SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2020, Vol. 63 222203:1–222203:13
https://doi.org/10.1007/s11432-020-3034-y

Prediction of COVID-19 spread by sliding mSEIR observer

Duxin CHEN[1†], Yifan YANG[1†], Yifan ZHANG[2] & Wenwu YU[1*]

[1]*Jiangsu Key Laboratory of Networked Collective Intelligence, School of Mathematics, Southeast University, Nanjing* 210096, *China;*
[2]*School of Information Science and Engineering, Southeast University, Nanjing* 210096, *China*

Received 12 March 2020/Revised 8 June 2020/Accepted 1 August 2020/Published online 12 November 2020

Abstract The outbreak of COVID-19 has brought unprecedented challenges not only in China but also in the whole world. Thousands of people have lost their lives, and the social operating system has been affected seriously. Thus, it is urgent to study the determinants of the virus and the health conditions in specific populations and to reveal the strategies and measures in preventing the epidemic spread. In this study, we first adopt the long short-term memory algorithm to predict the infected population in China. However, it gives no interpretation of the dynamics of the spread process. Also the long-term prediction error is too large to be accepted. Thus, we introduce the susceptible-exposed-infected-removed (SEIR) model and further the metapopulation SEIR (mSEIR) model to capture the spread process of COVID-19. By using a sliding window algorithm, we suggest that the parameter estimation and the prediction of the SEIR populations are well performed. In addition, we conduct extensive numerical experiments to show the trend of the infected population for several provinces. The results may provide some insight into the research of epidemics and the understanding of the spread of the current COVID-19.

Keywords epidemic spread, prediction, sliding window algorithm, COVID-19

Citation Chen D X, Yang Y F, Zhang Y F, et al. Prediction of COVID-19 spread by sliding mSEIR observer. Sci China Inf Sci, 2020, 63(12): 222203, https://doi.org/10.1007/s11432-020-3034-y

1 Introduction

Human society has been severely influenced by epidemic outbreaks many times in history. The spread of the epidemic is full of accidental factors, which is a complex dynamical phenomenon [1, 2]. Infectious and infected individuals are all proactive individuals whose behaviors have a serious impact on the development of the disease, and they also adapt to changes in the external situation. This adds more complexity to the research on epidemic spreading dynamics and the prediction [3, 4]. With the rapid development of modern transportation and communication tools, the human activity radius has grown rapidly, and the contact between any two individuals has become possible. However, the development of transportation and communication tools has brought convenience yet also brought certain disasters. On the one hand, the small-world network feature of our living society makes the radius of the spread of diseases or rumors bigger and bigger, which makes it easier to spread diseases [5,6]. On the other hand, the real society network has a high clustering coefficient. Concerning disease spread, contact between families and communities accelerates the speed of transmission and expands the scope of transmission [7, 8]. As human beings are facing long-term and severe threats of various diseases, such as H1N1, H7N9,

* Corresponding author (email: wwyu@seu.edu.cn)
† Duxin CHEN and Yifan YANG have the same contribution to this work.

and the current COVID-19, this has seriously affected the lives of our life and the national economy over the recent years [9–11]. Thus, the study of the mechanism of epidemic spreading is urgent and important, which would be the basis for intervening in the epidemic and also an important way to suppress the epidemic. Recently, many epidemiological models have been proposed to reveal the spreading dynamics of virus and disease in different population structures, such as the compartment models [12] for small-scale, well-mixed populations, and the network epidemiology models [13, 14] for individuals with complex contact relationships in a regional population. For epidemic spread in large-scale spatial regions, the most widely adopted model is the metapopulation model, which refers to a group of separated subspecies of the same species connected by an interacting network. Concerning large-scale epidemic outbreaks, such as the global spread of influenza, the dynamics can be simulated for the transmission of pathogens through a metapopulation network [15], in which the cities in different countries are described as subpopulations, and human flows between cities are modeled as edges connecting subpopulations. The metapopulation model has been successful in the study of large-scale pandemic spread. For instance, previous studies [16, 17] used the worldwide cite global aviation networks to analyze the spread of SARS and H1N1 in the global urban population. The data of the mobility network of mobile phone users are investigated [18] to analyze the malaria transmission process in Kenya. However, because the access to the detailed flow data for all cities in the world and even in a country is usually impossible in practice, most of these previous studies can only use coarse-grained flow data to establish epidemic transmission networks between cities, under the assumption that all contacts and infections between individuals are homogeneous in the same city. With the fast development of metropolis around the world, the social structure within the city becomes more and more complex, and the assumption of homogeneous mixing of the population within the city is no longer valid. Also, it is not clear whether the physical network can approximate the general infection network of all different diseases, which further limits the application value of the exhaustive existing methods. Therefore, the method that can achieve fine-grained urban infectious epidemic spread analysis without the need for detailed empirical data on resident mobility is still ideal. Only if the detailed empirical data are open to the researchers, more valid and deep research results can be obtained. Deterministic models based on differential equations are difficult to systematically simulate the virus propagation process effectively in both cyberspace and real physical world [19–21]. In recent years, thanks to the fast development of AI techniques and the network science tools, revealing the evolutionary rules of such complex systems has become traceable. For instance, Fraser et al. [22] judged the pandemic potential of H1N1 with limited data and made an early assessment of transmissibility and severity of the outbreak of international spread, and viral genetic diversity. Wang et al. [23] proposed a network inference model, which can reduce the individual-based network into a sub-population network without loss of information, and incorporated the power-law distribution prior and data prior for better performance. It extended the widely accepted susceptible-infected-removed (SIR) model to a metapopulation SIR model. Concerning cybersecurity, the previous study [24] investigated the identification problem of potential malware propagation via the Internet, which proposed a tree-shaped deep neural network and performed flow detection on real data with the imbalanced distribution. Moreover, to uncover the dynamical mechanisms of the propagations of diseases, many recent researches have been conducted to predict and control epidemic spreading in social systems via a network analysis strategy [25–29]. As we know, large-scale population movements activated long-range links in the network, making COVID-19 rapidly spread from Wuhan to other cities across China. Therefore, we aim to study the dynamics and prediction of the trend of COVID-19 and to provide some understanding of the disaster. In this paper, we first test and verify that the current prediction method based on a neural network can predict the trend of the evolution of the different types of population. However, it gives no interpretation, and the accuracy becomes lower for the long-term prediction. Thus, we introduce the widely accepted susceptible-exposed-infected-removed (SEIR) model and identify the parameters therein to fit and predict future trends. The prediction error is much lower than the neural network-based method. Furthermore, we extend and propose a metapopulation SEIR (mSEIR) model with consideration of the mobility and interaction of the urban population in different regions. We suggest that by using the sliding window method based on the proposed mSEIR model, we can predict the inflection point and the end of the

COVID-19. Throughout this paper, considering the real spread of the COVID-19 is determined by multiple factors, we adopt the following assumptions to simplify the study, but without loss of generality.

