PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Feng M, Deng L-J, Chen F, Perc M, Kurths J. 2020 The accumulative law and its probability model: an extension of the Pareto distribution and the log-normal distribution. *Proc. R. Soc. A* **476**: 20200019. http://dx.doi.org/10.1098/rspa.2020.0019

Received: 10 January 2020
Accepted: 24 March 2020

Subject Areas:
complexity

Keywords:
accumulative law, complex network, network model, probability density function, Pareto distribution, log-normal distribution

Author for correspondence:
Matjaž Perc
e-mail: matjaz.perc@gmail.com

The accumulative law and its probability model: an extension of the Pareto distribution and the log-normal distribution

Minyu Feng[1], Liang-Jian Deng[2], Feng Chen[1], Matjaž Perc[3,4] and Jürgen Kurths[5,6]

[1]College of Artificial Intelligence, Southwest University, Chongqing 400715, People's Republic of China
[2]School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China
[3]Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, 2000 Maribor, Slovenia
[4]Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 404, Taiwan
[5]Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany
[6]Department of Physics, Humboldt University, 12489 Berlin, Germany

MF, 0000-0001-6772-3017; MP, 0000-0002-3087-541X

The divergence between the Pareto distribution and the log-normal distribution has been observed persistently over the past couple of decades in complex network research, economics, and social sciences. To address this, we here propose an approach termed as the accumulative law and its related probability model. We show that the resulting accumulative distribution has properties that are akin to both the Pareto distribution and the log-normal distribution, which leads to a broad range of applications in modelling and fitting real data. We present all the details of the accumulative law, describe the properties of the distribution, as well as the allocation and the accumulation of variables. We also show how the proposed accumulative law can be applied to generate complex networks, to describe the accumulation of personal wealth, and to explain the scaling of internet traffic across different domains.

THE ROYAL SOCIETY PUBLISHING

1. Introduction

During the development of the probability theory, Pareto distribution named after the Italian economist and sociologist Vilfredo Pareto, which is also known as the power-law distribution for a specific case, has become an indispensable component in research fields. It is a continuous probability distribution of a random variable whose curve is long-tailed, and specifically, the zeta distribution is its discrete case. The observable phenomena presented by the Pareto distribution are commonly referred to as the Pareto principle, or '80-20 rule'. The rule says that, e.g. for networks, 80% of the degree of a complex network is held by 20% of its vertices [1]. This distribution has already been found empirically to fit a wide range of situations including the frequency of occurrence of unique words in a novel [2], the size distribution of gene families [3], the asymptotic decay of the total conductance of subcritical trees [4], the sizes of human settlements [5], the protein sequence alignments [6], even the distribution of artists by the average price of their artworks [7], etc.

Another frequently used continuous probability distribution is the log-normal distribution consisting of a random variable whose logarithm is normally distributed. A log-normal distribution describes numerous growing processes based on the accumulation of small percentage changes that is a log scale. Since a log-normally distributed variable only takes real positive values and is non-symmetric, it better simulates a non-negative and non-uniform distribution than the normal distribution, especially in various phenomena of economics and social sciences [8]. Some of the usual log-normally distributed cases include fire sizes [9], the size of cities [10], the fatigue lifetime of a maintainable systems [11], the trip duration for taking a taxi [12] and indeed many more.

Both the Pareto distribution and the log-normal distribution play a significant role in the probability theory and statistical applications. However, some important unaddressed issues persist, leading to a considerable divergence in many cases. Taking the network science as an example, Barabási highlighted the degree distribution, the probability distribution of degrees over the whole network, as a fundamental concept [13]. He used the mean-field and continuous theory to calculate the degree distribution of scale-free networks, and the outcome is the very famous power-law distribution independent of time, a particular case of the Pareto distribution [14]. However, due to the method Barabási used to obtain the power-law distribution, he ignored the discrete and continuous problems and idealized the variations of the networks, and some queries came along. The central question is the concept and derivation of a power-law distribution. Soon after the scale-free networks were proposed, Bollobás *et al.* suggested that the modelling process of scale-free networks by Barabási is not precise, and they presented a more mathematical-based method to construct networks [15]. Later, Li *et al.* claimed that there was no consistent, precise definition of scale-free networks and only a few rigorous proofs of many of their claimed properties [16]. Then, Krioukov *et al.* used a geometric approach to obtain the distribution considering hyperbolic spaces [17], and others applied it to real-world cases [18,19]. May *et al.* indicated that the subnetworks of scale-free networks are not scale-free, i.e. not following a power-law distribution [20]. Instead of the power-law distribution, Fang *et al.* proposed a double Pareto log-normal distribution for complex networks [21].

Besides the degree distribution, there are also outstanding challenges in economics. For example, Robert Gibrat proposed a principle to describe that the proportional rate of growth of a firm is independent of its absolute size [22]. Generally, processes characterized by the Gibrat's Law converge to a limiting distribution, often proposed to be the log-normal, or a power-law, depending on more specific assumptions about the stochastic growth process. Nevertheless, the tail of the log-normal sometimes drops fast and its probability density function is not monotonic, but rather has a Y-intercept of zero probability at the beginning. The Pareto distribution's tail cannot simulate the decline in the tail when the size is large, and which does not extend downwards to zero, should be truncated at some positive minimum value instead. Thus, many researchers proposed other distributions they claimed as a better result for the Gibrat's process, such as the Weibull distribution [23]. Considering the insufficiency of rigorous proof on the

Gibrat's Law, Stanley *et al.* even disagreed with the theory and result of Gibrat's Law and proposed that the outcome only works empirically in the study of firms [24]. Rozenfeld *et al.* also pointed out that the city size is partially against the Gibrat's Law [25]. As we can see, both the Pareto and log-normal distributions also have some problems in fitting practical distributions.

In this article, we aim to address the issue on the inaccuracy of fitting certain distributions in real situations. For that purpose, we first investigate those practical situations that the Pareto distribution and the log-normal distribution can not perfectly fit, e.g. the degree distribution, firm's size, personal wealth distribution, etc. All of them have a common characteristic: they have a continuous growing process. Taking the scale-free networks as an example, the connection of a vertex keeps growing since the new vertices generate and bring connections allocating to the network, then the degree will be an accumulation of connections. Another common feature displays as the unequal allocation of the increments or income. For people in a company, the total income appeals to inequality in allocation, thus different people obtain different wealth, usually the rich are getting richer, and the poor are getting poorer. Taking both processes into consideration, our solution is not simply to use the Pareto distribution or log-normal distribution to empirically fit the practical distribution, instead, we propose a mathematical model based on the growing process and allocation process. The algorithm of this model concludes as the accumulative law since the model shows an accumulative property to describe the unequally accumulative phenomenon. Through this law, we derive a novel probability distribution and its variant form to widely fit unequally allocated distribution. The statistical properties are studied to better understand the behaviour of the accumulative law. The goodness of fit of the accumulative distribution is an application as well as an evidence to show that our distribution should have a better result in describing real systems, e.g. the degree distribution, personal wealth distribution, website visit distribution, etc.

The organization of this paper is as follows: A concrete theory of the accumulative law and its probability model is discussed in §2. Accumulative function and some statistical properties are calculated in §3. The real systems based on the accumulative law are described in §4. Besides, we display some simulation studies in §5. Finally, some conclusions are drawn in §6.

2. Accumulative law and its probability model

As described above, the controversy over the log-normal and power-law distribution has never ended in economics and various other fields. However, they both do not strictly describe the allocated and accumulative process. In this section, we propose a linearly growing model by the time whose allocation is decided by the 'Matthew effect' which is proven useful in empirical data [26] and calculate its accumulative distribution. It is worth noting that the accumulative distribution in this paper is different from the cumulative distribution function, which is an accumulation of increments instead of probabilities.

(a) Accumulative law

In essence, our probability model under the accumulative law describes a growing system by time that contains large individuals and keeps receiving new individuals. Each new individual arrives at the system at a specific rate and some increments come along, and these increments are allocated among the existing individuals by the rule of the 'Matthew effect', i.e. those possessing more accumulative increments are more likely to be distributed than the fewer ones. These increments gradually constitute the accumulation of distributed individuals. As time passes, the stationary accumulative distribution of the system is our object function.

In detail, the accumulative law consists of two featured processes of growth and allocation:

(1) *Growing process*: the number of individuals in this system keeps growing at a specific rate and makes the size of the system also grow; the increments carried by new

individuals also keep growing along with the number of new individuals and turn into the accumulation of the existing individuals.

(2) *Allocating process*: the increments of each new individual are divided into portions, each portion is more possibly allocated to an existing individual with a higher accumulation.

Based on this law, we calculate the accumulative probability distribution.

To simplify and clarify the calculation process, some fundamental conceptions are required and described as follows:

As the object function, we primarily propose the definition of the accumulative distribution in this paper.

