Physica A 606 (2022) 128059

Contents lists available at ScienceDirect

Physica A

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

Power laws, the Price model, and the Pareto type-2 distribution

Grzegorz Siudem [a,*], Przemysław Nowak [a], Marek Gagolewski [b,c]

[a] Warsaw University of Technology, Faculty of Physics, ul. Koszykowa 75, 00-662 Warsaw, Poland
[b] Deakin University, School of IT, Geelong, VIC 3220, Australia
[c] Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland

ARTICLE INFO

Article history:
Received 9 March 2022
Received in revised form 27 July 2022
Available online 27 August 2022

Keywords:
Price model
Pareto distribution
Power laws
Rich get richer
Complex networks
Citations

ABSTRACT

We consider a version of D. Price's model for the growth of a bibliographic network, where in each iteration, a constant number of citations is randomly allocated according to a weighted combination of the accidental (uniformly distributed) and the preferential (rich-get-richer) rule. Instead of relying on the typical master equation approach, we formulate and solve this problem in terms of the rank–size distribution. We show that, asymptotically, such a process leads to a Pareto-type 2 distribution with a new, appealingly interpretable parametrisation. We prove that the solution to the Price model expressed in terms of the rank–size distribution coincides with the expected values of order statistics in an independent Paretian sample. An empirical analysis of a large repository of academic papers yields a good fit not only in the tail of the distribution (as it is usually the case in the power law-like framework), but also across a significantly larger fraction of the data domain.

1. Introduction

Citing papers is a common way to express appreciation of someone else's research or to acknowledge the relevance thereof with regard to our own results [1]. Regardless of the true motivation [2], the what-to-cite decisions of individual authors can be averaged over the whole bibliographic networks. This way, we may study the underlying mechanisms governing the emergence of particular citation distributions. Understanding these is of interest not only to network science, but also to the entire research community, especially bearing in mind that citation counts are widely used as a (controversial [3,4]) proxy for articles', authors', and journals' impact.

One such mechanism is known as rich-get-richer, success-breads-success [5], the Matthew effect [6], or the preferential attachment rule [7]. It assumes that highly cited papers are most likely to receive even more citations in the future (for the possible bibliometric applications of this rule see [8,9]). Yet, recent research [10,11] into the origins of success in science and beyond highlights the role of other factors — such as chance. Some [12,13] claim that luck contributes more than the beneficiaries are eager to admit.

This leads to the question: how to measure the level of randomness in a citation network? It is noted in [14] that when citations follow the rich-get-richer rule, it is virtually impossible to distinguish between the merit- or non-merit-driven motivations. The literature knows several other approaches to modelling the structure of the networks in general [15],

* Corresponding author.
E-mail addresses: grzegorz.siudem@pw.edu.pl (G. Siudem), m.gagolewski@deakin.edu.au (M. Gagolewski).
URLs: http://if.pw.edu.pl/~siudem (G. Siudem), https://www.gagolewski.com (M. Gagolewski).

https://doi.org/10.1016/j.physa.2022.128059

and citation networks in particular, including those based on recursive searching, node duplicating, and fitness factors [16–20]. They usually employ a significant number of parameters which reduces their interpretability. What is more, they do not allow for a direct measurement of the level of accidentality, which is our intention herein.

For these reasons, in this paper, we shall consider a classical D. Price's model [21] which explicitly combines the two said mechanisms. We shall discuss a few methods for identifying the preferential-to-accidental ratio from real-world data. The Price model was studied in, amongst others, [22, Chap. 14] and frequently appears in the literature under different names and modifications [23–28] and in different contexts: e.g., analysis of resistance to random failures and intentional attacks in complex networks [23] or computation of longest paths in random graphs [29].

The most typical approach (e.g., [26]) to deriving the citation distribution and thus the preferential-to-accidental ratio in the Price model is via master equations (see [30]). In this work, however, we shall apply the rank–size (order statistics) approach which was inspired by our earlier work [28], where we studied citation vectors of individual scientists, i.e., in small scale, but with similar accidental and preferential contributors. Here we shall modify the model's boundary conditions so that we can focus on papers which have obtained a sufficiently large number of citations (e.g., 1, to avoid problems with computing and drawing on the log scale). Moreover, we shall consider citation networks in their entirety, i.e., study the model's asymptotic behaviour. We shall show that this leads to a well-known generalisation of the Pareto distribution [31–33] albeit with a new, very appealing parametrisation. This way we make an interesting addition to the list [31] of processes from which such a distribution emerges. Revealing the connection between the Price model and the Pareto-type 2 distribution will allow us to estimate the preferential-to-accidental ratio using some more statistically reliable methods than those previously mentioned in the literature.

This work is set out as follows. In Section 2 we introduce the rank distribution approach to the preferential-accidental attachment process, which allows us to establish the relation between the Price model and the Pareto-type 2 distribution. In Section 3 we discuss three methods for estimating the model parameters from data and quantify their usability. Further, in Section 4 we apply them on the DBLP [34] repository of computer science papers. Lastly, in Section 5 we discuss the implications of our findings and propose some directions for future research.

2. Rank distribution approach to Price's model

Let us consider a process where in every time step, a citation network grows by one new paper with $\delta \geq 0$ initial[1] citations. This new paper features $m - \delta$ references to the articles that already exist in the system:

1. $a = (1-\rho)(m-\delta)$ citations are allocated completely at random,
2. $p = \rho(m-\delta)$ citations follow the preferential attachment (i.e., rich-get-richer rule; compare [7]) rule,

with $a + p = m - \delta$ and $\rho \in (0, 1)$ representing the extent to which the rich-get-richer rule dominates over pure luck. Note that $m > \delta$ gives the total number of citations added into the system in every time step.

2.1. Exact solution

Let $X_k(t)$ denote the number of citations to the kth most cited paper at a time step t. We assume that $X_k(k-1) = \delta$ for every k, i.e., the kth publication enters the system with $\delta \in [0, m)$ citations.

According to the above description, the number of citations grows between the iterations as follows:

$$X_k(t) = \underbrace{X_k(t-1)}_{\text{previous value}} + \underbrace{\frac{a}{t}}_{\text{accidental income}} + \underbrace{p\,\frac{X_k(t-1) + \frac{a}{t}}{\delta + (t-1)m + a}}_{\text{preferential income}} = X_k(t-1)\frac{t}{t+\phi} + \frac{a}{t+\phi}, \tag{1}$$

where a and p were defined above and for brevity we denote:

$$\phi = -\frac{p}{m} = \rho\left(\frac{\delta}{m} - 1\right). \tag{2}$$

Let us iterate Eq. (1) in a similar manner as we did in [28]:

$$\begin{aligned} X_k(t) =& X_k(t-2)\frac{t-1}{t-1+\phi}\frac{t}{t+\phi} + \frac{a}{t-1+\phi}\frac{t}{t+\phi} + \frac{a}{t+\phi} = \underbrace{X_k(k-1)}_{=\delta}\frac{k\cdot(k+1)\cdot\dots\cdot t}{(k+\phi)\cdot(k+1+\phi)\cdot\dots\cdot(t+\phi)} + \\ &+ \frac{a\cdot(k+1)\cdot\dots\cdot t}{(k+\phi)\cdot(k+1+\phi)\cdot\dots\cdot(t+\phi)} + \frac{a\cdot(k+2)\cdot\dots\cdot t}{(k+1+\phi)\cdot\dots\cdot(t+\phi)} + \dots + \frac{at}{(t-1+\phi)(t+\phi)} + \frac{a}{t+\phi}. \end{aligned}$$

[1] Note that in [28], where we considered a single author's record and not the whole citation network, we assumed $\delta = 0$. Our more general setting might now mean, for example, that we become interested in a paper's existence not immediately after it has been published, but only when it has reached some level of maturity/recognisability/attention-worthiness. Furthermore, we abstract from the many possible origins of the citations that have emerged before the paper has appeared in the simulation.

