The share of items of highly productive sources in function of the size of the system

The share of items of highly productive sources as a function of the size of the system

by

L. Egghe

Limburgs Universitair Centrum (LUC), Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

leo.egghe@luc.ac.be

ABSTRACT

The research in this paper is based on the paper "D.W. Aksnes and G. Sivertsen. The effect of highly cited papers on national citation indicators. Scientometrics 59(2), 213-224, 2004" where one states that "the few highly cited papers account for the highest share of the citations in the smallest fields".

This, at first sight, evident property is examined in the theoretical models that exist in the literature. We first define exactly what we mean by "size of a field" (i.e. when is a field "smaller" or "larger" than another one). We show that there are two, non-equivalent possible definitions. Next we define exactly the possible property under study. This leads us again to

---

[1] Permanent address

Key words and phrases: highly productive source, field size, Lotka, Zipf

two possible, non-equivalent formulations. Hence, in total, there are four different formulations to consider.

We show, by giving counterexamples, that none of these four formulations are true in general. We also express conditions (in Lotkaian and Zipfian informetrics), under which the property of Aksnes and Sivertsen is true.

All these results are not only valid in the papers-citations relationships but in any informetric source-item relationship. In this connection we present formulae describing the share of items of highly productive sources as a function of the parameters of the system (e.g. the size of the system).

I. Introduction

It is clear and well-known that informetric (and more generally sociometric, econometric, webometric,…) data are highly skewed in the sense that few sources produce many items and many sources produce few items. This is expressed in informetrics by authors writing ("producing") papers, by journals containing articles or even by articles receiving (or giving) citations. In webometrics (to be considered as that part of informetrics devoted to the metrics of the web and other social networks) one has skewness of data on websites and their in- or outlinks (i.e. hyperlinks pointing to or from the website, respectively). The earliest notices (at the end of the 19th century and in the beginning of the 20th century) of this type of skewness were made in econometrics expressing inequality in production, income or wealth. We refer to Lotka (1926) (author-paper relation), Bradford (1934) (journal-article relation) or even Pareto (1985) (in econometrics) for historical references and to Egghe and Rousseau (1990a) and Egghe (2005) for general references.

In econometrics one expresses this skewness by defining so-called good concentration (or inequality) measures such as the ones of Gini and Theil (see e.g. Gini (1909), Theil (1967)) or again Egghe and Rousseau (1990a) or Egghe (2005) for applications in the field of informetrics. Another way of studying this inequality is to look at the top-sources, i.e. at the sources that produce most items and to look at their impact on the total production. This

approach (and in the context of articles-citations (to these articles)) was followed in Aksnes and Sivertsen (2004) (see also Aksnes (2003)) where one studies the impact of highly cited papers on national citation indicators. More concretely one studies the effect (or rather the share) of the most cited article (or 5 most cited articles) in the totality of citations received by all articles in the field. Here several fields were selected.

Basic to both approaches (with inequality measures or by looking at the impact of the top – say k – sources) are the underlying frequency distributions of sources versus items in each field. By this we mean, for each field

- their rank-frequency function g : for each $r = 1,...,T$ (the total number of sources), $g(r)$ is the number of items in (or produced by) the source on rank r, where sources are ranked in decreasing order of number of items. Classical examples for g are a decreasing power law (i.e. the law of Zipf) or its generalization : the law of Mandelbrot. We do not go into this matter now but we will do so in the sequel.
- Their size-frequency function f : for each $j = 1,...\rho_m$, $f(j)$ is the number of sources with j items ($j = \rho_m$ being the maximum number of items in a source). The most classical example of such a function is a decreasing power law, i.e. the law of Lotka. Also here we do not go into this matter now but we will do so in the sequel.

