned graphs and the rest of the data.

Suppose that the number of observable countries is K ($K \leq N$), i.e., the signal bandwidth. Based on the inferred graphs, e.g., those in Fig. 3, and the observed COVID-19 data of K randomly selected European countries, we reconstruct the number of confirmed cases per million population in the remaining $(N-K)$ countries. The recovered graph signals, denoted by $\hat{\mathbf{x}}_p$, can be obtained as [43]

$$\hat{\mathbf{x}}_P = \mathbf{U}_{\mathcal{K}}\mathbf{U}_{\mathcal{K}}^T\mathbf{\Psi}^T\mathbf{\Psi}\mathbf{D}^2\mathbf{\Psi}^T\mathbf{y}_P, \tag{17}$$

where $\mathbf{y}_p \in \mathbb{R}^K$ is sampled $K \times 1$-dimensional COVID-19 data on the p-th day, which is chosen from $\mathbf{x}_p$ randomly and independently [43]. $\mathbf{\Psi} \in \mathbb{R}^{K\times N}$ stands for a sampling operator. $\mathbf{\Psi}_{ij} = 1$ if $j = \mathcal{K}_i$; and 0, otherwise. Here, $\mathcal{K}_i$ is the i-th element of $\mathcal{K}$, indicating the i-th of the $K = |\mathcal{K}|$ European countries with COVID-19 data available. $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a diagonal rescaling matrix with $D_{ii} = 1/\sqrt{K\pi_i}$ and π_i being the probability of choosing the i-th $K \times N$-dimensional sample of the K countries on the p-th day of the considered period. Since a uniform sampling process is considered, the sampling score for each node is $\pi_i = 1/N$.

The RMSE and the R^2 are adopted to quantify the accuracy of the recovered data with respect to the ground-truth COVID-19 data, as given by

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N}(\hat{x}_{pi} - x_{pi})^2/N}; \tag{18}$$

$$R^2 = 1 - \frac{\|\hat{\mathbf{x}_p} - \mathbf{x}_p\|_2^2}{\|\hat{\mathbf{x}_p} - \bar{\mathbf{x}_p}\|_2^2}. \tag{19}$$

Here, $\hat{\mathbf{x}}_p$ and $\bar{\mathbf{x}}_p$ are the reconstructed signals and the average of the ground-truths of $\mathbf{x}_p$.

Fig. 10 plots the correlations of determination, i.e., R^2, of the proposed algorithm with different regularizer α and data bandwidth K under the pandemic network during Period 1. We see that R^2 reaches its peak at $\alpha = 1$ and $K = 26$; indicating that the optimal regularizer is $\alpha = 1$ for a data bandwidth of $K = 26$. We can similarly determine the optimal values of α for Periods 2 to 4. Fig. 11 shows the R^2 of the considered graph learning algorithms in four different periods, where $K = 26$. The proposed algorithm obtains the largest

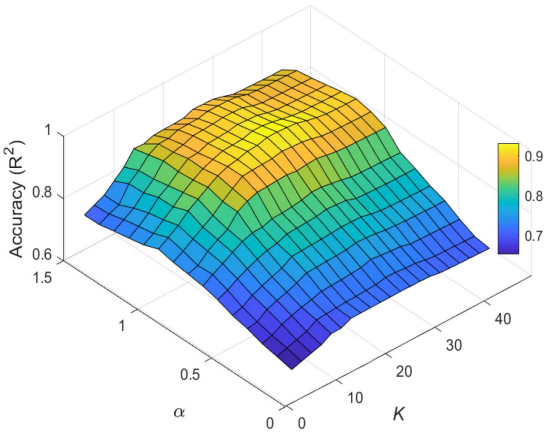


Fig. 10. The accuracy vs. bandwidth K and α in Period 1.

