Systems without low-productive sources

Non Peer-reviewed author version

EGGHE, Leo & ROUSSEAU, Ronald (2006) Systems without low-productive sources. In: INFORMATION PROCESSING & MANAGEMENT, 42(6). p. 1428-1441.

DOI: 10.1016/j.ipm.2006.03.009
Handle: http://hdl.handle.net/1942/1782

Systems without low-productive sources

Leo Egghe [1,2] and Ronald Rousseau [1,2,3]

[1] Universiteit Hasselt (UHasselt), Agoralaan, B-3590 Diepenbeek, Belgium,
[2] Universiteit Antwerpen (UA), IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
[3] KHBO (Association K.U.Leuven), IWT, Zeedijk 101, B-8400 Oostende, Belgium

E-mails: leo.egghe@uhasselt.be and ronald.rousseau@khbo.be

Abstract

Information Production Processes (IPPs) without low-productive sources are studied. A success-breeds-success or preferential attachment mechanism is established in which, from some point in time on, no new sources are created. Such systems are called mature systems. When time increases in mature systems the expected number of sources with a low number of items strictly decreases. An adaptation of the Naranan-Egghe model indicates that IPPs without low-productive sources must have small alpha exponents ($\alpha < 2$) in their size-frequency power law descriptions.

A positive reinforcement model explains all the essential properties. Using this approach it is shown that, when time increases in mature systems the alpha exponent of the power size-frequency function decreases, while, moreover, the minimum source size increases.

Examples related to country and city sizes illustrate the concepts and results discussed in this article.

Introduction

Classical informetrics deals with sources producing (or having) items. Such systems are referred to as Information Production Processes (IPPs) or as conglomerates (a slightly more general framework). Well-known cases are

- authors as sources and publications as items;
- journals as sources and published articles about a given subject as items;
- articles as sources and received citations (from a given pool of journals) as items;
- articles as sources and their references as items;
- web sites as sources and links (inlinks or outlinks) as items.

In all the above mentioned cases it is normal that the lowest number of sources considered is one (although sometimes also systems including sources without items are studied, citation studies being a case in point). In most cases f(1), this is the number of sources producing one item, is the largest among the numbers f(n), n = 1,2,3, …Recall that the function f is called the size-frequency function (Egghe, 2005a). It is indeed often the case that in a set of scientific authors the group consisting of authors with exactly one publication is the largest one. Generally, all IPPs that can be described by a decreasing power law have the property that f(1) > f(2) > f(3) > … For other IPPs, such as web sites and links, or articles in chemistry and the number of authors of these articles, f(1) is usually not zero, but it is not the largest value in the set {f(1), f(2), … } either. Systems that can be described by a lognormal, Weibull, negative binomial, or general Poisson distribution are examples of this case (Rousseau, 1994; Rao, 1995; Egghe & Rao, 2002).

IPPs with a power law as size-frequency continuous density function f(j) will be described as (Egghe, 2005a, p.128):

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

with $j\in[1,\rho_m]$, C > 0 and where ρ_m denotes the maximum item density in a source. Its equivalent description in rank-frequency continuous density form is (Egghe, 2005a, p. 128):

$$g(r)=\frac{E}{(1+Fr)^{\beta}}\ ,r\in[0,T] \tag{2}$$

The equation

$$\beta=\frac{1}{\alpha-1}\ or\ \alpha=1+\frac{1}{\beta} \tag{3}$$

expresses the relation between the two exponents involved in these equations.

IPPs where f(1) ≠ 0, and where the size-frequency function f increases first and then shows a long decreasing tail can be considered as describing an intermediate situation between IPPs with a decreasing size-frequency function, and those which are the focus of our investigations in this article. Consider texts with words (as types) and their occurrences (as tokens). Such IPPs are classically described by a decreasing power law as rank-frequency function (Zipf, 1949). Note that it is not required that the size-frequency function corresponding to this 'Zipf-function' has sources with production 1. Indeed, it often happens that most words (excluding misprints) occur several times and that f(1) = 0 (Tuldava, 1996). This is to be expected when a restricted vocabulary is used, such as for children's books, or computer programs. Baayen (2001, p. 51-52 and Fig.2.3) gives even examples where n_{min}, the minimum number of times that a word is used, deviates more and more from 1. An abstract example can be given by considering N random draws with replacement from an urn containing M balls. As this experiment can only have M possible outcomes, $f(n_{min})$, the lowest value among the number of times a ball has been drawn, is first zero, but will increase as N increases. A concrete example is an IPP describing the sizes of databases. Here n_{min} certainly is quite large. The same is true for human settlements (usually containing much more than one inhabitant) or countries and their population sizes.

IPPs where $n_{min} >> 1$, and hence f(n) = 0 for n = 1, …, n_{min} -1 will be referred to as IPPs without low-productive sources. Such IPPs are often characterized by the fact that f(n) – n a specified value - is either 0 or 1: there is only one city in a country with exactly n = 395,465 inhabitants and there is none with exactly n = 395,466 inhabitants.

The above mentioned examples of systems without low-productive sources will be interpreted in the next section within the "success-breeds-success" (preferential attachment) formalism. We will reveal the mechanics behind such processes and will show that f(n), for n fixed, is decreasing as a function of system size.

The third section deals with an interpretation of the Naranan-Egghe formalism (fractal theory of the growth of the number of sources and the number of corresponding items, leading to Lotkaian informetrics) in the framework of IPPs without low-productive sources. We show that this model leads to low values of the Lotka exponent (denoted as α).