(1) It is assumed that the transmission of the virus happens in an enclosed environment, regardless of natural birth rate and natural mortality, which means the total population is constant.

(2) The reported cases of confirmed, cured, and dead data are accurate.

(3) Patients are not infectious during the incubation period, and there are no super-spreaders.

(4) Healers will be able to produce antibodies, which prevent them from getting infected again.

2 SEIR model

In this section, we introduce the basic SEIR model [30] that is adopted to describe the recent outbreak of COVID-19 in China. Let the total population be N, and we use $S(t)$ to represent the susceptible population, $E(t)$ to represent the exposed population, $I(t)$ to represent the infected population, and $R(t)$ to be the removed population, which consists of those who are healed or dead owing to the COVID-19. Let the time-dependent variables be simplified without index t, and the dynamics of the model is described as follows:

$$\begin{cases} \dfrac{\mathrm{d}S}{\mathrm{d}t} = -\dfrac{\beta SI}{N}, \\ \dfrac{\mathrm{d}E}{\mathrm{d}t} = \dfrac{\beta SI}{N} - \sigma E, \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I, \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} = \gamma I, \end{cases} \tag{1}$$

where $N = S(t) + E(t) + I(t) + R(t)$, and the contact rate β, the incubation rate σ, the recovery rate γ are model parameters to be determined.

The discrete format of (1) is shown as

$$\begin{cases} S_{t+\triangle t} = S_t - \dfrac{\beta S_t I_t}{N} \cdot \triangle t, \\ E_{t+\triangle t} = E_t + \left(\dfrac{\beta S_t I_t}{N} - \sigma E \right) \cdot \triangle t, \\ I_{t+\triangle t} = I_t + (\sigma E_t - \gamma I_t) \cdot \triangle t, \\ R_{t+\triangle t} = R_t + \gamma I_t \cdot \triangle t. \end{cases} \tag{2}$$

3 Prediction of infection cases via LSTM

A basic idea is to predict the evolution of these variables by using the popular deep learning technique [31]. In this case, the dynamics of the SEIR model is not required, since we can roughly predict the curves by choosing a proper structure of the neural network. Long short-term memory (LSTM) network [32], which is a special recurrent neural network, is designed to solve the problem of long dependencies. In this part, we simply predict the cases of infection with LSTM to gain some cognition of the trend of the COVID-19. The input X of LSTM is the data of 5 consecutive days, $X = \{x(t), x(t+1), \ldots, x(t+4)\}$, where $x(t) = \{S(t), E(t), I(t), R(t)\}$ and $S(t)$, $E(t)$, $I(t)$, $R(t)$ denote the number of susceptible, exposed, infected and recovered individuals of day t, respectively. The output is the prediction value of $\{S, E, I, R\}$ for the 6th day.

The construction of our model is shown in Figure 1. The input data X goes through LSTM layer with 16 hidden units, and the output of the last LSTM cell passes through a fully connected layer to yield the four-dimensional prediction value $\{S, E, I, R\}$ for the next day.

Prior to the training of the model, min-max normalization is applied to the data to constrain the value of the data between 0 and 1. The data of recent 10 and 20 days are split as the test set in prediction,

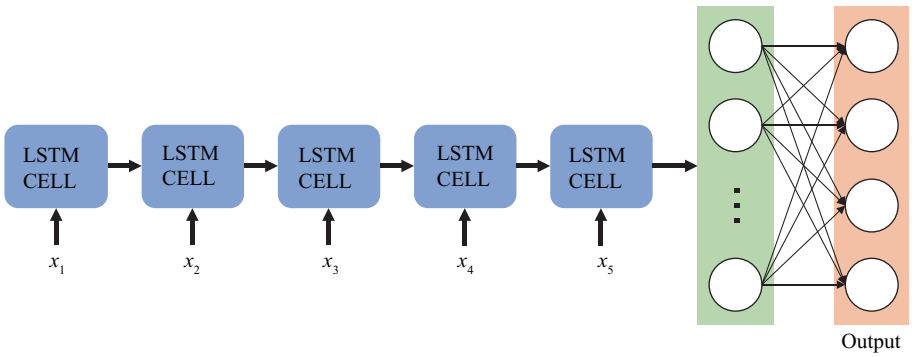

Figure 1 (Color online) The structure of LSTM model.

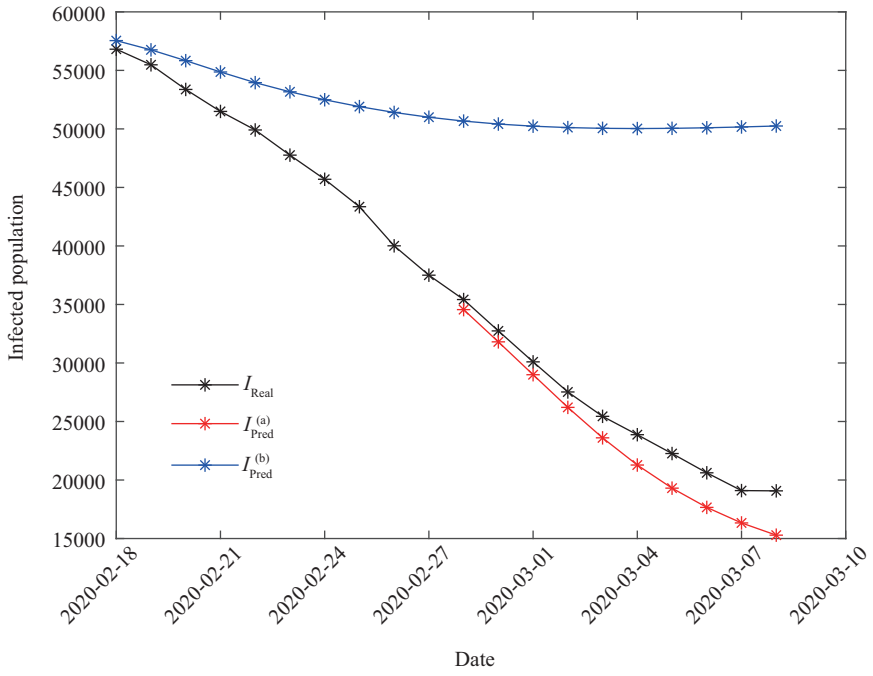

Figure 2 (Color online) The prediction of infected population via LSTM, where $I^{(a)}_{\text{Pred}}$ and $I^{(b)}_{\text{Pred}}$ denote the prediction of 10 days and 20 days, respectively.

while the rest is for the training set. We choose mean square error (MSE), MSE $= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ where $\hat{y}_i$ denotes the estimation of a data point y_i, as the loss function. We use Adam optimizer and set the learning rate as 0.001. The LSTM model is trained for 1000 epochs and the batch size is 16.