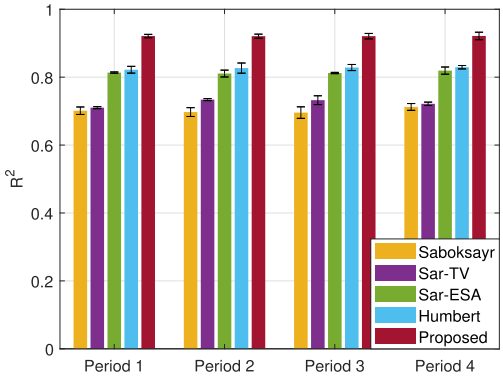


Fig. 11. Efficiency of reconstruction of different methods upon four periods when $K = 26$.

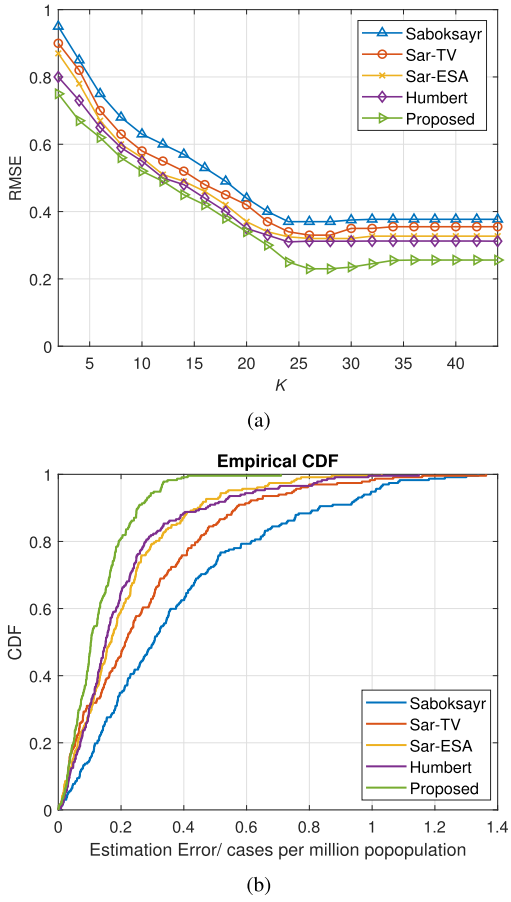


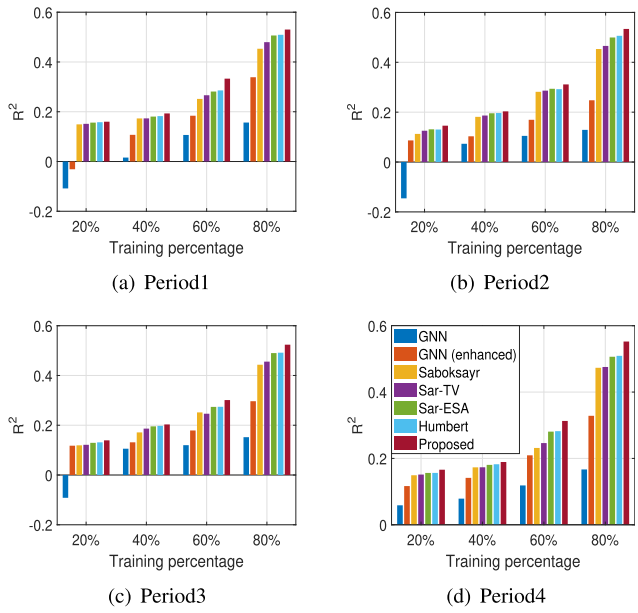
Fig. 12. (a) The RMSE vs. the bandwidth K. (b) The CDFs of estimation error under different graph learning methods.

R^2. For example, the improvements of the algorithm are about 29.36%, 27.71%, 12.46%, and 11.11%, compared to Saboksayr's [26], Sar-TV [28], Sar-ESA [28], and Humbert's [29], respectively. To ensure a fair comparison, the parameters are individually tested and optimized for each benchmark in these figures.