Both functions f and g are different expressions of the same phenomenon: the production distribution in a field. Their relation is as follows: let the source on rank r have j items. Then, by definition of f and g (g^{-1} denotes the inverse function of g and primes (such as in j' and r' also further on) denote dummy variables)

$$r = g^{-1}(j) = \sum_{j'=j}^{\rho_m} f(j') \qquad (1)$$

(see also formula (I.3) in Egghe (2005)). Indeed,

$$\sum_{j'=j}^{\rho_m} f(j') \qquad (2)$$

is the cumulative number of sources with a number of items larger than or equal to j. Since sources are ranked decreasingly according to their number of items and since the source on rank r has j items (by notation) we indeed have that (2) equals r. But, by definition,

$$j = g(r), \tag{3}$$

the number of items in the source on rank r, hence, supposing g to be injective (we will see in the sequel that we can take g to be strictly decreasing, hence injective),

$$r = g^{-1}(j). \tag{4}$$

Hence (1) follows.

The functions f and g lead to inequality studies as invented in econometrics; for this see Egghe and Rousseau (1990a,b) or Egghe (2005), Chapter IV. The functions f and g also yield the other approach of studying inequality by looking at the top sources (as done in Aksnes and Sivertsen (2004)). Indeed we have the following formulae: the total number of items:

$$A = \sum_{r'=1}^{T} g(r') = \sum_{j'=1}^{\rho_m} j' f(j') \tag{5}$$

The total number of items in the r top sources equals

$$\sum_{r'=1}^{r} g(r') = \sum_{j'=j}^{\rho_m} j' f(j') \tag{6}$$

since (3) (or (4)) is valid. Hence the division of (6) by (5) yields the share of items of the top r sources in the total item production.

It is now clear that, with the above machinery, we can study the impact of highly cited papers on the national citation output in different fields, as was done in Aksnes and Sivertsen (2004) and therefore conjecturing the following property:

Conjecture (Aksnes and Sivertsen (2004) p. 217):
The few highly cited papers account for the highest shares of the citations in the smallest fields.

This conjecture, intuitively, is true since the smallest fields have the least papers. It is clear that the above conjecture can be, more generally, rephrased as follows:

Generalized conjecture:
The few highly productive sources account for the highest shares of the items in the smaller systems.

With "system" we mean any source-item production situation as described above and called in Egghe (1990), Egghe and Rousseau (1990a) and Egghe (2005) an information production process (IPP).

The study of the above conjecture is the topic of this paper. First of all, in the next section, the used terms and the conjecture itself will be rephrased in order to obtain a mathematically exact framework: we will define "field size" by using sources or items (non-equivalent definitions) and we will give two – as we will show in section III – non-equivalent, exact formulations of the conjecture, one using the function f and one using the function g. Hence, in total, we consider four different formulations of the above conjecture.

We will also use, from now on, the continuous equivalent formulations of formulae (1)-(6), the validity of which we will prove in the Appendix: there we will present self-contained proofs of all the results needed in the sequel.

In the third section we show, by presenting counterexamples using Lotkaian and Zipfian functions, that none of the four conjectures is true in general. We also, however, provide mathematical conditions (on the parameters of the Lotkaian and Zipfian functions) in order to have the validity of these conjectures, hereby also showing that the rank-frequency version of the conjecture is not equivalent with the size-frequency version.

II. Exact formulations

II.1 Comparing field (or IPP) sizes

It is clear that there are two natural ways of expressing the size of a field or an IPP: by expressing the size of T, the total number of sources or by expressing the size of A, the total number of items. We define

Definition II.1.1 (source-sense)

Given two IPPs, we say that the first one is smaller than the second one in the source-sense if

$$T_1 < T_2 \tag{7}$$

where $T_i (i = 1,2)$ denotes the total number of sources in IPP i.

Definition II.1.2 (item-sense)

Given two IPPs, we say that the first one is smaller than the second one in the item-sense if

$$A_1 < A_2 \tag{8}$$

where $A_i (i = 1,2)$ denotes the total number of items in IPP i.

