Title

On the emergence of a power law in the distribution of COVID-19 cases.

Authors

Beare, Brendan K

Toda, Alexis Akira

Publication Date

2020-11-01

DOI

10.1016/j.physd.2020.132649

Peer reviewed

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Physica D 412 (2020) 132649

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

On the emergence of a power law in the distribution of COVID-19 cases

Brendan K. Beare [a], Alexis Akira Toda [b,*]

[a] School of Economics, University of Sydney, Sydney NSW 2006, Australia
[b] Department of Economics, University of California San Diego, La Jolla, CA 92093, USA

ARTICLE INFO

Article history:
Received 24 June 2020
Received in revised form 14 July 2020
Accepted 14 July 2020
Available online 16 July 2020
Communicated by V.M. Perez-Garcia

Keywords:
Coronavirus
COVID-19
Gibrat's law
Mathematical modeling of epidemics
Power law
Tauberian theorem

ABSTRACT

The first confirmed case of Coronavirus Disease 2019 (COVID-19) in the US was reported on January 21, 2020. By the end of March, 2020, there were more than 180,000 confirmed cases in the US, distributed across more than 2000 counties. We find that the right tail of this distribution exhibits a power law, with Pareto exponent close to one. We investigate whether a simple model of the growth of COVID-19 cases involving Gibrat's law can explain the emergence of this power law. The model is calibrated to match (i) the growth rates of confirmed cases, and (ii) the varying lengths of time during which COVID-19 had been present within each county. Thus calibrated, the model generates a power law with Pareto exponent nearly exactly equal to the exponent estimated directly from the distribution of confirmed cases across counties at the end of March.

1. Introduction

In the natural and social sciences, a variety of size distributions exhibit a power law in the right tail, meaning that the fraction of observations whose size S exceeds a threshold s decays like a power function as that threshold increases: $\mathrm{P}(S > s) \sim s^{-\zeta}$ for large s. The parameter $\zeta > 0$ is called the Pareto (or power law) exponent. Examples of distributions exhibiting power laws include income [1–4], wealth [5,6], consumption [7,8], city populations [9–12], firm size [13,14], family names [15–18], stock returns [19–21], and numerous others [22–25].

A common feature of models that purport to explain the prevalence of power laws is the presence of random multiplicative growth. Loosely, the size S_t of a quantity of interest at time t (e.g., wealth, population, firm size, etc.) is said to exhibit random multiplicative growth if its (random) growth factor $G_{t+1} = S_{t+1}/S_t$ is independent of the current size S_t. This is known among economists as Gibrat's law of proportional growth. On its own, Gibrat's law is not sufficient to generate a power law. For instance, a geometric Brownian motion (the continuous-time analogue of a random multiplicative growth process with lognormal growth) stopped at a fixed time has a lognormal distribution, which does not exhibit a power law. On the other hand, a geometric Brownian motion stopped at an exponentially distributed time has a double Pareto distribution, which exhibits a power law in both tails [26]. We may thus expect to observe a power law in the size distribution of a population whose members have been growing like geometric Brownian motions since birth, and whose distribution of ages is exponential. The combination of Gibrat's law with an exponential age distribution as a generative mechanism for power laws has been used extensively in recent economics literature [7,8,27–41]. Related techniques have also been employed in the physics literature [42–45].

In this paper we study the distribution of confirmed cases of Coronavirus Disease 2019 (COVID-19) across US counties. As we will see, by the end of March 2020, a power law had emerged in the right tail of that distribution. We investigate whether the combination of Gibrat's law (for growth in the number of cases within a county) and a suitable age distribution (for the length of time since the outbreak in each county) can explain this power law. Using daily county-level data from the onset of COVID-19 in the US in January 2020 until the end of March 2020, we estimate the distributions of growth rates and ages, employing a gamma parametrization of the former and a truncated logistic parametrization of the latter. Our primary finding is that the Pareto exponent implied by the estimated growth rate and age distributions, which is 0.936, nearly exactly matches the Pareto exponent estimated directly from the distribution of cases across counties at the end of March, which is 0.930. This indicates that the combination of Gibrat's law with a suitable age distribution can explain the power law observed in the right tail of the distribution of COVID-19 cases across US counties.

A nice aspect of the COVID-19 data we analyze is that they span the entire history of confirmed cases in the US population, thus permitting us to observe the distribution of ages (days since outbreak) across counties. While there is limited empirical

* Corresponding author.
E-mail addresses: brendan.beare@sydney.edu.au (B.K. Beare), atoda@ucsd.edu (A.A. Toda).

https://doi.org/10.1016/j.physd.2020.132649
0167-2789/

evidence that the age distributions of cities [46] and firms [47] may be approximately exponential, it is rarely the case that data used in economics and related fields allow a reliable estimate of the relevant age distribution. Indeed, it can be difficult to unambiguously define the age of a city, firm, or household, the latter of which is frequently interpreted as a dynastic unit spanning multiple generations. This confounds validation of the dynamic generative mechanism. Conveniently, our COVID-19 data reveal the entire shape of the age distribution, which allows us to provide what appears to be the first empirical analysis in which a Pareto exponent is obtained from direct estimates of both the growth rate and age distributions.

The remainder of our paper is organized as follows. Section 2 contains theoretical background material. In Section 2.1 we explain how the combination of Gibrat's law and an exponential age distribution determines a Pareto exponent. In Section 2.2 we discuss the connection between Gibrat's law and a simple model of epidemics. Section 3 contains our empirical findings. In Section 3.1 we describe our dataset. In Section 3.2 we display the distribution of COVID-19 cases across US counties at the end of March, and report a Pareto exponent estimated directly from this distribution. In Section 3.3 we assess the empirical plausibility of COVID-19 cases evolving according to Gibrat's law. In Sections 3.4 and 3.5 we report our estimates for the distributions of growth rates and ages respectively. In Section 3.6 we show how to compute the Pareto exponent implied by those growth rate and age distributions, and observe that it is close to the Pareto exponent reported in Section 3.2. Section 4 contains brief remarks in nontechnical language summarizing the practical import of our findings. Our data and replication files are available online.[1]

2. Theoretical background

2.1. Power laws via Gibrat's law and exponentially distributed age

Suppose that a unit (say, a county) starts with initial size (number of COVID-19 cases) $S_0 = 1$ at $t = 0$ and grows randomly according to Gibrat's law of motion $S_t = G_t S_{t-1}$ for integer-valued $t \geq 1$, where $\{G_t\}_{t=1}^{\infty}$ is a sequence of independent and identically distributed (i.i.d.) positive random variables. Let T be a random integer-valued time (days since COVID-19 outbreak in the county), independent of the sequence $\{G_t\}_{t=1}^{\infty}$, at which the unit size S_T is observed. Suppose for now that T has the geometric distribution (i.e., the discrete-time analogue of the exponential distribution), meaning that for $n \geq 1$ we have $\mathrm{P}(T = n) = p(1-p)^{n-1}$ for some parameter $p \in (0, 1)$ called a success probability.

