# Accuracy of predicting epidemic outbreaks

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

**DOI**
10.1103/PhysRevE.105.014302

**Publication date**
2022

**Document Version**
Final published version

**Published in**
Physical Review E

**Citation (APA)**
Prasse, B., Achterberg, M. A., & Van Mieghem, P. (2022). Accuracy of predicting epidemic outbreaks. *Physical Review E*, *105*(1), 014302-1 - 014302-16. Article 014302. https://doi.org/10.1103/PhysRevE.105.014302

# Accuracy of predicting epidemic outbreaks

Bastian Prasse,* Massimo A. Achterberg,† and Piet Van Mieghem‡
*Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, P.O. Box 5031, 2600 GA Delft, The Netherlands*

(Received 4 May 2021; revised 22 July 2021; accepted 10 December 2021; published 7 January 2022)

During the outbreak of a virus, perhaps the greatest concern is the future evolution of the epidemic: How many people will be infected and which regions will be affected the most? The accurate prediction of an epidemic enables targeted disease countermeasures (e.g., allocating medical staff and quarantining). But when can we trust the prediction of an epidemic to be accurate? In this work we consider susceptible-infected-susceptible (SIS) and susceptible-infected-removed (SIR) epidemics on networks with time-invariant spreading parameters. (For time-varying spreading parameters, our results correspond to an optimistic scenario for the predictability of epidemics.) Our contribution is twofold. First, accurate long-term predictions of epidemics are possible only after the peak rate of new infections. Hence, before the peak, only short-term predictions are reliable. Second, we define an exponential growth metric, which quantifies the predictability of an epidemic. In particular, even without knowing the future evolution of the epidemic, the growth metric allows us to compare the predictability of an epidemic at different points in time. Our results are an important step towards understanding when and why epidemics can be predicted reliably.

DOI: 10.1103/PhysRevE.105.014302

## I. INTRODUCTION

Forecasting the evolution of an infectious virus is crucial for deploying appropriate, and timely, lockdown measures. Sophisticated predictions of epidemic outbreaks are based on mathematical epidemiology. The vast majority of epidemic models assumes that every individual is in either one compartment [1–5]. Every compartment describes another stage of the disease. The two most fundamental compartments are *susceptible* $\mathcal{S}$ (healthy) and *infected* $\mathcal{I}$. Susceptible individuals can get infected by contact with infectious individuals.

We consider a population of $N$ groups of individuals, which could be households, cities or whole provinces. We denote the curing rate of group $i$ by $\delta_i > 0$. Furthermore, for every group $i, j = 1, \ldots, N$, we denote the infection rate from group $j$ to group $i$ by $\beta_{ij}$. If individuals of group $j$ are in contact with individuals of group $i$, then it holds that $\beta_{ij} > 0$. The more probable an infection from group $j$ to group $i$, the greater the infection rate $\beta_{ij}$. If individuals of group $j$ are *not* in contact with individuals of group $i$, then it holds that $\beta_{ij} = 0$. The $N \times N$ infection rate matrix $B$ consists of the elements $\beta_{ij}$ and specifies the contact network of the whole population.

Conceptually, there are two kinds of compartmental epidemic models. First, the susceptible-infected-susceptible (SIS) epidemic model, which assumes that infected individuals can cure and become susceptible again.

*Definition 1 [Susceptible-infected-susceptible (SIS) epidemic model [1,6–8]].* Consider a population of $N$ groups of individuals, which are either susceptible $\mathcal{S}$ or infected $\mathcal{I}$ at every time $t \geqslant 0$. Then, for every group $i = 1, \ldots, N$, the fraction of infected individuals $\mathcal{I}_i(t)$ evolves according to

$$\frac{d\mathcal{I}_i(t)}{dt} = -\delta_i \mathcal{I}_i(t) + \mathcal{S}_i(t) \sum_{j=1}^{N} \beta_{ij} \mathcal{I}_j(t), \tag{1}$$

and the fraction of susceptible individuals in group $i$ follows as $\mathcal{S}_i(t) = 1 - \mathcal{I}_i(t)$.

The susceptible-infected-removed (SIR) model is the second kind of compartmental epidemic model. The SIR model assumes that cured individuals are immune to the disease, which is modelled by the *removed* compartment $\mathcal{R}$.

*Definition 2 [Susceptible-infected-removed (SIR) epidemic model [9,10]].* Consider a population of $N$ groups of individuals, which are either susceptible $\mathcal{S}$, infected $\mathcal{I}$ or removed $\mathcal{R}$ at every time $t \geqslant 0$. Then, for every group $i = 1, \ldots, N$, the fraction of infected individuals $\mathcal{I}_i(t)$ evolves according to

$$\frac{d\mathcal{I}_i(t)}{dt} = -\delta_i \mathcal{I}_i(t) + \mathcal{S}_i(t) \sum_{j=1}^{N} \beta_{ij} \mathcal{I}_j(t), \tag{2}$$

the fraction of removed individuals $\mathcal{R}_i(t)$ evolves according to

$$\frac{d\mathcal{R}_i(t)}{dt} = \delta_i \mathcal{I}_i(t), \tag{3}$$

and the fraction of susceptible individuals follows as $\mathcal{S}_i(t) = 1 - \mathcal{I}_i(t) - \mathcal{R}_i(t)$.

Hence, the key difference between the SIS and the SIR model is that, after curing from the disease, individuals either can be reinfected or are immune, respectively. There are vari-

*b.prasse@tudelft.nl
†m.a.achterberg@tudelft.nl
‡p.f.a.vanmieghem@tudelft.nl

 

ations to both the SIS and the SIR model [3]. For instance, the susceptible-infected-removed-susceptible (SIRS) model and the susceptible-exposed-infected-removed (SEIR) model consider time-limited immunity and an incubation period, respectively.

Both the SIS model and the SIR model in Definitions 1 and 2 consider time-invariant spreading rates $\delta_i$, $\beta_{ij}$. However, in many epidemics, the spreading rates $\delta_i(t)$ and $\beta_{ij}(t)$ do depend on time $t$, at least to some extent, which is due to several factors, including seasonality of the virus and time-varying mitigation measures and vaccinations [4,11], which result in multiple "waves" of the epidemic (see Fig. 1 of Ref. [12]) or fluctuations in the human mobility pattern, depending on the day of the week or the season of the year. In this work, we focus on the epidemics with constant spreading rates $\beta_{ij}$, $\delta_i$ for three reasons. First, the SIS and SIR models with constant rates $\beta_{ij}$, $\delta_i$ are among the most fundamental and well-studied epidemic models on networks. Hence, the predictability of the models in Definitions 1 and 2 is of interest in its own right. Second, in some scenarios, it is reasonable to assume approximately constant spreading rates, i.e., $\beta_{ij}(t) \approx \beta_{ij}$ and $\delta_i(t) \approx \delta_i$ from time $t = 0$ until some time $t = t_{\text{end}}$. For instance, the time interval $[0, t_{\text{end}}]$ could correspond to (parts of) a single wave of COVID-19. Hence, the predictability of the models in Definitions 1 and 2 is relevant to epidemics with time-varying rates $\beta_{ij}(t)$, $\delta_i(t)$ over bounded time intervals $t \in [0, t_{\text{end}}]$. Third, time-varying rates $\delta_i(t)$, $\beta_{ij}(t)$ are more general than constant rates $\beta_{ij}$, $\delta_i$. Even if the epidemic closely follows the epidemic model with time-invariant parameters $\beta_{ij}$, $\delta_i$, we show that accurate predictions cannot be obtained. *For time-dependent rates $\beta_{ij}(t)$, $\delta_i(t)$, predicting epidemics is even harder, arguably leading to an even worse prediction accuracy than for time-invariant parameters $\beta_{ij}$, $\delta_i$.* As a result, by considering constant rates $\beta_{ij}$, $\delta_i$, we obtain optimistic results for predicting epidemics with time-varying rates $\beta_{ij}(t)$, $\delta_i(t)$.

Besides variations over time $t$, the contact network might be adaptive [13,14]: For instance, individuals might avoid contacts with symptomatic individuals, which results in infection rates $\beta_{ij}(\mathcal{I}_i(t), \mathcal{I}_j(t))$ that depend on the infection states $\mathcal{I}_i(t)$, $\mathcal{I}_j(t)$ of the respective individuals. Definition 1 and Definition 2 are a deterministic, mean-field, description of the virus spread, which might be more accurately described stochastically. Furthermore, both Definition 1 and Definition 2 consider Markovian viral dynamics. Non-Markovian viral dynamics can be dramatically different to Markovian viral dynamics [15]. We refer the reader to [3,4] for an overview of more complex epidemic models.

For the SIS and SIR model, we denote the *prevalence* (average fraction of infections) at time $t$ by

$$y(t) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{I}_i(t). \tag{4}$$

Then, predicting the course of the epidemic translates to estimating the prevalence $y(t)$ at future times $t$. For both the SIS and SIR model, we argue that the prediction of an epidemic is inherently difficult, independently of the particular prediction algorithm.

## II. RELATED WORK

Several studies approach the prediction limits of epidemic outbreaks from different angles. Cirillo and Taleb [16] demonstrate that the number of fatalities of various past epidemics is strongly fat-tailed, which renders long-term predictions of epidemics outbreaks impossible. Castro *et al.* [17] and Paggi [18] study extensions of the SIR model and show that, even though the respective model accurately fits the past epidemic outbreak, a reliable prediction is not possible. The same conclusion is drawn by Alberti and Faranda [19], who directly fit a logistic function to the number of infections. In this work, we aim to *quantify* the predictability of an epidemic. We show that the predictability is limited by the initial exponential growth of the epidemic, and we propose a metric to quantify exponential growth. Based on the growth metric, it is indeed possible to obtain quantitative statements on the predictability of an epidemic.

## III. THE LOGISTIC FUNCTION FOR EPIDEMICS ON NETWORKS

Of crucial importance to both the SIS and the SIR epidemic model is the logistic function[1] $f(t)$, which has been introduced by Verhulst [20] as

$$f(t) = \frac{y_\infty}{1 + e^{-K(t-t_0)}}. \tag{5}$$

Here we denote the *steady-state* prevalence by $y_\infty > 0$, the *inflection point* (time when the peak of the number of new infections occurs) by $t_0$ and the logistic *growth rate* by $K > 0$. Acknowledging recent results [21,22], the objective of this section is to summarize and motivate the use of the logistic function for epidemic models on networks, introduced in Sec. I. Building upon the results summarized in this section, the novelty of this work lies in the results presented in Sec. IV.

Initially, for small times $t$, the logistic function $f(t)$ increases exponentially. At the inflection point $t = t_0$, the slope of the logistic function $f(t)$ reaches its maximum, and, as $t \to \infty$, the logistic function converges to $f(t) \to y_\infty$. For most epidemic models, the initial phase of the epidemic outbreak is approximated by a branching process [23–25]. In agreement with the logistic function $f(t)$, the branching process results in exponential, or Malthusian, growth, and the growth rate $K$ corresponds to the Malthusian growth parameter. Fairly rapidly, the pure exponential growth ceases due to finite size effects and resistive mechanisms [26].

### A. SIS epidemics

For the SIS epidemic model (1), the basic reproduction number $R_0$ equals the spectral radius of the $N \times N$ effective infection rate matrix $W = \text{diag}(1/\delta_1, \ldots, 1/\delta_N)B$. Here the $N \times N$ diagonal matrix with an $N \times 1$ vector $v$ on its diagonal is denoted by $\text{diag}(v)$. We denote the principal eigenvector of the matrix $W$ by $x_1$. The steady-state fraction of infections is

[1]Equivalently, we can describe the number of infections by a hyperbolic tangent $\tanh(t)$, which equals to a shifted logistic function $f(t)$.