For evaluation, the model is used to predict the infection cases of the test set (days after the training set). The prediction is made in an iterative method. We use the last 5-day's data in the training set to predict the first day in the test set, and then use the last 4-day's data in the training set and the prediction of the first day in the test set to predict the second day in the test set and so on. We introduce the mean absolute percentage error (MAPE), MAPE $= \frac{100\%}{n}\sum_{i=1}^{n}|\frac{\hat{y}_i - y_i}{y_i}|$ with $\hat{y}_i$ denoting the estimation of y_i, to quantify the prediction accuracy. The MAPE values for the prediction of 10 and 20 days are 10.8% and 30.0% on average, respectively. The curve of the prediction on test set is shown in Figure 2. Clearly, it can be observed that the LSTM technique may predict the short-term trend with a high accuracy, while it fails to predict the long-term evolution.

4 Parameter identification and numerical simulation

4.1 Identification of the SEIR parameters

Although by using the LSTM tool, we may predict the trend of the evolution of these types of population, it gives no interpretation on the dynamics of the epidemic spreading process. Moreover, LSTM is not suitable to predict the long-term evolution of the infected case owing to the large value of MAPE. Thus, we further seek to identify the parameters in the SEIR model to test whether it is suitable to describe the virus propagation. Suppose we have four-dimensional time series data $X = \{S(t), E(t), I(t), R(t)\}$ generated by the SEIR model. X can be rewritten as $X_{t+\triangle t} = \Phi(X_t)$, which indicates that the state of $X_{t+\triangle t}$ can be obtained based on the state of X_t, if $\Phi(\cdot)$ is known.

Our objective function is as follows:

$$\max_{\theta_{i-1}}\{P(\hat{X}(t+\triangle t) = X(t+\triangle t)|X(t))\}. \tag{3}$$

Therein, $\hat{X}$ denotes the estimation of X. We can fit the data based on the discrete SEIR model mentioned above. However, in the SEIR model, the number of exposed individuals $E(t)$ is not exactly equal to the number of reported suspected cases. Therefore, the problem turns into identifying four parameters $\Theta = \{E(0), \beta, \sigma, \gamma\}$, and the optimization objective function is transformed into

$$\min_{\Theta}\max_{\forall i}\left\{\frac{I(t_i,\Theta) - \hat{I}(t_i,\Theta)}{I(t_i,\Theta)}, \frac{R(t_i,\Theta) - \hat{R}(t_i,\Theta)}{R(t_i,\Theta)}\right\}, \tag{4}$$

where I and R denote published data of confirmed and recovered cases including deaths, respectively. As only small values of I and R are reported in the early stage, we choose the maximum relative error of I and R as the objective function, and minimize it to obtain the parameters' values.

In order to identify unknown parameters of the SEIR model and capture the initial value of the exposed number $E(0)$, we propose an online parameter identification method based on the idea of sliding window, and make the method forgetful to the past information by using the dynamic window size. Meanwhile, the intensity of forgetfulness can be adjusted by setting the size of the last window.

To identify the parameters, $\hat{E}_1, \hat{S}_1, I_1, R_1, \theta_0$ should be initialized at first. Then, the predicted values in the sliding window are calculated based on the initial parameters, and use the trust region reflective algorithm to minimize the maximum relative error in order to obtain the suboptimal parameter θ_1. Next, the window slides to the second variables $\hat{E}_2, \hat{S}_2, I_2, R_2$ and uses their values and the initial value of parameter θ_1 to predict x_3, x_4, x_5 in the sliding window. Repeat the optimization method of the first step, from which we can get the prediction of $\hat{S}_i, \hat{E}_i$ and the parameter θ_{i-1} that optimize the prediction results. Repeat the above steps until there are no more known I_i, R_i on the right of the sliding window. Here, we can use the predicted values of $\hat{E}_i, \hat{S}_i$ and the true data I_i, R_i in the window with parameter θ_{i-1} to predict the variables on the right of the sliding window. However, this ignores the variable information in the window. Therefore, in order to make full use of limited time-series information, we choose to constrain the window to a smaller value and make the next prediction based on the information in the new window to update the variables and parameters. The advantage of this method is that the prediction of the future can be based on the newer information and the prediction function has a robust prediction ability.

The objective function of each sliding window is defined as

$$\max_{\Theta}\{P(\hat{X}_j = X_j|X_i(t)), X_j \in \text{Window}\}. \tag{5}$$

The illustration of our method is shown in Figure 3.

Since the parameters change smoothly and slowly, the accumulating influence of the previous data on the model gradually attenuates as the intensity of forgetfulness increases. Also, for the addition of new data, the method can dynamically adjust the parameters of SEIR based on previous information to modify the future prediction of the model, which not only enhances the prediction ability, but reduces computation complexity as well.

