Network-inference-based prediction of the COVID-19 epidemic outbreak in the Chinese province Hubei

Prasse, Bastian; Achterberg, Massimo A.; Ma, Long; Van Mieghem, Piet

DOI
10.1007/s41109-020-00274-2

Publication date
2020

Document Version
Final published version

Published in
Applied Network Science

Citation (APA)
Prasse, B., Achterberg, M. A., Ma, L., & Van Mieghem, P. (2020). Network-inference-based prediction of the COVID-19 epidemic outbreak in the Chinese province Hubei. *Applied Network Science*, *5*(1), 1-11. Article 35. https://doi.org/10.1007/s41109-020-00274-2

Prasse *et al. Applied Network Science* (2020) 5:35
https://doi.org/10.1007/s41109-020-00274-2

Applied Network Science

RESEARCH Open Access

Network-inference-based prediction of the COVID-19 epidemic outbreak in the Chinese province Hubei

Bastian Prasse*, Massimo A. Achterberg, Long Ma and Piet Van Mieghem

*Correspondence:
b.prasse@tudelft.nl
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, P.O Box 5031, 2600 GA Delft, The Netherlands

Abstract

At the moment of writing, the future evolution of the COVID-19 epidemic is unclear. Predictions of the further course of the epidemic are decisive to deploy targeted disease control measures. We consider a network-based model to describe the COVID-19 epidemic in the Hubei province. The network is composed of the cities in Hubei and their interactions (e.g., traffic flow). However, the precise interactions between cities is unknown and must be inferred from observing the epidemic. We propose the Network-Inference-Based Prediction Algorithm (NIPA) to forecast the future prevalence of the COVID-19 epidemic in every city. Our results indicate that NIPA is beneficial for an accurate forecast of the epidemic outbreak.

Keywords: Network inference, Epidemiology, COVID-19, Coronavirus, SIR model

Introduction

In December 2019, the novel coronavirus SARS-CoV-2 emerged in the Chinese city Wuhan (Munster et al. 2020). The SARS-CoV-2 virus causes the COVID-19 disease. Contrary to initial observations (Cheng and Shan 2020), the COVID-19 virus does spread from person to person, as confirmed in Chan et al. (2020). On March 19, 2020, there were more than 215,000 confirmed infections, and more than 8500 people died (World Health Organization 2020; 'Situation Update Worldwide, as of 18 March 2020', www.ecdc.europa.eu/en/geographical-distribution-2019-nCoV-cases, unpublished; 'Coronavirus (COVID-19)', www.cdc.gov/coronavirus/2019-nCoV/index.html, unpublished). Assessing the further spread of the COVID-19 epidemic poses a major public health concern.

Many studies aim to estimate the basic reproduction number R_0 of the COVID-19 epidemic (Zhao et al. 2020; Majumder and Mandl 2020; Li et al. 2020; Yang et al. 2020; Imai et al. 2019; Liu et al. 2020; Riou and Althaus 2020; Read et al. 2020; Wu et al. 2020). The basic reproduction number R_0 is a crucial quantity to evaluate the hostility of a virus (Hethcote 2000; Heesterbeek 2002). The basic reproduction number R_0 is defined (Diekmann et al. 1990) as "The expected number of secondary cases produced, in a completely susceptible population, by a typical infective individual during its entire period of infectiousness".

The greater the basic reproduction R_0, the more individuals are infected in the long-term endemic state of the virus. If $R_0 < 1$, then the virus dies out. The estimates for the basic reproduction number R_0 of the COVID-19 epidemic range from $R_0 = 2.0$ to $R_0 = 3.77$.

The basic reproduction number R_0 only coarsely assesses the quantitative behaviour of the epidemic. To obtain a more detailed picture of the epidemic, the development of epidemic outbreak prediction methods is focal. A diverse body of research considers the prediction of general epidemics. For instance, prediction methods are based on Kalman filtering (Yang et al. 2014), Bayesian model averaging (Yamana et al. 2017), basic regression (Brooks et al. 2015) and kernel density estimation (Ray and Reich 2018). Recent work focussed on the dependency of population flow and the viral spread (Colizza et al. 2006; Balcan et al. 2009; Belik et al. 2011; Brockmann and Helbing 2013). As shown by (Pei et al. 2018), the spread of influenza can be more accurately predicted by taking the population flow between cities into account. Read et al. (2020) predicted the COVID-19 epidemic by using the Official Aviation Guide (OAG) Traffic Analyser dataset. Additionally to the OAG dataset, (Wu et al. 2020) used the Tencent database to predict the COVID-19 viral spread.

The population flow clearly has an impact on the evolution of an epidemic. However, the exact population flow is unknown, and epidemic prediction methods must account for inaccuracies of population flow data. In this work, we consider the most extreme case by assuming no prior knowledge of the population flow. To forecast the COVID-19 epidemic, we design the network-based prediction method NIPA that estimates the interactions between cities as an intermediate step. On February 14th, 2020, approximately 75% of the global COVID-19 infections are located in the Chinese province Hubei. Thus, we focus on the COVID-19 epidemic in Hubei. More precisely, our goal is to predict the COVID-19 outbreak for every city in Hubei.