Fig. 12(a) shows the RMSE of the considered algorithms with the increase in the signal bandwidth K. We see that under all the considered algorithms, the RMSEs decrease quickly with the growth of K and then converge to constant values. Our proposed algorithm has the smallest RMSE under all values of K. It has the minimum RMSE around 0.23 at $K = 26$ and achieves performance improvements by about 60.87%, 43.48%, 34.78%, and 33.33%, compared to Saboksayr's [26], Sar-TV [28], Sar-ESA [28], and Humbert's [29], respectively. Fig. 12(b) plots the cumulative distribution function (CDF) of the errors undergone by the considered algorithms. As shown, the proposed algorithm has much lower estimation errors than the rest of the algorithms. In particular, over 80% of the estimation errors are smaller than 0.2 case per million population under our algorithm. By contrast, 38.3%, 48.4%, 59.5%, and 64.6% of the estimation errors are smaller than 0.2 case per million population under Saboksayr's [26], Sar-TV [28], Sar-ESA [28], and Humbert's [29], respectively.

Next, we proceed to assess the accuracy (R^2) of the considered graph learning algorithms when predicting future missing data based on the graph topologies extracted in the past. In addition to the graph learning techniques, we also consider the state-of-the-art GNN [42]. The COVID-19 dataset of each period is divided into a training set (e.g., the first 80% of the dataset) and a test set (e.g., the remaining 20% of the dataset). In the training phase, the graph learning algorithms extract the graph topology of the training set. In the test phase, the test data of Ukraine is assumed to be missing and is predicted based on the graph topologies extracted from the training set and the available test data of the other countries. By adjusting the ratio between the training and test sets, we show the robustness of the algorithms to the small training set.

As shown in Fig. 13(a)–(d), the graph learning methods, including our proposed algorithm, outperform the GNN under different ratios between the training and test sets. When the training set is set to 80% and the testing set is 20%, our algorithm achieves the highest R^2 values with the improvements of about 70.49%, 75.85%, 70.99%, and 68.11% in the four periods, compared to the GNN. Notice that the R^2 value of the GNN can yield negative values, especially when the training set is small. This is the case when even the mean of the data can provide a better fit to the data than the fitted function, e.g., the GNN, when the training set is small, i.e., 20%.



(a) Period1 (b) Period2 (c) Period3 (d) Period4

Fig. 13. Efficiency of reconstruction of different methods of different periods.

On the other hand, the proposed graph learning algorithm can enhance the state-of-the-art GNN by providing more accurate graph topologies, compared to a direct calculation of adjacency matrices (as done in the GNN [42]). By inputting the weighted adjacency matrices of the graphs learned by the algorithm,the GNN can be enhanced and consistently outperform the original GNN in the experiments. Nevertheless, the enhanced GNN is still not as good as the state-of-the-art graph learning techniques, primarily due to the relatively small size of the training set, i.e., the COVID-19 data set.

VI. Conclusion

In this article, we proposed a new graph-learning technique to analyze the evolution of the COVID-19 pandemic and reveal the underlying relationship and spreading pattern among different countries. The new technique estimates the graph Laplacian of the COVID-19 data by first deriving the closed-form expression for its eigenvectors and then estimating its eigenvalues with convex optimization. Based on the COVID-19 data, the accuracy of the estimated graph Laplacian was shown to outperform the existing approaches by 33.3% in RMSE and 11.11% in correlation of determination. The new technique helped identify a different set of the most influential and representative European countries, compared to the existing techniques. Given the superior accuracy of the technique, the set of identified influential countries is expected to be sensible and deserves dedicated research efforts to help understand the COVID-19 spread.

Appendix
Proof of Theorem 1

Proof: The solution for (12) is derived in two cases:

1) In the Case of $\mathbf{u}_1 \notin \mathbf{U}_\mathcal{K}$: Let $\mathbf{P}_{\mathbf{U}_{\mathcal{K}\setminus\{1\}}}$ refer to the orthogonal projection of the subspace consisting of the column vectors of $\mathbf{U}_\mathcal{K}$, except for the first column, i.e., $\mathbf{P}_{\mathbf{U}_{\mathcal{K}\setminus\{1\}}} = \mathbf{P}_{\mathbf{U}_\mathcal{K}}(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T$. Then, the objective function of (12) is converted as

$$\max_{\mathbf{U},\mathcal{K}} \operatorname{tr}\left(\mathbf{P}_{\mathbf{U}_\mathcal{K}}\left(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T\right)\mathbf{X}\mathbf{X}^T\left(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T\right)^T\right). \tag{20}$$

