Scaling rules for pandemics: Estimating infected fraction from identified cases for the SARS-CoV-2 Pandemic

Mingyang Ma[1], Mary Zsolway[2], Ayushya Tarafder[2], Gyan Bhanot[1,3]

[1] Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey, USA
[2] School of Arts and Sciences, Rutgers University, New Jersey, USA
[3] Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey, USA

Keywords: Pandemics, Scaling Laws, SIR Model, Simulations, infected fraction of symptomatic and asymptomatic cases among susceptible, interacting individuals, Worldwide Covid-19 analysis.

ABSTRACT

Using a modified form of the SIR model, we show that, under general conditions, all pandemics exhibit certain scaling rules. Using only daily data for symptomatic, confirmed cases, these scaling rules can be used to estimate: (i) r_{eff}, the effective pandemic R-parameter; (ii) f_{tot}, the fraction of *exposed* individuals that were infected (symptomatic and asymptomatic); (iii) $L_{eff,}$ the effective latency, the average number of days an infected individual is able to infect others in the pool of susceptible individuals; and (iv) α, the probability of infection per contact between infected and susceptible individuals. We validate the scaling rules using an example and then apply our method to estimate r_{eff}, f_{tot}, L_{eff} and α for the first phase of the SARS-Cov-2, Covid-19 pandemic for thirty-four countries where there was a well separated first peak in identified infected daily cases after the outbreak of this pandemic in early 2020. Our results are general and can be applied to any pandemic.

INTRODUCTION

A pandemic occurs when a new pathogen enters a naïve population. The recent SARS-Cov-2 pandemic was caused by a Coronavirus, one of a family of large, enveloped, single-stranded RNA viruses that are widespread in animals and usually cause only mild respiratory illnesses in humans [1-5]. In 2003, a new coronavirus emerged, and was named SARS-CoV (Severe Acute Respiratory Syndrome – Corona Virus). This virus caused a life-threatening respiratory disease in humans, with a fatality rate of almost 10% [6,7]. Unfortunately, after an initial burst of interest in development of treatment options, interest in this virus waned. The emergence of the novel coronavirus SARS-CoV-2, identified in December 2019 in Wuhan, China, has since caused a worldwide pandemic [8-13]. SARS-CoV-2 is the seventh known coronavirus to cause pathology in humans [1]. The associated respiratory illness, called COVID-19, ranges in severity from a symptomless infection [8], to common-cold like symptoms, to viral pneumonia, organ failure, neurological complications, and death [9-11]. While the mortality in SARS-CoV-2 infections is lower than in SARS-CoV [9-12], it has more favorable transmission characteristics, a higher reproduction number, a long latency period and an asymptomatic infective phase [13].

The governments of several countries took significant measures to slow the infection rate of Covid-19, such as social distancing, quarantine, identification, tracking and isolation. However, there was no uniform policy, some governments reacted later than others, and some (e.g. Sweden)

decided to keep the country open, leaving counter-measures up to individuals. A large amount of consistent public data is now available on the number of tests performed, the number of confirmed infected cases, and the number of deaths in different contexts, such as locations and health conditions [14]. These provide important sources of information for the development and testing of models to estimate pandemic characteristics, guide public policy and assess the efficacy of interventions [15].

It is well known that in most pandemics, confirmed infected cases often seriously underestimate the actual number of infections [16,17]: not everyone who is infected is symptomatic, and not everyone who dies from the disease has been tested [18]. Even the number of reported deaths may be underestimated because of co-mortalities; i.e. COVID-19 increases susceptibility to other diseases and conditions [19]. Moreover, the virus can be transmitted by asymptomatic individuals, who can comprise a substantial portion of the infected population [20], militating against accurate estimates of total infection rates. In this context, as indicated in [21], analytical models can provide useful information.

Dynamical (mechanistic) models have been used for forecasting and for making projections. For example, projections and forecasting models of various types were used as early as February 2020 to determine a reproductive number for SARS-CoV-2 [13]. More generally, multiple research groups have models to estimate Case Fatality Ratios (CFRs) [22], to forecast and project the need for hospital beds [23] and to project and forecast mortality [24]. Among the many applications of models to COVID-19, four variable Susceptible-Exposed-Infective-Recovered (SEIR) models have been used to project the impact of social distancing on mortality [25], three variable Susceptible-Infective-Recovered (SIR) models have been used to estimate case fatality and recovery ratios early in the pandemic [26], and a time delayed SIR has been used to evaluate the effectiveness of suppression strategies [27]. One of the most ambitious dynamical models, which includes 8 state variables, and 16 parameters, was fruitfully applied to evaluate intervention strategies in Italy, in spite of the fact that parameter identifiability could not be assured [28]. There is also some model based evidence that the transmission of the SARS-Cov-2 virus is regulated by temperature and humidity [29]. In this paper, we model the Covid-19 pandemic using an extension of the SIR model [30], which partitions the population into three compartments: Susceptibles (S), Infectious (I) and Removed R. This and other models to study the global spread of diseases have been used in a variety of contexts (For some recent reviews, see [31-33]). The SIR model based method developed in this paper differentiates itself from earlier studies in that it provides a way to make an a-posteriori estimate of several useful epidemiological parameters for any pandemic, using only data on confirmed, identified cases.

The question we ask in this paper is the following: Using only daily recorded case data of symptomatic individuals, is it possible to estimate the actual fraction of infected individuals from among the pool of susceptible individuals who contributed to the recorded cases in the region from which the data was collected? We will show that this question can be answered in the affirmative, at least within the context of an extension of the standard epidemiological SIR model [30]. The reason this is possible is that in this model there is a connection between the identified daily cases and the actual number of individuals who remain infected in the population on that day. We will show that this connection leads to general scaling rules for the location of the peak (days from start of the pandemic to the peak in daily cases) and the half width at full maximum in identified daily cases. We will further show that these scaling rules allow an estimate of an "effective" pandemic R-parameter r_{eff}, the fraction f_{tot} of exposed individuals who got infected (both symptomatic and asymptomatic), the effective latency L_{eff}, the average number of days an infected individual is able

to infect others and α, the probability of infection per contact between infected and susceptible individuals. These results are general and can, in principle, be applied to any pandemic. After demonstrating the internal consistency of our approach on model data, we apply our method to worldwide daily case data for the first phase of the SARS-Cov-2 (Covid-19) pandemic in 2020 to derive estimates of these parameters for a number of countries where there was a well separated first peak in identified infected daily cases after the outbreak of this pandemic in early 2020

We note that our results for f_{tot} represent only the fraction of infected individuals in the "exposed population" in a given region – i.e., it only applies to the set of susceptible individuals who came into sufficiently close contact with infected individuals for the virus to transmit. This value should not be taken to represent the fraction of infected individuals in the population as a whole, because our analysis does not include those individuals who were sufficiently isolated in some way (e.g., self-quarantined, wore masks etc.), so as to avoid contact with the virus.

METHODS: The Extended SIR Model

We assume that each country is a region where a subset of the population consists of interacting individuals who are equally susceptible to infection and once infected, are responsible for virus transmission. We also assume that there are two types of individuals: those who become symptomatic after infection, and those who do not. Daily counts of infected individuals reflect only those who become symptomatic. We also assume that identified symptomatic individuals are no longer able to infect others because, once identified as infected (possibly after confirmatory testing), they would be isolated, confined, or quarantined. On the other hand, asymptomatic individuals, being unaware of their infected state, would continue to infect others until they become non-infective (cured/recovered). We define the start of the pandemic as the day when the number of recorded daily cases begins to rise exponentially towards a well-defined peak (a more useful practical definition will be provided later), before decreasing to less than half the peak and possibly continuing to decrease further.

Let L_0 be the average number of days an asymptomatic individual is infective and L_1 be the average number of days a symptomatic individual is infective. An asymptomatic individual becomes non-infective (recovered/cured) after an average of L_0 days, while a symptomatic individual would have symptoms before L_0. Consequently, we assume that $L_1 < L_0$. Let $\gamma_1 = \frac{1}{L_1}$ and $\gamma_0 = \frac{1}{L_0}$ be the rates at which these two sets of infected individuals leave the infective pool. Let α be the probability of infection when an infected individual meets a susceptible (non-infected) individual. Under these assumptions, we can write down a simple extension of the SIR model [1] for the pandemic dynamics:

Let $S(t), I(t), R(t)$ to be the number of Susceptible, Infected and Removed individuals at time t, with $S(t) + I(t) + R(t) = N$. Here N is the pool of susceptible individuals who were exposed to the virus. At any given time, the I(t) compartment consists of two parts, $I_0(t)$ and $I_1(t)$ where the first is a fraction 1-ω of individuals who remain asymptomatic until they recover and the second is a fraction ω of individuals who become symptomatic, are identified and are no longer able to infect others (they move to the "Removed" compartment). Consequently, the R compartment consists of two sets of individuals, a set $R_1(t)$ derived from $I_1(t)$, and a set $R_0(t)$ derived from $I_0(t)$. The extension of the SIR model that applies in such a situation is defined by the equations:

$$\frac{dS}{dt} = -(\frac{\alpha}{N})SI \quad (1)$$

$$\frac{dI}{dt} = \frac{dI_1}{dt} + \frac{dI_0}{dt} = \left(\frac{\alpha}{N}\right)SI - (\omega\gamma_1 + (1-\omega)\gamma_0)\,I = \left(\frac{\alpha}{N}\right)SI - \gamma_{eff}I \quad (2)$$

$$\frac{dR}{dt} = \frac{dR_1}{dt} + \frac{dR_0}{dt} = (\omega\gamma_1 + (1-\omega)\gamma_0)\,I = \gamma_{eff}I \quad (3)$$

Here

$$\gamma_{eff} = (\omega\gamma_1 + (1-\omega)\gamma_0). \quad (4a)$$

The quantity α is the probability of infection in a single encounter between an infected and susceptible individual. The reciprocal L_{eff} of γ_{eff} is the average effective latency, the average number of days that an infected individual (symptomatic or not) is infective. Thus,

$$L_{eff} = 1/\gamma_{eff} = \frac{1}{\omega\gamma_1 + (1-\omega)\gamma_0} \quad (4b)$$

Note that L_{eff} can sometimes be estimated from monitoring and testing of individuals. However, in the general case, it is quite difficult to estimate because its value depends on the fraction of asymptomatic infected cases.

The key quantity in our approach is X(t), the rate at which symptomatic individuals are identified. Thus,

$$X(t) = \frac{dR_1}{dt} = \omega\gamma_1 I(t). \quad (5)$$

For any pandemic, X(t) is the *observed* daily cases reported from hospitals and testing sites from symptomatic and/or tested individuals. The key observation that leads to the results in this paper is Eq. 5, which asserts that X(t) is proportional to I(t). This proportionality means that the width and location of the peak in X(t) and I(t) are the same.

