Modelling the epidemic growth of preprints on COVID-19 and SARS-CoV-2

Giovani L. Vasconcelos [1,*], **Luan P. Cordeiro** [1], **Gerson C. Duarte-Filho** [2] **and Arthur A. Brum** [3]

[1] *Departamento de Física, Universidade Federal do Paraná, 81531-990 Curitiba, Paraná, Brazil*
[2] *Departamento de Física, Universidade Federal de Sergipe, 49100-000, São Cristóvão, Sergipe, Brazil*
[3] *Departamento de Física, Universidade Federal de Pernambuco, 50670-901 Recife, Pernambuco, Brazil*

Correspondence*:
Giovani L. Vasconcelos
giovani.vasconcelos@ufpr.br

ABSTRACT

The response of the scientific community to the global health emergency caused by the COVID-19 pandemic has produced an unprecedented number of manuscripts in a short period of time, the vast majority of which has been shared in the form of preprints posted before peer review on preprint repositories. This surge in preprint publications has in itself attracted considerable attention, although mostly in the bibliometric literature. In the present study we apply a mathematical growth model, known as the generalized Richards model, to describe the time evolution of the cumulative number of COVID-19 related preprints. This mathematical approach allows us to infer several important aspects concerning the underlying growth dynamics, such as its current stage and its possible evolution in the near future. We also analyze the rank-frequency distribution of preprints servers, ordered by the number of COVID-19 preprints they host, and find that it follows a power-law decay. This Zipf-like law indicates the presence of a cumulative advantage effect, whereby servers that already have more preprints receive more submissions.

1 INTRODUCTION

The COVID-19 pandemic is undoubtedly the most serious public health crisis in over a century. It has claimed the lives of nearly 900,000 people worldwide, as of this writing, and it has seemingly touched, in one way or another, every aspect of our daily lives. Not surprisingly, given the health risks represented by the novel coronavirus (SARS-CoV-2) and the many scientific challenges involved, the COVID-19 pandemic has had a huge impact on the scientific community. On the downside, social isolation and lockdown measures, for example, have caused disruption (if only temporary) of research projects that were being undertaken before and made it difficult for researchers to collaborate and discuss their work in person during the pandemic. But on the up side, the scientific community responded quickly to the challenges posed by the unprecedented crisis by producing an unprecedented number of scholarly works in a short period of time.

A considerable proportion of this scientific output has been disseminated in the form of preprints posted before peer review on open and publicly available online platforms. Despite some concerns regarding the quality of preprints [1, 2, 3, 4, 5], in comparison to peer-reviewed articles, there seems to be a growing consensus that the benefits of the rapid sharing of information allowed by preprints "far outweigh

the disadvantages" [2]. For example, a recent bibliometric analysis [5] of COVID-19 related preprints concluded that the faster rate with which they are posted on preprint repositories have only "limited impact on the quality of preprints that are subsequently published." Many other studies of both peer-reviewed papers and unrefereed preprints related to COVID-19 research have recently been conducted, but mostly with emphasis on the bibliometric aspects of the topic; see, e.g., Refs. [6, 7, 8, 9, 10, 11, 12] and references therein.

In the present paper we take a different approach from these previous studies. Here we seek to understand the growth dynamics underlying the rapid surge of preprints related to COVID-19 research. To this end, we employ a generalized logistic growth model to describe the time evolution of the cumulative number of preprints deposited on preprint repositories. Since the pioneering work by Verhulst on the standard logistic model [13], phenomenological growth models have been successfully applied to many growth processes, ranging from human [13] and animal [14] population dynamics to epidemics [15], including the COVID-19 epidemic itself [16, 17]. It is thus natural to expect that growth models should also be applicable to this "epidemic in an epidemic," as the rapid surge of preprints on COVID-19 has been called [10]. More specifically, here we shall apply the generalized Richards model (GRM) to describe the cumulative curve of COVID-19 preprints. We show that the GRM does give a very good fit to the empirical curve. Furthermore, the model allows us to extract relevant information about the social dynamics driving this quick growth of preprints. We show, for instance, that the early growth in the number of COVID-19 preprints follows a subexponential regime—rather than an exponential behavior, as was initially thought [6, 12, 18]. The model also predicts that the growth profile is currently in a phase of decreasing deceleration.