A simple example, from Lotkaian informetrics, shows that the above definitions are not equivalent

Example II.1.3:

Given $A_1 = 1,000$ and $T_1 = 500$ we have, by Proposition A.3 in the Appendix that the Lotka function $(j \hat{I} [1,+¥)^{o} \{x \| 1£ x\})$

$$f_1(j) = \frac{1,000}{j^3}$$

satisfies the requirements (A10) and (A11). Indeed:

$$\alpha_1 = \frac{2A_1 - T_1}{A_1 - T_1} = 3$$

$$C_1 = \frac{A_1 T_1}{A_1 - T_1} = 1{,}000$$

We now look for an IPP satisfying $A_2 = 900 < A_1$ and $T_2 = 600 > T_1$. For this we need, again according to Proposition A.3

$$\alpha_2 = \frac{2A_2 - T_2}{A_2 - T_2} = 4$$

$$C_2 = \frac{A_2 T_2}{A_2 - T_2} = 1{,}800$$

Hence the Lotka function $(j \hat{I} [1, +¥))$

$$f_2(j) = \frac{1{,}800}{j^4}$$

satisfies the requirements (A10) and (A11).

Conclusion:

The above example shows that there exist IPPs for which $A_1 > A_2$ but for which $T_1 < T_2$.

Notice that in the above example $C_1 < C_2$. The next proposition shows that if $C_1 > C_2$ we have that Definition II.1.1 implies Definition II.1.2 (if we use functions of the type (A9), i.e. Lotkaian functions with $\rho_m = ¥$).

Proposition II.1.4:

If

$$C_1 > C_2 ,$$

then

$$T_1 < T_2 \Þ A_1 < A_2 \tag{9}$$

Proof:

Using formulae (A14) and (A15) we have that

$$T_1 < T_2 \Û \frac{C_1}{\alpha_1 - 1} < \frac{C_2}{\alpha_2 - 1}$$

$$A_1 < A_2 \Û \frac{C_1}{\alpha_1 - 2} < \frac{C_2}{\alpha_2 - 2}$$

which is equivalent with (since $\alpha_1, \alpha_2 > 2$ and since $C_1, C_2 > 0$ obviously)

$$T_1 < T_2 \Û C_1\alpha_2 - C_1 < C_2\alpha_1 - C_2 \tag{10}$$

$$A_1 < A_2 \Û C_1\alpha_2 - 2C_1 < C_2\alpha_1 - 2C_2 \tag{11}$$

Using that $-C_1 < -C_2$ we have that $T_1 < T_2 \Þ A_1 < A_2$. ~

We now give examples that the reverse direction in (9) is not true, given $C_1 > C_2$ and also that the reverse of Proposition II.1.4 is not true.

Examples II.1.5:

1. In Proposition II.1.4, relation (9) is not an equivalency. Indeed take $T_1 = 2,000$, $C_1 = 4,000$, so, by (A14), $\alpha_1 = 3$. Take $T_2 = 1,000$, $C_2 = 1,100$, so, by (A14),

$\alpha_2 = 2{,}1$. By (A15) we have $A_1 = 4{,}000$ and $A_2 = 11{,}000$. Hence $C_1 > C_2$, $A_1 < A_2$ but $T_1 > T_2$.

2. The reverse of Proposition II.1.4 is not true. Indeed take $T_1 = 500$, $A_1 = 1{,}000$, $T_2 = 600 > T_1$, $A_2 = 1{,}200 > A_1$. By (A13) we have $C_1 = 1{,}000 < C_2 = 1{,}200$.

From the above results it is clear that "size" needs to be defined exactly since there are many cases where $T_1 < T_2$ and $A_1 > A_2$ or vice-versa. Definitions II.1.1 and II.1.2 provide this exactness.

We will now express, in a mathematically exact way, the (generalized) conjecture given in the previous section.

II.2 Conjectures

Let us have a first IPP with T_1 sources, A_1 items, with rank-frequency function g_1 and size-frequency function f_1. Let us have a second IPP with corresponding notation T_2, A_2, g_2 and f_2.