Definition 2.1. Let $Z(t)$ denote the accumulation in the system at time t and define the accumulative distribution $f(z)$, $z > 0$, by

$$f(z) = \lim_{t\to\infty} P_z(t) = \lim_{t\to\infty} P\{Z(t) < z\} \tag{2.1}$$

where we assume the limit exists, and $P_z(t)$ is the probability that the accumulation value is z at time t.

The function $f(z)$ is a continuous probability distribution of a random variable whose value is distributed by the increments unequally, and the function $F(Z)$ is its cumulative distribution.

We then introduce the definition of the transition probability for the accumulation $Z(t)$.

Definition 2.2. The transition probability $P_{ij}(t)$ is the probability that the process will, when in the accumulation i, after a time t, next make a transition into the accumulation j, and is expressed as

$$P_{ij}(t) = P\{Z(s+t) = j | z(s) = i\}. \tag{2.2}$$

The transition probability is useful for understanding the change of the system. Next, we also propose five basic assumptions to support our model and calculation.

(1) The individuals in this system are generated by a rate λ, which we suppose to be an intensity function of a homogeneous Poisson process. The process is described as $\{N(t), t \geq 0\}$, which denotes the number of individuals in the system at time t, often quoted as the scale of the system. The expectation of $N(t)$ is λt, which means we shall expect that λt individuals generate for each unit time. We make that assumption in order to adapt different accumulative models with different reactions to the time. In that case, the generation interval of new individuals is changeable depending on the external environment.

(2) We suggest that every selecting individual denoted as a variable X is totally random in the system, i.e. the individual is distributed uniformly in an interval $[0, \lambda t]$, i.e. $X \sim U(0, \lambda t)$, where t is the time recorder of the growing system and individuals are generated at rate λ.

(3) Then, we assume that the increment Y is a variable following a log-normal distribution with the parameters μ and σ, i.e. $\ln(Y) \sim N(\mu, \sigma^2)$, since there is evidence that most increments are distributed log-normally [27]. The benefits of this suggestion are that the log-normal distribution is non-uniform and non-monotonic which is a typical characteristic of increments, e.g. the income in economics. Besides, the log-normal variables are positive which follows the non-negativity of increments and simplifies the calculation. Considering all the benefits, the log-normal is better than the power-law and normal to describe increments.

(4) The allocating rule following the 'Matthew effect' is controlled by the probability,

$$P = \frac{z}{\sum_{i=1}^{n} z_i}, \tag{2.3}$$

where z is the accumulative variable. This probability ensures that the higher value of z is more likely to be allocated and in fact realizes that 'the rich get richer'.

(5) The last assumption says that the total accumulation is in a directed proportion to the increments, since the increments are allocated and collected to become the accumulation, and the proportion rate is α.

Significantly, the above definitions and assumptions influence our model and contribute to the derivation.

(b) Calculation of the accumulative probability model

In this subsection, we mainly deduce the accumulative distribution of our model.

Referring to the assumptions (a) and (d), we start calculating the accumulative distribution denoted as a variable Z. At a specific time t, the total accumulation is

$$\sum_{i=1}^{n} z_i = \alpha y \lambda t. \tag{2.4}$$

Based on the *Growing process* and assumption (b), we only consider that λ new individual generates each time and each one's increment is denoted as y. According to the *Allocating process* and assumption (c), we first assume that one portion of one individual's increment is distributed to an existing individual denoted as x, i.e. only one in y can be obtained by x, and we also know that λ individuals generate each time, therefore this distribution should be multiplied by λ. Obviously, the binomial distribution fits this situation, and the time tends to be large, consequently

$$P = \lambda \binom{y}{1} \left(\frac{z}{\sum_{i=1}^{n} z_i} \right) \left(1 - \frac{z}{\sum_{i=1}^{n} z_i} \right)^{y-1} \approx \frac{z}{\alpha t}. \tag{2.5}$$

This probability P explains not only the allocation of the new individual, more importantly, it denotes the growing probability of the existing individual x. We interpret it as the growing rate in a very short time.

Obviously, for the individual x, the probability of two or more increments in a time Δt is $o(\Delta t)$. Therefore, considering definition 2.1, $1 - P_{zz}(\Delta t)$, the probability that a process in the accumulation z at time 0 will not be in the accumulation z at time Δt, equals to the probability that a transition occurs within time Δt plus something small compared to t. Thus, we say,

$$1 - P_{zz}(\Delta t) = P_x \Delta t + o(\Delta t). \tag{2.6}$$

The probability $P_{zj}(\Delta t)$, the process goes from accumulation z to j in a time Δt, equals to the probability that a transition occurs in this time multiplied by the probability that the transition is into the accumulation j, plus something small compared to Δt, that is

$$P_{zj}(\Delta t) = P_x P_{z+1j} \Delta t + o(\Delta t). \tag{2.7}$$

However, noting that assumption (a), our model generates λ individuals in a very short time, and we take the shortest unit time Δt for the change of system is 1. Then, from equations (2.6) and (2.7), we get probabilities at a certain time t,

$$P_{zz}(t) = 1 - \frac{z}{\alpha t} \tag{2.8}$$

and

$$P_{z-1z}(t) = \frac{(z-1)}{\alpha t}. \tag{2.9}$$

On the basis of the total probability formula, equations (2.8) and (2.9), the accumulation probability at next time $t+1$ $P_z(t)$ is obtained by

$$P_z(t+1) = \frac{(z-1)}{\alpha t} P_{z-1}(t) + \left(1 - \frac{z}{\alpha} t \right) P_z(t). \tag{2.10}$$

We then change the order of equation (2.10), and obtain the difference equation

$$P_z(t+1) - P_z(t) = \frac{1}{\alpha t}[(z-1)P_{z-1}(t) - zP_z(t)]. \tag{2.11}$$

Since the differences of z and t are both 1, the equation can be regarded as a partial differential equation, that is

$$\frac{\partial P_z}{\partial t} = -\frac{1}{\alpha t}\frac{\partial zP_z}{\partial z} \tag{2.12}$$

whose left and right sides can be multiplied by z and integrated over z in the domain. Then we get

$$\frac{\partial \int_0^{+\infty} zP_z\,\mathrm{d}z}{\partial t} = -\frac{1}{\alpha t}\int_0^{+\infty} z\frac{\partial zP_z}{\partial z}\,\mathrm{d}z. \tag{2.13}$$

The definition of the expectant accumulation z_x and its integral transformation are

$$z_x = \int_0^{+\infty} zP_z(t)\,\mathrm{d}z = -\int_0^{+\infty} z\,\mathrm{d}z\frac{\partial[zP_z(t)]}{\partial z}. \tag{2.14}$$

Substituting equation (2.14) in (2.13), and considering one of the special solutions that an individual just arrives at the system, we yield the new differential equation

$$\left.\begin{aligned}\frac{\partial z_x(t)}{\partial t} &= \frac{z_x}{\alpha t}\\ \text{and}\qquad z_x\left(\frac{x}{\lambda}\right) &= y\end{aligned}\right\} \tag{2.15}$$

whose solution is

$$z_x(t) = y\left(\frac{\lambda t}{x}\right)^{1/\alpha}. \tag{2.16}$$

From equation (2.16), we know that $f(x)$ and $f(y)$ are the marginal probability densities for the probability density $f(z)$. Obviously, no matter which individual is selected, the increments are always uncertain, i.e. the increment Y is mutually independent of the individual X. Thus we have $f(z) = f(x)f(y)$. Considering the integration limit and assumptions (b) and (c), the cumulative distribution function $F(z)$ is derived as

$$\begin{aligned}F(z) &= \iint_{z_x(t)<z} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint_{x>\lambda t(y/z)^{\alpha}} f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_0^z\int_{\lambda t(y/z)^{\alpha}}^{\lambda t} \frac{1}{\sqrt{2\pi}\sigma y\lambda t}\,\mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_0^z \frac{1-(y/z)^{\alpha}}{\sqrt{2\pi}\sigma y}\,\mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}y.\end{aligned} \tag{2.17}$$

The probability density function $f(z)$ is the derivation of equation (2.17) leading to

$$\begin{aligned}f(z) = F'(z) &= \int_0^z \frac{\partial}{\partial z}\frac{1-(y/z)^{\alpha}}{\sqrt{2\pi}\sigma y}\,\mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}y\\ &= \frac{\alpha}{z^{\alpha+1}}\int_0^z \frac{y^{\alpha-1}}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}y.\end{aligned} \tag{2.18}$$

Considering both equations (2.17) and (2.18), the limitation of this probability is irrelevant to the time if the time is large enough. Therefore, referring to definition 2.1, the accumulative function is displayed as equation (2.18).

