Dynamical aspects of Information Production Processes

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

In this paper and talk we present a short proof of a theorem of Naranan stating that if sources grow exponentially and if items in sources also grow exponentially, then the system is Lotkaian, i.e. its size-frequency function is the law of Lotka.

We apply this technique to the case of power law growth of sources and of items in sources and determine the size- and rank frequency functions in this case. These functions have a greater variety of shapes than in the classical Naranan case and we give practical examples. We also show that, in this context, the law of Heaps can be proved.
We also further generalise this technique to general growth models of sources and items in sources.

---

[1] Permanent address
Key words and phrases: Dynamical aspects, Naranan, growth

Introduction

Information Production Processes (IPPs) can be generally described by sources that have (or produce) items. Many examples (in informetrics and beyond) can be given.

sources	→	items
authors		articles
journals		articles
articles		citations, references
articles		authors
books		borrowings
words (types)		occurence in a text (tokens)
websites		hyperlinks (in/out)
cities, villages		inhabitants
employees		production
employees		salaries
...		

To measure these IPPs we use two related functions.

(i) The size-frequency function *f*:

for each $n = 1, 2, 3, \ldots$ $f(n)$ is the number of sources with *n* items.

If we rank the sources in decreasing order of the number of items they have, we can define the second function:

(ii) The rank-frequency function *g*:

for each $r=1,2,...,T$, $g(r)$ is the number of items in the source on rank r (T = total number of sources).

In informetric models we use continuous variables. Here, for $j \geq 1$, $f(j)$ is the density of the sources with item density j and, for $r \in [0,T]$, $g(r)$ is the item density in the source density r. In this context, the size-frequency function f and the rank-frequency function g are related as in (1) and (2)

$$r = \int_{g(j)}^{\infty} f(j)\,dj \tag{1}$$

where $j = g(r)$ and

$$f(j) = \frac{1}{g(g^{-1}(j))} \tag{2}$$

where g^{-1} is the inverse function of the (injective) function g (g is injective since it strictly decreases, by (1)) and g' is the derivative of g.

Obviously (1) implies (2) and (2) implies (1) given that $g(0) = \infty$ (hence $g^{-1}(\infty) = 0$). For these introductory results, we refer to Egghe (2005).

General shape-relations between f and g can be proved, based on (1) and (2). They are given in Egghe and Waltman (2011):

Proposition 1:

(A) f is decreasing if and only if g is convex.

(B) f is first increasing and then decreasing if and only if g has an S-shape: first convex and then concave.

Also in Egghe and Waltman (2011) examples of both cases are given, see Fig. 1.

