Power-law distribution in the number of confirmed COVID-19 cases

Bernd Blasius*
Institute for Chemistry and Biology of the Marine Environment,
University of Oldenburg, Oldenburg, Germany
ORCID: 0000-0002-6558-1462
and
Helmholtz Institute for Functional Marine Biodiversity (HIFMB),
Carl von Ossietzky University Oldenburg, Oldenburg, Germany
(Dated: April 3, 2020)

COVID-19 is an emerging respiratory infectious disease caused by the coronavirus SARS-CoV-2. It was first reported on in early December 2019 in Wuhan, China and within three month spread as a pandemic around the whole globe. Here, we study macro-epidemiological patterns along the time course of the pandemic. We compute the distribution of confirmed COVID-19 cases and deaths for countries worldwide and for counties in the US, and provide *prima facie* evidence that both distributions follow a power-law over five orders of magnitude. We are able to explain the origin of this scaling behavior as a dual-scale process: the large-scale spread of the virus between countries and the small-scale accumulation of case numbers within each country. Assuming exponential growth on both scales, the critical exponent of the power-law is determined by the ratio of large-scale to small-scale growth rates. We confirm this theory in numerical simulations in a simple meta-population model, describing the epidemic spread in a network of interconnected countries. Our theory gives a mechanistic explanation why most COVID-19 cases occurred within a few epicenters, at least in the initial phase of the outbreak. Assessing how well a simple dual-scale model predicts the early spread of epidemics, despite the huge contrasts between countries, could help identify critical temporal and spatial scales of response in which to mitigate future epidemic threats.

Introduction

COVID-19 is an emerging infectious disease caused by the coronavirus SARS-CoV-2. It was first reported on in Hubei, mainland China on 31 December 2019 and has spread well outside China in a matter of a few weeks, reaching countries in all parts of the globe within a time span of three month. As of 29 March 2020, the disease has arrived in 177 countries, with more than 700,000 confirmed cases and 30,000 deaths worldwide [33]. Despite the drastic, large-scale containment measures implemented in most countries these numbers are rapidly growing every day - posing an unprecedented threat to the global health and economy of interconnected human societies.

One of the most powerful tools to understand the laws of epidemic growth is mathematical modeling, going back to Daniel Bernoulli's work [3] on the spread of smallpox in 1760. Epidemiological models can be roughly divided into two classes. The first class of models is focused on describing the temporal development of the epidemic within a localized region or country. These models are often variants of the well known SIR-model [17, 18] and have recently been adapted to the situation of COVID-19, taking into account non-pharmaceutical interventions (e.g., quarantine, hospitalization, and containment policies) and allowing first predictions of healthcare demand [11, 20, 31].

The second class of models is concerned with the geographic spread of the epidemic around the globe. For these aims spatially explicit models have been developed that leverage information on the topology of transport networks. For example, the global network of cargo ship movements [16] was used to model the dispersal of invasive species [28]. Similarly for infectious diseases, in a pioneering study, the 2003 spread of SARS in the global aviation network [32] was modeled [15]. Based on these approaches, conceptual frameworks have been developed to estimate epidemic arrival times as effective distances [6]. At the same time, these models have been refined to highly detailed simulation frameworks for predicting the spread of disease and are able to include factors such as vaccination, multiple susceptibility classes, seasonal forcing, and the stochastic movement of individual agents [10, 34]. Reacting rapidly to the emergent pandemic, spatial epidemiological models have been developed to describe and anticipate the spread of COVID-19 [1, 7, 12, 25]. These models allow to predict the incidence of the epidemics in a spatial population through time, permitting to study the impact of travel restrictions and other control measures.

Despite this theoretical progress, not much is known about the biogeography of COVID-19, neither from empirical studies nor from mathematical models. This is astonishing, as one prominent characteristic of the pandemic is the huge variation in the number of cases that have been reported from different countries of the world. As of March 2020, some countries were already badly affected by the pandemic, while others had just confirmed the first few cases. This geographic variation in COVID-

* blasius@icbm.de; http://staff.uol.de/bernd.blasius

19 prevalence might be explained by several arguments: A first obvious possibility would be that the variation is caused by the idiosyncratic circumstances of the individual countries which differ largely in their geography and population size, but also in the way they are combatting the disease. Alternatively, parts of the variation could simply be due to reporting errors, reflecting disparate national testing regimes, with countries such as China, Japan, South Korea, or Germany having high testing rates, in contrast to other countries with much poorer testing. Here, we argue, however, that a dominant part of this variation may be a direct consequence of the dynamics of the spreading process itself. Thereby, the epidemic prevalence in a country should be directly correlated to the arrival time of the disease: countries that were invaded very early by the virus have accumulated many cases in time, while countries with a late invasion naturally still have smaller prevalence.

To test this hypothesis, we use empirical data [8] to compute the country-level distribution, P, of confirmed COVID-19 cases, n, worldwide and find that it follows a power-law

$$P(n) \sim n^{-\mu} \tag{1}$$

over five orders of magnitude.

Power-law distributions characterize a large range of phenomena in natural, economic, and social systems, which is known as Zipf- or Pareto law [9, 21, 22, 30]. Examples range from the number of species in biological taxa [35], the number of cities with a given size [36], the number of different words in human language [36], the distribution of wealth [24], the number of scientific citations [26], the frequency of earthquakes [14], and the popularity of chess openings [5].

Our study shows that epidemic prevalence, at least in the emerging stage of a pandemic, is another system that falls into this class. We provide a conceptual dual-scale model that explains the emergence of the power-law distribution by the 'superposition' of two concurrent processes: large-scale spread of the virus between countries and small-scale snowballing of case numbers within each country. Assuming exponential growth on both scales, the critical exponent is simply determined by the ratio of large-scale to small-scale growth rates. We confirm this theory in numerical simulations in a simple meta-population model, describing the epidemic spread in a network of interconnected countries. By combining real world data, modeling, and numerical simulations we make the case that that the distribution of epidemic prevalence, and possibly that of spreading processes in general, might follow universal rules.

Power-law distribution in empirical data

Our research builds on the COVID-19 data repository operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) [8]. The database contains information about the daily number of confirmed COVID-19 cases and confirmed deaths in various countries worldwide.