If we rescale time to $\tau = \gamma_{eff}t = \frac{t}{L_{eff}}$ Eq. 1-3 can be rewritten in terms of the fractions s = S/N, i = I/N, r= R/N, $r_1 = R_1/N$ and x = X/N as follows:

$$\frac{ds(\tau)}{d\tau} = -r_{eff}\,s(\tau)i(\tau) \quad (6)$$

$$\frac{di(\tau)}{d\tau} = r_{eff}\,s(\tau)i(\tau) - i(\tau) \quad (7)$$

$$\frac{dr(\tau)}{d\tau} = i(\tau) \quad (8)$$

$$x(\tau) = \frac{dr_1(\tau)}{d\tau} = \omega\gamma_1 i(\tau) \quad (9)$$

$$\text{with } r_{eff} = \alpha L_{eff} \quad (10)$$

At the start of the pandemic, i.e., at τ=0, both $i(\tau)$ and $x(\tau)$ are near zero, since a very small fraction of the population is initially infected. It is easy to show that, starting with a small fraction ε of infected cases at $\tau = 0$, $i(\tau)$ and $x(\tau)$ increase exponentially as $e^{(r_{eff}-1)\tau}$ in the interval $0 < \tau \leq$

$\frac{\log(\varepsilon)}{(1-r_{eff})}$ (Appendix A, Eq. A16a,b). Eventually (as we will see in the data and the solution to the model equations), both quantities reach a peak when the fraction of susceptible individuals decreases sufficiently to slow the growth of the pandemic. Finally, $i(\tau)$ and $x(\tau)$ diminish to a value near zero when the likelihood of further infections becomes negligible. It is easy to show that for a pandemic to take place at all, r_{eff} must exceed unity. In other words, for $r_{eff} \leq 1$, there is no pandemic and $s(\infty) = 1$ (Appendix A (Eq. A8). Thus, r_{eff} is identified as the so called "Pandemic R-parameter", the single parameter that controls the pandemic dynamics in this model.

To facilitate further discussion, we define the following quantities:

(i) $L_I = L_X =$ locations of the peaks in I(t) and X(t) (11a)
(ii) $W_I = W_X =$ widths of the peaks in I(t) and X(t) (11b)
(iii) $H_I =$ maximum value of I(t) (11c)
(iv) $H_X =$ maximum value of $X(t) = \omega\gamma_1 H_I$ (11d)
(v) $f_{tot} = (1 - S(\infty)/N)$ (11e)

f_{tot} is the total fraction of exposed individuals who become infected, including both symptomatic and asymptomatic cases. This quantity is generally difficult to estimate. However, as noted, we can exploit the fact (Eq. 5 and Eq. 9) that there is a connection between the time dependence of identified symptomatic cases X(t) and $x(\tau)$, and the time dependence of the total number of cases I(t) and $i(\tau)$, which includes both symptomatic and asymptomatic cases. Specifically, Eq. 5 says that the location and widths of the peaks in X(t) and I(t) are the same, and Eq. 9 says that location and width of the peaks in $x(\tau)$ and $i(\tau)$ are also the same. *The key idea of this paper is that this fact allows one to relate properties of X(t) and I(t) (or $x(\tau)$ and $i(\tau)$) to estimate r_{eff}, f_{tot}, L_{eff} and* α *using only data for X(t).*

RESULTS:

I. Universal Scaling Rules for Pandemics

Since time t in physical units (seconds, hours, days) is related to dimensionless time τ by $\tau = \frac{t}{L_{eff}}$, we can relate properties of I(t) and X(t) in Eq. 11 to properties of $i(\tau)$, and $x(\tau)$. Thus:

(i) $L_X/L_{eff} = L_I/L_{eff}$ = locations of the peaks in $x(\tau)$ and $i(\tau)$. (12a)
(ii) $W_X/L_{eff} = W_I/L_{eff}$ = widths of the peaks in $x(\tau)$ and $i(\tau)$. (12b)
(iii) H_I/N = maximum value of $i(\tau)$ (12c)
(iv) H_X/N = maximum value of $x(\tau) = \omega\gamma_1 H_I/N$ (12d)
(v) $f_{tot} = (1 - s(\infty))$ (12e)

In the limit of large N, it is easy to find an exact formula for H_I/N (see Eq. A10 and the discussion preceding it in Appendix A):

$$\frac{H_I}{N} = 1 - \frac{[1+\log(r_{eff})]}{r_{eff}}, \; r_{eff} > 1 \qquad (13)$$

However, although this is interesting, it is not very useful, because relating this quantity to the measurable quantity H_X/N requires the values of ω, γ_1. On the other hand, the relationships in Eq. 12a,b and the fact that all the quantities in Eqs. 11 and 12 are controlled by a single parameter r_{eff} lead to universal scaling rules that can be exploited to estimate f_{tot}, r_{eff}, L_{eff} and α using only data for X(t). The simplest way to do this is to note that the ratio L_X/W_X is independent of L_{eff} and can be estimated from the measured daily cases X(t). Figures 1a,b show the dependence f_{tot}, r_{eff} on L_X/W_X (data in Supplementary Table 1). These results were obtained by numerically solving Eq. 6-8 using the stiff ODE solver ode15s in Matlab for r_{eff} in the range 0.5-6.5. Once r_{eff} (or f_{tot}) is known, L_{eff} can be estimated from the functional dependence of L_X/L_{eff} and W_X/L_{eff} on these quantities (Figure 1 c-f, data in Supplementary Table 1,).

Figure 1a-f and the data in Supplementary Table 1 are the main results of this paper. Within the limits of the SIR model, these results are universal and apply to any pandemic. For any pandemic, once L_X and W_X are estimated from data for X(t) in a given region, these data can be used to estimate pandemic parameters.

II. Inferring f_{tot}, r_{eff}, L_{eff} and α using only data for X(t)

Appendix B shows an example of the use of the data in Figure 1 and Supplementary Table 1 to estimate f_{tot}, r_{eff}, and L_{eff} and α from L_X and W_X for one specific set of test parameters used to generate numerical solution of Eq. 1-5. For use in general, we used a minimization procedure that generates initial estimates of f_{tot} and r_{eff} using the experimental value $y_e = L_X/W_X$ from the time dependence of X(t). The data in Figure 1b (and Supplementary Table 1) was then used to make an initial estimate r^0_{eff} for r_{eff} which was iteratively improved by choosing nearby values of r_{eff} to solve Eq. 6-9, compute $y(r_{eff}) = L_X(r_{eff})/W_X(r_{eff})$ and minimize $(y_e - y(r_{eff}))^2$ as a function of r_{eff}.

III. Application to data for the SARS-CoV-2/Covid-19 pandemic:

Worldwide data for confirmed Covid-19 cases and deaths from January 3, 2020 was downloaded from the World Health Organization (WHO) website: https://covid19.who.int/data (Supplementary Table 2). This data estimates the function X(t) in our analysis. Before performing any analysis, the data for daily cases was averaged over eleven days to reduce noise. Averaging over seven or three days did not change the results given below.

Our model assumes that there was a single circulating strain of the virus that infected a homogeneous set of individuals in a given region who were equally susceptible to infection (uniform immune response). The model also assumes that exposed individuals observed the same rules regarding the use of masks/isolation/quarantine, there was no significant variation in population density among them, little variation in their movements, and equal vaccination status. Symptomatic cases were equally likely to be identified across the region, and consistently obeyed (or disobeyed) rules regarding quarantine, testing, etc. Since we cannot correct for these effects in this paper, we present the results only as proof of concept, and apply our method only to the first wave of the Covid-19 pandemic. For this first wave of the pandemic, the world population was naïve to the virus (no immunity) so that everyone was susceptible. Moreover, at least some of the other assumptions of the model did apply in some countries, such as homogeneity of response, lack of vaccines resulting in no innate immunity, standard medical protocols (and in some cases testing for viral RNA) used in identifying cases, and a single circulating strain of the virus.

We also apply the model only to countries where the data showed a clear exponential rise from a few cases followed by a clear peak in daily cases with a measurable half width at full

maximum for the first phase of the pandemic, which took place in most countries between January 1, 2020, and August 31, 2020. We also require that this initial peak not overlap with subsequent peaks. Thirty-four countries satisfied these conditions. For these, the values of L_X and W_X were determined for the first peak in daily cases and f_{tot}, r_{eff}, L_{eff} and α were estimated as described below and in Appendix B.

The results from solving Eqs. 6-9 generates values for $x(\tau)$ (Eq. 9) as a function of τ for each value of r_{eff}. The peak in $x(\tau)$ as a function of τ from this solution was mapped to the actual data by scaling it to match the observed peak in the measured X(t) for each country. The scaling from $\tau \rightarrow$ t was performed by making a linear map of the location of the positions of the half width of the maximum in $x(\tau)$ (as a function of τ) to the positions of these locations in real time t (in days) in the measured function X(t) for each country. Note that the scaling required for this mapping provides an independent estimate of L_{eff} which, in all cases reported here, agreed with the estimate of L_{eff} directly from the data in Supplementary Table 1 (Figs 1a-f). Errors in the parameters were determined by varying L_X and W_X by +/- 1 and recomputing them as described above. To identify the "start" date of the pandemic, which affects the estimate of L_X, we used the procedure described in Appendix B and which was also used in generating the data in Supplementary Table 1: The start date was chosen as the day when the measured daily cases numbered approximately 1% of the peak. We also checked that in the days following this start date, the daily cases fit well to an exponential function, as would be expected at the start of a pandemic (Appendix A).

The results for r_{eff}, f_{tot}, L_{eff} and α for six countries which had r_{eff} varying from 1.23 to 6.04 are shown in Figure 2 a-f. The results for the thirty-four countries where we could apply our methods are given in Table 1. Supplementary Figures 1 show plot of the data for X(t) and the fits. Also shown in the plots are the location of the start day (caseload = 1% of peak), the location of the peak and of the half width at full maximum as well as the inferred values of L_X, W_X and H_X for each country. Some notable exclusions in the list of countries are the United States, the Russian Federation, Canada, India, and Pakistan. The reason for this is that these countries (and others) either had very broad first peaks or had multiple subpeaks within the first peak, making estimates of L_X and W_X problematic. This is presumably because they cannot be considered homogeneous for a variety of reasons, the most important likely being non-uniform response from authorities regarding the use of masks and variable rules across the country regarding movement of people, quarantine etc. In cases such as the United States and Canada, where the response from the authorities was state or province specific, it should be possible to do a state-by-state or province-by-province analysis. We plan to analyze data for the United States and Canada (and possibly other countries) where such compartmentalization is possible in a subsequent paper.