From the derivation, we successfully confirm the relationship among the accumulation, individuals and increments in this growing system. Their function defines that the accumulative distribution should be a probability distribution displayed as a function of two variables. The derivation also indicates that the inequality of the 'Matthew effect' will finally have an indirect influence on the boundary of the probability density function. By the principle of accumulative

behaviour, we can better understand and simulate real accumulative systems and discover the consequence of the action of the accumulation of increments and their unbalanced allocations to individuals.

(c) The variations of the accumulative law

Based on the previous description, the accumulative distribution depends on the fact that the increment variables follow a log-normal distribution which is described as the assumption (3) in §2a. However, the increments may be other potential probability models, even not a variable. Therefore, we hereby discuss an alternative form of the accumulative distribution.

(i) Constant increment

The first extension of the accumulative function is very special, since its increment remains unchanged instead of being a variable, i.e. y is a constant. In that case, based on the results in §2b, we have the cumulative distribution

$$F(z) = \int_{z_x(t)<z} f(x)\,\mathrm{d}x = \int_{x>\lambda t(y/z)^\alpha} f(x)\,\mathrm{d}x = 1 - \left(\frac{y}{z}\right)^\alpha. \tag{2.19}$$

Then, the probability density function is

$$f(z) = F'(z) = \frac{\alpha y^\alpha}{z^{\alpha+1}}. \tag{2.20}$$

Obviously, this function is the probability density of a Pareto distribution, also known as a power-law distribution. Thus we say, a Pareto distribution is a special case of an accumulative distribution following the accumulative law.

(ii) Exponential increment

In some situations, the increment may be memoryless, which means that we will not get any information about them over time. Therefore, we consider that the increment distributes exponentially as another extension of the accumulative function and assume that the increment Y is a variable following an exponential distribution with a rate parameter β, i.e. $Y \sim \exp(\beta)$.

In that case, the cumulative distribution is derived as

$$\begin{aligned} F(z) &= \int_0^z \int_{\lambda t(y/z)^\alpha}^{\lambda t} \frac{\beta}{\lambda t}\,\mathrm{d}x\,\mathrm{d}y = \int_0^z \left[1 - \left(\frac{y}{z}\right)^\alpha\right] \beta\,\mathrm{e}^{-\beta y}\,\mathrm{d}y \\ &= 1 - \frac{\alpha}{z^\alpha} \int_0^z y^{\alpha-1}\,\mathrm{e}^{-\beta y}\,\mathrm{d}y. \end{aligned} \tag{2.21}$$

Then, we have the probability density function

$$f(z) = F'(z) = \frac{\alpha}{\beta z^{\alpha+1}}(1 - \mathrm{e}^{-\beta z} - \beta z\,\mathrm{e}^{-\beta z}). \tag{2.22}$$

This accumulative distribution function is different from the original one, since the exponential distribution has a distinct property from the log-normal distribution. As a result, this exponential-based accumulative distribution is decreasing, which is a variation of the power-law distribution but more complicated.

In practical, the accumulative distribution and its extension distributions can describe many real phenomena, we will discuss it in §4.

3. Accumulative function and some statistical properties

The derived process of accumulative distribution is fuzzy, and in the practical situations, its definition and statistical properties are more serviceable. In this section, for a clear investigation of the accumulative distribution and a potential application on the statistical study on the real

systems, we display its strict definition, variant expression and some useful statistical properties. Particularly, we employ the original accumulative law, i.e. the increments are log-normally distributed.

On the basis of the previous research, we show the fundamental description of the accumulative probability as follow:

Definition 3.1. The function $f(x)$, $x \geq 0$ is called to be an accumulative distribution of a continuous random variable X having a rate parameter α, $\alpha > 0$, if

(1) $f(0) = 0$.
(2) The variable has independent increments following a log-normal distribution with parameters μ and σ.
(3) The probability of the accumulation of X is the probability density function, that is, for all $x \geq 0$

$$f(x) = \int_0^x \frac{\alpha y^{\alpha-1}}{\sqrt{2\pi}\sigma x^{\alpha+1}} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2} \mathrm{d}y. \tag{3.1}$$

The definition includes all conditions which an accumulative distribution has to follow and the precise probability density function. Condition (i) simply shows that the accumulation of the system starts at $t = 0$, and the condition (ii) is another significant description that the increments in the process of accumulation should be independent and log-normally distributed. Condition (iii) is the final mathematical expression of this probability distribution. To ensure the correctness of this expression, we next prove that it is a probability density function in statistics.

Theorem 3.2. *The function $f(x)$ described in definition* 3.1 *is a probability density function.*

Proof. In order to prove that $f(x)$ is a natural probability density function, we need to proof that this function is non-negative everywhere, and its integral over the entire space is equal to one.

Based on condition (i), for $x \leq 0$, $f(x) = 0$. Otherwise, by equation (3.1), $f(x) > 0$. Therefore, the non-negative feature holds.

For all y, having $0 < y < x$, thus

$$\begin{aligned}
\int_{-\infty}^{+\infty} f(x) &= \int_0^{+\infty} \int_0^x \frac{\alpha y^{\alpha-1}}{\sqrt{2\pi}\sigma x^{\alpha+1}} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2} \mathrm{d}y \, \mathrm{d}x \\
&= \int_0^{+\infty} \int_y^{+\infty} \frac{\alpha y^{\alpha-1}}{\sqrt{2\pi}\sigma x^{\alpha+1}} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2} \mathrm{d}x \, \mathrm{d}y \\
&= \int_0^{+\infty} \frac{1}{y\sqrt{2\pi}\sigma} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2} \mathrm{d}y \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-((\ln y-\mu)/\sigma)^2/2} \mathrm{d}\frac{\ln y - \mu}{\sigma} = 1.
\end{aligned} \tag{3.2}$$

the integration is equal to one.

This ends the proof. ∎

This proves that the accumulative distribution is beyond a function derived from our model, but technically a probability density function as well, which has the potential to describe and simulate many practical situations. Besides, the cumulative distribution function is also useful and expressed as follows.

Theorem 3.3. *The cumulative distribution function of an accumulative variable X is, for all $x \geq 0$,*

$$F(x) = \int_0^x \frac{x^\alpha - y^\alpha}{\sqrt{2\pi}\sigma y x^\alpha} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2} \mathrm{d}y. \tag{3.3}$$

Proof. The cumulative distribution function is the integration of the probability density function over $[0, +\infty]$. We change the order of the integration, that is

$$\begin{aligned} F(x) &= \int_0^x f(x)\,\mathrm{d}x = \int_0^x \int_y^x \frac{\alpha y^{\alpha-1}}{\sqrt{2\pi}\sigma x^{\alpha+1}} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^x \frac{x^\alpha - y^\alpha}{\sqrt{2\pi}\sigma y x^\alpha} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}y. \end{aligned} \tag{3.4}$$

The result follows. ∎

In statistics, our probability density function and cumulative distribution function may not present in a convenient form, thus we show its alternative statistical expression.

Remark 3.4. The probability density function of an accumulative distribution $f(x)$, $x <\geq< /0$, introducing $s = (\mu + \alpha\sigma^2 - \ln y)/\sqrt{2}\sigma$, is interpreted as follows:

$$\begin{aligned} f(x) &= \frac{\alpha}{x^{\alpha+1}} \int_{(\mu+\alpha\sigma^2-\ln x)/\sqrt{2}\sigma}^{+\infty} \frac{\mathrm{e}^{\alpha(\alpha\sigma^2+\mu-\sqrt{2}\sigma s)}}{\sqrt{\pi}\sigma} \mathrm{e}^{-(\alpha\sigma-\sqrt{2}s)^2/2}\,\mathrm{d}s \\ &= \frac{\alpha}{x^{\alpha+1}} \mathrm{e}^{\alpha(\mu+(1/2)\alpha\sigma^2)} \int_{(\mu+\gamma\sigma^2-\ln x)/\sqrt{2}\sigma}^{+\infty} \frac{1}{\sqrt{\pi}} \mathrm{e}^{-s^2}\,\mathrm{d}s \\ &= \frac{\alpha}{x^{\alpha+1}} \mathrm{e}^{\alpha(\mu+(1/2)\alpha\sigma^2)} \Phi\left(\frac{\ln x - \mu - \alpha\sigma^2}{\sigma}\right), \end{aligned} \tag{3.5}$$

where $\Phi(x)$ is a cumulative distribution function of the standard normal distribution.

