Study of the rank- and size-frequency functions in case of power law growth of sources and items and proof of Heaps' law

Study of the rank- and size-frequency functions in the case of power law growth of sources and items and proof of Heaps' law

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]

and

Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium

leo.egghe@uhasselt.be

ABSTRACT

Supposing that the number of sources and the number of items in sources grow in time according to power laws, we present explicit formulae for the size- and rank-frequency functions in such systems. Size-frequency functions can decrease or increase while rank-frequency functions only decrease. The latter can be convex, concave, S-shaped (first convex, then concave) or reverse S-shaped (first concave, then convex). We also prove that, in such systems, Heaps' law on the relation between the number of sources and items is valid.

---

[1] Permanent address

Key words and phrases: power law growth, rank-frequency function, size-frequency function, Heaps' law

Acknowledgement: The author is grateful to two anonymous referees for valuable suggestions to improve this paper. He is especially grateful to one referee who pointed out a mistake in Theorem 2 in an earlier version of this paper.

Introduction

In 1970, Naranan (1970) proved his famous theorem in the journal Nature: if the number of sources (e.g. journals) grow exponentially in time and if the number of items (e.g. articles) in these sources also grow exponentially in time (according to a possibly different exponential function than the one of the sources), then the size-frequency function $f(j)$ (the number of sources with j items) is a decreasing power law, i.e. is Lotka's law (see Egghe (2005a)).

Stated more exactly the theorem of Naranan is as follows.

Theorem (Naranan (1970)):

Suppose

(i) Sources grow exponentially in time t:

$$T(t) = c_1 a_1^t \tag{1}$$

$(c_1 > 0,\ a_1 > 1)$

(ii) Items in sources grow exponentially in time and their growth rate in sources is the same for every source:

$$p(t) = c_2 a_2^t \tag{2}$$

$(c_2 > 0,\ a_2 > 1)$

then the size-frequency function $f(j)$ has the form

$$f(j) = \frac{C}{j^{\alpha}} \tag{3}$$

where $C > 0$ is dependent on t, $j \geq 1$ and

$$\alpha = 1 + \frac{\ln a_1}{\ln a_2} \tag{4}$$

Formula (3) is called the law of Lotka and α is called Lotka's exponent ($\alpha > 1$). Formula (4) is linked with the fractal dimension of this system (see Egghe (2005a,b)).

It is well-known that Lotka's law is equivalent with Zipf's law $g(r)$ (being the number of items in the source on rank r, where sources are ranked in decreasing order of their number of items):

$$g(r) = \frac{B}{r^{\beta}} \tag{5}$$

$B > 0$, $\beta > 0$ and $r \in [0,T]$, T being the total number of sources.

The relation between α and β is

$$\beta = \frac{1}{\alpha - 1} \tag{6}$$

Note that all arguments are in continuous setting (all quantities are densities, which is the same in continuous probability theory). The equivalence between (3) and (5) is described in Egghe (2005a) or in Egghe and Rousseau (2006), Appendix, where also formula (6) is proved.

The basis of the proof of the equivalence of (3) and (5) is the relation

$$r = g^{-1}(j) = \int_j^{\infty} f(j')dj' \tag{7}$$

Which describes the relation between the source rank density r and the item density j: formula (7) describes the number of sources with item density j or higher; hence the source with this item density j has rank density r. Formula (7) is universal and not related to Lotkaian informetrics. It will be used in this paper as well.

Clearly, both laws (3) and (5) are convexly decreasing functions. Recently this author was able to give a short algebraic proof of the theorem of Naranan – see Egghe (2010): Naranan proved (3) and (4) and Egghe proved (5) which is equivalent with (3) (as mentioned above) and where we showed that $\beta = \frac{\ln a_2}{\ln a_1}$ which yields (4) by (6).