DISCUSSION:

In this paper we have developed a method, applicable to any pandemic, to identify the fraction of infected individuals from among the pool of interacting susceptible individuals in a given region, using a simple extension of the epidemiological SIR model [30]. We show that in this model, there is a universal scaling function that relates the ratio of the location L_X, and the width W_X of the peak in daily identified cases X(t) to the effective Pandemic R-parameter r_{eff} and the fraction f_{tot} of infected exposed individuals (including both symptomatic and asymptomatic infected individuals) (see Figure 1 and Supplementary Table 1). This in turn allows an estimate of the effective latency L_{eff} (average number of days an infected individual is able to infect others) and the infection probability α of transmission from an infected individual to a susceptible individual

in a single encounter (see Appendix B for details). Within the limits of the SIR model, our results are general and apply to any pandemic. We apply our method to worldwide country specific data to find r_{eff}, f_{tot}, L_{eff} and α for the first phase (first peak in daily cases) for the SARS-COV-2 pandemic for thirty-four countries which had a clear, well separated peak in daily cases (Table 1, Figure 2, Supplementary Figure 2).

It is important to note that our result for f_{tot} represents only the fraction of infected individuals in the "exposed population" in a given region – i.e., it only applies to the set of susceptible individuals who came into sufficiently close contact with infected individuals for the virus to transmit. This value should not be taken to represent the fraction of infected individuals in the population as a whole, because our analysis does not include those individuals who were sufficiently isolated in some way (e.g., self-quarantined, wore masks etc.), so as to avoid any contact with the virus.

With this caveat in mind, we note that our results suggest that in the SARS-COV-2 pandemic, the fraction of infected individuals who were exposed to the virus was very high in most countries that met our analysis criteria, suggesting that in its early stages, when countries did not impose quarantines and the use of masks was limited, this virus was highly effective in transmission. In some of the developed countries, our results suggest that almost all exposed individuals were infected in the first phase of the pandemic (Table 1). The only countries where f_{tot} was less than 0.5 were those with a low population density (Australia), low mobility rates of citizens (Afghanistan) or where the use of masks was common, even in the absence of a pandemic (Japan).

Several countries, notably the United States, Canada, The Russian Federation, India, and Pakistan did not meet our criterion of a clear, well separated first peak in daily cases in 2020. This is most likely due to the fact that they cannot be thought of as homogeneous in the sense of response from local authorities regarding the use of masks, quarantine etc. In the United States for example, the response was state specific. Wherever reliable data is available, we plan to apply our model to these countries by stratifying them into appropriate compartments with uniform rules of containment of the virus from local authorities.

Our method can also be applied to subsequent recurrences of the SARS-COV-2 virus (second, third, fourth peaks in daily cases), as the virus evolved into less virulent and more infective strains. Comparing changes in the inferred parameters across countries would provide a country specific overall estimate of preventive measures, such as the effectiveness/efficacy of vaccination, changes in behavior (mask use, testing/quarantine, work-from-home, social distancing, travel restrictions) etc. Our method can, in principle, also be applied to other viral pandemics, such as the SARS-COV pandemic of 2003, and Influenza pandemics of the past, such as the H1N1 Spanish Flu pandemic of 1918-19 which recurred in 1950 and 1977, the H2N2 Asian Flu pandemic of 1957, the H3N2 Hongkong pandemic of 1968 and the more deadly H5N1 East Asian pandemic of 1997.

Data and Software Availability: The data and Matlab codes used in this paper are available on request to gyanbhanot@gmail.com. The data used to fit the model to actual pandemic data is in Supplementary Table 1. The World Health Organization country specific data we used for the SARS-CoV-2 pandemic is in Supplementary Table 2. The data in Supplementary Figures and Tables is available online at:
https://drive.google.com/file/d/1P6emrvTBMC0uo-dD6I21U2_iv1tjK2hS/view?usp=sharing

Author Approval: All authors have seen and approved the manuscript.

Competing Interests: The authors declare there are no competing interests.

Acknowledgements: GB was partly supported by grants from DoD (KC180159) and NIH (P01CA250957). GB thanks Professor Charles DeLisi from Boston University for many discussions and collaboration on earlier (unpublished) work on the SARS-CoV-2 pandemic using the SIR model.

Figure and Table Captions:

Figure 1: (**a,b):** Universal scaling curves in the SIR model for f_{fot} and r_{eff} as functions of the ratio L_X/W_X, where L_X is the number of days from the start of the pandemic to the location of the peak in daily observed cases X(t) (Eq. 5) and W_X is the width of that peak. Note that these functions are independent of L_{eff} and apply to any pandemic. These can be used to find f_{tot} and r_{eff} using the ratio L_X/W_X from data for X(t). (**c-f):** Universal scaling curves in the SIR model for L_X/L_{eff} and W_X/L_{eff} as functions of f_{fot} and r_{eff}. L_X and W_X are the location and width of the peak in daily observed cases X(t) (Eq. 5). These data can be used to estimate L_{eff} once r_{eff} and f_{tot} are estimated using Figure 1 (a,b) (data in Supplementary Table 1). Note that these functions are universal and apply to any pandemic.

Figure 2 a-f: Fits of our model to data for X(t) from the World Health Organization website https://covid19.who.int/data for six of the thirty-four countries for which analysis was possible. These fits were made by using the data in Supplementary Table 1 to find r_{eff} from the ratio L_X/W_X measured from the data followed by rescaling the time axis by mapping the locations of the half width at full maximum for $x(\tau)$ as a function of scaled time τ to the location of the half width at full maximum in X(t) from the data for real time t (details in the text). The red dots are the data averaged over eleven days and the blue curve is the fit obtained. The locations of the "start" of the pandemic (daily cases = 1% of peak) and of the maximum are shown as a yellow circle and a blue mark on the time axis respectively. The green dots are the location of the half maximum. The values of the fitted parameters are shown in the text above each plot.

Table 1: Results for r_{eff}, f_{tot}, L_{eff} and α (Columns O, Q, S, U) from applying our methods to analyse WHO data (https://covid19.who.int/data) for the first peak in X(t) (daily identified cases) for thirty-four countries, which had a clear, well separated peak in X(t) starting January 3, 2020.

Supplementary Figure and Table Captions:

Supplementary Figure 1: Fits of our model to data for X(t) from the World Health Organization website https://covid19.who.int/data for thirty-four countries for which analysis was possible. The red dots are the data averaged over eleven days and the blue curve is the fit obtained as described in the text. The locations of the "start" of the pandemic (daily cases = 1% of peak) and of the maximum are shown as a yellow circle and a blue mark on the time axis respectively. The green dots are the location of the half maximum. The values of the fitted parameters are shown in the text above each plot.

Supplementary Table 1: Results obtained by numerically solving Eq. 6-8 using the stiff ODE solver ode15s in Matlab for r_{eff} in the range 0.5-6.5. These data were used to derive all the results in the paper.

Supplementary Table 2: World Health Organization data for the SARS-CoV-2 pandemic from https://covid19.who.int/data that was used in our analysis.

References:

1. Thiel V, *Coronaviruses: Molecular and Cellular Biology*, Caister Academic Press 2007, ISBN:978-1-904455-16-5.

2. Rabadan R, *Understanding Coronavirus*, University Printing House, Cambridge, UK (2020), ISBN: 978-1-108-920254.

3. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF, *Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses*, Trends Microbiol. 2016 Jun; 24(6):490-502. doi: 10.1016/j.tim.2016.03.003. Epub 2016 Mar 21.

4. Lai MM, Cavanagh D, *The molecular biology of coronaviruses*, Adv Virus Res. 1997;48:1-100.

5. Masters PS, *The molecular biology of coronaviruses*, Adv Virus Res. 2006; 66:193-292; Artika IM, Dewantari AK, Wiyatnoc A, *Molecular biology of coronaviruses: current knowledge.* Heliyon. 2020 Aug; 6(8): e04743.

6. Cherry JD, Krogstad P, *SARS: the first pandemic of the 21st century*, Pediatric Research 2004 Jul; 56(1):1-5. Epub 2004 May 19.

7. Peiris J, Guan Y and Yuen K, *Severe acute respiratory syndrome*. Nat Med 10, S88–S97 (2004). https://doi.org/10.1038/nm1143.

8. J. A. Al-Tawfiq, *Asymptomatic coronavirus infection: Mers-cov and sars-cov-2 (covid-19)*, Travel Med Infect Dis. 2020 May-June; 35: 101608.

9. Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP, *The trinity of COVID-19: immunity, inflammation and intervention*, [published online ahead of print, 2020 Apr 28]. Nat Rev Immunol 20, 363–374 (2020). https://doi.org/10.1038/s41577-020-0311-8

10. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, Liu S, Zhao P, Liu H, Zhu L, *et. al*., *Pathological findings of covid-19 associated with acute respiratory distress syndrome*, The Lancet Respiratory Medicine, 2020 Apr;8(4):420-422.

11. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X, Guan L, Wei Y, Li H, Wu X, Xu J, Tu S, Zhang Y, Chen H, Cao B, *Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study*. Lancet. 2020 Mar 28;395(10229):1054-1062. doi: 10.1016/S0140-6736(20)30566-3.

12. Sørensen MD, Sørensen B, Gonzalez-Dosal R, *et. al., Severe acute respiratory syndrome (SARS): development of diagnostics and antivirals*. Ann N Y Acad Sci. 2006;1067(1):500-505. doi:10.1196/annals.1354.072

13. Inglesby TV, *Public Health Measures and the Reproduction Number of SARS-CoV-2* [published online ahead of print, 2020 May 1], JAMA. 2020 Jun 2;323(21):2186-2187. doi: 10.1001/jama.2020.7878.; Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J, *The reproductive number of COVID-19 is higher compared to SARS coronavirus*, Journal of Travel Medicine, 2020 Mar 13;27(2):taaa021. doi: 10.1093/jtm/taaa021.

14. For example, see: https://Ourworldindata.org/coronavirus from The *European Center for Disease Control and Prevention*.

15. Flaxman, S, Mishra, S, Gandy, A et al. *Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe*. Nature 584, 257–261 (2020). https://doi.org/10.1038/s41586-020-2405-7; Davies NG, Kucharski AJ,Eggo RM, Gimma A, Edmunds WJ, *Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study*, The Lancet, July 01, 2020, Volume 5, Issue 7, e375-e385; Ferguson et al., Imperial College Covid-19 Response Team, Report 9(https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/).