Analogously, introducing another $s = (\mu - \ln t)/\sqrt{2}\sigma$, the cumulative distribution function is alternatively denoted as

$$\begin{aligned} F(x) &= \int_{(\mu-\ln x)/\sqrt{2}\sigma}^{+\infty} \frac{1}{\sqrt{\pi}} \mathrm{e}^{-t^2}\,\mathrm{d}t - \frac{1}{x^\alpha} \mathrm{e}^{\alpha(\mu+(1/2)\alpha\sigma^2)} \int_{(\mu+\gamma\sigma^2-\ln x)/\sqrt{2}\sigma}^{+\infty} \frac{1}{\sqrt{\pi}} \mathrm{e}^{-s^2}\,\mathrm{d}s \\ &= \Phi\left(\frac{\ln x - \mu}{\sigma}\right) - \frac{1}{x^\alpha} \mathrm{e}^{\alpha(\mu+(1/2)\alpha\sigma^2)} \Phi\left(\frac{\ln x - \mu - \alpha\sigma^2}{\sigma}\right). \end{aligned} \tag{3.6}$$

These alternative expressions of the accumulative distribution are much easier to calculate both probability density function and cumulative distribution function, making them useful in a practical situation.

For the furthermore application in statistics, we display some statistical properties of the accumulative distribution, e.g. expected value, standard deviation, etc. One of the centre properties, the expected value is a weighted average of all possible values, in our case, i.e. an integral of the variable with respect to its probability density. Thus, we have the following theorem.

Theorem 3.5. *The expected value of an accumulative variable is* $(\alpha/(\alpha-1))\,\mathrm{e}^{\mu+\sigma^2/2}$, *if* $\alpha > 1$.

Proof. Considering the continuity of the accumulative variable and the integration order, also introducing $t = (\ln y - \mu)/\sqrt{2}\sigma$, we have the expected value

$$\begin{aligned} E[x] &= \int_0^{+\infty} x f(x)\,\mathrm{d}x = \int_0^{+\infty} \int_y^{+\infty} \frac{\alpha y^{\alpha-1}}{\sqrt{2\pi}\sigma x^\alpha} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{\alpha}{\alpha-1} \int_0^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}y \\ &= \frac{\alpha}{\alpha-1} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-t^2/2+t\sigma+\mu}\,\mathrm{d}t \\ &= \frac{\alpha}{\alpha-1} \mathrm{e}^{\mu+\sigma^2/2}. \end{aligned} \tag{3.7}$$

The result follows. ∎

Apparently, if we let $\alpha \leq 1$, then the value of $E[x]$ is equal to $+\infty$. We say that the expected value no longer converges in this case. In other words, the expected value of an accumulative variable exists only if $\alpha > 1$. Another fact we can learn from the expression of the expected value is that it is positively correlated with the parameters μ and σ, while negatively correlated with α. Practically speaking, from the derivation, we know that the value of μ and σ is related to the value of the increment, and α indicates the proportion rate of the total accumulation to the increment. We can then infer that the more increments and its less proportion rate lead to a larger accumulation of an individual, and vice versa. The result perfectly follows equation (2.16) in §2, which displays the relationship of the accumulation variable, the increment variable and rate. Based on this, we may easily estimate the possible value of different accumulation distributions.

Based on the expected value which is also called the first moment, we can extend it to another centre statistical properties, the nth moment. The nth moment is known as a powerful tool to study the shape of our function, its expression is as follows:

Theorem 3.6. *The nth moment of an accumulative variable is* $(\alpha/(\alpha - n))\, \mathrm{e}^{n\mu+(1/2)n^2\sigma^2}$, *if* $\alpha > n$.

Proof. Analogously to theorem 3.5, and introducing $t = (\ln y - \mu)/\sqrt{2}\sigma$, we have then the expression of the nth moment as

$$\begin{aligned}\mu_n &= \int_0^{+\infty} x^n f(x)\,\mathrm{d}x = \int_0^{+\infty}\int_y^{+\infty} \frac{\alpha y^{\alpha-1}}{\sqrt{2\pi}\sigma x^{\alpha+1-n}}\,\mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{\alpha}{\alpha-n}\int_0^{+\infty} \frac{y^{n-1}}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-(\ln y-\mu)^2/2\sigma^2}\,\mathrm{d}y \\ &= \frac{\alpha}{\alpha-n}\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-t^2/2+nt\sigma+n\mu}\,\mathrm{d}t \\ &= \frac{\alpha}{\alpha-n}\,\mathrm{e}^{n\mu+n^2\sigma^2/2}. \end{aligned} \tag{3.8}$$

The result follows. ∎

From this derivation, we learn that the nth moment exists only if $\alpha > n$. Obviously, the moment is also positively correlated with the parameters μ and σ of the increment variable, while negatively related to the proportion rate of α. The moment shows a more general way to study the accumulative distribution, since we can easily get the skewness, kurtosis, and even higher-order moment from it.

Apart from the centre properties, we present a dispersion property, standard deviation, which measures the dispersion of a set of the accumulative data from the expected value.

Theorem 3.7. *The standard deviation of an accumulative variable is* $\sqrt{\alpha\,\mathrm{e}^{2\mu+\sigma^2}[(1/(\alpha-2))\mathrm{e}^{\sigma^2} - \alpha/(\alpha-1)^2]}$, *if* $\alpha > 2$.

Proof. By using theorems 3.5 ($n = 2$) and 3.6, the standard deviation is calculated as

$$\begin{aligned} SX &= \sqrt{E[(x - E[x])]^2} = \sqrt{\mu_2 - E[x]^2} \\ &= \sqrt{\alpha\,\mathrm{e}^{2\mu+\sigma^2}\left[\frac{1}{\alpha-2}\mathrm{e}^{\sigma^2} - \frac{\alpha}{(\alpha-1)^2}\right]} \end{aligned} \tag{3.9}$$

The result follows. ∎

Corollary 3.8. *The standard deviation of an accumulative variable for* $\alpha > 2$ *exits only if* $\sigma > \ln(\alpha(\alpha-2)/(\alpha-1)^2)$.

Proof. The existence of the standard deviation is related to the parameter σ. The reason is that the value under the square root should be larger than zero, thus

$$\alpha\,\mathrm{e}^{2\mu+\sigma^2}\left[\frac{1}{\alpha-2}\,\mathrm{e}^{\sigma^2} - \frac{\alpha}{(\alpha-1)^2}\right] > 0. \tag{3.10}$$

Obviously, it equals to

$$\frac{1}{\alpha-2}\,\mathrm{e}^{\sigma^2}-\frac{\alpha}{(\alpha-1)^2}>0. \tag{3.11}$$

Then we have

$$\sigma>\ln\frac{\alpha(\alpha-2)}{(\alpha-1)^2} \tag{3.12}$$

which is another condition that the standard deviation will exit.

The result follows. ■

Since the standard deviation, which measures how far a set of random numbers are spread out from their average value, uses the second moment and the value under the square root should be positive, it exists only if $\alpha>2$ and $\sigma>\ln(\alpha(\alpha-2)/(\alpha-1)^2)$. Different from the above statistical properties, the standard deviation is positively correlated with all parameters α, μ and σ, i.e. the larger increment and its proportion rate will let the value of the accumulative data more dispersed. Based on this result, the variance is $\alpha\mathrm{e}^{2\mu+\sigma^2}[(1/(\alpha-2))\,\mathrm{e}^{\sigma^2}-\alpha/(\alpha-1)^2]$, which is the square of the standard deviation.

Furthermore, we show the skewness of the accumulative distribution to measure the asymmetry of the distribution about its expected value.

Theorem 3.9. *The skewness of an accumulative variable is expressed as*

$$\frac{\frac{1}{\alpha-3}\,\mathrm{e}^{3\sigma^2}-\frac{3\alpha}{(\alpha-1)(\alpha-2)}\,\mathrm{e}^{\sigma^2}+\frac{2\alpha^2}{(\alpha-1)^3}}{\alpha^{1/2}[\frac{1}{\alpha-2}\mathrm{e}^{\sigma}-\frac{\alpha}{(\alpha-1)^2}]^{3/2}},\quad \textit{if } \alpha>3.$$

Proof. By the definition of Pearson's moment coefficient of skewness, theorems 3.6 (n=2 and 3) and (3.7), we have

$$\begin{aligned}\gamma&=E\left[\left(\frac{x-E[x]}{s}\right)^3\right]=\frac{E[x^3]-3E[x]E[x^2]+2E[x]^2}{s^3}\\&=\frac{\mu_3-3E[x]\mu_2+2E[x]^2}{s^3}\\&=\frac{\frac{1}{\alpha-3}\,\mathrm{e}^{3\sigma^2}-\frac{3\alpha}{(\alpha-1)(\alpha-2)}\,\mathrm{e}^{\sigma^2}+\frac{2\alpha^2}{(\alpha-1)^3}}{\alpha^{1/2}[\frac{1}{\alpha-2}\mathrm{e}^{\sigma}-\frac{\alpha}{(\alpha-1)^2}]^{3/2}}.\end{aligned} \tag{3.13}$$

The result follows. ■

Obviously, the skewness involves the third moment, we thus require $\alpha>3$ to let it exist. What we care about our distribution is that whether it has a positive or a negative skew, which appears as a left-leaning curve or a right-leaning curve. To address this issue, we have the following corollary.