Materials and methods

Data on the COVID-19 epidemic outbreak in Hubei

The time series of reported infections in Hubei forms the basis for the epidemic outbreak prediction. Hubei is divided into 17 cities (more precisely, prefecture-level divisions) and contains the city Wuhan, as illustrated by Fig. 1. We do not consider the city Shennongjia, since the number of infections in Shennongjia is small. We denote the number of considered cities by $N = 16$. The number of newly reported infections for each city in Hubei is openly accessible via the website of the Hubei Province Health Committee (http://www.hubei.gov.cn/, unpublished). The data is updated daily and follows the standard time offset of UTC+08:00. Except for Wuhan, the total number of reported infections is small before January 21, 2020. Hence, we consider the COVID-19 epidemic outbreak starting from January 21. From February 13 on, a new diagnosing method on the basis of chest scans has been used for reporting the infections in Hubei ('Coronavirus Latest: China's Epicentre Records No N ew Cases', www.nature.com/articles/d41586-020-00154-w, unpublished). The new diagnosing method resulted in an erratic spike in the number of reported infections. We focus on predicting the number of infections of the initial diagnosing method, which is based on genetic tests. The number of reported infections of the initial diagnosing method is accessible from (http://www.hubei.gov.cn/, unpublished) until February

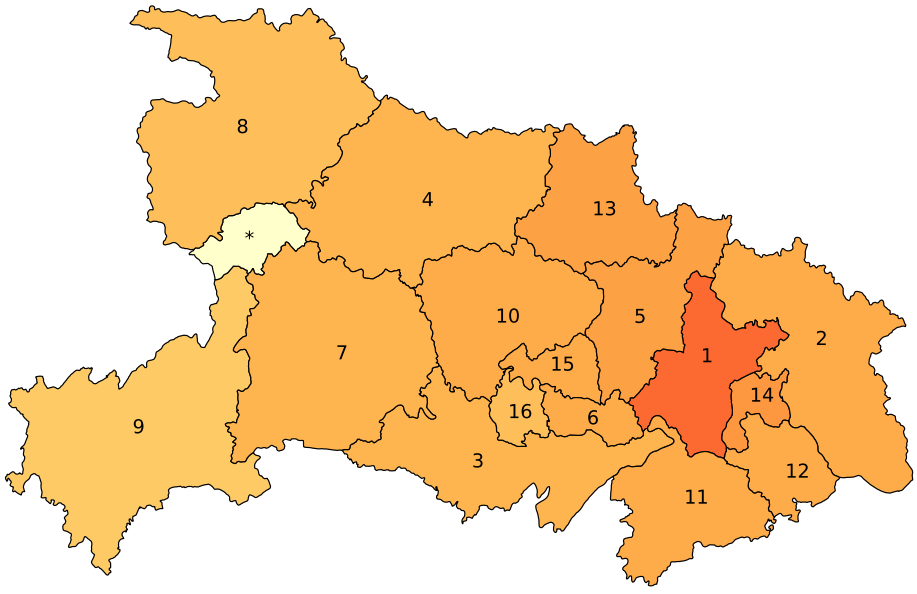

Fig. 1 Map of cities in Hubei. The 17 cities (prefecture-level divisions) of the Chinese province Hubei. The names of the cities are stated in Supplementary Table S2. The greater the fraction of people that have been infected by COVID-19 on February 14, the darker the city. We do not consider the city Shennongjia in this work, which is marked with a star (*) and a light colour

14, 2020. Thus, we focus on the COVID-19 epidemic in Hubei from January 21 until February 14, 2020.

We denote the discrete time by $k \in \mathbb{N}$. The difference of time k to $k+1$ equals one day, and the initial time $k = 1$ corresponds to January 21, 2020. The website (http://www.hubei.gov.cn/, unpublished) states the number of reported infections $N_{rep,i}[k]$ at every time k in every city $i = 1, ..., N$. We obtain the population size p_i of each city i from the Hubei Statistical Yearbook (Li and Xu 2016). The reported fraction of infected individuals in city i at time k follows as

$$\mathcal{I}_{rep,i}[k] = N_{rep,i}[k] / p_i. \tag{1}$$

Supplementary Table S2 states the population size p_i and the complete time series of the number of infections $N_{rep,i}[k]$ for each city in Hubei.

Modelling the COVID-19 epidemic between cities

We model the spread of the COVID-19 virus by the SIR-model: At any discrete time k, every individual is in either one of the compartments *susceptible* (healthy), *infectious* or *removed*. Susceptible individuals can get infectious due to contact with infectious individuals. Due to curing, hospitalisation, quarantine measures or death, infectious individuals become removed individuals, which cannot infect susceptible individuals any longer. For every city i, we denote the 3×1 *viral state* vector at time k by

$$v_i[k] = \begin{pmatrix} \mathcal{S}_i[k] \\ \mathcal{I}_i[k] \\ \mathcal{R}_i[k] \end{pmatrix}. \tag{2}$$