In this article we are interested in the "power law" variant of the theorem of Naranan: sources and items grow according to a power law (exact formulation will be given in the next section). We consider this to be an important problem since it is known that not all growth functions are exponential (Price (1963), Egghe an Rao (1992)). In addition, power laws are one of the most important models in mathematics and furthermore they are very simple. What size- and rank-frequency function do we have in such a system? This is the topic of this paper. In the next section we will prove explicit formulae for the size-frequency function $f(j)$ and the rank-frequency function $g(r)$. Contrary to the Naranan case we now can have (besides

decreasing) increasing size-frequency functions (or even non-monotonic functions). The rank-frequency function always decreases (as in Naranan's case) but is not always convex (as in Naranan's case): $g(r)$ can be convex, concave, or have an S-shape (first convex, then concave) or a reverse S-shape (first concave, then convex).

In the third section we show that Heaps' law (in linguistics also called the law of Herdan) is valid in this system. Heaps' law states (see also Heaps (1978), Herdan (1960, 1964) or Egghe (2007)) that, in a growing system of sources and items, denoting by T the total number of sources and by A the total number of items we have

$$T = KA^{\gamma} \tag{8}$$

where $K > 0$ is a constant and $0 < \gamma < 1$ is a fixed exponent (independent of time). In other words the number of sources is a concavely (also called sublineary) increasing function of the number of items. In linguistics (Herdan) this is described as: vocabulary size (distinct words, also called "type") is a concavely increasing function of text size (used words, also called "token").

The paper closes with some conclusions and open problems.

Study of the system in which the number of sources and number of items in sources grow according to a power law

Similar to the Naranan formalism we suppose the following assertions

(i) Sources grow according to a power law in t:

$$T(t) = c_1 t^{a_1} \tag{9}$$

where $c_1 > 0$, $a_1 > 0$

(ii) Items in sources grow according to a power law in t and their growth rate in sources is the same for every source:

$$p(t) = c_2 t^{a_2} \tag{10}$$

where $c_2 > 0$, $a_2 > 0$.

Note that the above also includes linear growth in case $a_1 = 1$ and/or $a_2 = 1$.

We adopt the method, developed in Egghe (2010) for the Naranan framework which yields a closed expression for the rank-frequency function $g(r)$.

Theorem 1:

In the framework of power law growth of the number of sources and of the number of items we have the following expression for the rank-frequency function $g(r)$ in function of time t:

$$g(r) = c_2 \left(t - \left(\frac{r}{c_1} \right)^{\frac{1}{a_1}} \right)^{a_2} \tag{11}$$

where $r \in \left[0, c_1 t^{a_1} \right]$ and where c_1, c_2, a_1 and a_2 are the parameters appearing in (i) and (ii) above.

Proof:

The defining equation for the rank-frequency function $g(r)$ is, by (9) and (10): for every $\theta \in [0,1]$:

$$g(r) = g\left(c_1 (\theta t)^{a_1} \right) = c_2 (t - \theta t)^{a_2} \tag{12}$$

where $r = c_1 (\theta t)^{a_1}$ from which we derive

$$\theta = \left(\frac{r}{c_1} \right)^{\frac{1}{a_1}} \frac{1}{t} \tag{13}$$

But (12) and (13) combined yield

$$g(r) = c_2 t^{a_2} \left(1 - \left(\frac{r}{c_1} \right)^{\frac{1}{a_1}} \frac{1}{t} \right)^{a_2}$$

from which (11) readily follows.

□

Now we study the shape of $g(r)$. It will turn out that $g(r)$ is decreasing (as in the Naranan case) but that it has several different shapes (in the Naranan case, $g(r)$ was always convex).

Theorem 2:

$g(r)$ in (11) is always strictly decreasing. It has an inflection point if and only if

$$1 \geq \frac{1-a_1}{a_2-a_1} > 0 \tag{14}$$

and it has no inflection point if and only if

$$\frac{1-a_1}{a_2-a_1} \leq 0 \text{ or } \frac{1-a_1}{a_2-a_1} > 1 \tag{15}$$

or if $a_1 = a_2$. In case of (14), $g(r)$ has an S-shape (first convex, then concave) if and only if $a_1 > 1$ and it has a reverse S-shape (first concave, then convex) if and only if $0 < a_1 < 1$. In case of the first inequality of (15), $g(r)$ is convex if and only if $a_2 > a_1$ or if $a_1 = a_2$ and $a_1 > 1$ and $g(r)$ is concave if and only if $a_2 < a_1$ or if $a_1 = a_2$ and $0 < a_1 < 1$. In case of the second inequality of (15), $g(r)$ is convex or concave for non-determined values of a_1 and a_2. If $a_1 = a_2 = 1$, then $g(r)$ is a decreasing straight line.