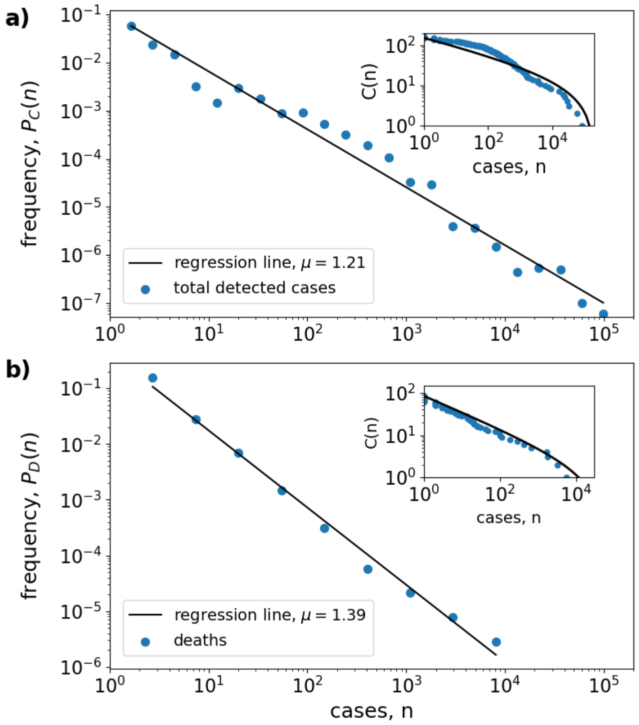

Figure 1. Power-law scaling in the country-level distribution of COVID-19 cases. The figures show the estimated probability $P_x(n)$ for a country to have a certain number n of **(a)** confirmed cases ($x = C$) and **(b)** confirmed deaths ($x = D$) on 22 March, 2020. Histogram bins are spaced equally on a logarithmic axis and only bins with a positive number of entries are shown. Black solid lines show straight-line fits with slope μ, indicated in the figure labels. Insets: Cumulative number $C(n) = \sum_{m=n+1}^{N} P(m)$ of countries with case number $m > n$. Solid lines show the cumulative distribution $C(n) = n^{1-\mu} - n_f^{1-\mu}$ of a truncated power law with cut-off value **(a)** $n_f = 2 \cdot 10^5$ and **(b)** $n_f = 5 \cdot 10^4$.

Using this data we computed the distribution $P_C(n)$ of confirmed cases and the distribution $P_D(n)$ of confirmed deaths at a given date, excluding all countries without cases (independently for confirmed cases and confirmed deaths). To estimate the distribution of case numbers that vary over many orders of magnitude, we first computed the histogram of log-transformed case numbers $\nu = \log(n)$ using equally-spaced bins, which, after normalization, yielded the distribution $\tilde{P}(\nu)$. Next, we used the back-transform $P(n) = \tilde{P}(\exp(\nu))/n$ to obtain the probability distribution $P(n)$ of non-logarithmic case numbers. This procedure results in a histogram with bins that are equally spaced on a logarithmic scale. We have checked that the resulting distribution is largely independent to the choice and number of used histogram bins and other numerical parameters (see also Appendix Fig. 6,

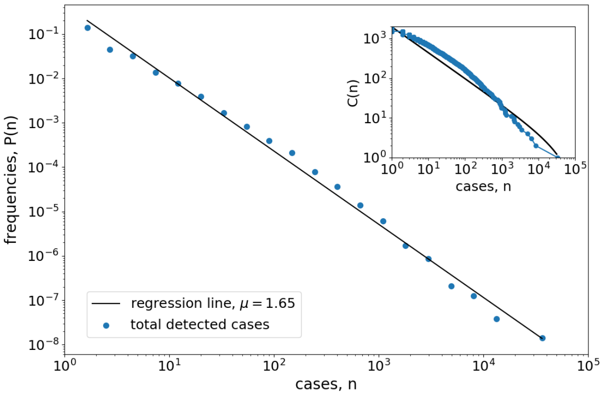

Figure 2. Power-law scaling in the distribution of confirmed COVID-19 cases in the 1962 US counties that have been invaded by the coronavirus on March 29, 2020. Details as in Fig. 1a. The cut-off value for the cumulative distribution in the inset was set to $n_f = 1 \cdot 10^5$.

where we tested the histogram algorithm on artificially generated random numbers). Additionally, we computed the cumulative number $C(n) = \sum_{m=n+1}^{N} P(m)$ of countries with case number $m > n$. This was obtained by taking a rank-plot of case numbers and inverting axes, i.e., sorting the array of case numbers in descending order and plotting for each country the rank as a function of the sorted case number on double-logarithmic axes.

The country-level prevalence distribution on March 22, 2020 is shown in Fig. 1. On that day 168 countries were invaded by the coronavirus. The number of confirmed cases varied between 81,435 cases in China (followed by 59,138 cases in Italy) and 1 case in 16 countries. The number of confirmed deaths varied between 5,476 in Italy (followed by 3,274 in China) and one or zero deaths in many countries. Fig. 1 clearly demonstrates that the probability for a country to have a certain number of COVID-19 cases follows a broad, long-tailed distribution that in very good approximation can be described by a power-law, spanning five orders of magnitude for the confirmed number of cases and four orders of magnitude for the confirmed number of deaths.

At this point it is important to stress that our aim is not to prove that the COVID-19 case distribution is a perfect power-law, an undertaking that would require sophisticated statistical analysis and a much larger sample size [9]. Instead, our claim is merely that the empirical data are highly *consistent with the hypothesis* that the number of reported cases are taken from a distribution of the form of Eq. (2) and therefore, our focus is on possible mechanistic explanations and the epidemiological implications of such a broad distribution (see Discussion).

To illustrate the robustness of our hypothesis to spatial scale, in Fig. 2 we depict the same analysis for the distribution of confirmed COVID-19 cases in US counties on March 29, 2020. On this day 1962 counties were invaded by the virus and epidemic prevalence varied between 33,768 confirmed cases in New York City and one case in 456 counties. Again, we find that the distribution of confirmed cases follows a power-law over several orders of magnitude. Thus, although the two data sets differ greatly in spatial scale and resolution (168 invaded countries in Fig. 1 vs. 1962 invaded US counties in Fig. 2) we obtain very similar prevalence patterns, confirming the robustness of our analysis.

