THEORETICAL ANALYSIS OF EMPIRICAL RELATIONSHIPS FOR PARETO-DISTRIBUTED SCIENTOMETRIC DATA

Vladimir Atanassov, Ekaterina Detcheva

Abstract: *In this paper we study some problems involved in analysis of Pareto-distributed scientometric data (series of citations versus paper ranks). The problems include appropriate choices of i) the distribution type (continuous, discrete or finite-size discrete) and ii) statistical methods to obtain unbiased estimates for the power-law exponent (maximum likelihood procedure or least square regression.). Since relatively low magnitudes of the power exponent (less than 2), are observed massively in scientometric databases, finite-size discrete Pareto distribution (citations, distributed to finite number of paper ranks) appears to be more adequate for data analysis than the traditional ones. This conclusion is illustrated with two examples (for synthetic and actual data, respectively). We also derive empirical relationships, in particular, for the maximum and the total number of citations dependence on the Hirsch index. The latter generalize results of previous studies.*

Keywords: *Scientometrics, Hirsch index, Pareto distributions, data analysis, empirical relationships*

ACM Classification Keywords: *H. Information Systems, H.2. Database Management, H.2.8. Database applications, subject: Scientific databases; I. Computing methodologies, I.6 Simulation and Modeling, I.6.4. Model Validation and Analysis*

Introduction

It is hard to deny that, although not always and not everywhere welcome, *scientometrics* has entered the life of scientific community worldwide. The reasons for that are in much extent the tools it provides for assessment of quantity and quality of scientific output. Nowadays *scientometric indicators* are widely used for variety of purposes, including decision making in projects approval and evaluation, team building, scientific careers promotion, development of university and educational programs, scientific journals ranking, to mention just a few. Almost all of these indicators are based on two primary quantities, namely *number of publications*, considered as a measure of *productivity* and *number of citations* (appearing in the form of references to these publications), as a measure of *impact*, or popularity among the scientists. The attempts to find out a single score of scientific activity have brought to the world a variety of (secondary) scientometric indicators, the Hirsch index [Hirsch, 2005] being probably one of the most popular among them. Based on a simple model, this index suggests some compromise between productivity and impact.

Due to their common origin, scientometric indicators are mutually related *via* theoretically derived or empirically obtained relationships. Studying these relationships is important not only for understanding which indicators indicate what and how (*i.e.* from pure academic point of view), but also for appropriate choice of indicators in various applications. The relations between the indicators essentially depend on the (form of the) ranked citations-papers distribution. The empirical dependence of the total number of citations on the Hirsch index appearing in [Hirsch, 2005] has been supported by the linear negative-slope distribution resulting from the model, while other theoretical studies [*e.g.* Glänzel, 2006; Egghe and Rousseau, 2006] are concentrated on Pareto-distributed scientometric quantities.

In this paper we study several problems associated with the analysis of Pareto-distributed data (citations to paper ranks), suggest solutions to these problems and obtain relationships among indicators that are easy to obtain, or provided by the scientometric databases: total number of citations, number of publications, number of citations of the most cited paper, Hirsch's *h*-index *etc*. The problem of fitting power-law data has been addressed, too. This problem is of crucial importance, in particular with the recent development in that field ([Goldstein *et al*, 2004], [Clauset *et al*, 2009]).

The paper is organized as follows: in the first section we discuss the essential properties of Hirsch's index and the associated model; the second section critically reviews the application of Pareto distributions in scientometrics and the problems that arise in fitting to power-law data. In the next two sections we use the discrete finite-size Pareto distribution for scientometric data analysis and obtain some useful relationships, including the dependence of the total number of citations on the Hirsch's index.

Hirsch's model considerations

We recall the well known *h*-index definition: '*A scientist has index h if h of his or her N papers have at least h citations each and the other N-h papers have no more than h citations each',* and continue with a brief description of the model [Hirsch, 2005] used as its basis. Under the assumption that an individual publishes N_{ppy} papers per year and each published paper is cited N_{cpppy} times every (subsequent) year, the total number of citations N_c earned after N_y ($N_y \gg 1$) years is:

$$N_c \approx \frac{1}{2} N_{ppy} N_{cpppy} (1/N_{ppy} + 1/N_{cpppy})^2 h^2 , \tag{1}$$

where the *h*-index depends linearly on N_y :

$$h \approx (1/N_{ppy} + 1/N_{cpppy})^{-1} N_y . \tag{2}$$

It is worth noting that the scientometric indicators involved in Eqs. (1) and (2) are provided by most bibliometric databases, *e.g.* Web of Knowledge (Thomson Reuters), Scopus (Elsevier), Google Scholar *etc*.

For a time interval of N_y years this simple model yields a linear negative-slope distribution $C(P)$ of citation number C to paper rank P, papers being arranged in descending order of number of citations, *i.e.* the most cited placed first. It follows from Eq. (1) that N_c depends quadratically on h :

$$N_c = Ah^2 . \tag{3}$$

It is easy to see that $A \geq 2$; it reaches its minimum for a straight line of slope -1 (Fig. 1). Moreover, for quite general negative slope concave type of $C(P)$ distribution (Fig. 2) it can be shown that

$$A > A_0 = 1 - \frac{1}{2}(C' + 1/C') \geq 2 , \tag{4}$$

where $C' = (dC/dP)_{P=h}$ is the slope of the distribution at $P = h$. Values of $A < A_0$ indicate negative slope convex type of distributions. It should be noted that Hirsch has empirically found $A \approx 3-5$ as a typical value.