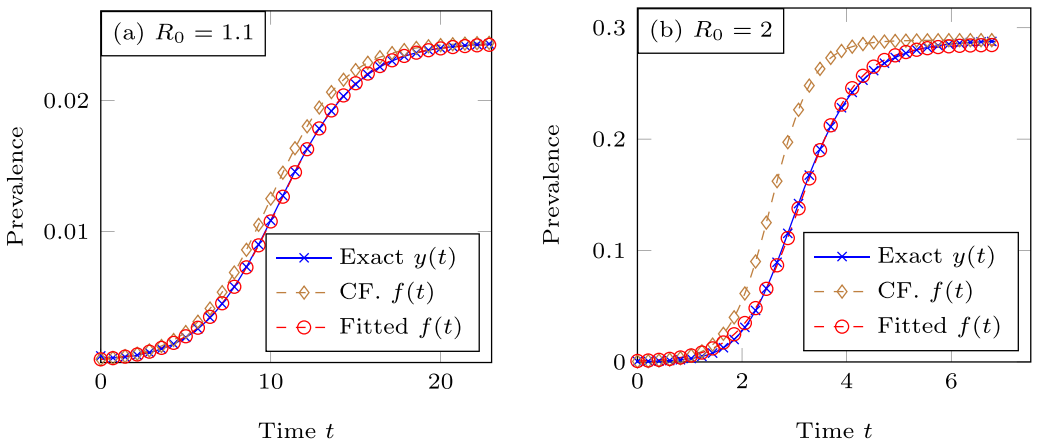


FIG. 1. The accuracy of approximating the prevalence $y(t)$ of the SIS epidemic model on networks (1) by a logistic curve $f(t)$, for a Barabási-Albert random graph [27] with $N = 500$ nodes, heterogeneous spreading rates $\delta_i$, $\beta_{ij}$, and small initial viral states $\mathcal{I}_i(0)$. The red curve ("Fitted") shows the logistic function $f(t)$, which is fitted to the prevalence $y(t)$. The brown curve ("CF.") shows the logistic function $f(t)$ whose parameters are given explicitly, in closed form, by Proposition 4, which is accurate as $R_0 \downarrow 1$.

denoted by $\mathcal{I}_{\infty,i} = \lim_{t\to\infty} \mathcal{I}_i(t)$ for every node $i$ and satisfies

$$(1 - \mathcal{I}_{\infty,i}) \sum_{j=1}^{N} \beta_{ij} \mathcal{I}_{\infty,j} = \delta_i \mathcal{I}_{\infty,i}, \quad i = 1, \ldots, N. \tag{6}$$

As shown in Theorem 4 of Ref. [22], the prevalence $y(t)$ of SIS epidemics on a complete graph is described by a logistic function:

*Proposition 3 ([22]).* Consider the SIS epidemic model (1) and suppose that, for some $\beta$, $\delta$, the spreading rates satisfy $\beta_{ij} = \beta$ and $\delta_i = \delta$ for all nodes $i$, $j$. Then, provided that $R_0 > 1$, the prevalence (4) equals a logistic function, $y(t) = f(t)$, where the logistic growth rate equals $K = \beta N - \delta$, the steady state equals $y_\infty = \frac{K}{\beta N}$, and the inflection point equals

$$t_0 = -\frac{2}{K} \operatorname{arctanh}\left( 2\frac{\beta}{K} \sum_{i=1}^{N} \mathcal{I}_i(0) - 1 \right).$$

Furthermore, as shown in Theorem 3 of Ref. [21], the prevalence $y(t)$ of SIS epidemics on an arbitrary, undirected graph follows a logistic function, provided that $R_0 \approx 1$ and the initial viral states $\mathcal{I}_i(0)$ are small:[2]

*Proposition 4 ([21]).* Consider the SIS epidemic model (1) with a symmetric and irreducible infection rate matrix $B$ and suppose that the basic reproduction number $R_0$ is close to, but above, 1. Then, provided the initial states $\mathcal{I}_i(0)$ are small, the evolution of the viral state $\mathcal{I}_i(t)$ for every node $i$ is approximated by a logistic function, $\mathcal{I}_i(t) \approx f_i(t)$ at every time $t$, where the logistic growth rate equals

$$K = (R_0 - 1) \sum_{i=1}^{N} \delta_i (x_1)_i^2, \tag{7}$$

the steady-state prevalence equals the solution of (6), $y_{\infty,i} = \mathcal{I}_{\infty,i}$, and the inflection point equals

$$t_0 = -\frac{2}{K} \operatorname{arctanh}\left( 2\frac{\sum_{i=1}^{N} \mathcal{I}_{\infty,i} \mathcal{I}_i(0)}{\sum_{i=1}^{N} \mathcal{I}_{\infty,i}^2} - 1 \right). \tag{8}$$

Proposition 4 implies that, for $R_0$ close to 1, the prevalence $y(t)$ in (4) is approximated by a logistic function with the steady state $y_\infty = \sum_{i=1}^{N} \mathcal{I}_{\infty,i}/N$ and the growth rate $K$ and inflection point $t_0$ given by (7) and (8), respectively.

We perform simulations of SIS epidemics on networks to illustrate the accuracy of approximating the prevalence $y(t)$ by a logistic curve $f(t)$. We consider a Barabási-Albert random graph [27] with $N = 500$ nodes and parameters $m = m_0 = 3$. If there is a link between node $i$ and $j$, then we set the infection rate $\beta_{ij}$ to a uniformly distributed random number in $[0.5, 0.6]$. The curing rates $\delta_i$ are set to a uniform random number in $[0.5c, 0.6c]$, where the scalar $c$ is set such that the basic reproduction number equals $R_0 = 1.1$ and $R_0 = 2$ for the respective subplot in Fig. 1. We consider small initial viral states $\mathcal{I}_i(0)$, which are set to a uniform random number in [0,0.001] for every node $i$. Since we interpret the nodes as groups, $\mathcal{I}_i(0) \in [0, 0.001]$ is equivalent to considering that at most 1 out of 1000 individuals are infected in group $i$. Figure 1 shows that the SIS prevalence $y(t)$ is accurately approximated by a logistic curve $f(t)$. We emphasize that the approximation by a logistic curve is less accurate if the initial viral states $\mathcal{I}_i(0)$ are large.

Propositions 3 and 4 demonstrate the applicability of the logistic curve $f(t)$ to the deterministic SIS process (1). Furthermore, the logistic curve gives an approximation and bounds for the prevalence of the *stochastic* SIS process [28]. In some settings [28,29], the logistic function $f(t)$ is not an accurate description of the stochastic SIS process. In particular, the prevalence of the stochastic SIS process can exhibit a local minimum, before converging to the metastable state. Such local minima generally resolve from strong heterogeneity in the network or result from highly heterogeneous transition rates between nodes [28]. For real-world epidemics, the human contact network is also highly heterogeneous. However,

[2]For ease of exposition, we deliberately choose to not present Proposition 4 rigorously, in particular, the condition on $R_0$, the initial states $\mathcal{I}_i(0)$, and the approximation accuracy $y(t) \approx f(t)$. We refer to [21] for a precise version of Proposition 4.

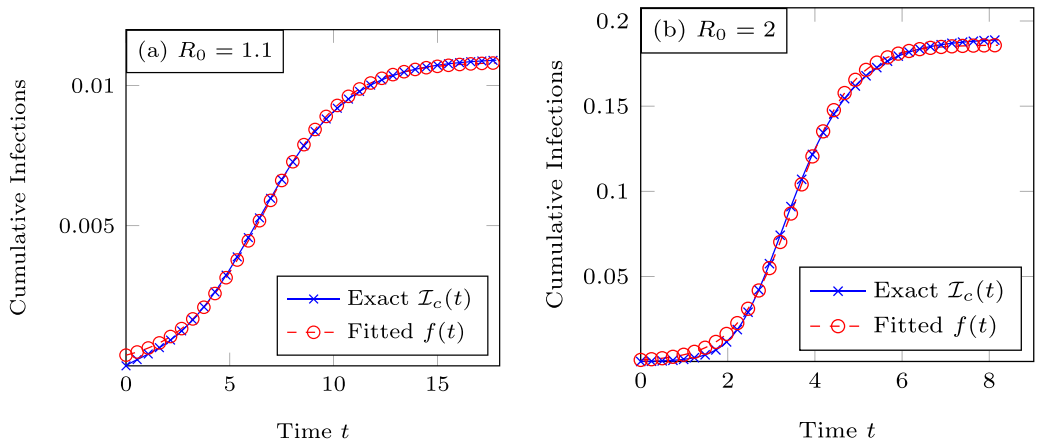


FIG. 2. The accuracy of approximating the prevalence $y(t)$ of the SIR epidemic model on networks, in Definition 2, by a logistic curve $f(t)$, for a Barabási-Albert random graph [27] with $N = 500$ nodes, heterogeneous spreading rates $\delta_i$, $\beta_{ij}$, and small initial viral states $\mathcal{I}_i(0)$. The red curve ("Fitted") shows the logistic function $f(t)$, which is fitted to the cumulative prevalence $\mathcal{I}_c(t)$ in (9), where we set $N_{\text{pop}} = 1$.

we consider the situation where the population separated into groups, which significantly reduces the heterogeneity of the population. Therefore, we argue that the first wave of most real-world epidemics on time-invariant networks is well described by a sigmoid curve, without local minima.

### B. SIR epidemics

Similarly to Proposition 3, in the SIR epidemic model, the solution for the removed compartment $\mathcal{R}(t)$ can be approximated by a logistic function, as shown in the seminal work of Kermack and McKendrick [9]. We define $\tau = \frac{N\beta}{\delta}$.

*Proposition 5 ([9]).* Consider the SIR epidemic model (2) and assume that $\mathcal{R}_i(0) = 0$ for all nodes $i$ and $\mathcal{I}_1(0) = \cdots = \mathcal{I}_N(0) > 0$. Denote the average fraction of removed by $\mathcal{R}(t) = 1/N \sum_{i=1}^{N} \mathcal{R}_i(t)$ and the initial fraction of susceptible by $s_0 = \mathcal{S}_i(0)$ for an arbitrary node $i$. Then, if $\mathcal{R}(t) \ll 1/\tau$ holds true at all times $t$, the removed individuals $\mathcal{R}(t)$ can be approximated by a logistic curve at all times $t \geqslant 0$ as

$$\mathcal{R}(t) \approx N r_1 + \frac{N(r_2 - r_1)}{1 - \frac{r_2}{r_1} e^{-\frac{1}{2}\tau^2 (r_2 - r_1)\delta t}}.$$

Here the constants $r_1$ and $r_2$ are equal to

$$r_l = \begin{cases} \frac{1}{s_0\tau^2}\left[(s_0\tau - 1) + \sqrt{(s_0\tau - 1)^2 + 2s_0(1 - s_0)\tau^2}\right] & \text{if } l = 1, \\ \frac{1}{s_0\tau^2}\left[(s_0\tau - 1) - \sqrt{(s_0\tau - 1)^2 + 2s_0(1 - s_0)\tau^2}\right] & \text{if } l = 2. \end{cases}$$

Proposition 5 states that the removed individuals $\mathcal{R}(t)$ is approximated by a logistic function *plus the offset* $Nr_1$. By the definition of the SIR model in (3), the prevalence $y(t)$ in (4) is proportional to the derivative of the removed individuals $\mathcal{R}(t)$. Thus, Proposition 5 implies that the *cumulative* prevalence

$$\mathcal{I}_c(t) = \int_0^t N_{\text{pop}} y(\tilde{t}) d\tilde{t}, \tag{9}$$

where $N_{\text{pop}}$ is the number of individuals in the whole population (consisting of all $N$ groups), is approximated by a logistic function (plus offset). Then the peak of the epidemic, i.e., the largest *increase* of infections, occurs at the inflection point $t_0$.

Figure 2 demonstrates the accuracy of approximating the cumulative prevalence $\mathcal{I}_c(t)$ of SIR epidemics on Barabási-Albert random graphs. The parameter settings are the same as for Fig. 1.

## IV. PREDICTING EPIDEMIC OUTBREAKS

Propositions 3–5, and variations thereof, motivate the application of the logistic function (5) to the prediction of an epidemic outbreak. In particular, the logistic function has been applied to forecast the coronavirus virus disease 2019 (COVID-19) outbreak in China [30–34], the Netherlands [35] and Italy [36]; see also [37,38]. Furthermore, the logistic function has been applied to predict other phenomena than COVID-19, including tuberculosis [39] and product sales [40,41]. We consider the prediction of the cumulative number of infections $\mathcal{I}_c(t)$, as defined in (9). In a real-world epidemic, the infections $\mathcal{I}_c(t)$ do not exactly follow a logistic function $f(t)$. Instead, the infections $\mathcal{I}_c(t)$ satisfy

$$\mathcal{I}_c(t) = f(t) + w(t) \tag{10}$$

for some logistic function $f(t)$ and the unknown *model error* $w(t)$. The motivation behind applying the logistic curve (10) for predicting epidemics is based on the connection to the SIS and SIR epidemic model on networks, as outlined in Sec. III. Real-world epidemic data are collected in a periodic manner, i.e., in discrete time intervals. For instance, the Dutch National Institute for Public Health and the Environment (RIVM) reports the number of COVID-19 infections in the Netherlands on a daily basis (see Ref. [42]). We assume that the cumulative number of infections $\mathcal{I}_c(t)$ has been observed

at discrete times $t = 1, 2, \ldots, t_{\text{obs}}$, where $t_{\text{obs}} \in \mathbb{N}$ denotes the *observation time*.

To predict the number of infections $\mathcal{I}_c(t)$ at times $t > t_{\text{obs}}$, we consider a two-step approach. First, we obtain parameter estimates $\hat{y}_\infty, \hat{t}_0, \hat{K}$ of the logistic function $f(t)$ by solving the nonlinear least-squares problem

$$(\hat{y}_\infty, \hat{t}_0, \hat{K}) = \underset{y_\infty, t_0, K}{\operatorname{argmin}} \sum_{t=1}^{t_{\text{obs}}} \left( \mathcal{I}_c(t) - \frac{y_\infty}{1 + e^{-K(t-t_0)}} \right)^2. \quad (11)$$

In line with Appendix F of Ref. [43], we solve the nonlinear least-squares problem (11) with the Matlab command `GlobalSearch`, with the initial conditions $y_\infty = \mathcal{I}_c(t_{\text{obs}})$, $K = 1$, $t_0 = t_{\text{obs}}$. Second, we predict the number of infections $\mathcal{I}_c(t)$ at times $t > t_{\text{obs}}$ by the logistic function (5) as $\hat{\mathcal{I}}_c(t) \approx \hat{f}(t)$, where the estimate of the logistic function $f(t)$ equals

$$\hat{f}(t) = \frac{\hat{y}_\infty}{1 + e^{-\hat{K}(t-\hat{t}_0)}}.$$

Schultz [44] analyzed the impact of errors of the parameters $\hat{y}_\infty, \hat{K}, \hat{t}_0$ on the deviation of the logistic function $f(t)$ to its estimate $\hat{f}(t)$. The remainder of this section consists of two parts. First, we focus on the simplified problem of fitting the logistic function $f(t)$ to three points in Sec. IV A. Second, we argue that the prediction of epidemics is ill-conditioned in Sec. IV B.