What can be said about the right tail of the distribution of S_T? It turns out that, under a regularity condition on the distribution of the growth rate $X_t = \ln G_t$, the tail exhibits a power law. Specifically, letting E denote the expected value operator, we shall assume that the distribution of X_t has Laplace transform $M(z) = \mathrm{E}(e^{zX_t})$ finite for real $z \in [0, \eta)$ and diverging to infinity as z increases to η, where η may be any positive real number or $+\infty$. Loosely, this means that X_t can take positive values and has a distribution with a right tail that decays to zero exponentially or faster. When this regularity condition is satisfied, we may argue as follows to establish that the right tail of the distribution of S_T exhibits a power law. Let $Y = \ln S_T = \sum_{t=1}^{T} X_t$. Observe that the distribution of Y has Laplace transform M_Y satisfying

$$M_Y(z) = \sum_{n=1}^{\infty} p(1-p)^{n-1} M(z)^n = \frac{pM(z)}{1-(1-p)M(z)} \tag{1}$$

for all positive real z such that $(1-p)M(z) < 1$. Noting that $M(z)$ is convex as a function of $z \in (0, \eta)$ and satisfies $M(0) < 1/(1-p) < M(\eta)$, we deduce that there is a unique $\zeta \in (0, \eta)$ at which

$$(1-p)M(\zeta) = 1, \tag{2}$$

and that $M(z)$ has strictly positive derivative at $z = \zeta$. It thus follows from Eq. (1) and an application of l'Hôpital's rule that

$$\lim_{z \to \zeta} (z - \zeta) M_Y(z) = -\frac{p}{(1-p)^2 M'(\zeta)} < 0,$$

implying that ζ is a simple pole of M_Y.

The fact that ζ is a positive real pole of M_Y means that the right tail of the distribution of Y decays to zero exponentially at rate ζ, in the sense that $\ln \mathrm{P}(Y > y) \sim -\zeta y$ for large y. This is a consequence of a Tauberian theorem proved in Ref. [48]. It follows that the right tail of the distribution of S_T exhibits a power law with Pareto exponent ζ: setting $y = \ln s$, we have

$$\lim_{s \to \infty} \frac{\ln \mathrm{P}(S_T > s)}{\ln s} = \lim_{y \to \infty} \frac{\ln \mathrm{P}(Y > y)}{y} = -\zeta.$$

Eq. (2) appears as Eq. (10) in Ref. [42]. It shows how the Pareto exponent ζ is uniquely determined by the interaction of the growth rate distribution (through its Laplace transform M) and the age distribution (through its parameter p). A more general version of Eq. (2) applicable in settings where the growth rates may not be i.i.d. but instead satisfy a weaker condition involving Markov modulation has been established in Ref. [49].

We assumed in the preceding discussion that T has the geometric distribution. This assumption was stronger than necessary; what matters is that the right tail of the distribution of T decays at an exponential rate. In our empirical analysis in Section 3 we employ a truncated logistic parametrization of the distribution of T, which has an exponentially decaying right tail but in other respects does not resemble the geometric distribution. We will see in Section 3.6 that Eq. (2) remains valid in this case, with p determined by the rate at which the right tail of the truncated logistic distribution decays exponentially to zero. We may also allow the initial size S_0 to be a positive random variable independent of $\{G_t\}_{t=1}^{\infty}$ and T, provided that it satisfies $\mathrm{E}S_0^{\zeta+\epsilon} < \infty$ for some $\epsilon > 0$; this can be shown using Breiman's lemma [50].

2.2. The Susceptible–Infected–Recovered model

Here we provide a brief discussion of the Susceptible–Infected–Recovered (SIR) model of epidemics [51] and the extent to which it is consistent with Gibrat's law. In the SIR model, a community (say, a county) consists of a mass of individuals who are either susceptible to an infectious disease (they are neither infected nor have immunity), infected, or immune (possibly because they are vaccinated, infected and recovered, or dead). Individuals meet each other randomly, and conditional on an infected individual meeting a susceptible individual, the disease is transmitted with some probability. Let $\beta > 0$ be the rate at which an infected individual meets a person and transmits the disease if susceptible. Let $\gamma > 0$ be the rate at which an infected individual recovers or dies. Letting x, y, z be the fractions of susceptible, infected, and recovered individuals in the community (so $x + y + z = 1$), the SIR model is described by the system of differential equations

$$\dot{x} = -\beta x y, \tag{3a}$$

$$\dot{y} = \beta x y - \gamma y, \tag{3b}$$

$$\dot{z} = \gamma y, \tag{3c}$$

where $\dot{x}, \dot{y}, \dot{z}$ are the rates of change of x, y, z.

[1] https://github.com/alexisakira/COVID-19_power_law.

