Understanding and predicting the spatio-temporal spread of COVID-19 via integrating diffusive graph embedding and compartmental models

Tong Zhang\(^1\) | Jing Li\(^2\)

\(^1\)State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
\(^2\)Department of Geography and the Environment, University of Denver, Denver, CO, USA

Correspondence
Tong Zhang, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China.
Email: zhangt@whu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 41871308; National Key R&D Program of China (International Scientific & Technological Cooperation Program), Grant/Award Number: 2019YFE0106500

Abstract
In order to find useful intervention strategies for the novel coronavirus (COVID-19), it is vital to understand how the disease spreads. In this study, we address the modeling of COVID-19 spread across space and time, which facilitates understanding of the pandemic. We propose a hybrid data-driven learning approach to capture the mobility-related spreading mechanism of infectious diseases, utilizing multi-sourced mobility and attributed data. This study develops a visual analytic approach that identifies and depicts the strength of the transmission pathways of COVID-19 between areal units by integrating data-driven deep learning and compartmental epidemic models, thereby engaging stakeholders (e.g., public health officials, managers from transportation agencies) to make informed intervention decisions and enable public messaging. A case study in the state of Colorado, USA was performed to demonstrate the applicability of the proposed transmission modeling approach in understanding the spatio-temporal spread of COVID-19 at the neighborhood level. Transmission path maps are presented and analyzed, demonstrating their utility in evaluating the effects of mitigation measures. In addition, integrated embeddings also support daily prediction of infected cases and role analysis of each area unit during the transmission of the virus.
This study is concerned with the modeling of spatio-temporal spreading patterns of the novel coronavirus disease (COVID-19), which is highly infectious and has been a global pandemic since March 2020. It has spread to almost every country in the world and killed more than 3,325,000 people as of late August 2020 (WHO, 2021). Compared with historical pandemics, such as the 1918 influenza pandemic, COVID-19 spreads at a much more rapid rate at the global scale, due to higher population density, much more advanced transportation infrastructure than ever before, as well as high human-to-human transmissibility, limited supply of vaccines, and the lack of consensus on cost-effective intervention measures worldwide (Hao et al., 2020; Haug et al., 2020; Kim, Marks, & Clemens, 2021).

The spatio-temporal modeling of COVID-19 spread has two research implications: (a) it may facilitate our understanding of the transmission patterns of the disease and how the spread relates to human mobility and local demographic/environmental factors; and (b) it can be used to forecast the most probable transmission paths in the near future and to evaluate the effectiveness of control and prevention measures, as well as the consequences of reopening policies (Zhang et al., 2020; Zhou et al., 2020). In order to find useful intervention strategies for COVID-19, it is vital to understand how the disease spreads and to predict how it will transmit across space and time (Chang et al., 2021; Chinazzi et al., 2020).

For the past few months, studies on the modeling and prediction of COVID-19 transmission have proliferated. Among them, the majority of work manages to estimate and predict basic epidemic parameters, including the basic reproductive number, the number of incident cases, the onset or the peak time for specific regions (Niehus, De Salazar, Taylor, & Lipsitch, 2020; Roosa et al., 2020; Wu, Leung, & Leung, 2020). Since quantifying spatio-temporal heterogeneity of these parameters is critical for discovering realistic transmission patterns and making accurate predictions, it has recently been explored with different epidemiological parameters, which are estimated using various compartmental epidemic models (Lai et al., 2020; Zhang et al., 2020; Zhou et al., 2020).

Studies on the transmission modeling of infectious diseases have produced an extensive volume of work, which covers three categories: (1) agent-based (Germann, Kadau, Longini, & Macken, 2006); (2) metapopulation-based (Balcan et al., 2009); and (3) machine learning-driven models (Senanayake, O’Callaghan, & Ramos, 2016). In particular, the estimation of the transmission path over the study region can contribute to the visualization of epidemic spreading patterns in a holistic view, such as the epidemic invasive tree map (Balcan et al., 2009) or source–sink risk density maps (Wesolowski et al., 2012). Historical models of infectious diseases, especially influenza epidemics, cannot readily be used for the exploration of COVID-19 spreading patterns because COVID-19 manifests fast-changing transmission patterns, given that massive intervention measures have been taken across the world.

In response to the challenges of modeling spatio-temporal spreading patterns for COVID-19 and many other infectious diseases, we need to account for space-time varied epidemic dynamics that are inherent to this particular epidemic and introduced by human interventions. We aim to build an integrated data-driven model to recover the strength of historical transmission links for COVID-19. The model is essential to formulate effective social distancing policies, inform the public about the exposure risks in communities, and allocate limited health resources efficiently. This study leverages graph embedding methods, which are popular representative techniques in machine learning (Goyal & Ferrara, 2018), to generate compact representations of areas and model mobility-driven epidemic mechanisms of COVID-19. The modeled dynamics, along with areal representations, can be used to predict future infections. The prediction of future transmission paths hinges on the availability of future traffic flow data. We focus on the estimation of historical transmission paths and leave the path prediction task for future studies.

This study makes the following contributions to spatial epidemiology:
1. An integrated data-driven learning model to capture the mobility-related spreading mechanism of COVID-19, utilizing multi-sourced mobility and attributed data. By integrating deep embedding and compartmental epidemic models, the hybrid model offers a powerful tool to conduct advanced analysis and to suggest effective mitigation measures.

2. A transmission path visual analytic approach that engages stakeholders (e.g., public health officials, managers from transportation agencies) to make informed intervention decisions and enables public messaging by depicting the strength of the historical transmission paths between areal units.

3. A case study of COVID-19 spread modeling and prediction in the state of Colorado, USA that demonstrates the applicability of the proposed approach in capturing space-time varied epidemic dynamics and supporting informed decision-making on disease intervention and prevention.

2 | RELATED WORK

2.1 | Modeling epidemic spread

The extensive volume of work on the modeling of epidemic spread covers three main categories: (1) agent-based (Germann et al., 2006); (2) metapopulation-based (Balcan et al., 2009); and (3) machine learning-driven models (Senanayake et al., 2016). In particular, the estimation and prediction of the transmission path over the study region can contribute to the visualization of epidemic spreading patterns in a holistic view, such as the epidemic invasive tree map (Balcan et al., 2009) or source–sink risk density maps (Wesolowski et al., 2012). Human mobility data are essential gradients for epidemic spread modeling and prediction, including airline flights (Colizza et al., 2006), intra-city commuting (Dalziel et al., 2013), mobile phones (Wesolowski et al., 2012), and household visit survey data (Stoddard et al., 2013). For example, Balcan et al. (2009) develop a structured metapopulation model, named Global Epidemic and Mobility (GLEaM), to simulate global epidemic spreading on the basis of a multi-scale mobility network. These aforementioned models of infectious diseases, especially influenza epidemics, cannot readily be used for the exploration of COVID-19 spreading patterns because COVID-19 exhibits versatile transmission patterns, given that massive intervention measures have been taken across the world. These fast-changing patterns render it challenging for traditional analytical-based models to adapt and make accurate predictions.