A crude estimation of the critical exponent can be obtained by measuring the slope of a regression line through the data on a double-logarithmic plot. Applying this method to the country-level distribution (Fig. 1), we obtain a value of $\mu_C = 1.21$ (slope of the distribution of confirmed cases) and $\mu_D = 1.39$ (confirmed deaths). For the US-county distribution (Fig. 2) we obtain a somewhat larger slope of $\mu_C = 1.65$.

A more accurate estimation of the critical exponent of a power-law distribution is given by the log-likelihood estimator [22]

$$\hat{\mu} = 1 + n \left[\sum_i \frac{n_i}{n_{min}} \right]^{-1} \tag{2}$$

with a standard error of

$$\sigma = \frac{\hat{\mu} - 1}{\sqrt{n}}. \tag{3}$$

Here, the 'hat' means that this is an estimated value. Applying this formula to the country-level COVID-19 distribution (where the minimal case number equals $n_{min} = 1$ individuals) yields critical exponents of $\hat{\mu}_C = 1.24 \pm 0.02$ and $\hat{\mu}_D = 1.53 \pm 0.04$. For the US-county distribution we obtain the value $\hat{\mu}_C = 1.54 \pm 0.01$. These exponents are slightly larger than those obtained from the regression analysis, but are still in the same ballpark.

Given a perfect power-law distribution, Eq. (1), the cumulative distribution function $C(n) = \int_n^\infty P(n')dn'$ should also follow a power-law $C(n) \sim n^{-\mu-1}$. As shown in the insets in Figs. 1 and 2, this is not the case for the distribution of COVID-19 cases, for which the cumulative numbers $C(n)$ do not really follow a straight line in a double logarithmic plot. Instead, they are better described by a truncated power law, that is, a distribution with a maximal case number n_f, for which the cumulative distribution function reads

$$C(n) = \int_n^{n_f} P(n')dn' \sim n^{(1-\mu)} - n_f^{(1-\mu)}. \tag{4}$$

This is not necessarily a strong evidence against the hypothesis of a power-law distribution because we observe similar behavior also when we analyze artificially generated random numbers, taken from a power-law distribution with a small sample size and a small critical exponent (see Appendix Fig. 6).

The presence of a power-law distribution means that global COVID-19 prevalence patterns are characterized

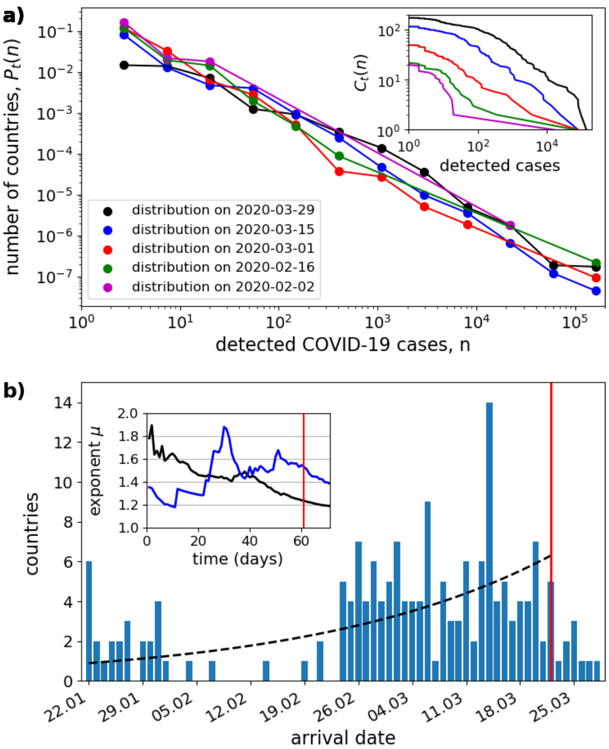

Figure 3. Temporal development of the COVID-19 pandemic. **a)** Evolution of the distribution of confirmed cases per country. The same as Fig.1a, but for five different time instances separated by 2 weeks (see figure legend) during the pandemic. **b)** Distribution of arrival times. The histogram shows the number of countries that were invaded by the virus on a certain day between Jan 22, 2020 and March 29, 2020 (blue bars). Further shown is an exponential increasing function, $\exp(st)$ (black dashed line) with growth rate $s = 0.03\ d^{-1}$, obtained by a least square fit to the histogram during the first 61 days. The inset shows the critical exponents $\hat{\mu}_C(t)$ (black) and $\hat{\mu}_D(t)$ (blue), estimated by Eq. (2), as a function of time. The vertical red line indicates 22 March, the date of the distribution shown in Fig. 1.

by a small number of countries with large epidemic prevalence and a large number of countries that are (yet) barely affected by the disease. In general, the obtained critical exponents are rather small. While for most natural power-law distribution critical exponents are around $\mu \approx 2$, here we estimate exponents that are clearly below two, $\mu < 2$, indicating a very broad distribution for which the mean value diverges.

Temporal development during the pandemic spread

While the present analysis considers the distribution of case numbers at a temporal snapshot, in reality the pandemic is a dynamic process successively invading countries worldwide. In Fig. 3 we investigate the temporal development of the COVID-19 distribution over the time course of the pandemic. The figure shows that the country-level distribution of confirmed cases is formed already within a few weeks from the start of the outbreak and remains roughly stationary over the considered time interval of 68 days. A closer inspection (see inset in Fig. 3b) reveals that the critical exponents in fact are not constant, but vary during the course of the pandemic. Thereby, the exponent $\hat{\mu}_C$ of the distribution of confirmed cases is decreasing in time, while the exponent $\hat{\mu}_D$ corresponding to the distribution of confirmed deaths, at first is increasing in time and starts to fall again after March 12. Over the whole time span, the two exponents are well below 2.

Fig. 3b further investigates the spatial spread of COVID-19 across countries worldwide more systematically. The figure plots the number of countries that were invaded by the coronavirus (i.e., having the first confirmed COVID-19 case) at a particular day in the time span from January 22 to March 29, 2020. On January 22, the first entry in the database, six countries (China, Japan, South Korea, Taiwan, Thailand, US) were already invaded by the virus. From this day, within roughly two months the pandemic spread to nearly every country in the world.