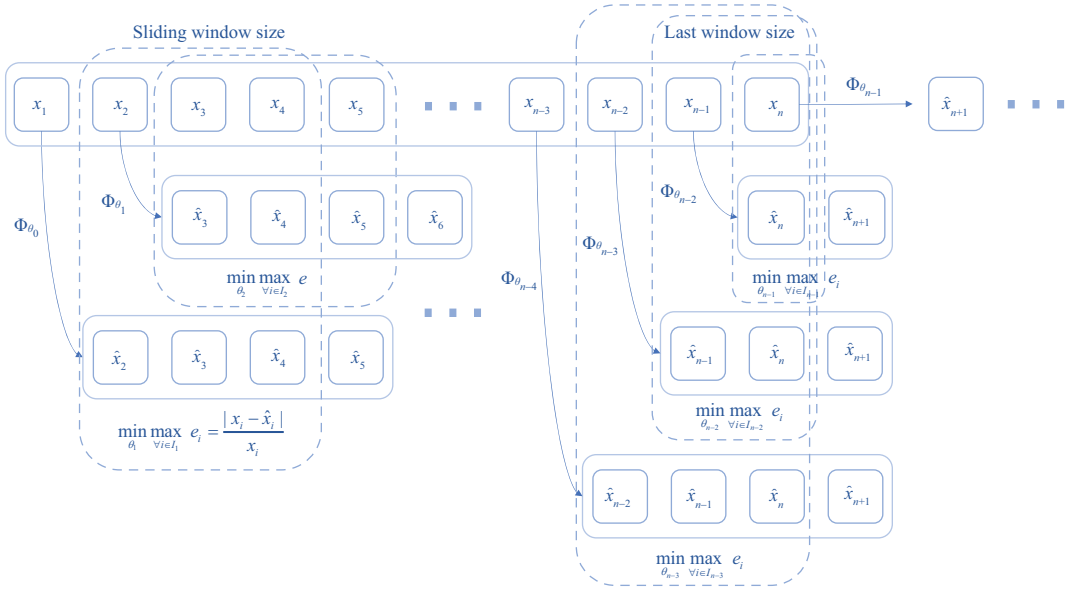

Figure 3 (Color online) The structure of the sliding window method. Therein, I_i denotes the ith window.

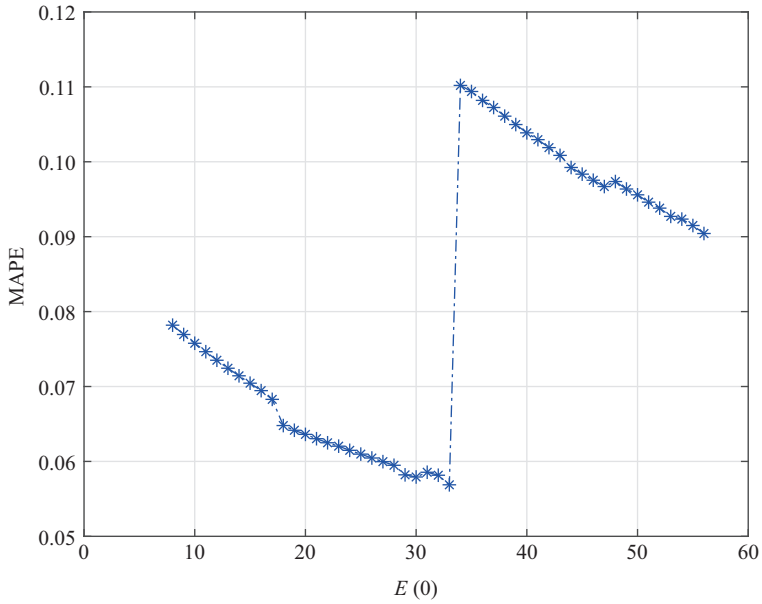

Figure 4 (Color online) MAPE with different values of $E(0)$ of the test set.

4.2 Numerical experiments of the SEIR model

Based on the proposed method above, we conducted numerical experiments with the COVID-19 data in China. We set the tunable parameter sliding window size as 7, the last window size as 3 and initialize the parameters to be identified between 0 and 1. Note that, generally, with a larger window size, the sensitivity of the prediction becomes weak. However, a smaller window size will result in an overfitting prediction.

The training data set of the country and some provinces starts from January 10th to February 16, and the test data set starts from February 17 to March 8, both of which are composed of numbers of the current confirmed cases, recovered patients and death. The estimated exposed case $\hat{E}_1$ is initialized between 8 and 56 when the step size is 1, and the population N is 1400000000. The optimization is accomplished with the trust region reflective algorithm [33].

For evaluation, the model is used to predict the infection cases of the test set analogous to LSTM. We

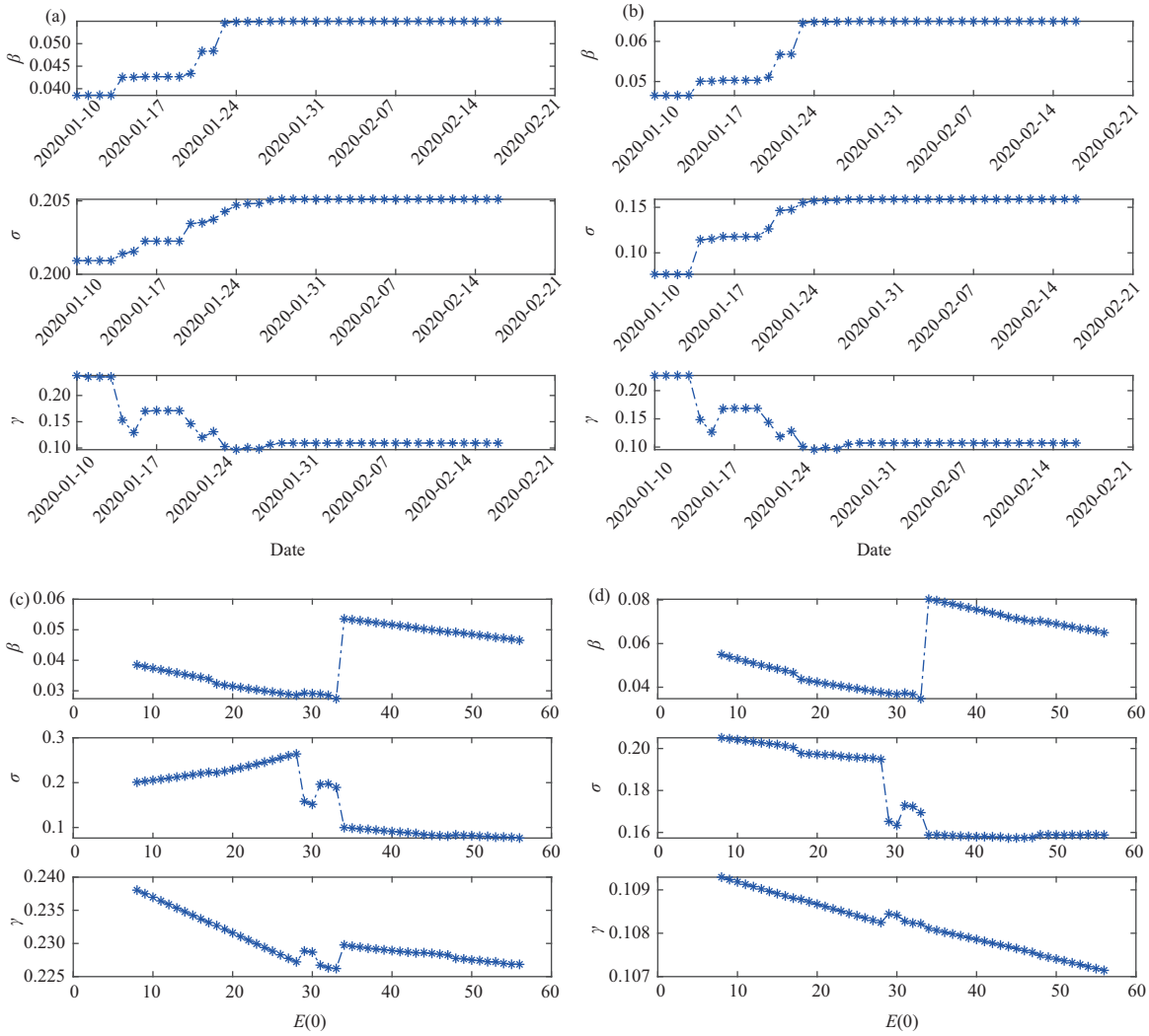

Figure 5 (Color online) Parameter identification with different values of $E(0)$ on different days. (a) $E(0) = 8$; (b) $E(0) = 56$; (c) parameters for January 10; (d) parameters for February 16.