The components $\mathcal{S}_i[k]$, $\mathcal{I}_i[k]$, and $\mathcal{R}_i[k]$ denote the fraction of susceptible, infectious, and removed individuals, respectively. Thus, it holds that $\mathcal{S}_i[k] + \mathcal{I}_i[k] + \mathcal{R}_i[k] = 1$ for

every city i at every time k. The discrete-time SIR model follows from applying Euler's method to the continuous-time mean-field SIR model of (Youssef and Scoglio 2011):

Definition 1 (SIR Epidemic Model (Youssef and Scoglio 2011; Prasse and Van Mieghem 2020)) *For every city i, the viral state* $v_i[k] = (\mathcal{S}_i[k], \mathcal{I}_i[k], \mathcal{R}_i[k])^T$ *evolves in discrete time* $k = 1, 2, ...$ *according to*

$$\mathcal{I}_i[k+1] = (1-\delta_i)\mathcal{I}_i[k] + (1 - \mathcal{I}_i[k] - \mathcal{R}_i[k]) \sum_{j=1}^{N} \beta_{ij}\mathcal{I}_j[k], \tag{3}$$

$$\mathcal{R}_i[k+1] = \mathcal{R}_i[k] + \delta_i\mathcal{I}_i[k],$$

and the fraction of susceptible individuals follows as

$$\mathcal{S}_i[k] = 1 - \mathcal{I}_i[k] - \mathcal{R}_i[k].$$

Here, β_{ij} *denotes the infection probability from city j to city i, and* δ_i *denotes the curing probability of city i.*

The SIR model (3) assumes that the spreading parameters δ_i, β_{ij} do not change over time k. The curing probability δ_i quantifies the capacity of individuals in city i to cure from the virus. The infection probability β_{ij} specifies the number of contacts of individuals in city j with individuals in city i. We emphasise that $\beta_{ii} \neq 0$ since individuals within one city i do interact with each other. The *contact network* between cities in Hubei is given by the $N \times N$ matrix

$$B = \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{N1} & \beta_{N2} & \dots & \beta_{NN} \end{pmatrix},$$

whose elements are probabilities $0 \leq \beta_{ij} \leq 1$. Neither the curing probabilities δ_i nor the infection probabilities β_{ij} are known for the COVID-19 epidemic. Potentially, it is possible to state bounds or estimates for the spreading parameters δ_i and β_{ij} by making use of the people flow or geographical distances between the respective cities. Nevertheless, there would remain an uncertainty regarding the precise value of the spreading parameters δ_i and β_{ij}. In this work, we consider the most extreme case: there is no a priori knowledge on the curing probabilities δ_i nor the infection probabilities β_{ij}.

Network-inference-based prediction algorithm (NIPA)

We propose the NIPA method to predict the outbreak of COVID-19 virus, which consists of three steps. First, we preprocess the raw data of the confirmed number of infected individuals to obtain an SIR time series $v_i[1], ..., v_i[n]$ of the viral state for every city i. Here, the number of observations is denoted by n. Second, based on the time series $v_i[1], v_i[2], ...$, we obtain estimates $\hat{\delta}_i$ and $\hat{\beta}_{ij}$ of the unknown spreading parameters δ_i and β_{ij}. Third, the estimates $\hat{\delta}_i$ and $\hat{\beta}_{ij}$ result in an SIR model (3), which we iterate for future times k to predict the evolution of the 2019-Cov virus. In the following, we give an outline of the first two steps of the prediction method. We refer the reader to Supplementary Information S1 for further details on NIPA.

Data preprocessing

We denote the number of observations by n, which equals the number of days since January 21, 2020. Based on the reported number of infections $N_{rep,i}[k]$, our goal is to obtain an SIR viral state vector $v_i[k] = (\mathcal{S}_i[k], \mathcal{I}_i[k], \mathcal{R}_i[k])^T$ for every city i at any time $k = 1, ..., n$. The fraction of susceptible individuals follows as $\mathcal{S}_i[k] = 1 - \mathcal{I}_i[k] - \mathcal{R}_i[k]$ at any time $k \geq 1$. Thus, it suffices to determine the fraction of infectious individuals $\mathcal{I}_i[k]$ and recovered individuals $\mathcal{R}_i[k]$.

The fraction of infectious individuals $\mathcal{I}_i[k]$ follows from the reported fraction of infections $\mathcal{I}_{rep,i}[k]$. To be precise, the reported data is the number $N_{rep,i}[k]$ of individuals that are *detected* to be infected by COVID-19. Upon detection of the infection, the respective individuals are hospitalised and, hence, not infectious any more to individuals outside of the hospital. We consider the reported fraction of infections $\mathcal{I}_{rep,i}[k]$ as an *approximation* for the number of infectious individuals $\mathcal{I}_i[k]$. In fact, the reported fraction of infections $\mathcal{I}_{rep,i}[k]$ lower-bounds the true fraction of infected individuals $\mathcal{I}_i[k]$ for two reasons. First, not all infectious individuals are aware that they are infected. Second, the diagnosing capacities in the hospitals are limited, particularly when the number of infections increases rapidly. Hence, not all infectious individuals that arrive at a hospital can be reported timely.