Interestingly, the invasion speed was not constant. Instead Fig. 3b clearly indicates two broad modes in the arrival time distribution. Many countries were invaded by the disease in the end of January. In contrast, in the first three weeks of February nearly no new arrivals were reported. Only after February 24 did a second wave of invasions appear, which lasted until the end of March. After this, the number of new arrivals began to fall again, probably reflecting the fact that the pandemic had reached basically all countries of the world. As of March 29, the first day without a new reported invasion after a series 34 days, a total 177 countries were invaded by the coronavirus.

There are several possible reasons why the disease arrival is not more evenly distributed. One explanation for the bimodal shape is related to the lockdown of airline transportation in China in the end of January 2020. According to this hypothesis, after the first pandemic bubble in January, the further spread of the pandemic came to a temporary standstill with the onset of travel restrictions, only to resurface in a second wave, starting end of February. Alternatively, it may be that many arrivals of the virus in countries all over the world simply went undetected during the first weeks of February and were detected only later with the increasing awareness and increased testing. This hypothesis is corroborated by the observation that end of February is also the time when the first PCR based tests became available.

Mechanistic explanation of the power-law distribution

Fig. 3 would suggest that the temporal development of the pandemic is characterized by two complementary processes: the successive invasion of more and more countries and the increasing number of cases within each affected country. Here we argue that the emergence of the power-law distribution could be related to the concurrent 'superposition' of these two processes. Thereby, on a large geographic scale, the pandemic is driven by the spread of the virus in the network of interconnected countries. On a small scale, case numbers are snowballing within each country, once it has been invaded, thereby further increasing the epidemic imbalance due to different arrival times between countries.

In the simplest approximation, at the begin of the pandemic both of these processes developed exponentially in time. A straightforward calculation shows that the combination of the two exponential processes generically yields a power-law distribution in the number of cases in countries. Let us first assume that the probability of a country to be invaded by the virus at time t_i grows exponential in t_i with spreading rate s,

$$P(t_i) \sim e^{st_i}. \tag{5}$$

This function corresponds to an exponential growth in the geographic distribution of the pandemic and would be the expectation if one modeled the spread in a network where nodes are countries (neglecting saturation when the pandemic has reached most countries).

Second, we assume that in each country the number of confirmed cases grows exponentially in the time since invasion with growth rate r (neglecting saturation after the epidemic peak)

$$n(t) \sim e^{r(t-t_i)}. \tag{6}$$

Combining these two equations, the probability distribution of confirmed cases $P(n)$ can be calculated as (see [22])

$$P(n) = P(t_i)\left|\frac{dt_i}{dn}\right| \sim \frac{e^{st_i}}{e^{-rt_i}} \sim n^{-(1+s/r)} \tag{7}$$

which is a power-law with critical exponent

$$\mu = 1 + \frac{s}{r}. \tag{8}$$

Thus, the critical exponent is simply determined by the ratio of large-scale to small-scale growth rates. In the symmetric case that both growth rates are identical, $s = r$, we would expect a power law with $\mu = 2$. In the limiting case that the large-scaling spreading process is linear in time, $s = 0$, we obtain a border-line distribution with critical exponent $\mu = 1$.

Obviously, this simple theory far from accurately describes a real-word pandemic. First of all, the theory is valid only in the initial phase of the pandemic, while both geographical spread and within-country epidemic growth are still exponential. As soon as saturation processes set in, the derivation of the power law breaks down. Next, as shown in Fig. 3b the arrival time distribution during the COVID-19 pandemic is not exponential, as discussed above. In gross oversimplification we may nevertheless fit an exponential function $P(t) \sim e^{st}$ through the data, yielding an 'average' spreading rate of $s = 0.03\ d^{-1}$ (black dashed line in Fig. 3b). Finally, epidemic growth rates during the COVID-19 pandemic have not been not identical in all countries (even in the initial stages). They have also not remained constant in time, but in most countries have fallen in the course of the epidemic. Furthermore, most countries were invaded multiple times, leading to different epidemic foci within countries. Neglecting all these observations, for the sake of argument, let us assume an average doubling time of case numbers of $T_{1/2} = 5\ d$ in all countries, yielding an exponential growth rate of $r = \log(2)/T_{1/2} = 0.14\ d^{-1}$. Then, according to our simple theory Eq. (8) we would expect a critical exponent of $\mu = 1 + 0.03/0.14 \approx 1.21$ in rather good agreement to the fitted exponents in Fig. 1.

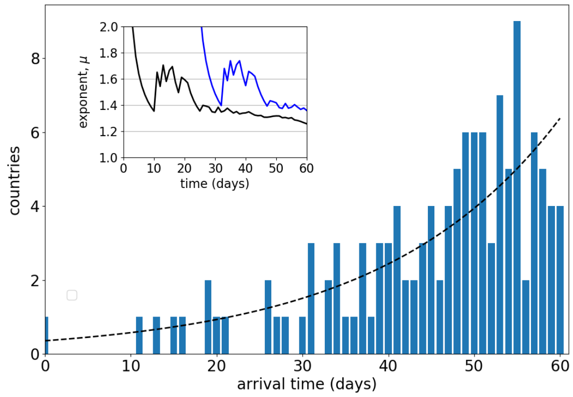

Figure 4. Spatial spread in the meta-population model. Similar to Fig. 3b, the plotted histogram shows the number of countries that were invaded by the virus on a certain day for the simulation time of 60 days. Parameter values: number of countries $M = 200$ and invasion probability $p = 5 \cdot 10^{-4}$. The black dashed line shows an exponentially increasing function, $\exp(st)$ with spreading rate $s = 0.048\ d^{-1}$, obtained by a least square fit to the data. The inset shows the time dependence of the critical exponents $\hat{\mu}_C(t)$ (black) and $\hat{\mu}_D(t)$ (blue) for the distribution of the number of cases and deaths, estimated by Eq. (2).