Conjecture II.2.1 (Rank-frequency version):
Let $T_1 < T_2$ (alternatively: let $A_1 < A_2$), then, for $r \in (0, T_1] \subset \{x \mid 0 < x \leq T_1\}$:

$$\varphi_1(r) > \varphi_2(r) \tag{12}$$

where for $i = 1, 2$

$$\varphi_i(r) = \frac{\int_0^r g_i(r')dr'}{A_i} \tag{13}$$

Comment:

It is clear from the preceding subsection that Conjecture II.2.1 comprises two, non-equivalent, assertions: one supposing $T_1 < T_2$ and one supposing $A_1 < A_2$. Each time we can prove the validity of this assertion we must clearly indicate whether the result is true in case $T_1 < T_2$ or in case $A_1 < A_2$. For a counterexample to the above conjecture we will try to give a counterexample comprising both cases $T_1 < T_2$ and $A_1 < A_2$.

The above conjecture could be "relaxed" by requiring (13) only for a small range of values of r (smaller than $(0, T_1]$) (expressing the highest productive sources) but we will see that we can prove (13) for all $r \in (0, T_1]$ or that we can give a counterexample also for small values of r.

Based on Proposition A.1 in the Appendix, we can also formulate another conjecture as follows.

Conjecture II.2.2 (Size-frequency version):

Let $T_1 < T_2$ (alternatively: let $A_1 < A_2$), then, for $j \in (1, \min(\rho_{m,1}, \rho_{m,2}))$, where $\rho_{m,i}$ is the maximal item-density in system $i (i = 1, 2)$:

$$\psi_1(j) > \psi_2(j) \tag{14}$$

where for $i = 1, 2$

$$\psi_i(j) = \frac{\int_j^{\rho_m} j' f_i(j') dj'}{A_i} \tag{15}$$

Comment:

It is clear from Proposition A.1 that Conjecture II.2.2 is a good alternative of expressing the conjectures mentioned in Section I. One might even think, because of formula (A5) that Conjectures II.2.1 and II.2.2 are the same. This is, however, not so. This can be seen as follows. Suppose that we have (12):

$$\varphi_1(r) > \varphi_2(r)$$

for a certain $r > 0$. According to Proposition A_1 we have

$$\int_0^r g_1(r')dr' = \int_{j_1}^{\rho_m} j' f_1(j')dj' \qquad (16)$$

for $j_1 = g_1(r)$ and

$$\int_0^r g_2(r')dr' = \int_{j_2}^{\rho_m} j' f_2(j')dj' \qquad (17)$$

for $j_2 = g_2(r)$. So, if $g_1 \neq g_2$ we do not find a single $j > 1$ such that

$$\psi_1(j) > \psi_2(j).$$

In fact, in the next section we will show that Conjecture II.2.1 and Conjecture II.2.2 are <u>not</u> equivalent !

We are now in a position to study both (or even all 4) conjectures in a mathematically exact way.

III. Study of Conjectures II.2.1 and II.2.2.

Since all our results (positive and negative) are based on Lotkaian and Zipfian models (cf. the Appendix) we will, first, calculate the fractions φ and ψ in such systems.

III.1 General Lotkaian and Zipfian forms of φ and ψ

We will restrict ourselves to the case of Proposition A.2 since in this case we can prove nice results on φ and ψ and since this case also yields all necessary counterexamples.

Theorem III.1.1:

If we have that our IPP is Lotkaian with size-frequency function

$$f : [1, +\infty) \to \mathbb{R}^{+}$$

$$j \to f(j) = \frac{C}{j^{\alpha}} \tag{18}$$

$(C > 0, \alpha > 2)$ (cf.(A7)) then we have that

$$\varphi(r) = \left(\frac{r}{T}\right)^{\frac{\alpha - 2}{\alpha - 1}} \tag{19}$$

for all $r \in [0, T] \setminus \{x \mid 0 \leq x \leq T\}$ (for the notation: see formula (A8)) and

$$\psi(j) = \frac{1}{j^{\alpha - 2}} \tag{20}$$

for all $j \in [1, +\infty)$.