Although the system of differential equations (3) is nonlinear, it admits an exact analytical solution [52]. It suffices for our purposes to study the system (3) heuristically at the beginning of the epidemic, where $x \approx 1$ and y, z are small. Setting $x = 1$, Eq. (3b) becomes $\dot{y} = (\beta - \gamma)y$, and hence $y(t) = y_0 e^{(\beta-\gamma)t}$. Integrating Eq. (3c), we obtain

$$z(t) = z_0 + \int_0^t \gamma y(s) ds = z_0 + y_0 \frac{\gamma}{\beta - \gamma}(e^{(\beta-\gamma)t} - 1).$$

The cumulative number of cases up to time t is therefore given by

$$c(t) := y(t) + z(t) = z_0 + \frac{y_0}{\beta - \gamma}(\beta e^{(\beta-\gamma)t} - \gamma).$$

Assuming that $\beta > \gamma$ (so that there is an epidemic), that time t is neither too large (so that the approximation $x \approx 1$ is valid) nor too small (so that $\gamma \ll \beta e^{(\beta-\gamma)t}$), and that z_0 is small relative to y_0, it follows that

$$c(t) \approx y_0 \frac{\beta}{\beta - \gamma} e^{(\beta-\gamma)t},$$

so cases grow exponentially at rate $g := \beta - \gamma > 0$. This implies that the growth factor for cases between day t and $t + 1$,

$$G_{t+1} := c(t + 1)/c(t) \approx e^g, \tag{4}$$

is independent of the current size $c(t)$. In practice, the transmission rate β may fluctuate over time, so it may be plausible to assume that the growth factor G_{t+1} is a random variable independent of the current size $c(t)$, as in Gibrat's law.

The point of the preceding heuristic discussion is that, in the SIR model, the growth of infections may be broadly consistent with Gibrat's law in the early stages of an epidemic. The same cannot be said for the later stages of an epidemic. In the exact analytical solution to the system (3) given in Ref. [52], the growth rate of infections falls as the proportion of the population that is infected or recovered rises, because a smaller proportion of the population remains susceptible to infection. Furthermore, the growth rate of infections may fall as individuals take more precautionary measures such as avoiding crowded spaces, washing hands, wearing face coverings, etc. Our empirical findings reported in Section 3 are based on data from the onset of COVID-19 in the US in January 2020 until the end of March 2020. At the end of that period the total number of confirmed COVID-19 cases in the US (182,308) was less than 0.06% of the total population (330 million). We provide evidence in Section 3.3 that the growth rate of cases remained independent of the number of cases up until the end of March 2020, consistent with Gibrat's law.

3. Empirical findings

3.1. Dataset

Our dataset consists of the daily numbers of confirmed COVID-19 cases in US counties reported by *The New York Times*,[2] based on reports from state and local health agencies. These numbers are cumulative. We use data from January 21, 2020, when the first case in the US was reported, through to March 31, 2020. There are a total of 3243 counties in the US (including both states and territories). We include the 2121 counties that reported at least one COVID-19 case by March 31 and exclude the remainder. Exceptionally, the dataset combines the five boroughs of New York City (New York, Kings, Queens, Bronx and Richmond counties) into a single unit called New York City.

[2] https://github.com/nytimes/covid-19-data.

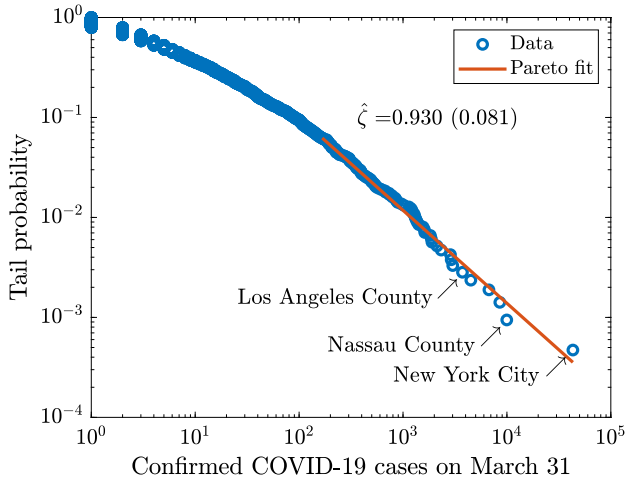


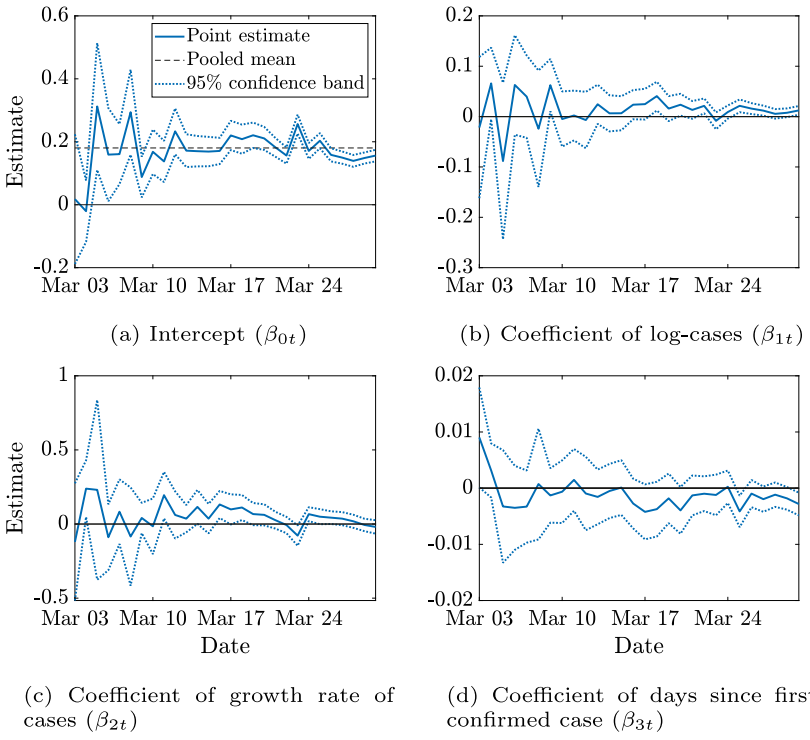
Fig. 1. Log–log plot of confirmed COVID-19 cases against tail probabilities across US counties on March 31, 2020. The tail probability of a county is the proportion of all counties matching or exceeding its number of COVID-19 cases. The Pareto fit was obtained by applying the Hill estimator to the top 6.2% of counties by number of cases. The estimated Pareto exponent is $\hat{\zeta} = 0.930$, with a standard error of 0.081.