Results from a meta-population model

To test the theory of the previous section, we develop a conceptual dual-scale meta-population model. The first level describes the large-scale spread of the virus in a network of M interconnected countries. The state of a

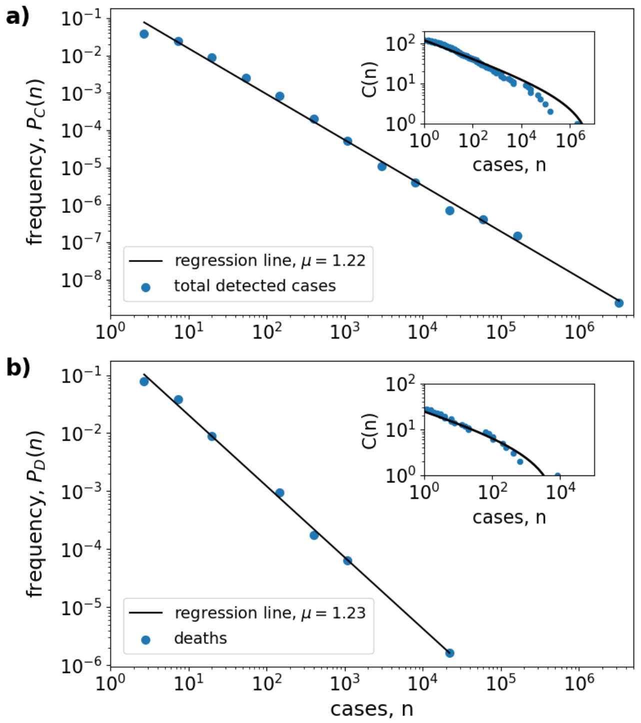

Figure 5. Power-law scaling in the simulated distribution of COVID-19 cases. Same as Fig. 1, but for the meta-population model after a simulation time of 60 days. Parameter values: country population size $N = 5 \cdot 10^7$, case fatality rate $m = 0.01$, infectious period $1/\gamma = 6d$, and contact rate $\beta = 0.4d^{-1}$; yielding a growth rate $r = \beta - \gamma = 0.23d^{-1}$, corresponding to a doubling time of $T_{1,2} = \log(2)/r = 3d$ and a basic reproduction number $R_0 = \beta/\gamma = 2.4$. Cut-off values of the cumulative distribution were set to **(a)** $n_f = 1 \cdot 10^7$ and **(b)** $n_f = 1 \cdot 10^4$. Other parameters as in Fig. 4.

country is given as a Boolean value, being either invaded by the virus or non-invaded. The model starts with a single invaded country. The geographic spread runs in discrete time, each step corresponding to one day of the time-continuous small-scale model. In each time step, every non-invaded country becomes infected by an invaded country with probability p. Thus, if at time t a number of m countries have already been invaded, the probability for a non-invaded country to receive the virus in this time step equals $1-(1-p)^m$. Thereby, the number of invaded countries grows stochastically and roughly follows a sigmoidal shape. The arrival time distribution is a unimodal function of time that starts exponentially. This is shown in the exemplary simulation run in Fig. 4 for the first 60 time steps, after which 123 out of the $M = 200$ countries were invaded by the virus.

The small-scale model is time-continuous and deterministically describes the epidemic dynamics within a country, which is started in each country from the time point of invasion by the virus. The model determines the time course of susceptible S, infected I, recovered R, and dead D from a standard SIR-model [17, 18]

$$\dot{S} = -\beta\frac{S}{N}I,\ \dot{R} = \beta\frac{S}{N}I - \gamma I,\ \dot{R} = (1-m)\gamma I,\ \dot{D} = m\gamma I. \tag{9}$$

Here, N is the constant population size in the country, β is the contact rate, $1/\gamma$ the infectious period, and m the case fatality rate. The total number of infected is determined as $C = I + R + D$. In the small-scale model countries are simulated independently from each other and are only coupled by the unique invasion event for each country, which starts the epidemic growth in that country with initial values $S(0) = 5 \cdot 10^7$, $I(0) = 1$ and $R(0) = D(0) = 0$. All infection state variables in a country are zero before invasion by the virus, $I = R = D = 0$. The resulting epidemic dynamics in a country is shown in the Appendix Fig. 7. The dynamics follows the well-known SIR-curve. With the chosen parameterization, it takes roughly 80 days until the epidemic peak is reached. After this time, the assumption of an exponential increase, Eq. (6), breaks down.

Combining the large-scale and small-scale model components allows to simulate the epidemic prevalence in each country as a function of time. Fig. 5 shows the resulting distribution of cases and deaths after a simulation time of 60 days. Again, the distributions are characterized by a power-law scaling. Comparison with Fig. 1 shows that the model is able to describe the characteristics of the empirical distribution of COVID-19 cases rather well. The log-likelihood estimation of the critical exponents yields values of $\hat{\mu}_C = 1.26 \pm 0.02$ and $\hat{\mu}_D = 1.36 \pm 0.03$. These exponents can be compared to our theory Eq. (8). From Fig. 4 we estimated a spatial spreading rate of $s = 0.048\ d^{-1}$. The initial growth rate of infected in the SIR-model equals $r = 0.23\ d^{-1}$. Thus, according to Eq. (8) we would expect a critical exponent of $\mu = 1 + 0.048/0.23 = 0.21$, in good agreement to the estimated value from the numerical simulation.

We want to note that the scaling relation is lost when the spatial spreading starts to saturate. Eventually, in the limit of large time, the distribution of cases must converge towards a delta function $P(n) = \delta(n - fN)$, with f the fraction of susceptible in a country that will be infected, when the epidemic has come to an end in every country. Interestingly, in our numerical simulations, we still obtained power-law distribution when the contact rate β was set to large value, so that the dynamics within a country rapidly reach a stationary state. In this case, with increasing β the critical exponents tended to $\mu \to 1$.

Discussion

Our finding of power-law distributions in the number of reported cases has important consequences for epidemiology. Most notably, the small values of the estimated critical power-law exponents are related to the strong inequality of case numbers that was frequently observed all over the world in the initial phase of the COVID-19 outbreak. Following a power law distribution means that this pattern prevails even as numbers grew and the scale of infection expanded globally. In particular, during the course of the pandemic, most cases were reported to have occurred in a few countries, sometimes even a single country - the so-called epicenters of the pandemic. The distribution of cases within countries followed a similar pattern. Often COVID-19 was peaking in a few localized foci (local regions or cities), while other parts of the country at the same time had experienced only a moderate number of cases. Our theory provides a mechanistic explanation why this might have been the case.