2.2 | Modeling COVID-19 transmission

Since the outbreak of COVID-19, numerous efforts have been dedicated to the modeling and prediction of how this epidemic will evolve. Some researchers focus on the prediction of infected case numbers in specific regions at specific times using the traditional Susceptible–Infectious–Recovered (SIR) model or its variants (e.g., Susceptible–Exposure–Infectious–Recovered [SEIR]; Roosa et al., 2020; Wu et al., 2020). These efforts are based on rough estimates of population flow and basic epidemic parameters, including the basic reproductive number, incubation period, and transmission rate. For example, Wu et al. (2020) utilize a SEIR metapopulation model to predict the spread of COVID-19 in China and worldwide using Tencent mobility data and monthly international air passenger data. Traditional compartment models are rarely used to reveal detailed spatio-temporal transmission paths of COVID-19 due to the challenges in explicit modeling of spatio-temporal heterogeneity of disease transmission. Metapopulation mobility models have the advantage of capturing infectious risk differences between different population groups (Balcan et al., 2009) and have been used for modeling the spatio-temporal spread of COVID-19 (Arenas et al., 2020; Chinazzi et al., 2020). Parsimonious macroscopic models, such as exponential growth, the self-exciting branching process, and the SIR model can be of use for the modeling and forecasting of COVID-19 transmission (Bertozzi et al., 2020). To derive time-dependent transmission paths, researchers should
develop spatially explicit models (e.g., network-based models; Prasse et al., 2020) to simulate space-time varied epidemic spreading processes based on multi-sourced georeferenced data. These models do not explicitly predict transmission paths, which is the focus of this study.

3 | METHODS

3.1 | Overview

We define the problem of COVID-19 spread modeling in this sub-section. The study region can be abstracted as a directed graph $G = \{V, E, W_t\}$, where $V$ represents a set of predefined areal units ($|V| = N$), $E$ is the set of links denoting traffic connections between areal units, and $W_t \in \mathbb{R}^{N \times N \times T}$ is a time-varying adjacency matrix that quantifies the transmission potential of the disease between any pair of nodes during the time interval $t$ in $T$. Each element in $W_t$ can be denoted as $W_{t}(i,j)$. Formally, the COVID-19 transmission modeling problem for the graph $G$ is to estimate the transmission potential matrix $\hat{W}_t \in \mathbb{R}^N$ given historical numbers of infections and human mobility data.

Mobility data are particularly useful for COVID-19 spread modeling (Buckee et al., 2020). We couple a modified SIR model with a spatio-temporal extended representation learning model to model the mobility-related epidemic spread and local transmission. The reason for combining the two models is that data-driven representation learning can better capture space-time variant transmission patterns than the SIR model, which however has the capability to derive key transmission parameters when the reported case data are error-prone and unreliable.

Figure 1 shows the pipeline of the proposed modeling methodology. Initially, a mobility network is constructed, based on a commute network. It can be extended to a multi-layered network that contains other mobility networks, such as airline networks at multiple scales. The inter-node transmission potential is encoded as an adjacency matrix and initialized by inter-node traffic flow count. To derive epidemic parameters, inter-region traffic flow count, census, and COVID-19 daily case data are fed into a dual SIR model, which is inspired by a recent mobility-based SIR model (Goel & Sharma, 2020). Our SIR model uses census and traffic flow data to model intra-node and inter-node transmission, respectively. Intra-node cases refer to community-acquired infections, whereas inter-node cases are defined as infected people who have traveled to other nodes or are exposed to infectious cases coming from other nodes in the 14 days before the confirmation of their infection. Based on simulated infectious numbers of the SIR model, we leverage graph embedding to produce compact vector representations for individual nodes and inter-node transmission links due to its powerful representative capabilities (Goyal & Ferrara, 2018). Node
embeddings encode intra-node transmission risk under the impact of local socio-economic and environmental factors. Similar to other dynamic spatio-temporal processes (Li, Yu, Shahabi, & Liu, 2018), the transmission of the virus can be effectively modeled as random-walk diffusion. Therefore, we use diffusion embeddings to capture inter-node transmission risk controlled by traffic flow.

To be specific, this study adopts a Diffusion Convolutional Recurrent Neural Network (DCRNN) model (Li et al., 2018) and a Graph Convolutional Network (GCN) model (Kipf & Welling, 2017) to capture inter- and intra-node transmissions, respectively. The two types of embeddings are integrated and used as input for a Long Short-Term Memory (LSTM) model (Hochreiter & Schmidhuber, 1997) to train a predictor for each areal unit. As the SIR learning, graph embedding, and prediction model are constantly trained and updated, we estimate and update the transmission path matrix for each date, which is used to recover daily historical transmission pathways. We also use integrated embeddings for further visual analytics, such as role analysis for graph nodes.

The core of the pipeline is at the integration of a SIR model and two graph embedding components, leveraging the simulation capabilities of compartmental epidemic models (the dual SIR model) and the expressive representation of graph embedding techniques. The two parts complement each other and contribute to the understanding and prediction of COVID-19 spread over space and time.

3.2 | SIR modeling

Different from the traditional SIR model, the developed dual model used in this study distinguishes two categories of susceptible population: individuals who are exposed to imported infected individuals and those that are susceptible due to intra-node transmission. Figure 2 presents the proposed SIR modeling workflow. For each node at specific time $t$, the proposed model estimates the size of intra-node susceptible population using local population data (Equation 5). The inter-node susceptible population is obtained using local population and traffic flow data (Equation 6). The two categories of susceptible population are then summed to obtain the number of total susceptible population $S_{\text{all}}(t)$. Based on daily COVID-19 statistics, we can compute the number of newly infected individuals (Equations 7–9), which is then used to estimate the transmission rate $\beta(t)$ (Equation 10). Note that $\beta(t)$ is not constant and varies over each date in the proposed dual SIR model. Finally, the numbers of intra- and inter-node infectious individuals can be predicted based on the estimated number of susceptible population and $\beta(t)$ (Equations 11 and 12).

![FIGURE 2 SIR modeling workflow. The inputs of the model are multi-sourced data. The model outputs predicted intra-node and inter-node infectious numbers](image-url)
We chose to extend the traditional SIR model rather than the SEIR model because the latter is more challenging to derive an accurate estimation for the extra parameter with limited data. The extra parameter is the probability of being infected for an exposed person. The details of the proposed dual SIR model are described in the following.