We do not know the fraction of removed individuals $\mathcal{R}_i[k]$. At the initial time $k = 1$, it is realistic to assume that $\mathcal{R}_i[1] = 0$ holds for every city i. At any time $k \geq 2$, the removed individuals $\mathcal{R}_i[k]$ could be obtained from (3), if the curing probability δ_i were known. However, we do not know the curing probability δ_i. Hence, we consider 50 equidistant *candidate values* for the curing probability δ_i, ranging from $\delta_{min} = 0.01$ to $\delta_{max} = 1$. We define the set of candidate values as $\Omega = \{\delta_{min}, ..., \delta_{max}\}$. For every candidate value $\delta_i \in \Omega$, the fraction of removed individuals $\mathcal{R}_i[k]$ follows from (3) at all times $k \geq 2$. Thus, we obtain 50 potential sequences $\mathcal{R}_i[1], ..., \mathcal{R}_i[n]$, each of which corresponding to one candidate value $\delta_i \in \Omega$. We estimate the curing probability δ_i, and hence implicitly the sequence $\mathcal{R}_i[1], ..., \mathcal{R}_i[n]$, as the element in Ω that resulted in the best fit of the SIR model (3) to the reported number of infections.

The raw time series $\mathcal{I}_{rep,i}[1], ..., \mathcal{I}_{rep,i}[n]$ exhibits erratic fluctuations. There is a single outlier in city $i = 1$ (Wuhan) at time $k = 8$ (January 28, 2020), which we replace by $\mathcal{I}_{rep,1}[8] = (\mathcal{I}_{rep,1}[7] + \mathcal{I}_{rep,1}[9])/2$. (Potentially, the outlier is due to the increase in the maximum number of individuals that can be diagnosed in Wuhan, from 200 to 2000 individuals per day as of January 27th (https://m.chinanews.com/wap/detail/zw/sh/2020/01-28/9071697.shtml, unpublished). To reduce the fluctuations, we apply a moving average, provided by the Matlab command `smoothdata`, to the time series $\mathcal{I}_{rep,i}[1], ..., \mathcal{I}_{rep,i}[n]$ of every city i. The preprocessed time series $\mathcal{I}_i[1], ..., \mathcal{I}_i[n]$ equals the output of `smoothdata`.

Network inference

For every city i, the curing probability δ_i is estimated as one of the candidate values in the set Ω, as outlined above. The remaining task is to estimate the infection probabilities β_{ij}. The goal of *network inference* (Peixoto 2019; Ma et al. 2019; Di Lauro et al. 2019; Timme and Casadiego 2014; Wang et al. 2016) is to estimate the matrix B of infection probabilities from the SIR viral state observations $v_i[1], ..., v_i[n]$. The matrix B can be interpreted as a weighted adjacency matrix. We adapt a network inference

approach (Prasse and Van Mieghem 2018; 2020), which is based on formulating a set of linear equations and the *least absolute shrinkage and selection operator* (LASSO) (Tibshirani 1996; Hastie et al. 2015). We remark that the network inference approach (Prasse and Van Mieghem 2020) is also applicable to general compartmental epidemic models (Sahneh et al. 2013), such as the Susceptible-Exposed-Infected-Removed (SEIR) epidemic model. The crucial observation from the SIR governing equations (3) is that β_{ij} appears linearly, whereas the state variables $\mathcal{S}_i$, $\mathcal{I}_i$ and $\mathcal{R}_i$ do not. From (3), the infection probabilities β_{ij} satisfy

$$V_i = F_i \begin{pmatrix} \beta_{i1} \\ \vdots \\ \beta_{iN} \end{pmatrix} \tag{4}$$

for all cities $i = 1, ..., N$. Here, the $(n-1) \times 1$ vector V_i and the $(n-1) \times N$ matrix F_i are given by

$$V_i = \begin{pmatrix} \mathcal{I}_i[2] - (1-\delta_i)\mathcal{I}_i[1] \\ \vdots \\ \mathcal{I}_i[n] - (1-\delta_i)\mathcal{I}_i[n-1] \end{pmatrix} \tag{5}$$

and

$$F_i = \begin{pmatrix} \mathcal{S}_i[1]\,\mathcal{I}_1[1] & \dots & \mathcal{S}_i[1]\,\mathcal{I}_N[1] \\ \vdots & \ddots & \vdots \\ \mathcal{S}_i[n-1]\,\mathcal{I}_1[n-1] & \dots & \mathcal{S}_i[n-1]\,\mathcal{I}_N[n-1] \end{pmatrix}. \tag{6}$$

If the SIR model (3) were an exact description of the evolution of the coronavirus, then the linear system (4) would hold with equality. However, the viral state vector $v_i[k]$ in city i does not exactly follow the SIR model (3). Instead, the evolution of the viral state vector $v_i[k]$ is described by

$$v_i[k+1] = f_{\text{SIR}}(v_1[k], ..., v_N[k]) + w_i[k],$$

where the 3×1 vector $f_{\text{SIR}}(v_1[k], ..., v_N[k])$ denotes the right-hand sides of the SIR model (3), and the 3×1 vector $w_i[k]$ denotes the unknown *model error* of city i at time k. Due to the model errors $w_i[k]$, the linear system (4) only holds approximately. Thus, we resort to estimating the infection probabilities β_{ij} by minimising the deviation of the left side and the right side of (4). We infer the network by the LASSO (Tibshirani 1996; Hastie et al. 2015) as follows:

$$\begin{aligned} \min_{\beta_{i1},...,\beta_{iN}} \quad & \left\| V_i - F_i \begin{pmatrix} \beta_{i1} \\ \vdots \\ \beta_{iN} \end{pmatrix} \right\|_2^2 + \rho_i \sum_{j=1, j\neq i}^{N} \beta_{ij} \\ s.t. \quad & 0 \leq \beta_{ij} \leq 1, \quad j = 1, ..., N. \end{aligned} \tag{7}$$