Corollary 3.10. *The accumulative distribution for $\alpha>3$, if its standard deviation exists, has a positive skew.*

Proof. To prove the skew that an accumulative distribution has is positive, we only need to prove that its expression of skewness is greater than 0.

First, we consider the non-negativeness of its numerator as

$$\begin{aligned}&\frac{1}{\alpha-3}\,\mathrm{e}^{3\sigma^2}-\frac{3\alpha}{(\alpha-1)(\alpha-2)}\,\mathrm{e}^{\sigma^2}+\frac{2\alpha^2}{(\alpha-1)^3}>\frac{1}{\alpha-2}\,\mathrm{e}^{3\sigma^2}-\frac{1}{(\alpha-2)^2}\,\mathrm{e}^{\sigma^2}\\&>\frac{1}{\alpha-2}\,\mathrm{e}^{3\sigma^2}-\frac{1}{(\alpha-2)^2}\,\mathrm{e}^{3\sigma^2}=\frac{(\alpha-3)}{(\alpha-2)^2}\,\mathrm{e}^{3\sigma^2}>0.\end{aligned} \tag{3.14}$$

Then, for the denominator, referring to remark 2, we know that the standard deviation exists

$$\alpha^{1/2}\left[\frac{1}{\alpha-2}\,\mathrm{e}^{\sigma}-\frac{\alpha}{(\alpha-1)^2}\right]^{3/2}>0. \tag{3.15}$$

Overall, we apparently conclude that $\gamma > 0$, i.e. the distribution has a positive skew. The result follows. ■

From this corollary, we say that the accumulation is right-skewed and right-tailed, right refers to the right tail being drawn out and the mean being skewed to the right of a typical centre of the data. Then, we easily infer that those individuals holding a large value of the accumulation is the minority, and most of the individuals only possess a low value, which is a consequence of Matthew effect and also a significant property of our distribution.

So far, we show the most statistical properties of the accumulative distribution frequently used. Our distribution has mathematical expressions for them and displays its inimitable characteristic.

4. The accumulative law applied to real systems

The accumulative law and its probability model display their characteristics in very many fields, letting a phenomenon of the accumulative distribution very common. From the practical perspective, we specially present the modelling of complex networks studied in natural sciences, the accumulated wealth investigated by economists and the Internet traffic in information science in this section. All of them follow the accumulative law to create their systems and accumulatively distributed in certain ways.

(a) Complex network and degree distribution

Complex networks, as a typical representative of complex systems, show various topological characteristics. One of the most significant ones is the non-uniform distribution of degrees of vertices. Here, we mainly propose a complex network model based on the extended scale-free networks which have the property of the vertex growth, variable connections and preferential attachment resulting in a specific degree distribution. We will demonstrate how exactly the network follows the accumulative law and the degree distribution follows an accumulative distribution.

(i) Modelling process

First, we demonstrate the modelling process of improved scale-free networks.

Initialization: Assuming that the number of total vertices of the initial network is n. Each of them links to k neighbours and has the probability p to link to others, which is a small-world network.

Growth is a significant step for an improved scale-free network, which consists of the vertex growing rate and its connection to existing vertices. At each time t, we add λ vertices. In the interval $[t, t + \bigtriangleup t]$, the probability of the number of new vertices is then

$$P\{N(t+\delta t) - N(t) = k\} = \frac{(\lambda \delta t)^k}{k!} \mathrm{e}^{\delta t}, \quad k = 0, 1, \ldots \tag{4.1}$$

Besides, for each vertex, we connect m edges to the m different vertices already existing in the network, where m follows a log-normal distribution with parameters μ and σ.

Connection is simply linearly dependent on the degree of the target based on the knowledge of scale-free networks, $\phi(i)$, the probability of a connection to a vertex i, is denoted as

$$\phi(i) = \frac{k_i}{\sum_j k_j}, \tag{4.2}$$

where k_i is the degree of vertex i.

Termination is controlled by time t, which directly affects the scale of networks.

Obviously, the degree of vertices is the accumulative variable of a network, which is denoted by $p(k)$. From the modelling process of improved scale-free networks, we can apparently see that it follows the accumulation law. The *Growth* step and *Connection* step, respectively, interpret

Table 1. The mapping between the accumulation law and improved scale-free networks.

the terms in accumulation law	the terms in improved scale-free networks
individual x	vertex i
increment y	connection m
proportion rate α	equal to 2
probability P	probability ϕ
accumulation z	degree k
accumulative distribution $f(z)$	degree distribution $p(k)$

the *growing process* and *allocating process* of the law. Thus we say, the vertices are considered as independent individuals of a network, and the connections to existing vertices are increments. For a better understanding, we present the mapping terms between the complex networks and accumulation law (described in §2a) in table 1. From this table, we illustrate the practical description of a network corresponding to the abstract terms in the accumulation law. Thus the network turns into an accumulation model and its accumulative distribution of degree can be calculated by our method.

(ii) Degree distribution

We apply our method from §2b to calculate the degree distribution, and it is worth noting that the proportion rate of complex networks are equal to 2 since each edge has two vertices. Then, the differential equation of degree is

$$\left.\begin{aligned} \frac{\partial k_i(t)}{\partial t} &= \frac{k_i}{2t} \\ \text{and} \qquad k_i\left(\frac{i}{\lambda}\right) &= m. \end{aligned}\right\} \tag{4.3}$$

Its solution is

$$k_i = m\sqrt{\left(\frac{\lambda t}{i}\right)} \tag{4.4}$$

which shows the functional relationship among the degree, connection and vertex number.

Furthermore, the cumulative function is expressed as

$$P\{k_i(t) < k\} = \iint_{k_i(t)<k} f(i,m)\,\mathrm{d}i\,\mathrm{d}m = \int_0^k \frac{1-(m/k)^2}{m\sqrt{2\pi}\sigma}\,\mathrm{e}^{-(\ln m-\mu)^2/2\sigma^2}\,\mathrm{d}m. \tag{4.5}$$

and the degree distribution is

$$p(k) = P'\{k_i(t) < k\} = \frac{2}{k^3}\int_0^k \frac{m}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-(\ln m-\mu)^2/2\sigma^2}\,\mathrm{d}m. \tag{4.6}$$

Hence, the degree distribution is a special case of the accumulative distribution which verifies that these complex networks follow the accumulation law. Besides, the original scale-free networks with the constant connection m having $p(k) = 2mk^{-3}$, which follows a Pareto distribution, another special case of our accumulative distribution. It is a similar way to other different types of complex networks, since they essentially work by the accumulation law resulting in an accumulative distribution.

(b) Income allocation and personal wealth distribution

Income accumulation is a significant research issue in economics, which involves the allocation of wealth as the income and obtains the wealth distribution for each individual. In this subsection,

Table 2. The mapping between the accumulation law and income allocation model.

the terms in accumulation law	the terms in income allocation model
individual x	person p
increment y	income i
proportion rate α	equal to 1
probability P	probability ϕ
accumulation z	wealth w
accumulative distribution $f(z)$	personal wealth distribution $p(w)$

we use the accumulation law to manifest a wealth model based on Matthew's effect for the wealth allocation and calculate the personal wealth distribution as an accumulative distribution.

(i) Modelling process

The detailed process of a wealth allocation is shown as follows.

Initial Endowment: We assume that the wealth of each person is equal, and the number of initial people is small enough.

Expansion and Income directly affects wealth accumulation. The people in the model keeps growing at a certain rate of λ, the probability of new people joining in follows a Poisson distribution. The income denoted as i constantly distributes among the people, and the income itself follows a log-normal distribution with the parameters μ and σ.

Income Allocation based on the Matthew effect is controlled by

$$\phi(p) = \frac{w_p}{\sum_s w_s} \tag{4.7}$$

where w_p denotes the wealth of a person.

Termination executes when time is t, determining the scale of the model.

For the *Initial Endowment*, the endowment specifically means wealth. Like complex networks, the *Expansion and Income* and *Income Allocation*, respectively, interpret the *growing process* and *allocating process* of the law. Again, we present the mapping terms between the wealth model and accumulation law in table 2. From this table, we illustrate the practical description of the model corresponding to the abstract terms in accumulation law. Thus it can be considered as an accumulation model and its accumulative distribution of personal wealth can be calculated by our method.