3.2. Distribution of COVID-19 cases on March 31

In Fig. 1 we plot the number of confirmed COVID-19 cases in each county on March 31 against the corresponding tail probabilities in log–log scale. The tail probability for county i is defined to be the fraction of counties whose number of COVID-19 cases is greater than or equal to that of county i. The county with the smallest tail probability, and therefore the highest number of COVID-19 cases, is New York City (though it is in fact an aggregate of five counties, as pointed out in Section 3.1). The county with the next highest number of COVID-19 cases is Nassau County, which is located in Long Island and borders Queens County in New York City. The county with the highest population is Los Angeles County, which has the sixth highest number of COVID-19 cases.

Setting aside the two data points for New York City and neighboring Nassau County, the data toward the lower-right of Fig. 1 appear to lie roughly on a straight line. This pattern is indicative of a power law in the right tail of the distribution of COVID-19 cases across US counties. To investigate further, we followed the approach recommended in Ref. [53]. Specifically, we used a version of the Hill estimator [54] to fit a Pareto distribution to the data exceeding a threshold selected using the algorithm described in Ref. [55]. Slightly more than 6% of counties had COVID-19 cases exceeding the selected threshold. The Hill estimate of the Pareto exponent was $\hat{\zeta} = 0.930$, with a standard error of 0.081. (Similar results were obtained using a 5% or 10% threshold.) The fitted Pareto tail is displayed in Fig. 1, where in log–log scale it appears as a straight line with slope $-\hat{\zeta}$.

The two data points for New York City and Nassau County, which recorded the highest numbers of COVID-19 cases, lie somewhat to the right and to the left of our estimated power law in Fig. 1. In its notes on methodology and definitions, *The New York Times* (Footnote 3.1) states that where possible it assigned cases to the county where they were treated, not where they resided. This could mean that a significant number of Nassau County residents who were confirmed as having COVID-19 but received treatment in New York City are classified as New York City cases rather than Nassau County cases. If we suppose that one third of Nassau County residents confirmed with COVID-19 were classified as New York City cases, and reassign those cases to



(a) Intercept (β_{0t}) (b) Coefficient of log-cases (β_{1t})

(c) Coefficient of growth rate of cases (β_{2t}) (d) Coefficient of days since first confirmed case (β_{3t})

Fig. 2. Ordinary least squares estimates of $\beta_{0t}, \beta_{1t}, \beta_{2t}, \beta_{3t}$ in Eq. (5) for the 28 days between March 3 and March 30 inclusive, with 95% confidence bands. In panel (2(a)) we also display the pooled mean growth rate of 0.180.

Nassau County, then the small circles representing Nassau County and New York City in Fig. 1 shift to the right and left respectively, such that both are touching our line of Pareto fit.

3.3. Empirical plausibility of Gibrat's law

The remainder of our empirical analysis focuses on determining whether the power law with estimated exponent $\hat{\zeta} = 0.930$ obtained in Section 3.2 can be explained by a combination of Gibrat's law and an age distribution with exponential right tail, as described in Section 2.1. We first assess the empirical plausibility of Gibrat's law as a description of the growth in COVID-19 cases within counties. To this end, for each day t between March 3 and March 30 inclusive, we estimate the cross-sectional regression equation

$$\Delta \ln c_{i,t+1} = \beta_{0t} + \beta_{1t} \ln c_{it} + \beta_{2t} \Delta \ln c_{it} + \beta_{3t} D_{it} + \varepsilon_{it} \tag{5}$$

by ordinary least squares, where c_{it} is the number of cases in county i up to day t, $\Delta \ln c_{i,t+1}$ is the growth rate in cases in county i between day t and $t+1$, D_{it} is the number of days (inclusive) between day t and the day of the first confirmed case in county i, ε_{it} is the regression residual, and $\beta_{0t}, \beta_{1t}, \beta_{2t}, \beta_{3t}$ are regression coefficients that are potentially time-varying. (Here β_{0t} corresponds to the growth rate $g = \beta - \gamma$ in Eq. (4).) The estimation of Eq. (5) on day t uses the data for all counties i reporting a positive number of cases ($c_{it} > 0$). We estimate Eq. (5) for the 28 days between March 3 and March 30 inclusive because these are the days on which at least 30 counties reported a positive number of cases. The number of counties used in each regression increases from 32 on March 3 to 1940 on March 30.

Fig. 2 displays our estimates of $\beta_{0t}, \beta_{1t}, \beta_{2t}, \beta_{3t}$ in Eq. (5) from March 3 to March 30, with accompanying 95% confidence bands. In each panel, the confidence bands narrow as we move from left to right, reflecting the fact that the regression sample size increases from 32 to 1940. In panels 2(b)–2(d), we see that the estimates of $\beta_{1t}, \beta_{2t}, \beta_{3t}$ are close to zero. This is exactly what we would expect under Gibrat's law: the growth rate in cases between days t and $t+1$ ought to be unrelated to the number of cases on day t, the growth rate in cases between days $t-1$ and t, and the time elapsed since the outbreak.

The estimated parameters are quite stable over time. The estimates of β_{0t} displayed in panel 2(a) indicate that the expected growth rate of cases during March was roughly stable at around 15%–20% per day. The pooled mean growth rate (i.e., the average over all days and counties with at least one reported case) was 18%, which falls outside the daily 95% confidence bands on only three days excluding the very end of March, when mitigation efforts may have begun to slow the epidemic. The pooled mean growth rate may be compared to estimates of related parameters obtained in prior research on COVID-19. In the context of the SIR model described in Section 2.2, the recovery rate γ is a biological parameter determined by the virus. In Ref. [56] the mean serial interval, which corresponds to $1/\gamma$, is estimated through contact tracing to be 7.5 days. Therefore setting $\gamma = 1/7.5 = 0.133$ and $g = \beta - \gamma = 0.180$, our pooled mean growth rate, we estimate the transmission rate β to be 0.314. This in turn implies that the reproductive number of COVID-19 (which plays an important role in epidemic dynamics) is $R_0 = \beta/\gamma = 2.35$, which is close to the estimate of 2.2 reported in Ref. [56].