Proof:

$$g'(r) = c_2 a_2 \left(t - \left(\frac{r}{c_1} \right)^{\frac{1}{a_1}} \right)^{a_2 - 1} \left(-\frac{1}{a_1 c_1} \left(\frac{r}{c_1} \right)^{\frac{1}{a_1} - 1} \right)$$

which is clearly negative (hence $g(r)$ strictly decreases which is also clear from (11) or (12)). So we have

$$g'(r) = D \left(t - \left(\frac{r}{c_1} \right)^{\frac{1}{a_1}} \right)^{a_2 - 1} \left(\frac{r}{c_1} \right)^{\frac{1}{a_1} - 1} \tag{16}$$

where $D = -\frac{c_2 a_2}{c_1 a_1} < 0$.

$$g''(r) = D(a_2 - 1) \left(t - \left(\frac{r}{c_1} \right)^{\frac{1}{a_1}} \right)^{a_2 - 2} \left(-\frac{1}{a_1 c_1} \left(\frac{r}{c_1} \right)^{\frac{1}{a_1} - 1} \right) \left(\frac{r}{c_1} \right)^{\frac{1}{a_1} - 1}$$

$$+ D \left(t - \left(\frac{r}{c_1} \right)^{\frac{1}{a_1}} \right)^{a_2 - 1} \left(\frac{1}{a_1} - 1 \right) \left(\frac{r}{c_1} \right)^{\frac{1}{a_1} - 2} \frac{1}{c_1}$$

Noting that $D<0$, $g"(r)$ has the sign of

$$(a_2-1)\frac{1}{a_1}\left(\frac{r}{c_1}\right)^{\frac{1}{a_1}}-\left(t-\left(\frac{r}{c_1}\right)^{\frac{1}{a_1}}\right)\left(\frac{1}{a_1}-1\right) \tag{17}$$

If we denote

$$x=:\left(\frac{r}{c_1}\right)^{\frac{1}{a_1}} \tag{18}$$

for simplicity, we see that $g"(r)$ has the sign of

$$x(a_2-a_1)-t(1-a_1) \tag{19}$$

Hence $g(r)$ has no inflection point if $a_1=a_2$: if $0<a_1<1$, then $g(r)$ is concave since (19) is <0; if $a_1>1$ then $g(r)$ is convex since (19) is >0. Let us now assume that $a_1\neq a_2$. Then $g(r)$ has an inflection point if and only if (x must be $\leq t$ by (18) and since $r\leq c_1t^{a_1}$)

$$t\geq x_0=\frac{t(1-a_1)}{a_2-a_1}>0 \tag{20}$$

(since $g(r)$ is not defined on negative values) and has no inflection if and only if

$$\frac{1-a_1}{a_2-a_1}\leq 0 \text{ or } \frac{1-a_1}{a_2-a_1}>1 \tag{21}$$

(20) and (21) prove (14) and (15) since $t>0$.

Let us now assume (15). Suppose that we have the first inequality in (15). If $a_2>a_1$, then by (15) $a_1>1$ and hence the sign of (19) is >0. Hence $g(r)$ is convex. If $a_2<a_1$ then by (15), $0<a_1<1$ and hence the sign of (19) is <0. Hence $g(r)$ is concave. Suppose that we have the second inequality in (15). If $a_2>a_1$ then $a_2<1$ hence $a_1<1$ and the sign of (19) is not determined. If $a_2<a_1$ then $a_2>1$ hence $a_1>1$ and the sign of (19) is not determined.