Further on it is worth mentioning another (although rather artificial) distribution $C(P)$ – the uniform distribution (Fig. 3 a,b) that could be considered as a *limiting case* of convex negative-slope type distribution. It clearly

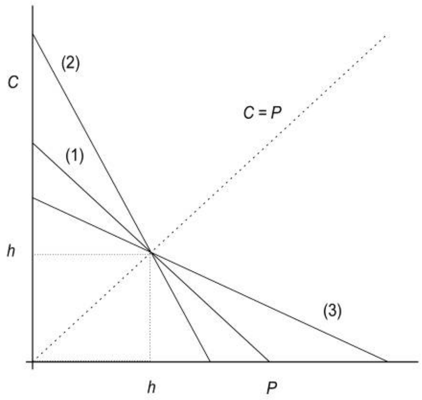

Fig. 1. Illustration of linear negative-slope distributions: (1) slope $= -1$, $A = 2$; (2) slope < -1, $A > 2$ (excess of citations); (3) slope > -1, $A > 2$ (excess of papers).

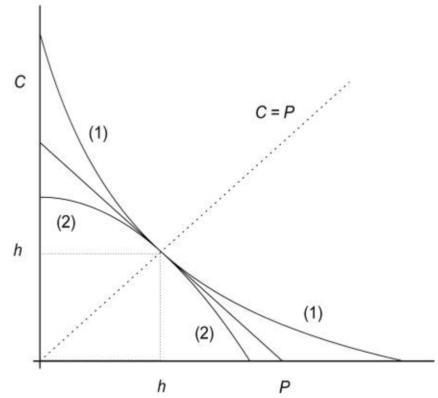

Fig. 2. Examples of negative slope concave (1) and convex (2) type of $C(P)$ distributions

demonstrates the limitations to *h, i.e. h* cannot exceed the maximum number of citations $C_{\max}$ and the maximum number of papers $P_{\max}$:

$$h = \min(C_{\max}, P_{\max}), \tag{5}$$

and the total number of citations is:

$$N_c = \max(C_{\max}, P_{\max})h \; . \tag{6}$$

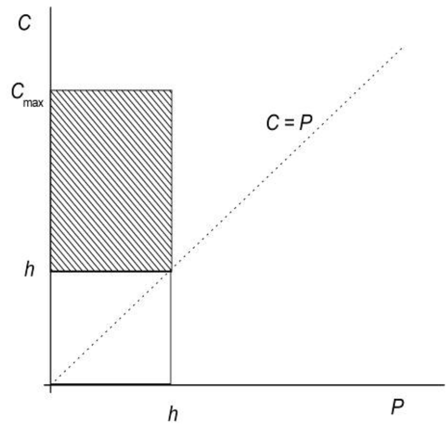

Fig. 3 a. Uniform $C(P)$ distribution example: excess of citations, *h*-index limited by maximum number of papers $P = h$

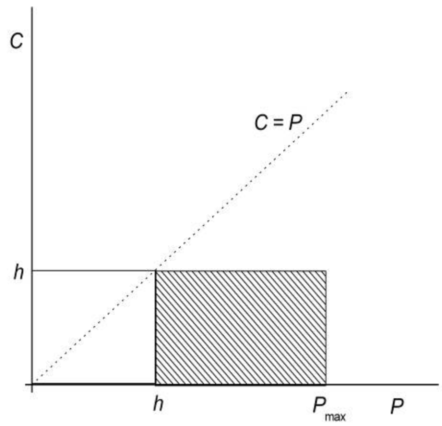

Fig. 3 b. Uniform $C(P)$ distribution example: excess of papers, *h*-index limited by maximum number of citations $C = h$

Pareto distributions in scientometrics

The distributions of number of citations to papers' rank $C(P)$ considered in the previous section concern rather sparse examples of data sets; the bulk of scientometric data is mainly characterized by negative slope concave type probability densities. One of the simplest types of such distributions has been used by *Alfredo Pareto* (1848-1923) for solving problems in economics, as allocation of wealth *etc*. *Pareto distribution* is a special case of *power-law distribution* with negative exponent, defined as strictly zero below some (positive) number that we shall assume to be 1:

$$PDF(X) = (\alpha - 1)X^{-\alpha}, CPF(X) = 1 - X^{1-\alpha}, X \geq 1 \quad (7)$$

Three examples for (continuous) Pareto distribution as well as one for its discrete version (called also *Zeta distribution*, defined for positive integers):

$$Probability(I) = (1/\zeta(\alpha))I^{-\alpha}, \alpha > 1, I = 1,2,3,\ldots \quad (8)$$

are demonstrated on Fig. 4 a,b. In Eq. 8 $\zeta(\alpha)$ is the Riemann zeta-function, tabulated in [Walther, 1926], [Janke *et al*, 1960] *etc*.