(ii) Personal wealth distribution

Different from the network model, the proportion rate α of the income allocation model is 1. The reason is the division of income belongs to one person and one person only. Thus, we have the wealth increasing rate

$$\left.\begin{aligned} \frac{\partial w_p(t)}{\partial t} &= \frac{w_t}{t} \\ \text{and} \qquad w_p\left(\frac{p}{\lambda}\right) &= i \end{aligned}\right\} \tag{4.8}$$

and the solution is

$$w_p = i\left(\frac{\lambda t}{p}\right). \tag{4.9}$$

Furthermore, the personal cumulative function is

$$P\{w_p(t) < w\} = \iint_{w_p(t)<w} f(p,i)\,\mathrm{d}p\,\mathrm{d}i = \int_0^w \frac{w-i}{\sqrt{2\pi}\sigma wi}\,\mathrm{e}^{-(\ln i-\mu)^2/2\sigma^2}\,\mathrm{d}i. \tag{4.10}$$

Finally, the personal wealth distribution is

$$p(w) = P'\{w_p(t) < w\} = w^{-2} \int_0^w \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(\ln i - \mu)^2/2\sigma^2} \,\mathrm{d}i = w^{-2} \Phi\left(\frac{\ln w - \mu}{\sigma}\right), \tag{4.11}$$

where Φ is the cumulative probability distribution function of the normal distribution $N(0, 1)$.

Obviously, the personal wealth distribution is another special case of the accumulative distribution which verifies that the income allocation follows the accumulation law. Specifically, it is a simple case of the accumulative distribution when the proportion rate $\alpha = 1$ and has a strong relationship with the normal distribution.

(c) Internet traffic and website visit distribution

A reliable Internet traffic model is very necessary for service providers to properly maintain the quality of service. The internet traffic is allocated on different websites as the website visiting times, and for each website, it is accumulated to be the website traffic distribution. Specifically, this model is different from the models above, since the Internet traffic is obviously memoryless, e.g. the fever of some breaking news online will often rapidly decrease over time, and we will hardly get any information. Therefore, we employ the exponential distribution instead of the log-normal distribution in this model. We hereby utilize the accumulation law to display an Internet traffic model based on the Matthew effect for the traffic allocation and calculate the website traffic distribution as an accumulative distribution.

(i) Modelling process

The details of this modelling process are displayed as follows:

Initial Website is a small group highly gathered as the initialization of the Internet, ARPANET (Advanced Research Projects Agency Network). We choose m websites, and each one already has the same traffic.

Expansion and Website Visits step includes the growth of websites and the increasing visits allocated as visiting times to the websites in the model. Similarly, λ websites are introduced to the model for each time t. In particular, the visits v distribute among the websites following an exponential distribution with a parameter β.

Traffic Allocation based on the Matthew effect is expressed as

$$\phi(w) = \frac{T_w}{\sum_s T_s}, \tag{4.12}$$

where T_w denotes the traffic of web w.

Termination in time t has a direct impact on the size of the model.

For the *Initial Website*, we use the ARPANET model which can be ignored if the model is large enough. *Expansion and Website Visits* and *Traffic Allocation* respectively, interpret the *growing process* and *allocating process* of the law, but we employ the exponential distribution to express the increment considering the memoryless property of Internet traffic. The mapping terms between the Internet traffic model and the accumulation law are shown in table 3. From this table, we illustrate the practical description of the model corresponding to the abstract terms in the accumulation law. Thus it can be considered an accumulation model and its accumulative distribution of website traffic can be calculated by our method.

(ii) Website visit distribution

To calculate the website traffic distribution, we first assume the proportion rate $\alpha = 1$, since the visit can be obtained by one website and one website only. The traffic increasing rate is expressed

Table 3. The mapping between the accumulation law and Internet traffic model.

	the terms in accumulation law	the terms in Internet traffic model
	individual x	website w
	increment y	visit v
	proportion rate α	equal to 1
	probability P	probability ϕ
	accumulation z	traffic T
	accumulative distribution $f(z)$	website visit distribution $p(T)$

as

$$\left.\begin{aligned}\frac{\partial T_w(t)}{\partial t} &= \frac{T_t}{t}\\ \text{and}\qquad T_w\left(\frac{w}{\lambda}\right) &= v\end{aligned}\right\} \tag{4.13}$$

and the solution is

$$T_w = i\left(\frac{\lambda t}{w}\right). \tag{4.14}$$

Then, based on this solution, we calculate the cumulative distribution of website visit as

$$P\{T_w(t) < T\} = \iint_{T_w(t)<T} f(w,v)\,\mathrm{d}w\,\mathrm{d}v = 1 + \frac{\mathrm{e}^{-\beta T} - 1}{T\beta} \tag{4.15}$$

and the website visit distribution is

$$p(T) = P'\{T_w(t) < T\} = \frac{1}{\beta T^2}(1 - \mathrm{e}^{-\beta T} - \beta T\,\mathrm{e}^{-\beta T}). \tag{4.16}$$

As we see, the website visit distribution is a variation of the accumulative distribution showing that Internet traffic follows the accumulation law. Specifically, it is a special case of the accumulative distribution when the proportion rate $\alpha = 1$ and has an exponential increment as described in §2c(iii).

Above all, the accumulation law applies in various practical models along with its different forms of accumulative distribution. It clearly shows that the accumulation law and distribution are useful to understand and describe accumulative behaviours.

5. Simulation studies

Based on our theory, the accumulative law can be widely applied to the accumulative systems affected by the Matthew effect. Therefore, in this section, we present some data from extended scale-free networks, personal wealth and Internet traffic, then use the accumulative distribution, i.e. the degree distribution, the personal wealth distribution and website traffic distribution described in §4, respectively, to simulate these data. The goodness of fit of the real data and theoretical accumulative distribution is shown to verify the application of our law on fitting. All the simulations are run on Matlab 2010.

(a) Measure of goodness of fit

Before the simulations are operated, we first give an evaluation standard for the goodness of fit of our distribution and real data. Considering that both of the distributions should be discrete when they are compared, we introduce the relative entropy, also known as the Kullback–Leibler divergence, which is a measure of the difference between two random variables or probability distribution.

Specifically, the evaluation of the goodness of fit is expressed as

$$d = \sum_{x \in \chi} p(x) \log \frac{p(x)}{f(x)}, \tag{5.1}$$

where x is the accumulative value, $p(x)$ is the frequency or probability of the real data and $f(x)$ is the accumulative distribution.

Obviously, it is the expectation of the logarithmic difference between the probabilities p and f, where the expectation is taken using the probabilities p. If the value is very close to 0, it means the goodness of fit is high.

Besides the divergence, we also employ the Pearson product-moment correlation coefficient measuring the linear correlation between two variables. In our case, it measures the correlation between a real distribution and our model distribution.

The detailed expression is

$$\rho = \frac{E[(X - E(X))(Y - E(Y))]}{SXSY}, \tag{5.2}$$

where X denotes the real data, Y denotes our accumulative variable, E is the expectation function and S is the standard deviation.

The coefficients have a value between $+1$ and -1, where 1 is the best result of total positive linear correlation, 0 is no linear correlation, and -1 is a total negative linear correlation.

(b) Goodness of fit to degree distribution

The first simulation is to model complex networks and use the accumulative distribution to fit their degree distribution. The modelling process is referred to as §4a. For the initialization, we let $k = 2$, $p = 0.8$ and $n = 20$ to ensure that the initial network is small enough and highly clustered. For the rest part, we first show how the data of nonhomogeneous Poisson process of vertex generates as below:

(1) we let the values of input rate $\lambda = 50$ and termination time $t = 1000$, thus the network is large enough;
(2) we generate exponential distribution random values with λ, denoted as t_i, $i = \{1, 2, 3, \ldots\}$;
(3) if the cumulative time $T_i \leq t$, let $T_i = T_i + t_i$, else terminate and output the temporal series. Then we get the temporal series of arrival vertices.

Next, for the connection, we use the *lognrnd* function in Matlab to produce the number of connections. The result is rounded by the *round* function, and we use the roulette algorithm to realize the preferential connection. Then, we let $\lambda = 50$, $\mu = 1, 2$ and $\sigma = 1/2, 1$, the parameters are small for a better vision of the presented figures, thus we shall have four different networks. We use the association matrix to record the network degree and calculate the degree distribution.

To fit these three networks, we employ the accumulative distribution referring to equation (4.6) with the same parameters $\mu = 1, 2$, and $\sigma = 1/2, 1$. The results are shown in figure 1. We compare the first 150 values of both distributions, and for a clear illustration of the peak, we only show the first 50 values in these figures.

From these figures, the results of the goodness of fit to the degree distribution of complex networks with different parameters are good because the Kullback–Leibler divergence is very low (all below 0.1) and the correlation coefficient is very high (all above 95%). The detailed results are listed in table 4. Then, we can say that the accumulative distribution almost perfectly fit the degree distribution and the model we propose is correct from the perspective of simulation. From the results, we can also argue that the degree distribution of complex networks is not a simply power law, it should have a peak when the degree is very low once the connection is a variable. In other words, the power-law distribution only fits the specific complex networks as a specific case described in §2c(i), but our distribution can more widely fit complex networks with variable connections.