We also analyze the distribution of the COVID-19 preprints among the many available servers. Here we find that the corresponding rank-frequency distribution follows a power law, similar to the Zipf law found in many preferential attachment processes, such as the frequency of words in a text [19], the size distribution of cities [20], the wealth distribution of individuals [20], and the distribution of nodes in social networks [21], among others. We thus argue that the distribution of manuscripts among repositories also seems to follow a similar preferential attachment dynamics, whereby servers that already have more preprints tend to get more submissions.

2 MATERIALS AND METHODS

2.1 Data Source

The data used in this study were obtained from the GitHub site maintained by Fraser and Kramer [22]. As explained in their site, preprints are considered to be related to COVID-19 on the basis of keywords matches in their titles or abstracts, according to the following search string: coronavirus OR covid-19 OR sars-cov OR ncov-2019 OR 2019-ncov OR hcov-19 OR sars-2. It is also explained there [22] that "only the earliest posted version" of a preprint is included in the dataset and that, in cases where a preprint is deposited to multiple repositories, "all preprint records are included." The first preprint included in their collection was posted on bioRxiv on January 15, 2020 [23]. The dataset used in the present study was updated up until August 30, 2020, and contains a total number of 24,644 preprints, distributed among 35 preprint servers, with 15 of them hosting more than 100 preprints. For the list of the fifteenth largest servers (as per the number of COVID-19 preprints), see Sec. 3. A complete list of the preprints servers in the dataset can be found in [5].

2.2 Mathematical Growth Model

We model the growth dynamics by means of the generalized Richards growth model (GRM), which is defined by the following ordinary differential equation:

$$\frac{dN}{dt} = r\,[N(t)]^q \left[1 - \left(\frac{N(t)}{K}\right)^{\alpha}\right], \tag{1}$$

where $N(t)$ is the growing quantity at time t, r is the growth rate at the early growth stage, q controls the initial growth regime and allows to interpolate from linear growth ($q = 0$) to sub-exponential growth ($q < 1$) to purely exponential growth ($q = 1$), α is the asymmetry parameter that controls the asymmetry of the growth profile with respect to the symmetric S-shaped curve of the logistic model, which is recovered for $q = \alpha = 1$, and K represents the total quantity at the end of the growth process (i.e., for $t \to \infty$). Equation (1) must be supplemented with the initial condition $N(0) = N_0$, for some given value of N_0.

Here we shall apply the GRM to the growing number of COVID-19 related preprints, so that $N(t)$ will represent the cumulative number of preprints in our dataset up to the time t, where t is measured in days since the first preprint (hence $N_0 = 1$). In adjusting the GRM to the empirical data we need to determine four free parameters, namely (r, q, α, K); the numerical fit is made quite easy by the existence of an analytical solution for the GRM, as discussed next.

Equation (1) admits an exact solution in implicit form given by [17]

$$t = f(N; r, q, \alpha, N_0), \tag{2}$$

where

$$f(N; r, q, \alpha, N_0) = \frac{N^{1-q}}{r(1-q)}\, {}_2F_1\left(1, \frac{1-q}{\alpha}; 1 + \frac{1-q}{\alpha}; \left(\frac{C}{K}\right)^{\alpha}\right) - t_i, \tag{3}$$

with ${}_2F_1(a, b; c; x)$ denoting the Gauss hypergeometric function [24] and

$$t_i = {N_0}^{1-q}\, {}_2F_1\left(1, \frac{1-q}{\alpha}; 1 + \frac{1-q}{\alpha}; \left(\frac{N_0}{K}\right)^{\alpha}\right).$$

The fact that the solution of the GRM is given implicitly as $t(N)$, rather than as an explicit function $N(t)$, does not represent any hurdle to its practical use. Indeed, the above solution can be directly applied for curve-fitting purposes by viewing the empirical data in the same 'implicit' form, namely t_k as a function of N_k, where N_k are the data points at times t_k. The availability of an exact solution also has the advantage that it allows us to compute explicitly the location of certain key characteristic points of the growth profile, as indicated below.