Proof:

Since we have (18) it follows from Proposition A.2 that

$$g(r) = \frac{C^{\frac{1}{\alpha - 1}}}{(r(\alpha - 1))^{\frac{1}{\alpha - 1}}}$$

Hence

$$\int_0^r g(r')dr' = Dr^{\frac{\alpha - 2}{\alpha - 1}} \tag{21}$$

where

$$D = \frac{C^{\frac{1}{\alpha-1}}(\alpha-1)}{(\alpha-1)^{\frac{1}{\alpha-1}}(\alpha-2)}$$

Hence

$$\varphi(r) = \frac{\int_0^r g(r')dr'}{\int_0^T g(r')dr'}$$

$$\varphi(r) = \left(\frac{r}{T}\right)^{\frac{\alpha-2}{\alpha-1}},$$

using (21). For the fraction ψ we have for all $j \in [1,+\infty)$:

$$\psi(j) = \frac{\int_j^{\infty} j' f(j')dj'}{\int_1^{\infty} j' f(j')dj'}$$

But, for all $j \geq 1$:

$$\int_j^{\infty} j' f(j')dj' = C\frac{j^{2-\alpha}}{\alpha-2} \qquad (22)$$

Hence

$$\psi(j) = j^{2-\alpha}$$

proving (20). ~

Since we do not need it in this paper we leave it as an easy exercise to prove the following extension of Theorem II.1.1, also valid for $\alpha < 2 (\alpha \neq 1)$ (but also for $\alpha > 2$).

Theorem III.1.2:

If we have (18) for $j \in [1, \rho_m]$, for $\alpha > 1$, $\alpha \neq 2$ we have that, for all $r \in [0, T]$

$$\varphi(r) = \frac{(1 + Fr)^{\frac{\alpha - 2}{\alpha - 1}} - 1}{(1 + FT)^{\frac{\alpha - 2}{\alpha - 1}} - 1} \tag{23}$$

where

$$F = \frac{\alpha - 1}{C\rho_m^{1-\alpha}} \tag{24}$$

and

$$\psi(j) = \frac{\rho_m^{2-\alpha} - j^{2-\alpha}}{\rho_m^{2-\alpha} - 1} \tag{25}$$

Hint: Use Theorem II.2.2.3 in Egghe (2005).

The cases $\alpha = 1$ and $\alpha = 2$ can also be treated; this is left to the reader, now using Theorem II.2.2.7 $(\alpha = 1)$ respectively Theorem II.2.2.1 $(\alpha = 2)$ in Egghe (2005).

One can also check easily that Theorem III.1.1 follows from Theorem III.1.2 by letting ρ_m go to $+\infty$ (and noting that, in this case, $\alpha > 2$).

We can now prove the validity of some conjectures under some parameter (e.g. α) conditions.

Theorem III.1.3:

Let us have two IPPs of Lotkaian type as in Theorem III.1.1, the first one with Lotkaian exponent $\alpha_1 > 2$ and the second one with Lotkaian exponent $\alpha_2 > 2$. Let

$$\alpha_1 \leq \alpha_2 \tag{26}$$

Let T_i, A_i denote the total number of sources, items in IPP i $(i = 1,2)$. Then

(i) If $T_1 < T_2$, then

$$\varphi_1(r) > \varphi_2(r) \tag{27}$$

for all $r \in (0, T_1]$, i.e. Conjecture II.2.1 is valid, given that the first IPP is smaller than the second IPP, in the source-sense (cf. Definition II.1.1).

(ii) If the inequality in (26) is strict, then

$$\psi_1(j) > \psi_2(j) \tag{28}$$

for all $j \in (1,+\infty) \setminus \{x \mid 1 < x\}$, i.e. Conjecture II.2.2 is valid in <u>all</u> cases where

$$\alpha_1 < \alpha_2 \tag{29}$$

applies and, in fact, (29) is necessary and sufficient for (28) to be valid.

<u>Proof</u>:

(i) We use Theorem III.1.1, formula (19) to express that (27) reads

$$\left(\frac{r}{T_1}\right)^{\frac{\alpha_1 - 2}{\alpha_1 - 1}} > \left(\frac{r}{T_2}\right)^{\frac{\alpha_2 - 2}{\alpha_2 - 1}} \tag{30}$$

Since $T_1 < T_2$ this inequality is valid if we can prove that

$$\left(\frac{r}{T_1}\right)^{\frac{\alpha_1 - 2}{\alpha_1 - 1}} \geq \left(\frac{r}{T_1}\right)^{\frac{\alpha_2 - 2}{\alpha_2 - 1}} \tag{31}$$

for all $r \in (0, T_1]$. Since $\frac{r}{T_1} \leq 1$, (31) is valid if and only if

$$\frac{\alpha_1 - 2}{\alpha_1 - 1} \leq \frac{\alpha_2 - 2}{\alpha_2 - 1} \tag{32}$$

and this is equivalent with

$$\alpha_1 \leq \alpha_2 \tag{33}$$

as is readily seen.