Assume the base population of a graph node at the end of date \( t \) is written as \( P_{\text{base}}(t) \) and the net traffic flow count is denoted as \( P_{\text{net}}(t) \) (calculated as incoming traffic flow count minus outgoing counterpart). We can update the base population by accounting for net traffic flow of date \( t \):

\[
P_{\text{base}}(t) = P_{\text{base}}(t-1) + P_{\text{net}}(t) \tag{1}
\]

The net flow ratio \( r_j \) can be calculated as:

\[
r_j = \frac{P_{\text{in}}(t)}{P_{\text{total}}(t) - P_{\text{base}}(t)} \tag{2}
\]

where \( P_{\text{in}}(t) \) is the incoming traffic flow count of date \( t \), \( P_{\text{total}}(t) \) is the population of the entire study region, and \( r_j \) denotes the ratio of incoming traffic flow over the population outside the node in the graph.

For each node, the newly infected population \( N(t) \) by the end of date \( t \) can be calculated as the difference between the number of cumulative infected population on date \( t \) and \( t-1 \) [i.e., \( I_c(t) \) and \( I_c(t-1) \)]:

\[
N(t) = I_c(t + t_{\text{delay}}) - I_c(t - 1 + t_{\text{delay}}) = N_{\text{intra}}(t) + N_{\text{inter}}(t) \tag{3}
\]

where \( t_{\text{delay}} \) is the average delayed reporting period (i.e., the number of delayed days before confirmation), and \( N_{\text{intra}}(t) \) and \( N_{\text{inter}}(t) \) represent numbers of intra-node transmission cases and imported cases from other nodes that happen during the date \( t \), respectively. \( N(t) \) can be written as the sum of \( N_{\text{intra}}(t) \) and \( N_{\text{inter}}(t) \).

The number of susceptible individuals at the beginning of date \( t \), which is denoted by \( S_{\text{all}}(t) \), can be divided into two components:

\[
S_{\text{all}}(t) = S_{\text{inter}}(t) + S_{\text{intra}}(t) \tag{4}
\]

where \( S_{\text{inter}}(t) \) and \( S_{\text{intra}}(t) \) are susceptible population exposed to locally infected or imported infected individuals, respectively. These two components can be calculated as:

\[
S_{\text{intra}}(t) = P_{\text{base}}(t - 1) - P_{\text{out}}(t) - I_c(t - 1 + t_{\text{delay}}) \tag{5}
\]

\[
S_{\text{inter}}(t) = P_{\text{in}}(t) - I_{ci}(t - 1 + t_{\text{delay}}) \times r_f \tag{6}
\]

where \( P_{\text{out}}(t) \) is the outgoing traffic flow count of date \( t \). \( I_{ci}(t) \) is the cumulative number of infected cases outside the node in the entire study region for the \( t \)th day.

Active infectious population induced by local intra-node transmission at the beginning of date \( t \), \( I_{\text{intra}}(t) \), can be also written as:

\[
I_{\text{intra}}(t) = I_c(t - 1 + t_{\text{delay}}) - R_c(t - 1) - D_c(t - 1)
= I_c(t - 1 + t_{\text{delay}}) - I_c(t - 1 + t_{\text{delay}} - t_{\text{recover}}) \tag{7}
\]

where \( R_c(t - 1) \) and \( D_c(t - 1) \) are numbers of cumulative recovered and deceased individuals from the previous day, respectively. \( t_{\text{recover}} \) is the mean number of days it takes to recover from COVID-19. The number of recovered population can be calculated as follows:
The active infectious population induced by imported cases at the beginning of date \( t \), \( l_{\text{inter}}(t) \), can also be calculated as:

\[
l_{\text{inter}}(t) = \left[ l_{c/i} \left( t - 1 + t_{\text{delay}} \right) - l_{c/i} \left( t - 1 + t_{\text{delay}} - t_{\text{recover}} \right) \right] \times r_f
\]

The time-varying transmission rate \( \beta(t) \) can be computed as:

\[
\beta(t) = \frac{NN(t)}{[l_{\text{intra}}(t) + l_{\text{inter}}(t)] \times S_{\text{all}}(t)}
\]

where \( l_{\text{intra}}(t) \) and \( l_{\text{inter}}(t) \) represent active infectious population induced by local intra-node transmission and by imported cases at the beginning of date \( t \), respectively.

Finally, we can write two equations to compute the newly infected individuals for each date from the two sources (local and imported transmission) as follows:

\[
N_{\text{intra}}(t) = \beta(t) \times l_{\text{intra}}(t) \times \frac{S_{\text{all}}(t)}{P_{\text{base}}(t)}
\]

\[
N_{\text{inter}}(t) = \beta(t) \times l_{\text{inter}}(t) \times \frac{S_{\text{all}}(t)}{P_{\text{base}}(t)}
\]

Following the traditional SIR model, we can summarize the above equations as follows:

\[
\frac{dS_{\text{all}}(t)}{dt} = -\beta(t) \times S_{\text{all}}(t) \times \left[ l_{c} \left( t + t_{\text{delay}} \right) + l_{c/i} \left( t + t_{\text{delay}} \right) \right]
\]

\[
\frac{dl_{c}(t)}{dt} = \beta(t) \times S_{\text{all}}(t) \times \left[ l_{c} \left( t + t_{\text{delay}} \right) + l_{c/i} \left( t + t_{\text{delay}} \right) \right] - \frac{dR_{c}(t)}{dt}
\]

The traditional SIR model can be written as three differential equations:

\[
\frac{dS_{\text{all}}(t)}{dt} = -\beta S_{\text{all}}(t) I_{\text{all}}(t)
\]

\[
\frac{dl_{\text{all}}(t)}{dt} = \beta S_{\text{all}}(t) I_{\text{all}}(t) - \gamma I_{\text{all}}(t)
\]

\[
\frac{dR_{\text{all}}(t)}{dt} = \gamma I_{\text{all}}(t)
\]

where \( R_{\text{all}}(t) \) is the total number of recovered population at time \( t \), respectively. \( \beta \) is the effective contact rate (i.e., likelihood of being infected) for susceptible individuals. \( P \) is the total population, \( P = S_{\text{all}}(t) + I_{\text{all}}(t) + R_{\text{all}}(t) \). The three equations illustrate how the three groups change with time.

There are three improvements of the proposed dual model over the traditional SIR model.
1. Different from the traditional model, we distinguish two categories of susceptible population: individuals who are exposed to imported infected individuals and those who are susceptible due to local transmission. We then predict two types of newly infected individuals: infected by local or imported transmission (denoted as intra- and inter-node infections). Based on inter-node infection data, transmission paths are estimated using graph diffusion embedding.

2. Rather than using a fixed value of $\beta$, we use a time-varying $\beta(t)$ to better capture the dynamics of transmission from the infected population to the susceptible individuals.

3. The proposed dual SIR model accounts for delayed reporting periods (i.e., the number of delayed days before confirmation), which play a similar role to the incubation period in the SEIR model.

Together, these three improvements not only deliver a better prediction performance than the traditional SIR model, but also enable the estimation of historical transmission paths.