For example, the inflection point t_c of the curve $N(t)$, where $\ddot{N}(t_c) = 0$, is given by [17]

$$t_c = \frac{K^{1-q}}{r(1-q)} \left(\frac{q}{q+\alpha}\right)^{(1-q)/\alpha} {}_2F_1\left(1, \frac{1-q}{\alpha}; 1 + \frac{1-q}{\alpha}; \frac{q}{q+\alpha}\right) - t_i. \tag{4}$$

Knowledge of the inflection point t_c is important because it divides the growth process into two main phases according to its acceleration, as follows: i) an accelerating phase, for $t < t_c$, when $\ddot{N}(t) > 0$; and ii) a decelerating phase, for $t > t_c$, during which $\ddot{N}(t) < 0$. Each of these two main phases can be further divided into two subphases, according to whether the corresponding acceleration/deceleration is increasing

or decreasing. More specifically, recalling that the rate of acceleration is known as the jerk, let us denote the points of zero jerk by $t_j^{\pm}$, with $\dddot{N}(t_j^{\pm}) = 0$, where t_j^- and t_j^+ correspond to the points of maximum acceleration and maximum deceleration, respectively, so that $t_j^+ > t_j^-$. After some tedious algebra one finds that $t_j^{\pm} = f(Ky_{\pm})$, where $f(x)$ is as given in (3) and

$$y_{\pm} = \left[\frac{\alpha^2 + 2q(-1+2q) \pm \alpha\left(1 - 4q + \sqrt{4q(-1+2q) + 1 - 2\alpha + \alpha^2 + 8\alpha q}\right)}{4\alpha^2 + 2q(-1+2q) + 2\alpha(-1+4q)} \right]^{1/\alpha} . \tag{5}$$

Now, comparing the time, t_f, of the last empirical datapoint (corresponding to the 'current time') with the characteristic points (t_j^-, t_c, t_j^+) of the theoretical curve allows us to classify the current stage of the growth process. More specifically, we can define four growth stages, as follows: i) increasing acceleration, if $t_f < t_j^-$; ii) decreasing acceleration, if $t_j^- < t_f < t_c$; iii) increasing deceleration, if $t_c < t_f < t_j^+$; and iv) decreasing deceleration, if $t_f > t_j^+$. Having such a finer 'diagnosis' of the growth process is useful not only because it provides valuable information about the current stage of the underlying dynamics, but also because it allows us to make predictions about its likely evolution within the near future. For instance, depending on how close the current time t_f is in comparison to the nearest phase-separation point, we may have an idea of how recently the growth curve has entered its current stage or how soon it may transition to the next one (if it is not yet in the last stage).

In particular, we note that in the first subphase, i.e., for $t < t_j^-$, the GRM predicts a polynomial growth of the form [17]

$$N(t) \approx At^{\mu}, \tag{6}$$

where $A = [r(1-q)]^{1/(1-q)}$ and $\mu = 1/(1-q)$, for $q < 1$. In contradistinction, early exponential growth is obtained only for $q = 1$, in which case one has $N(t) \approx N_0 \exp(rt)$. Similarly, in the late-time dynamics, i.e., for $t > t_j^+$, the GRM predicts an exponential rise to the plateau of the form [17]: $N(t) - K \propto \exp(-\gamma t)$, where $\gamma = r\alpha/K^{1-q}$.

2.3 Rank-Frequency Distribution

To analyze the size distribution of the preprint repositories, we first order the repositories according to the number of COVID-19 preprints they hold, where $n = 1$ is attributed to the repository with the highest number of preprints, $n = 2$ for the second largest repository, and so on. The relative frequency of preprints in the n-th repository will be denoted by $P(n)$, where $P(n) = NP(n)/\sum_j NP(j)$, with $NP(j)$ denoting the number of preprints in the j-th largest repository.

Preprint servers operate largely under similar principles. They aim to provide a publicly and freely accessible platform for rapid dissemination and sharing of scientific manuscripts that were not yet certified by peer review. There are preprint repositories that specialize in certain areas, such as: arXiv for physics and mathematics; bioRxiv and medRxiv for biomedical sciences; SSRN for the social sciences; and RePEc for economics research. There are also multidisciplinary preprint platforms, such as Research Square and others. As already mentioned, there are 35 preprint servers in our dateset, with fifteen of them hosting more than 100 COVID-19 manuscripts.