This can be simplified further thanks to the notion of the Euler gamma function:

$$X_k(t) = \delta\frac{\Gamma(t+1)\Gamma(k+\phi)}{\Gamma(k)\Gamma(t+1+\phi)} + a\sum_{\ell=k}^{t}\frac{\Gamma(t+1)\Gamma(\ell+\phi)}{\Gamma(\ell+1)\Gamma(t+1+\phi)}$$
$$= \delta\frac{\Gamma(t+1)\Gamma(k+\phi)}{\Gamma(k)\Gamma(t+1+\phi)} - \frac{a}{\phi}\left(\frac{\Gamma(t+1)\Gamma(k+\phi)}{\Gamma(k+1)\Gamma(t+1+\phi)} - 1\right),$$

which leads to:

$$X_k(t) = \left(\delta + m\frac{1-\rho}{\rho}\right)\frac{\Gamma(t+1)}{\Gamma(k)}\frac{\Gamma(k+\phi)}{\Gamma(t+1+\phi)} - m\frac{1-\rho}{\rho}. \tag{3}$$

Let us stress that, in each iteration t, the above formula preserves the average number of citations, i.e., it holds:

$$\frac{1}{t}\sum_{k=1}^{t} X_k(t) = m. \tag{4}$$

2.2. Limiting case

Let us recall the Gautschi inequality (see Eq. (7) in [35]) which states that for any $k \geq 0$ and $\psi \in [-1, 0]$, it holds:

$$k^{\psi} \leq \frac{\Gamma(k+\psi)}{\Gamma(k)} \leq (k-1)^{\psi}. \tag{5}$$

Eq. (3) can be rewritten as a function of $y = k/t$. Applying the above inequality yields in the limit as $t \to \infty$:

$$X(y) := \lim_{t\to\infty} X_{yt}(t) = m\frac{\rho-1}{\rho} + \frac{m+\delta\rho-\rho m}{\rho}y^{\phi} = \left(\delta + m\frac{1-\rho}{\rho}\right)y^{-\rho(1-\delta/m)} - m\frac{1-\rho}{\rho}. \tag{6}$$

The inverse of X, denoted $S = X^{-1}$, is given for $x \geq \delta$ by:

$$S(x) = \left(\frac{x+m/\rho-m}{\delta+m/\rho-m}\right)^{1/\phi} = \left(1+\frac{x-\delta}{m/\rho-(m-\delta)}\right)^{1/\phi} = \left(\frac{m-\rho(m-\delta)}{\rho x+(1-\rho)m}\right)^{\frac{m}{\rho(m-\delta)}}. \tag{7}$$

and of course $S(x) = 1$ for $x < \delta$.

It is a strictly decreasing continuous function onto (0, 1]. Hence, we can treat it as some complementary cumulative distribution function (CCDF; also known as the tail distribution or the survival function). Generally, if a random variable Z has a CCDF S, then $S(x) = \Pr(Z > x)$. Its inverse $X = S^{-1}$ is referred to as the complementary quantile function.

From now on shall refer to S given by Eq. (7) as the CCDF of the δ-*truncated Price distribution* with parameters $m > \delta \geq 0$ and ρ. This way we honour the contributions of D. Price who, as we mentioned in the introduction, studied a similar model in [21].

2.3. Price meets Pareto

Let us re-express the CCDF given by Eq. (7):

$$S(x) = \left(1+\frac{x-\delta}{\lambda}\right)^{-\alpha} = \left(\frac{\lambda}{x-\delta+\lambda}\right)^{\alpha} \qquad (x \geq \delta), \tag{8}$$

where:

$$\alpha = \frac{m}{\rho(m-\delta)}\ (\alpha > 1) \quad \text{and} \quad \lambda = (m-\delta)(\alpha-1)\ (\lambda > 0). \tag{9}$$

This is nothing else than the standard parametrisation of the Pareto-type 2 distribution (e.g., [31]). In some sources, it is also called the Lomax distribution shifted by δ.

Hence, the above derivations form an interesting addition to the catalogue of processes from which the Pareto-type 2 distribution emerges; see [31] for an overview of other ones.

Furthermore, let us emphasise that in the special case of $\rho = 1$ (purely preferential attachment), Eq. (7) reduces to:

$$S_1(x) = \left(\frac{x}{\delta}\right)^{-\alpha} \qquad (x \geq \delta), \tag{10}$$

with: $\alpha = m/(m-\delta)$ and $\alpha > 1$, i.e., Eq. (8) with $\lambda = \delta$. This is known in the literature as the (type-1) Pareto distribution; see [31].

2.4. Back to the ranks

Despite our model's stemming from a deterministic setting, its asymptotic expansion gives a description of a whole population of papers, from which we can then pick items at random. It might be interesting to see how the distribution we have just derived relates to Eq. (3).

First, however, let us note (e.g., [36, Eq. 21.9]) that the expected value of a random variable $X \geq \delta \geq 0$ with a continuous CCDF S is given by the tail-sum formula:

$$\mathbb{E}[X] = \int_0^\infty S(x)\,dx = \delta + \int_\delta^\infty S(x)\,dx. \tag{11}$$

This is equivalent to:

$$\mathbb{E}[X] = \int_0^1 S^{-1}(y)\,dy. \tag{12}$$

In our case, where S^{-1} is given by Eq. (6), it holds $\mathbb{E}[X] = m$, which is consistent with our model's assumptions: a sample of randomly selected papers will have m citations on average (compare Eq. (4)).

Further, given an independent, identically distributed sample of random variables $X_1, \ldots, X_n$ following a CCDF S, let $X_{r:n}$ denote the rth order statistic, i.e., the rth smallest value therein. Hence, $X_{1:n}$ is the sample minimum and $X_{n:n}$ is the maximum.