### A. Fitting the logistic function to three equidistant points

As shown below, a central quantity for fitting the logistic function $f(t)$ is the *growth metric* $\Phi(y_1, y_2, y_3)$:

*Definition 6.* For some function $g(t)$, with $g(t) > 0$ at all times $t$, consider three equidistant points $y_1 = g(0)$, $y_2 = g(\Delta t)$, $y_3 = g(2\Delta t)$. Then the growth metric is defined by

$$\Phi(y_1, y_2, y_3) = \frac{y_2}{y_3} - \frac{y_1}{y_2}. \quad (12)$$

The growth metric $\Phi(y_1, y_2, y_3)$ can be interpreted in two ways. First, consider the sign of the growth metric $\Phi(y_1, y_2, y_3)$. It holds that $\Phi(y_1, y_2, y_3) > 0$ if and only if[3] to $y_3/y_2 < y_2/y_1$. In other words, the relative increase $y_3/y_2$ from time $t = \Delta t$ to $t = 2\Delta t$ must be smaller than the relative increase $y_2/y_1$ from time $t = 0$ to $t = \Delta t$. By definition of exponential growth, it would hold that $y_3/y_2 = y_2/y_1$ if the three points $y_1$, $y_2$ and $y_3$ were on an exponential function, i.e., $y_1 = b^t$, $y_2 = b^{t+\Delta t}$ and $y_3 = b^{t+2\Delta t}$ for some basis $b \geqslant 0$. Thus, $\Phi(y_1, y_2, y_3) > 0$ and $\Phi(y_1, y_2, y_3) < 0$ indicates that the function $g(t)$ grows slower or faster, respectively, than an exponential function from time $t = 0$ to time $t = 2\Delta t$.

Second, if the time spacing $\Delta t$ is small, then the growth metric $\Phi(y_1, y_2, y_3)$ is related to the logarithmic derivative of the function $g(t)$. Denote the logarithm of the function $g(t)$ as $h(t) = \log(g(t))$. The first derivative of $h(t)$ equals $h'(t) = g'(t)/g(t)$. For small time spacings $\Delta t$, the derivative $h''(t)$ is approximated by the difference quotient

$$h''(t) \approx \frac{1}{\Delta t} \left[ \frac{g'(t + \Delta t)}{g(t + \Delta t)} - \frac{g'(t)}{g(t)} \right].$$

Analogously, both derivatives $g'(t + \Delta t)$ and $g'(t)$ can be approximated by difference quotients, which yields that

$$\begin{aligned} h''(t) &\approx \frac{1}{\Delta t^2} \left[ \frac{g(t + \Delta t) - g(t)}{g(t + \Delta t)} - \frac{g(t) - g(t - \Delta t)}{g(t)} \right] \\ &= -\frac{1}{\Delta t^2} \left[ \frac{g(t)}{g(t + \Delta t)} - \frac{g(t - \Delta t)}{g(t)} \right]. \end{aligned}$$

Hence, by identifying $y_1 = g(t - \Delta t)$, $y_2 = g(t)$ and $y_3 = g(t + \Delta t)$, the growth metric $\Phi(y_1, y_2, y_3)$ is related to the second logarithmic derivative of the function $g(t)$ as

$$\Phi(y_1, y_2, y_3) \approx -\Delta t^2 h''(t).$$

Particular, if $\Phi(y_1, y_2, y_3) > 0$ and the time spacing $\Delta t$ is sufficiently small, then the function $g(t)$ is strictly logarithmically concave [45].

Pearl and Reed [46] showed that the logistic function $f(t)$ can be fitted in closed form to three points $y_1$, $y_2$ and $y_3$ at equidistant time points[4] $t = 0$, $t = \Delta t$ and $t = 2\Delta t$, respectively. We observe that the results in [46] can be stated in dependency on the growth metric $\Phi(y_1, y_2, y_3)$ as:

*Proposition 7.* Consider three points $y_3 > y_2 > y_1 > 0$ and a time spacing $\Delta t > 0$. Then there exists a logistic function $f(t)$ with $f(0) = y_1$, $f(\Delta t) = y_2$ and $f(2\Delta t) = y_3$ if and only if

$$\Phi(y_1, y_2, y_3) > 0. \quad (13)$$

Furthermore, the logistic function $f(t)$ is unique, and the steady state equals

$$y_\infty = y_1 + \frac{(y_1 - y_2)^2}{y_2} \frac{1}{\Phi(y_1, y_2, y_3)}, \quad (14)$$

the logistic growth rate equals

$$K = -\frac{1}{\Delta t} \log \left[ \frac{y_1}{y_2} + \frac{y_1}{y_1 - y_2} \Phi(y_1, y_2, y_3) \right], \quad (15)$$

and the inflection point equals

$$t_0 = \frac{1}{K} \log \left[ \frac{(y_1 - y_2)^2}{y_1 y_2} \frac{1}{\Phi(y_1, y_2, y_3)} \right]. \quad (16)$$

*Proof.* Appendix A. ■

We emphasize that condition (13) implies that a logistic function $f(t)$ can be fitted exactly only to three points $y_1$, $y_2$, $y_3$ whose relative increase is slower than exponential. Figure 3 shows that the growth metric $\Phi(f(0), f(t/2), f(t))$ is close to zero for small times $t$. Thus, the logistic function $f(t)$ is practically indistinguishable[5] from an exponential function

[3] Furthermore, $y_3/y_2 = y_2/y_1$ is equivalent to $\log(y_3) - \log(y_2) = \log(y_2) - \log(y_1)$. Thus, $\Phi(y_1, y_2, y_3) = 0$ if and only if the three equidistant points $y_1, y_2, y_3$ lie on a line in a semilogarithmic plot (see also Fig. 3).

[4] Without loss of generality, we assume that the first point $y_1$ corresponds to time $t = 0$. Otherwise, if the first point $y_1$ corresponds to some time $\tilde{t} > 0$, then consider a time shift by formally replacing $t$ with $t + \tilde{t}$.

[5] Here we consider logistic functions $f(t)$ whose inflection point $t_0 \gg 0$, such that $f(t) \approx y_\infty \exp[K(t - t_0)]$ when $t$ is small. For

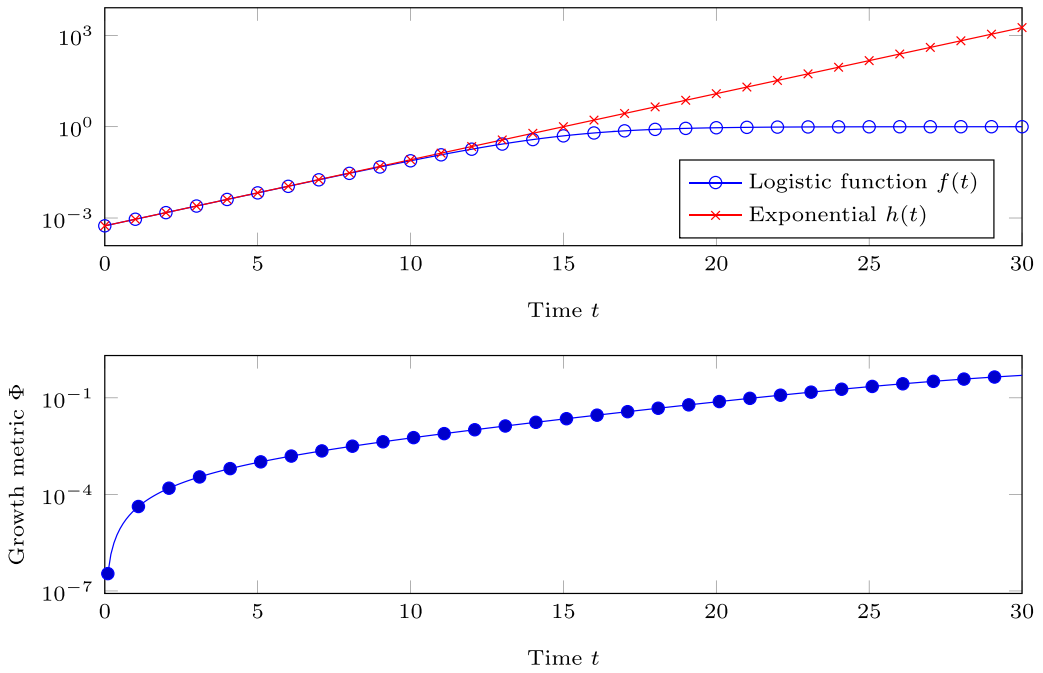


FIG. 3. The growth metric for a logistic function. Upper subplot: The logistic function $f(t)$ with parameters $K = 0.5$, $t_0 = 15$, and $y_\infty = 1$ and the exponential function $h(t) = y_\infty e^{K(t-t_0)}$ on a semilogarithmic scale. Lower subplot: The growth metric $\Phi(y_1, y_2, y_3)$ for the points $y_1 = f(0)$, $y_2 = f(t/2)$, $y_3 = f(t)$ vs time $t$ on a semilogarithmic scale.

at small times $t$. As we argue in Sec. IV B, the strong resemblance of the logistic function $f(t)$ and an exponential function is decisive for the prediction limits of an epidemic outbreak.

### B. Ill-conditioning of predicting epidemic outbreaks

If the model errors $w(t)$ in (10) are sufficiently small, then the solution $\hat{y}_\infty, \hat{t}_0, \hat{K}$ to the least-squares problem (11) approximately equals to the true parameters $y_\infty, t_0, K$. However, it is not clear what "sufficiently small" means. Thus, we face the fundamental question: *How much do small, but nonzero, model errors $w(t)$ affect the accuracy of the estimate $\hat{f}(t)$?*

To quantify the deviation of the estimated logistic function $\hat{f}(t)$ to the true function $f(t)$, we apply Proposition 7, which states that every logistic function can be parameterized by specifying three points $y_1, y_2$, and $y_3$. We choose three points $y_1, y_2$, and $y_3$ in the observation time interval $[0, t_{\text{obs}}]$. More precisely, we set the three points of the true logistic function $f(t)$ in (10) to $y_1 = f(0)$, $y_2 = f(t_{\text{obs}}/2)$, and $y_3 = f(t_{\text{obs}})$. Analogously, we denote the corresponding points of the estimate $\hat{f}(t)$, obtained by (11), as $\hat{y}_1 = \hat{f}(0)$, $\hat{y}_2 = \hat{f}(t_{\text{obs}}/2)$, and $\hat{y}_3 = \hat{f}(t_{\text{obs}})$. The points $\hat{y}_1, \hat{y}_2, \hat{y}_3$ depend on the unknown model error $w(t)$. If the model error $w(t) \to 0$ at every time $t \in [0, t_{\text{obs}}]$, then it holds that $\hat{y}_i \to y_i$ for $i = 1, 2, 3$, which implies that $\hat{f}(t) \to f(t)$ at every time $t$.

We consider the *best case*[6] and assume that, due to nonzero model errors $w(t)$, the estimate $\hat{f}(t)$ differs from the true function $f(t)$ in only one of the points $y_1, y_2, y_3$. More precisely, we consider that $\hat{y}_1 = y_1$, $\hat{y}_2 = y_2$, and $\hat{y}_3 = y_3 + \epsilon$ for some small $\epsilon > 0$. Thus, $\epsilon \downarrow 0$ implies that $\hat{f}(t) \to f(t)$ at every time $t$. For now, we focus on the sensitivity of estimating the steady state $y_\infty$. We define $\hat{y}_\infty(\epsilon)$ as the estimate of the steady state $y_\infty$, given the perturbation $\hat{y}_3 = y_3 + \epsilon$. By applying Taylor's Theorem to (14), we obtain for a small $\epsilon > 0$ that

$$\hat{y}_\infty(\epsilon) = y_\infty + \epsilon \kappa_1(t_{\text{obs}}) + O(\epsilon^2), \tag{17}$$

where we define[7] the *condition number* $\kappa_1(t_{\text{obs}})$ as

$$\kappa_1(t_{\text{obs}}) = \frac{\partial}{\partial y_3}\left[y_1 + \frac{(y_1 - y_2)^2}{y_2}\frac{1}{\Phi(y_1, y_2, y_3)}\right]. \tag{18}$$

The condition number $\kappa_1(t_{\text{obs}})$ depends on the observation time $t_{\text{obs}}$, since the three points are given by $y_1 = f(0)$, $y_2 = f(t_{\text{obs}}/2)$, and $y_3 = f(t_{\text{obs}})$. From (17) it follows that the condition number $\kappa_1(t_{\text{obs}})$ describes the impact, or the amplification, of a small error $\epsilon = \hat{y}_3 - y_3$ on the estimate

instance, a short computation yields that the relative approximation error $|f(t) - y_\infty \exp[K(t - t_0)]|/y_\infty \leqslant 0.01$ is attained for all $t \leqslant t_0/2$ if the inflection point satisfies $t_0 \geqslant 2\log(99)/K \approx 9.19/K$.

[6] If instead all three points $i = 1, 2, 3$ are perturbed as $\hat{y}_i = y_i + \epsilon_i$, then the estimate $\hat{f}(t)$ depends on multiple error terms $\epsilon_i$. The distortions on the estimate $\hat{f}(t)$ due to perturbing multiple points might cancel out for specific values of the errors $\epsilon_1, \epsilon_2, \epsilon_3$. However, in most situations the impact of the errors $\epsilon_i$ accumulates, and considering the perturbation of only one point is an optimistic scenario.

[7] For a matrix $A$, the most common definition of the condition number is $\kappa(A) = \sigma_{\max}/\sigma_{\min}$, where $\sigma_{\max}$ and $\sigma_{\min}$ denote the greatest and smallest singular value of the matrix $A$, respectively. Analogously to (18), the condition number $\kappa(A)$ describes the sensitivity the solution $x$ of the linear system $Ax = b$ when the vector $b$ is perturbed [47].