Authors thus have a substantial array of preprint servers to choose from when submitting a manuscript on COVID-19 related research. This is in quite contrast to non-COVID-19 preprints, where authors often

prefer to use servers that specialize in their disciplines. It is therefore interesting to investigate how authors decide to which server to submit their COVID-19 preprints. For instance, preprint repositories may differ somewhat in their screening procedures and other policy requirements [1, 4], but it is fair to argue that these eventual differences do not represent a significant decision factor. In other words, the 'cost' of submission (say, in terms of time and extra work involved in the submission process) and the eventual 'risk' of rejection (if the manuscript is deemed not appropriate for the chosen server) are roughly the same among the different platforms. It is therefore reasonable to expect that authors will preferentially seek those platform that may provide greater visibility for their work.

A reasonable strategy decision in such situation is of course to favor those repositories that already have a substantial number of preprints. In the context of the COVID-19 emergency, this strategy makes particular sense for authors whose principal areas of expertise are not directly related to, say, epidemiology, infectious diseases, virology, etc. This selection dynamics naturally leads to the so-called cumulative advantage [25] or preferential attachment effect [21], so that repositories that already have more preprints receive more submissions. Preferential attachment processes usually lead to rank-frequency distributions, $P(n)$, that exhibit power-law decay or the so-called Zipf law [19, 26, 20, 25]:

$$P(n) \propto \frac{1}{n^{\rho}}, \tag{7}$$

for $n \geq 1$ and $\rho > 0$. In the next section we shall investigate the evidence of power-law behavior in the frequency distribution of COVID-19 preprints by repositories.

2.4 Statistical Fits

To perform the statistical fit for the GRM, we employed the Levenberg-Marquardt algorithm to solve the non-linear least square optimization problem, as implemented in the *lmfit* package for Python [27], which provides the parameter estimates and their respective errors. Here we have set $N(0) = N_0 = 1$, so that according to (2) we are left with four parameters, namely (r, q, α, K), to determine numerically. In the case of the rank-frequency distribution, we fitted the selected data with a simple power-law function, $P(n) = Bn^{-\rho}$, and also applied *lmfit* to determine the parameters B and ρ. The computer codes for the statistical fits were written in the *Python* language, and the plots were produced with the data visualisation library *Matplotlib*.

3 RESULTS

In Fig. 1 we show the cumulative number (red circles) of preprints on Covid-19 in our dataset, which we recall covers from January 15, 2020, to August 30, 2020, together with the GRM best fit (black solid curve). One sees from this figure that the theoretical curve describes very well the empirical data. Also shown in Fig. 1 are the point of maximum acceleration (orange vertical line), the inflection point (yellow vertical line), and the point of maximum deceleration (green vertical line), as obtained from the theoretical fit. The legend box in the figure shows the parameter estimates from the best fit.

In Fig. 2 we show the daily number of preprints (green curve) and the theoretical daily curve, which corresponds to the time derivative of the model cumulative curve shown in Fig. 1. Although the daily curve is quite noisy, as expected for a random-like process such as preprint submissions, we clearly see that the theoretical curve captures rather well the general trend of the daily data. In particular, the theoretical estimate for the "peak" of the daily curve (corresponding to the inflection point t_c of the cumulative curve) matches quite well the location of the region of largest values in the empirical curve.

In Fig. 3 we show the rank-frequency distribution for the preprint repositories that have at least 100 preprints; there are 15 of them, which account for the near totality (98.1%) of all COVID-19 related preprints. We see furthermore that the first ranked server (medRxiv) alone contributes with over one quarter (25.6%) of all submissions, which is nearly the double of the second ranked repository (SSRN, with 13.2%). Furthermore, the six largest preprint servers (as per the number of preprints on COVID-19) together account for more than three quarters (76.6%) of all preprints. Moreover, one sees that there is a significant 'gap' in size between the sixth (bioRxiv) and the seventh (JMIR) largest repositories, which seems to suggest a change of dynamics at this point. These evidences taken together thus indicate that the first six preprint server dominate the 'preferential attachment' submission process. This is further corroborated by the fact that the rank-size distribution for these first sixth serves does follow a Zipf law, with exponent $\rho = 0.65$, as shown in the inset of Fig. 3.