3.4. Distribution of growth rate of confirmed cases

In Fig. 3 we display a histogram of the growth rates of confirmed COVID-19 cases, obtained by pooling our data across all days t up to March 30 and all counties i with at least 10 confirmed cases ($c_{it} \geq 10$) and a positive growth rate ($\Delta \ln c_{i,t+1} > 0$). Overlaying the histogram we plot a gamma distribution fit to our data by the method of maximum likelihood. The gamma distribution has density

$$f(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0, \tag{6}$$

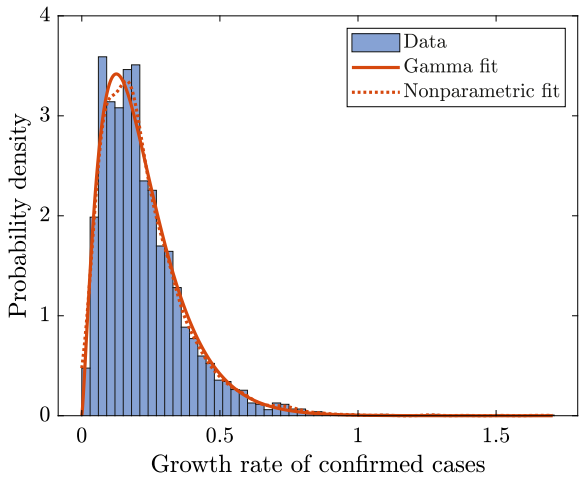


Fig. 3. Histogram of growth rates of confirmed COVID-19 cases, using data from all days up to the end of March and all counties with at least 10 confirmed cases and a positive growth rate of cases. The gamma fit was obtained by the method of maximum likelihood. The nonparametric fit was obtained by Gaussian kernel smoothing.

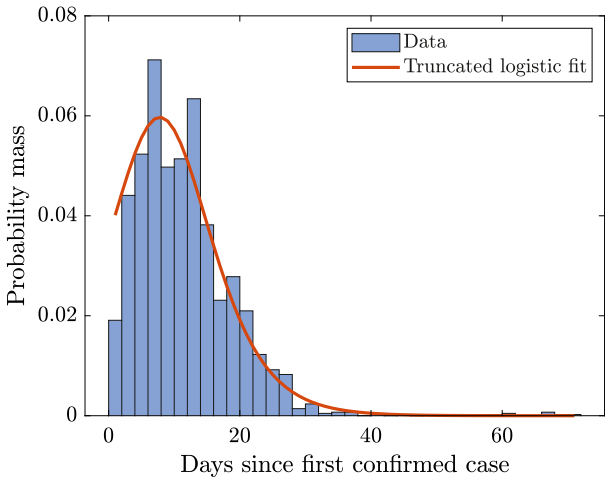


Fig. 4. Histogram of days-since-outbreak on March 31, using data from all counties reporting at least one confirmed COVID-19 case by March 31. The truncated logistic fit was obtained by the method of maximum likelihood.

where $\alpha, \lambda > 0$ are called the shape and rate parameters. The maximum likelihood parameter estimates are $\hat{\alpha} = 2.30$ and $\hat{\lambda} = 10.4$. The fit of the gamma distribution appears to be excellent, particularly in the right tail of the data. It closely matches a nonparametric estimate obtained using Gaussian kernel smoothing, which we plot alongside it in Fig. 3.

Our data up to March 30 include a total of 5687 day-county pairs with at least 10 confirmed cases ($c_{it} \geq 10$). Of those, 725 day-county pairs had a zero growth rate of cases ($\Delta \ln c_{i,t+1} = 0$), and were therefore excluded from the computation of the histogram in Fig. 3 and corresponding gamma fit. An observed growth rate could be zero because there were indeed no new cases, or because the data were not updated on a particular day. The proportion of growth rates observed to be zero is $\pi = 725/5687 = 0.128$, which is substantial. For this reason, in the analysis in Section 3.6, we model the distribution of the growth rate of confirmed cases as a mixture of our maximum likelihood estimate of the gamma distribution plotted in Fig. 3 and a point mass at zero, with proportions $1 - \pi$ and π respectively.

3.5. Distribution of days since first confirmed case

In Fig. 4 we display a histogram of the number of days (inclusive) between the day of the first confirmed case of COVID-19 in a county, and March 31. The histogram is computed from the 2121 counties in our dataset that reported at least one confirmed case by March 31.

As discussed earlier, the combination of Gibrat's law with an exponential age distribution has been widely used as a generative mechanism for power laws. It is apparent from the histogram in Fig. 4, however, that the exponential distribution (or its discrete counterpart, the geometric distribution) cannot provide an acceptable approximation to the age distribution we observe in our data. The problem is that the density of the exponential distribution is monotonically decreasing, whereas the histogram in Fig. 4 is roughly hump shaped.

In nature, an exponential distribution of ages arises when a population grows exponentially over time, as in the Yule model of speciation discussed in Refs. [10,26]. By analogy, we may expect to see an exponential distribution of days-since-outbreak in our data if the number of counties that have reported at least one confirmed COVID-19 case is growing exponentially over time. The analogy fails because there are only 3243 counties in the US, so that exponential growth cannot be maintained. Once COVID-19 has spread to a substantial proportion of counties, the rate of growth in the number of counties reporting at least one case ought to fall, eventually vanishing as saturation is approached. Given that nearly two thirds of US counties had reported at least one confirmed case by March 31, we would expect the number of newly infected counties to be declining by this time. This qualitative argument could, in principle, explain the hump shape in the histogram in Fig. 4. We now model this argument and show that it does explain the hump shape quantitatively.

The logistic function was introduced in the 19th century as a model of population growth that commences at an exponential rate but tapers off as a saturation point is approached due to competition for resources [57]. In the SIR model discussed in Section 2.2, in the absence of recovery (so that $\gamma = 0$ in Eq. (3b)), a logistic function describes the growth of the infected population over time, and the distribution of time-since-infection for the infected population at any given time is the truncated logistic distribution. (Truncation is always necessary because saturation is not reached in finite time.) By analogy, when considering the spread of an infection across a population of counties, we might expect the truncated logistic distribution to well-approximate the distribution of days-since-outbreak across counties at any given time. We therefore truncate a discrete version of the logistic distribution introduced in Ref. [58]. Without truncation, for any integer n, the discrete logistic distribution has probability mass

$$\mathrm{P}(T = n) = \frac{(1-q)q^{n-\mu}}{(1+q^{n-\mu})(1+q^{n-\mu+1})}, \tag{7}$$

where $q \in (0, 1)$ is a parameter determining the rate of exponential decay in the tails, and μ is a location parameter. After truncating all mass on nonpositive integers and rescaling so that the remaining mass sums to one, the probability mass assigned to each integer $n \geq 1$ is

$$\mathrm{P}(T = n) = \frac{(1+\phi)(1-q)q^{n-1}}{(1+\phi q^{n-1})(1+\phi q^{n})}, \tag{8}$$

where we have reparametrized the distribution in terms of q and $\phi = q^{1-\mu}$, the latter equal to the ratio of probability masses included and excluded by truncation.