The first term in the objective function of (7) measures the deviation of the left side and the right side of (4). The sum in the objective of (7) is an ℓ_1–norm regularisation term which avoids overfitting. We choose to not penalise the probabilities β_{ii}, since we expect the infections among individuals within the same city i to be dominant. The regularisation parameter $\rho_i > 0$ is set by cross–validation. The LASSO network inference (7) allows for the incorporation of a priori knowledge of the contact network B by adding further

constraints to the infection probabilities β_{ij}. We emphasise that an accurate prediction of an SIR epidemic outbreak does not require an accurate network inference (Prasse and Van Mieghem 2020), see also Supplementary Information S1. If the observed viral state sequence $v_i[1], ..., v_i[n]$ is generated by the SIR model (3), then NIPA accurately predicts the infection state $\mathcal{I}_i[k]$. Furthermore, NIPA provides accurate short-term predictions, also when the viral state $v_i[k]$ does not exactly follow the SIR model (3), i.e., in the presence of model errors $w_i[k]$. We refer the reader to Supplementary Information S1 for further details on NIPA.

Logistic regression

The accuracy of NIPA is evaluated by comparison to a simple prediction method. Qualitatively, the virus spread in many epidemiological models follows a sigmoid function, see also (Van Mieghem 2016). A particular sigmoid function is obtained by logistic regression. As a comparison to NIPA, we apply logistic regression on the reported fractions $\mathcal{I}_{rep,i}[1]$, ..., $\mathcal{I}_{rep,i}[n]$ of infection individuals, *independently* for each city i in Hubei. Logistic regression is advantageous because a logistic function is a closed-form expression. Moreover, the logistic function is an approximation to the exact solution of some epidemiological models and population growth models (Verhulst 1838; Van Mieghem 2016; Prasse and Van Mieghem 2019).

A logistic curve is given by the following equation

$$y(t) = \frac{y_\infty}{1 + e^{-K(t-t_0)}}. \tag{8}$$

In our formulation, $y(t)$ is the time-dependent fraction of infectious individuals, t is the time in days, where January 21 serves as initial condition ($t = 0$), y_∞ is the fraction of infected individuals when time approaches infinity, K is the logistic growth rate and t_0

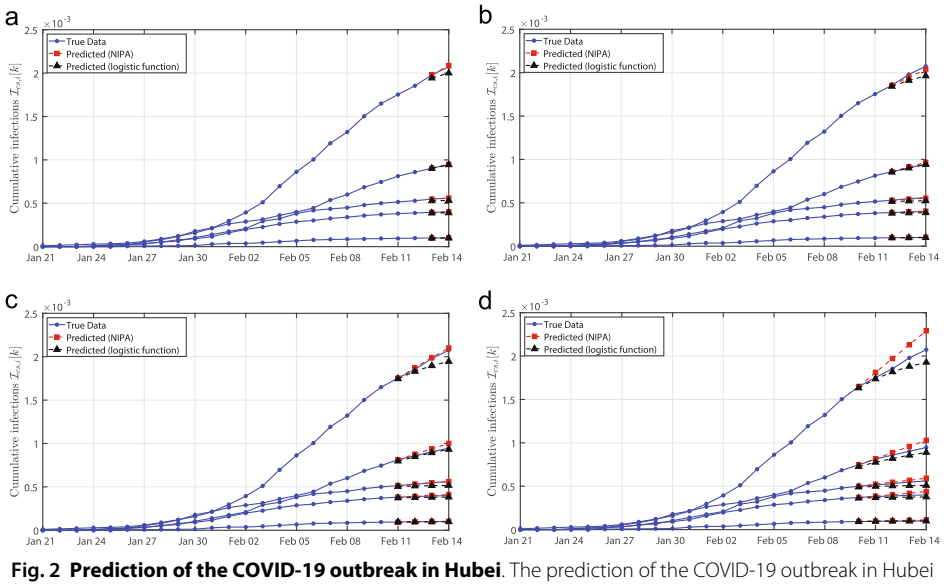

Fig. 2 Prediction of the COVID-19 outbreak in Hubei. The prediction of the COVID-19 outbreak in Hubei by NIPA and by simple logistic regression. For clarity, only five of the $N = 16$ cities are depicted. Each subfigure is obtained by omitting a number $m = 1, 2, 3, 4$ of days prior to February 14, 2020, and subsequently predicting the same number of days ahead in time. The omitted number of data points is equal to: **a** $m = 1$ day, **b** $m = 2$ days, **c** $m = 3$ days and **d** $m = 4$ days. The first prediction data point, for instance February 13 in subfigure **a**), coincides with the last day that has been observed

indicates the inflection point of the logistic equation. For each city in Hubei, we have applied the Matlab command `lsqcurvefit` to fit the reported cumulative fraction

$$\mathcal{I}_{rep,cs,i}[k] = \sum_{\tau=1}^{k} \mathcal{I}_{rep,i}[\tau]$$

of infected individuals to Eq (8).