Let us now assume (14). Hence in this case, $g(r)$ has an inflection point. In $r=x=0$, (19) (hence $g"(0)$) has the sign of

$$a_1 - 1 \tag{22}$$

since $t > 0$. If $0 < a_1 < 1$ then (22) is < 0 and hence $g(r)$ starts concave and then, after the inflection point (20), is convex. Hence $g(r)$ has a reverse S-shape. If $a_1 > 1$, then (22) is > 0 and hence $g(r)$ starts convex and then, after the inflection point (20), is concave. Hence $g(r)$ has an S-shape.

□

Further properties of the rank-frequency function are given in Theorem 3.

Theorem 3:

(i) $g(0) = c_2 t^{a_2}$ (23)

(ii) $g\left(c_1 t^{a_1}\right) = 0$ (24)

If $g(r)$ in $r = 0$ is convex then $g'(0) = -\infty$. If $g(r)$ in $r = 0$ is concave then $g'(0) = 0$. If $g(r)$ in $r = c_1 t^{a_1}$ is convex then $g'\left(c_1 t^{a_1}\right) = 0$. If $g(r)$ in $r = c_1 t^{a_1}$ is concave then $g'\left(c_1 t^{a_1}\right) = -\infty$.

Proof:

(23) and (24) follow readily from (11). From (16) we have that $g'(0) = 0$ if $0 < a_1 < 1$ and $g'(0) = -\infty$ for $a_1 > 1$ (since $D < 0$). Hence this are the only two possible values for $g'(0)$. From (16) we have that $g'\left(c_1 t^{a_1}\right) = 0$ if $a_2 > 1$ and that $g'\left(c_1 t^{a_1}\right) = -\infty$ if $0 < a_2 < 1$, since $D < 0$. Hence this are the only two possible values for $g'\left(c_1 t^{a_1}\right)$.

These results prove Theorem 3 since, if $g(r)$ is convex in $r = 0$, $g'(0) \neq 0$ and if $g(r)$ is concave in $r = 0$, $g'(0) \neq -\infty$. The same argument if $r = c_1 t^{a_1}$. □

This means that, according to the values of a_1 and a_2, described in Theorem 2, we have the possible shapes of $g(r)$ as shown in Figs. 1, 2, 3, 4.