Overlaying the histogram in Fig. 4 we plot a truncated logistic distribution fit to our data by the method of maximum likelihood.

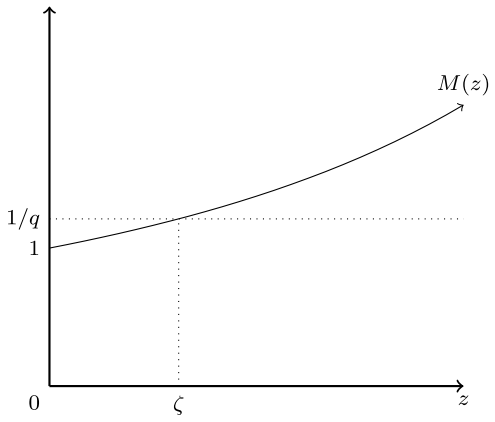


Fig. 5. The Pareto exponent ζ is the unique positive real z at which the Laplace transform $M(z)$ is equal to $1/q$.

The maximum likelihood parameter estimates are $\hat{q} = 0.825$ and $\hat{\phi} = 4.06$. The fit captures the general shape of the histogram reasonably well, particularly toward the right tail, which is the more important region for our purposes.

3.6. Implied Pareto exponent

In Section 2.1 we discussed how the combination of Gibrat's law and an exponential age distribution can generate a power law, with Pareto exponent ζ solving Eq. (2). It remains for us to determine the Pareto exponent thus obtained when the distributions of growth rates and ages are as estimated in Sections 3.4 and 3.5. A complicating factor is that our age distribution is not exactly exponential, but instead belongs to the family of truncated logistic distributions defined by Eq. (8). This leads us to replace Eq. (1) with

$$M_Y(z) = \sum_{n=1}^{\infty} \frac{(1+\phi)(1-q)q^{n-1}}{(1+\phi q^{n-1})(1+\phi q^n)} M(z)^n, \tag{9}$$

valid for all positive real z such that $qM(z) < 1$. Define

$$r_n = [(1+\phi q^{n-1})(1+\phi q^n)]^{-1} - 1,$$

and let $p = 1 - q$. It is straightforward to show that $|r_n| \leq \phi(1+q)q^{n-1}$, so we may rewrite Eq. (9) as

$$M_Y(z) = (1+\phi)p \left[\frac{M(z)}{1-qM(z)} + \sum_{n=1}^{\infty} q^{n-1} r_n M(z)^n \right],$$

where the first term in square brackets has a pole at the unique $\zeta \in (0, \eta)$ solving Eq. (2) and the second term in square brackets is analytic in a neighborhood of ζ. This shows that ζ is a positive real pole of M_Y and so, as in Section 2.1, we deduce from the Tauberian theorem proved in Ref. [48] that the right tail of the distribution of S_T exhibits a power law with Pareto exponent ζ. Fig. 5 displays visually how ζ is determined by the parameter q and Laplace transform $M(z)$, with our empirical estimates for q and $M(z)$.

Our estimate of the distribution of the growth rate of confirmed COVID-19 cases (i.e., the distribution of X_t) reported in Section 3.4 was a mixture of a point mass at zero and a gamma distribution with weights $\pi = 0.128$ and $1 - \pi$ respectively. The particular form of this distribution allows us to obtain the solution ζ to Eq. (2) in analytic form. Specifically, the Laplace transform of the distribution of X_t is given for real $z < \lambda$ by

$$M(z) = \pi + (1-\pi) \int_0^{\infty} e^{zx} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx$$

$$= \pi + (1-\pi)(1 - z/\lambda)^{-\alpha}, \tag{10}$$

and so the unique solution to Eq. (2) is

$$\zeta = \lambda \left[1 - \left(\frac{1-\pi}{1/q - \pi} \right)^{1/\alpha} \right]. \tag{11}$$

Substituting the empirical estimates $\pi = 0.128$, $\hat{\alpha} = 2.30$, $\hat{\lambda} = 10.4$ and $\hat{q} = 0.825$ into Eq. (11), we obtain the implied Pareto exponent $\zeta = 0.936$, which is nearly exactly equal to (and well within the 95% confidence interval of) the estimate $\hat{\zeta} = 0.930$ reported in Section 3.2. (If we use the nonparametric distribution in Fig. 3 in place of the gamma distribution then the implied Pareto exponent is $\zeta = 0.928$.) Thus our simple model involving Gibrat's law generates precisely the power law we observe in the distribution of COVID-19 cases at the end of March.

4. Final remarks

We conclude with some brief remarks in nontechnical language to summarize the primary import of our results to policymakers dealing with the COVID-19 epidemic, or to historians seeking to understand the early weeks of the COVID-19 epidemic in the US. An empirical feature of the distribution of COVID-19 cases across US counties at the end of March 2020 is that case loads are dramatically higher in some counties than in others. That is, the distribution of COVID-19 cases across US counties at the end of March has a heavy right tail. While it may be natural to look for county-specific characteristics to explain why this is the case, our results indicate that this is not necessary. The very high case loads observed in some counties are accurately predicted by a simple empirically calibrated model combining (i) random multiplicative growth within each county, and (ii) variation across counties in the duration of the spread of COVID-19. There is no need to attribute the highest case loads to other idiosyncratic factors. New York City, where confirmed cases substantially exceed our model's prediction, may be an exception.

CRediT authorship contribution statement

Brendan K. Beare: Methodology, Formal Analysis, Writing, Visualization, Project administration. **Alexis Akira Toda:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Patrick Huggins for helpful comments.