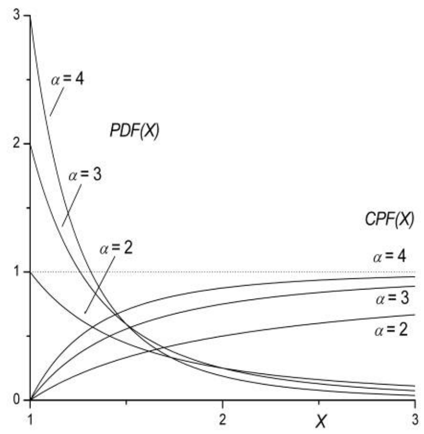

Fig. 4a. Pareto distribution examples: probability density $PDF(X) = (\alpha - 1)x^{-\alpha}$ and cumulative probability function $CPF(x) = 1 - x^{-\alpha}$ for $\alpha = 2$, 3 and 4.

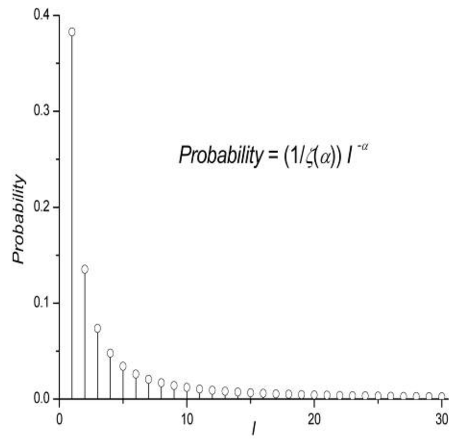

Fig. 4b. Discrete Pareto (*Zeta*) distribution example: $Probability(I) = (1/\zeta(\alpha))I^{-\alpha}, \alpha = 1.5$

Both (continuous and discrete) Pareto distributions should be handled with care bearing in mind that all moments of order *k* and above simply *do not exist* for $k \geq \alpha - 1$**.** For instance, a (rather popular, see [Goldstein *et al*, 2004] and [Clauset *et al*, 2009]) Pareto distribution with $\alpha = 2.5$ has infinite dispersion and does not meet the requirements of the *Central Limit Theorem*. Probability densities with $1 < \alpha \leq 2$ have no mean, and those with $\alpha \leq 1$ are no distributions at all**.**

There are some questions and problems that appear in connection with Pareto distribution applications in scientometrics and we address them further on. The first one is: are citation numbers really Pareto-distributed to paper ranks? An acceptable answer to this important question could be found in [Clauset *et al*, 2009]: it states that *power law* is a plausible choice, just as *log-normal* and *stretched exponential* (see, *e.g.* [Hirsch, 2005]) are;

all other statistical hypotheses could be rejected. Testing of power-law hypothesis is a tough problem that remains beyond the scope of this study and further on we assume that data is *apriori* Pareto distributed.

Another point of Pareto distributed data analysis, namely data fit and derivation of unbiased power-law exponent estimate seems to be a 'hot potato', too (see [Goldstein *et al*, 2004], [Clauset *et al*, 2009]). The analyses of synthetic data described in above mentioned papers reveal that the straightforward log-log *ordinary least square regression* (OLSR, see *e.g.* [Weisberg, 2005]) gives essentially distorted estimates for the power-law exponent; the authors recommend the use of *maximum likelihood estimate* (MLE) instead. For continuous Pareto-distributed data the latter looks like:

$$\alpha_{est} = 1 + (\frac{1}{n}\sum_{i=1}^{n} \ln x_i)^{-1}, \tag{9}$$

while for discrete Pareto-distributed data one has to solve a transcendental equation:

$$\zeta'(\alpha_{est})/\zeta(\alpha_{est}) = -(1/n)\sum_{i=1}^{n} \ln x_i \ . \tag{10}$$

In Eqs. (9) and (10) $\zeta'(\alpha)$ is the first derivative of the Riemann zeta-function $\zeta(\alpha)$, (see Fig. 5), α_{est} is the MLE for the power-low exponent and $\{x_i; i = 1, 2, \dots, n\}$ represents a *sample* containing *n independent data*. Again we refer to [Clauset *et al*, 2009] for estimates of the MLE standard error.