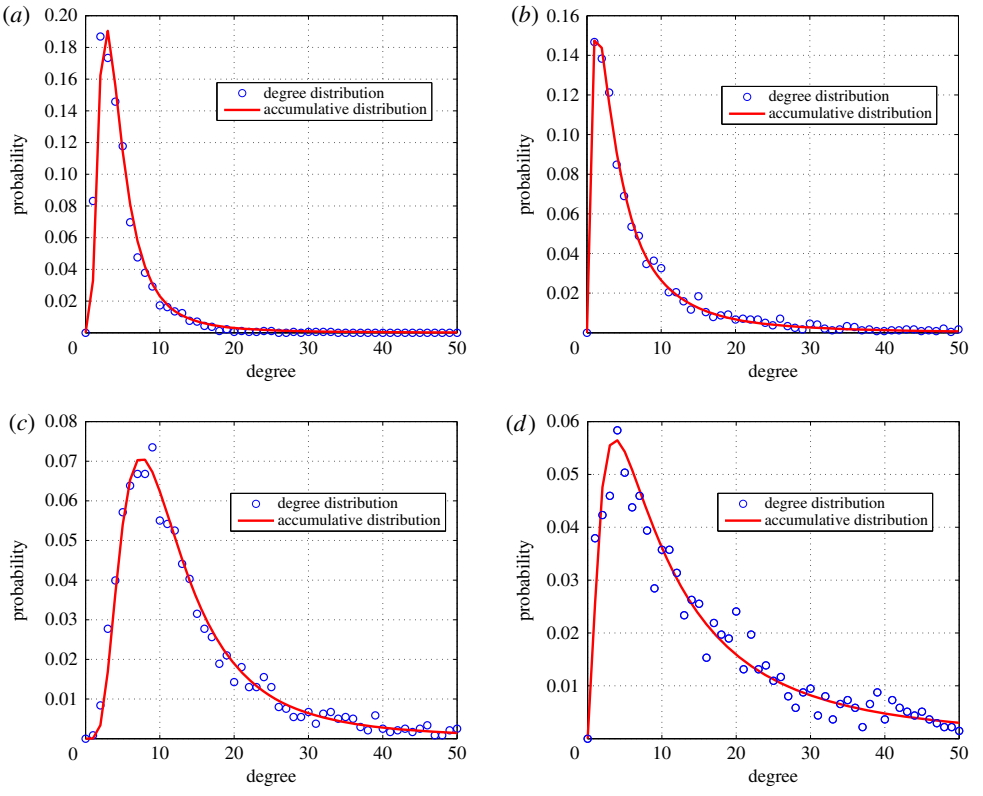

Figure 1. Goodness of fit to the degree distribution of complex networks with different parameters: the blue points are the degree distribution of complex networks while the red line is the simulated accumulative distribution. (*a*) $\mu = 1$ and $\sigma = 0.5$, (*b*) $\mu = 1$ and $\sigma = 1$, (*c*) $\mu = 2$ and $\sigma = 0.5$, (*d*) $\mu = 2$ and $\sigma = 1$. (Online version in colour.)

Table 4. The goodness of fit to the degree distribution of complex networks.

	parameter			
measure	$\mu = 1 \quad \sigma = 0.5$	$\mu = 1 \quad \sigma = 1$	$\mu = 2 \quad \sigma = 0.5$	$\mu = 2 \quad \sigma = 1$
d	0.0918	0.0632	0.0222	0.0535
ρ	95.17%	95.80%	98.74%	96.14%

(c) Goodness of fit to personal wealth of UK

The accumulative law suggests that the personal wealth distribution can be fit by the distribution function as equation (4.11). Therefore, in this simulation, we use the theoretical function to fit the personal wealth distribution of the United Kingdom.

It is common knowledge that people avoid reporting their total wealth routinely making it very difficult to collect personal wealth data and the statistical data are often quartered or more making it difficult to seek for a precise value for each wealth level. However, when a person dies, all assets must be reported for the purpose of inheritance tax. Using these data and an adjustment procedure, the British tax agency, the Inland Revenue, reconstructed wealth distribution of the whole UK population. In this simulation, we mainly use the UK residential buildings, other

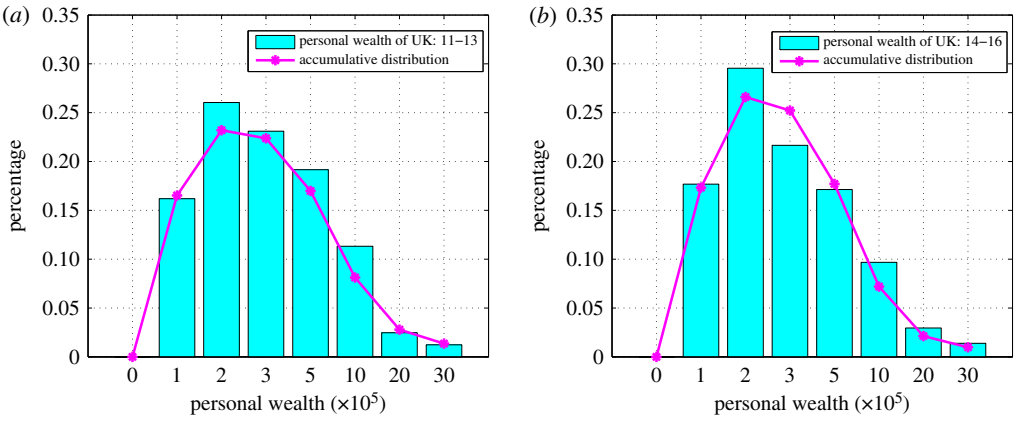

Figure 2. Goodness of fit to the personal wealth of UK from 2011 to 2013 and 2014 to 2016: the blue histogram is the distribution of personal wealth while the purple curve is the simulated accumulative distribution. (*a*) UK 2011 to 2013, (*b*) UK 2014 to 2016. (Online version in colour.)

buildings and land as the real identified estate which are closer to our model described in §4), for this estate, is slowly accumulated than spent from time to time.[1]

The datasets are collected from 2011 to 2013 and 2014 to 2016, we divide the people into seven levels of net estate, $ 0 to 100 000, 100 000 to 200 000, 200 000 to 300 000, 300 000 to 500 000, 500 000 to 1 000 000, 1 000 000 to 2 000 000, and 2 000 000 to 3 000 000, the maximum bound to represent each level, then the value of each level is 1, 2, 3, 5, 10, 20, 30, the unit is 10^5 pounds. Refer to the law of large numbers, we consider their frequencies as their probabilities, thus we can get the personal wealth percentage of UK as probability distribution during these two periods. As the data are sparse, the histogram is applied to illustrate the distribution. To fit these two wealth distribution, we estimate the parameters for equation (4.11), then let $\mu = 1.20, \sigma = 0.90$ for the data from 2011 to 2013 and $\mu = 1.00, \sigma = 0.78$ for the data from 2014 to 2016, and substitute each level into equation (4.11). The results are shown in figure 2.

Apparently, the tendency of our distribution fits the wealth distribution well. In detail, the goodness of fit to the personal wealth distribution is displayed at table 5, the Kullback–Leibler divergence is low and the correlation coefficient is high indicating that our distribution is very close to the real wealth distribution.

Thus, from this simulation, we show that our personal wealth model and its distribution can be applied to simulate the real personal wealth, and in this case of UK, the real personal wealth follows neither the Pareto distribution nor the log-normal distribution, but the general accumulative distribution, which may be the most potential distribution to describe the personal wealth and reduce the dispute over Pareto and log-normal distribution in economics.

(d) Goodness of fit to the website visit distribution

The last practical model we mention in §4c is the Internet traffic and website visit distribution. To verify this kind of model, we employ a real website visit distribution and fit by equation (4.16).

The dataset we use is retrieved from the US Government's open data, it provides web traffic statistics for the top 2000 most visited pages on nyc.gov in New York by month which is collected by the Department of Information Technology & Telecommunications (DoITT) from July 2015 to October 2016.[2] The number of websites is 32 015 in total, and a visit is a series of page views, beginning when a visitor's browser requests the first page from the server and ending when

[1]The dataset can be found at https://www.gov.uk/government/collections/distribution-of-personal-wealth-statistics.

[2]The dataset can be viewed at https://catalog.data.gov/dataset/nyc-gov-web-analytics.

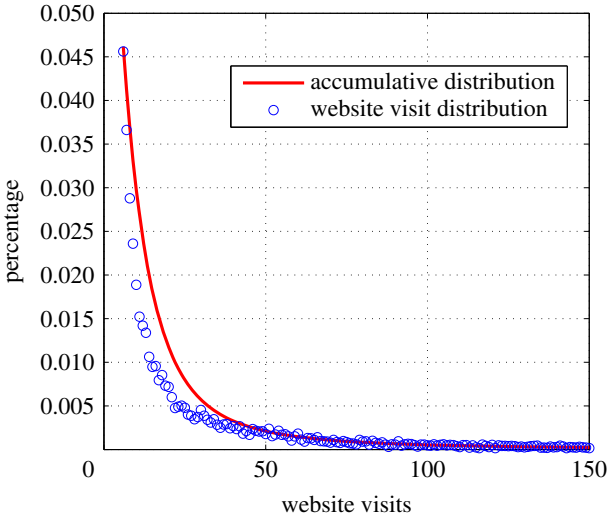

Figure 3. Goodness of fit to the website of nyc.ogve visit distribution from 2015 to 2016: the blue point denotes the distribution of website visits while the red curve denotes the simulated accumulative distribution. (Online version in colour.)