Results and discussion

To evaluate the prediction accuracy, we remove the data for a fixed number of days, say m, prior to February 14. The prediction model is determined by the observation from 21 January up to 14 − m February, 2020. Then, we predict the course of the disease up to February 14. The course of the disease is shown in Fig. 2 for the removal of $m = 1, 2, 3, 4$ days. For most predictions shown in Fig. 2, the logistic curve appears to underestimate the true fraction of infected individuals, whereas NIPA seems to overestimate the true value.

We quantify the prediction accuracy by the Mean Absolute Percentage Error (MAPE)

$$e[k] = \frac{1}{N} \sum_{i=1}^{N} \frac{\left| \hat{\mathcal{I}}_{cs,i}[k] - \mathcal{I}_{cs,i}[k] \right|}{\mathcal{I}_{cs,i}[k]},$$

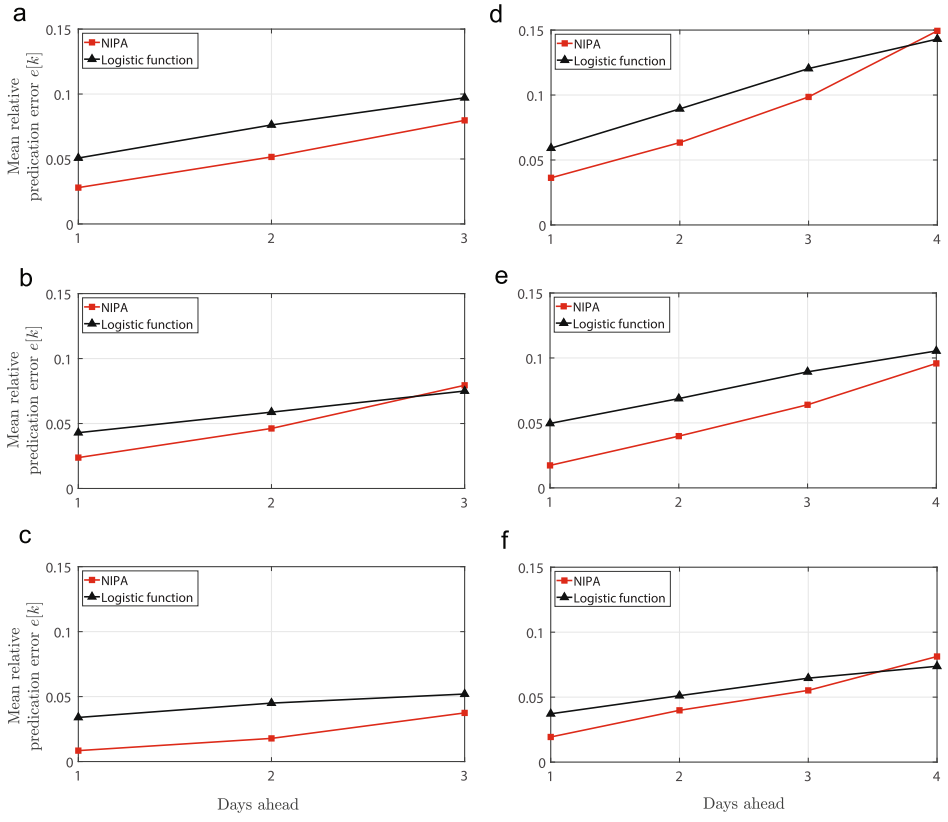

Fig. 3 Prediction error versus prediction time. The prediction accuracy of NIPA and logistic regression to forecast the COVID-19 outbreak in Hubei. Each subfigure is obtained by omitting a number of days prior to February r and subsequently predicting the same number of days ahead in time. The subfigures of the three rows (**a**, **d**), (**b**, **e**) and (**c**, **f**) correspond to February $r = 10$, $r = 12$ and $r = 14$, respectively. The subfigures of the two columns (**a**, **b**, **c**) and (**d**, **e**, **f**) correspond to $m = 3$ and $m = 4$ omitted days before February r, respectively

at any prediction time $k \geq n+1$. Here, the predicted cumulative fraction of individuals of city i at time k equals

$$\hat{\mathcal{I}}_{cs,i}[k] = \sum_{\tau=1}^{k} \hat{\mathcal{I}}_i[\tau]\,. \tag{9}$$

Figure 3 depicts the MAPE prediction error for the data shown in Fig. 2. Two observations are worth mentioning. First, as expected, the prediction error increases when predicting more days ahead. Second, the prediction accuracy of NIPA is almost always better than the logistic regression. In particular, NIPA provides more accurate *short-term* predictions.