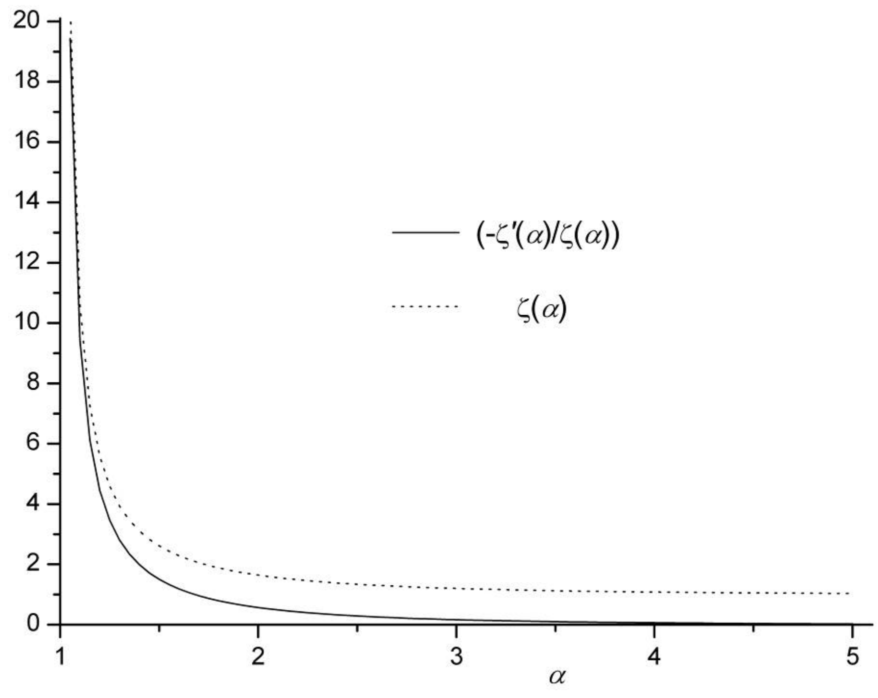

Fig. 5. Riemann zeta-function $\zeta(\alpha)$ and its logarithmic derivative $-\zeta'(\alpha)/\zeta(\alpha)$.

The last but not least important item of our list of questions and problems is the (estimated) value of the power-law exponent α **itself.** Theoretical studies on Pareto-distributed scientometric data (*e.g.* [Glänzel, 2006]) consider samples with essentially infinite number of papers *N* and finite mean (*i.e.* first moment, $k = 1$). As we have already mentioned, this imposes a restriction $\alpha > 2$ for most results obtained there. The point is, whether citation data resulting from such distributions exist in any extent of importance. One could argue that data sets with 90 percent of citations concentrated in the first four papers (as it follows from a distribution with $\alpha = 2$) are extremely rare to find in bibliometric databases. Usually $\alpha < 2$, and values close to (and even less than) 1 could be found, too. For example, Schreiber's dataset [Schreiber, 2007] used for analysis of self-citations effect on the *h*-index is characterized by $\alpha = 1.35$ (MLE) and $\alpha = 0.70$ (log-log OLSR estimate with $R^2 = 0.98$). Hence it is necessary to find out a scheme that can be used to analyze scientometric data in the 'twilight zone' $1 \le \alpha \le 2$, evading the pole of $\zeta(\alpha)$ at $\alpha = 1$ and moments that diverge even for the lowest orders.

Finite size discrete Pareto distributions

Introducing some kind of upper limit (cut-off) for all summations or integrations is usually considered as universal remedy ('painkiller') for the problems similar to those we discussed in the previous section. It is not a secret for anyone trying to compute directly the zeta function that millions of terms must be summed up to achieve somewhat acceptable accuracy. This, however, does not correspond to the simple fact that individual scientist's production normally does not exceed several hundred papers. Therefore, it does not make much sense in merging slowly converging tails to the *actual* probability density by approximating it with discrete infinite-sized Pareto distribution. The latter could be replaced by a finite size distribution:

$$Probability(l,N) = (1 / S(\alpha,N))l^{-\alpha}, l = 1,2,\ldots,N \,, \tag{11}$$

where N is the number of papers *cited at least once* (or its estimate) and the incomplete zeta function (Fig. 6) is:

$$S(\alpha,N) = \sum_{l=1}^{N} l^{-\alpha} \tag{12}$$

Since $S(\alpha,N)$ depends slowly (within an interval from several tens to several hundreds) on N, we can perform the standard maximum likelihood optimization to obtain the MLE for α :

$$S'(\alpha_{est},N)/S(\alpha_{est},N) = -(1/n)\sum_{i=1}^{n} \ln x_i \,, \tag{13}$$

where $S'(\alpha,N) = \partial S(\alpha,N) / \partial\alpha$ (Fig. 7.). Figs 6 and 7 give a notion for why and where the standard discrete Pareto analysis fails to give acceptable (unbiased) estimates for the power exponent α .

In order to illustrate how this scheme works we have chosen two examples: a synthetic data one (Fig. 8) and one with actual data (Fig. 9). Synthetic data have been computed by using

$$l_c(l) = \text{nint}(l_{c\max} l^{-\alpha}), l = 1,2,\ldots,N \,, \tag{14}$$

where the maximum number of citations, *i.e.* the number of citations gained by the first rank paper is estimated for given total number of citations N_c and power-law exponent α as

$$l_{c\max} = l_c(1) = N_c / S(\alpha,N) \,. \tag{15}$$

Thus rounding error was the only one introduced by the computation. The actual data have been adopted from Thomson Reuters Web of Knowledge database. Four kinds of analysis have been performed in both cases, as follows: 1. C+U MLE denotes maximum likelihood estimate for continuous, infinite-size (*i.e*. unlimited) argument Pareto distribution, by using Eqs. (7) and (9); 2. D+U MLE denotes MLE for discrete infinite-size (unlimited) argument discrete Pareto distribution, obtained *via* Eqs. (8) and (10); 3. D+L MLE corresponds to MLE for discrete finite size (limited argument) Pareto distribution according to Eqs. (11)-(13) and 4. log-log OLSR stays for