4 DISCUSSION

We have seen above that the time evolution of the number of COVID-19 related preprints is well described by the generalized Richards growth model. The application of such growth model allows us to infer several interesting aspects of the dynamics underlying the epidemic-like growth of COVID-19 preprints. First, we saw that that early in the epidemic the number of preprints increased in a subexponential manner, as indicated by the value $q = 0.71$ obtained from the GRM fit; see Fig. 1. This means, more concretely, that initially the number of preprints grows polynomially in time according to (6), with an exponent $\mu = 3.4$, rather than exponentially fast as was claimed in some early bibliometric studies on the subject [6, 12, 18]. (We recall that pure exponential growth occurs only for $q = 1$.) This subexponential spreading is also found in many real epidemics [28], including COVID-19 itself [29]. Polynomial epidemic growth is usually attributed to heterogeneous mixing [28, 30], where clustering effects in the underlying propagation network can lead to polynomial spreading [31, 32]. So it is quite likely that such complex dynamics also takes place in the early rapid growth of COVID-19 publications, leading to a subexponential regime. (Developing a 'microscopic' epidemic model for the growth of COVID-19 preprints, where such effects could be studied in more detail, is an interesting problem but one that is beyond the scope of the present article.)

Another interesting result obtained from Fig. 1 is the fact that the inflection point of the growth profile was reached slightly over four months after the first preprint in our dataset, i.e., $t_c = 130$ days, corresponding to May 24, 2020, after which the curve has entered a deceleration phase. This deceleration regime can be explained by a combination of factors. First, after a few months of an exceedingly rapid growth in the number of preprints, it becomes naturally more difficult for researchers to obtain novel results at the same pace. Second, it is also possible that after a few months of intensive work on COVID-19, some researchers (especially those whose main areas of expertise are not directly related to epidemics and infectious diseases) may have shifted their focus back to previous research problems or moved on to new ones. Third, starting in late April and early May, repository administrators began to screen more closely COVID-19 preprints against "poor science" [4], and this more stringent vetting processes may have had an impact (however modest) on the rate of accepted submissions. Furthermore, it may also have discouraged authors from submitting manuscripts that they feared would not pass the stricter screening. In other words, enhancing the screening procedures was the equivalent, to some extent, of a 'mitigation' intervention in epidemic outbreaks. These factors combined (and possibly others) have lead to a slowing down in the rate of new preprints—an effect that is effectively captured by the saturation term in the GRM, as represented by the term in square brackets in Eq. (1).

We have found that the point of maximum deceleration (and hence zero jerk) of the theoretical growth curve happened on July 23, 2020, as indicated by the green vertical line in Fig. 1. From the GRM fit, we have also computed the point of maximum jerk (not shown in Fig. 1) and obtained that it is predicted for September 11, 2020. This implies that the growth curve is currently in a regime of decreasing deceleration and increasing jerk. One important effect of this increasing jerk is that it contributes to "bend the curve" away from the near-linear growth that is typical of intermediate region around the inflection point t_c; see Fig. 1. The model thus predicts that the growth curve should from now on develop a more curved profile and that in the near future it will likely enter a saturation phase, where one should see a rather slower growth towards a plateau. It is not yet clear, however, how this approach to the plateau will take place: if exponentially fast as predicted by the GRM or in a subexponential fashion, in which case one would have to consider more general growth models [17]. The dynamical nature of this 'mature phase' will depend, of course, on the intensity level of the continuing research on themes related to COVID-19. As more data is accumulated over the next months, it will be possible to investigate in more detail this late-time regime.

We have also analyzed the size distribution of preprint repositories as ordered by the number of COVID-19 related manuscripts they host. In particular, we found that this distribution is highly peaked at the front-runner (medRxiv)—a property that is typical of the so-called cumulative advantage processes. Such processes often exhibit power-law distributions, and we have indeed verified that the rank-size distribution of preprint servers does have a power-law decay (at least for the largest repositories). This seems to indicate that a sort of preferential attachment dynamics takes place when authors are considering to which platform to send their manuscripts, as servers that already have more preprints tend to receive more submissions.