Let us derive the formula for $\mathbb{E}[X_{r:n}]$. Of course, from Eq. (4.5.1) in [37] we get that for large n it approximately holds:

$$\mathbb{E}[X_{r:n}] \simeq 1 - S^{-1}\left(\frac{r}{n+1}\right). \tag{13}$$

But we can do better than this: let us provide an exact formula. Namely, knowing that the CCDF of the rth order statistic is (e.g., [37]):

$$S_{r:n}(x) = 1 - \sum_{j=r}^{n} \binom{n}{j} (1-S(x))^j (S(x))^{n-j}, \tag{14}$$

we obtain:

$$\begin{aligned}\mathbb{E}[X_{r:n}] &= \int_0^\infty S_{r:n}(x)\,dx = \delta + \int_\delta^\infty \left[\sum_{j=0}^{n}\binom{n}{j}(1-S(x))^j S^{n-j}(x) - \sum_{j=r}^{n}\binom{n}{j}(1-S(x))^j S^{n-j}(x)\right] dx \\ &= \delta + \sum_{j=0}^{r-1}\binom{n}{j}\int_\delta^\infty (1-S(x))^j S^{n-j}(x) dx.\end{aligned} \tag{15}$$

Applying the binomial theorem to $(1-S(x))^j$ and performing some elementary integration, we get:

$$\begin{aligned}\mathbb{E}[X_{r:n}] &= \delta + (m-\delta)\sum_{j=0}^{r-1}\left[\binom{n}{j}\sum_{\ell=0}^{j}\binom{j}{\ell}(-1)^\ell \tfrac{\delta-m+m/\rho}{\delta-m+(n-j+\ell)m/\rho}\right] \\ &= m\frac{\rho-1}{\rho} + (-1)^r\frac{m+\rho\delta-\rho m}{\rho}\frac{\Gamma(n+1)}{\Gamma(n-r+1)}\frac{\Gamma(-(n+\phi))}{\Gamma(-(n-r+\phi))}.\end{aligned} \tag{16}$$

Let us also note that:

$$\frac{\Gamma(-(n+\phi))}{\Gamma(-(n-r+\phi))} = (-1)^r\frac{\Gamma(n-r+1+\phi)}{\Gamma(n+1+\phi)}. \tag{17}$$

Combining the above yields:

$$\mathbb{E}[X_{r:n}] = m\frac{\rho-1}{\rho} + \frac{m+\rho\delta-\rho m}{\rho}\frac{\Gamma(n+1)}{\Gamma(n-r+1)}\frac{\Gamma(n-r+1+\phi)}{\Gamma(n+1+\phi)}. \tag{18}$$

Eq. (18) is identical to Eq. (3) with $k = n - r + 1$ and $t = n$. This observation strengthens the rationale behind our asymptotic expansion even further: the rank–size approach gives the expected values of the order statistics from any finite sample therefrom, including a very small one.

2.5. Price meets power laws

Newman in [38] refers to the CCDF of the Pareto-type 1 distribution given by Eq. (10) as a power law distribution (compare Eq. (4) therein). This is due to the fact that, when plotted on a double log-scale, $S_1(x)$ is a straight line with slope $-\alpha$.

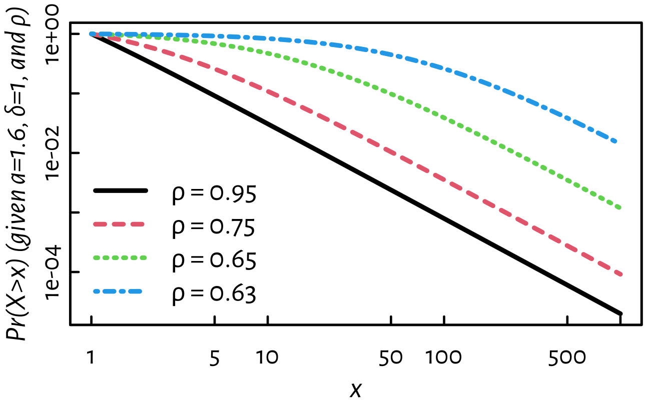

Fig. 1. CCDF of the Pareto-type 2 distribution for fixed $\alpha = 1.6$, $\delta = 1$, and different $\lambda = \frac{\delta(\alpha-1)}{\rho\alpha-1}$ for given ρs; note the double log-scale. The tail of each distribution follows the power law with the same exponent.

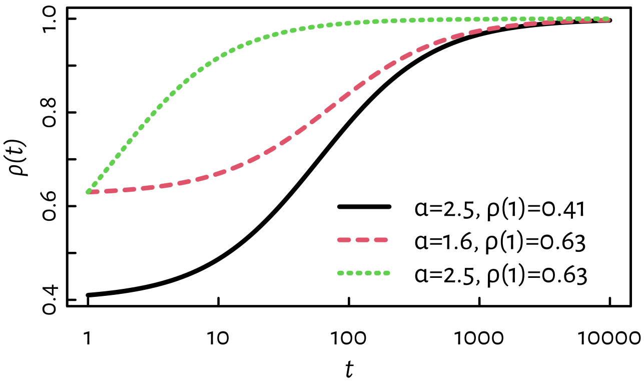

Fig. 2. ρ as a function of the truncation threshold t for different fixed αs and initial ρs at $\delta = 1$. It holds $\rho(t) \to 1$ as $t \to \infty$.

More generally, i.e., for any ρ, the Pareto-type 2 distribution follows a power law at least in its tail. Namely, the plot of $x \mapsto S(x)$ given by Eq. (8) on a double logarithmic scale looks like the plot of:

$$\ell \mapsto (-\alpha \log(\exp(\ell) - \delta + \lambda) + \alpha \log(\lambda)) \overset{\ell \text{ large}}{\simeq} -\alpha\ell + \alpha \log \lambda. \tag{19}$$

See Fig. 1 for an illustration. On a side note, if the slope of the CCDF's tail is $-\alpha$, then the slope the probability density function's tail is $-\alpha + 1$.

It is not rare in the data analysis practice to consider only a truncated version of a given dataset, and focus solely on its right tail; compare [38,39]. This is particularly the case when we are merely interested in describing the behaviour of "significant" objects in our database (excess), e.g., highly-cited papers.

For a fixed α, λ, and δ, let S be the CCDF of the corresponding Pareto-type 2 distribution, $\Pr(X > x)$. Consider any threshold $t > \delta$. Then, the CCDF of the truncated distribution, $\Pr(X > x | X > t)$ is given by:

$$S|_t(x) = \frac{S(x)}{S(t)} = \left(\frac{t - \delta + \lambda}{x - \delta + \lambda}\right)^{\alpha} = \left(\frac{\lambda'}{x - \delta' + \lambda'}\right)^{\alpha} \qquad (x \geq t). \tag{20}$$

It is thus still a Pareto-type 2 distribution with the same α, but with different $\lambda' = t - \delta + \lambda$ and $\delta' = t$.