3.3 | Modeling the spread induced by inter-areal traffic flow

As we have previously mentioned, the infected population is divided into two groups: individuals who are infected by imported cases and individuals who are infected by local transmission. In this sub-section, we will present how we model the spread of the virus induced by inter-area traffic flow. A network-based diffusion process has been studied in statistical physics (Gomez et al., 2013; Tejedor, Longjas, Foufoula-Georgiou, Georgiou, & Moreno, 2018).

Due to the presence of spatio-temporal heterogeneity during epidemic spreading, we argue that data-driven models are more appropriate than analytical models with respect to modeling the effects of this heterogeneity on disease diffusion. Specially, we consider the spread of the disease as a diffusion process within a network and apply diffusion convolutions to capture heterogeneous transmission patterns over space and time (Figure 3) (Li et al., 2018).

We compute diffusion embeddings to represent inter-node links, supervised by the estimated size of inter-node infected population. The embeddings use compact vectors to account for inter-node traffic dynamics, epidemic spreading dynamics, thereby forcing links with similar spreading patterns to be close to each other in the embedding space. Assume the disease spreads for $k$ time intervals (or steps); the diffusion process is simulated by a multi-step random walk on the directed graph $G$. The stationary distribution of the random walk starting from node $i$ can be computed as (Teng, 2016):

$$\text{Pr}_j = [\text{Pr}_{i \rightarrow 1}, \text{Pr}_{i \rightarrow 2}, \ldots, \text{Pr}_{i \rightarrow N}]^T = \alpha \sum_{k=0}^{\infty} \left(1 - \alpha \right)^k (\text{W}_k^T (\text{D}_{out}^{-1} \text{W}_k)^{-1})^k \text{1}_i$$

**Figure 3** Sequence-to-sequence diffusion embedding. The diffusion embeddings are generated by an LSTM-based diffusion encoder, which relies on $K$-order diffusion convolutions over the $K$-order neighboring nodes of each node.
where \( \alpha \) is the restart probability at \( i \). Recall that \( W_k \) is the adjacent matrix at time \( k \) that describes the inter-node transmission potential. Each element of \( W_k \) is the traffic flow count between each OD pair. \( D_{W}^{\text{out}} \) is a diagonal matrix of out degrees. \( \mathbf{1} \) is a vector in which the \( i \)th position is filled with value 1, while others are 0.

To facilitate the computation, we can rewrite Equation (15) as an infinite \( K \)-step truncation form (Li et al., 2018):

\[
F_\theta[N_{\text{inter}}(t)] = \sum_{k=0}^{K-1} \theta^k \left( W_k \left( D_{W}^{\text{out}} \right)^{-1} \right)^k N_{\text{inter}}(t)
\]

The above equation is then used to define the diffusion convolution operation that describes the disease transmission process.

### 3.4 Modeling community transmission regulated by local attributes

In addition to imported cases, intra-node local transmission needs to be accounted for. This type of infected individual is not directly related to inter-node traffic flow. Rather, they are affected by local socio-demographic factors. We omit environmental and weather factors because these factors are assumed to be homogeneous across the study region. Since fine-grained intra-node mobility data are not available, we apply node embedding to socio-demographic data to model the contribution of intra-node activities to community-based transmission. Using socio-demographic data (see Table 1 for a list) and intra-node infected number as the inputs, we employ graph embedding to encode complex interactions between these factors and local transmission for each graph node via compact one-dimensional vectors (Figure 4). Supervised by the predicted intra-node infectious numbers, the embedding is performed by a GCN encoder via multiple graph convolutions over multiple time steps in parallel. The outputs of these GCN encoders are then concatenated together as the embeddings for graph nodes.

### 3.5 Prediction via the integration of diffusion and node embeddings

For each node, daily reported case numbers can be recorded as time-series data, \( \mathbf{X}_i = \{ x_1^i, x_2^i, \ldots, x_T^i \} \), where \( T \) is the size of the time series. Typically, given \( \mathbf{X} \) as the set of time-series data in the study region, we aim to predict the numbers of confirmed infected cases for all areas over a prediction horizon \( h \) based on node and diffusion embeddings \( \mathbf{E}_{\text{node}} \) and \( \mathbf{E}_{\text{diffusion}} \). Formally, the prediction problem can be formulated as:

### TABLE 1 Data description and sources

| Description               | Data source                                                                 |
|---------------------------|----------------------------------------------------------------------------|
| Population                | Population by county or neighborhood                                       |
| Health status             | Percentage of population in poor health by county, converted to number of   |
|                           | unhealthy population by county or neighborhood                              |
| Poverty                   | Number of persons in poverty                                                |
| Unemployment              | Unemployment in civil labor force by county or neighborhood                 |
| Mobility                  | Weekday and weekend traffic flows since March 2020                          |
| COVID-19 cases            | Daily COVID cases by county or neighborhood since March 2020               |
|                           | American Community Survey (ACS) 2018, neighborhood level.                    |
|                           | Neighborhoods are constructed by merging census tracts according to local   |
|                           | development plans                                                           |
|                           | TomTom Move                                                                 |
|                           | Colorado Department of Public Health and Environment, Tri-county Health      |
|                           | Department                                                                  |
The dimensions of node embeddings and diffusion embeddings are set as the same number $D$ to facilitate the integration of the two types of embeddings. For each date, we construct a matrix $Y \in \mathbb{R}^{N \times N}$ to concatenate the two types of embeddings. Each row of $Y$ contains diffusion embeddings that correspond to all edges connected to a node, except that the diagonal element is occupied by embedding for the node.

For each row in $Y$ (i.e., concatenated embedding), a weighting mechanism is applied to produce an integrated embedding that condenses node embedding and diffusion embeddings in this row into a $d$-dimensional vector. For example, for the $i$th row of $Y$, the integrated embedding $I_i$ can be computed as:

$$I_i = \sum_{j=1}^{N} w_{ij} \times Y_{ij}$$

Based on the integrated embeddings, an LSTM model is used to predict the number of new daily infected population for each area over a time period (prediction horizon) (Figure 5). The incubation period of COVID-19 is also considered and modeled as a gap period between the end date of input data and the initial date to be predicted. Given $t$ days of data as inputs, the predictor forecasts the number of new cases over the period $[t + z, t + z + n]$ days. To address the gap issue, we use a moving-window training scheme to better capture spatio-temporal heterogeneity in disease transmission (Figure 6).
Within each batch, we feed $n$ sequences of time-series data $\{X_1[i], X_2[i], ..., X_n[i]\}$ as inputs into the model and use time-series data $\{X_1[i+z+t+z+h], X_2[i+z+t+1+z+h], ..., X_n[i+z+t+n+z+h]\}$ as outputs to train the predictor, where $z$ is the number of gap days. The moving window-based training scheme replaces previous training data by newly confirmed case numbers to adjust the prediction model. We set the step size as 1 day. Since node embeddings are assumed to be time-dependent, the model is constantly updated when new node embeddings are available. Nevertheless, the predictor refreshes diffusion embeddings by re-initiating model weights and retraining from scratch because imported cases are dependent on traffic flow of the same date. The predictor makes predictions on the time horizons of $[T, T + h], [T + 1, T + 1 + h], ..., [T + n' + z, T + h + n' + z]$ successively, given $n'$ sequences of known time-series data $\{X_1[T-z-T-z], X_2[T+1-z-T-T-z], ..., X_n[T+n'-z-T-T+n'-z]\}$, where $T - z > t + n + z$.