A graphical representation for the inequality of a distribution is given by the Lorenz curve [22] which in the case of the COVID-19 case distribution is a plot of the fraction of the total number of confirmed cases in dependence of the fraction of the most affected countries. This is shown in the Appendix Fig. 8 for the number of confirmed COVID-19 cases and confirmed deaths on 22 March 2020. The Lorenz curve shows that on this day 95.7% of confirmed cases and 97.6% of the confirmed deaths had been reported in the 20% most affected countries (while the top 5% most affected countries had accumulated 82.3% of all confirmed cases and 84.4% of all confirmed deaths). With 81,435 out of 336,953 confirmed cases on that day China alone had accumulated a fraction of 24% of all cases. The two most affected countries, China and Italy, together had accumulated a fraction of 41% of the worldwide reported cases.

This inequality can also be measured by the Gini-coefficient G [13], which ranges between $G = 0$ for perfect equality, i.e., all countries having the same number of cases, and $G = 1$, corresponding to maximal inequality, where all cases appear in a single country. For the distribution of confirmed COVID-19 cases on March 22 (Fig. 1) we obtain a Gini-coefficient of $G = 0.92$ and for the number of confirmed death of $G = 0.94$. These large values are a direct consequence of the small critical exponents of the estimated power-law distributions. In fact, for $\mu < 2$ one would theoretically expect a Gini-coefficient of $G = 1$ [22].

The emergence of power-law distributions with a small critical exponent and the associated inequality of the distribution, with Gini coefficients close to one is also observed in the developed meta-population model. Consequently, also in the model case numbers are mostly concentrated in a few countries. In the simulations, these epicenters of the pandemic, i.e., the countries with most cases, are always the countries in which the diseases originated or which were first invaded by the virus. In other words, the prevalence rank order among countries remains unchanged during the course of the pandemic. This is akin to the "rich-get-richer process" or "first-mover-advantage" [23, 29], a well-studied process to generate power-law distributions. In the real COVID-19 pandemic, this was not the case. During the begin of the pandemic most cases were observed in China, later the "leading role" changed next to Italy and finally to the USA. This reflects different mitigation strategies and circumstances in different countries, a factor that is not considered in the simple model. Nevertheless, despite these changes in the rank order, the distribution of cases in the empirical data was always closely represented by a power-law.

Remarkably, we obtained power-law distributions in the absolute number of cases in each country. At first guess, one might have expected such scaling only after case numbers have been normalized by population sizes. Our preliminary investigations show that such normalized case numbers become even more unequally distributed, with even smaller estimated values of the critical exponent, and the distributed values do not line-up any more so well on a straight line on a double logarithmic plot. Thus 'folding' the distribution of population sizes over the COVID-19 case distribution does not flatten, but rather tends to further increase, the inequality of the resulting distribution. This indicates that absolute (non-normalized) case numbers may be the natural variables to describe the patterns of the pandemic in its initial stage. In all likelihood, the role of country sizes and population numbers will become increasingly important with the further spread of the pandemic.

It is well known from the literature (e.g., [9, 22]) that caution is in order when trying to identify power-law distributions in real data and, in particular, that a straight line in a double-logarithmic plot does not suffice to prove the existence of a power law distribution. Therefore we repeat that the aim of this study is not to proof that the COVID-19 case distribution is a perfect power-law, nor do we intend to rule-out other likely candidate distributions (e.g., log-normal or stretched exponential distributions). Instead, our claim is merely to demonstrate that the empirical data are highly consistent with the hypothesis that the number of reported cases are taken from a power-law distribution of the form Eq. (2). This is consistent with our simple theory which predicts power-law distributions only as an approximation in the initial phase of the pandemic, while in all likelihood the distribution of case numbers will drastically change as soon as saturation effects start to become important. This motivates our focus on possible mechanistic explanations for the observed distribution, and the epidemiological implications of such a broad distribution.

Nevertheless, the scaling in the distributions shown in (Figs. 1 and 2) is remarkably constant over the whole range of case numbers, stretching several orders of magnitude with no obvious signs of saturation for either the range of small or large case numbers. One might argue that the bend in the cumulative distribution is a sign that the growth in some countries (e.g., Italy, China, Korea) had already become sub-exponential. However, this is contradicted by the observation that a similar bend is also exhibited by the cumulative distribution obtained from the meta-population model (Fig. 5) and from artificially generated random numbers with a small sample size (Fig. 6).

We would like to remark that the available database only provides information on the reported COVID-19 cases in each country. In all likelihood, the real number of cases will be much larger. Not much is known about the reporting rates, but first estimates indicate that a substantial fraction (possible 86%) of infections might go undetected [19]. Reporting rates probably vary strongly between countries and may change in time with the awareness of national health institutions and available testing capabilities. Further uncertainties arise because the criteria by which a person is classified as active case (and even more so for being classified as recovered) vary between countries and not uncommonly have been modified during the course of the pandemic within a country.

We have shown that a simple conceptual model yields an accurate description of the COVID-19 prevalence distribution in the initial phase of the pandemic. This is remarkable in that many important epidemiological aspects of the spreading process are not captured by the model. Most notably, the model does not take into account variability in country sizes, population numbers, or testing rates. Furthermore, the model does not consider the heterogeneity of intra- and inter-country connectivity, as well as the corresponding changes due to social distancing, lock-down measures, closing of airline connections and shut-down of borders.

These simplifications leave much room for future investigations and model improvements. For, example, while we have assumed constant and identical population sizes, in reality country sizes are highly heterogenous. In general, the epidemic growth rate and also the maximal prevalence in a country should be correlated to the population number. This effect may not be so strong in the initial stages of a pandemic, but should become increasingly important, the further the spread has been going. In the limit of large time, the distribution $P(n)$ should become stationary and approach the distribution of population numbers. These ideas could be readily checked in numerical simulations in a metapopulation which considers heterogeneously distributed country sizes. Furthermore, the model assumes only single infections in each country. One obvious improvement would be to make this initial number of infected individuals a random number, as would be a better description of what happened in many countries.