use the data and parameter θ in the training set to predict the situation of the following 21 days. The MAPE of the 21-day prediction is 8.15% on average, as shown in Figure 4. The parameter identification results of different days and different values of $E(0)$ on training set are shown in Figure 5.

Then, we set the sliding window size and the last window size as 2, to make the most use of the data. We analyze the situation of China and make prediction by our sliding window model. For comparison, the curves of the prediction on the training set, the test set, and the next 100 days from March 17 are shown in Figure 6. The MAPE for the prediction of 21 days is 3.88%. The forecast result suggests that we can predict the inflection point of the curve and COVID-19 will cease in the end of April in China.

4.3 Modified SEIR model

Next, we consider the case that patients are infectious during the incubation period, and the contact incubation rate β_2 during the incubation period is relatively lower than the contact infection rate β_1 during the infectious period, since 12.6% of the reports indicated that pre-symptomatic transmission existed [34]. Thus, the susceptible population has the possibility to be transferred into exposed population

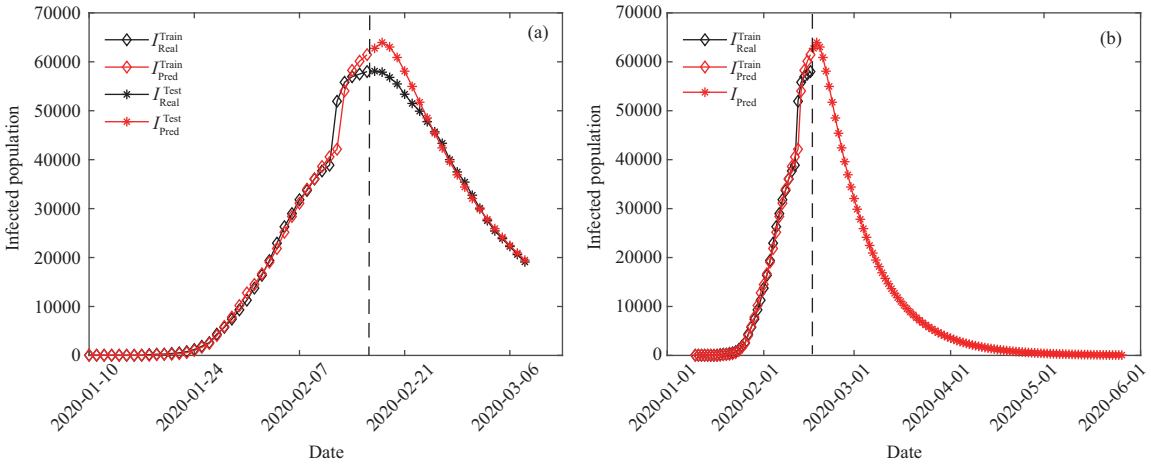

Figure 6 (Color online) Prediction of the infected population. On the left part of the dashed straight line, we show the training of the data, whereas on the right part, we give the prediction results of the test data. (a) Prediction of the training set and test set; (b) prediction of the next 100 days from March 17.

after the contact. The dynamics of the modified SEIR model is described as follows:

$$\begin{cases} \dfrac{\mathrm{d}S}{\mathrm{d}t} = -\dfrac{\beta_1 SI}{N} - \dfrac{\beta_2 SE}{N}, \\ \dfrac{\mathrm{d}E}{\mathrm{d}t} = \dfrac{\beta_1 SI}{N} + \dfrac{\beta_2 SE}{N} - \sigma E, \\ \dfrac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I, \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} = \gamma I, \end{cases} \tag{6}$$

where $N = S(t) + E(t) + I(t) + R(t)$, and the contact infection rate β_1, the contact incubation rate β_2, the incubation rate σ, and the recovery rate γ are model parameters to be determined.

In order to test the difference of the modified SEIR model considering the infectivity during incubation period on the prediction results, we choose different training data sets and compare the results with that obtained by the traditional SEIR model.

We select the data set ranging from January 10 to March 8, where the training data sets are from January 10 to February 13, February 14, February 15, and February 16, respectively, and the data from February 17 to March 8 belong to the test set. We initialize the tunable parameter of sliding window size and the last window size as 6 and 3, respectively, and initialize the model parameters to be identified between 0 and 1. We perform parameter identification on the training data sets based on the traditional SEIR model and the modified SEIR model, and calculate the MAPE values on the test data set. The results are recorded in Table 1 and shown in Figure 7. Note that, subscript 1 represents the traditional SEIR model, and subscript 2 represents the modified SEIR model. For instance, MAPE_1 refers to the results of the traditional SEIR model, and MAPE_2 stands for the result of the modified SEIR model.

It can be observed that as the training data set becomes larger, MAPE values gradually decrease and then indicate an accurate predicting result. At the beginning, the prediction results of the modified SEIR model are better than the traditional SEIR model, but the increase of the number of parameters is more likely to cause overfitting, resulting in poor prediction results sometimes. In order to simplify the model parameters, it can be observed that our assumption that the incubation period is not infectious will not have much impact on the prediction results.