The corresponding m' and ρ' can also be recreated:

$$m' = \frac{t\alpha - \delta + \lambda}{\alpha - 1} = m + (t - \delta)\frac{\alpha}{\alpha - 1}, \tag{21}$$

$$\rho' = \frac{t\alpha - \delta + \lambda}{\alpha(t - \delta + \lambda)}. \tag{22}$$

The linear increase of m' (as a function of t) is unsurprising. What *is* particularly remarkable, though, is that the ratio of preferentially to accidentally allocated citations increases as the cutoff threshold increases; see Fig. 2 for an illustration.

2.6. Closer to real-world data: Discretisation

Typically, real-world empirical data are not continuous (for instance: raw citation counts). We should therefore study a discretised version of our distribution as well.

Let us assume that data come from the original distribution, X, but what we observe is its truncated version, $\lfloor X \rfloor$, being the greatest integer less than or equal to X. The CCDF of such a random variable is simply given by:

$$\Pr(\lfloor X \rfloor > x) = \Pr(X > \lfloor x \rfloor + 1) = S(\lfloor x \rfloor + 1). \tag{23}$$

Remark 1. On a side note, in the limit as $N \to \infty$, for $\delta = 0$, the master equation approach (see [30]) applied to the same random process as above, yields the probability mass function $\underline{p}(x)$ for any $x \in \mathbb{N}_0$ given by:

$$\underline{p}(x) = \frac{1}{m+1-\rho m} \frac{\Gamma(\frac{m}{\rho} - m + 1 + \frac{1}{\rho})}{\Gamma(\frac{m}{\rho} - m)} \frac{\Gamma(x + \frac{m}{\rho} - m)}{\Gamma(x + \frac{m}{\rho} - m + 1 + \frac{1}{\rho})} = \frac{1}{m+1-\rho m} \frac{(m/\rho - m)_{1+1/\rho}}{(x + m/\rho - m)_{1+1/\rho}}, \tag{24}$$

where $(a)_x = \Gamma(a+x)/\Gamma(a)$ denotes the Pochhammer symbol. Let us compare the distribution-domain results via the master equation approach, with the CCDF-domain rank approach considered in this paper. For this purpose, let us compute the CCDF from the above distribution. And thus, it may be shown that for any integer k:

$$\sum_{\ell=0}^{k} \underline{p}(\ell) = 1 - \frac{\Gamma\left(-m + \frac{m}{\rho} + \frac{1}{\rho}\right)}{\Gamma\left(-m + \frac{m}{\rho}\right)} \frac{\Gamma\left(k - m + \frac{m}{\rho} + 1\right)}{\Gamma\left(k - m + \frac{m}{\rho} + \frac{1}{\rho} + 1\right)}. \tag{25}$$

Hence, the complementary cumulative distribution function of this distribution is defined for any $x \geq 0$ as:

$$\underline{S}(x) = \sum_{\ell=\lfloor x \rfloor + 1}^{\infty} p(\ell) = \frac{(m/\rho - m)_{1/\rho}}{(\lfloor x \rfloor + 1 + m/\rho - m)_{1/\rho}}. \tag{26}$$

By the aforementioned Gautschi inequality, our discretisation leads to:

$$S\left(\lfloor x \rfloor + 1 \middle| \delta = 0\right) = \frac{(m/\rho - m)^{1/\rho}}{(\lfloor x \rfloor + 1 + m/\rho - m)^{1/\rho}} \simeq \underline{S}(x), \tag{27}$$

which gives a nice correspondence between the master equation- and the rank distribution-approach.

3. Estimating model parameters

Fitting different heavy-tailed distributions to real-world data is quite commonly exercised in the complex systems practice. In particular, [38,40] feature a comprehensive discussion related to the fitting of the power law distributions.

3.1. All parameters are interpretable

The model we have landed at is in fact a Pareto-type 2 distribution, which by itself has of course been studied extensively; see [31]. However, in this paper we have arrived at a non-classic parametrisation thereof: all its parameters are easily interpretable. Namely, δ gives the cut-off (distribution shift), m is the expected number of citations of a randomly selected paper (and also the number of citations allocated per iteration), whereas ρ gives the ratio of preferentially to accidentally attached citations.

The truncation threshold $\delta \geq 0$ is usually known a priori; we therefore consider it fixed. For instance, we might only be recording the incidence of extreme values ($\Pr(X > x | X > \delta)$; compare [39]). Moreover, if data modellers wish to investigate only what is happening in the tail of the distribution, they trim the observations themselves (compare, e.g., [38,40] where only large-enough cities, the richest Americans, the most popular surnames, etc., are taken into account in the analysis). Otherwise, just like in the case of the power law distribution in [38], we can estimate it by simply taking the smallest value in the sample.

We should thus be interested in studying the properties of various basic estimators of m and ρ so that we can make sure that we are able to fit our model reliably to empirical data.

3.2. Different estimators of m and ρ

From a continuous CCDF we can easily obtain the underlying probability density function:

$$f(x|m, \rho) = \frac{d(1 - S(x|m, \rho))}{dx}. \tag{28}$$

This enables us to compute the maximum likelihood estimator (MLE) of the model parameters, being the solution to:

$$\max_{m,\rho} \sum_{i=1}^{n} \log f(X_i|m, \rho), \tag{29}$$

However, instead of solving directly with respect to m and ρ, we can consider the well-known, computationally less complicated maximum likelihood estimator of λ, which is determined by numerically solving the following nonlinear equation:

$$\frac{1}{n}\sum_{i=1}^{n} \log(x_i + \lambda - \delta) + 1 - \log\lambda - \frac{n}{\lambda \sum_{i=1}^{n}(x_i + \lambda - \delta)^{-1}} = 0. \tag{30}$$

Then we get the MLE of α:

$$\alpha = \left(\frac{1}{n}\sum_{i=1}^{n} \log(x_i + \lambda - \delta) - \log\lambda\right)^{-1}. \tag{31}$$

From this we obtain:

$$m = \frac{\lambda}{\alpha - 1} + \delta, \tag{32}$$

$$\rho = \frac{m}{\alpha(m - \delta)}. \tag{33}$$

For the sake of comparison, we will consider two other estimators based on the empirical CCDF,

$$\hat{S}(x|X_1, \ldots, X_n) = \frac{|\{i : X_i > x\}|}{n}. \tag{34}$$

The first estimator minimises the maximal difference between the empirical and the theoretical CCDF (from now on called SUP):

$$\min_{m,\rho} \max_{x\in\{X_1,\ldots,X_n\}} \left|\hat{S}(x|X_1, \ldots, X_n) - S(x|m, \rho)\right|, \tag{35}$$

and the second one minimises the sum of squared errors (SSE) at all points of discontinuity of the empirical CCDF:

$$\min_{m,\rho} \sum_{x\in\{X_1,\ldots,X_n\}} \left(\hat{S}(x|X_1, \ldots, X_n) - S(x|m, \rho)\right)^2. \tag{36}$$

The estimators will be computed using the Nelder–Mead method with a logarithmic barrier (via `constrOptim` in R) to enforce the constraints on m and ρ. We observed that if the algorithm starts from m equal to the 90% trimmed mean and $\rho = 0.5$, then it always converges nicely.