In conclusion, it is fair to say that, alongside the public health crisis, the COVID-19 pandemic also triggered a scientific emergency. The scientific community has met this challenge by producing an unprecedented number of scholarly works in a very short period of time, so much so that the phenomenon has been dubbed "an epidemic in an epidemic" [10]. To better understand this "scidemic," we have applied a generalized logistic growth model to describe the time evolution of the cumulative number of unrefereed preprints on COVID-19 and SARS-CoV2. Our analysis shows that the quick surge in COVID-19 related preprints can be seen as a sort of contagion process, where existing preprints tend to spur more preprints, and so on, in a cascade-like effect. Eventually this rapid growth is tamed by the system's own dynamics, as it takes more time and effort for researchers to obtain new results, leading to a deceleration in the growth of the preprint 'epidemic curve.' (Other factors, such as closer screening of preprints by the repositories and the return of researchers to pre-epidemic projects, may also contribute to the slowing down in the rate of publication of COVID-19 preprints.) This initial deceleration regime is expected to be followed by a saturation phase of slow growth, as both the COVID-19 and the publication epidemics subside. It is possible, however, that in view of the huge amount of data that has been, and continues to be, generated by the COVID-19 pandemic and the fact that there are still many unanswered scientific questions, a somewhat steady level of COVID-19 publications may continue for a longer period, thus delaying the late-time saturation regime.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

GLV designed the study. LPC was responsible for collecting and curating the data. LPC, GCDF, and AAB developed the codes. LPC performed the numerical analyses. All authors discussed the results. GLV wrote

the first draft. All authors revised and contributed to subsequent drafts. All authors read and approved the final manuscript.

FUNDING

This work was partially supported by the National Council for Scientific and Technological Development (CNPq) in Brazil, through a research fellowship for GLV (grant No. 303772/2017-4), a PhD fellowship for AAB (grant No. 167348/2018-3), and a Science Initiation fellowship for LPC (grant No. 104166/2020-7). GLV also acknowledges partial funding from UFPR through the COVID-19/PROIND-2020 Research Program.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study, together with the respective numerical codes, can be found in the website of our research group (`http://fisica.ufpr.br/redecovid19/software.html`) or can be requested from the authors.

REFERENCES

[1] Kirkham JJ, Penfold N, Murphy F, Boutron I, Ioannidis JP, Polka JK, et al. A systematic examination of preprint platforms for use in the medical and biomedical sciences setting. *bioRxiv* (2020). doi:10.1101/2020.04.27.063578.

[2] Kupferschmidt K. Preprints bring 'firehose'of outbreak data. *Science* **367** (2020) 963 – 964. doi:10.1126/science.367.6481.963.

[3] Majumder MS, Mandl K. Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. *The Lancet* **8** (2020) E627 – E630. doi:10.1016/S2214-109X(20)30113-3. Published Online March 24, 2020.

[4] Kwon D. How preprint server are blocking bad coronavirus research. *Nature* **581** (2020) 130 – 131.

[5] Fraser N, Brierley L, Dey G, Polka JK, Pálfy M, Coates JA. Preprinting a pandemic: the role of preprints in the COVID-19 pandemic. *bioRxiv* (2020). doi:10.1101/2020.05.22.111294.

[6] Torres-Salinas D. Ritmo de crecimiento diario de la producción científica sobre COVID-19. análisis en bases de datos y repositorios en acceso abierto. *El profesional de la información* **29** (2020). doi:10.3145/epi.2020.mar.15.

[7] Al-Zaman MS. Bibliometric analysis of COVID-19 literature. *medRxiv* (2020). doi:10.1101/2020.07.15.20154989.

[8] Aristovnik A, Ravšelj D, Umek L. A bibliometric analysis of COVID-19 across science and social science research landscape. *Preprints* (2020). doi:10.20944/preprints202006.0299.v3.

[9] Homolak J, Kodvanj I, Virag D. Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders. *Scientometrics* (2020). doi:10.1007/s11192-020-03587-2.

[10] Odone A, Salvati S, Bellini L, Bucci D, Capraro M, Gaetti G, et al. The runaway science: a bibliometric analysis of the COVID-19 scientific literature: How COVID-19 has changed academic publishing. *Acta Bio Medica Atenei Parmensis* **91** (2020) 34–39. doi:10.23750/abm.v91i9-S.10121.

[11] Nowakowskaa J, Sobocińskaa J, Lewickib M, Lemańskaa Z, Rzymskic P. When science goes viral: The research response during three months of the COVID-19 outbreak. *Biomed. Pharmacother.* **129** (2020) 110451. doi:10.1016/j.biopha.2020.110451.

[12] Bobrowski T, Melo-Filho CC, Korn D, Alves VM, Popov KI, Auerbach S, et al. Learning from history: do not flatten the curve of antiviral research! *Drug Discovery Today* (2020). doi:10.1016/j.drudis.

2020.07.008.