### 3.6 Implementation

The proposed modeling approach is implemented in Python. For each date, we build a directed mobility graph and use NetworkX 2.4 to build and store graph data. Numpy is used for matrix computation. Scikit-learn is used for performing clustering and evaluating prediction performance (i.e., computing MAE and MAPE). Embedding and prediction models are developed using Keras 2.1. We used the graph convolutional network model developed by Kipf and Welling (2017) to model inter-community transmission. The LSTM-based prediction model was trained using the mean squared error loss function. To avoid overfitting, we added a dropout layer after each neural layer with a dropout rate of 0.2 and use L2-norm regularization with a small hyperparameter (5E-4) for all embedding and prediction models. The node and diffusion embedding dimensions are all set as 128. For the LSTM model, we enforce positive kernel weight constraints to ensure positive contributions of nodes. For the GCN model, we use the tanh activation function to ensure the outputs fall between 0 and 1. We use the ReLu activation function for all other deep learning models. All embedding and prediction models are trained for 200 iterations before stopping.

### 4 RESULTS

#### 4.1 Study region and data

The study region covers the Denver metropolitan area, including seven counties in the state of Colorado: Denver, Boulder, Broomfield, Jefferson, Douglas, Adams, and Arapahoe (Figure 7). For the last three counties...
(tri-county), we obtain COVID-19 daily case data at the neighborhood level. Areas with sparse population in the counties of Adams and Arapahoe were removed. For the remaining four counties, only county-level data are accessible. The study region is the most populated area in Colorado, with very minor traffic connections to neighboring counties and other states. After the onset of the virus, both commute and air traffic drop dramatically, implying that imported infections from areas outside the study region would be insignificant. To model the impact of areas beyond the boundary of the study region, we incorporate the data of 12 counties and use one external node to represent these 12 counties. The external node is also involved in the proposed spread model as a regular graph node. In Table 1, we list the datasets used in this study, including census, mobility, and COVID-19 case data. In particular, we obtained inter-node trip data (origin-destination [OD] pairs) from TomTom’s traffic analytic service, which produces high-quality mobility data based on multiple types of sensors over road networks, covering different types of travel purposes (commute, shopping, and other social activities). The coverage and accuracy of our trip data are arguably better than mobility data derived from mobile phone data. We tested the proposed model on a Microsoft Windows 10 operating system with a desktop machine, which has a 4-core Xeon E5-1607 processor @3.1 GHz, 8 GB of memory, and a NVIDIA TITAN Xp GPU (12 GB VRAM). It took less than 2 min to finalize the training and test for each round, which is reasonable for public health applications.

The first infected case was reported on March 5, 2020 in the study region. There are two critical dates: March 26, when the stay-at-home order was issued, and May 29, when the state was initially reopened. In the following, we chose to present the modeling results for March 12 because the virus started to spread within 1 week of the first case. We also selected April 12 and June 12 for mapping and analysis because these two dates were approximately 2 weeks after the above mentioned two critical dates, a time period in which the effectiveness of the orders can be fully observed. Different visual analytics tools can be developed and used to examine the spreading patterns from different perspectives, thus helping validate the analysis results and promoting understanding of the transmission mechanism.

![Study region in the state of Colorado (orange-colored areas)](image)
4.2 Estimation of historical transmission paths

As we have already computed the diffusion matrix for each date, transmission paths between each areal unit can readily be reconstructed. The accumulated diffusion weights for each node can be categorized into three classes: high, medium, and low risk. To avoid clutter, we set a threshold of 0.07 to filter trivial paths and to retain the most significant paths. According to Figure 8, we can observe distinct transmission patterns for the three critical dates. On March 12, we identify Denver and Jefferson as two primary sources of transmission in the study region. Denver County posed higher risks for the study region than Jefferson, as indicated by the width of its paths, which exhibit a hub and spoke form. One month later, the map of April 12 shows that Jefferson County developed into another significant source of infection. Many other neighborhoods became medium-risk areas and started to infect neighboring areas. Although inter-area traffic within the entire study region had decreased by 50% compared to that of March 12, the pandemic still managed to spread over the study region. The traffic flow recovered to 90% of the level of March 12 but Denver County only recovered to 72%. Compared to April 12, the transmission path map of June 12 depicts a slightly better scenario: the number of medium-risk neighborhoods decreased and

**FIGURE 8** Transmission path map for the study region. To reveal the most significant paths, we remove inter-node links where the number of introduced infections is less than a threshold. Green, orange, and red represent high, medium, and low risk of the nodes. Path width is proportional to the weight of the diffusion matrix (i.e., potential risk of COVID-19 transmission)
the path widths (i.e., transmission potential) also decreased. The improvement over April 12 might be attributed to behavioral changes of residents (e.g., social distancing or wearing masks).

In addition to the holistic transmission path map, we also develop a transmission path tree for specific nodes to examine their impact on other nodes. Figure 9 shows two of the most extended transmission path trees, which originated from Denver County and a neighborhood in Douglas County. Again, we chose March 12, April 12, and June 12 for analysis to examine the effectiveness of preventive measures (e.g., stay-at-home, mask recommendation, and reopening). After the stay-at-home order was implemented, both the strength (cumulated diffusion weights) and scale of the transmission path (measured as the number of impacted areal units along the path, i.e., white nodes on the map) became larger (April 12 vs. March 12), thus implying that mitigation measures did not stop the virus from spreading. After the state started to reopen for 2 weeks on June 12, the strength of transmission

**FIGURE 9** Transmission path trees originating from two nodes. The map was created with the ArcGIS Pro 2.5 Distributive Flow Lines tool, which merges transmission routes starting from the source node to sink nodes and delays the generation of branches until routes approach sink nodes. Different transmission path trees are visualized with different colors. The two most significant transmission path trees are selected for visualization.
The number of infected nodes also decreased. Overall, the Denver-originated transmission pathways dominated other paths in terms of strength and scale over the period from March to June.

4.3 | Daily prediction of confirmed cases

The prediction errors of the proposed approach and the traditional SIR model are depicted in Figure 10. The training size $n$ has a minor influence on the performance since the prediction errors of different training sizes (7, 10, and 15) are similar. The spike errors are usually caused by the inclusion of data backlog in the report.
15 days) are quite close to each other. Our model outperforms the traditional SIR model by a large margin over the four months. For most dates, the MAE remains at an acceptable level and gradually decreases, indicating that our model can adapt to the data when more training samples are used.