Lastly, Fig. 4 illustrates the prediction accuracy versus the time that the epidemic outbreak has been observed. As the epidemic evolves over time, the prediction accuracy of both methods increases. For nearly all forecasts, the NIPA method outperforms logistic regression. Also, as expected, forecasting more days ahead always decreases the prediction accuracy for both prediction methods.

Conclusion

We applied a network-based SIR epidemic model to predict the outbreak of the COVID-19 virus for each city in the Chinese province Hubei. The epidemic model allows to explicitly specify the interactions of individuals of different cities, for instance by using traffic patterns between cities. However, the precise interactions between cities is unknown and must be inferred from observing the evolution of the epidemic.

We proposed the NIPA prediction method, which estimates the interactions between cities as an intermediate step. We did not assume any prior knowledge on the interactions between cities. The prediction method is evaluated on past data of the COVID-19 outbreak in Hubei. Our results indicate that a network-based modelling approach may yield more accurate predictions than modelling the epidemic for each city independently. We believe that the prediction accuracy of NIPA could be further improved, e.g., by using traffic flow patterns as prior knowledge.

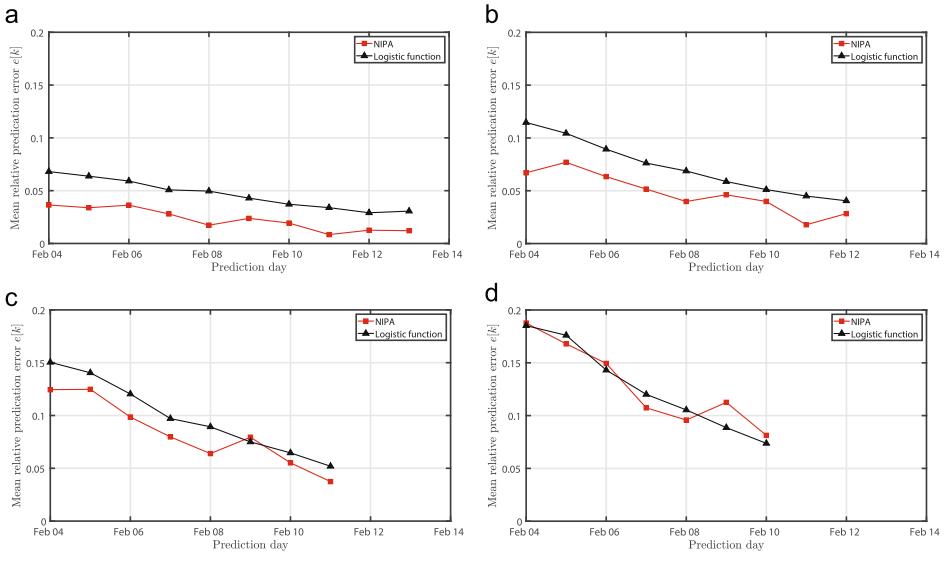

Fig. 4 Prediction error versus number of observed days. The accuracy of both prediction methods for the COVID-19 outbreak versus the date until the data is available. The subfigures correspond to a prediction of : (**a**) 1-day ahead, (**b**) 2-days ahead, (**c**) 3-days ahead, (**d**) 4-days ahead

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1007/s41109-020-00274-2.

Additional file 1: Appendix S1 – Details of NIPA. The details and pseudocode of the Network-Inference-based Prediction Algorithm (NIPA). Furthermore, the prediction accuracy of NIPA is evaluated on the SIR epidemic model.

Additional file 2: Table S2 – Data of the COVID-19 epidemic outbreak in Hubei. The time series of the reported number of infections and the population size for every city in Hubei.

Abbreviations

COVID-19: Coronavirus disease 2019; LASSO: Least absolute shrinkage and selection operator; MAPE: Mean absolute percentage error; NIPA: Network inference-based prediction algorithm; OAG: Official aviation guide; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2; SIR: Susceptible infected removed (epidemic model)

Acknowledgements

We are grateful to Fenghua Wang for helping with collecting the data.

Authors' contributions

BP and MA developed the mathematical and algorithmic framework. MA carried out the simulations. LM has made substantial contributions to the design of the work and collected the epidemic data. PVM initiated and supervised the research. All authors read and approved the manuscript.

Funding

LM is supported by the China scholarship council.

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests.

Received: 20 March 2020 Accepted: 6 June 2020
Published online: 08 July 2020

References

Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci 106(51):21484–21489

Belik V, Geisel T, Brockmann D (2011) Natural human mobility patterns and spatial spread of infectious diseases. Phys Rev X 1(1):011001

Brockmann D, Helbing D (2013) The hidden geometry of complex, network-driven contagion phenomena. Science 342(6164):1337–1342

Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R (2015) epiforecast: Tools for forecasting semi-regular seasonal epidemic curves and similar time series

Chan JF-W, Yuan S, Kok K-H, To KK-W, Chu H, Yang J, Xing F, Liu J, Yip CC-Y, Poon RW-S, et al. (2020) A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet 395(10223):514–523

Cheng JC, Shan J (2020) 2019 novel coronavirus: Where we are and what we know. Infection 48

Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proc Natl Acad Sci 103(7):2015–2020