[13] Verhulst PF. Recherches mathématiques sur la loi d'accroissement de la population. *Nouv. mém. de l'Academie Royale des Sci. et Belles-Lettres de Bruxelles* **18** (1845) 1 – 41.

[14] Richards F. A flexible growth function for empirical use. *Journal of experimental Botany* **10** (1959) 290–301.

[15] Wang XS, Wu J, Yang Y. Richards model revisited: Validation by and application to infection dynamics. *Journal of Theoretical Biology* **313** (2012) 12–19.

[16] Vasconcelos GL, Macêdo AM, Ospina R, Almeida FA, Duarte-Filho GC, Brum AA, et al. Modelling fatality curves of COVID-19 and the effectiveness of intervention strategies. *PeerJ* **8** (2020) e9421. doi:10.7717/peerj.9421.

[17] Vasconcelos GL, Macêdo AM, Duarte-Filho GC, Araújo AA, Ospina R, Almeida FA. Complexity signatures in the COVID-19 epidemic: power law behaviour in the saturation regime of fatality curves. *medRxiv medRxiv 2020.07.12.20152140* (2020).

[18] Zhu X, Jin Q, Jiang X, Dan Y, Zhang A, Qiu G, et al. Global pattern of COVID-19 research. *medRxiv* (2020). doi:10.1101/2020.07.04.20146530.

[19] Zipf GK. *Human Behavior and the Principle of Least Effort* (Addison-Wesley) (1949).

[20] Simon HA. On a class of skew distribution functions. *Biometrika* **42** (1995) 425 – 440.

[21] Barabási AL, Albert R. Emergence of scaling in random networks. *Science* **286** (1999) 509 – 511.

[22] [Dataset] Fraser N, Kramer B. COVID-19 preprints. Available at: `https://github.com/nicholasmfraser/covid19_preprints` (2020). Accessed: September 01, 2020.

[23] Wang Y, Zhang W, Jefferson M, Sharma P, Bone B, Kipar A, et al. The WD and linker domains of ATG16L1 required for non-canonical autophagy limit lethal respiratory infection by influenza a virus at epithelial surfaces. *bioRxiv* (2020). doi:10.1101/2020.01.15.907873.

[24] Aomoto K, Kita M, Kohno T, Iohara K. *Theory of Hypergeometric Functions* (Springer) (2011).

[25] Price DdS. A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Inf. Sci.* **27** (1976) 292 – 306.

[26] Yule GU. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B* **213** (1925) 21 – 87.

[27] Newville M, Stensitzki T, Allen D, Ingargiola A. Non-linear least-squares minimization and curve-fitting for Python. *Chicago, IL* (2015).

[28] Viboud C, Simonsen L, Chowell G. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics* **15** (2016) 27–37.

[29] Manchein C, Brugnago EL, da Silva RM, Mendes CF, Beims MW. Strong correlations between power-law growth of COVID-19 in four continents and the inefficiency of soft quarantine strategie. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **30** (2020) 041102.

[30] Chowell G, Sattenspiel L, Bansal S, Viboud C. Mathematical models to characterize early epidemic growth: A review. *Phys. of Life Reviews* **18** (2016) 66–97.

[31] Szendroi B, Csányi G. Polynomial epidemics and clustering in contact networks. *Proc. R. Soc. Lond. B (Suppl.)* **271** (2004) S364 – S366. doi:10.1098/rsbl.2004.0188.

[32] Vazquez A. Polynomial growth in branching processes with diverging reproductive number. *Phys. Rev. Lett.* **96** (2006) 038702. doi:10.1103/PhysRevLett.96.038702.