Table 1 Prediction of the infected population on different training data sets

Training data set	Test data set	MAPE_1 (%)	MAPE_2 (%)
01-10–02-13 (35 d)	02-17–03-08 (21 d)	5.5004	5.4720
01-10–02-14 (36 d)	02-17–03-08 (21 d)	4.2543	2.1176
01-10–02-15 (37 d)	02-17–03-08 (21 d)	2.0342	2.1118
01-10–02-16 (38 d)	02-17–03-08 (21 d)	2.3528	2.6669

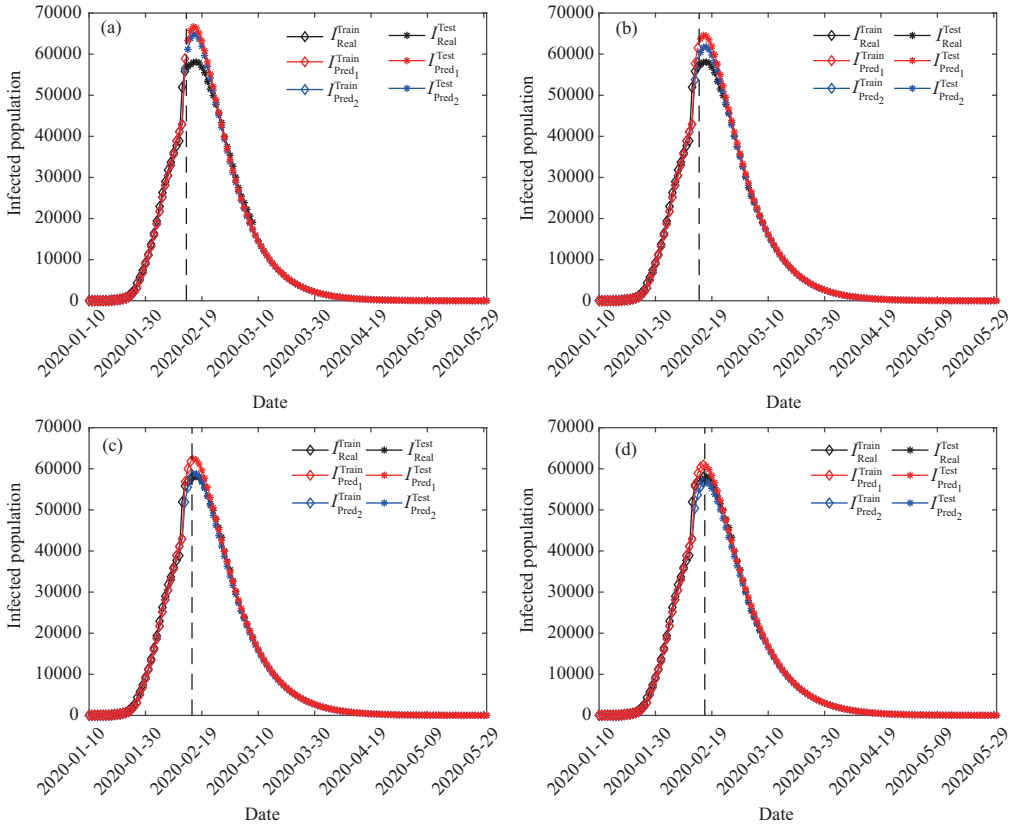

Figure 7 (Color online) Prediction of the infected population based on the traditional SEIR model and the modified SEIR model on different training data sets: (a) from January 10 to February 13, (b) from January 10 to February 14, (c) from January 10 to February 15, and (d) from January 10 to February 16.

5 mSEIR model and numerical simulation

5.1 mSEIR model description

Although SEIR model captures the dynamics of the epidemic spreading process to a great degree, it cannot reveal the social mobility and contact of different regions. As we know, the social network has a small-world feature, which means the contact of human beings is frequent and may accelerate the spread of diseases. It would be more important to consider the interaction of different regions within the large social network. Thus, we extend the SEIR model into the following mSEIR model:

$$\begin{cases} \dfrac{\mathrm{d}S_n}{\mathrm{d}t} = -\dfrac{\beta S_n \sum_{m=1}^{N}(\frac{h_{mn}}{P_m} + \frac{h_{nm}}{P_n})I_m}{P_n}, \\ \dfrac{\mathrm{d}E_n}{\mathrm{d}t} = \dfrac{\beta S_n \sum_{m=1}^{N}(\frac{h_{mn}}{P_m} + \frac{h_{nm}}{P_n})I_m}{P_n} - \sigma E_n, \\ \dfrac{\mathrm{d}I_n}{\mathrm{d}t} = \sigma E_n - \gamma I_n, \\ \dfrac{\mathrm{d}R_n}{\mathrm{d}t} = \gamma I_n, \end{cases} \tag{7}$$

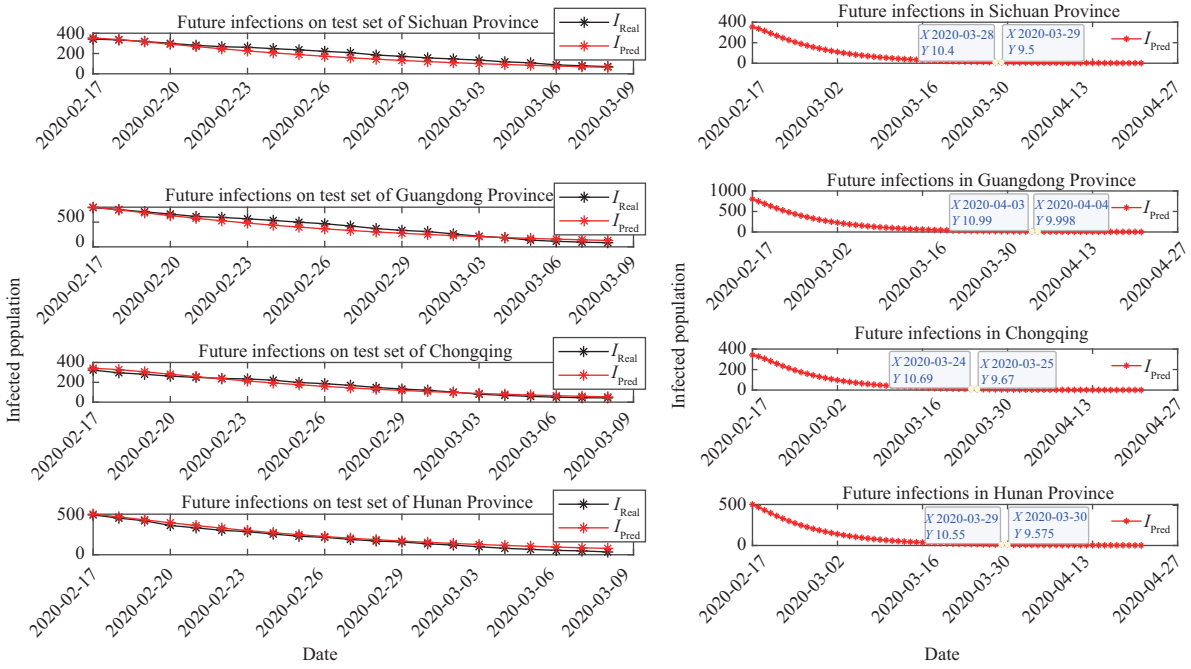

Figure 8 (Color online) Prediction of the infected populations in provinces around Hubei.

where we use P_n and P_m to denote the total population in the n-th and m-th subregion of a considered society, respectively. The variables S_n, E_n, I_n, R_n represent the population of the subregion n within the total numbers of S, E, I, R. Between two subregions n and m, the interaction strength is defined as h_{nm}, which may be correlated to the average population flow from n to m. Note that $h_{nm} \neq h_{mn}$, because the population flow from subregion n to m may be different from that from subregion m to n.