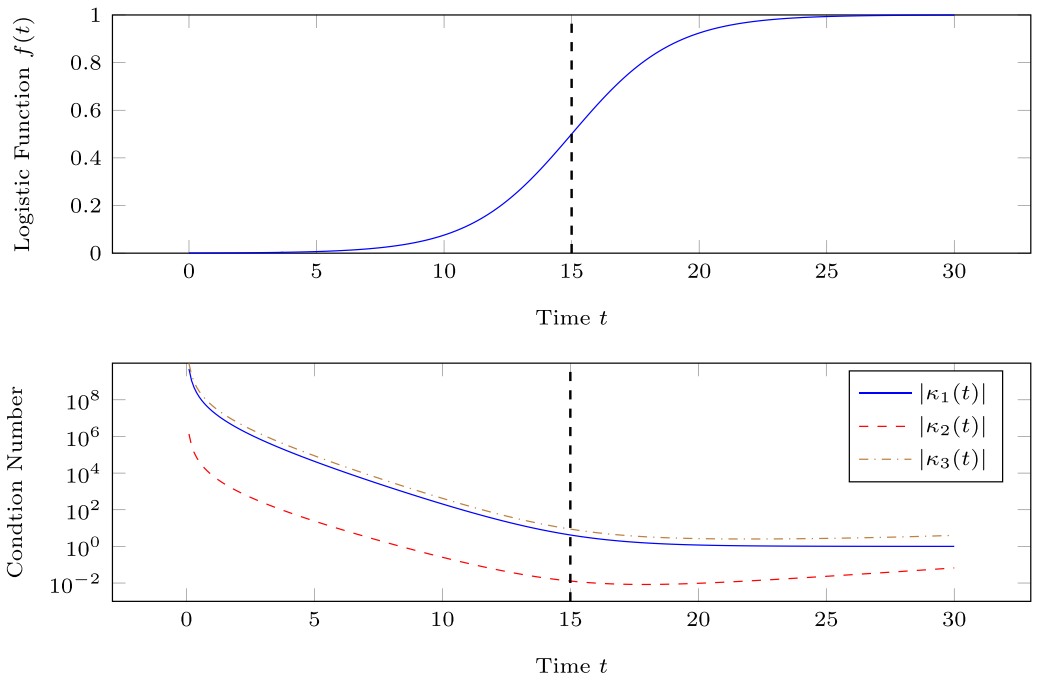


FIG. 4. The condition numbers of estimating the parameters of a logistic function. Upper subplot: The logistic function $f(t)$ vs time $t$. Lower subplot: The absolute value of the condition numbers $\kappa_1(t)$, $\kappa_2(t)$, and $\kappa_3(t)$ vs time $t$ on a semilogarithmic plot. The dashed lines indicate the inflection point $t_0 = 15$.

$\hat{y}_\infty(\epsilon)$. *The greater the condition number $\kappa_1(t_{obs})$, the harder it is to estimate the steady state $y_\infty$.* Analogously to the condition number $\kappa_1(t_{\rm obs})$ for the estimate of the steady state $y_\infty(\epsilon)$, we define the condition numbers $\kappa_2(t_{\rm obs})$ and $\kappa_3(t_{\rm obs})$ for the growth-rate estimate $\hat{K}(\epsilon)$ and the inflection-point estimate $\hat{t}_0(\epsilon)$, respectively. (See also Appendix B.)

*Proposition 8 (Condition numbers of estimating the logistic function parameters).* Consider three points $y_1 = f(0)$, $y_2 = f(t_{\rm obs}/2)$, and $y_3 = f(t_{\rm obs})$ on the logistic function $f(t)$. With respect to a small perturbation $\epsilon$ of the point $y_3$, the condition number of the steady-state estimate $\hat{y}_\infty(\epsilon)$ equals

$$\kappa_1(t_{\rm obs}) = \frac{(y_1 - y_2)^2}{y_3^2} \frac{1}{\Phi^2(y_1, y_2, y_3)}, \tag{19}$$

the condition number of the growth-rate estimate $\hat{K}(\epsilon)$ equals

$$\kappa_2(t_{\rm obs}) = \frac{2}{t_{\rm obs}} \frac{y_2^2}{y_3^2} \frac{1}{y_1 - y_2 + y_2\Phi(y_1, y_2, y_3)}, \tag{20}$$

and the condition number of the inflection-point estimate $\hat{t}_0(\epsilon)$ equals

$$\kappa_3(t_{\rm obs}) = \frac{1}{K} \frac{y_2}{y_3^2} \left[ \frac{1}{\Phi(y_1, y_2, y_3)} - \frac{2t_0 y_2}{t_{\rm obs}} \frac{1}{y_1 - y_2 + y_2\Phi(y_1, y_2, y_3)} \right]. \tag{21}$$

*Proof.* Appendix B. ■

To assess the difficulty of estimating the parameters $y_\infty$, $K$, $t_0$, we consider an exemplary logistic function $f(t)$ with $K = 0.5$, $t_0 = 15$, and $y_\infty = 1$. Figure 4 shows that the condition numbers $\kappa_1(t)$, $\kappa_2(t)$, and $\kappa_3(t)$ are very large. For instance, at time $t = 5 = t_0/3$, the magnitude of the condition number $|\kappa_1(5)|$ is greater than $10^4$. Thus, the steady-state estimate $\hat{y}_\infty(\epsilon)$ is distorted by the error $\epsilon$ times a factor of $10^4$. Furthermore, Fig. 4 indicates that the estimation of the growth-rate parameter $K$ is most robust against model errors $w(t)$, since the condition number $\kappa_2(t)$ is the smallest. We emphasize that, for simplicity, Proposition 8 considers the best case: the perturbation of only one point $y_3$. If the points $y_1$ and $y_2$ are also perturbed, then the condition numbers can become even greater than the expressions in Proposition 8.

The condition numbers in Proposition 8 are given by rather complicated expressions. To obtain a better understanding of the condition numbers, we derive lower bounds as:

*Proposition 9 (Lower bounds on the condition numbers).* Consider three points $y_1 = f(0)$, $y_2 = f(t_{\rm obs}/2)$ and $y_3 = f(t_{\rm obs})$ on the logistic function $f(t)$, whose inflection point $t_0 \geqslant 0$. For every observation time $t_{\rm obs} > 0$, the condition number of the steady-state estimate $\hat{y}_\infty(\epsilon)$ is lower bounded by $\kappa_1(t_{\rm obs}) > \kappa_{1,\rm lb}(t_{\rm obs})$, where

$$\kappa_{1,\rm lb}(t_{\rm obs}) = 1 + \frac{4}{K^2} \frac{1}{t_{\rm obs}^2} e^{2K(t_0 - t_{\rm obs})}, \tag{22}$$

the condition number of the growth-rate estimate $\hat{K}(\epsilon)$ is lower bounded by $|\kappa_2(t_{\rm obs})| > \kappa_{2,\rm lb}(t_{\rm obs})$, where

$$\kappa_{2,\rm lb}(t_{\rm obs}) = \frac{y_2^2}{y_3^2} \frac{1}{y_\infty} \frac{K}{1 + \frac{1}{2}Kt_{\rm obs}} \frac{1}{\Phi(y_1, y_2, y_3)}, \tag{23}$$

and the condition number of the inflection-point estimate $\hat{t}_0(\epsilon)$ is lower bounded by $\kappa_3(t_{\rm obs}) > \kappa_{3,\rm lb}(t_{\rm obs})$, where

$$\kappa_{3,\rm lb}(t_{\rm obs}) = \frac{1}{K} \frac{y_2}{y_3^2} \frac{1}{\Phi(y_1, y_2, y_3)}. \tag{24}$$

*Proof.* Appendix C. ■

Figure 5 shows that the lower bounds of Proposition 9 are accurate, where we use the same parameters for the

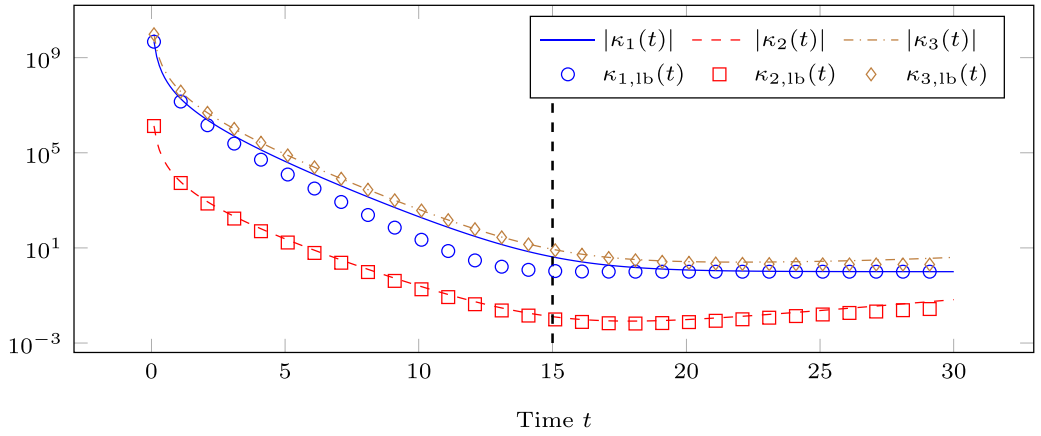


FIG. 5. Lower bounds on the condition numbers: The absolute value of the condition numbers $\kappa_i(t)$, where $i = 1, 2, 3$, and the respective lower bounds $\kappa_{i,\mathrm{lb}}(t)$ vs time $t$ on a semilogarithmic plot. The dashed line indicates the inflection point $t_0 = 15$.

logistic function as in Figure 3. From Proposition 9, we obtain two statements on the prediction limits of epidemic outbreaks. First, the lower bound (22) grows exponentially with $(t_0 - t_{\mathrm{obs}})$. *Thus, only if the epidemic has been observed until the inflection point $t_{obs} \approx t_0$ (or longer), the steady state $y_\infty$ can be estimated accurately.* Second, the lower bounds (23) and (24) depend on the reciprocal of the growth metric $\Phi(y_1, y_2, y_3)$. The more the epidemic growth from $y_1 = f(0)$ to $y_3 = f(t_{\mathrm{obs}})$ resembles an exponential, the smaller the growth metric $\Phi(y_1, y_2, y_3)$; see Sec. IV A. But real epidemics grow practically exponentially in the beginning of the outbreak. *Hence, the growth rate $K$ and the inflection point $t_0$ cannot be estimated accurately at early stages of an epidemic.* Or, as a simple rule of thumb: as long as the infections $\mathcal{I}_c(t)$ are on a straight line in a semilogarithmic plot, the epidemic outbreak cannot be predicted accurately.

We perform numerical simulations to illustrate the sensitivity of predicting an epidemic outbreak subject to model errors $w(t)$. We generate the model errors $w(t)$ in (10) as Gaussian random variables with zero mean and standard deviation $\sigma$. The model errors $w(t)$ and $w(\tilde{t})$ are stochastically independent for all times $t \neq \tilde{t}$. If the cumulative number of infections $\mathcal{I}_c(t)$, resulting from (10), is negative, then we set $\mathcal{I}_c(t) \leftarrow |\mathcal{I}_c(t)|$. Figure 6 shows that small model errors $w(t)$ have a severe impact on the accuracy of the estimated number of infections $\hat{I}_c(t)$ and the inflection-point estimate $\hat{t}_0$. The prediction of the number of infections $I_c(t)$ is accurate only in the short term. We emphasize that, for real epidemics, the model errors $w(t)$ are significantly larger than in Fig. 6.

Figure 7 shows the distribution of the logistic curve parameter estimates $\hat{y}_\infty$, $\hat{K}$ and $\hat{t}_0$. The distribution of the steady-state estimate $\hat{y}_\infty$ and the inflection-point estimate $\hat{t}_0$ is not unimodal, which might be due to multiple local minima of the nonconvex optimization problem (11).

### C. Illustration of the predictability of epidemics for COVID-19 outbreaks

We consider the prediction of the COVID-19 prevalence in several countries: Belgium, Italy, the Netherlands and South Africa. We emphasize that, since the SIS and SIR models in Definitions 1 and 2 assume constant spreading rates $\beta_{ij}$ and $\delta_i$, there are more suitable methods for predicting the COVID-19 prevalence, which take into account time-varying lockdown measures, seasonality, human behavior, and mobility patterns. The main motivation of the following COVID-19 predictions is to provide an illustration of our predictability results. To reduce the impact of large variations in the spreading rates $\beta_{ij}(t)$, $\delta_i(t)$, we confine ourselves to predicting a single COVID-19 wave from $t = 0$ to $t = t_{\mathrm{end}}$. In particular, we assume that from time $t = 0$ to $t = t_{\mathrm{end}}$, the spreading parameters $\beta_{ij}(t)$, $\delta_i(t)$ do not change significantly. Hence, we assume that $\beta_{ij}(t) \approx \beta_{ij}$ and $\delta_i(t) \approx \delta_i$ for $t \in [0, t_{\mathrm{end}}]$ to approximate the outbreak a single COVID-19 wave with time-invariant SIS and SIR models in Definitions 1 and 2.

We obtain the infection data from the COVID-19 Dashboard of the Johns Hopkins University [48] and determine the period of the first wave to be from March 1 until June 16 (Belgium), from February 16 until June 17 (Italy), from February 22 until July 11 (the Netherlands), and from April 30 until October 7 (South Africa). In the subsequent plots, time $t = 0$ denotes the first day of the respective country. For instance, time $t = 0$ corresponds to February 22, 2020, for the Netherlands. We choose the observation time $t_{\mathrm{obs}}$ to be before the peak of the epidemic, i.e., before the inflection point $t_0$. More precisely, we set the observation time $t_{\mathrm{obs}}$ to $t_{\mathrm{end}}/3$, rounded to the next largest integer, where $t_{\mathrm{end}}$ denotes the last considered day. For instance, $t_{\mathrm{end}} = 108$ corresponds to the last day, June 16, for Belgium. In Appendix D we show the impact of choosing the observation time $t_{\mathrm{obs}}$ differently, in particular the predictions become more accurate for a larger observation time $t_{\mathrm{obs}}$.