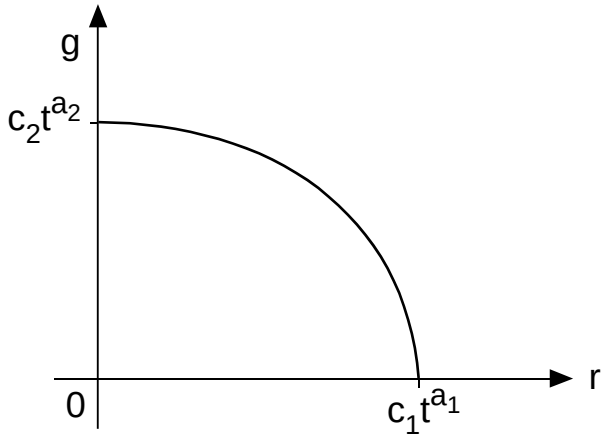

Fig.1. Concave shape of $g(r)$

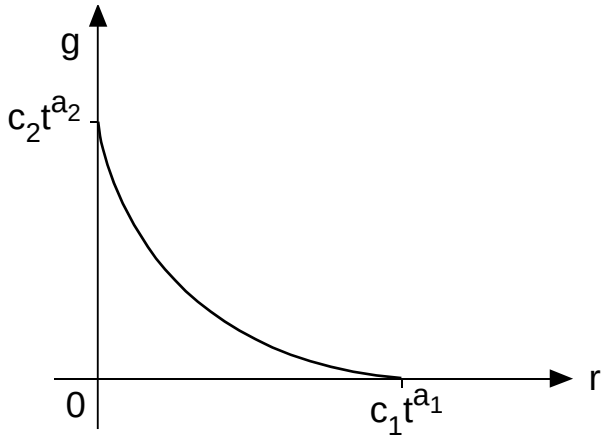

Fig.2. Convex shape of $g(r)$

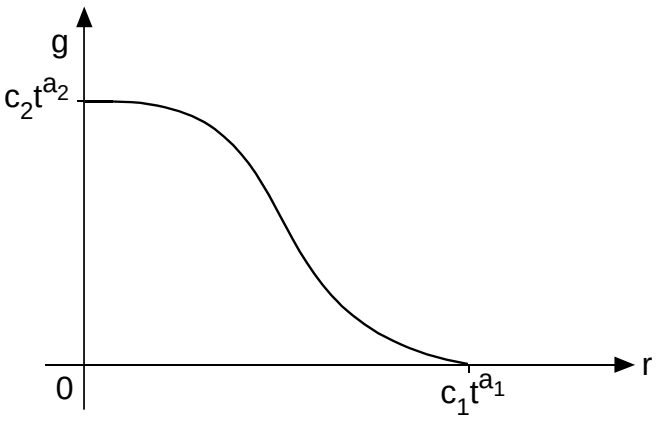

Fig.3. Reverse S-shape of $g(r)$

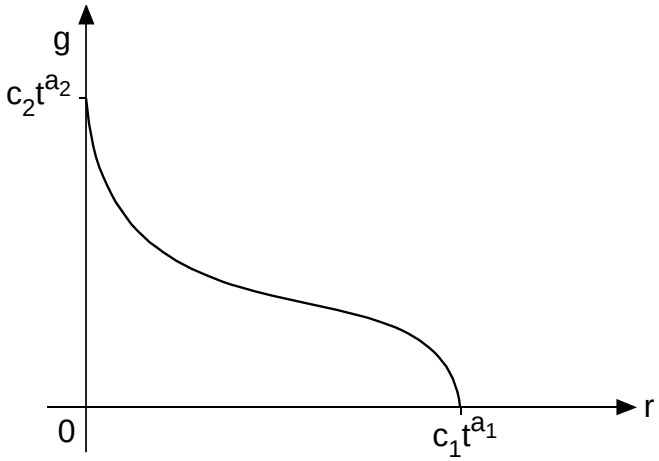

Fig.4. S-shape of $g(r)$

If $g(r)$ has an inflection point (Figs. 3, 4) then this point is $(r_0, g(r_0))$, where

$$r_0 = c_1\left(t\frac{1-a_1}{a_2-a_1}\right)^{a_1} \tag{25}$$

$$g(r_0) = c_2\left(t\frac{a_2-1}{a_2-a_1}\right)^{a_2} \tag{26}$$

This follows from (20) and (11). Note that here $a_1 \neq a_2$ since in this case $g(r)$ has no inflection point (see Theorem 2).

Fig. 2 is a classical shape for a rank-frequency function. The S-shape in Fig. 4 occurs, e.g., in Mansilla et al. (2007) Martinez-Mekler et al. (2009), Lavalette (1996), Campanario (2010), Egghe (2009) and Egghe and Waltman (2011) both for the rank-order function of the impact factor $IF(r)$ and its logarithm $\ln(IF(r))$.

Now we turn our attention to the size-frequency function f that is equivalent with the already studied rank-frequency function g.

The equivalence between the rank-frequency function $g(r)$ and the size-frequency function $f(j)$ is expressed by the equation (7), from which it follows that

$$f(j) = -\frac{1}{g'(g^{-1}(j))} \tag{27}$$

From (27) we can deduce the function $f(j)$ based on the results that we have obtained on the function $g(r)$. This is described in the next theorem.