Problem (20) can be transformed into a problem defined on a Grassmann manifold, making it an unconstrained optimization problem. Since the Grassmann manifold is a closed set, the maximum or minimum of a continuous function defined on this set, such as the optimal solution to (20), exists [44]. Having established the existence of the optimal solution, we can prove that the K-dimensional subspace of $\mathbf{U}_\mathcal{K}$ is composed of eigenvectors that correspond to the K largest eigenvalues of $(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T$.

Suppose that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N$ are the eigenvalues of $(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T$, corresponding to the eigenvectors $\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_N\}$. Let $S_1=\operatorname{span}\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_K\}$ correspond to the K largest eigenvalues $\sigma_1,\sigma_2,\ldots,\sigma_K$. Let S_2 $(S_2\neq S_1)$ be another K-dimensional subspace, and $E_0=S_1\cap S_2$. Suppose that $S_1=E_0\oplus E_1$ and $S_2=E_0\oplus E_2$, where E_1 is the subset of S_1 and E_2 is the subset of $S_1^\perp$; i.e., $E_1\subset \operatorname{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_K\}$ and $E_2\subset\operatorname{span}\{\mathbf{v}_{K+1},\ldots,\mathbf{v}_N\}$, and $S_1^\perp$ is the orthogonal complement space of S_1. Let $\dim(E_1)=\dim(E_2)=t$. According to the Minimax theorem [45], we have $\operatorname{tr}(\mathbf{P}_{E_1}(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T)\geq t\sigma_K$ and $\operatorname{tr}(\mathbf{P}_{E_2}(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T)\leq t\sigma_{K+1}$. As a result, $\operatorname{tr}(\mathbf{P}_{S_1}(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T)\geq \operatorname{tr}\left(\mathbf{P}_{S_2}(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T\right)$ based on $\sigma_K\geq\sigma_{K+1}$, $S_1=E_0\oplus E_1$ and $S_2=E_0\oplus E_2$. Here, $\mathbf{P}_{S_1}$ is the projection of the subspace S_1 spanned by the K largest eigenvalues of $(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T$. In other words, the projection of $(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T$ is the largest on the span of the eigenvectors corresponding to the K largest eigenvalues. Thus, the solution to (20) is given by

$$\mathbf{U}_\mathcal{K}^* = \operatorname{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_K\}, \tag{21}$$

which corresponds to the K largest eigenvalues of $(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T$. Accordingly, $\mathbf{U}_{\mathcal{K}^c}^*$ consists of the rest of the eigenvectors, leading to (13).

2) In the Case of $\mathbf{u}_1\in\mathbf{U}_\mathcal{K}$: By writing $\mathbf{U}_\mathcal{K}=[\mathbf{u}_1,\mathbf{U}_{\mathcal{K}\setminus\{1\}}]$, the objective of (12) turns to $\arg\max_{\mathbf{U},\mathcal{K}} \operatorname{tr}(\mathbf{P}_{\mathbf{U}_\mathcal{K}}(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T)$, which is identical to (20). Thus, the solution, $\mathbf{U}_\mathcal{K}^*$, comprises $\mathbf{u}_1$ and the eigenvectors associated with the $(K-1)$ largest eigenvalues of $(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{u}_1\mathbf{u}_1^T)^T$, as can be proved in the same way as done above. Accordingly, $\mathbf{U}_{\mathcal{K}^c}^*$ consists of the rest of the eigenvectors, leading to (13).

References

[1] World Health Organization, "Coronavirus disease (COVID-19) pandemic," Nov. 2020. Accessed: Sep. 2022. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] *World Health Org.*, "Coronavirus disease (COVID-19): Variants of SARS-CoV-2," Nov. 2020. Accessed: Sep. 2022. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease

[3] G. Guo, Z. Liu, S. Zhao, L. Guo, and T. Liu, "Eliminating indefiniteness of clinical spectrum for better screening COVID-19," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1347–1357, May 2021.