Figure 8 shows a crucial contrast: the logistic function $f(t)$ fits the number of infections until the observation time $t_{\mathrm{obs}}$. But the logistic function $f(t)$ does not yield accurate predictions for the number of infections. Only short-term predictions, until day $t \approx t_{\mathrm{obs}} + 4$, are possible. One reason for the inaccurate predictions in Fig. 8 is, as we pointed out in the beginning of Sec. III, that the logistic function $f(t)$ does not consider time-varying human mobility patterns, risk-aware adaptive behavior of individuals, the increase of testing capacities, or disease countermeasures that tighten over time $t$. In particular, we observe that the steady state $y_\infty$ underestimates the true long-term number of infections, which might be due to underreported infections in the early stages of the

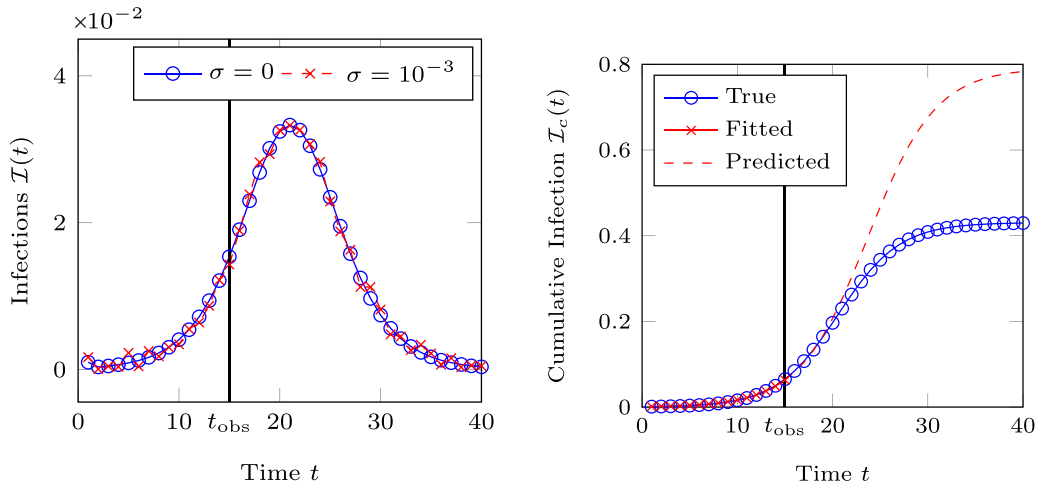


FIG. 6. The sensitivity of predicting an epidemic outbreak. The left subplot shows the logistic function (10) with and without Gaussian model errors $w(t)$ with a standard deviation of $\sigma = 10^{-3}$. The randomly generated parameters of the logistic function $f(t)$ are $t_0 = 20.5$, $K = 0.31$, and $y_\infty = 0.43$. The right subplot shows the cumulative number of infections $\mathcal{I}_c(t)$ and the predicted value $\hat{\mathcal{I}}_c(t)$, based on the logistic function plus model errors $w(t)$.

pandemic because of limited testing capacities. However, we stress that even if the logistic function $f(t)$ were *almost exact*, the prediction of the future course of the outbreak would be difficult, due to the fundamental limits in Sec. IV B.

*When can we trust the predictions to be accurate?* Proposition 8 and Proposition 9 suggest that the growth metric $\Phi(y_1, y_2, y_3)$ is decisive for the prediction accuracy. We compute the prediction accuracy in three steps. First, by fitting a logistic function to the number of infections of the *complete* first wave, we obtain the "exact" steady state $y_\infty$, growth rate $K$, and inflection point $t_0$. Second, to reduce erratic fluctuations, we apply a moving average of window length five to the estimates $\hat{y}_\infty$, $\hat{K}$, $\hat{t}_0$ and the growth metric $\Phi(y_1, y_2, y_3)$. For instance, we replace the steady-state estimate $\hat{y}_\infty(t_{\text{obs}})$ at observation time $t_{\text{obs}}$ by the average of the steady-state estimates $\hat{y}_\infty(t_{\text{obs}})$, $\hat{y}_\infty(t_{\text{obs}} - 1)$,..., $\hat{y}_\infty(t_{\text{obs}} - 4)$. Third, we define the absolute error of the steady-state estimate $\hat{y}_\infty$ as $\varepsilon(y_\infty, \hat{y}_\infty) = |\hat{y}_\infty - y_\infty|$. Analogously, for the growth-rate estimate $\hat{K}$ and the inflection-point estimate $\hat{t}_0$, the respective absolute errors are denoted by $\varepsilon(K, \hat{K})$ and $\varepsilon(t_0, \hat{t}_0)$.

Figures 9 and 10 show that there is a strong correlation between the estimation errors $\varepsilon(K, \hat{K})$, $\varepsilon(t_0, \hat{t}_0)$ and inverse growth metric $\Phi(y_1, y_2, y_3)$, which is in line with Proposition 9. The red lines in Figs. 9–11 are obtain by robust linear regression with the Matlab command `fitlm`. The linear regression is performed without intercept, i.e., the red lines go through the origin. Here we define the *relative error* $\Delta_K$ of the linear regression as the average of the absolute deviation of $\varepsilon(K, \hat{K})$ to the linear curve, divided by the maximum value of the error $\varepsilon(K, \hat{K})$. The relative error $\Delta_{y_\infty}$ and $\Delta_{t_0}$ are defined analogously. Furthermore, Fig. 11 shows[8] that the estimation error $\varepsilon(y_\infty, \hat{y}_\infty)$ of the steady state $y_\infty$ is reasonably correlated with the inverse growth metric $\Phi(y_1, y_2, y_3)$, except for South Africa. An interesting observation is that the fit of the linear regression in Figs. 9–11 seems better for small values of the

[8]For clarity, we removed four outliers from Fig. 11(a) because the axis range would be too large. The linear regression and the relative error $\Delta_{y_\infty}$ consider all points including the outliers.

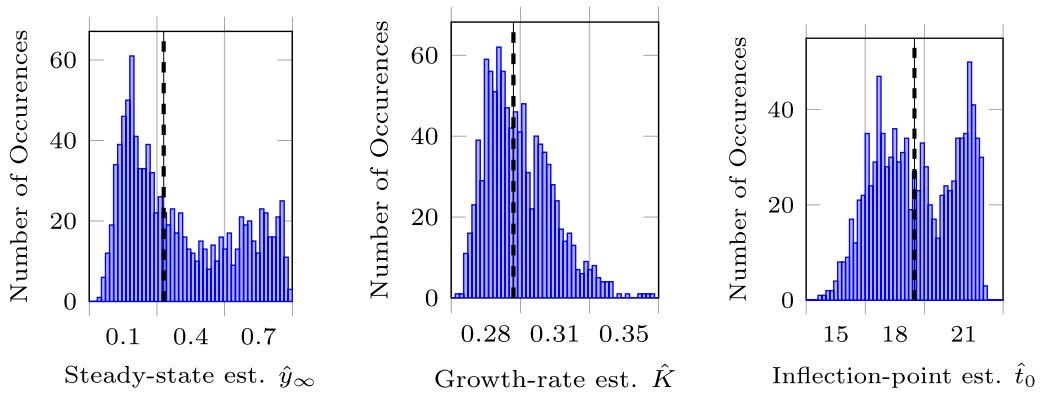


FIG. 7. The distribution of the logistic curve parameter estimates. The left, middle, and right subplots depict the histograms of the steady-state estimate $\hat{y}_\infty$, the growth-rate estimate $\hat{K}$, and the inflection-point estimate $\hat{t}_0$, respectively, which have been obtained by repeating the prediction for 1000 realizations of the model errors $w(t)$. The real values of the parameters $y_\infty$, $K$, and $t_0$ are shown by dashed lines.

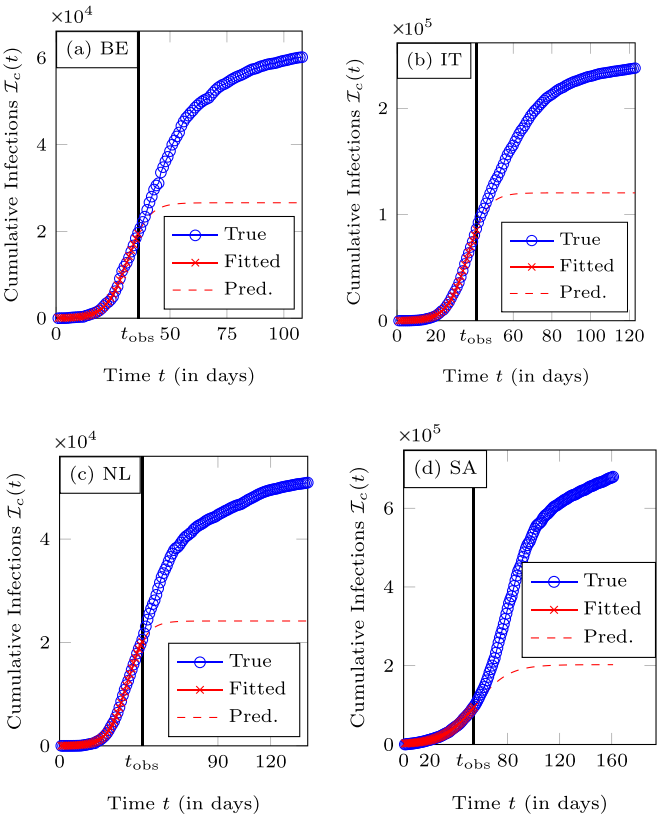


FIG. 8. The difficulty of predicting COVID-19 for four countries: (a) Belgium, (b) Italy, (c) the Netherlands, and (d) South Africa. The blue curves show the cumulative number of the first wave of confirmed infections with SARS-CoV-2. The red curves shows the logistic curve, which is fitted to the infections from day $t = 0$ until day $t_{\text{obs}}$ and used for predictions at times $t > t_{\text{obs}}$.

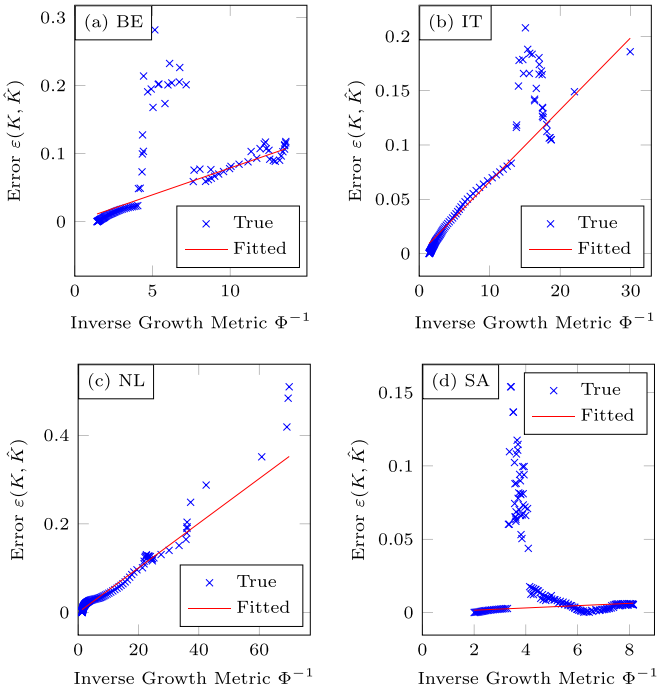


FIG. 9. Assessing the growth-rate estimation accuracy via the growth metric $\Phi$ for four countries: (a) Belgium, (b) Italy, (c) the Netherlands, and (d) South Africa. In blue: The error $\varepsilon(K, \hat{K})$ of growth-rate estimate $\hat{K}$ at different times $t_{\text{obs}}$ vs the inverse of the growth metric $\Phi(y_1, y_2, y_3)$, where $y_1 = \mathcal{I}_c(0)$, $y_2 = \mathcal{I}_c(t_{\text{obs}}/2)$ and $y_3 = \mathcal{I}_c(t_{\text{obs}})$. In red: The curve obtained by linear regression. The relative error $\Delta_K$ of the linear regression equals (a) $\Delta_K = 0.11$, (b) $\Delta_K = 0.03$, (c) $\Delta_K = 0.01$, (d) $\Delta_K = 0.15$.

inverse growth metric $\Phi(y_1, y_2, y_3)$, and the underlying reason is an open question.

We emphasize that the growth metric $\Phi(y_1, y_2, y_3)$ is computed solely based on past data until the observation time $t_{\text{obs}}$. *Hence, it is possible to quantify the prediction accuracy only based on past data.* For instance, suppose that the growth metric $\Phi(y_1, y_2, y_3)$ increases by a factor $\mu$ from time $t_{\text{obs}}$ to $\tilde{t}_{\text{obs}} > t_{\text{obs}}$. Then we can expect that the accuracy of the estimates $\hat{K}$, $\hat{t}_0$ and $\hat{y}_\infty$ increases by the factor $\mu$.

## V. CONCLUSIONS

For many epidemic models on time-invariant networks, the cumulative number of infections resembles a logistic function, at least approximately. In this work, we showed that the prediction of a logistic function is ill-conditioned. More specifically, a good fit of a logistic function $\hat{f}(t)$ to the epidemic data until some observation time $t_{\text{obs}}$ does *not* imply that the function $\hat{f}(t)$ yields accurate predictions at times $t > t_{\text{obs}}$. Hence, even under idealized conditions, the prediction of an epidemic is inherently difficult, regardless of the particular prediction algorithm.