In the mSEIR model, susceptible crowd in the n-th subregion may convert to the exposed in three potential ways as follows:

(1) The internal propagation in subregion n: $\beta \cdot S_n I_n$.

(2) The infected people in subregion m with the number of $\frac{h_{mn}}{P_m} \cdot I_m$ come to subregion n with the probability of β to spread to the susceptible person S_n.

(3) The susceptible people in subregion n with the number of $\frac{h_{nm}}{P_n} \cdot S_n$ come to subregion m, and are infected with a probability of β by the infected people I_m.

5.2 Numerical experiments of the mSEIR model

We take the provinces adjacent to Hubei Province and the Yangtze River Delta as examples for analysis and prediction. The training data set starts from January 10 to February 16, and the test data set starts from February 17 to March 8, both of which are composed of the numbers of confirmed cases, recovered patients, and the death. We set both sliding window size and the last window size as 2, initialize $E(0)$ of each province based on the proportion of confirmed cases in each province i multiplied by the exposed cases in China, i.e., $E_i(0) = (I_i(0) \div I_{\text{China}}(0)) \cdot E_{\text{China}}(0)$. The population of each province is the resident population of provinces surveyed in 2019. To reduce the number of parameters, we let $\frac{h_{nn}}{P_n} = \frac{1}{2}$ and $g_{nm} = \beta(\frac{h_{mn}}{P_m} + \frac{h_{nm}}{P_n}), \forall n, m$, and this will lead to a reduction of $\frac{N(N-1)}{2}$ parameters in the total interaction matrix for a considered region. The optimization is accomplished with the trust region reflective algorithm [33].

The MAPE values of the first 10-day prediction in Sichuan, Guangdong, Chongqing, Hunan are 9.29%, 10.20%, 8.50% and 5.97%, respectively, and the MAPE values of the last 10-day prediction in Sichuan, Guangdong, Chongqing, Hunan are 21.32%, 21.11%, 16.56% and 34.17%, respectively. The prediction result shows that COVID-19 in these area will cease in the end of March, as shown in Figure 8.

The MAPE values of the first 10-day prediction in Zhejiang, Jiangsu, Shanghai are 13.95%, 5.26% and

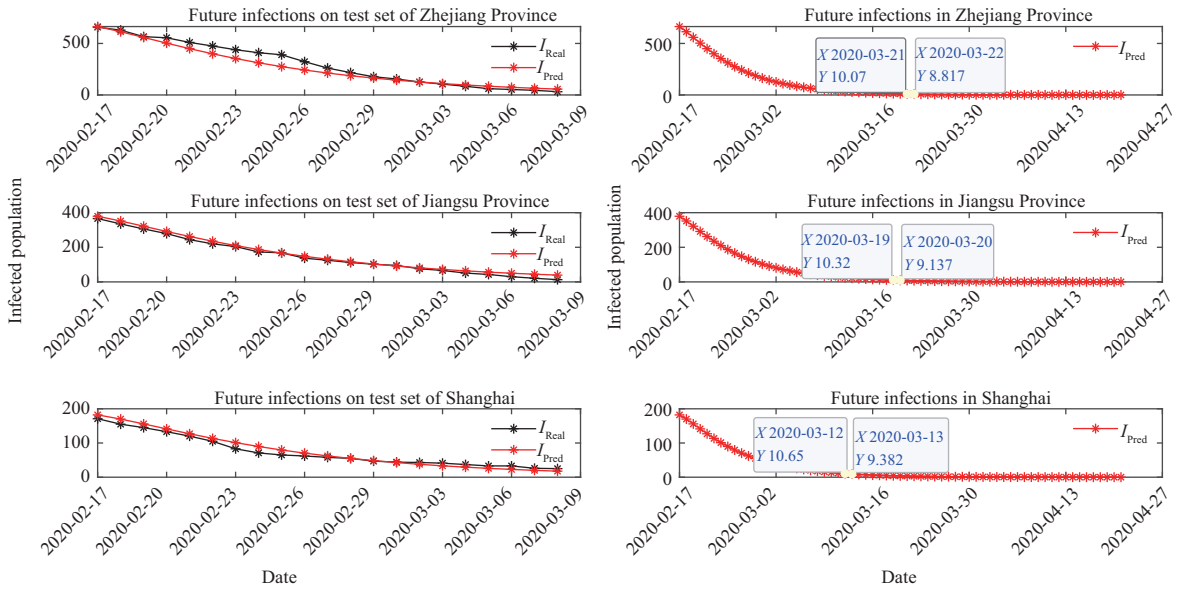

Figure 9 (Color online) Prediction of the infected populations in provinces in Yangtze River Delta.

12.80%, respectively, and the MAPE values of the last 10-day prediction in Zhejiang, Jiangsu, Shanghai are 19.06%, 26.36% and 14.07%, respectively. Although in the last 10 days, MAPE values are larger, this is due to the fact that the real populations are small, it will indeed turn out that the MSE values will be small. The prediction result shows that COVID-19 in these area will cease in March, as shown in Figure 9.