Similarly, we will also consider the MLE, SSE, and SUP estimators for the discretised version of the original random variable, $\lfloor X \rfloor$, with the MLE being the (numerical) solution to:

$$\max_{m,\rho} \sum_{i=1}^{n} \log p(X_i|m, \rho), \tag{37}$$

where the probability mass function p is given by $p(x) = S(\lfloor x \rfloor) - S(\lfloor x \rfloor + 1)$. Note that this time we are optimising the likelihood directly with respect to m and ρ.

3.3. Assessing estimator quality

Thanks to the principle of inverse transform sampling and the exact formula for the inverse of the CCDF given by Eq. (6), i.e., S^{-1}, we can easily generate realisations of independent samples like $X_1, \ldots, X_n$ following our distribution for specific δ, m, and ρ. Namely, we set $X_i = S^{-1}(U_i)$, where U_i is uniformly distributed on the unit interval.

To study the quality of the estimators via the Monte Carlo approach, we need to generate many samples, where each sample consists of $n \in \{1000, 100{,}000\}$ observations from the Pareto-type 2 distribution with $\delta = 1$, $m = 25$, and a range of ρs from the set $\{0.5, 0.75, 0.9\}$.

From the evaluation perspective and for the purpose of this study, the choice of δ is not really important, because we have already observed that a truncated Pareto-type 2 distribution is still a Pareto-type 2 distribution. Moreover, m scales linearly with δ, therefore the results we obtain are expected to be quite representative of other parameters sets as well.

Let us first consider the case of $n = 100{,}000$. This indicates that we are interested in a large-sample behaviour of the estimators.

Table 1 gives the approximate bias and the root mean squared error of the various estimators of the m parameter. Recall that for a given estimator $\hat{m}(X_1, \ldots, X_n)$ (which is a function of a sequence of random variables) of the true parameter

Table 1
Bias (and root mean squared error in parentheses) of different estimators of the m parameter (true $n = 100{,}000$, $\delta = 1$, $m = 25$, and different ρ; 10,000 Monte Carlo samples); the values that the t-test deems not significantly different from 0 (at the 0.01 significance level) are greyed out (in theory, the mean is an unbiased estimator of m).

	$\rho = 0.50$		$\rho = 0.75$		$\rho = 0.90$	
Continuous data						
Mean	−0.00	(0.27)	−0.12	(5.86)	−1.67	(47.10)
SSE	0.01	(0.25)	0.02	(0.58)	0.10	(1.36)
SUP	0.01	(0.29)	0.03	(0.65)	0.11	(1.51)
MLE	0.00	(0.18)	0.02	(0.41)	0.06	(0.99)
Discretised data						
SSE	0.00	(0.18)	0.02	(0.43)	0.07	(1.07)
SUP	0.01	(0.26)	0.02	(0.56)	0.09	(1.26)
MLE	0.00	(0.18)	0.01	(0.41)	0.03	(0.99)

Table 2
Bias (and root mean squared error in parentheses) of different estimators of the ρ parameter (the same simulation set-up as in Table 1).

	$\rho = 0.50$		$\rho = 0.75$		$\rho = 0.90$	
Continuous data						
SSE	0.0001	(0.0072)	0.0001	(0.0074)	0.0001	(0.0065)
SUP	0.0001	(0.0081)	0.0001	(0.0081)	0.0001	(0.0071)
MLE	0.0001	(0.0048)	0.0001	(0.0052)	0.0001	(0.0047)
Discretised data						
SSE	0.0001	(0.0050)	0.0001	(0.0056)	0.0001	(0.0052)
SUP	0.0001	(0.0072)	0.0001	(0.0071)	0.0001	(0.0060)
MLE	0.0001	(0.0048)	0.0000	(0.0052)	−0.0001	(0.0048)

m (which is a fixed value), its bias is defined as $\mathbb{E}[\hat{m}(X_1, \ldots, X_n) - m]$. Moreover, the root mean squared error (RMSE) is given by $\sqrt{\mathbb{E}[(\hat{m}(X_1, \ldots, X_n) - m)^2]}$.

We estimate the bias and RMSE based on $M = 10{,}000$ (for such a large sample, the computations took a few hours to compute on a modern PC) independent Monte Carlo samples of size n like $x_1^{(i)}, \ldots, x_n^{(i)}$, $i = 1, \ldots, M$, using, respectively, the sample mean and the standard deviation of a vector $(e_1, \ldots, e_M)$ with $e_i = \hat{m}(x_1^{(i)}, \ldots, x_n^{(i)}) - m$. In other words, we generate M pseudorandom vectors of size n from the Pareto-type 2 distribution with parameters m, ρ, δ, estimate the model parameters for each data vector, compare the estimates to the true values, and then aggregate the results.

The one sample Student t-test with a significance level of 0.01 indicates that for smaller ρs, the e_i values are, on average, not significantly different from 0. This indicates that the estimators might be asymptotically unbiased (note again the large n we use).

It might be tempting to use the methods of moments estimator for the m parameter, i.e., the arithmetic mean. Even though it is an unbiased estimator, it unfortunately tends to have very high variance and hence it is a practically useless measure. We note that $\alpha > 1$ guarantees the existence of the expected value (which always holds in our case), but the variance is only defined if $\alpha > 2$, which for $m = 25$ and $\delta = 1$ holds whenever $\rho < 25/48 \simeq 0.521$ (but it is not the case in our empirical study in Section 4).

Both in the continuous and in the discretised case, the MLE estimators work very well. The discretised case seems even more well-behaving. The MLE estimator should definitely be chosen if we suspect that data might really come from the distribution studied herein. Nevertheless, any deviations from the model (such as data contamination) might affect its performance. In such a case, the SSE estimator could also be noteworthy. SUP, on the other hand, is in theory a consistent (i.e., with asymptotic guarantees; compare the Glivenko–Cantelli theorem) estimator, but in our case it has a higher root mean squared error.

Table 2 provides the approximations to the bias and root mean squared error of the different estimators of ρ. In all cases, the bias is close to 0 and the root mean squared error is quite low.

Unfortunately, for smaller ns, the quality of the estimators deteriorates, particularly in the case of the m parameter, see Tables 3 and 4. In the case of larger ρ, the SUP estimator of m actually becomes quite trustworthy.

We should notice that there are of course many other estimators of α and λ (which we could use for estimating m and ρ) for the continuous case known in the literature, e.g., [41–44]. Many of them are better than MLE for samples of smaller sizes. However, to the best of our knowledge, there are no discrete counterparts thereof. This is why we restrict ourselves only to the three above methods. By the time better methods are developed (which is left for further research), we suggest that for small sample sizes, the estimated values should be used with caution.

Table 3
Bias (and root mean squared error in parentheses) of different estimators of the m parameter (true $n = 1000$, $\delta = 1$, $m = 25$, and different ρ; 10,000 Monte Carlo samples).