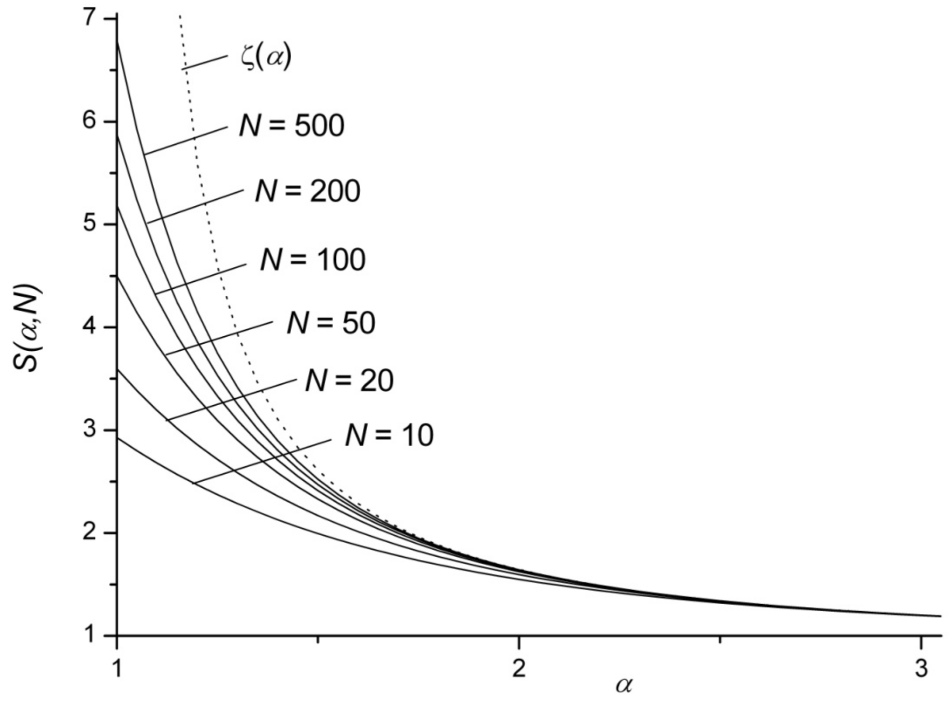

Fig. 6. The function $S(\alpha,N)$ versus α for various N . Note that it closely approaches $\zeta(\alpha)$ for $\alpha > 2$.

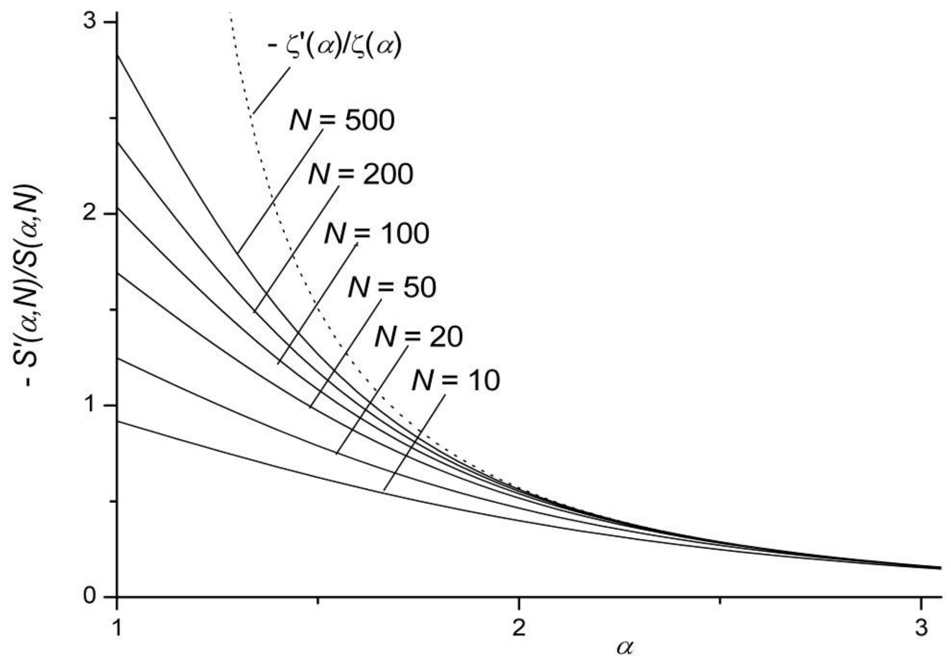

Fig. 7. The function $-S'(\alpha,N) / S(\alpha,N)$. Note that it closely follows $-\zeta'(\alpha)/\zeta(\alpha)$ for $\alpha > 2.5$.

log-log ordinary linear regression. The results of these analyses are summarized in Tables 1 and 2 for the synthetic and actual data, respectively.

Table 1. Estimates of α and $I_{c\max}$ for synthetic data $I_c = \mathrm{nint}(I_{c\max} I^{-\alpha})$, $\alpha = 1.5$, $I_{c\max} = 213$, $N = 56$ (Fig. 8)

	C+U MLE	D+U MLE	D+L MLE	log-log OLSR
α_{est}	1.98	1.67	1.48	1.37
$I_{c\max}$	495	233	210	158
n	506	506	506	$(R^2 = 0.97)$

Table 2. Estimates of α and $I_{c\max}$ for real data $I_C(I)$ with $I_{c\max} = 62$, $N = 16$, $N_c = 234$, $h = 8$ (Fig. 9.)