Table 5. The goodness of fit to the personal wealth of UK.

	parameter	
measure	2011 to 2013	2014 to 2016
d	0.0906	0.0390
ρ	99.17%	98.32%

the visitor leaves the site or remains idle beyond the idle-time limit. Then, in order to obtain the frequency of each website visit time, we sum up all the visit times as the total visit and let the time of each possible visit value from six times to 551941 divide the total visit. Refer to the law of large numbers, we consider their frequencies as their probabilities, thus we have the website visit percentage on nyc.gov as a probability distribution. We show the first 150 values of the website visit for a clear vision, the result is shown as the blue points in figure 3.

Based on the knowledge in §4c, referring to equation (4.16), we estimate the parameter $\beta = 1.90$ and get the accumulative distribution as shown in figure 3, the red curve. As we discussed before, this accumulative distribution has no peak and drops monotonously, which is close to the real visit distribution in tendency. In detail, the Kullback–Leibler divergence of these two distributions is 0.1143 and their correlation coefficient is 84.53%, which also shows that our distribution fits the real visit distribution well.

Again, the result shows that the accumulative distribution fits the website visit, i.e. the accumulative law can describe the behaviour of Internet traffic. However, from the tendency of the curve, we see that this specific accumulative distribution is more close to the Pareto distribution which has a long tail. The reason is that the increasing visit follows an exponential distribution with the property of memoryless. Therefore, it is an exceptional case of the accumulative distribution to fit the accumulation increasing memoryless instead of log-normally.

6. Conclusion and future work

In this paper, in order to deal with the controversy over the goodness of fit of the Pareto distribution and log-normal distribution in various research fields, we create a novel probability

model based on the accumulative law and calculate its probability density function as the accumulative distribution, which has the properties of both Pareto distribution and log-normal distribution making it superior to them and fit more real situations. The statistical properties of one specific accumulative distribution, e.g. *n*th moment, skewness, etc., are calculated to investigate its intrinsic factor. We show that the accumulative distribution and its variations have a wide range of applications to describe real systems, e.g. complex networks and degree distribution, income allocation and personal wealth distribution, Internet traffic and website visit distribution, solving the modelling and distribution issues in multiple disciplines. We display their modelling processes in detail, deduce the corresponding probability density function, and furthermore, fit them by real datasets which obtains in a convincing result.

In future work, this promising approach via the accumulative law should be dug deeply in other disciplines, and it should be used in more areas than fitting. Besides, the allocation rule in this paper mainly conducted based on Matthew's effect, different allocation methods, however, may result in different functions. Thus we should have numerous forms of the accumulative distribution. It also remains a fundamental challenge to study further the general theory and possible applications of the accumulative law.

Data accessibility. The source code is available at github.com/fengminyu1987/new.git.

Authors' contributions. M.F., L.-J.D., F.C., M.P. and J.K. designed and performed the research as well as wrote the paper.

Competing interests. We declare we have no competing interests.

Funding. M.F. is supported by the Fundamental Research Funds for the Central Universities (grant no. SWU019029). L.D. is supported by the National Science Foundation of China (grant no. 61702083). M.P. is supported by the Slovenian Research Agency (grant nos. J4-9302, J1-9112 and P1-0403).

Acknowledgements. M.F. express a special gratitude to his wife Qin Li for her encouragement and inspiration guiding him to the right way during those difficult times.

References

1. Estrada E. 2012 *The structure of complex networks: theory and applications*. Oxford, UK: Oxford University Press.
2. Newman ME. 2005 Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351. (doi:10.1080/00107510500052444)
3. Ryszard R, Jerzy T. 2014 Size distribution of gene families in a genome. *Math. Models Methods Appl. Sci.* **24**, 697–717. (doi:10.1142/S0218202513500644)
4. Arous GB, Hammond A. 2012 Randomly biased walks on subcritical trees. *Commun. Pure Appl. Math.* **65**, 1481–1527. (doi:10.1002/cpa.21416)
5. Reed WJ, Jorgensen M. 2004 The double Pareto-lognormal distribution. A new parametric model for size distributions. *Commun. Stat.* **33**, 1733–1753. (doi:10.1081/STA-120037438)
6. Qin C, Colwell LJ. 2018 Power law tails in phylogenetic systems. *Proc. Natl Acad. Sci. USA* **115**, 690–695. (doi:10.1073/pnas.1711913115)
7. Etro F, Stepanova E. 2018 Power-laws in art. *Physica A* **506**, 217–220. (doi:10.1016/j.physa.2018.04.057)
8. Stefano G, Giuseppe T. 2019 Human behavior and lognormal distribution. A kinetic description. *Math. Models Methods Appl. Sci.* **29**, 717–753. (doi:10.1142/S0218202519400049)
9. Hantson S, Pueyo S, Chuvieco E. 2016 Global fire size distribution: from power law to log-normal. *Int. J. Wildland Fire* **25**, 403–412. (doi:10.1071/WF15108)
10. González-Val R. 2019 Lognormal city size distribution and distance. *Econ. Lett.* **181**, 7–10. (doi:10.1016/j.econlet.2019.04.026)
11. O'Connor P, Kleyner A. 2012 *Practical reliability engineering*. New York, NY: John Wiley & Sons.
12. Xia F, Wang J, Kong X, Wang Z, Li J, Liu C. 2018 Exploring human mobility patterns in urban scenarios: a trajectory data perspective. *IEEE Commun. Mag.* **56**, 142–149. (doi:10.1109/MCOM.2018.1700242)
13. Barabási AL, Albert R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)
14. Barabási AL, Albert R, Jeong H. 1999 Mean-field theory for scale-free random networks. *Physica A* **272**, 173–187. (doi:10.1016/S0378-4371(99)00291-5)

15. Bollobás BE, Riordan O, Spencer J, Tusnády G. 2001 The degree sequence of a scale-free random graph process. *Random Struct. Algor.* **18**, 279–290. (doi:10.1002/rsa.1009)
16. Li L, Alderson D, Doyle JC, Willinger W. 2005 Towards a theory of scale-free graphs: definition, properties, and implications. *Internet Math.* **2**, 431–523. (doi:10.1080/15427951.2005.10129111)
17. Krioukov D, Papadopoulos F, Kitsak M, Vahdat A, Boguñá M. 2010 Hyperbolic geometry of complex networks. *Phys. Rev. E* **82**, 036106. (doi:10.1103/PhysRevE.82.036106)
18. Javarone MA, Armano G. 2013 Perception of similarity: a model for social network dynamics. *J. Phys. A* **46**, 455102. (doi:10.1088/1751-8113/46/45/455102)
19. Bu Z, Li H-J, Zhang C, Cao J, Li A, Shi Y. 2020 Graph K-means based on leader identification, dynamic game and opinion dynamics. *IEEE Trans. Knowledge Data Engng.* (doi:10.1109/TKDE.2019.2903712).
20. Stumpf MP, Wiuf C, May RM. 2005 Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl Acad. Sci. USA* **102**, 4221–4224. (doi:10.1073/pnas.0501179102)
21. Fang Z, Wang J, Liu B, Gong W. 2012 Double Pareto lognormal distributions in complex networks. In *Handbook of optimization in complex networks*, pp. 55–80. New York, NY: Springer.
22. Mansfield E. 1962 Entry, Gibrat's law, innovation, and the growth of firms. *Am. Econ. Rev.* **52**, 1023–1051.
23. Levy M. 2009 Gibrat's law for (all) cities: comment. *Am. Econ. Rev.* **99**, 1672–1675. (doi:10.1257/aer.99.4.1672)
24. Stanley MH, Amaral LA. 1996 Scaling behaviour in the growth of companies. *Nature* **379**, 804–806. (doi:10.1038/379804a0)
25. Rozenfeld HD, Rybski D, Andrade JS, Batty M, Stanley HE, Makse HA. 2008 Laws of population growth. *Proc. Natl Acad. Sci. USA* **105**, 18 702–18 707. (doi:10.1073/pnas.0807435105)
26. Perc M. 2014 The Matthew effect in empirical data. *J. R. Soc. Interface* **11**, 20140378. (doi:10.1098/rsif.2014.0378)
27. Clementi F, Gallegati M. 2005 Pareto's law of income distribution: evidence for Germany, the United Kingdom, and The United States. *Econophys. Wealth Distrib.*, pp. 3–14.