Di Lauro F, Croix J-C, Dashti M, Berthouze L, Kiss I (2019) Network inference from population-level observation of epidemics. arXiv preprint arXiv:1906.10966

Diekmann O, Heesterbeek JAP, Metz JA (1990) On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. J Math Biol 28(4):365–382

Hastie T, Tibshirani R, Wainwright M (2015) Statistical Learning with Sparsity: the Lasso and Generalizations. CRC press, Florida

Heesterbeek JAP (2002) A brief history of R_0 and a recipe for its calculation. Acta Biotheor 50(3):189–204

Helbing D, Brockmann D, Chadefaux T, Donnay K, Blanke U, Woolley-Meza O, Moussaid M, Johansson A, Krause J, Schutte S, et al (2015) Saving human lives: What complexity science and information systems can contribute. J Stat Phys 158(3):735–781

Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42(4):599–653

Imai N, Cori A, Dorigatti I, Baguelin M, Donnelly CA, Riley S, Ferguson NM (2019) Report 3: Transmissibility of 2019-nCoV. Reference Source

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A, Containing Pap Math Phys Character 115(772):700–721

Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KS, Lau EH, Wong JY, et al. (2020) Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New Engl J Med

Li T, Xu X (2016) Hubei Statistical Yearbook. China Statistics Press, China

Liu T, Hu J, Kang M, Lin L, Zhong H, Xiao J, He G, Song T, Huang Q, Rong Z, et al. (2020) Transmission dynamics of 2019 novel coronavirus (2019-nCoV). bioRxiv
Ma L, Liu Q, Van Mieghem P (2019) Inferring network properties based on the epidemic prevalence. App Netw Sci 4(1):93
Maier BF, Brockmann D (2020) Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. Science 368(6492):742–746
Majumder M, Mandl KD (2020) Early transmissibility assessment of a novel coronavirus in Wuhan, China. Elsevier BV, China. January 23, 2020
Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E (2020) A novel coronavirus emerging in China — key questions for impact assessment. New Engl J Med 382(8):692–694. https://doi.org/10.1056/NEJMp2000929
Pei S, Kandula S, Yang W, Shaman J (2018) Forecasting the spatial transmission of influenza in the United States. Proc Natl Acad Sci 115(11):2752–2757
Peixoto TP (2019) Network reconstruction and community detection from dynamics. Phys Rev Lett 123:128301. https://doi.org/10.1103/PhysRevLett.123.128301
Perc M, Gorišek Miksić N, Slavinec M, Stožer A (2020) Forecasting COVID-19. Front Phys 8:127
Prasse B, Van Mieghem P (2018) Network reconstruction and prediction of epidemic outbreaks for NIMFA processes. arXiv preprint arXiv:1811.06741
Prasse B, Van Mieghem P (2019) Time-dependent solution of the NIMFA equations around the epidemic threshold. Submitted
Prasse B, Van Mieghem P (2020) Network reconstruction and prediction of epidemic outbreaks for general group-based compartmental epidemic models. IEEE Trans Netw Sci Eng
Ray EL, Reich NG (2018) Prediction of infectious disease epidemics via weighted density ensembles. PLoS Comput Biol 14(2):1005910
Read JM, Bridgen JR, Cummings DA, Ho A, Jewell CP (2020) Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. medRxiv
Riou J, Althaus CL (2020) Pattern of early human-to-human transmission of Wuhan 2019-ncov. bioRxiv
Sahneh FD, Scoglio C, Van Mieghem P (2013) Generalized epidemic mean-field model for spreading processes over multilayer complex networks. IEEE/ACM Trans Netw (TON) 21(5):1609–1620
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 58(1):267–288
Timme M, Casadiego J (2014) Revealing networks from dynamics: an introduction. J Phys A Math Theor 47(34):343001
Van Mieghem P (2016) Universality of the SIS prevalence in networks. arXiv preprint arXiv:1612.01386
Verhulst P-F (1838) Notice sur la loi que la population suit dans son accroissement. Corresp Math Phys 10:113–126
Wang W-X, Lai Y-C, Grebogi C (2016) Data based identification and prediction of nonlinear and complex dynamical systems. Phys Rep 644:1–76
World Health Organization (2020) Coronavirus Disease (COVID-2019) Situation Reports. www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. Accessed 18 Mar 2020
Wu JT, Leung K, Leung GM (2020) Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 395(10225):689–697
Yamana TK, Kandula S, Shaman J (2017) Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. PLoS Comput Biol 13(11):1005801
Yang W, Karspeck A, Shaman J (2014) Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. PLoS Comput Biol 10(4):e1003583
Yang Y, Lu Q, Liu M, Wang Y, Zhang A, Jalali N, Dean N, Longini I, Halloran ME, Xu B, Zhang X, Wang L, Liu W, Fang L (2020) Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China. medRxiv. https://doi.org/10.1101/2020.02.10.20021675 https://www.medrxiv.org/content/early/2020/02/11/2020.02.10.20021675.full.pdf
Youssef M, Scoglio C (2011) An individual-based approach to SIR epidemics in contact networks. J Theor Biol 283(1):136–144
Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, Lou Y, Gao D, Yang L, He D, et al. (2020) Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. Int J Infect Dis 92:214–217

Publisher's Note