16. Lu FS, Nguyen AT, Link NB, Davis JT, Chinazzi M, Xiong X , Vespignani A, Lipsitch M, Santillana M, *Estimating the Cumulative Incidence of COVID-19 in the United States Using Four Complementary Approaches*, medRxiv 2020.04.18.20070821; doi: https://doi.org/10.1101/2020.04.18.20070821.

17. COVID-19 Antibody Seroprevalence in Santa Clara County, California
Bendavid E, Mulaney B, Sood N, Shah S, Ling E, Bromley-Dulfano R, Lai C, Weissberg Z, Saavedra-Walker R, Tedrow J, Tversky D, Bogan A, Kupiec T, Eichner, Gupta R, Ioannidis J, Bhattacharya J, *COVID-19 Antibody Seroprevalence in Santa Clara County, California*,
Sood N, Simon P, Ebner P, et al. *Seroprevalence of SARS-CoV-2–Specific Antibodies Among Adults in Los Angeles County, California, on April 10-11*, JAMA. 2020;323(23):2425–2427. doi:10.1001/jama.2020.8279

18. https://www.washingtonpost.com/investigations/coronavirus-death-toll-americans-are-almost-certainly-dying-of-covid-19-but-being-left-out-of-the-official-count/2020/04/05/71d67982-747e-11ea-87da-77a8136c1a6d_story.html.

19. https://www.nytimes.com/interactive/2020/06/01/us/coronavirus-deaths-new-york-new-jersey.html

20. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J, *Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)*, Science, 01 May 2020, Vol. 368, Issue 6490, pp. 489-493.

21. https://covid-19.bsvgateway.org

22. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et. al. *Estimates of the severity of coronavirus disease 2019: a model-based analysis*, The Lancet, June 01, 2020, Volume 20, Issue 6, p669-677.

23. https://penn-chime.phl.io/

24. https://www.kff.org/policy-watch/covid-19-models/

25. Branas CC, Rundle A, Pei S, Yang W, Carr BG, Sims S, Zebrowski A, Doorley R, Schluger N, Quinn JW, Shaman J, *Flattening the curve before it flattens us: hospital critical care capacity limits and mortality from novel coronavirus (SARS-CoV2) cases in US counties*, medRxiv 2020.04.01.20049759; doi:https://doi.org/10.1101/2020.04.01.20049759

26. Anastassopoulou C, Russo L, Tsakris A, Siettos C. *Data-based analysis, modelling and forecasting of the COVID-19 outbreak.* PLoS One. 2020, 15(3):e0230405. doi:10.1371/journal.pone.0230405.

27. Casella F. *Can the COVID-19 epidemic be controlled on the basis of daily test reports?* Preprint 2020: https://arxiv.org/abs/2003.06967.

28. Giordano, G., Blanchini, F., Bruno, R. et al. *Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy.* Nat Med 26, 855–860 (2020). https://doi.org/10.1038/s41591-020-0883-7

29. Raines K, Doniach S, Bhanot G, *The transmission of SARS-CoV-2 is likely comodulated by temperature and by relative humidity*, PLoS One. 2021 Jul 29;16(7):e0255212. doi: 10.1371/journal.pone.0255212. PMID: 34324570; PMCID: PMC8321224.

30. Kermack W, McKendric A (1991), *Contributions to the mathematical theory of epidemics – I. Bulletin of Mathematical Biology*. 53 (1–2): 33–55; ibid. *Contributions to the mathematical theory of epidemics – II. The problem of endemicity*", Bulletin of Mathematical Biology. 53 (1–2): 57–87. doi:10.1007/BF02464424. PMID 2059742; ibid. *Contributions to the mathematical theory of epidemics – III. Further studies of the problem of endemicity*. Bulletin of Mathematical Biology. 53 (1–2): 89–118. doi:10.1007/BF02464425. PMID 2059743.

31. Huppert A, Katriel G. *Mathematical modelling and prediction in infectious disease epidemiology*. Clin Microbiol Infect. 2013;19(11):999-1005. doi:10.1111/1469-0691.12308

32. Brauer F, *Mathematical epidemiology: Past, present, and future*. Infect Dis Model (2017) 2(2):113-127. Published 2017 Feb 4. doi:10.1016/j.idm.2017.02.001

33. Colizza V, Barrat A, Barthelemy M, Valleron AJ, Vespignani A, *Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions*. PLoS Med. 2007;4(1):e13. doi:10.1371/journal.pmed.0040013.

Appendices:

Appendix A

The rescaled equations for the pandemic dynamics are:

$$\frac{ds(\tau)}{d\tau} = -r_{eff}\, s(\tau) i(\tau) \quad \text{(A1)}$$

$$\frac{di(\tau)}{d\tau} = r_{eff}\, s(\tau) i(\tau) - i(\tau) \quad \text{(A2)}$$

$$\frac{dr(\tau)}{d\tau} = i(\tau) \quad \text{(A3)}$$

The scaled quantities $s(\tau)$, $i(\tau)$ are related to S(t), I(t) of the SIR model by:

$$s(\tau) = S(t)/N \quad \text{(A4)}$$

$$i(\tau) = I(t)/N \quad \text{(A5)}$$

$$\text{with, } \tau = \gamma_{eff} t = \frac{t}{L_{eff}} \quad \text{(A6)}$$

Dividing (A2) by (A1) gives:

$$\frac{di(\tau)}{ds(\tau)} = \frac{1}{r_{eff} s} - 1 \quad \text{(A6)}$$

Using the large N boundary conditions s(0) = 1, i(0) = 0 generates the exact result:

$$i(\tau) = 1 - s(\tau) + \log(s(\tau))/r_{eff} \quad \text{(A7)}$$

At $t=\infty$, $i(\tau) = 0$. Hence,

$$r_{eff}(s(\infty)) = -\frac{\log(s(\infty))}{1-s(\infty)} \quad \text{(A8)}$$

When $s(\infty) = 1$ (no pandemic), l'Hospital's rule gives $r_{eff}(s(\infty) = 1) = 1$.

It is easy to see for $0 \leq s(\infty) < 1$, $r_{eff} > 1$. Hence, a pandemic requires $r_{eff} > 1$.

From (A2), the maximum in $i(\tau)$ happens when $s(\tau) = 1/r_{eff}$. Hence:

$$\text{Maximum value of } i(\tau) = 1 - \frac{[1+\log(r_{eff})]}{r_{eff}}, \; r_{eff} > 1 \quad \text{(A9)}$$

Note that because of (A5) this quantity is the same as the maximum of I(t)/N which is the quantity H_I/N in Eq. 12c in the main text. Hence,

$$H_I/N = 1 - \frac{[1+\log(r_{eff})]}{r_{eff}},\ r_{eff} > 1 \qquad \text{(A10)}$$

For small τ, $s(\tau)\sim 1$. Hence, we can expand the right-hand side of (A7) in powers of $(1- s(\tau))$. To lowest order,

$$\log(s(\tau)) = \log\left[1 - \left(1 - s(\tau)\right)\right] \cong -\left(1 - s(\tau)\right) \qquad \text{(A11)}$$

Substituted into (A7) gives,

$$i(\tau) = \frac{r_{eff}-1}{r_{eff}}\left(1 - s(\tau)\right) \qquad \text{(A12)}$$

Substituting from (A12) into (A1) gives the Logistic Equation:

$$\frac{ds(\tau)}{d\tau} = -(r_{eff} - 1)s(\tau)(1 - s(\tau)) \qquad \text{(A13)}$$

whose solution, with the boundary condition $s(0) = 1-\varepsilon$ is:

$$s(\tau) = \frac{1}{\left[1+\varepsilon\, e^{(r_{eff}-1)\tau}\right]} \qquad \text{(A14)}$$

Hence, for $\tau \leq \frac{\log(\varepsilon)}{(1-r_{eff})}$,

$$s(\tau) = [\,1 - \varepsilon\, e^{(r_{eff}-1)\tau}], \qquad \text{(A15)}$$

Combining (A12) and A(15) shows that for $\tau \leq \frac{\log(\varepsilon)}{(1-r_{eff})}$,

$$i(\tau) = \frac{r_{eff}-1}{r_{eff}}\,\varepsilon\, e^{(r_{eff}-1)\tau} \qquad \text{(A16a)}$$

Hence, from Eq. 9,

$$x(\tau) = \omega\gamma_1 i(\tau) = \omega\gamma_1\,\frac{r_{eff}-1}{r_{eff}}\,\varepsilon\, e^{(r_{eff}-1)\tau} \qquad \text{(A16b)}$$

Appendix B

We will show an example of the use of the data in Supplementary Table 1 and Figure 1 a-f to find f_{tot}, r_{eff}, L_{eff} and α using only data for X(t), the symptomatic/identified cases per day (Eq. 5). Consider a numerical solution of Eq. 1-5 for parameter values: $N = 10^4$, $L_0 = 10$ days, $L_1 = 5$ days, $L_{eff} = 8$ days, and $r_{eff} = 1.5$ which, from Eq. 4b and Eq. 10, corresponds to $\omega = \frac{\left(\frac{L_0}{L_{eff}}-1\right)}{\left(\frac{L_0}{L_1}-1\right)} = 0.25$ and $\alpha = 0.1875$ respectively. The functional form of S(t), I(t), R(t) and X(t) in Figs. B1a-d was obtained by numerically solving Eq. 1-5 using the stiff ODE solver ode15s in Matlab with boundary condition: S(0) = 1- ε N, I(0) = ε N and R(0) = 0, with $\varepsilon = 0.001$. The result obtained are shown in Figs. B1 a-d. The measured total infected fraction was found to be $f_{tot} = 0.584$. To make contact with real data and to measure L_X, we need an objective definition of "the start day of the pandemic." We define this as the day when the number of recorded daily cases is 1% of the peak in daily cases, which was also the criterion used for the data in Supplementary Table 1.