6 Conclusion

The new type of coronavirus spread has been epidemic for months at home and abroad, which has caused a huge and painful loss of health and life. Previous deterministic models in the format of differential equations are difficult to systematically describe the virus propagation process effectively because the real circumstances have more unexpected influential factors. Thanks to the development of artificial intelligence techniques and the network science tools, revealing the evolutionary rules of such complex systems has become traceable. To investigate the dynamics of COVID-19 spread and help people gain some understanding on the evolution trend, we first use an LSTM model to predict the infected population in China. The MAPE results have shown that it is accurate for short-term prediction; however, it fails to predict the long-term evolution. Moreover, because it gives no interpretation of the strategy of dynamics of the spread process, we introduce the widely accepted SEIR model to describe the epidemic spread process. By using the proposed sliding window method, we identify the parameters in the SEIR model and predict the trend of the evolution of COVID-19. We test the assumption that patients are infectious during the incubation period and introduce another contact infection rate in the modified SEIR model for comparison. It is observed that by introducing more parameters, the predicting results may be better; thus, it would be easier to cause overfitting. Because the infectious number of patients during the incubation period is quite small, the results of both the SEIR and modified SEIR models do not have much difference. The results suggest that the prediction error is quite small, and the proposed method may help predict the end of the epidemic disease. However, the typical SEIR model takes no consideration of the mobility of the urban population and the interaction of inter-city population flow. We then extend the SEIR model and propose the mSEIR model to describe the spread process of COVID-19. We use the sliding window algorithm for parameter estimation in the model and further predict the infected populations to obtain the trend. Numerical experiments have been conducted to show the evolution curves of the infected population for several provinces. With a proper selection of the parameters in the

sliding window prediction, we suggest that we may predict the trend of the populations accurately in different cases. By using a small window size in the method, we may sensitively acquire the inflection point of the trend of COVID-19 and predict the end of COVID-19 for different provinces and the entire country. Although intervention and regulation by our government are effective in preventing the disaster, they may also bring difficulty in modeling the natural dynamics of epidemic spread and prediction. In short, the real spreading process of COVID-19 shall be much more complicated, which cannot be entirely captured by the limited data and any theoretical model. In future work, we will further investigate the general dynamics of COVID-19 and the general spread strategies of more epidemic viruses.

Acknowledgements This work was supported by Fundamental Research Funds for the Central Universities (Grant No. 2242019K40111), National Natural Science Foundation of China (Grant Nos. 61903079, 61673107), and Jiangsu Provincial Key Laboratory of Networked Collective Intelligence (Grant No. BM2017002).

References

1 Merrell D S, Butler S M, Qadri F, et al. Host-induced epidemic spread of the cholera bacterium. Nature, 2002, 417: 642–645
2 Lopman B, Vennema H, Kohli E, et al. Increase in viral gastroenteritis outbreaks in Europe and epidemic spread of new norovirus variant. Lancet, 2004, 363: 682–688
3 Newman M E J. Spread of epidemic disease on networks. Phys Rev E, 2002, 66: 016128
4 Barthélemy M, Barrat A, Pastor-Satorras R, et al. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. Phys Rev Lett, 2004, 92: 178701
5 Gang Y, Tao Z, Jie W, et al. Epidemic spread in weighted scale-free networks. Chin Phys Lett, 2005, 22: 510–513
6 Keeling M J, Eames K T D. Networks and epidemic models. J R Soc Interface, 2005, 2: 295–307
7 Lee H W, Malik N, Shi F, et al. Social clustering in epidemic spread on coevolving networks. Phys Rev E, 2019, 99: 062301
8 Ratmann O, Grabowski M K, Hall M, et al. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. Nature Commun, 2019, 10: 1–13
9 Funk S, Gilad E, Watkins C, et al. The spread of awareness and its impact on epidemic outbreaks. Proc Natl Acad Sci USA, 2009, 106: 6872–6877
10 He M, Miyajima F, Roberts P, et al. Emergence and global spread of epidemic healthcare-associated Clostridium difficile. Nat Genet, 2013, 45: 109–113
11 Valdano E, Fiorentin M R, Poletto C, et al. Epidemic threshold in continuous-time evolving networks. Phys Rev Lett, 2018, 120: 068302
12 Angstmann C N, Erickson A M, Henry B I, et al. Fractional order compartment models. SIAM J Appl Math, 2017, 77: 430–446
13 Galea S, Riddle M, Kaplan G A. Causal thinking and complex system approaches in epidemiology. Int J Epidemiol, 2010, 39: 97–106
14 Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. Science, 2013, 342: 1337–1342
15 Arino J, Ducrot A, Zongo P. A metapopulation model for malaria with transmission-blocking partial immunity in hosts. J Math Biol, 2012, 64: 423–448
16 Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. Proc Natl Acad Sci USA, 2004, 101: 15124–15129
17 Bajardi P, Poletto C, Ramasco J J, et al. Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. Plos One, 2011, 6: e16591
18 Wesolowski A, Eagle N, Tatem A J, et al. Quantifying the impact of human mobility on malaria. Science, 2012, 338: 267–270
19 Gao C, Liu J. Modeling and restraining mobile virus propagation. IEEE Trans Mobile Comput, 2013, 12: 529–541
20 Wang X, Ni W, Zheng K, et al. Virus propagation modeling and convergence analysis in large-scale networks. IEEE Trans Inform Forensic Secur, 2016, 11: 2241–2254
21 Li Q, Brass A L, Ng A, et al. A genome-wide genetic screen for host factors required for hepatitis C virus propagation. Proc Natl Acad Sci USA, 2009, 106: 16410–16415
22 Fraser C, Donnelly C A, Cauchemez S, et al. Pandemic potential of a strain of influenza a (H1N1): early findings. Science, 2009, 324: 1557–1561
23 Wang J, Wang X, Wu J. Inferring metapopulation propagation network for intra-city epidemic control and prevention. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018. 830–838

24 Chen Y C, Li Y J, Tseng A, et al. Deep learning for malicious flow detection. In: Proceedings of 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017. 1–7

25 Wang W, Liu Q H, Liang J, et al. Coevolution spreading in complex networks. Phys Rep, 2019, 820: 1–51

26 Ogura M, Mei W, Sugimoto K. Synergistic effects in networked epidemic spreading dynamics. IEEE Trans Circuits Syst II, 2020, 67: 496–500

27 Chen S, Small M, Fu X. Global stability of epidemic models with imperfect vaccination and quarantine on scale-free networks. IEEE Trans Netw Sci Eng, 2020, 7: 1583–1596

28 Koher A, Lentz H H K, Gleeson J P, et al. Contact-based model for epidemic spreading on temporal networks. Phys Rev X, 2019, 9: 031017

29 Chang L, Duan M, Sun G, et al. Cross-diffusion-induced patterns in an SIR epidemic model on complex networks. Chaos, 2020, 30: 013147

30 Li M Y, Muldowney J S. Global stability for the SEIR model in epidemiology. Math Biosci, 1995, 125: 155–164

31 Yang Z, Zeng Z, Wang K, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis, 2020, 12: 165–174

32 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

33 Yuan Y. Recent advances in trust region algorithms. Math Program, 2015, 151: 249–281

34 Du Z W, Xu X K, Wu Y, et al. The serial interval of COVID-19 from publicly reported confirmed cases. medRxiv, 2020. doi: 10.1101/2020.02.19.20025452