[4] D. Tsiotas and V. Tselios, "Understanding the uneven spread of COVID-19 in the context of the global interconnected economy," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022.

[5] Y. Han et al., "Spatial distribution characteristics of the COVID-19 pandemic in Beijing and its relationship with environmental factors," *Sci. Total Environ.*, vol. 761, 2021, Art. no. 144257.

[6] Z. Feng, C. Xiao, P. Li, Z. You, X. Yin, and F. Zheng, "Comparison of spatio-temporal transmission characteristics of COVID-19 and its mitigation strategies in China and the US," *J. Geographical Sci.*, vol. 30, no. 12, pp. 1963–1984, 2020.

[7] D. C. d. S. Gomes and G. L. d. O. Serra, "Machine learning model for computational tracking and forecasting the COVID-19 dynamic propagation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 3, pp. 615–622, Mar. 2021.

[8] A. Kuzdeuov et al., "A network-based stochastic epidemic simulator: Controlling COVID-19 with region-specific policies," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2743–2754, Oct. 2020.

[9] M. Milano, "Cctv: A new network-based methodology for the analysis and visualization of COVID-19 data," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2021, pp. 2000–2001.

[10] K. Demertzis, D. Tsiotas, and L. Magafas, "Modeling and forecasting the COVID-19 temporal spread in Greece: An exploratory approach based on complex network defined splines," *Int. J. Environ. Res. Public Health*, vol. 17, no. 13, 2020, Art. no. 4693.

[11] M. K. So, A. M. Chu, A. Tiwari, and J. N. Chan, "On topological properties of COVID-19: Predicting and assessing pandemic risk with network statistics," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, 2021.

[12] Y. Pan, L. Zhang, J. Unwin, and M. J. Skibniewski, "Discovering spatial-temporal patterns via complex networks in investigating COVID-19 pandemic in the United States," *Sustain. Cities Soc.*, vol. 77, 2022, Art. no. 103508.

[13] S. Azad and S. Devi, "Tracking the spread of COVID-19 in India via social networks in the early phase of the pandemic," *J. Travel Med.*, vol. 27, no. 8, 2020, Art. no. taaa130.

[14] W. Jo, D. Chang, M. You, and G.-H. Ghim, "A social network analysis of the spread of COVID-19 in South Korea and policy implications," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, 2021.

[15] A. M. Chu, J. N. Chan, J. T. Tsang, A. Tiwari, and M. K. So, "Analyzing cross-country pandemic connectedness during COVID-19 using a spatial-temporal database: Network analysis," *JMIR Public Health Surveill.*, vol. 7, no. 3, 2021, Art. no. e27317.

[16] A. M. Chu, T. W. Chan, M. K. So, and W.-K. Wong, "Dynamic network analysis of COVID-19 with a latent pandemic space model," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, 2021, Art. no. 3195.

[17] S. Syga, D. David-Rus, Y. Schälte, H. Hatzikirou, and A. Deutsch, "Inferring the effect of interventions on COVID-19 transmission networks," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021.

[18] T. McMahon, A. Chan, S. Havlin, and L. K. Gallos, "Spatial correlations in geographical spreading of COVID-19 in the United States," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, 2022.

[19] A. Neşe and H. Bakir, "Spatiotemporal analysis of COVID-19 in Turkey," *Sustain. Cities Soc.*, vol. 76, 2022, Art. no. 103421.

[20] R. Alguliyev, R. Aliguliyev, and F. Yusifov, "Graph modelling for tracking the COVID-19 pandemic spread," *Infect. Dis. Model.*, vol. 6, pp. 112–122, 2021.

[21] C. Ieracitano et al., "A fuzzy-enhanced deep learning approach for early detection of COVID-19 pneumonia from portable chest X-ray images," *Neurocomputing*, vol. 481, pp. 202–215, 2022.