(ii) We use Theorem III.1.1, formula (20) to express that (28) reads

$$\frac{1}{j^{\alpha_1 - 2}} > \frac{1}{j^{\alpha_2 - 2}} \tag{34}$$

for $j \in (1, +\infty)$. But this is equivalent with

$$\alpha_1 < \alpha_2 . \tag{35}$$

as is readily seen. ~

Note that condition (35) is equivalent with (use formula (A12))

$$\mu_1 > \mu_2 \tag{36}$$

where $\mu_i = \frac{A_i}{T_i}$ is the average number of items per source in IPP i $(i = 1, 2)$.

Theorem III.1.3 contains a proof (under certain restricting conditions) of the conjecture of Aksnes and Sivertsen:

Corollary III.1.4:

If we have two Lotkaian systems with the same Lotka exponent α such that $T_1 < T_2$ then $\varphi_1(r) > \varphi_2(r)$ for all $r \in (0, T_1]$.

Proof:

This follows from Theorem III.1.3 (i).

Of course, the above corollary is more restrictive than Theorem III.1.3 itself since we assume now $\alpha = \alpha_1 = \alpha_2$. Nevertheless, Corollary III.1.4 is a partial explanation and confirmation of the conjecture of Aksnes and Sivertsen since, in comparing different fields concerning papers and their citations we can assume (as a first approximation) that $\alpha_1 \gg \alpha_2$ and hence, in this case, Corollary III.1.4 confirms Conjecture II.2.1.

From Theorem III.1.3 another important Corollary follows.

Corollary III.1.5:

Conjectures II.2.1 and II.2.2 are not equivalent.

Proof:

From the above theorem we see that, if we can make an example where $T_1 < T_2$ and $\alpha_1 = \alpha_2$ that then Conjecture II.2.1 is satisfied but Conjecture II.2.2 is not. So let us take $T_1 = 500 < T_2 = 1{,}000$ and $\alpha_1 = \alpha_2 = 3$. It follows from (A14) and (A15) that

$$A = \frac{T(\alpha - 1)}{\alpha - 2} \tag{37}$$

So $A_1 = 1{,}000 < A_2 = 2{,}000$ (so both conditions on IPP size are satisfied). We have, by Theorem III.1.1 that, for all $r \in (0, T_1]$

$$\varphi_1(r) = \left(\frac{r}{T_1}\right)^{\frac{1}{2}} > \varphi_2(r) = \left(\frac{r}{T_2}\right)^{\frac{1}{2}}$$

but

$$\psi_1(j) = \psi_2(j) = \frac{1}{j}$$

for all $j \geq 1$. ~

We close this section by providing further counterexamples, showing that, in general, none of the formulated conjectures are true.

Counterexamples III.1.6:

Let $A_1 = 1{,}000$, $T_1 = 500$, $\rho_{m,1} = +\infty$. Then by (A12) and (A13) we have $\alpha_1 = 3$, $C_1 = 1{,}000$, hence we have the size-frequency function

$$f_1(j) = \frac{1{,}000}{j^3}$$

satisfying (A10) and (A11). By (A8), the corresponding rank-frequency function is $(r \in [0, T_1])$

$$g_1(r) = \sqrt{\frac{1{,}000}{2r}}$$

Let $A_2 = 10{,}000$, $T_2 = 1{,}000$, $\rho_{m,2} = +\infty$. Then by (A12) and (A13) we have $\alpha_2 = \frac{19}{9}$, $C_2 = \frac{10{,}000}{9}$, hence we have the size-frequency function

$$f_2(j) = \frac{10,000}{9 j^{\frac{19}{9}}}$$

satisfying (A10) and (A11). By (A8), the corresponding rank-frequency function is $(r \in [0, T_2])$

$$g_2(r) = \frac{\left(\frac{10,000}{9}\right)^{\frac{9}{10}}}{\left(r \frac{10}{9}\right)^{\frac{9}{10}}}$$