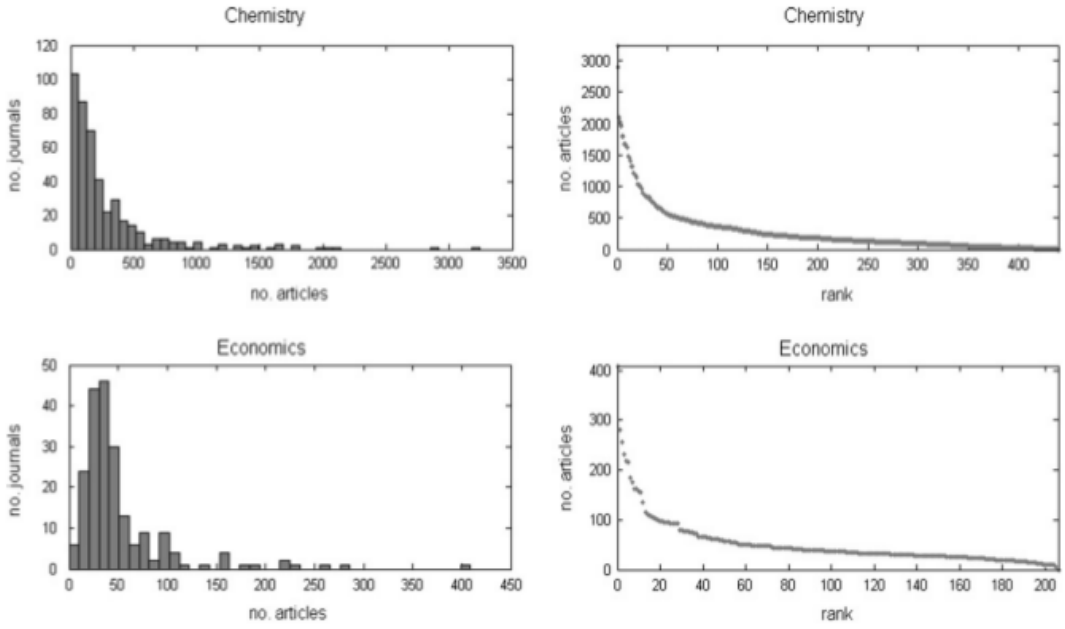

Fig. 1. Size-(left) and rank-(right) frequency functions in chemistry (type (A)) and in economics (type (B)), taken from Egghe and Waltman (2011), p. 242.

The most classical examples of size- and rank-frequency functions are the laws of Lotka and Zipf, respectively, being decreasing power laws:

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

$C>0$, $\alpha>1$, $j\geq 1$ and

$$g(r)=\frac{B}{r^{\beta}} \tag{4}$$

$B,\beta>0$, $r\in]0,T]$.

We have the following well-known result (see Egghe (2005), Egghe and Rousseau (2006) where a proof is available in the Appendix).

Proposition 2: The following assertions are equivalent:

(i) We have Lotka's law (3)

(ii) We have Zipf's law (4)

In this case the exponents α and β relate as in (5)

$$\beta = \frac{1}{\alpha - 1} \tag{5}$$

In Naranan (1970) there is the following rationale for Lotka's law given:

Proposition 3: Let us have an IPP in which

(i) The number of sources $\varphi(t)$ grows exponentially in time *t*, denoted as

$$\varphi(t) = c_1 a_1^t \tag{6}$$

(ii) The number of items $\psi(t)$ in each source grows exponentially in time *t* and the growth is the same in each source, denoted as

$$\psi(t) = c_2 a_2^t \tag{7}$$

($c_1, c_2 > 0$, $a_1, a_2 > 1$). Then this IPP satisfies Lotka's law (3) and we have formula (8) for the relation between Lotka's exponent α and the growth rates a_1 and a_2:

$$\alpha = 1 + \frac{\ln a_1}{\ln a_2} \tag{8}$$

The proof is long and clarified in Egghe (2005). In Egghe (2010) we presented a short proof of this result whereby we prove Zipf's law. Hence, by Proposition 2 we have also shown Lotka's law. In the next section we will present this short proof in the even shorter version as given in Egghe (2012b).

The third section deals with an important variant of Naranan's result: we assume that the growth function φ and ψ are power laws (instead of exponential laws in Naranan) – see Egghe (2012a). Now we receive other laws for the size- and rank-frequency functions $f(j)$ and $g(r)$. Contrary to the Naranan case we do not only have case (A) in Proposition 1 (for *f* being Lotka's law and g being Zipf's law) but we can also have case (B) in Proposition 1 (and we also have two other possibilities but for which we have no empirical evidence).

In this case we can also give a proof of Heaps' law (or Herdan's law) (see Heaps (1978), Herdan (1960,1964), Egghe (2007)) relating the total number *T* of sources to the total number *A* of items as in formula (9):

$$T = K.A^{\gamma} \tag{9}$$

where $K > 0$ is a constant and $0 < \gamma < 1$ is a fixed exponent (independent of time) – see also Egghe (2012a). Such a result is not valid in the Lotkaian case.

In the last section we further generalise the Naranan formalism by assuming general growth functions $\varphi(t)$ (for the sources) and $\psi(t)$ (for the items) and we present the general relation with the size- and rank-frequency functions $f(j)$ and $g(r)$.

Short proof of the Theorem of Naranan (Egghe (2010, 2012b))

Let *t* be a fixed time period (e.g. the present). The further we look back in the past, the longer time sources have to grow and hence their rank densities *r* are described by

$$r = c_1 a_1^{\theta t} \tag{10}$$

for $\theta \leq 1$ (by (6)).