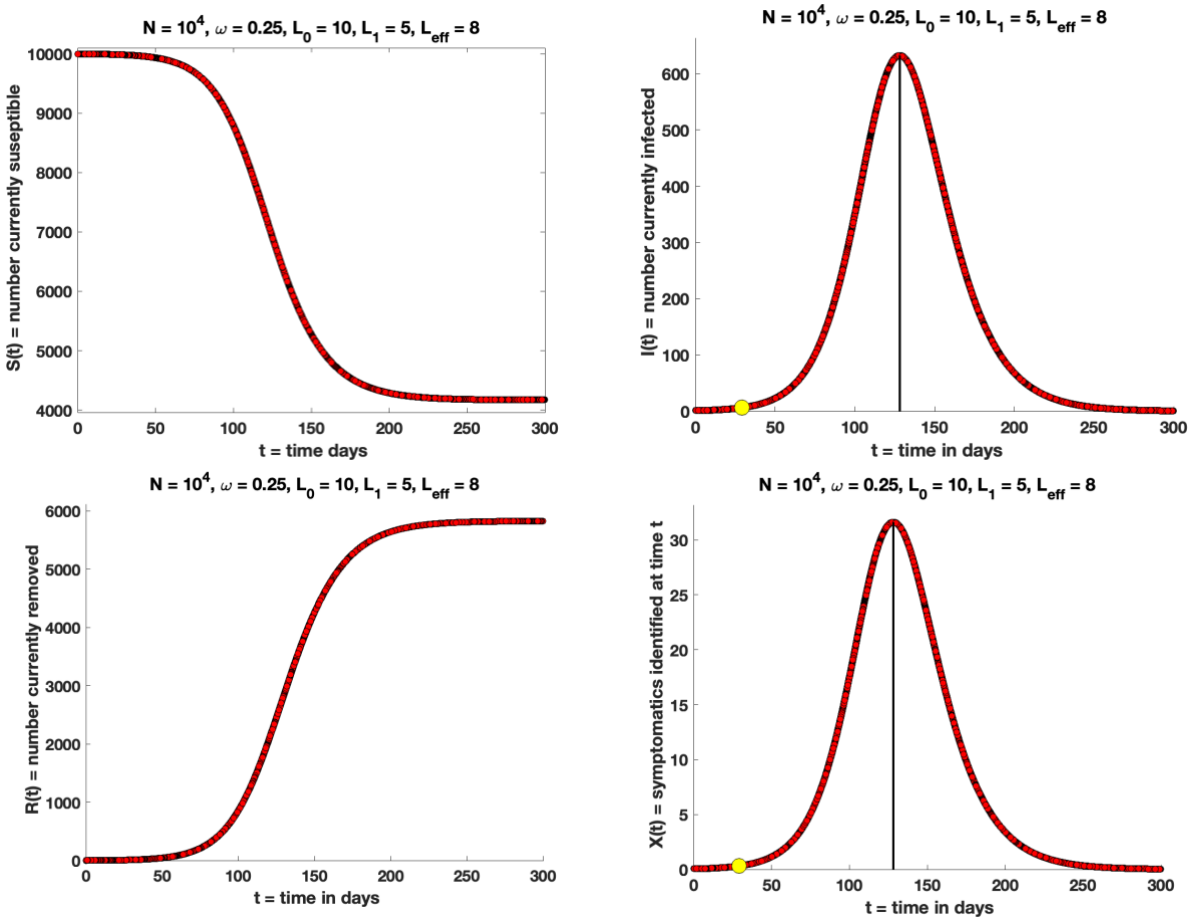


Fig. B1a-d: Solution of Eq 1-5 using the parameters shown. *Note that the yellow dots in Figs. B1b and B1d are our objective definitions of the "start of the pandemic," i.e., the day when the number of cases is 1% of the peak.* The same definition was used in generating the data in Supplementary Table 1. L_X is the number of days from the yellow dot in Figure B1d to the location of the peak. From Fig. B1d, we find, $L_X = 98.9$ days, $W_X = 64.9$ days, and $f_{tot} = 0.58$.

Now imagine that we only know X(t) (Fig B1d but without the parameters in the title), and not the other data in Fig B1a-c (shown in Fig B1e). Using this information alone, we want to estimate f_{tot}, r_{eff}, L_{eff} and α. From the data for X(t), we can find the location of the peak and W_X = 64.87 days, the half width at full maximum. We can also find L_X = 98.85 days from the number of days from when X(t) was 1% of the peak to the location of the peak.Thus L_X/W_X = 1.53. Using this value in Supplementary Table 1 gives the correct values f_{tot} =0.58 and r_{eff} = 1.5. Using the values L_X/L_{eff} = 12.37 and W_X/L_{eff} = 8.10 for r_{eff} = 1.5 in Supplementary Table 1, we get two estimated values 8.01 days and 7.99 days respectively for L_{eff}. Finally, we can estimate $\alpha = \frac{r_{eff}}{L_{eff}} = 0.1875$.

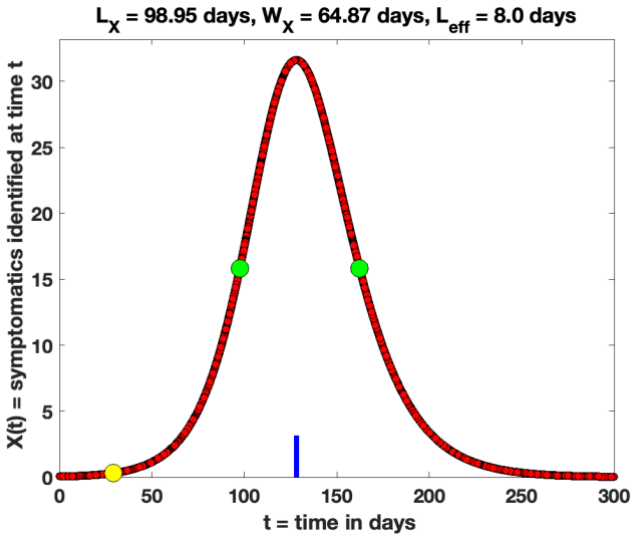
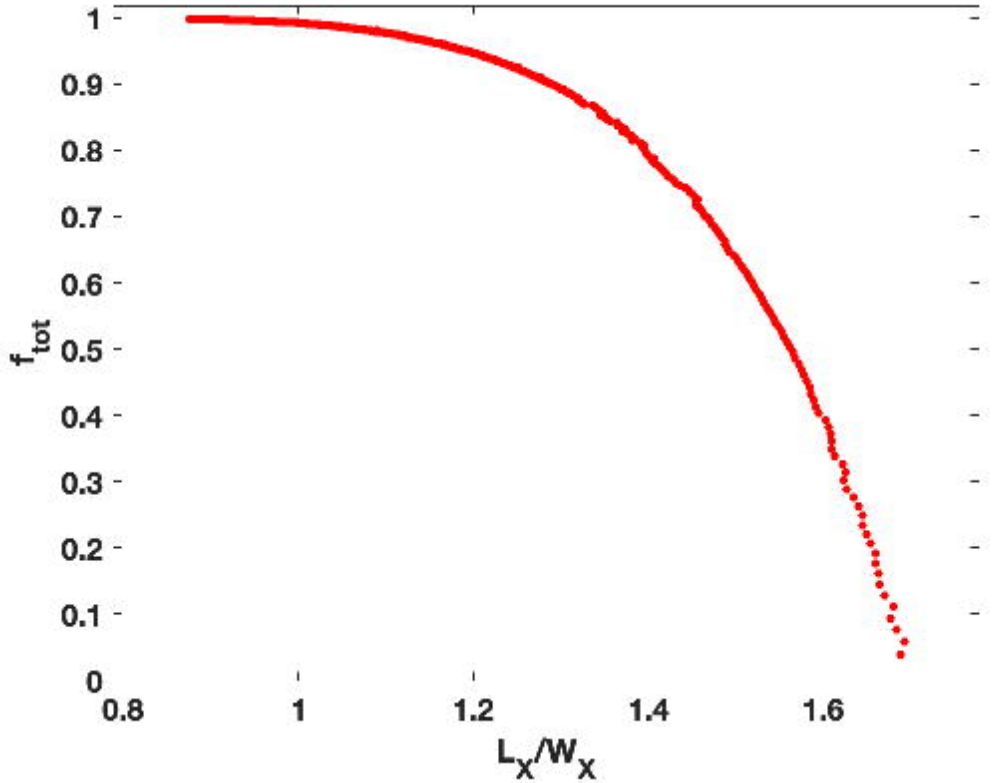
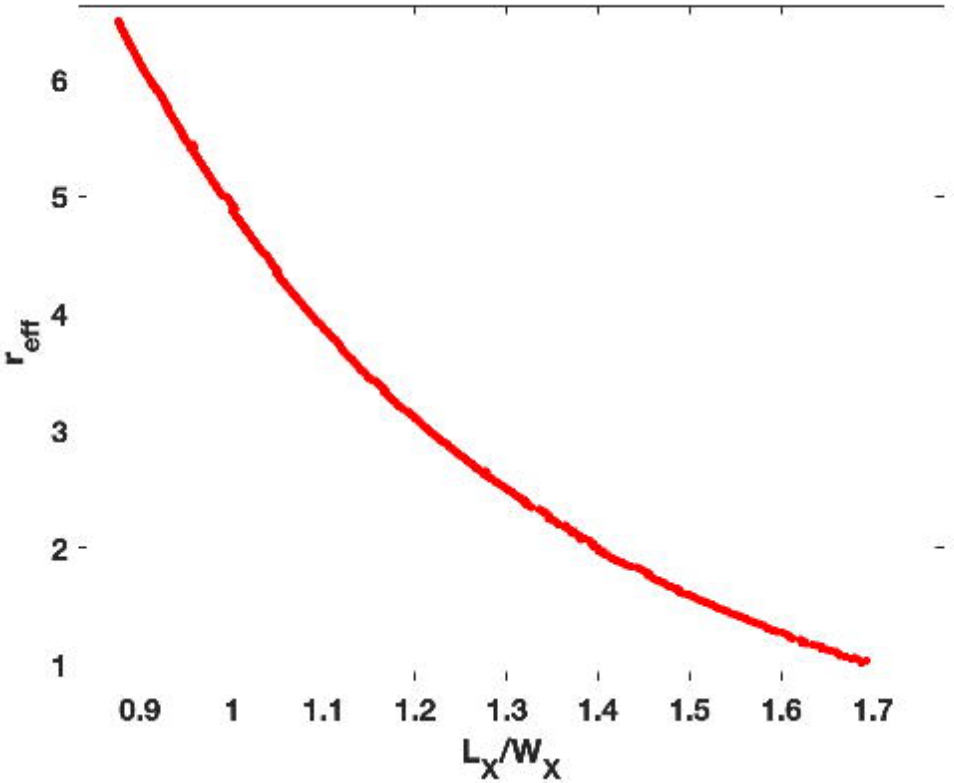
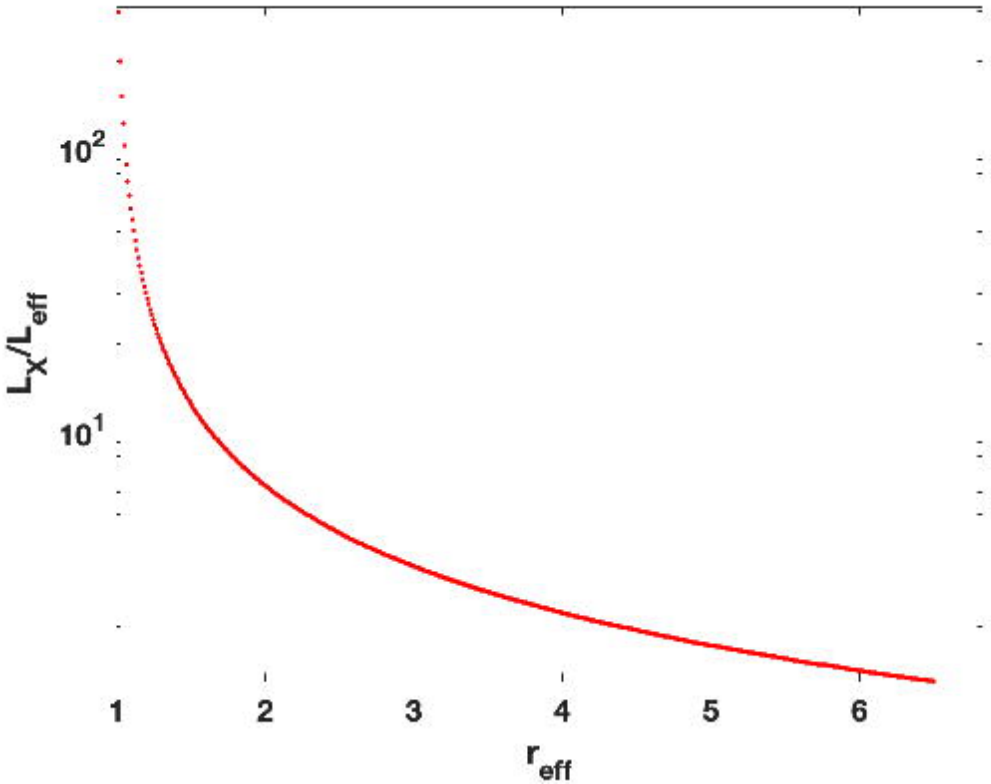
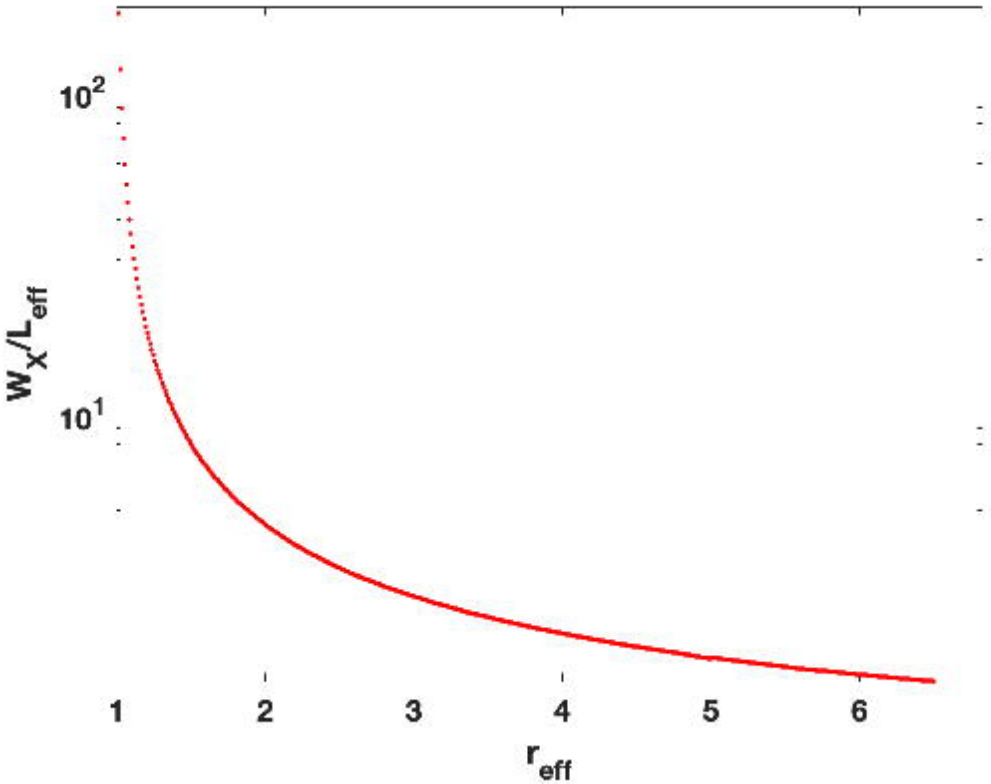