	C+U MLE	D+U MLE	D+L MLE	log-log OLSR
α_{est}	1.94	1.65	1.09	1.37
$I_{c\max}$	220	108	76	112
n	234	234	234	$(R^2 = 0.85)$

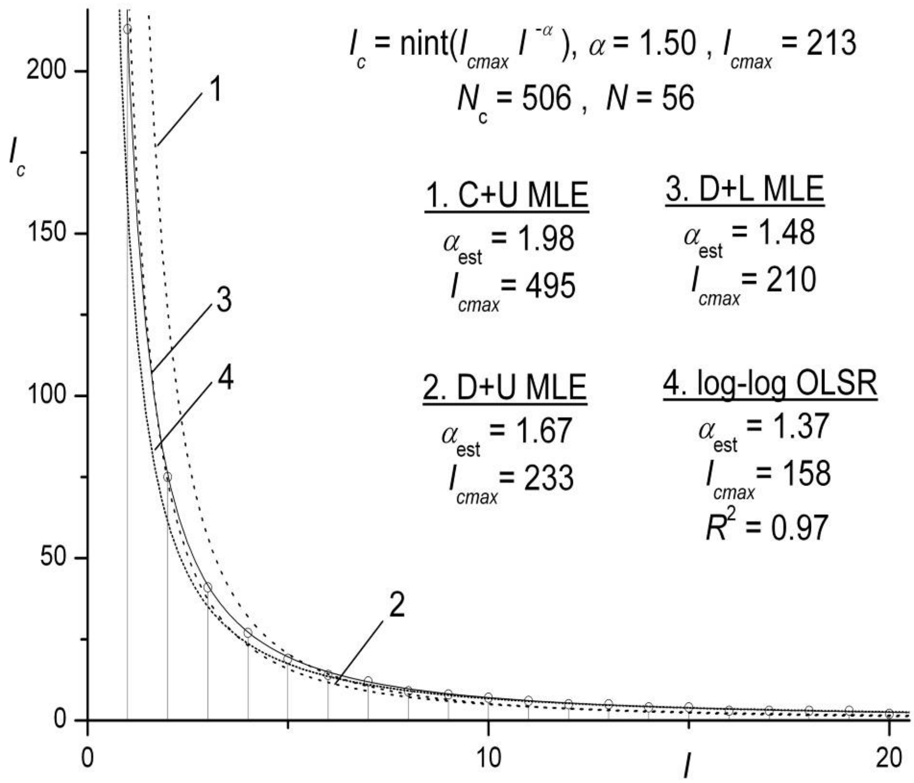

Fig. 8. Synthetic data example: number of citations I_c versus paper rank I.

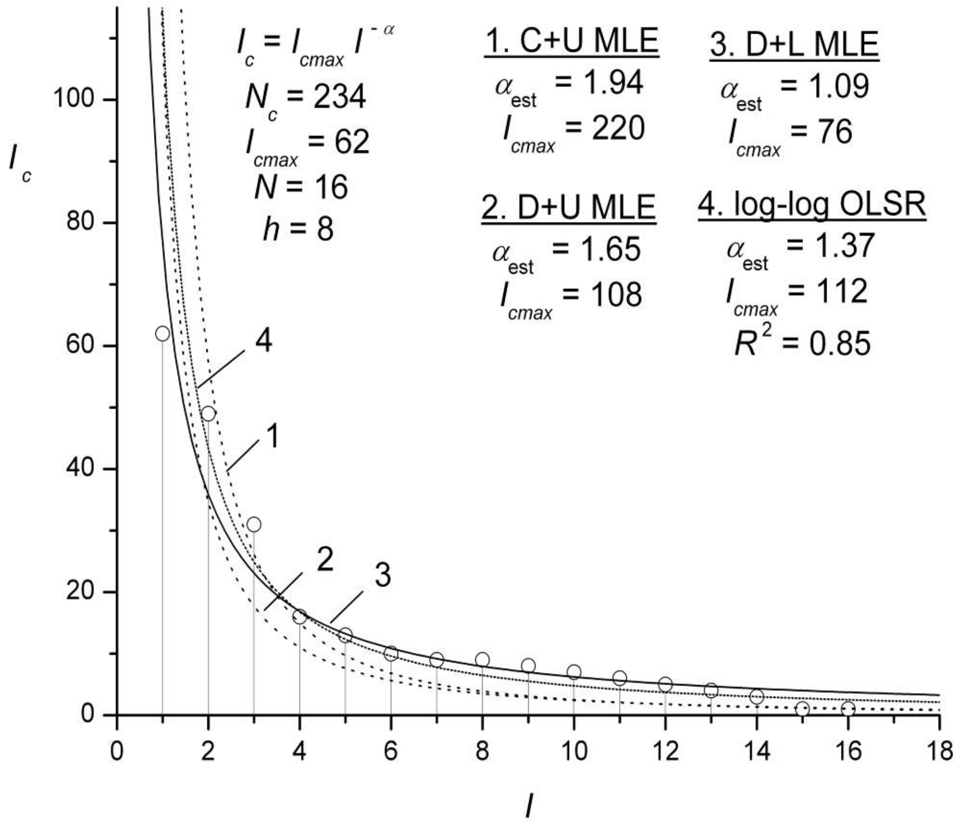

Fig. 9. Real world data example: number of citations I_c versus paper rank I.