By (7) we have for the item density on rank density *r*:

$$g(r) = c_2 a_2^{t-a} \tag{11}$$

Combining (10) and (11), using that

$$\theta = \frac{\ln\left(\frac{r}{c_1}\right)}{t \ln a_1} \tag{12}$$

yields

$$g(r) = \frac{c_2 a_2^{t}}{a_2^{\frac{\ln\left(\frac{r}{c_1}\right)}{\ln a_1}}} \tag{13}$$

$$g(r)=\frac{c_2 d_2^{\,i}}{e^{\frac{\ln a_2}{\ln a_1}\ln\left(\frac{r}{c_1}\right)}} \tag{14}$$

$$g(r)=\frac{B}{r^{\beta}} \tag{15}$$

where

$$\beta=\frac{\ln a_2}{\ln a_1} \tag{16}$$

and

$$B=c_2 d_2^{\,i}\left(c_1^{\frac{\ln a_2}{\ln a_1}}\right) \tag{17}$$

Hence we have shown that Zipf's law is valid. By Proposition 2 we hence have proved Lotka's law and, using (5) and (16) we have that

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

So also (8) is proved. □

Because of formula (8),

$$D=\alpha-1=\frac{\ln q}{\ln 2}$$

Hence a Lotkaian IPP can be viewed as a self-similar fractal with fractal dimension $D=\alpha-1$.

Naranan dynamics in case of power law growth of sources and items

If we replace in Proposition 3 the exponential growth of sources and items by power law growth we obtain the result in Proposition 4.

Proposition 4 (Egghe (2012a)): Let us have an IPP in which

(i) The number of sources $\varphi(t)$ grows according to a power law in time *t*, denoted as

$$\varphi(t)=c_1 t^{a_1} \tag{18}$$

(ii) The number of items $\psi(t)$ in each source grows according to a power law in time *t* and the growth is the same in each source, denoted as

$$\psi(t)=c_2 t^{a_2} \tag{19}$$

$(c_1, c_2, a_1, a_2 > 0)$.

Then this IPP has the rank-frequency function $g(r)$ as in (20)

$$g(r)=c_2\left(t-\left(\frac{r}{c_1}\right)^{\frac{1}{a_1}}\right)^{a_2} \tag{20}$$

where $r \in \left[0, c_1 t^{a_1}\right]$.

Proof: Let *t* be a fixed time period (e.g. the present). The further we look back in the past, the longer time sources have to grow and hence their rank densities *r* are described by

$$r = c_1 (\theta t)^{a_1} \tag{21}$$

for $\theta \in [0,1]$ (by(18)). By (19) we have for the item density on rank density *r*:

$$g(r) = g_1(\theta t) \tag{22}$$

From (21) we have

$$\theta t = \left(\frac{r}{c_1}\right)^{\frac{1}{a_1}} \tag{23}$$

(22) and (23) yield (20). □

Contrary to the classical Naranan case of exponential growth of sources and items, here $g(r)$ can have convex and non-convex shapes, e.g. including the S-shape described in Proposition 1 (case (B)). The size-frequency function $f(j)$ corresponding with (20) is described in Proposition 5.

Proposition 5 (Egghe (2012a)): Let φ and ψ be as in Proposition 4. Then this IPP has the size-frequency function $f(j)$ as in (24)

$$f(j)=\frac{D}{\left(t-\left(\frac{j}{c_2}\right)^{\frac{1}{a_2}}\right)^{1-a_1} j^{1-\frac{1}{a_2}}} \tag{24}$$

with

$$D=\frac{a_1 c_1}{a_2 c_2^{\frac{1}{a_1}}} \tag{25}$$

This is easily proved using formula (2)

Based on the possible shapes of $g(r)$ and Proposition 1 we have that $f(j)$ can have the shapes of Fig. 1.

In this framework of power law growth of sources and items, we can give a proof of the famous law of Herdan, also called Heaps' law (Heaps 1978, Herdan (1960,1964) – see also Egghe (2007))

Proposition 6 (Egghe(2012a)): Let φ and ψ be as in Proposition 4.