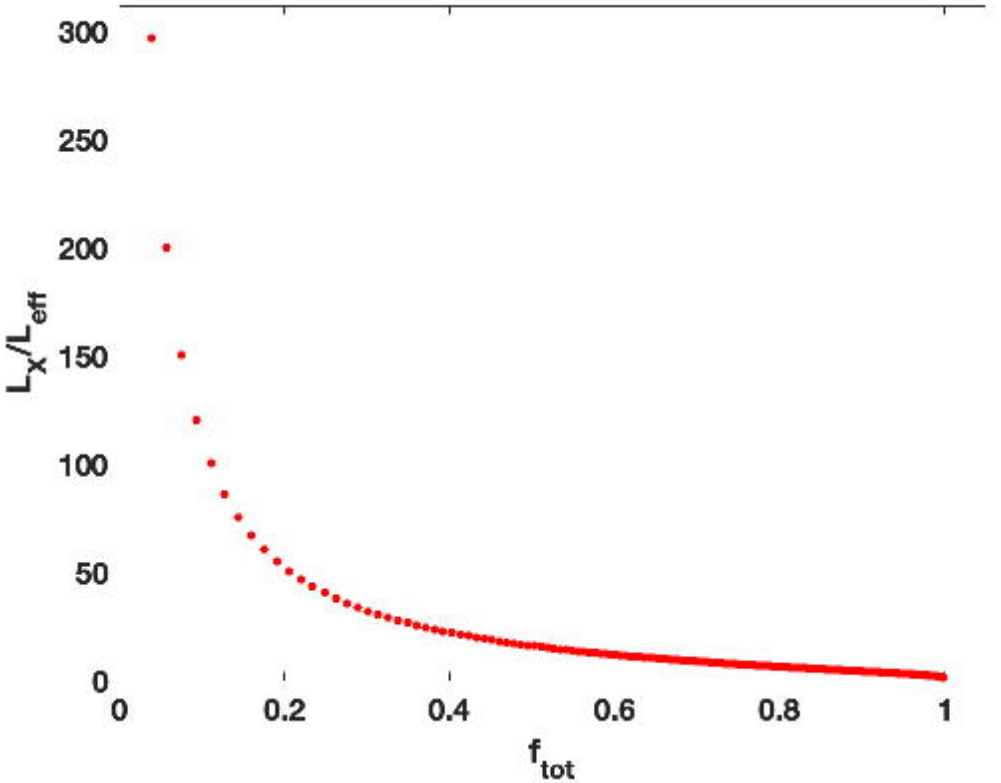
Fig. B1e: Inferring parameters from only X(t). The yellow dots defines the "start" of the pandemic, the day the number of cases is 1% of the peak. The same definition was used in generating the data in Supplementary Table 1. L_X is the number of days from the yellow dot to the location of the maximum (shown as a blue mark) and W_X is the half width of the peak in X(t) (the time between the green dots). This gives L_X/W_X = 1.53, which from the data in Supplementary Table 1 gives f_{tot} =0.58 and r_{eff} = 1.5. Two estimates for L_{eff} can be obtained from the data for L_X/L_{eff} and W_X/L_{eff} for this value of r_{eff} in Supplementary Table 1.

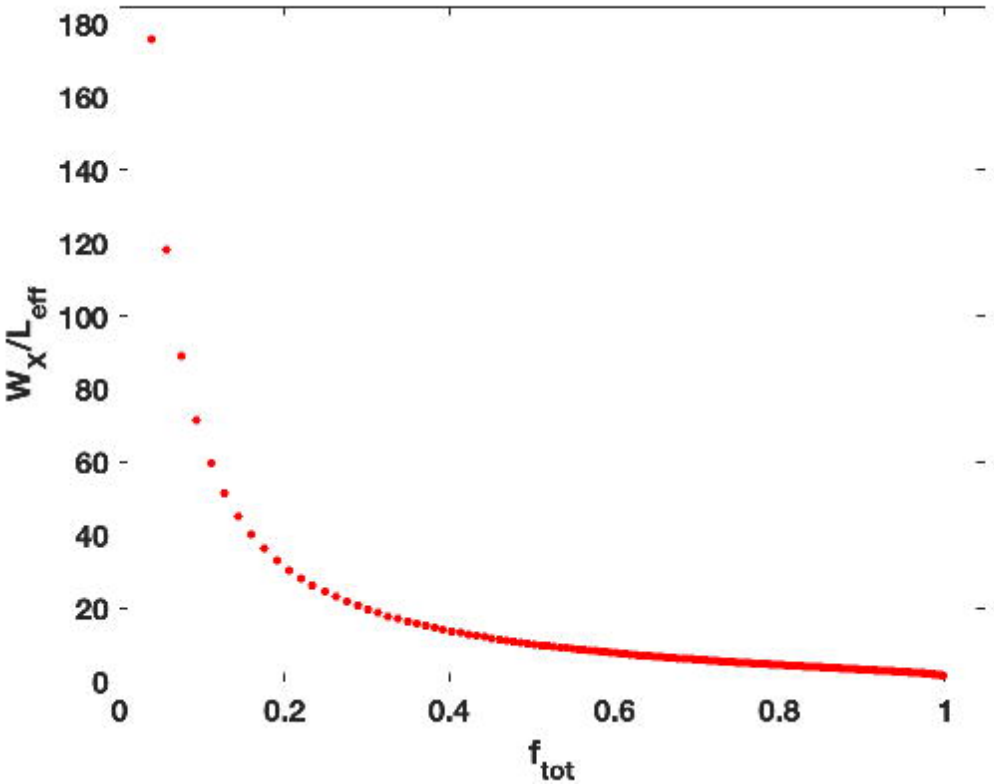

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
f_{tot}
0.8
1
1.2
1.4
1.6
L_X/W_X