Both examples demonstrate the complete failure of the continuous infinite-sized Pareto distribution analysis (Eqs. (7) and (9)) to provide an acceptable unbiased estimate for α and $I_{c\max}$ (for the synthetic data example) and $I_{c\max}$ (for the actual data). The MLE, resulting from an infinite-size discrete Pareto distribution -Eqs. (8) and (10), is usually considered to be the best one and this might be true for $\alpha > 2$. In our synthetic data example, however, it tends to overestimate the prescribed power-law exponent ($\alpha = 1.5$), while the log-log ordinary least square regression underestimates it. The latter phenomenon has been reported earlier in the simulation experiments of [Goldstein *et al*, 2004] and [Clauset *et al*, 2009]. We point out that, without any doubt, the discrete finite size MLE (Eqs. (11)-(13)) performs best for both synthetic and real life data.

Empirical relationships for Pareto-distributed scientometric data

Let us now consider some relationships between scientometric parameters following from the assumption for citations that are (discrete) Pareto distributed to paper ranks. From the definition of *h*-index $I_c(h) \approx h$ we obtain

$$I_{c\max} \approx h^{1+\alpha}. \tag{16}$$

This relation holds for all types of Pareto distributions – continuous, discrete, infinite or finite-sized; it also matches quite well the actual scientometric data (Fig. 9). Note that in the Hirsch's model $I_{c\max}$ depends *linearly* on *h:*

$$I_{c\max} \approx (1 + N_{cpppy} / N_{ppy})h \,. \tag{17}$$

By replacing $I_{c\max}$ in (15) we arrive at

$$N_c \approx S(\alpha, N)h^{1+\alpha} \,. \tag{18}$$

Eq. (18) relates the total number of citations N_c the $(1+\alpha)$-th power of the *h*-index *via* the (slowly depending on the power law exponent α and on the number of publications cited at least once, N) function $S(\alpha, N)$. It represents a *generalization* of Hirsch's relationship (3) for Pareto-distributed scientometric data. Moreover, one can see on Fig. 10 that for reasonable choice of N (50-350) and α (1.05-1.50) the coefficient $S(\alpha, N)$ in (17) varies between 2.5 and 5.5, *i.e.* close to the Hirsch's empirical result $A = 3 - 5$ [Hirsch, 2005].

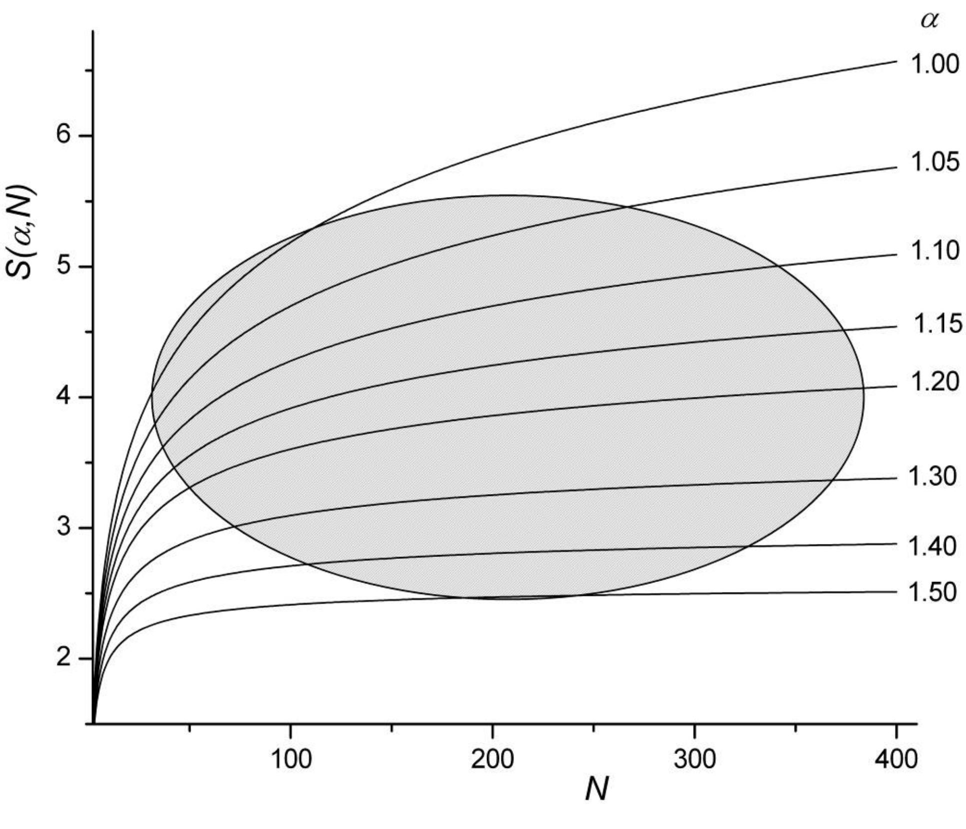

Fig. 10. $S(\alpha, N)$ versus number of publications N for various α .The area in grey roughly pictures the range of Hirsch's empirical evaluation of A .