Theorem 4:

Under the assumptions (i) and (ii) in this paper we have that the size-frequency function $f(j)$ has the following expression

$$f(j)=\frac{a_1c_1}{a_2c_2^{\frac{1}{a_2}}\left(t-\left(\frac{j}{c_2}\right)^{\frac{1}{a_2}}\right)^{1-a_1}j^{1-\frac{1}{a_2}}} \tag{28}$$

Proof:

According to (27) we need expressions for g' and for $g^{-1}(j)$. The derivative g' has already been calculated – see (16). For $g^{-1}(j)$, note that, by (11),

$$j=g(r)=c_2\left(t-\left(\frac{r}{c_1}\right)^{\frac{1}{a_1}}\right)^{a_2}$$

where

$$r=g^{-1}(j)=c_1\left(t-\left(\frac{j}{c_2}\right)^{\frac{1}{a_2}}\right)^{a_1} \tag{29}$$

Note that (29) has the same form as (11) upon replacing r by j and the index 1 by 2 and vice-versa. Now (16), (29) and (27) yield

$$f(j)=\frac{a_1c_1}{a_2c_2}\frac{1}{\left(\frac{j}{c_2}\right)^{1-\frac{1}{a_2}}\left(t-\left(\frac{j}{c_2}\right)^{\frac{1}{a_2}}\right)^{1-a_1}}$$

from which (28) follows. □

In the Appendix we present a second proof of (28), again based on (27) but now using the implicit defining relation (12) for g, hence confirming the correctness of (28).

We will now study the increasing and decreasing aspects of the size-frequency function $f(j)$ in (28).

(1) Let $a_1 = 1$.

Then, by (28)

$$f(j) = \frac{C}{j^{1-\frac{1}{a_2}}} \tag{30}$$

where $C = \frac{a_1 c_1}{a_2 c_2^{1/a_2}}$. If $a_2 > 1$ we then have that $f(j)$ decreases strictly. We could say that (30) is then Lotka's law but with the unusual exponent α (Egghe (2005))

$$0 < \alpha = 1 - \frac{1}{a_2} < 1 \tag{31}$$

If $0 < a_2 < 1$, then $f(j)$ strictly increases in j. For the cases $a_1 > 1$ and $0 < a_1 < 1$ we need the following Lemma, proved in Egghe and Waltman (2010) (see Lemma II.3 there).

Lemma 1 (Egghe and Waltman (2011)):

Let f and g be general size- (rank-)frequency functions. Let g be decreasing. Then we have the following equivalence: f is first increasing and then decreasing if and only if g has an S-shape, first convex and then concave.

We also need the following Lemma which is proved in the same way

Lemma 2:

Let f and g be as above and let g be decreasing. Then we have the following equivalence: f is first decreasing and then increasing if and only if g has a reverse S-shape, first concave and then convex.

The proof of Lemma 2 follows the lines of Theorem II.5 (using (6)) in Egghe and Waltman (2011) and Lemma II.3 (with g replaced by g^{-1}) in Egghe and Waltman (2011).

We now start the study of the cases $a_1 > 1$ and $0 < a_1 < 1$.

(2) Let $a_1 > 1$.

If $a_2 > 1$, then it is clear from (28) that $f(j)$ decreases strictly in j.

If $0 < a_2 < 1$, then (20) is valid. Hence $g(r)$ has an inflection point. But, since $a_1 > 1$, (22) is positive. As argued below (22), g has an S-shape (first convex, then concave). By Lemma 1, f is first increasing and then decreasing.

(3) Let $0 < a_1 < 1$.

If $0 < a_2 < 1$, then it is clear from (28) that $f(j)$ strictly increases in j.

If $a_2 > 1$, then (20) is positive and hence $g(r)$ has an inflection point. But, since $0 < a_1 < 1$, (22) is negative. As argued below (22), g has a reverse S-shape (first concave, then convex). By Lemma 2, f is first decreasing, and then increasing.

Proof of the validity of Heaps' law (or Herdan's law) in systems where sources and items grow according to a power law

So we suppose that assumptions (i) and (ii) of the previous section are valid. In formula (9), $T(t)$ denotes the total number of sources at time t. Let us denote by $A(t)$ the total number of items at time t. In the next theorem we show the validity of Heaps' (or Herdan's) law.