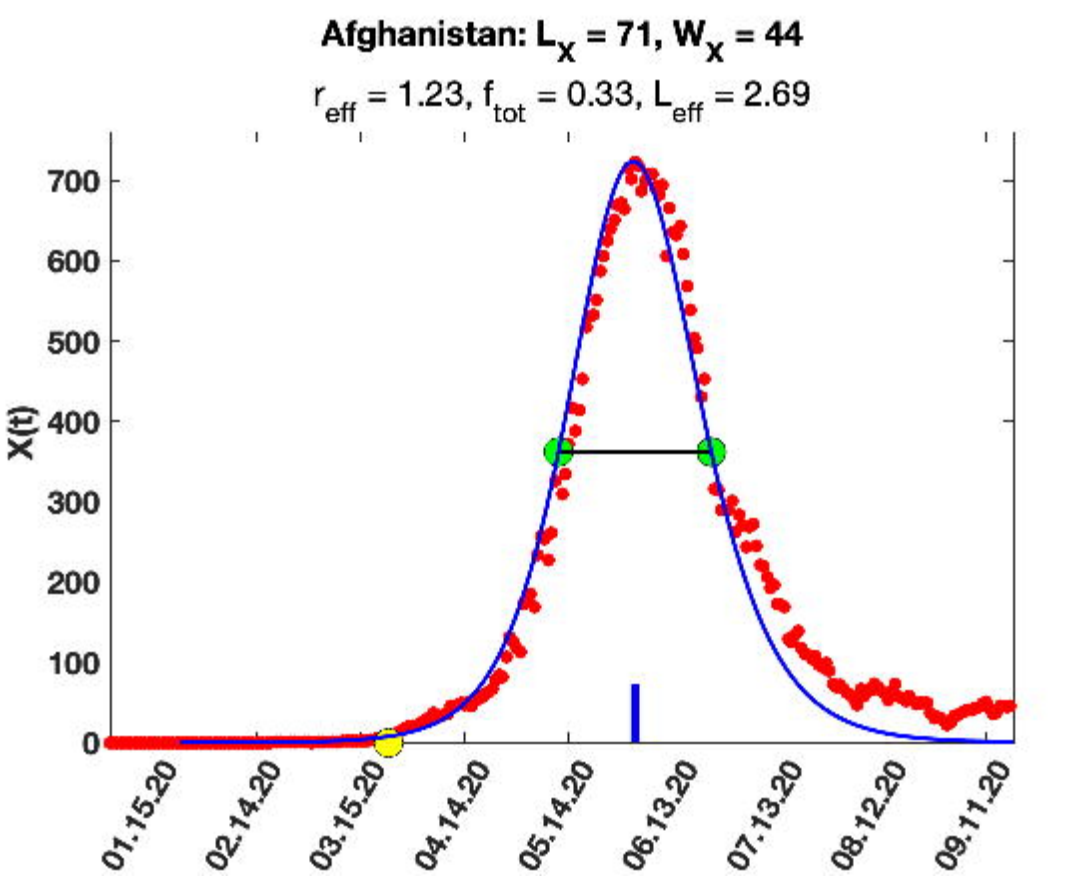

6
5
4
3
2
1
r_{eff}
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
L_X/W_X