One basic assumption of the developed model is the separation of the pandemic into two spatial scales, the large-spatial spread over a rather small number ($M < 200$) of interconnected countries and the small-scale growth within a population of much larger size ($N = 5 \cdot 10^7$). This separation, obviously is somewhat arbitrary. For the virus countries are, of course, quasi-arbitrary entities. Therefore, it would be important to check whether both the data analysis (Fig. 1) and the mathematical model are robust to arbitrarily subdividing or lumping countries. The very similar scaling observed among US counties (Fig. 2) lends credence to the model's generality. Similarly, one can readily ascertain that the model result is not an artifact of artificial lumping. Suppose a virus that is spreading in an all-to-all, or randomly coupled, network of a number of NM individuals. If we would artificially subdivide individuals into a small number M of classes (or countries), at the time point when the disease has spread to all countries, within each country we would still have only a few cases (of the order of $M \ll N$). Thus, the assumed simultaneous spread on both spatial scales requires a real physical separation in the network structure. It would be an interesting perspective for future research to study the spread in multi-scale hierarchies or in more realistic models of interconnected societies.

Finally, we would like to remark that the model's strong simplicity is at the same time a strength: being rather generic, it should be applicable to very different systems, to describe the spread of commodities as a process with two spatial scales. The fact that the distribution of COVID-19 resembles a model where only the initial infection 'counts' reflects the intrinsic difficulty in containing epidemics at global and local scales when unilateral measures (e.g., travel bans and lockdowns) are impractical or non-enforceable, i.e., where other countries or regions will step up and continue the spread. Thus, assessing how a well simple dual-scale model predicts the early spread of epidemics, despite the huge contrasts between countries, could help identify critical temporal and spatial scales of response in which to mitigate future epidemic threats.

Acknowledgments I would like to thank Christoph Feenders, Thilo Gross, Alastair Jamieson-Lane, Cora Kohlmeier, James McLaren, and Alexey Ryabov for helpful comments regarding the manuscript.

Software and data availability Code for data analysis and numerical simulations was written in Julia [4]. The differential equations were solved with the package DifferentialEquations.jl [27]. The source code of the used algorithms will be made publicly available at `https://github.com/berndblasius/Covid19`. The used data are available at `https://github.com/CSSEGISandData`.

Appendix: supplementary figures

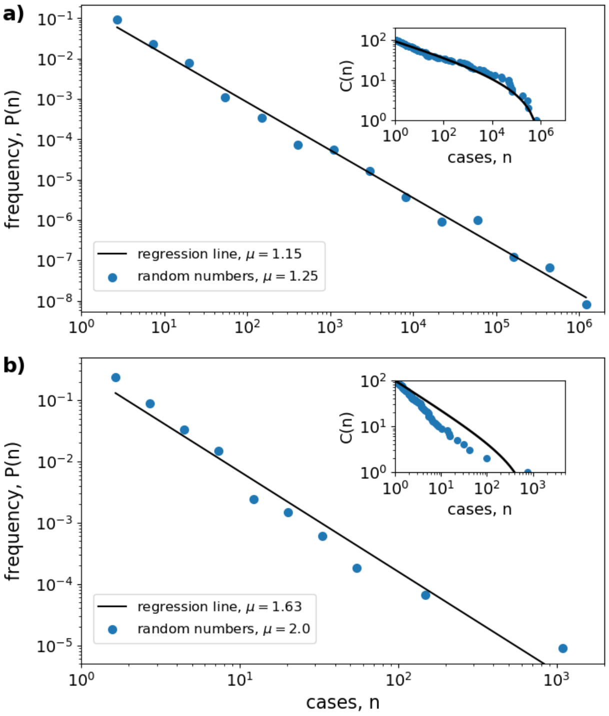

Figure 6. Robustness of the algorithm for computing log-binned histograms. The same analysis as in Fig.1, but for a small number of 100 random numbers that were generated from a power-law distribution with **(a)** $\mu = 1.25$ and **(b)** $\mu = 2.0$. The estimated distribution roughly follows a straight line on the double-logarithmic plot with equally spaced bins. Note, that even though only 100 random numbers were drawn, the estimated probabilities vary over many orders of magnitude (which is numerically possible since in order to compute the probability distribution, the counted histogram numbers are divided by the bin widths). Using the log-likelihood estimation of critical exponents, Eq. (2), yields **(a)** $\mu = 1.25$ and **(b)** $\mu = 1.99$ in good agreement with the actually used exponents. In contrast, the estimation by regression lines (see figure labels) yields exponents that are too small. For the small exponent of $\mu = 1.25$ in **(a)** the cumulative distribution $C(n)$ does not follow a straight line on the double logarithmic axes and instead is better described by a truncated power-law (cut-off values were chosen as **(a)** $n_f = 1 \cdot 10^6$ and **(b)** $n_f = 1 \cdot 10^3$). This is similar to the behavior observed in the COVID-19 distributions. Note that here we have chosen a very small sample size, in accord to the COVID-19 data. Therefore, the actual shape of $C(n)$ varies between different realizations of the random numbers.

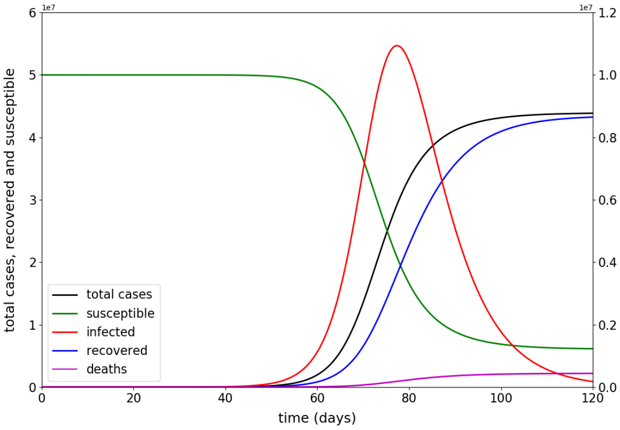

Figure 7. Simulation of the SIR-model (9) within a country. The plot shows the numerically obtained values of the total number of cases C (black), the number of susceptible S (green), and the number of recovered R (blue) on the left axis, as well as the number of infected I (red) and deaths D (magenta) and the right axis as a function of time. For the used initial values of $S(0) = 5 \cdot 10^7$, $I(0) = 1$ and $R(0) = D(0) = 0$, the epidemic peak is reached after 77 days. Parameter values as in Fig. 5.