Theorem 5:

Heaps' (or Herdan's) law is valid: there exists a constant $K > 0$ and a constant γ, $0 < \gamma < 1$, independent of time t such that, for every $t > 0$ we have

$$T(t) = KA(t)^{\gamma} \tag{32}$$

Furthermore

$$\gamma = \frac{1}{1 + \frac{a_2}{a_1}} \tag{33}$$

Proof:

Formula (9) yields

$$T(t) = c_1 t^{a_1} \tag{34}$$

The equation (10) or (12) imply that the total number of items in all sources at time t equals

$$A(t) = \int_0^1 c_1 (\theta t)^{a_1} c_2 (t - \theta t)^{a_2} \, d\theta \tag{35}$$

Indeed θt goes from 0 to t, covering the growth of sources and items up to time t. Hence (35) equals

$$A(t) = c_1 c_2 t^{a_1 + a_2} \int_0^1 \theta^{a_1} (1 - \theta)^{a_2} \, d\theta \tag{36}$$

The integral in (36) cannot be evaluated in an analytical way. In fact it is the beta function

$$B(a_2 + 1, a_1 + 1) = \int_0^1 \theta^{a_1} (1 - \theta)^{a_2} \, d\theta \tag{37}$$

However, what is important here is that (37) is a constant. Denote

$$D = c_1 c_2 B(a_2 + 1, a_1 + 1)$$

then we have that

$$A(t) = D t^{a_1 + a_2} \tag{38}$$

Formula (34) implies

$$t = \left(\frac{T(t)}{c_1} \right)^{\frac{1}{a_1}} \tag{39}$$

(39) in (38) yields

$$A(t) = D \left(\frac{T(t)}{c_1} \right)^{\frac{a_1 + a_2}{a_1}}$$

Hence

$$T(t) = c_1 \left(\frac{A(t)}{D} \right)^{\frac{a_1}{a_1 + a_2}} \tag{40}$$

Denote

$$K = \frac{c_1}{D^{\frac{a_1}{a_1 + a_2}}}$$

, then (40) reads

$$T(t) = KA(t)^{1/\left(1+\frac{a_2}{a_1}\right)} \tag{41}$$

hence Heaps' (or Herdan's) law with constant exponent

$$0 < \gamma = \frac{1}{1+\frac{a_2}{a_1}} < 1 \qquad \square$$

Note that γ only depends on the quotient $\frac{a_2}{a_1}$ of the power law exponents in (9) and (10).

Conclusions and open problems

In this paper we studied a variant of the Naranan formalism: instead of exponential growth of the number of sources and items (by Naranan (1970)), we suppose that the number of sources and items grow according to a power law. We prove explicit formulae for the rank- and size-frequency functions in such systems. The latter can decrease and/or increase (in Naranan's model, there is only a power law decrease: Lotka's law) while the former can only decrease (by definition (7)) but it can have the convex, concave, S-shape or reverse S-shape (while in Naranan's case it can only decrease convexly: Zipf's law).

We also proved the validity of Heaps' law (from information retrieval) (or equivalently Herdan's law from linguistics) in these systems.

This paper showed that informetric systems do not always have Lotka's law as a size-frequency function or Zipf's law as a rank-frequency function: based on a simple assumption of power law growth of items and sources we could prove (11) for the rank-frequency function and (28) for the size-frequency function. This implies that many results from Lotkaian informetrics (see Egghe (2005a)) could be re-studied in the present context. Besides these open theoretical problems, we also need empirical results to show the use-ability of the present results.

The fact that Heaps' law is a consequence of power law growth of sources and items and the fact that Heaps' law (Herdan's law) occurs in many cases in informetrics and linguistics,

shows that the growth model studied here has some advantage over other source-item growth models.

It would be very interesting to study other variants of the Naranan formalism, e.g. where sources grow exponentially and items (in sources) grow according to a power law (or vice-versa).

References

J.M. Campanario (2010). Distribution changes in impact factors over time. Scientometrics, 84(1), 35-42.

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

L. Egghe (2005b). The power of power laws and the interpretation of Lotkaian informetric systems as self-similar fractals. Journal of the American Society for Information Science and Technology, 56(7), 669-675.

L. Egghe (2007). Untangling Herdan's law and Heaps' law: mathematical and informetric arguments. Journal of the American Society for Information Science and Technology, 58(5), 702-709.