Furthermore, we introduced the growth metric $\Phi(y_1, y_2, y_3)$, which quantifies the exponential growth of the epidemic. The more exponential the epidemic growth, the more difficult the prediction of the epidemic, provided that the epidemic approximately follows a logistic function $f(t)$. In particular, the estimation error of the epidemic parameters correlates strongly with the inverse of the growth metric $\Phi(y_1, y_2, y_3)$, which enables quantitative statements on the prediction accuracy: Suppose that the epidemic is predicted at two different observation time $t_{\text{obs}}$ and $\tilde{t}_{\text{obs}} > t_{\text{obs}}$. Then the fraction of the respective growth metrics $\Phi(y_1, y_2, y_3)$, $\tilde{\Phi}(y_1, y_2, y_3)$ approximates the change of the prediction accuracy from time $t_{\text{obs}}$ to $\tilde{t}_{\text{obs}}$.

Last, the SIS and SIR epidemic models from Definitions 1 and 2 can be extended to more realistic models. Particular extensions include: considering more compartments, such as in the susceptible-exposed-infected-removed (SEIR) epidemic model; accounting for *time-varying* spreading rates $\delta_i(t)$, $\beta_{ij}(t)$; non-Markovian epidemic models; or adaptive networks, where the contact between two individuals depend on the probability that the respective individuals are infected. While extending the SIS and SIR model admittedly may result in more realistic models, the extensions complicate the epidemic models. Hence, in view of the prediction limits for the simpler SIS and SIR epidemic models in Definitions 1 and 2, we do not expect that considering more complex epidemic models resolves the problem of obtaining accurate, long-term predictions.

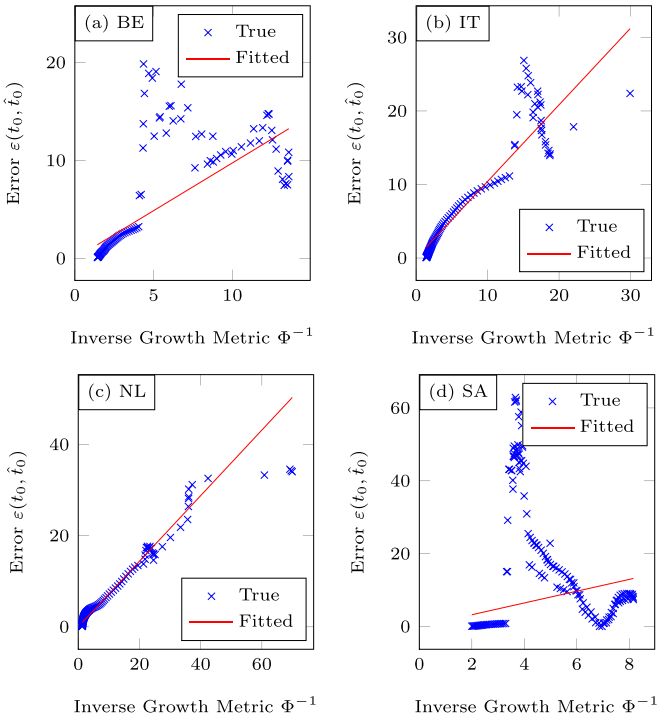


FIG. 10. Assessing the inflection-point estimation accuracy via the growth metric $\Phi$ for four countries: (a) Belgium, (b) Italy, (c) the Netherlands, and (d) South Africa. In blue: The error $\varepsilon(t_0, \hat{t}_0)$ of inflection-point estimate $\hat{t}_0$ at different times $t_{\text{obs}}$ vs the inverse of the growth metric $\Phi(y_1, y_2, y_3)$, where $y_1 = \mathcal{I}_c(0)$, $y_2 = \mathcal{I}_c(t_{\text{obs}}/2)$ and $y_3 = \mathcal{I}_c(t_{\text{obs}})$. In red: The curve obtained by linear regression. The relative error $\Delta_{t_0}$ of the linear regression equals (a) $\Delta_{t_0} = 0.14$, (b) $\Delta_{t_0} = 0.05$, (c) $\Delta_{t_0} = 0.03$, (d) $\Delta_{t_0} = 0.24$.

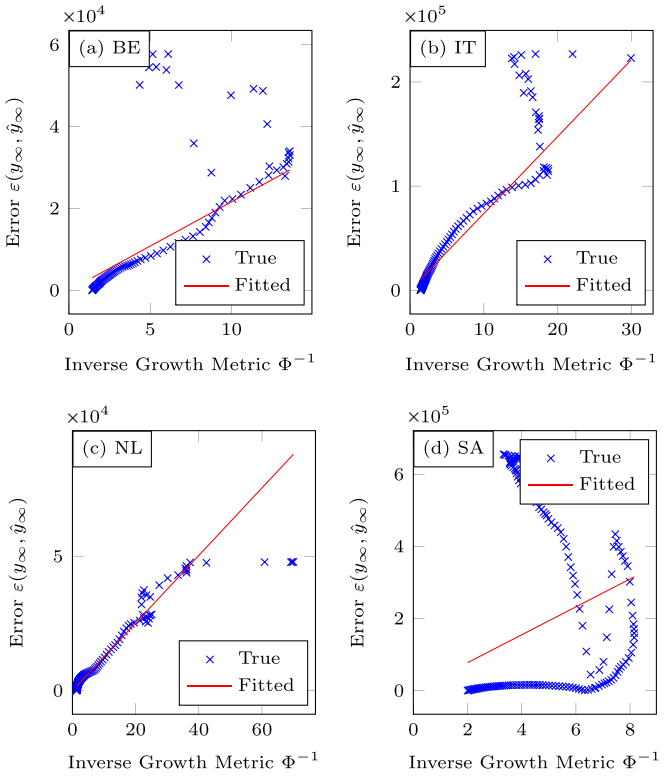


FIG. 11. Assessing the steady-state estimation accuracy via the growth metric $\Phi$ for four countries: (a) Belgium, (b) Italy, (c) the Netherlands, (d) and South Africa. In blue: The error $\varepsilon(y_\infty, \hat{y}_\infty)$ of steady-state estimate $\hat{y}_\infty$ at different times $t_{\text{obs}}$ vs the inverse of the growth metric $\Phi(y_1, y_2, y_3)$, where $y_1 = \mathcal{I}_c(0)$, $y_2 = \mathcal{I}_c(t_{\text{obs}}/2)$ and $y_3 = \mathcal{I}_c(t_{\text{obs}})$. In red: The curve obtained by linear regression. The relative error $\Delta_{y_\infty}$ of the linear regression equals (a) $\Delta_{y_\infty} = 0.01$, (b) $\Delta_{y_\infty} = 0.08$, (c) $\Delta_{y_\infty} = 2.8 \times 10^{-4}$, (d) $\Delta_{y_\infty} = 0.36$.

## APPENDIX A: PROOF OF PROPOSITION 7

Proposition 7 states there exists a logistic function $f(t)$ with $f(0) = y_1$, $f(\Delta t) = y_2$ and $f(2\Delta t) = y_3$ if and only if $\Phi(y_1, y_2, y_3) > 0$. Appendix A 1 shows the "only if" direction: if the three points $y_1, y_2, y_3$ satisfy $f(0) = y_1$, $f(\Delta t) = y_2$ and $f(2\Delta t) = y_3$ for some logistic function $f(t)$, then it holds that $\Phi(y_1, y_2, y_3) > 0$. In Appendix A 2 we prove the "if" direction: for any three points $y_1, y_2, y_3$ that satisfy $\Phi(y_1, y_2, y_3) > 0$, we construct a logistic function $f(t)$ with $f(0) = y_1$, $f(\Delta t) = y_2$ and $f(2\Delta t) = y_3$.

### 1. First part

*Lemma 10.* For some time spacing $\Delta t > 0$, consider three points $y_1 = f(0)$, $y_2 = f(\Delta t)$ and $y_3 = f(2\Delta t)$ on a logistic function $f(t)$. Then the growth metric $\Phi(y_1, y_2, y_3)$ defined in (12) equals

$$\Phi(y_1, y_2, y_3) = \frac{e^{Kt_0}}{1 + e^{Kt_0}} \frac{(1 - e^{-K\Delta t})^2}{1 + e^{-K(\Delta t - t_0)}}, \tag{A1}$$

which implies that $0 < \Phi(y_1, y_2, y_3) < 1$. ■

*Proof.* Since $y_1 = f(0)$, $y_2 = f(\Delta t)$ and $y_3 = f(2\Delta t)$, we obtain from the definition of the logistic function $f(t)$ in (5) that

$$y_1 = \frac{y_\infty}{1 + e^{Kt_0}},$$

$$y_2 = \frac{y_\infty}{1 + e^{-K(\Delta t - t_0)}},$$

and

$$y_3 = \frac{y_\infty}{1 + e^{-K(2\Delta t - t_0)}}. \tag{A2}$$

We define the two constants $\alpha$ and $c$ as

$$\alpha = e^{Kt_0} \tag{A3}$$

and

$$c = e^{-K\Delta t}. \tag{A4}$$

Thus, we can write the three points $y_1$, $y_2$ and $y_3$ more compactly as

$$y_1 = \frac{y_\infty}{1 + \alpha}, \tag{A5}$$

$$y_2 = \frac{y_\infty}{1 + \alpha c}, \tag{A6}$$

and

$$y_3 = \frac{y_\infty}{1 + \alpha c^2}.$$

From the definition of the growth metric $\Phi(y_1, y_2, y_3)$ in (12), we obtain that

$$\Phi(y_1, y_2, y_3) = \frac{1+\alpha c^2}{1+\alpha c} - \frac{1+\alpha c}{1+\alpha} = \frac{(1+\alpha c^2)(1+\alpha) - (1+\alpha c)^2}{(1+\alpha c)(1+\alpha)}.$$

Hence, it holds that

$$\Phi(y_1, y_2, y_3) = \frac{1+\alpha+\alpha c^2+\alpha^2 c^2 - 1 - 2\alpha c - \alpha^2 c^2}{(1+\alpha c)(1+\alpha)},$$

which simplifies to

$$\Phi(y_1, y_2, y_3) = \frac{\alpha}{1+\alpha}\frac{1}{1+\alpha c}(1-c)^2. \tag{A7}$$

Since $\alpha > 0$ and $c > 0$, we obtain that $\Phi(y_1, y_2, y_3) > 0$. Furthermore, $\Delta t > 0$ implies that $c < 1$. Thus, we obtain from (A7) that $\Phi(y_1, y_2, y_3) < 1$. To finish the proof, we substitute $\alpha$, $c$ in (A7) and arrive at (A1). ■

### 2. Second part

For $i = 1, 2, 3$, the point $y_i$ is on the logistic function (5) if and only if

$$y_i + y_i e^{-K((i-1)\Delta t - t_0)} - y_\infty = 0.$$

Dividing by $y_i$ yields that

$$e^{-K((i-1)\Delta t - t_0)} - \frac{1}{y_i}y_\infty + 1 = 0.$$

Thus, we arrive at a set of three nonlinear equations

$$e^{Kt_0}e^{-K(i-1)\Delta t} - \frac{1}{y_i}y_\infty + 1 = 0, \quad i = 1, 2, 3. \tag{A8}$$

With (A3) and (A4), we can express the second exponential in (A8) as

$$e^{-K(i-1)\Delta t} = \begin{cases} 1 & \text{if } i = 1, \\ c & \text{if } i = 2, \\ c^2 & \text{if } i = 3. \end{cases}$$

Then we obtain from (A8) a set of nonlinear equations for the three unknowns $\alpha$, $c$ and $y_\infty$ as

$$\alpha - \frac{1}{y_1}y_\infty + 1 = 0, \tag{A9}$$

$$\alpha c - \frac{1}{y_2}y_\infty + 1 = 0, \tag{A10}$$

$$\alpha c^2 - \frac{1}{y_3}y_\infty + 1 = 0. \tag{A11}$$

The first equation (A9) yields that

$$y_\infty = y_1(\alpha + 1). \tag{A12}$$

Combining (A12) with the second equation (A10) gives that

$$\alpha c - \frac{y_1}{y_2}(\alpha+1) + 1 = 0,$$

from which we obtain that

$$c = \frac{1}{\alpha}\left(\frac{y_1}{y_2}(\alpha+1) - 1\right).$$

Hence, it holds that

$$c = \frac{1}{\alpha}\left(\frac{y_1}{y_2} - 1\right) + \frac{y_1}{y_2}. \tag{A13}$$

Combining the expressions for $y_\infty$ and $c$ in (A12) and (A13), respectively, with the third equation (A11) yields that

$$\alpha\left[\frac{1}{\alpha}\left(\frac{y_1}{y_2} - 1\right) + \frac{y_1}{y_2}\right]^2 - \frac{y_1}{y_3}(\alpha+1) + 1 = 0,$$

which is equivalent to

$$\frac{1}{\alpha}\left(\frac{y_1}{y_2} - 1\right)^2 + 2\left(\frac{y_1}{y_2} - 1\right)\frac{y_1}{y_2} + \alpha\frac{y_1^2}{y_2^2} - \frac{y_1}{y_3}(\alpha+1) + 1 = 0.$$

Multiplication with $\alpha$ and rearranging gives that

$$\alpha^2\left(\frac{y_1^2}{y_2^2} - \frac{y_1}{y_3}\right) + \alpha\left[2\frac{y_1}{y_2}\left(\frac{y_1}{y_2} - 1\right) - \frac{y_1}{y_3} + 1\right] + \left(\frac{y_1}{y_2} - 1\right)^2 = 0. \tag{A14}$$

The quadratic equation (A14) has two solutions. The first solution is $\alpha = -1$ leads to a contradiction, since $\alpha$, defined in (A3), is positive. The second solution of (A14) is

$$\alpha = -\frac{\left(\frac{1}{y_2} - \frac{1}{y_1}\right)^2}{\frac{1}{y_2^2} - \frac{1}{y_1 y_3}},$$

which is equivalent to

$$\alpha = \frac{(y_1 - y_2)^2}{y_1 y_2}\frac{1}{\frac{y_2}{y_3} - \frac{y_1}{y_2}}.$$

Thus, we obtain with the definition of the growth metric $\Phi(y_1, y_2, y_3)$ in (12) that

$$\alpha = \frac{(y_1 - y_2)^2}{y_1 y_2}\frac{1}{\Phi(y_1, y_2, y_3)}. \tag{A15}$$

Since $y_1 > 0$ and $y_2 > 0$, the expression (A15) for $\alpha$ is positive only if

$$\Phi(y_1, y_2, y_3) > 0.$$

Hence, if and only if (13) holds true, there is a solution for the unknown $\alpha$, and, hence, for the logistic growth rate $K$ and the inflection point $t_0$. From (A15) and (A12), we obtain the steady state $y_\infty$ as

$$y_\infty = y_1 + \frac{(y_1 - y_2)^2}{y_2}\frac{1}{\Phi(y_1, y_2, y_3)}.$$

From (A13) and (A15), it follows that the unknown $c$ equals

$$c = \frac{y_1}{y_2} + \left(\frac{y_1}{y_2} - 1\right)\frac{y_1 y_2}{(y_1 - y_2)^2}\Phi(y_1, y_2, y_3),$$

which simplifies to

$$c = \frac{y_1}{y_2} + \frac{y_1}{y_1 - y_2}\Phi(y_1, y_2, y_3). \tag{A16}$$

The definition of $c$ in (A4) implies that

$$K = -\frac{1}{\Delta t}\log(c),$$

which yields with (A16) that

$$K = -\frac{1}{\Delta t}\log\left[\frac{y_1}{y_2} + \frac{y_1}{y_1 - y_2}\Phi(y_1, y_2, y_3)\right].$$

Finally, we obtain the inflection point $t_0$ from (A3) as

$$t_0 = \frac{1}{K}\log(\alpha) = \frac{1}{K}\log\left[\frac{(y_1 - y_2)^2}{y_1 y_2}\frac{1}{\Phi(y_1, y_2, y_3)}\right],$$

where the last equality follows from (A15).