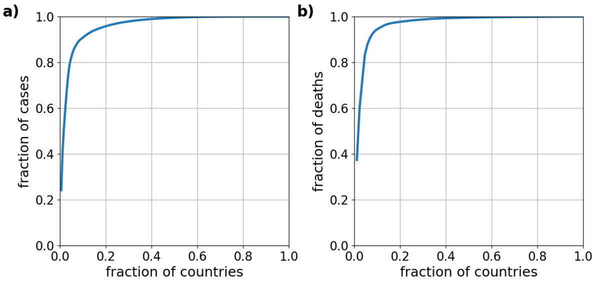

Figure 8. Lorenz curves, depicting the inequality in the distribution of confirmed COVID-19 cases. The plots show the fraction of the number of confirmed cases **(a)** and of the number of confirmed deaths **(b)** as a function of the fraction of most affected countries on March 22, 2020 (compare to Fig. 1). This inequality corresponds to a Gini-coefficient of $G = 0.92$ for the distribution of confirmed cases and of $G = 0.94$ for the number of confirmed deaths.

[1] Arenas A, Cota W, Gomez-Gardenes JG, Gomez S, Granell C, Matamalas JT, Soriano D, Steinegger B (2020) A mathematical model for the spatiotemporal epidemic spreading of COVID19. medRxiv 2020.03.21.20040022.

[2] Barabási AL, Albert R, (1999). Emergence of scaling in random networks. Science 286: 509-512.

[3] Bernoulli, D. (1760) M'em. Math. Phys. Acad. R. Sci. Paris, 1?45

[4] Bezanson J, Edelman A, Karpinski S, Shah VB (2017) Julia: A fresh approach to numerical computing. SIAM Review 59: 65-98.

[5] Blasius B, Tönjes R. (2009) Zipf's law in the popularity distribution of chess openings. Physical Review Letters 103: 218701.

[6] Brockmann D, Helbing D (2013) The hidden geometry of complex, network-driven contagion phenomena. Science 342: 1337-42.

[7] Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, Piontti AP, Rossi L, Sun K, Viboud C, Xiong X, Yu H, Halloran ME, Longini IM, Vespignani A (2020) The effect of travel restrictions on the spread of the 2019 novel coronavirus (2019-nCoV) outbreak. medRxiv 2020.02.09.20021261. Science eaba9757. DOI: 10.1126/science.aba9757.

[8] Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases. 2020 Feb 19. Doi: 10.1016/S1473-3099(20)30120-1

[9] Clauset A, Shalizi CR, Newman ME (2009). Power-law distributions in empirical data. SIAM Review 51:661-703.

[10] Colizza V, Barrat A, Barthelemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proceedings of the National Academy of Sciences 103: 2015-20.

[11] Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, Bhatia S, Boonyasiri A, Cucunuba Z, Cuomo-Dannenburg G, Dighe A (2020) Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. preprint, Doi:10.25561/77482.

[12] Gilbert M, Pullano G, Pinotti F, Valdano E, Poletto C, Boelle PY, D'Ortenzio E, Yazdanpanah Y, Eholie SP, Altmann M, Gutierrez B (2020) Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. The Lancet.

[13] Gini C (1912) Variabilita e mutabilita. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi.

[14] Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. Bulletin of the Seismological Society of America 34: 185-8.

[15] Hufnagel L, Brockmann D, Geisel T (2004) Forecast and control of epidemics in a globalized world. Proceedings of the National Academy of Sciences 101: 15124-9.

[16] Kaluza P, Kölzsch A, Gastner MT, Blasius B (2010) The complex network of global cargo ship movements. Journal of the Royal Society Interface 7: 1093-103.

[17] Keeling MJ, Rohani P (2008) Modeling infectious diseases in humans and animals. Princeton University Press, Princeton. OCLC: ocn163616681.

[18] Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proceedings Royal Society of London A 115: 700-21.

[19] Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science. 2020 Mar 16.

[20] Maier BF, Brockmann D (2020) Effective containment explains sub-exponential growth in confirmed cases of recent COVID-19 outbreak in Mainland China. arXiv preprint arXiv:2002.07572.

[21] Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. Internet Mathematics 1: 226-51.

[22] Newman ME (2005) Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46: 323-351.

[23] Newman ME (2009) The first-mover advantage in scientific publication. EPL (Europhysics Letters) 86: 68001.

[24] Pareto V (1964) Cours d'economie politique. Librairie Droz.

[25] Pullano G, Pinotti F, Valdano E, Boelle PY, Poletto C, Colizza V (2020) Novel coronavirus (2019-nCoV) early-stage importation risk to Europe, January 2020. Eurosurveillance: 25.

[26] DeSolla Price DJ (1965). Networks of scientific papers. Science 149: 510-515.

[27] Rackauckas C, Nie Q (2017) DifferentialEquations. jl ? a performant and feature-rich ecosystem for solving differential equations in Julia. Journal of Open Research Software May 25;5(1).

[28] Seebens H, Gastner MT, Blasius B (2013) The risk of marine bioinvasion caused by global shipping. Ecology Letters 6: 782-90.

[29] Simon HA, (1955). On a class of skew distribution functions. Biometrika 42: 425-440.

[30] Sornette D (2003) Critical Phenomena in Natural Sciences. Springer, Heidelberg, 2nd edition.

[31] Wang C, Liu L, Hao X, Guo H, Wang Q, Huang J, He N, Yu H, Lin X, Pan A, Wei S (2020) Evolving Epidemiology and Impact of Non-pharmaceutical Interventions on the Outbreak of Coronavirus Disease 2019 in Wuhan, China. medRxiv. 2020. Doi: 10.1101/2020.03.03.20030593

[32] Woolley-Meza O, Thiemann C, Grady D, Lee JJ, Seebens H, Blasius B, Brockmann D (2011) Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements. The European Physical Journal B 84: 589-600.

[33] World Health Organization, Coronavirus disease (COVID-2019) situation reports; https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/

[34] Van den Broeck W, Gioannini C, Goncalves B, Quaggiotto M, Colizza V, Vespignani A (2011) The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BMC Infectious Diseases 11: 37.

[35] Yule GU (1925). A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. Philos. Trans. R. Soc. London B 213: 21-87.

[36] Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley, Reading, MA.