References

[1] W.J. Reed, The Pareto law of incomes—an explanation and an extension, Physica A 319 (1) (2003) 469–486, http://dx.doi.org/10.1016/S0378-4371(02)01507-8.

[2] A.A. Toda, Income dynamics with a stationary double Pareto distribution, Phys. Rev. E 83 (4) (2011) 046122, http://dx.doi.org/10.1103/PhysRevE.83.046122.

[3] A.A. Toda, The double power law in income distribution: Explanations and evidence, J. Econ. Behav. Organ. 84 (1) (2012) 364–381, http://dx.doi.org/10.1016/j.jebo.2012.04.012.

[4] M. Ibragimov, R. Ibragimov, Heavy tails and upper-tail inequality: The case of Russia, Empir. Econ. 54 (2) (2018) 823–837, http://dx.doi.org/10.1007/s00181-017-1239-0.

[5] O.S. Klass, O. Biham, M. Levy, O. Malcai, S. Solomon, The Forbes 400 and the Pareto wealth distribution, Econom. Lett. 90 (2) (2006) 290–295, http://dx.doi.org/10.1016/j.econlet.2005.08.020.

[6] P. Vermeulen, How fat is the top tail of the wealth distribution?, Rev. Income Wealth 64 (2) (2018) 357–387, http://dx.doi.org/10.1111/roiw.12279.

[7] A.A. Toda, K. Walsh, The double power law in consumption and implications for testing Euler equations, J. Political Econ. 123 (5) (2015) 1177–1200, http://dx.doi.org/10.1086/682729.

[8] A.A. Toda, A note on the size distribution of consumption: More double Pareto than lognormal, Macroecon. Dyn. 21 (6) (2017) 1508–1518, http://dx.doi.org/10.1017/S1365100515000942.

[9] X. Gabaix, Zipf's law for cities: An explanation, Q. J. Econ. 114 (3) (1999) 739–767, http://dx.doi.org/10.1162/003355399556133.

[10] W.J. Reed, On the rank–size distribution for human settlements, J. Reg. Sci. 42 (1) (2002) 1–17, http://dx.doi.org/10.1111/1467-9787.00247.

[11] K.T. Soo, Zipf's law for cities: A cross-country investigation, Reg. Sci. Urban Econ. 35 (3) (2005) 239–263, http://dx.doi.org/10.1016/j.regsciurbeco.2004.04.004.

[12] K. Giesen, A. Zimmermann, J. Suedekum, The size distribution across all cities—double Pareto lognormal strikes, J. Urban Econ. 68 (2) (2010) 129–137, http://dx.doi.org/10.1016/j.jue.2010.03.007.

[13] R.L. Axtell, Zipf distribution of U.S. firm sizes, Science 293 (5536) (2001) 1818–1820, http://dx.doi.org/10.1126/science.1062081.

[14] Y. Fujiwara, H. Aoyama, C. Di Guilmi, W. Souma, M. Gallegati, Gibrat and Pareto-Zipf revisited with European firms, Physica A 344 (1–2) (2004) 112–116, http://dx.doi.org/10.1016/j.physa.2004.06.098.

[15] S. Miyazima, Y. Lee, T. Nagamine, H. Miyajima, Power-law distribution of family names in Japanese societies, Physica A 278 (1–2) (2000) 282–288, http://dx.doi.org/10.1016/S0378-4371(99)00546-4.

[16] D.H. Zanette, S.C. Manrubia, Vertical transmission of culture and the distribution of family names, Physica A 295 (1–2) (2001) 1–8, http://dx.doi.org/10.1016/s0378-4371(01)00046-2.

[17] W.J. Reed, B.D. Hughes, From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature, Phys. Rev. E 66 (6) (2002) 067103, http://dx.doi.org/10.1103/PhysRevE.66.067103.

[18] W.J. Reed, B.D. Hughes, On the distribution of family names, Physica A 319 (2003) 579–590, http://dx.doi.org/10.1016/S0378-4371(02)01455-3.

[19] X. Gabaix, P. Gopikrishnan, V. Plerou, H.E. Stanley, A theory of power-law distributions in financial market fluctuations, Nature 423 (6937) (2003) 267–270, http://dx.doi.org/10.1038/nature01624.

[20] X. Gabaix, P. Gopikrishnan, V. Plerou, H.E. Stanley, Institutional investors and stock market volatility, Q. J. Econ. 121 (2) (2006) 461–504, http://dx.doi.org/10.1162/qjec.2006.121.2.461.

[21] Z. Gu, R. Ibragimov, The cubic law of the stock returns in emerging markets, J. Empir. Financ. 46 (2018) 182–190, http://dx.doi.org/10.1016/j.jempfin.2017.11.008.

[22] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, Internet Math. 1 (2) (2004) 226–252, http://dx.doi.org/10.1080/15427951.2004.10129088.

[23] M.E.J. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. 46 (5) (2005) 323–351, http://dx.doi.org/10.1080/00107510500052444.

[24] X. Gabaix, Power laws in economics and finance, Annu. Rev. Econ. 1 (2009) 255–293, http://dx.doi.org/10.1146/annurev.economics.050708.142940.

[25] I. Eliazar, Power Laws: A Statistical Trek, Springer, 2020, http://dx.doi.org/10.1007/978-3-030-33235-8.

[26] W.J. Reed, The Pareto, Zipf and other power laws, Econom. Lett. 74 (1) (2001) 15–19, http://dx.doi.org/10.1016/S0165-1765(01)00524-9.

[27] E.G.J. Luttmer, Selection, growth, and the size distribution of firms, Q. J. Econ. 122 (3) (2007) 1103–1144, http://dx.doi.org/10.1162/qjec.122.3.1103.

[28] M. Nirei, W. Souma, A two factor model of income distribution dynamics, Rev. Income Wealth 53 (3) (2007) 440–459, http://dx.doi.org/10.1111/j.1475-4991.2007.00242.x.

[29] J. Benhabib, A. Bisin, S. Zhu, The distribution of wealth and fiscal policy in economies with finitely lived agents, Econometrica 79 (1) (2011) 123–157, http://dx.doi.org/10.3982/ECTA8416.