In the limit $N \to \infty$ we have $S(\alpha, N) \to \zeta(\alpha)$. A brief inspection of Fig. 5 allows us to conclude that $\zeta(\alpha)$ might be considered as slowly varying for $\alpha > 2$. This however, implies dependence stronger than $N_c \sim h^3$. In addition, the continuous version of Eq. (17):

$$N_c = (\alpha - 1)^{-1} h^{1+\alpha}, \tag{19}$$

has severe problems for α close to unity and is obviously not applicable to analyze scientometric data.

The Pareto assumption allows us to obtain another relation that might be useful when the sample under consideration contains publications, all of them with nonzero number of citations, *i.e.* $I_c(l) \neq 0$ for l=1,2,...,N. This means that the actual number of papers cited at least once is $N_a \geq N$. In order to estimate the actual upper limit of the discrete finite-size Pareto distribution one could take into account the fact that by definition the number of citations I_c is always a positive integer. Bearing in mind the **nint** (nearest integer) convention as well as Eqs. (14) and (16) we obtain

$$N_a = (2I_{c\max})^{1/\alpha} = 2^{1/\alpha} h^{(1+\alpha)/\alpha}, \tag{20}$$

as a crude estimate for the actual number of publications with at least one citation. One could compare it with the maximum number of publications that follows from the Hirsch's model:

$$N_{\max} = (1 + N_{ppy} / N_{cpppy}) h \quad , \tag{21}$$

and linearly depends on the h-index.

Summary and conclusions

In this paper we have studied some problems appearing in the analysis of Pareto-distributed scientometric data in the form of series of citations versus paper ranks. These problems include appropriate choice of the distribution type (continuous, discrete or finite-size discrete) and of statistical methods to obtain unbiased estimates for the power-law exponent (maximum likelihood procedure or least square regression). Further on, we have theoretically derived relationships, in particular, the total number of citations dependence on the Hirsch index, that generalize results of previous studies and may be proved empirically.

Our conclusions are summarized as follows:

- Pareto-distribution analysis of citations versus paper rank data requires use of probability densities with power-law exponent **of** magnitude 2 or less. Therefore it is necessary to use finite-sized (*i.e.* defined for a finite number of papers) discrete Pareto distribution;
- The maximum likelihood estimate of the power-law exponent for the finite-size discrete Pareto distribution seems to provide best fit to the data, while those resulting from the maximum likelihood procedure for infinite series discrete distribution and from the log-log ordinary least square regression give overestimated and underestimated values. The continuous Pareto analysis proves to be inappropriate for this kind of data;
- The maximum number of citations gained by a paper (this is the paper of rank 1) is a power function of Hirsch's index; the same holds for the total number of citations, however, with constant of proportionality weakly depending on power exponent and number of papers;
- Papers of zero citation count are irrelevant for Pareto distribution analysis of scientometric data. The number of papers with at least one citation is power function of the Hirsch's index, too.

Acknowledgments:

The authors wish to acknowledge the kind invitation and support of Dr K. Markov and Dr S. Poryazov. This paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Clauset et al, 2009] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, J. SIAM Review 51 (4) 661-703 (2009)

[Egghe and Rousseau, 2006] L. Egghe and R. Rousseau. An informetric model for the Hirsch-index, Scientometrics 69 (1) 121-129 (2006)

[Glänzel, 2006] W. Glänzel. On the h-index – A mathematical approach to a new measure of publication activity and citation impact, Scientometrics 67 (2) 315-321 (2006)

[Goldstein et al, 2004] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution, Eur. Phys. J. B – Condensed Matter and Complex Systems 41 (2) 255-258 (2004)

[Hirsch, 2005] J.E. Hirsch. An index to quantify an individual's scientific research output, Proc. Nat. Acad. Sci. 102 (46) 16569-16572 (2005)

[Janke et al, 1960] E. Janke, F. Emde and F. Lösh, Tafeln Höherer Funktionen, B. G. Teubner Verlagsgesellschaft, Stuttgart, 1960

[Schreiber, 2007] M. Schreiber. Self-citation corrections for the Hirsch index, Eur. Phys. Lett. 78 30002 (2007)

[Walther, 1926] A. Walther. Anschauliches zur Riemannischen Zetafunktion, Acta Mathematica 48 (3-4) 393-400 (1926)

[Weisberg, 2005] S. Weisberg, Applied Linear Regression, Wiley-Interscience, 2005

Authors' Information

Vladimir Atanassov *– Institute of Electronics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria; e-mail: v.atanassov@abv.bg*

Major Fields of Scientific Research: Plasma Physics and Gas Discharges, Radars & Ocean Waves, Nonlinearity & Chaos, Scientometrics

Ekaterina Detcheva *– Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,1113 Sofia, Bulgaria; e-mail: detcheva@math.bas.bg*

Major Fields of Scientific Research: Web-based applications, Image processing, analysis and classification, Knowledge representation, Business applications, Applications in Medicine and Biology, Applications in Psychology and Special Education, Computer Algebra.