In Figure 11, we depict the decomposition of predicted cases for three dates. For many neighborhoods on March 12, almost all confirmed cases were imported. Denver and Jefferson counties, along with a few neighborhoods located in the southeast of Denver, are characterized by a high percentage of intra-node infected cases, suggesting that they would become a source of importation later if they emit significant traffic flow. If we evaluate public health interventions until April 12 based on total confirmed cases, it seems that the effects were disappointing since the number was still growing. However, the map of April 12 in Figure 11 reveals that local transmission started to prevail for most areas close to Denver County, while only a few neighborhoods were infected by imported cases, implying that the stay-at-home order and other preventive measures helped slow the spread of the disease. Interestingly, imported cases on June 12 gained momentum again as reopening policies were announced. Neighborhoods located in the south of Denver (also in the east of Jefferson) were the most hit areas.
presumably owing to their strong connections to the two major source counties (Denver and Jefferson). These findings are reasonable, following the transmission patterns of the virus.

4.4 | Role analysis

Figure 12 indicates that the clustering of integrated embeddings typically results in four or five clusters, which exhibit distinct transmission patterns and play different roles during the pandemic. We can identify different roles played by each areal unit via the examination of maps in Figures 10–12 and Table 2.

Table 2 summarizes the population, flow, and transmission characteristics of typical clusters. Interestingly, the intermediate sink nodes have the highest incidence rate for March 12 and June 12 among all types of node. Boulder and Jefferson counties are categorized as minor hubs, probably due to their relatively large sizes of population, traffic flow, and infection cases compared to neighborhoods of other counties. The two counties then pose significant transmission risks on nearby areas, as we model them as two nodes. Rather, if neighborhood-level data are available, different neighborhoods in these two counties may take different roles.

The clustering structure and roles assumed by different clusters can be examined along with local economic-demographic environments and inter-areal traffic interconnections. Clusters 1 and 2 can be categorized as transmission source hubs, which constantly infect many other areas in the study region during the pandemic. With its central geographical position and large amount of emitting traffic flow, Cluster 1 (Denver County) facilitates the rapid spread of COVID-19 given that it infects most neighborhood nodes: 27, 32, and 23 nodes on March 12, April 12, and June 12, respectively. We classify Cluster 1 as a major source hub and Cluster 2 as a minor hub. Both clusters are characterized by large sizes of population, incoming flow, and outgoing flow, as well as infected cases.

Many factors contributed to the formation of hubs during the pandemic. Among them, we speculate that population, geographical location, and intervening measures are the most influential ones. A large population usually corresponds to a large size of infection. Being located in the central hub of the transportation network certainly facilitates the spread of the disease. Other clusters can be classified as transmission sink nodes, which can be further divided into intermediate (Cluster 3) and peripheral sink nodes (Clusters 4 and 5). Peripheral sink nodes are less vulnerable to the virus as they are infected at a slower rate than intermediate sink nodes. Intermediate sink nodes are located between source hubs and peripheral sink nodes on transmission paths. They are characterized by a high susceptibility rate since they are closely connected to source hubs and have relatively less population than hubs. Intermediate sink nodes are not the final stops on transmission paths, as they also infect a number of peripheral sink nodes in their vicinity. As indicated by Figure 12 and Table 2, intermediate sink nodes still have significant large inflow traffic. From the TomTom data, we can find that they also emit some outgoing traffic flow. This incoming and outgoing traffic justifies the intermediate role played by these nodes. Peripheral sink nodes can be further classified depending on their geographical locations and traffic connections to the central nodes of the study region. Nodes that are located in the outskirts and have the least traffic flow can be labelled as remote peripheral sinks. The map of June 12 shows that Cluster 5 is merged with Cluster 4, both of which are still peripheral sink nodes. Note that the definitions of sink and source are not absolute and depend on the context: if we examine the problem at the national level, Denver County may change to other roles.

Most of the nodes fall into the two categories of peripheral sink nodes, which is similar to the analysis results of some artificial networks such as air transportation (Guimera, Mossa, Turtschi, & Amaral, 2005). Intermediate sink nodes are also common in our case study, suggesting that transmission of COVID-19 somehow resembles the signal transduction in biological networks (Guimera & Amaral, 2005). Interestingly, these two observations imply that the transmission network may result from the joint effects of both human mobility dynamics and virus transmission mechanisms. The clustering structure is consistently stable over the three dates, indicating that the roles played by these areal units in spreading the disease are rather stable.
FIGURE 12  Clustering of neighborhoods using integrated embeddings. Different colors indicate different types of cluster. Each dot corresponds to a neighborhood or a county. The size of dot is proportional to the number of newly confirmed cases for each specific day. The scatterplot illustrates the relationship between inflow trip counts and newly infected intra-node cases for each date. The external node is excluded in the figures.
| TABLE 2  | Statistics of clusters |
|----------------|------------------------|
|               | No. | Average nodes | Average population | Average cumulative case | Average incidence (%) | Average predicted daily cases | Average outflow | Average inflow |
| March 12      | 1   | 1             | 663,303             | 125.0                   | 1.88E-02               | 27.0                   | 251,695        | 248,569       |
|               | 2   | 6             | 172,866             | 17.3                    | 9.31E-03               | 3.8                    | 42,958         | 43,513         |
|               | 3   | 12            | 31,778              | 2.8                     | 1.29E-02               | 0.6                    | 10,848         | 10,976         |
|               | 4   | 21            | 19,800              | 0.9                     | 8.88E-03               | 0.3                    | 5,033          | 5,034          |
|               | 5   | 47            | 10,575              | 1.3                     | 1.31E-02               | 0.3                    | 2,660          | 2,625          |
| April 12      | 1   | 1             | 663,303             | 2,071.0                 | 3.12E-01               | 67.0                   | 97,045         | 94,598         |
|               | 2   | 2             | 436,386             | 740.5                   | 1.62E-01               | 40.0                   | 49,547         | 50,023         |
|               | 3   | 15            | 43,307              | 101.3                   | 2.18E-01               | 5.5                    | 6,208          | 6,269          |
|               | 4   | 32            | 18,310              | 35.6                    | 2.01E-01               | 1.0                    | 2,544          | 2,553          |
|               | 5   | 37            | 6,026               | 14.4                    | 2.33E-01               | 0.4                    | 739            | 737            |
| June 12       | 1   | 1             | 663,303             | 6,663.0                 | 1.00E-00               | 33.0                   | 184,232        | 179,456        |
|               | 2   | 2             | 436,386             | 1,901.0                 | 4.28E-01               | 13.5                   | 100,242        | 101,357        |
|               | 3   | 30            | 30,834              | 205.0                   | 7.19E-01               | 0.5                    | 8,937          | 9,103          |
|               | 4   | 54            | 9,880               | 67.6                    | 6.05E-01               | 0.3                    | 2,323          | 2,267          |
The proposed spread modeling approach is constrained by the availability of disaggregate COVID-19 case data, such as individual contact tracing data. For areas outside the tri-county, we can only access state-level data, which limits our ability to provide high-resolution transmission path estimates for the other four counties. Genome sequencing data are also helpful for the tracing of transmission paths (Rockett et al., 2020). However, the sparsity of genomic data makes it impossible to recover high-resolution complete transmission pathways. The proposed modeling approach enables near real-time tracking of virus spread at any desired spatial and temporal granularities if data are available. The verification of estimated transmission paths is challenging because collecting genome sequencing data for each infectious individual is almost impossible. The results can be verified by evaluating the total infections. The proposed approach predicts inter- and intra-node infections separately. If the two predictions are accurate, then the total numbers of daily infections should be correct. Since we have daily reported case data, it is easy to verify if predicted daily infection cases are consistent with the reported data. The given evaluation results demonstrate that our results can produce accurate predictions. Meanwhile, we analyze the decomposition of prediction cases within the study time period (Figure 11). For many regions, inter-node transmissions decrease after March 12 but rebound after June 12, after 2 weeks of reopening. These findings generally follow the transmission patterns of the virus. We also investigate and justify the estimated transmission paths based on the characteristics of the virus and socio-economic features of the studied regions. The estimated diffusion weights (i.e., strengths of transmission paths) are intuitively related to incoming traffic flow from the origin nodes, infectious rates of the origin nodes, as well as the population of the origin/destination nodes. For each destination node, we visualize the correlations between the diffusion weights and potential infectious cases from the incoming flow as well as the ratio of incoming flow over the population size of the origin node. Figure 13 shows such correlations of an intermediate sink node (Exposition Park and Aurora Hills) for three critical dates, showing that the estimated diffusion weights are positively correlated to potential incoming infectious cases (computed as incoming flow multiplied by infectious rate) and incoming traffic density (computed as incoming flow divided by population). The estimated diffusion weights are positively correlated to potential incoming infectious cases (computed as incoming flow multiplied by infectious rate) and incoming traffic density (computed as incoming flow divided by population). The estimated paths are consistent with our knowledge and observation of the examined nodes and neighboring nodes, including local socio-economic features, the sizes of incoming infectious cases, and traffic flow intensities. For most of the nodes, the most influential nodes were located in the vicinity and were the primary sources of a large number of imported cases on April 12 and June 12. We argue that the proposed embedding approach can identify such paths in a much more effective manner than manually examining each OD pair and associated attributes. The estimated paths can thus be justified based on our understanding of local transmission dynamics.