[30] A.A. Toda, Incomplete market dynamics and cross-sectional distributions, J. Econom. Theory 154 (2014) 310–348, http://dx.doi.org/10.1016/j.jet.2014.09.015.

[31] D. Acemoglu, D. Cao, Innovation by entrants and incumbents, J. Econom. Theory 157 (2015) 255–294, http://dx.doi.org/10.1016/j.jet.2015.01.001.

[32] C. Arkolakis, A unified theory of firm selection and growth, Q. J. Econ. 131 (1) (2016) 89–155, http://dx.doi.org/10.1093/qje/qjv039.

[33] J. Benhabib, A. Bisin, S. Zhu, The distribution of wealth in the Blanchard-Yaari model, Macroecon. Dyn. 20 (2016) 466–481, http://dx.doi.org/10.1017/S1365100514000066.

[34] X. Gabaix, J.-M. Lasry, P.-L. Lions, B. Moll, The dynamics of inequality, Econometrica 84 (6) (2016) 2071–2111, http://dx.doi.org/10.3982/ECTA13569.

[35] S. Aoki, M. Nirei, Zipf's law, Pareto's law, and the evolution of top incomes in the United States, Am. Econ. J.: Macroecon. 9 (3) (2017) 36–71, http://dx.doi.org/10.1257/mac.20150051.

[36] A.A. Toda, K.J. Walsh, Fat tails and spurious estimation of consumption-based asset pricing models, J. Appl. Econometrics 32 (6) (2017) 1156–1177, http://dx.doi.org/10.1002/jae.2564.

[37] D. Cao, W. Luo, Persistent heterogeneous returns and top end wealth inequality, Rev. Econ. Dyn. 26 (2017) 301–326, http://dx.doi.org/10.1016/j.red.2017.10.001.

[38] T. Mukoyama, S. Osotimehin, Barriers to reallocation and economic growth: The effects of firing costs, Am. Econ. J.: Macroecon. 11 (4) (2019) 235–270, http://dx.doi.org/10.1257/mac.20170170.

[39] A.A. Toda, Wealth distribution with random discount factors, J. Monetary Econ. 104 (2019) 101–113, http://dx.doi.org/10.1016/j.jmoneco.2018.09.006.

[40] J. Stachurski, A.A. Toda, An impossibility theorem for wealth in heterogeneous-agent models with limited heterogeneity, J. Econom. Theory 182 (2019) 1–24, http://dx.doi.org/10.1016/j.jet.2019.04.001.

[41] Q. Ma, J. Stachurski, A.A. Toda, The income fluctuation problem and the evolution of wealth, J. Econom. Theory 187 (2020) 105003, http://dx.doi.org/10.1016/j.jet.2020.105003.

[42] S.C. Manrubia, D.H. Zanette, Stochastic multiplicative processes with reset events, Phys. Rev. E 59 (5) (1999) 4945–4948, http://dx.doi.org/10.1103/PhysRevE.59.4945.

[43] M. Montero, J. Villarroel, Monotonic continuous-time random walks with drift and stochastic reset events, Phys. Rev. E 87 (1) (2013) 012116, http://dx.doi.org/10.1103/PhysRevE.87.012116.

[44] M.R. Evans, S.N. Majumdar, Diffusion with resetting in arbitrary spatial dimension, J. Phys. A 47 (28) (2014) 285001, http://dx.doi.org/10.1088/1751-8113/47/28/285001.

[45] M. Montero, J. Villarroel, Directed random walk with random restarts: The Sisyphus random walk, Phys. Rev. E 94 (3) (2016) 032132, http://dx.doi.org/10.1103/PhysRevE.94.032132.

[46] K. Giesen, J. Suedekum, City age and city size, Eur. Econ. Rev. 71 (2014) 193–208, http://dx.doi.org/10.1016/j.euroecorev.2014.07.006.

[47] A. Coad, The exponential age distribution and the Pareto firm size distribution, J. Ind. Compet. Trade 10 (3–4) (2010) 389–395, http://dx.doi.org/10.1007/s10842-010-0071-4.

[48] K. Nakagawa, Application of Tauberian theorem to the exponential decay of the tail probability of a random variable, IEEE Trans. Inform. Theory 53 (9) (2007) 3239–3249, http://dx.doi.org/10.1109/TIT.2007.903114.

[49] B.K. Beare, A.A. Toda, Geometrically stopped Markovian random growth processes and Pareto tails. URL https://arxiv.org/abs/1712.01431.

[50] L. Breiman, On some limit theorems similar to the arc-sin law, Theory Probab. Appl. 10 (2) (1965) 323–331, http://dx.doi.org/10.1137/1110037.

[51] W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 115 (772) (1927) 700–721, http://dx.doi.org/10.1098/rspa.1927.0118.

[52] T. Harko, F.S.N. Lobo, M.K. Mak, Exact analytical solutions of the susceptible-infected-recovered (SIR) epidemic model and of the SIR model with equal death and birth rates, Appl. Math. Comput. 236 (2014) 184–194, http://dx.doi.org/10.1016/j.amc.2014.03.030.

[53] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703, http://dx.doi.org/10.1137/070710111.

[54] B.M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Statist. 3 (5) (1975) 1163–1174, http://dx.doi.org/10.1214/aos/1176343247.

[55] A. Clauset, M. Young, K.S. Gleditsch, On the frequency of severe terrorist events, J. Confl. Resolut. 51 (1) (2007) 58–87, http://dx.doi.org/10.1177/0022002706296157.

[56] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K.S. Leung, E.H. Lau, J.Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T.T. Lam, J.T. Wu, G.F. Gao, B.J. Cowling, B. Yang, G.M. Leung, Z. Feng, Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia, New England J. Med. 382 (2020) 1199–1207, http://dx.doi.org/10.1056/NEJMoa2001316.

[57] J.S. Cramer, Logit Models from Economics and Other Fields, Cambridge University Press, 2003.

[58] S. Chakraborty, D. Chakravarty, A new discrete probability distribution with integer support on $(-\infty, \infty)$, Comm. Statist. Theory Methods 45 (2) (2016) 492–505, http://dx.doi.org/10.1080/03610926.2013.830743.