## APPENDIX B: PROOF OF PROPOSITION 8

### 1. Condition number of estimating the steady state

From the definition of the condition number $\kappa_1(t_{\rm obs})$ in (18), we obtain that

$$\kappa_1(t_{\rm obs}) = -\frac{(y_1 - y_2)^2}{y_2}\frac{1}{\Phi^2(y_1, y_2, y_3)}\frac{\partial\Phi(y_1, y_2, y_3)}{\partial y_3}.$$

The definition of the growth metric $\Phi(y_1, y_2, y_3)$ in (12) yields that

$$\frac{\partial\Phi(y_1, y_2, y_3)}{\partial y_3} = -\frac{y_2}{y_3^2}. \tag{B1}$$

Thus, the condition number $\kappa_1(t_{\rm obs})$ follows as

$$\kappa_1(t_{\rm obs}) = \frac{(y_1 - y_2)^2}{y_3^2}\frac{1}{\Phi^2(y_1, y_2, y_3)}.$$

### 2. Condition number of estimating the logistic growth rate

With (15), we define the condition number $\kappa_2(t_{\rm obs})$ with respect to the growth-rate estimate $\hat{K}(t_{\rm obs})$ as

$$\kappa_2(t_{\rm obs}) = \frac{\partial}{\partial y_3}\left[-\frac{1}{\Delta t}\log\left(\frac{y_1}{y_2} + \frac{y_1}{y_1 - y_2}\Phi(y_1, y_2, y_3)\right)\right],$$

where $\Delta t = t_{\rm obs}/2$. Hence, it holds that

$$\kappa_2(t_{\rm obs}) = -\frac{1}{\Delta t}\frac{1}{\frac{y_1}{y_2} + \frac{y_1}{y_1 - y_2}\Phi(y_1, y_2, y_3)}\frac{y_1}{y_1 - y_2}\frac{\partial}{\partial y_3} \times \Phi(y_1, y_2, y_3).$$

Thus, we obtain with (B1) that

$$\kappa_2(t_{\rm obs}) = \frac{1}{\Delta t}\frac{1}{\frac{y_1}{y_2} - 1 + \Phi(y_1, y_2, y_3)}\frac{y_2}{y_3^2},$$

which simplifies to

$$\kappa_2(t_{\rm obs}) = \frac{1}{\Delta t}\frac{y_2^2}{y_3^2}\frac{1}{y_1 - y_2 + y_2\Phi(y_1, y_2, y_3)}. \tag{B2}$$

The expression (20) for the condition number $\kappa_2(t_{\rm obs})$ follows from $\Delta t = t_{\rm obs}/2$.

### 3. Condition number of estimating the inflection point

With (16), we define the condition number $\kappa_3(t_{\rm obs})$ with respect to the inflection-point estimate $\hat{t}_0(t_{\rm obs})$ as

$$\kappa_3(t_{\rm obs}) = \frac{\partial}{\partial y_3}\left[\frac{1}{K}\log\left(\frac{(y_1 - y_2)^2}{y_1 y_2}\frac{1}{\Phi(y_1, y_2, y_3)}\right)\right],$$

which becomes

$$\kappa_3(t_{\rm obs}) = -\frac{1}{K^2}\log\left[\frac{(y_1 - y_2)^2}{y_1 y_2}\frac{1}{\Phi(y_1, y_2, y_3)}\right]\frac{\partial K}{\partial y_3} - \frac{1}{K}\frac{1}{\frac{(y_1 - y_2)^2}{y_1 y_2}\frac{1}{\Phi(y_1, y_2, y_3)}}\frac{(y_1 - y_2)^2}{y_1 y_2}\frac{1}{\Phi^2(y_1, y_2, y_3)}\frac{\partial}{\partial y_3} \times \Phi(y_1, y_2, y_3).$$

Thus, it holds that

$$\kappa_3(t_{\rm obs}) = -\frac{1}{K^2}\log\left[\frac{(y_1 - y_2)^2}{y_1 y_2}\frac{1}{\Phi(y_1, y_2, y_3)}\right]\frac{\partial K}{\partial y_3} - \frac{1}{K}\frac{1}{\Phi(y_1, y_2, y_3)}\frac{\partial}{\partial y_3}\Phi(y_1, y_2, y_3).$$

With (16), (B1), and (B2), we obtain that

$$\kappa_3(t_{\rm obs}) = -\frac{1}{K}t_0\frac{1}{\Delta t}\frac{y_2^2}{y_3^2}\frac{1}{y_1 - y_2 + y_2\Phi(y_1, y_2, y_3)} + \frac{1}{K}\frac{1}{\Phi(y_1, y_2, y_3)}\frac{y_2}{y_3^2},$$

which simplifies to

$$\kappa_3(t_{\rm obs}) = \frac{1}{K}\frac{y_2}{y_3^2}\left[\frac{1}{\Phi(y_1, y_2, y_3)} - \frac{t_0 y_2}{\Delta t}\frac{1}{y_1 - y_2 + y_2\Phi(y_1, y_2, y_3)}\right].$$

The expression (21) for the condition number $\kappa_3(t_{\rm obs})$ follows from $\Delta t = t_{\rm obs}/2$.

## APPENDIX C: PROOF OF PROPOSITION 9

### 1. Auxiliary lemmas

*Lemma 11.* For some observation time $t_{\rm obs} > 0$, consider three points $y_1 = f(0)$, $y_2 = f(t_{\rm obs}/2)$ and $y_3 = f(t_{\rm obs})$ on a logistic function $f(t)$. Then the difference of the points $y_2$ and $y_1$ equals

$$y_2 - y_1 = \frac{y_\infty}{1 - e^{-\frac{1}{2}Kt_{\rm obs}}}\Phi(y_1, y_2, y_3).$$

*Proof.* From (A5) and (A6), we obtain that

$$y_2 - y_1 = y_\infty\left(\frac{1}{1 + \alpha c} - \frac{1}{1 + \alpha}\right),$$

where $\alpha$ and $c$ are defined by (A3) and (A4). We simplify and obtain that

$$y_2 - y_1 = y_\infty\frac{\alpha}{1 + \alpha}\frac{1 - c}{1 + \alpha c}.$$

Comparing with (A7) yields that

$$y_2 - y_1 = \frac{y_\infty}{1 - c}\Phi(y_1, y_2, y_3).$$

■

### 2. Lower bound for the condition number of estimating the steady state

From Lemma 11, we obtain that the condition number $\kappa_1(t_{\text{obs}})$ in (19) is equal to

$$\kappa_1(t_{\text{obs}}) = \frac{y_\infty^2}{y_3^2} \frac{1}{\left(1 - e^{-\frac{1}{2}Kt_{\text{obs}}}\right)^2}.$$

From the expression for $y_3$ in (26) and $2\Delta t = t_{\text{obs}}$, it follows that

$$\kappa_1(t_{\text{obs}}) = \left(\frac{1 + e^{-K(t_{\text{obs}} - t_0)}}{1 - e^{-\frac{1}{2}Kt_{\text{obs}}}}\right)^2.$$

Hence, we obtain that

$$\begin{aligned}\kappa_1(t_{\text{obs}}) &\geqslant \left(\frac{1}{1 - e^{-\frac{1}{2}Kt_{\text{obs}}}}\right)^2 + \left(\frac{e^{-K(t_{\text{obs}} - t_0)}}{1 - e^{-\frac{1}{2}Kt_{\text{obs}}}}\right)^2 \\ &\geqslant 1 + \left(\frac{e^{-K(t_{\text{obs}} - t_0)}}{1 - e^{-\frac{1}{2}Kt_{\text{obs}}}}\right)^2. \end{aligned} \tag{C1}$$

A basic inequality [49] for the exponential function is $e^{-x} \geqslant 1 - x$ for all $x \in \mathbb{R}$. Hence, the denominator in (C1) is bounded by

$$1 - e^{-\frac{1}{2}Kt_{\text{obs}}} \leqslant 1 - \left(1 - \frac{1}{2}Kt_{\text{obs}}\right) = \frac{1}{2}Kt_{\text{obs}},$$

which finally implies that

$$\kappa_1(t_{\text{obs}}) \geqslant 1 + \frac{4}{K^2 t_{\text{obs}}^2} e^{-2K(t_{\text{obs}} - t_0)}.$$

### 3. Lower bound for the condition number of estimating the logistic growth rate

We consider the denominator of the last factor in (20), which equals

$$\begin{aligned} &y_1 - y_2 + y_2\Phi(y_1, y_2, y_3) \\ &\quad = -(y_2 - y_1)\left[1 - \frac{y_2}{y_2 - y_1}\Phi(y_1, y_2, y_3)\right]. \end{aligned}$$

With Lemma 11 we obtain that

$$\begin{aligned} &y_1 - y_2 + y_2\Phi(y_1, y_2, y_3) \\ &\quad = -(y_2 - y_1)\left[1 - \frac{y_2}{y_\infty}\left(1 - e^{-\frac{1}{2}Kt_{\text{obs}}}\right)\right]. \end{aligned}$$

Since $y_2 > y_1$, $y_\infty > y_2$ and $t_{\text{obs}} > 0$, it holds that

$$y_1 - y_2 + y_2\Phi(y_1, y_2, y_3) < 0.$$

Thus, it follows from (20) that

$$\begin{aligned} |\kappa_2(t_{\text{obs}})| &= -\kappa_2(t_{\text{obs}}) \\ &= \frac{2}{t_{\text{obs}}} \frac{y_2^2}{y_3^2} \frac{1}{y_2 - y_1 - y_2\Phi(y_1, y_2, y_3)}. \end{aligned}$$

With Lemma 11

$$\begin{aligned} |\kappa_2(t_{\text{obs}})| &= \frac{2}{t_{\text{obs}}} \frac{y_2^2}{y_3^2} \left(\frac{y_\infty}{1 - e^{-\frac{1}{2}Kt_{\text{obs}}}} - y_2\right)^{-1} \frac{1}{\Phi(y_1, y_2, y_3)} \\ &= \frac{2}{t_{\text{obs}}} \frac{y_2^2}{y_3^2} \frac{1 - e^{-\frac{1}{2}Kt_{\text{obs}}}}{y_\infty - y_2\left(1 - e^{-\frac{1}{2}Kt_{\text{obs}}}\right)} \frac{1}{\Phi(y_1, y_2, y_3)}. \end{aligned}$$

Since

$$y_2\left(1 - e^{-\frac{1}{2}Kt_{\text{obs}}}\right) > 0,$$

it holds that

$$|\kappa_2(t_{\text{obs}})| > \frac{2}{t_{\text{obs}}} \frac{y_2^2}{y_3^2} \frac{1}{y_\infty} \left(1 - e^{-\frac{1}{2}Kt_{\text{obs}}}\right) \frac{1}{\Phi(y_1, y_2, y_3)}. \tag{C2}$$

To further bound (C2), we consider the term

$$\frac{2}{t_{\text{obs}}}\left(1 - e^{-\frac{1}{2}Kt_{\text{obs}}}\right) = K\frac{1 - e^{-\xi}}{\xi}, \tag{C3}$$

where $\xi = \frac{1}{2}Kt_{\text{obs}}$. Since $\xi > -1$, we obtain that

$$K\frac{1 - e^{-\xi}}{\xi} > K\frac{1}{1 + \xi}.$$

Thus, with (C3) and the definition of $\xi$, we obtain that

$$\frac{2}{t_{\text{obs}}}\left(1 - e^{-\frac{1}{2}Kt_{\text{obs}}}\right) \geqslant K\frac{1}{1 + \frac{1}{2}Kt_{\text{obs}}}.$$