We propose a hybrid approach that combines data-driven representative capabilities of graph embedding and simulation advantages of SIR models. We develop a dual SIR model that distinguishes intra- and inter-node susceptible population, thereby facilitating the estimation of historical transmission paths. This study contributes to spatial epidemiology by introducing an integrated data-driven learning model to capture the mobility-related spread mechanism of infectious diseases, utilizing multi-sourced mobility and attributed data: a case study of COVID-19 spread modeling and prediction in the state of Colorado, USA that demonstrates the applicability of the proposed approach in capturing space-time varied epidemic dynamics and supporting informed decision-making on disease intervention and prevention.

Different regions exert different effects during the transmission of viruses, which has been revealed by the "role analyses" in Section 3.4. Regions may assume different roles, such as transmission source hubs, intermediate and peripheral sink nodes. The heterogeneity of different regions is mainly accounted for by the proposed data-driven embedding method, which takes flow and socio-economic data as inputs and produces vectorized embeddings to represent heterogeneous transmission potential for each node. In addition, the developed SIR model is used to compute intra- and inter-node infectious numbers, which are based on region-specific socio-demographic,
COVID case, and flow data. Therefore, the heterogeneity of different regions has been modeled via the proposed approach that integrates graph embedding and SIR modeling.

The implications of this study are multi-fold. The transmission path maps can be used to promote our understanding of the spatio-temporal spread of the virus, including the most influential factors and the evolution of transmission patterns, thereby contributing to the development of appropriate and precise preventive measures. The findings are also beneficial for the evaluation of preventive measures at any specific spatial and temporal scale. The role analysis offers another informative tool to enable particular scrutiny of individual areas, which help identify the most critical high-risk nodes along the transmission routes and suggest role-targeting control strategies to achieve the most effective prevention with limited resources and response time. Informative decisions such as the estimation and allocation of necessary resources, enforcing control measures on specific roads and junctions, and public health messaging, can be made based on these maps.

Based on previous discussions, the advantages of the proposed integrated epidemiological modeling approach over existing related methods can be summarized as follows.

1. By accounting for transmission dynamics and local attributes via representation learning, our approach outperforms the classical compartment model in prediction performance.
2. Our approach enables more advanced visual analytics tasks such as transmission path mapping, decomposition of predicted cases, and role analysis of areal units, than the classical compartment model, which only make numerical predictions.
3. Through the integration of compartment models and machine learning models, our approach offers more powerful transmission modeling capabilities than some relevant visual analytics methods, such as the epidemic invasive tree map (Balcan et al., 2009) or source–sink risk density maps (Wesolowski et al., 2012).

Our transmission modeling and analysis results suggest that intervention measures (including travel restrictions, mask wearing and stay-at-home orders) have not been successful in the study region. Intra-node local transmissions are strongly affected by the levels of residents’ compliance with various measures, which are arguably correlated to socio-economic profiles of local residents and could have been partially modeled by the proposed approach.
approach. Explicit modeling of residents’ compliance with intervention measures may be challenging, since a large-scale survey is required. We will examine this possible improvement in future studies. Due to the challenges of data collection over the large study region, we omit the impacts of built-environment features, which are intuitively relevant to the spread of COVID-19. In the future, we will acquire the necessary data to compute built-environment features and build more fine-grained transmission models.

ACKNOWLEDGMENTS
This work was supported by Grant No. 41871308 from the National Natural Science Foundation of China and Grant No. 2019YFE0106500 from the National Key Research and Development Program of China (International Scientific & Technological Cooperation Program).

CONFLICT OF INTEREST
None.

DATA AVAILABILITY STATEMENT
The socio-demographic data that support the findings of this study are openly available at https://data-cdphe.opendata.arcgis.com/. The COVID-19 report data are from the Colorado Department of Public Health and Environment, Tri-county Health Department (https://www.tchd.org/823/COVID-19-Data and https://covid19.colorado.gov/data). Restrictions apply to the availability of TomTom Move data, which were used under license for this study.