Denote, for all $t>0$, $\varphi(t)=T(t)=$the total number of sources at *t* and by $A(t)$ the total number of items at *t*. Then there exist constants $K>0$ and γ, $0<\gamma<1$ (independent of *t*) such that

$$T(t)=KA(t)^{\gamma} \tag{26}$$

Proof: By (18),

$$\phi(t) = I(t) - c_1^{a_1} \tag{27}$$

and by definition of $g(r)$ and $A(t)$

$$A(t) = \int_0^{g(t)} g(r)\,dr$$

$$A(t) = \int_0^{c_1} c_2 \left(t - \left(\frac{r}{c_1} \right)^{\frac{1}{a_1}} \right)^{a_2} dr \tag{28}$$

by (20) and (27). But, by (21), (28) equals

$$A(t) = \int_0^{c_1} c_2 \left(t - \phi(r) - c_1^{a_1} \right)^{a_2} dr$$

$$A(t) = c_2 t^{a_2} \int_0^{c_1} \left(1 - \frac{\phi(r)}{t} - \left(\frac{c_1}{t} \right)^{a_1} \right)^{a_2} dr \tag{28}$$

Denote

$$D = c_2 \int_0^{c_1} \left(1 - \frac{\phi(r)}{t} - \left(\frac{c_1}{t} \right)^{a_1} \right)^{a_2} dr$$

, a constant only depending on c_1, c_2, a_1 and a_2. Then (28) becomes

$$A(t) = D t^{a_1 + a_2} \tag{29}$$

Formula (27) implies

$$t = \left(\frac{T(t)}{c_1} \right)^{\frac{1}{a_1}} \tag{30}$$

Now (30) and (29) yield

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

or

$$T(t) = K A(t)^{\gamma}$$

with

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

and

$$\gamma = \frac{1}{1 + \frac{a_2}{a_1}} \tag{31}$$

It is clear from (31) that $0 < \gamma < 1$ (since $a_1, a_2 > 0$). □

Note that γ only depends on the ratio $\frac{a_2}{a_1}$ of the power law exponents in (18) and (19).

Naranan dynamics for general growth functions of sources and items.

Propositions 3 and 4 can be generalized as in Proposition 7.

Proposition 7 (Egghe (2012b)): Let us have an IPP in which

(i) The number of sources grow according to a differentiable injective function $\varphi(t)$

(ii) The number of items in each source grow according to a differentiable injective function $\psi(t)$ which is the same for every source.

Then this IPP has the rank-frequency function $g(r)$ as in (32)

$$g(r) = \psi(t - \varphi^{-1}(r)) \tag{32}$$

Proof: Let $t > 0$ be fixed and $\theta \in [0,1]$.

By definition of the generalized Naranan framework we have that sources that are born at θt have a period equal to $t - \theta$ to "grow" items. Hence, ranking sources in decreasing order of their number of items, we have, by definition of the rank-frequency function $g(r)$:

$$r = \varphi(\theta t) \tag{33}$$

and

$$g(r) = g(\varphi(\theta t)) = \psi(t - \theta t) \tag{34}$$

Formula (33) implies

$$\theta = \frac{1}{t}\varphi^{-1}(r) \tag{35}$$

So (35) in (34) yields (32). □

The size-frequency function is given by Proposition 8.

Proposition 8 (Egghe (2012b)): Let φ and ψ as in Proposition 7. Then this IPP has the size-frequency function $f(j)$ as in (36)

$$f(j) = \frac{c'(t - \psi^{-1}(j))}{\psi'(\psi^{-1}(j))} \tag{36}$$

This result follows easily from (32) and (2).

With (32) and (36), different growth models of IPPs can be studied. This is left to the reader.

References

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

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 (2010). A new short proof of Naranan's theorem, explaining Lotka's law and Zipf's law. Journal of the American Society for Information Science and Technology 61(12), 2581-2583.

L. Egghe (2012a). Study of the rank- and size-frequency functions in the case of power law growth of sources and items and proof of Heaps' law. Information Processing and Management, to appear.

L. Egghe (2012b). Study of the rank- and size-frequency functions and their relations in a generalized Naranan framework. Mathematical and Computer Modelling 55, 1898-1903.

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 47(2), 238-245.

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

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