	$\rho = 0.50$		$\rho = 0.75$		$\rho = 0.90$	
Continuous data						
Mean	0.00	(2.86)	0.05	(49.53)	−0.62	(269.32)
SSE	0.28	(2.67)	1.97	(9.89)	16.51	(50.57)
SUP	0.39	(3.06)	2.51	(10.90)	11.16	(31.17)
MLE	0.11	(1.80)	0.80	(4.91)	0.57	(1292.72)
Discretised data						
SSE	0.13	(2.01)	0.90	(5.61)	10.29	(39.31)
SUP	0.31	(2.72)	1.80	(8.51)	10.87	(31.69)
MLE	0.11	(1.80)	0.80	(4.91)	7.39	(27.40)

Table 4
Bias (and root mean squared error in parentheses) of different estimators of the ρ parameter (the same simulation set-up as in Table 3).

	$\rho = 0.50$		$\rho = 0.75$		$\rho = 0.90$	
Continuous data						
SSE	−0.0012	(0.0711)	−0.0020	(0.0721)	−0.0055	(0.0600)
SUP	−0.0011	(0.0798)	−0.0025	(0.0795)	−0.0087	(0.0633)
MLE	−0.0019	(0.0475)	−0.0019	(0.0517)	−0.0025	(0.0469)
Discretised data						
SSE	−0.0030	(0.0546)	−0.0037	(0.0574)	−0.0052	(0.0503)
SUP	−0.0022	(0.0721)	−0.0027	(0.0705)	−0.0053	(0.0564)
MLE	−0.0019	(0.0475)	−0.0020	(0.0518)	−0.0033	(0.0461)

4. Analysis of empirical data

Let us fit our model to an example data set. We shall study the DBLP v12 bibliography database [34] which covers over 4,894,081 papers and 45,564,149 citation relationships. DBLP includes the most important outlets related to computer science. As the data consist of natural numbers, we will only fit the discretised versions of the MLE, SSE, and SUP estimators.

Fig. 3 depicts how the estimated m and ρ depends on the cutoff threshold δ.

First, let us note that for $\delta < 20$, the discrete Kolmogorov–Smirnov test rejects the hypothesis of Paretianity for all the estimated parameter pairs (at significance level of 1%). The SUP estimator uses the same criterion as this goodness-of-fit test, hence this is the method that recognises some Pareto-type 2 distribution the fastest. For MLE, we accept the Paretianity hypothesis for all $\delta \geq 29$, and in the case of SSE, this happens for all $\delta \geq 39$.

Furthermore, for all $\delta \geq 162$, the SUP-based fitting of the Pareto-type 1 model (pure power law, i.e., with $\rho = 1$) becomes legitimised by the Kolmogorov–Smirnov test. This is also where the quality of our ρ estimators deteriorates, which is of course an expected behaviour, because here the simpler model becomes applicable.

Furthermore, as δ increases, the size of the sample drastically decreases. For $\delta = 20$, we have $n = 803{,}721$. For $\delta = 100$, it holds $n = 135{,}982$. And for $\delta = 1000$, we already get $n = 4437$. This also affects the quality of the obtained estimates.

To conclude, our model enables a good description of the data over a large part of the domain, namely, for $\delta \in [\sim 20, \sim 162]$. We note that all the three estimators yield quite similar parameter values, which might indicate that they are well behaving in this interval. The behaviour of m and ρ as a function of the cutoff threshold is consistent with what we predict from Eq. (21).

Let us take a closer look at the fitted distributions in the $\delta = 20$ case (average number of citations equal to 82.97). Fig. 4 presents the CCDFs and Fig. 5 depicts the logarithmically binned histogram as well as the fitted probability mass functions. In terms of the CCDF reproduction accuracy, the quality of the MLE-fitted model ($m = 87.3534102$, $\rho = 0.8406532$) is somewhere in-between the SUP ($m = 90.0302823$, $\rho = 0.8545986$) and the SSE ($m = 87.0142624$, $\rho = 0.8354341$) ones. The estimated parameter values are similar, though.

Note that an inexperienced eye might be tempted to conclude that in the middle subfigure of Fig. 4, the models are not fitted well. However, we should emphasise that this graph utilises the double log-scale. Hence, the impressions of error magnitudes in the tails are exaggerated. Notably, even for simulated data (that come from the true distribution), such order statistics (close to the maximum) are subject to high variability. Therefore, they will from time to time deviate considerably from the theoretical curve. It seems, though, that our (asymptotic) model expects the most highly cited papers in the (finite) sample to be cited more frequently than they really are.

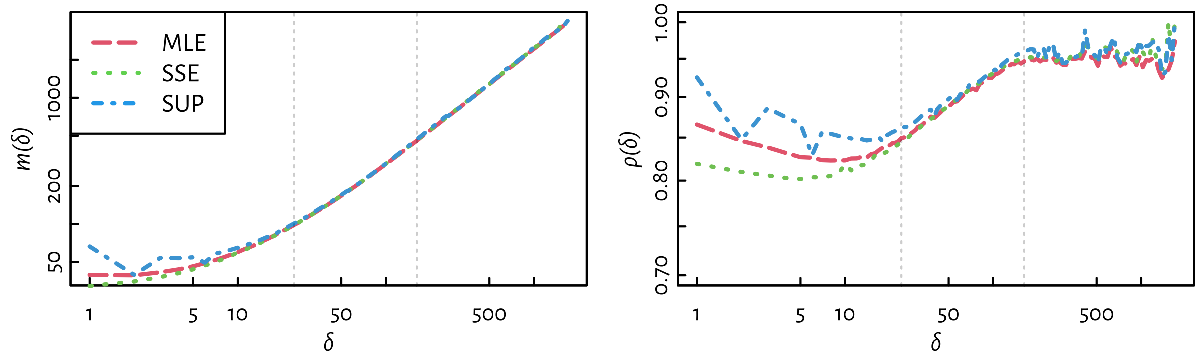

Fig. 3. Estimated parameters m (left), ρ (right) as functions of the cut-off threshold δ for DBLP data and different estimators. Minimal threshold for which the hypothesis of Paretianity is not rejected is equal to $\delta = 20$ (for the SUP estimator at significance level $\alpha = 1\%$). For $\delta \geq 162$, the estimated ρs become degenerated, but this is where fixing $\rho = 1$ (pure power law) leads to a good fit.