ORCID
Tong Zhang https://orcid.org/0000-0002-0683-4669

REFERENCES
Arenas, A., Cota, W., Gmez-Gardenes, J., Gomez, S., Granell, C., Matamalas, J., ... Steinegger, B. (2020). A mathematical model for the spatiotemporal epidemic spreading of COVID 19. Retrieved from https://www.medrxiv.org/content/10.1101/2020.03.21.20040022v1
Balcan, D., Colizza, V., Goncalves, B., Hu, H., Ramasco, J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences of the United States of America, 106, 21484–21489. https://doi.org/10.1073/pnas.0906910106
Bertozi, A., Franco, E., Mohler, G., Short, M., & Sledge, D. (2020). The challenges of modeling and forecasting the spread of COVID-19. Proceedings of the National Academy of Sciences of the United States of America, 117, 16732–16738. https://doi.org/10.1073/pnas.2006520117
Buckee, C., Balsari, S., Chan, J., Cross, M., Dominici, F., Gasser, U., ... Schroeder, A. (2020). Aggregated mobility data could help fight COVID-19. Science, 368(6487), 145–146. https://doi.org/10.1126/science.abb8021
Chang, S., Pierson, E., Koh, P., Gerardin, J., Redbird, B., Grusky, B., & Leskovec, J. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. Nature, 589, 82–87. https://doi.org/10.1038/s41586-020-2923-3
Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science, 368(6489), 395–400. https://doi.org/10.1126/science.aba9757
Colizza, V., Barrat, A., Barthélemy, M., & Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. Proceedings of the National Academy of Sciences of the United States of America, 103, 2015–2020. https://doi.org/10.1073/pnas.0510525103
Dalziel, B., Pourbohloul, B., & Ellner, S. (2013). Human mobility patterns predict divergent epidemic dynamics among cities. Proceedings of The Royal Society B, 280, 20130763. https://doi.org/10.1098/rspb.2013.0763
Germann, T., Kadau, K., Longini, I., & Macken, C. (2006). Mitigation strategies for pandemic influenza in the United States. Proceedings of the National Academy of Sciences of the United States of America, 103, 5935–5940. https://doi.org/10.1073/pnas.0601266103
Goel, R., & Sharma, R. (2020). Mobility based SIR model pandemics: With case study of COVID-19. Preprint, arXiv:2004.13015.
Gomez, S., Diaz-Guilera, A., Gomez-Gardenes, J., Perez-Vicente, C., Moreno, Y., & Arenas, A. (2013). Diffusion dynamics on multiplex networks. Physical Review Letters, 110, 028701. https://doi.org/10.1103/PhysRevLett.110.028701
Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems, 151, 78–94. https://doi.org/10.1016/j.knosys.2018.03.022
Guimerà, R., & Amaral, L. (2005). Functional cartography of complex metabolic networks. *Nature*, 433, 895–900. https://doi.org/10.1038/nature3288

Guimerà, R., Mossa, S., Turtschi, A., & Amaral, L. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities’ global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 7794–7799. https://doi.org/10.1073/pnas.0407994102

Hao, X., Cheng, S., Wu, D., Wu, C., Lin, X., & Wang, C. (2020). Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature*, 584, 420–424. https://doi.org/10.1038/s41586-020-2554-8

Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars-Larrive, A., Loreto, V., ... Klimek, P. (2020). Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature Human Behaviour*, 4, 1303–1312. https://doi.org/10.1038/s41562-020-01009-0

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kim, J., Marks, F., & Clemens, J. (2021). Looking beyond COVID-19 vaccine phase 3 trials. *Nature Medicine*, 27, 205–211. https://doi.org/10.1038/s41591-021-01230-y

Kim, J., Marks, F., & Clemens, J. (2021). Looking beyond COVID-19 vaccine phase 3 trials. *Nature Medicine*, 27, 205–211. https://doi.org/10.1038/s41591-021-01230-y

Kipf, T., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Proceedings of the Fifth International Conference on Learning Representations*, Toulon, France (pp. 1–14).

Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., ... Tatem, A. J. (2020). Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature*, 585, 410–413. https://doi.org/10.1038/s41586-020-2293-x

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, BC, Canada (pp. 1–16).

Niehus, R., De Salazar, P., Taylor, A., & Lipsitch, M. (2020). Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. *The Lancet: Infectious Diseases*, 20, 803–808. https://doi.org/10.1016/S1473-3099(20)30229-2

Prasse, B., Achterberg, M., Ma, L., & van Mieghem, P. (2020). Network-based prediction of the 2019-nCoV epidemic outbreak in the Chinese province Hubei. Retrieved from https://arxiv.org/abs/2002.04482

Rockett, R. J., Arnott, A., Lam, C., Sadsad, R., Timms, V., Gray, K.-A., ... Sintchenko, V. (2020). Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature Medicine*, 26, 1398–1404. https://doi.org/10.1038/s41586-020-1000-7

Roos, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J., ... Chowell, G. (2020). Real-time forecasts of the 2019-nCoV epidemic in China from February 5th to February 24th, 2020. *Infectious Disease Modelling*, 5, 256–263. https://doi.org/10.1016/j.idm.2020.02.002

Senanayake, R., O’Callaghan, S., & Ramos, F. (2016). Predicting spatio-temporal propagation of seasonal influenza using variational Gaussian Process Regression. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, Portland, OR (pp. 3901–3907). Menlo Park, CA: AAAI.

Stoddard, S., Forshey, B., Morrison, A., Paz-Soldan, V., Vazquez-Prokopec, G., Astete, H., ... Scott, T. (2013). House-to-house human movement drives dengue virus transmission. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 994–999. https://doi.org/10.1073/pnas.1213349110

Tejedor, A., Longías, A., Foufoula-Georgiou, E., Georgiou, T., & Moreno, Y. (2018). Diffusion dynamics and optimal coupling in multiplex networks with directed layers. *Physical Review X*, 8, 031071. https://doi.org/10.1103/PhysRevX.8.031071

Teng, S. (2016). Scalable algorithms for data and network analysis. *Foundations and Trends in Theoretical Computer Science*, 12(1–2), 1–274. https://doi.org/10.1561/0400000051

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on Malaria. *Science*, 338(6104), 267–270. https://doi.org/10.1126/science.1223467

WHO. (2021). Coronavirus disease (COVID-19) dashboard. Retrieved from https://covid19.who.int/

Wu, J., Leung, K., & Leung, G. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *The Lancet*, 395, 689–697. https://doi.org/10.1016/S0140-6736(20)30260-9

Zhang, T., & Li, J. (2021). Understanding and predicting the spatio-temporal spread of COVID-19 via integrating diffusive graph embedding and compartmental models. *Transactions in GIS*, 25, 3025–3047. https://doi.org/10.1111/tgis.12803

How to cite this article: Zhang, T., & Li, J. (2021). Understanding and predicting the spatio-temporal spread of COVID-19 via integrating diffusive graph embedding and compartmental models. *Transactions in GIS*, 25, 3025–3047. https://doi.org/10.1111/tgis.12803