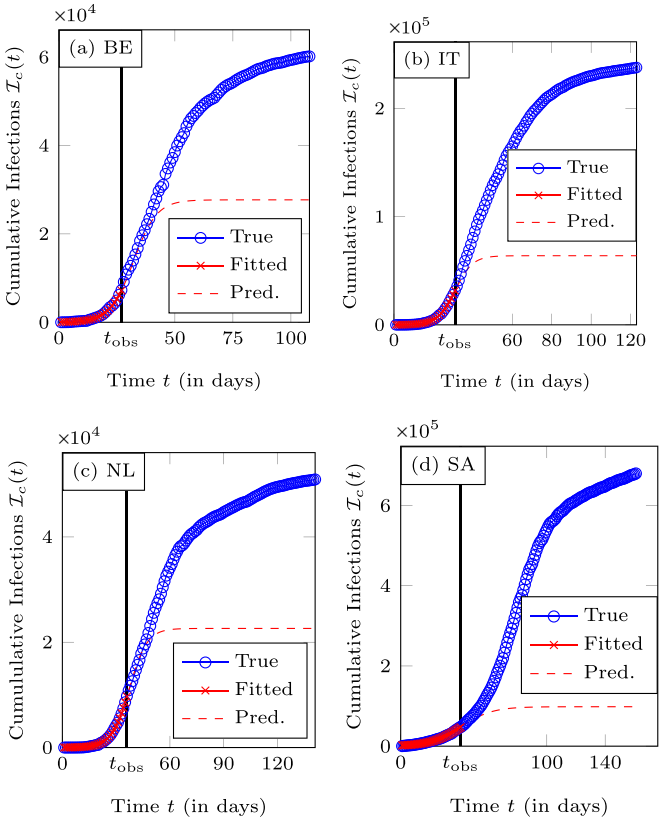


FIG. 12. Predicting COVID-19 with fewer observations for four countries: (a) Belgium, (b) Italy, (c) the Netherlands, and (d) South Africa. The blue curves show the cumulative number of the first wave of confirmed infections with SARS-CoV-2. The red curves show the logistic curve, which is fitted to the infections from day $t = 0$ until day $t_{\text{obs}}$ and used for predictions at times $t > t_{\text{obs}}$, where the observation time $t_{\text{obs}}$ equals $t_{\text{end}}/4$, rounded to the next largest integer.

Finally, (C2) yields that

$$|\kappa_2(t_{\rm obs})| > \frac{y_2^2}{y_3^2}\frac{1}{y_\infty}\frac{K}{1+\frac{1}{2}Kt_{\rm obs}}\frac{1}{\Phi(y_1,y_2,y_3)}.$$

#### 4. Lower bound for the condition number of estimating the inflection point

With (20), the expression for the condition number $\kappa_3(t_{\rm obs})$ in (21) is equivalent to

$$\kappa_3(t_{\rm obs}) = \frac{1}{K}\frac{y_2}{y_3^2}\left(\frac{1}{\Phi(y_1,y_2,y_3)} - \frac{t_0 y_3^2}{y_2}\kappa_2(t_{\rm obs})\right).$$

Since $\kappa_2(t_{\rm obs}) < 0$, we obtain a lower bound as

$$\kappa_3(t_{\rm obs}) > \frac{1}{K}\frac{y_2}{y_3^2}\frac{1}{\Phi(y_1,y_2,y_3)}.$$

## APPENDIX D: VARYING THE OBSERVATION TIME

We repeat the COVID-19 predictions in Sec. IV C with different choices for the observation time $t_{\rm obs}$. Similarly to Fig. 8 for which $t_{\rm obs} = t_{\rm end}/3$, Figs. 12 and 13 correspond to a shorter observation time of $t_{\rm obs} = t_{\rm end}/4$ and a longer observation time $t_{\rm obs} = t_{\rm end}/2$ respectively. Figures 8, 12, and 13 show that the longer the observation time $t_{\rm obs}$, the more accurate the prediction of the viral outbreak.

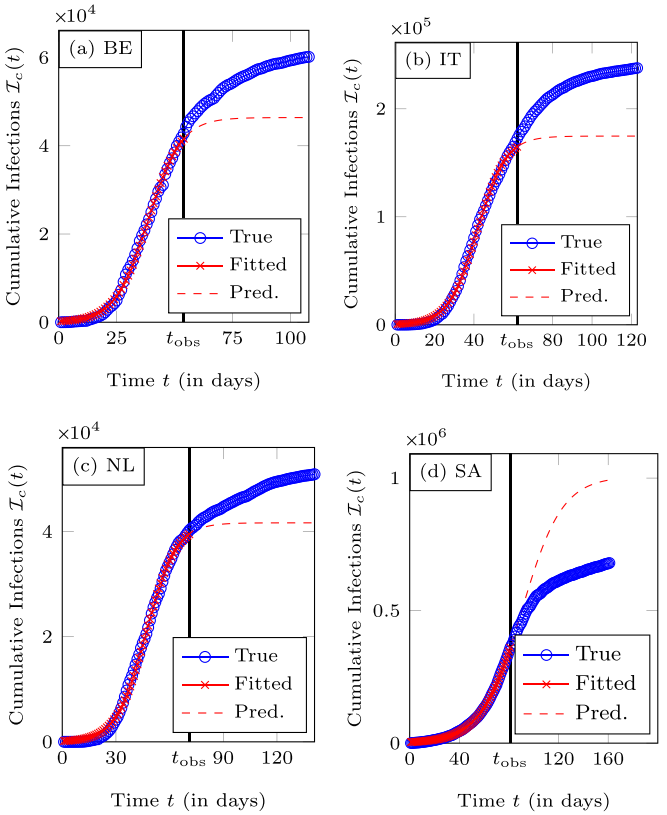


FIG. 13. Predicting COVID-19 with more observations for four countries: (a) Belgium, (b) Italy, (c) the Netherlands, and (d) South Africa.The blue curves show the cumulative number of the first wave of confirmed infections with SARS-CoV-2. The red curves show the logistic curve, which is fitted to the infections from day $t = 0$ until day $t_{\rm obs}$ and used for predictions at times $t > t_{\rm obs}$, where the observation time $t_{\rm obs}$ equals $t_{\rm end}/2$, rounded to the next largest integer.


[1] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases and Its Applications*, 2nd ed. (Charles Griffin, London, 1975).

[2] R. M. Anderson and R. M. May, *Infectious Diseases in Humans* (Oxford University Press, Oxford, 1992).

[3] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. **87**, 925 (2015).

[4] C. Nowzari, V. M. Preciado, and G. J. Pappas, Analysis and control of epidemics: A survey of spreading processes on complex networks, IEEE Control Syst. Mag. **36**, 26 (2016).

[5] I. Z. Kiss, J. C. Miller, and P. L. Simon, *Mathematics of Epidemics on Networks* (Springer, Cham, 2017).

[6] H. W. Hethcote, Qualitative analyses of communicable disease models, Math. Biosci. **28**, 335 (1976).

[7] A. Lajmanovich and J. A. Yorke, A deterministic model for gonorrhea in a nonhomogeneous population, Math. Biosci. **28**, 221 (1976).

[8] P. Van Mieghem, J. Omic, and R. Kooij, Virus spread in networks, IEEE/ACM Trans. Netw. **17**, 1 (2009).

[9] W. O. Kermack and A. G. McKendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. London A **115**, 700 (1927).

[10] M. Youssef and C. Scoglio, An individual-based approach to SIR epidemics in contact networks, J. Theor. Biol. **283**, 136 (2011).

[11] A. Khanafer and T. Başar, On the optimal control of virus spread in networks, in *2014 7th International Conference on Network Games, Control and Optimization (NetGCoop)* (IEEE, New York, 2014), pp. 166–172.

[12] P. Van Mieghem, SIS epidemics with time-dependent rates describing ageing of information spread and mutation of pathogens, Technical Report (Delft University of Technology, 2014), https://nas.ewi.tudelft.nl/people/Piet/papers/TUD20140615_SIS_aging.pdf.

[13] D. Guo, S. Trajanovski, R. van de Bovenkamp, H. Wang, and P. Van Mieghem, Epidemic threshold and topological structure of susceptible-infectious-susceptible epidemics in adaptive networks, Phys. Rev. E **88**, 042802 (2013).

[14] M. A. Achterberg, J. L. A. Dubbeldam, C. J. Stam, and P. Van Mieghem, Classification of link-breaking and link-creation updating rules in susceptible-infected-susceptible epidemics on adaptive networks, Phys. Rev. E **101**, 052302 (2020).

[15] P. Van Mieghem and R. van de Bovenkamp, Non-Markovian Infection Spread Dramatically Alters the Susceptible-Infected-

Susceptible Epidemic Threshold in Networks, Phys. Rev. Lett. **110**, 108701 (2013).
[16] P. Cirillo and N. N. Taleb, Tail risk of contagious diseases, Nat. Phys. **16**, 606 (2020).
[17] M. Castro, S. Ares, J. A. Cuesta, and S. Manrubia, The turning point and end of an expanding epidemic cannot be precisely forecast, Proc. Natl. Acad. Sci. USA **117**, 26190 (2020).
[18] M. Paggi, Simulation of Covid-19 epidemic evolution: are compartmental models really predictive? arXiv:2004.08207.
[19] T. Alberti and D. Faranda, On the uncertainty of real-time predictions of epidemic growths: A COVID-19 case study for China and Italy, Commun. Nonlinear Sci. Numer. Simul. **90**, 105372 (2020).
[20] P.-F. Verhulst, Notice sur la loi que la population suit dans son accroissement, Corresp. Math. Phys. **10**, 113 (1838).
[21] B. Prasse and P. Van Mieghem, Time-dependent solution of the NIMFA equations around the epidemic threshold, J. Math. Biol. **81**, 1299 (2020).
[22] B. Prasse, K. Devriendt, and P. Van Mieghem, Clustering for epidemics on networks: A geometric approach, Chaos **31**, 063115 (2021).
[23] P. Jagers, *Branching Processes with Biological Applications* (John Wiley, London, 1975).
[24] T. Britton and D. Lindenstrand, Epidemic modelling: Aspects where stochasticity matters, Math. Biosci. **222**, 109 (2009).
[25] P. Van Mieghem, *Performance Analysis of Complex Networks and Systems* (Cambridge University Press, Cambridge, 2014).
[26] M. Gatto, On Volterra and D'Ancona's footsteps: The temporal and spatial complexity of ecological interactions and networks, Ital. J. Zool. **76**, 3 (2009).
[27] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).
[28] P. Van Mieghem, Universality of the SIS prevalence in networks, arXiv:1612.01386.
[29] Q. Liu and P. Van Mieghem, Evaluation of an analytic, approximate formula for the time-varying SIS prevalence in different networks, Physica A **471**, 325 (2017).
[30] K. Biswas, A. Khaleque, and P. Sen, Covid-19 spread: Reproduction of data and prediction using a SIR model on Euclidean network, arXiv:2003.07063.
[31] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu, A time-dependent SIR model for COVID-19 with undetectable infected persons, IEEE Transactions on Network Science and Engineering **7**, 3279 (2020).
[32] B. F. Maier and D. Brockmann, Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China, Science **368**, 742 (2020).
[33] K. P. S. S. Hembram and J. Kumar, Epidemiological study of novel coronavirus (COVID-19), International Journal of Community Medicine and Public Health **8**, 1369 (2021).
[34] B. Prasse, M. A. Achterberg, L. Ma, and P. Van Mieghem, Network-inference-based prediction of the COVID-19 epidemic outbreak in the Chinese province Hubei, Appl. Netw. Sci. **5**, 35 (2020).
[35] E. van den Heuvel, M. Regis, and Z. Zhan, Statistical approach for making predictions of confirmed infection and deaths on corona virus, Technical report (TU Eindhoven), https://assets.tue.nl/fileadmin/content/pers/2020/03%20March/TUe%20-%20Technical_Report_Prediction_Corona_Virus.pdf.
[36] G. Vattay, Predicting the ultimate outcome of the COVID-19 outbreak in Italy, arXiv:2003.07912.
[37] K. Wu, D. Darcet, Q. Wang, and D. Sornette, Generalized logistic growth modeling of the COVID-19 outbreak: Comparing the dynamics in the 29 provinces in China and in the rest of the world, Nonlin. Dyn. **101**, 1561 (2020).
[38] E. Pelinovsky, A. Kurkin, O. Kurkina, M. Kokoulina, and A. Epifanova, Logistic equation and COVID-19, Chaos, Solitons Fractals **140**, 110241 (2020).
[39] M. Löytönen and P. Maasilta, Multi-drug resistant tuberculosis in Finland—A forecast, Social Sci. Med. **46**, 695 (1998).
[40] Y. Yang and E. Williams, Logistic model-based forecast of sales and generation of obsolete computers in the U.S., Tech. Forecast. Social Change **76**, 1105 (2009).
[41] A. C. Harvey, Time series forecasting based on the logistic curve, J. Operat. Res. Soc. **35**, 641 (1984).
[42] https://www.rivm.nl/coronavirus-covid-19/actueel.
[43] M. A. Achterberg, B. Prasse, L. Ma, S. Trajanovski, M. Kitsak, and P. Van Mieghem, Comparing the accuracy of several network-based COVID-19 prediction algorithms, Int. J. Forecast. (2020).
[44] H. Schultz, The standard error of a forecast from a curve, J. Am. Stat. Assoc. **25**, 139 (1930).
[45] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004).
[46] R. Pearl and L. J. Reed, On the rate of growth of the population of the United States since 1790 and its mathematical representation, Proc. Nat. Acad. Sci. USA **6**, 275 (1920).
[47] G. H. Golub and C. F. Van Loan, *Matrix Computations* (Johns Hopkins University Press, Baltimore, 2013), Vol. 3.
[48] E. Dong, H. Du, and L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis. **20**, 533 (2020).
[49] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1972).