Note that $T_1 < T_2$ as well as $A_1 < A_2$ so that in both the source- and item-sense, the first IPP is smaller than the second. Now we will show that Conjectures II.2.1 and II.2.2 are false. We have, by Theorem III.1.1 that

$$\psi_1(j) = \frac{1}{j} < \psi_2(j) = \frac{1}{j^{\frac{1}{9}}} \tag{38}$$

for all $j > 1$. Also by Theorem III.1.1 it follows that

$$\varphi_1(r) = \left(\frac{r}{500}\right)^{\frac{1}{2}}$$

and

$$\varphi_2(r) = \left(\frac{r}{1,000}\right)^{\frac{1}{10}}.$$

Now

$$\left(\frac{r}{500}\right)^{\frac{1}{2}} < \left(\frac{r}{1,000}\right)^{\frac{1}{10}}$$

for all

$$r \leq 420.$$

Hence, certainly for the most productive sources (compare with $r = 1$ or 5 in Aksnes and Sivertsen (2004)) we have that

$$\varphi_1(r) < \varphi_2(r) \tag{39}$$

Formulae (38) and (39) contradict both conjectures.

References

D.W. Aksnes (2003). Characteristics of highly cited papers. Research Evaluation 12(3), 159-170, 2003.

D.W. Aksnes and G. Sivertsen (2004). The effect of highly cited papers on national citation indicators. Scientometrics 59(2), 213-224, 2004.

S.C. Bradford (1934). Sources of information on specific subjects. Engineering 137, 85-86, 1934. Reprinted in: Collection Management 1, 95-103, 1976-1977. Also reprinted in: Journal of Information Science 10, 148 (facsimile of the first page) and 176-180, 1985.

L. Egghe (1990). The duality of informetric systems with applications to the empirical laws. Journal of Information Science 16(1), 17-27, 1990.

L. Egghe (2004). The source-item coverage of the Lotka function. Scientometrics, 61(1), 103-115, 2004.

L. Egghe (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, 2005.

L. Egghe and R. Rousseau (1990a). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam, 1990.

L. Egghe and R. Rousseau (1990b). Elements of concentration theory. IN: Informetrics 89/90. Proceedings of the second international Conference on Bibliometrics, Scientometrics and Informetrics, London (Canada), L. Egghe and R. Rousseau (eds.). Elsevier, Amsterdam, 97-137, 1990.

C. Gini (1909). Il diverso accrescimento delle classi sociali e la concentrazione della richezza. Giornale degli Economisti, serie 11, 37, 1909.

A.J. Lotka (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16, 317-323, 1926.

V. Pareto (1895). La legge della domanda. Giornale degli Economisti, 12, 59-68, 1895.

H. Theil (1967). Economics and Information Theory. North-Holland, Amsterdam, 1967.

Appendix

The following results are needed from continuous informetrics (cf. Egghe (1990, 2005)).

The continuous version of the size-frequency function is the strictly decreasing function

$$f : [1, \rho_m] \to \mathbb{R}^+ \tag{A1}$$

where $f(j)$ is the density of sources with item-density $j \in [1, \rho_m]$. The continuous version of the rank-frequency function is the function

$$g : [0, T] \to \mathbb{R}^+ \tag{A2}$$

where $g(r)$ is the item-density j:

$$j = g(r) \tag{A3}$$

in the continuous rank $r \in [0, T]$; T denotes the total number of sources.

The basic defining relation (but with intuition given by (1)) between f and g is

$$r = g^{-1}(j) = \int_j^{\rho_m} f(j')dj' \tag{A4}$$

for $r \in [0, T]$, $j \in [1, \rho_m]$.