[22] N. Absar et al., "Development of a computer-aided tool for detection of COVID-19 pneumonia from cxr images using machine learning algorithm," *J. Radiat. Res. Appl. Sci.*, vol. 15, no. 1, pp. 32–43, 2022.

[23] J. Dai, K. Huang, Y. Liu, C. Yang, and Z. Wang, "Global reconstruction of complex network topology via structured compressive sensing," *IEEE Syst. J.*, vol. 15, no. 2, pp. 1959–1969, Jun. 2021.

[24] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, Dec. 2016.

[25] V. Kalofolias, "How to learn a graph from smooth signals," in *Proc. 19th Int. Conf. Artif. Intell. Statist., Cadiz, Spain,*,, 2016, pp. 920–929.

[26] S. S. Saboksayr et al., "Online discriminative graph learning from multi-class smooth signals," *Signal Process.*, vol. 186, 2021, Art. no. 108101.

[27] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 825–841, Sep. 2017.

[28] S. Sardellitti, S. Barbarossa, and P. D. Lorenzo, "Graph topology inference based on sparsifying transform learning," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1712–1727, Apr. 2019.

[29] P. Humbert et al., "Learning Laplacian matrix from graph signals with sparse spectral representation," *J. Mach. Learn. Res.*, vol. 22, no. 195, pp. 1–47, 2021.

[30] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, Dec. 2015.

[31] Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[32] X. Tang, J. Wang, J. Zhong, and Y. Pan, "Predicting essential proteins based on weighted degree centrality," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 2, pp. 407–418, Mar.-Apr. 2014.

[33] P. Crescenzi, G. D'angelo, L. Severini, and Y. Velaj, "Greedily improving our own closeness centrality in a network," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 1, pp. 1–32, 2016.

[34] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.

[35] J. Zhou, L. Liu, W. Wei, and J. Fan, "Network representation learning: From preprocessing, feature extraction to node embedding," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–35, 2022.

[36] K. Berahmand, E. Nasiri, M. Rostami, and S. Forouzandeh, "A modified deepwalk method for link prediction in attributed social network," *Computing*, vol. 103, pp. 2227–2249, 2021.

[37] H. V. Lierde, T. W. Chow, and G. Chen, "Scalable spectral clustering for overlapping community detection in large-scale networks," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 754–767, Apr. 2020.

[38] Y. Gong, S. Liu, and Y. Bai, "Efficient parallel computing on the game theory-aware robust influence maximization problem," *Knowl. Based Syst.*, vol. 220, 2021, Art. no. 106942.

[39] C.-C. Yen, M.-Y. Yeh, and M.-S. Chen, "An efficient approach to updating closeness centrality and average path length in dynamic networks," in *Proc. IEEE 13th Int. Conf. Data Mining*., 2013, pp. 867–876.

[40] H. Shirouyehzad, J. Jouzdani, and M. K. Karimvand, "Fight against COVID-19: A global efficiency evaluation based on contagion control and medical treatment," *J. Appl. Res. Ind. Eng.*, vol. 7, no. 2, pp. 109–120, 2020.

[41] Y. Deng, Y. Zhang, and K. Wang, "An analysis of the Chinese scheduled freighter network during the first year of the COVID-19 pandemic," *J. Transp. Geogr.*, vol. 99, 2022, Art. no. 103298.

[42] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," in *Proc. Int. Conf. Learn. Representations*, New Orleans, LA, USA, 2019.

[43] S. Chen, R. Varma, A. Singh, and J. Kovačević, "Signal recovery on graphs: Fundamental limits of sampling strategies," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 4, pp. 539–554, Dec. 2016.

[44] S. Mittal et al., "Conjugate gradient on Grassmann manifolds for robust subspace estimation," *Image Vis. Comput.*, vol. 30, no. 6-7, pp. 417–427, 2012.

[45] M. Sion et al., "On general minimax theorems," *Pacific J. Math.*, vol. 8, no. 1, pp. 171–176, 1958.