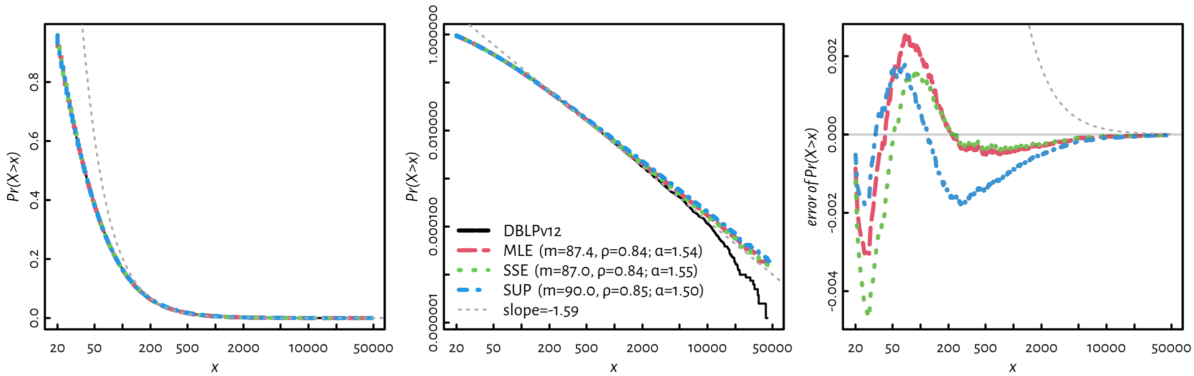

Fig. 4. Empirical (DBLP data, truncated at $\delta = 20$) and fitted complementary cumulative distribution functions; left: log scale on *Ox*, middle: log scale on both axes, right: error between the empirical and the theoretical CCDFs; the power law model (straight line on the log–log scale) gives a particularly bad fit for smaller observations (which are the most prevalent); MLE, SSE, and SUP seem to give poorer fits at the distribution tail in the middle plot but note that the probabilities therein are on the log scale, which exaggerates how the error magnitudes are perceived.

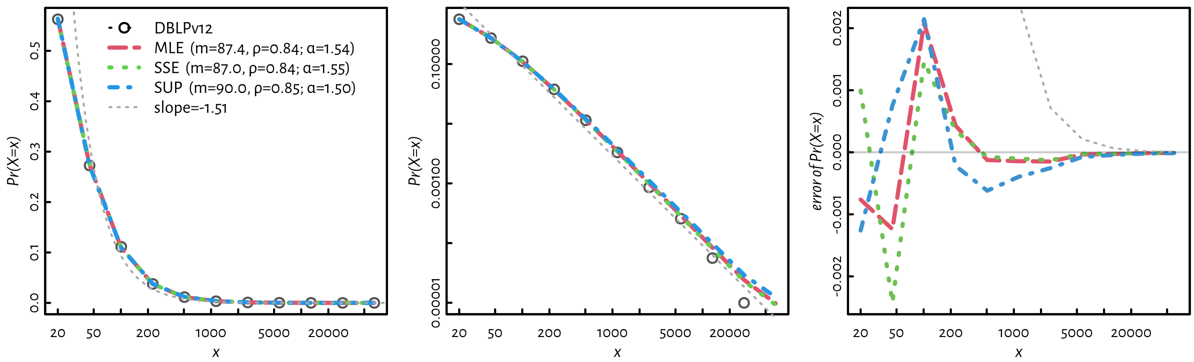

Fig. 5. Histogram (logarithmically-binned DBLP data, truncated at $\delta = 20$) and the fitted probability mass functions; left: log scale on *Ox*, middle: log scale on both axes, right: error between the histogram and the theoretical PMFs; note again how the power law model overestimates the density at smaller *x*s.

5. Conclusion

Our rank distribution approach marries the Price model and Pareto-type 2 distributions. We have shown how a combination of the rich-get-richer rule and sheer chance yields, in the long run, a well-known statistical distribution with a new, interpretable parametrisation. Reversely, the expected values of order statistics of any finite i.i.d. Paretian sample are consistent with the baseline Price ranks.

The considered model fits the DBLP citation data quite well. In fact, the Pareto distribution and its generalisations have been used in bibliometric modelling a few times already; see, e.g., [45–48].

Interestingly, in our case, for $\delta = 20$, the estimated ρ parameter ($\hat{\rho}_{\text{MLE}} = 0.84$) corresponds to a very high fraction of preferentially-attached citations — other studies of different real-world databases (e.g., [26,49]) suggested that success might be more accidental.

We noted that for $\rho = 1$, our model reduces to the ordinary power law (the Pareto-type 1 distribution) and that a δ-truncated Pareto-type 2 distribution yields $\rho \to 1$ as $\delta \to \infty$. Therefore, we expect that for all the power law-like datasets (e.g., considered in [38,40]) we enjoy a fit of the same or better quality, because we can always truncate at δ which yields a good power law-fit and assume $\rho = 1$. Still, we leave the empirical study of data from different domains (e.g., data from linguistics, economics, cellular biology, communication networks, ecology, transport, modelling of extreme weather events) for further research.

Future work will also involve the construction of usable estimators of m and ρ for small, discretised samples. Another interesting research direction is the analysis of the out-degrees of the network nodes in the Price model, since in this paper we have only focused on the in-degrees.

CRediT authorship contribution statement

Grzegorz Siudem: Conceptualization, Formal analysis, Methodology, Writing – original draft. **Przemysław Nowak:** Investigation, Formal analysis, Validation. **Marek Gagolewski:** Conceptualization, Formal analysis, Methodology, Software, Visualisation, Writing – original draft.

Declaration of competing interest

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

Data availability

The data are already publicly available.

Acknowledgements

This research was supported by the Australian Research Council (Australia) Discovery Project ARC DP210100227 (MG) and by the POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology (Poland) within the Excellence Initiative Program – Research University (ID-UB) (GS and PN).

The authors would also like to thank for valuable remarks the three anonymous reviewers as well as Anna Cena, Maciej J. Mrowiński, and Barbara Żogała-Siudem.