FIGURE CAPTIONS

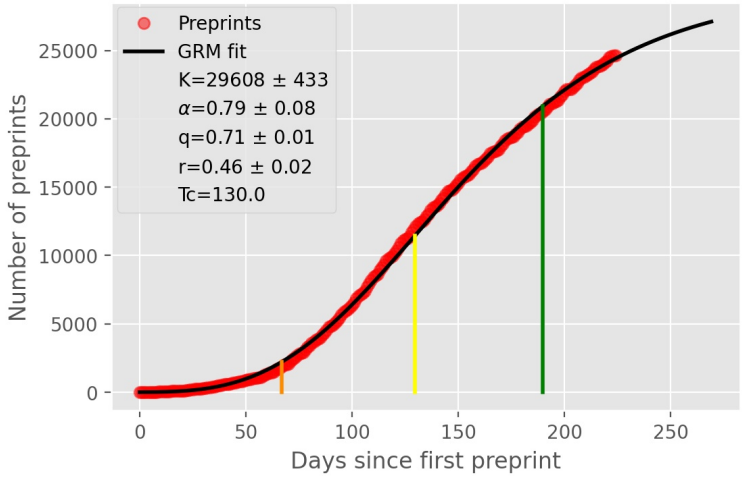

Figure 1. Cumulative number (red circles) of COVID-19 related preprints deposited on online preprint repositories up to August, 30, 2020. The solid black curve is the fit to the empirical data by the generalized Richards model, with the parameters shown in the legend box. The vertical lines indicate the location of some key characteristic points of the theoretical curve, as follows: i) point of maximum acceleration $t_j^- = 67$ (orange line); ii) inflection point $t_c = 130$ (yellow line); and iii) point of maximum deceleration $t_j^+ = 190$ (green line).

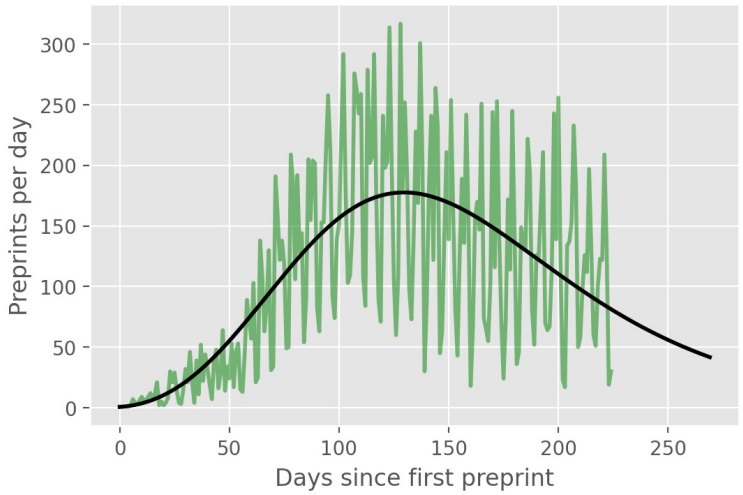

Figure 2. Daily number (green line) of COVID-19 preprints posted on online preprint repositories up to August, 30, 2020. The solid black line is the derivative of the theoretical cumulative curve shown in Fig. 1.

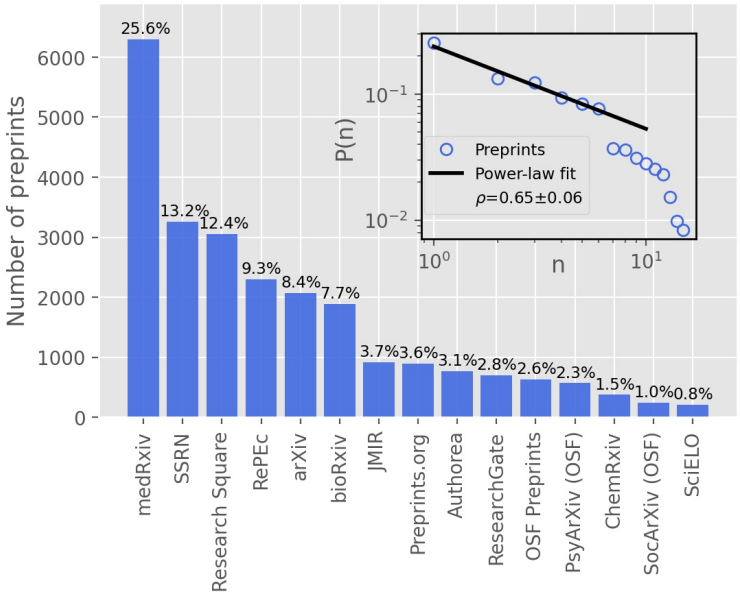

Figure 3. Ranking of the fifteenth largest preprint repositories by number of COVID-19 related preprints, up to August, 30, 2020. The inset shows the rank-frequency distribution in log-log scale, where the straight line is a power-law fit, $P(n) \propto n^{-\rho}$, to the first sixth largest repositories, which yields an exponent $\rho = 0.65$.