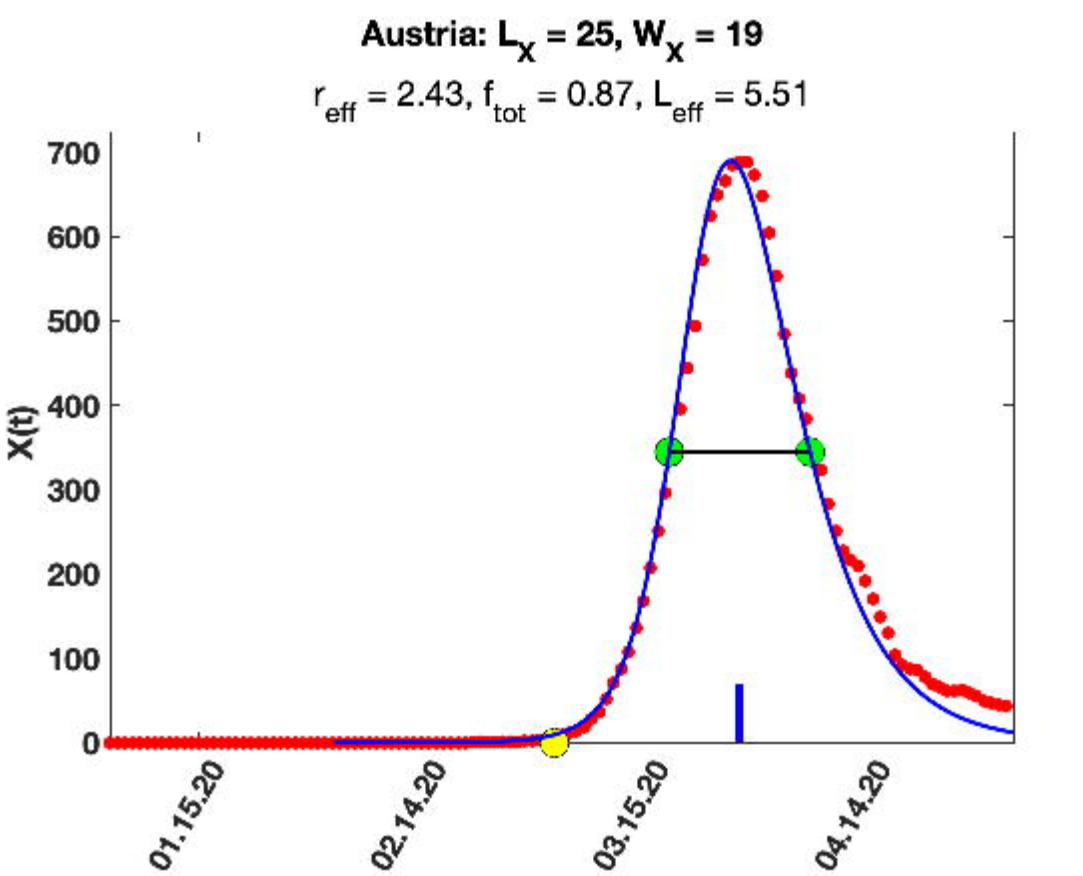

10^2
10^1
L_X/L_{eff}
1
2
3
4
5
6
r_{eff}

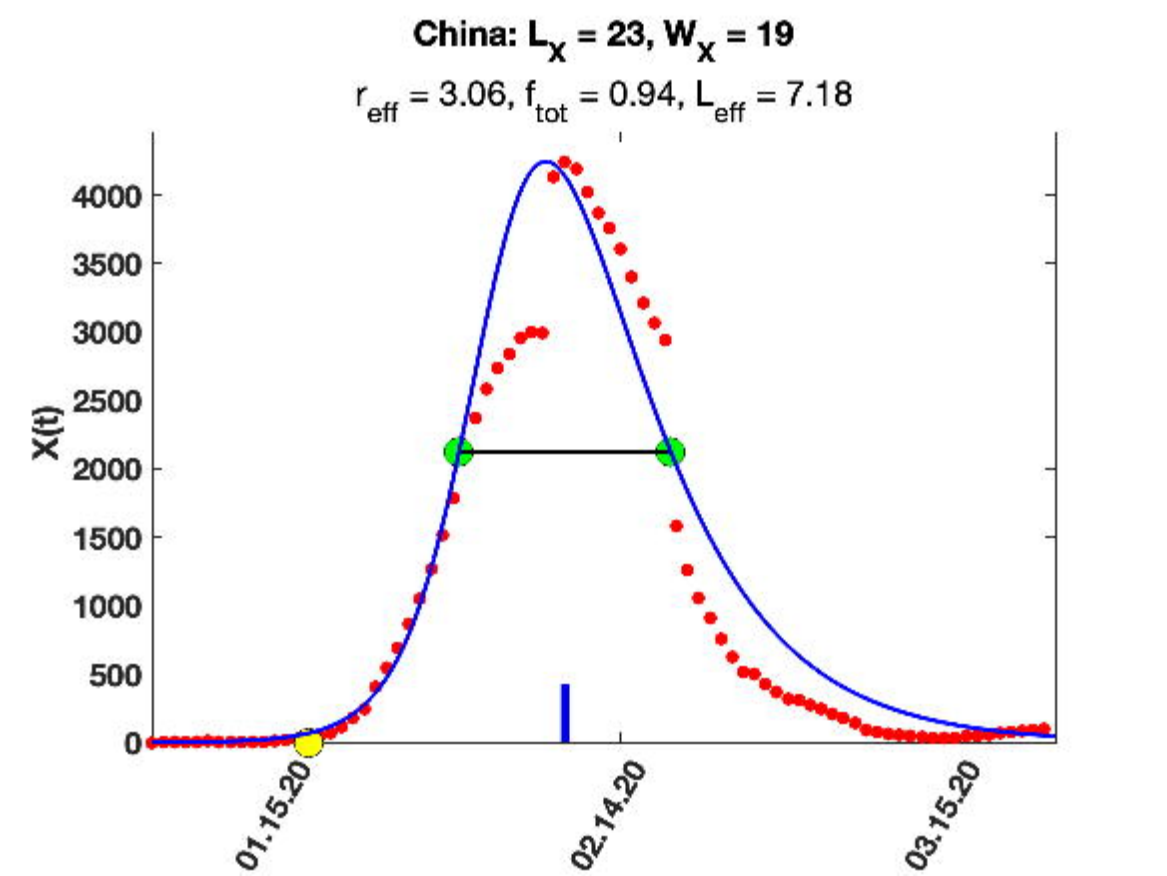

10^2
10^1
W_X/L_{eff}
1
2
3
4
5
6
r_{eff}

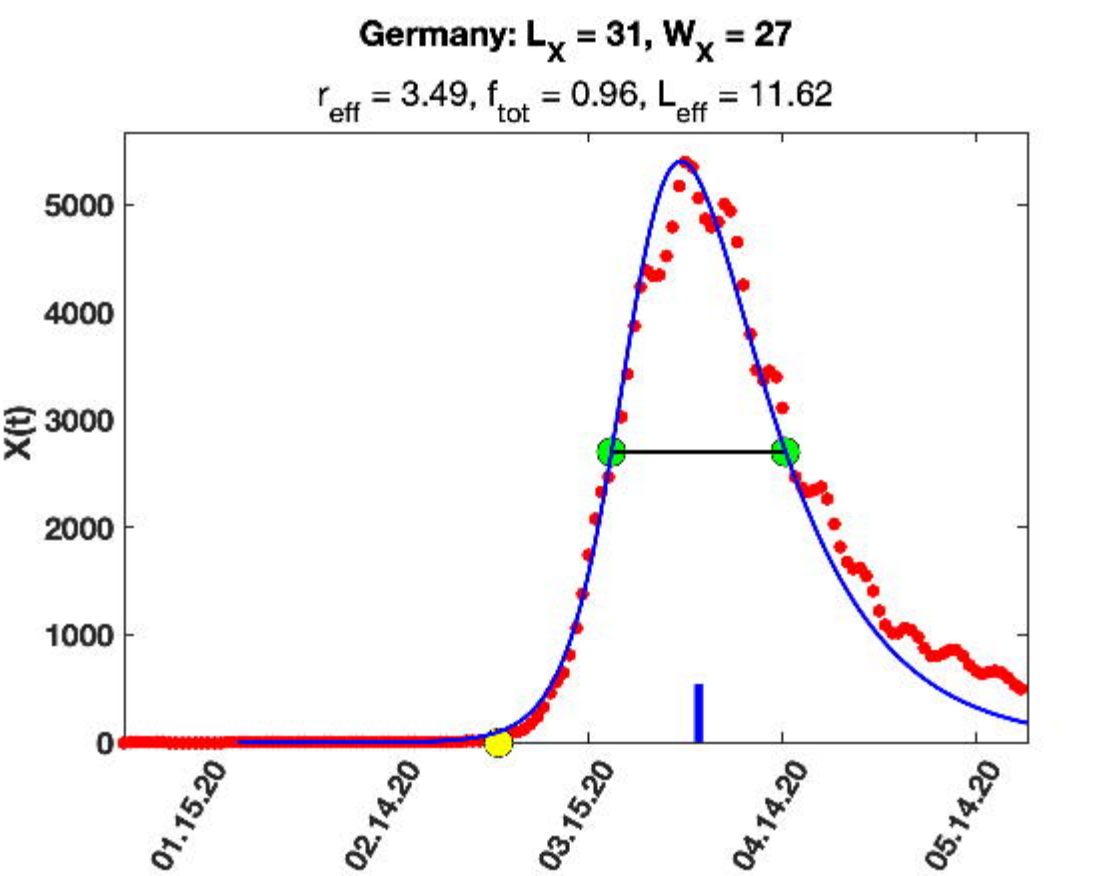

300
250
200
150
100
50
0
0
0.2
0.4
0.6
0.8
1
L_X/L_{eff}
f_{tot}

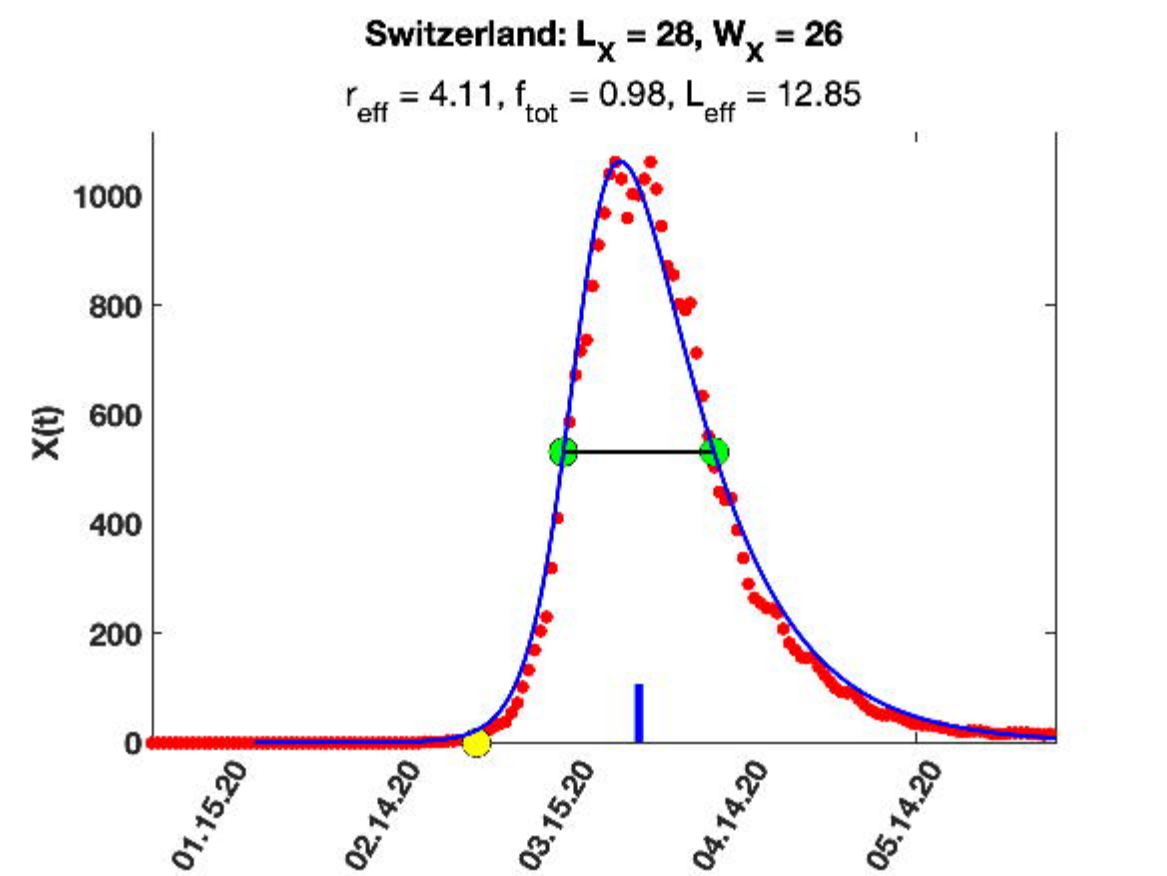

180
160
140
120
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
1
W_X/L_{eff}
f_{tot}

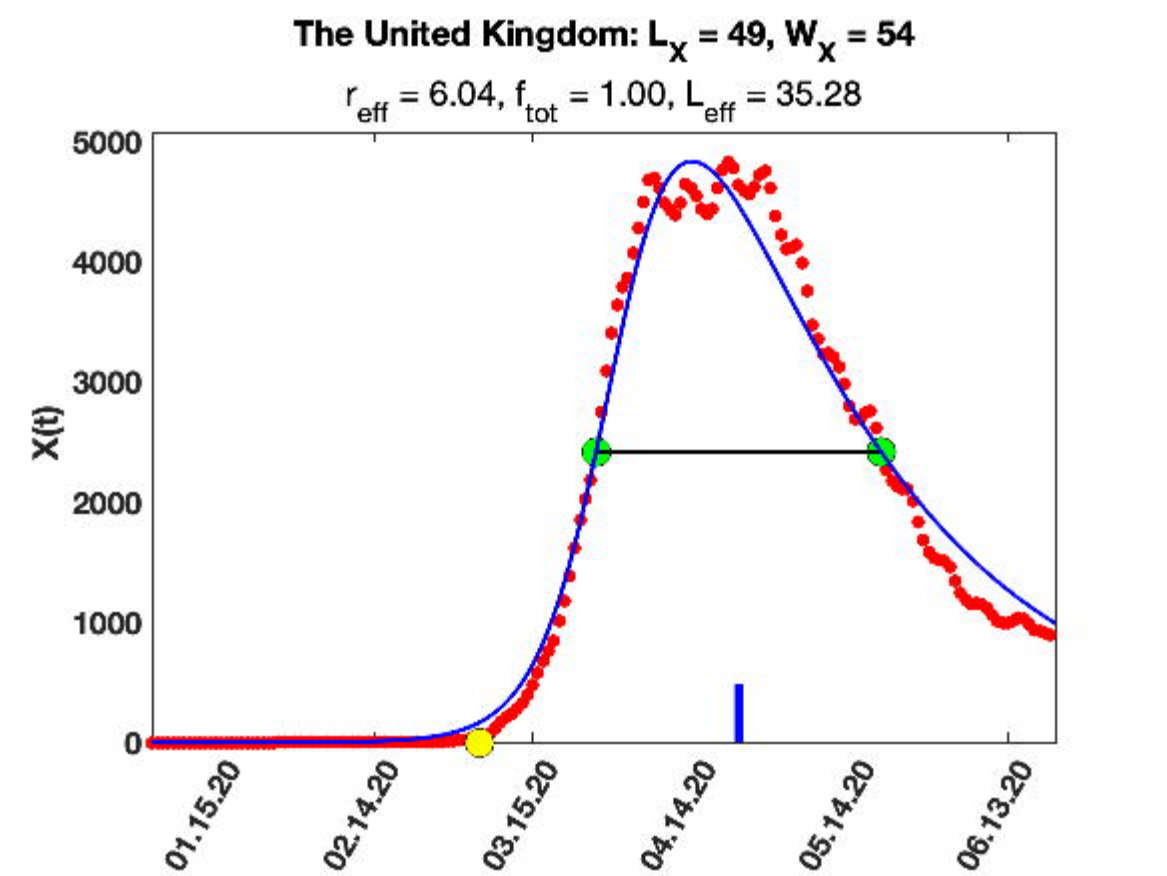

Afghanistan: $L_X = 71$, $W_X = 44$
$r_{eff} = 1.23$, $f_{tot} = 0.33$, $L_{eff} = 2.69$
X(t)
0
100
200
300
400
500
600
700
01.15.20
02.14.20
03.15.20
04.14.20
05.14.20
06.13.20
07.13.20
08.12.20
09.11.20


Austria: L_X = 25, W_X = 19
r_{eff} = 2.43, f_{tot} = 0.87, L_{eff} = 5.51
X(t)
700
600
500
400
300
200
100
0
01.15.20
02.14.20
03.15.20
04.14.20


China: $L_X = 23$, $W_X = 19$
$r_{eff} = 3.06$, $f_{tot} = 0.94$, $L_{eff} = 7.18$
X(t)
4000
3500
3000
2500
2000
1500
1000
500
0
01.15.20
02.14.20
03.15.20


Germany: $L_X = 31$, $W_X = 27$
$r_{eff} = 3.49$, $f_{tot} = 0.96$, $L_{eff} = 11.62$
X(t)
5000
4000
3000
2000
1000
0
01.15.20
02.14.20
03.15.20
04.14.20
05.14.20


Switzerland: $L_X = 28$, $W_X = 26$
$r_{eff} = 4.11$, $f_{tot} = 0.98$, $L_{eff} = 12.85$
X(t)
1000
800
600
400
200
0
01.15.20
02.14.20
03.15.20
04.14.20
05.14.20


The United Kingdom: L_X = 49, W_X = 54
r_{eff} = 6.04, f_{tot} = 1.00, L_{eff} = 35.28
X(t)
5000
4000
3000
2000
1000
0
01.15.20
02.14.20
03.15.20
04.14.20
05.14.20
06.13.20