References

[1] M. Thelwall, P. Sud, Do new research issues attract more citations? A comparison between 25 Scopus subject categories, J. Assoc. Inf. Sci. Technol. 72 (2021) 269–279.
[2] D. Lyu, X. Ruan, J. Xie, Y. Cheng, The classification of citing motivations: A meta-synthesis, Scientometrics 126 (2021) 3243–3264.
[3] M. Gagolewski, Scientific impact assessment cannot be fair, J. Informetr. 7 (2013a) 792–802.
[4] S.N. Dorogovtsev, J.F. Mendes, Ranking scientists, Nat. Phys. 11 (2015) 882–883.
[5] D. Price, Little Science, Big Science, Columbia Univ. Press, New York, 1963.
[6] R.K. Merton, The Matthew effect in science, Science 159 (1968) 56–63, http://dx.doi.org/10.1126/science.159.3810.56.
[7] M. Perc, The Matthew effect in empirical data, J. R. Soc. Interface 11 (2014) 20140378.
[8] G. Ionescu, B. Chopard, An agent-based model for the bibliometric h-index, Eur. Phys. J. B 86 (426) (2013).
[9] B. Żogała-Siudem, G. Siudem, A. Cena, M. Gagolewski, Agent-based model for the h-index – exact solution, Eur. Phys. J. B (21) (2016).
[10] R. Sinatra, D. Wang, P. Deville, C. Song, A.L. Barabási, Quantifying the evolution of individual scientific impact, Science (354) (2016).
[11] M. Janosov, F. Battiston, R. Sinatra, Success and luck in creative careers, EPJ Data Sci. 9 (9) (2020).
[12] A. Pluchino, A.E. Biondo, A. Rapisarda, Talent versus luck: The role of randomness in success and failure, Adv. Complex Syst. 21 (2018) 1850014.
[13] A. Pluchino, G. Burgio, A. Rapisarda, A.E. Biondo, A. Pulvirenti, A. Ferro, T. Giorgino, Exploring the role of interdisciplinarity in physics: Success, talent and luck, PLoS One 14 (2019) e0218793.
[14] R. Heesen, Academic superstars: Competent or lucky? Synthese 194 (2017) 4499–4518.
[15] S.R. Kharel, T.R. Mezei, S. Chung, P.L. Erdős, Z. Toroczkai, Degree-preserving network growth, Nat. Phys. (2021) 1–7.
[16] M. Golosovsky, S. Solomon, Growing complex network of citations of scientific papers: Modeling and measurements, Phys. Rev. E 95 (2017) 012324, http://dx.doi.org/10.1103/PhysRevE.95.012324.
[17] C. Steinbock, O. Biham, E. Katzav, Distribution of shortest path lengths in a class of node duplication network models, Phys. Rev. E 96 (2017) 032301, http://dx.doi.org/10.1103/PhysRevE.96.032301.
[18] M. Golosovsky, Mechanisms of complex network growth: Synthesis of the preferential attachment and fitness models, Phys. Rev. E 97 (2018) 062310, http://dx.doi.org/10.1103/PhysRevE.97.062310.
[19] C. Steinbock, O. Biham, E. Katzav, Analytical results for the distribution of shortest path lengths in directed random networks that grow by node duplication, Eur. Phys. J. B 92 (2019) 1–16.

[20] C. Steinbock, O. Biham, E. Katzav, Analytical results for the in-degree and out-degree distributions of directed random networks that grow by node duplication, J. Stat. Mech. Theory Exp. 2019 (2019) 083403.
[21] D. Price, A general theory of bibliometric and other cumulative advantage processes, J. Am. Soc. Inf. Sci. 27 (1976) 292–306.
[22] M. Newman, Networks, Oxford University Press, 2018.
[23] Z. Liu, Y.C. Lai, N. Ye, P. Dasgupta, Connectivity distribution and attack tolerance of general networks with both preferential and random attachments, Phys. Lett. A 303 (2002) 337–344.
[24] C. Bedogne, G.J. Rodgers, Complex growing networks with intrinsic vertex fitness, Phys. Rev. E 74 (2006) 046115.
[25] Z.G. Shao, X.W. Zou, Z.J. Tan, Z.Z. Jin, Growing networks with mixed attachment mechanisms, J. Phys. A: Math. Gen. 39 (2035) (2006).
[26] G.J. Peterson, S. Pressé, K.A. Dill, Nonuniversal power law scaling in the probability distribution of scientific citations, Proc. Natl. Acad. Sci. 107 (2010) 16023–16027.
[27] Z. Néda, L. Varga, T.S. Biró, Science and facebook: The same popularity law!, PLoS One 12 (2017) 1–11, http://dx.doi.org/10.1371/journal.pone.0179656.
[28] G. Siudem, B. Żogała-Siudem, A. Cena, M. Gagolewski, Three dimensions of scientific impact, Proc. Natl. Acad. Sci. 117 (2020) 13896–13900, http://dx.doi.org/10.1073/pnas.2001064117.
[29] T.S. Evans, L. Calmon, V. Vasiliauskaite, The longest path in the price model, Sci. Rep. 10 (2020) 1–9.
[30] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Structure of growing networks with preferential linking, Phys. Rev. Lett. 85 (2000) 4633–4636, http://dx.doi.org/10.1103/PhysRevLett.85.4633.
[31] B.C. Arnold, Pareto Distributions, Chapman and Hall/CRC, New York, NY, USA, 2015, http://dx.doi.org/10.1201/b18141.
[32] M. Bourguignon, H. Saulo, R.N. Fernandez, A new pareto-type distribution with applications in reliability and income data, Physica A 457 (2016) 166–175.
[33] F.C. Figueira, N.J. Moura Jr., M.B. Ribeiro, The gompertz–pareto income distribution, Physica A 390 (2011) 689–698.
[34] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and mining of academic social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 990–998.
[35] W. Gautschi, Some elementary inequalities relating to the gamma and incomplete gamma function, J. Math. Phys. 38 (1959) 77–81, http://dx.doi.org/10.1002/sapm195938177.
[36] P. Billingsley, Probability and Measure, Wiley, 1995.
[37] H.A. David, H.N. Nagaraja, Order Statistics, Wiley, 2003.
[38] M.E. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. 46 (2005) 323–351, http://dx.doi.org/10.1080/00107510500052444.
[39] J. Beirlant, Y. Goegebeur, J. Teugels, J. Segers, Statistics of Extremes: Theory and Applications, Wiley, 2004, http://dx.doi.org/10.1002/0470012382.
[40] A. Clauset, C.R. Shalizi, M.E. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (2009) 661–703.
[41] J. Zhang, M.A. Stephens, A new and efficient estimation method for the generalized Pareto distribution, Technometrics 51 (2009) 316–325.
[42] A. Luceno, Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators, Comput. Statist. Data Anal. 1 (2006) 904–917.
[43] P.Z. Bermudez, S. Kotz, Parameter estimation of the generalized Pareto distribution. Part II, J. Statist. Plann. Inference 140 (2010) 1374–1388.
[44] J. Zhang, Improving on estimation for the generalized Pareto distribution, Technometrics 52 (2010) 335–339.
[45] A. Schubert, W. Glänzel, A systematic analysis of Hirsch-type indices for journals, J. Informetr. 1 (2007) 179–184.
[46] Glänzel W., On the h-index – a mathematical approach to a new measure of publication activity and citation impact, Scientometrics 67 (2006) 315–321.
[47] M. Gagolewski, Statistical hypothesis test for the difference between Hirsch indices of two Pareto-distributed random samples, in: R. Kruse, et al. (Eds.), Synergies of Soft Computing and Statistics for Intelligent Data Analysis, in: Advances in Intelligent Systems and Computing, vol. 190, Springer, 2013, pp. 359–367, http://dx.doi.org/10.1007/978-3-642-33042-1_39.
[48] M. Gagolewski, Sugeno integral-based confidence intervals for the theoretical h-index, in: P. Grzegorzewski, et al. (Eds.), Strengthening Links Between Data Analysis and Soft Computing, in: Advances in Intelligent Systems and Computing, vol. 315, Springer, 2015, pp. 233–240, http://dx.doi.org/10.1007/978-3-319-10765-3_28.
[49] N.A. Arnold, R.J. Mondragón, R.G. Clegg, Likelihood-based approach to discriminate mixtures of network models that vary in time, Sci. Rep. 11 (2021) 1–13.