We will prove the following Proposition (see Egghe (2005), Chapter II):

Proposition A.1 :

For every $r \in [0, T]$, and $j \in [1, \rho_m]$ such that $j = g(r)$ we have

$$\int_0^{r} g(r')dr' = \int_j^{\rho_m} jf(j')dj' . \qquad \text{(A5)}$$

Proof :

By (A4) we have

$$-\frac{1}{g'(g^{-1}(j))} = f(j)$$

hence

$$-\frac{j}{g'(g^{-1}(j))} = jf(j).$$

Consequently

$$\int_j^{\rho_m} \frac{-j'dj'}{g'(g^{-1}(j'))} = \int_j^{\rho_m} j'f(j')dj' . \qquad \text{(A6)}$$

Making the substitution (A3) gives $j = g(r)$, $r = g^{-1}(j)$, $dj = g'(r)dr$ and, since $g(0) = \rho_m$ (by (A4)), we have, using (A6):

$$\int_j^{\rho_m} j'f(j')dj' = \int_r^{0} \frac{-g(r')dg(r')}{g'(g^{-1}(g(r')))}$$

$$= \int_0^{r} g(r')dr'$$

which is (A5). ~

For our counterexample we also need the following result (see Egghe (2005), Chapter II) on the equivalency of the Lotka function and Zipf's function.

Proposition A.2 :

Let $\rho_m = \infty$. We have that the following assertions are equivalent:

(i) $f : [1, +\infty) \to \mathbb{R}^+$

$$j \to f(j) = \frac{C}{j^{\alpha}} \tag{A7}$$

where $C > 0$, $\alpha > 2$

(ii) $g : [0, T] \to \mathbb{R}^+$

$$r \to g(r) = \frac{C^{\frac{1}{\alpha - 1}}}{(r(\alpha - 1))^{\frac{1}{\alpha - 1}}} \tag{A8}$$

where $C > 0$, $\alpha > 2$.

Proof :

(i) $\Rightarrow$ (ii)

Using (A4) again yields

$$r = g^{-1}(j) = \int_j^{\infty} \frac{C}{j'^{\alpha}} dj'$$

$$r = g^{-1}(j) = \frac{C}{\alpha - 1} j^{1-\alpha}$$

So

$$j = g(r) = \left(r \frac{\alpha - 1}{C} \right)^{\frac{1}{1-\alpha}}$$

from which (A8) follows.

(ii) Þ (i)

By (A4) we have that

$$f(j) = -\frac{1}{g'(g^{-1}(j))}$$

$$f(j) = \frac{\alpha - 1}{\frac{C^{\frac{1}{\alpha-1}}}{(\alpha-1)^{\frac{1}{\alpha-1}}} r^{-\frac{\alpha}{\alpha-1}}}$$

but where we have to substitute (A8). This yields

$$f(j) = \frac{C}{j^{\alpha}}$$

hence (A7). ~

Finally we show the following result (also proved in Egghe (2005), Proposition II.2.1.1.1 (see also Egghe (2004)).

Proposition A.3 :

The Lotka function

$$f : [1, +\infty) \to \mathbb{R}^{+}$$

$$j \to f(j) = \frac{C}{j^{\alpha}} \qquad \text{(A9)}$$

(hence given $\rho_m = +\infty$) satisfies

$$\int_1^{+\infty} f(j)dj = T \qquad \text{(A10)}$$

(a given total number of sources) and

$$\int_1^{\infty} j f(j)\,dj = A \tag{A11}$$

(a given total number of items, $A > T$) if

$$\alpha = \frac{2A - T}{A - T} \tag{A12}$$

and

$$C = \frac{AT}{A - T} \tag{A13}$$

(which trivially implies $\alpha > 2$ for given $A < \infty$ (hence $T < \infty$)).

Proof :

Requirements (A10) and (A11) yield, for $\alpha > 2$:

$$\int_1^{\infty} f(j)\,dj = \frac{C}{\alpha - 1} = T \tag{A14}$$

$$\int_1^{\infty} j f(j)\,dj = \frac{C}{\alpha - 2} = A \tag{A15}$$

proving (A12) and (A13). ~