L. Egghe (2009) Mathematical derivation of the impact factor distribution. Journal of Informetrics, 3(4), 290-295.

L. Egghe (2010). A new short proof of the Theorem of Naranan, explaining the laws of Lotka and Zipf. Journal of the American Society for Information Science and Technology, to appear.

L. Egghe and I.K.R. Rao (1992). Classification of growth models based on growth rates and its applications. Scientometrics 25(1), 5-46.

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

L. Egghe and L. Waltman (2011). Relations between the shape of a size-frequency distribution and the shape of a rank-frequency distribution. Information Processing and Management, to appear.

H.S. Heaps (1978). Information Retrieval. Computational and theoretical Aspects. Academic Press, New York, USA.

G. Herdan (1960). Type-token Mathematics. A textbook of mathematical Linguistics. Mouton, 's Gravenhage, the Netherlands.

G. Herdan (1964). Quantitative Linguistics. Butterworths, London, UK.

D. Lavalette (1996). Facteur d'impact: impartialité ou impuissance? Report INSERM U350, Institut Curie-Recherche, Bât. 112, Centre Universitaire, 91405, Orsay, France.

R. Mansilla, E. Köppen, G. Cocho and P. Miramontes (2007). On the behavior of journal impact factor rank-order distribution. Journal of Informetrics, 1(2), 155-160.

G. Martinez-Mekler, R. Alvarez Martinez, M. Beltrán del Rio, R. Mansilla, P. Miramontes and G. Cocho (2009). Universality of rank-order distributions in the arts and sciences. PLosONE, 4(3), e4791.

S. Naranan (1970). Bradford's law of bibliography of science: an interpretation. Nature, 227, 631-632.

D.D.S. Price (1963). Little science, big science. Columbia University Press, New York, USA.

Appendix

Second proof of formula (28).

The defining relation for $g(r)$ is (12): for $r = c_1(\theta t)^{a_1}$ ($\theta \in [0,1]$),

$$g(r) = g\left(c_1 t^{a_1}\theta^{a_1}\right) = c_2 t^{a_2}(1-\theta)^{a_2} \tag{A1}$$

Since $r = c_1(\theta t)^{a_1}$ we have

$$\theta = \left(\frac{r}{c_1}\right)^{\frac{1}{a_1}} \frac{1}{t} \tag{A2}$$

Furthermore

$$\begin{aligned} g'(r) &= \frac{dg(r)}{dr} = \frac{dg(r)}{d\theta}\frac{d\theta}{dr} \\ &= -c_2 t^{a_2} a_2 (1-\theta)^{a_2-1} \frac{1}{t}\frac{1}{c_1^{1/a_1}}\frac{1}{a_1} r^{\frac{1}{a_1}-1} \\ &= \frac{-c_2 t^{a_2-1} a_2}{c_1^{1/a_1} a_1} r^{\frac{1}{a_1}-1}\left(1-\left(\frac{r}{c_1}\right)^{\frac{1}{a_1}}\frac{1}{t}\right)^{a_2-1} \end{aligned} \tag{A3}$$

by (A1) and (A2). By (A1),

$$\begin{aligned} r = g^{-1}(j) &= g^{-1}\left(c_2 t^{a_2}(1-\theta)^{a_2}\right) \\ &= c_1 t^{a_1}\theta^{a_1} \end{aligned} \tag{A4}$$

Since $j = g(r)$, we have, by (A1),

$$j = c_2 t^{a_2}(1-\theta)^{a_2}$$

whence

$$\theta = 1 - \left(\frac{j}{c_2}\right)^{\frac{1}{a_2}} \frac{1}{t} \tag{A5}$$

Hence (A4) and (A5) yield

$$r = g^{-1}(j) = c_1 t^{a_1}\left(1 - \left(\frac{j}{c_2}\right)^{\frac{1}{a_2}}\frac{1}{t}\right)^{a_1} \tag{A6}$$

Finally (27), (A3) and (A6) yield (28) after some elementary calculation.