In the fourth section we adopt the positive reinforcement model of IPPs (Egghe, 2005a) to the case of systems without low-productive sources. We generalize the

results obtained in (Egghe, 2004a, 2005a) by showing that the evolution of such systems can be described by

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

defined on [a, +∞ [, where the exponent α decreases and a increases over time. This yields a complete description of the growth of such systems and agrees with the result about α, as obtained in the previous section.

In the fifth section we present some results related to fitting real data. These data deal with country sizes, city sizes, and database sizes. They confirm the main features of our theoretical models. The paper closes with some open problems and conclusions.

The success-breeds-success formalism and sources without low-productive sources

General aspects of success-breeds-success

Success-breeds-success (SBS), also known as preferential attachment, is a formalism describing how sources and items grow in time. It originates with Herbert Simon (1955) and was introduced in the information sciences by Price (1976). For further details we refer to Egghe (2005a).

SBS is a discrete approach. At each tick a new item is created. At any moment, the total time passed is set equal to the total number of ticks and hence to the total number of items created. The formalism describes how the new item t+1 (hence at time t+1) is linked to a source. Creation of a new item leads to the following alternatives:

(i) Source creation alternative: with probability $c(t) \in [0,1]$ this new item is created (or attached to) a new source, i.e. a source that did not yet exist at time t; (ii) General SBS alternative: with probability 1- c(t) the new item is created (or attached to) an already existing source. Then, with probability $x(t,n) \in]0,1]$ this item is created by a source that had n items at time t ($n = n_{min}(t), .., n_{max}(t)$).

Formulated in this generality the formalism is more general than pure success-breeds-success. Indeed, in pure SBS the probability x(t,n) is required to increase with n. Note also that for a fixed situation at time t, there is a lower and an upper bound for the number of items that actually occur: $1 \leq n_{min}(t) < n_{max}(t) \leq t$. We next generalize a result shown in (Egghe & Rousseau, 1996), see also (Egghe, 2005a, Chapter I).

Theorem. For every t = 1, 2,… let f(t,n) denote the number of sources with n items at time t. Let $E_t(f(t+1,n))$ denote the conditional expectation of f(t+1,n) with respect to t, i.e. the average of all possibilities at time t+1, given the situation at time t. Then, for $n = n_{min}(t)+1, .., n_{max}(t)$

$$E_t(f(t+1,n)) = f(t,n) + (1-c(t))(x(t,n-1)-x(t,n)) \quad (4)$$

$$\text{for } n = n_{min}(t) = 1: \; E_t(f(t+1,1)) = f(t,1) + c(t) - x(t,1)\,(1-c(t)) \quad (5)$$

$$\text{and for } n = n_{min}(t) \neq 1: \; E_t(f(t+1, n_{min}(t))) = f(t, n_{min}(t)) - x(t, n_{min}(t))\,(1-c(t)) \quad (6)$$

Proof. Clearly, for $n = n_{min}(t)+1, .., n_{max}(t)$:

$E_t(f(t+1,n)) = c(t)f(t,n)$
$+ (1-c(t)).[x(t,n)(f(t,n)-1) + x(t,n-1)(f(t,n)+1)+(1-x(t,n)-x(t,n-1))f(t,n)],$

yielding equality (4). For $n = n_{min}(t) = 1$, we have:

$E_t(f(t+1,1)) = c(t).(f(t,1)+1)+(1-c(t)).[x(t,1)(f(t,1)-1)+(1-x(t,1))f(t,1)],$

yielding equation (5). Finally, for $n = n_{min}(t) \neq 1$

$E_t(f(t+1,n_{min}(t))) =$
$c(t).(f(t, n_{min}(t)))+(1-c(t)).[x(t, n_{min}(t))(f(t, n_{min}(t))-1)+(1-x(t, n_{min}(t)))f(t, n_{min}(t))],$

leading to equation (6).

We now present an interpretation of this result in the context of IPPs without low-productive sources.

SBS for systems without low-productive sources

It is not possible to derive analytical forms for f(t,n), even for concrete c(t) and x(t,n), due to the occurrence of the conditional expectation E_t. Nonetheless, the simple theorem shown above helps us interpreting the evolution of general IPPs and, especially, those without low-productive sources.

In the introductory section we briefly described an evolution starting with systems where the majority of sources had low productions (a typical case being scientific author productions), over systems where sources with one item exist, but they are not the majority (e.g. articles in chemistry and the number of authors of these articles), leading finally to systems without low-productive sources, i.e. f(n) = 0 for $n = 1, \ldots, n_{min} -1$ (e.g., cities and their inhabitants). The natural number n_{min} is the smallest one for which f(n) ≠ 0. We refrain from calling this evolution as one

going from ‘young’ to ‘old’ systems. While this may be true for systems without low-productive sources (any city has very humble origins), this is certainly not the case for the classical cases studied in informetrics, where there are always sources (authors, journals, articles) producing a small number of items (publications, articles, citations). The study of the differences between these systems, and the causes leading to these differences is left as an open problem.

When a system without low-productive sources reaches the end of its evolution c(t), the probability of the creation of a new source is almost zero. Yet, originally c(t) must have been very high (in the beginning many new sources must have been created). This is true for cities and countries: very few new countries or municipalities are created nowadays). So we conclude that systems without low-productive sources are characterized by

$$c(t) \approx 0 \tag{7}$$

for t large. Interpreting equation (7) as c(t) = 0 (such systems will be called stable systems) yields the following forms for equations (4), (5) and (6):

$$E_t(f(t+1,n)) = f(t,n) + x(t,n-1) - x(t,n) \tag{8}$$

for $n = n_{min}(t)+1, .., n_{max}(t)$,

$$\text{and } E_t(f(t+1,n)) = f(t,n) - x(t,n) \tag{9}$$

for $n = n_{min}(t)$.

We derive from equation (9) that, as $x(t, n_{min}(t)) > 0$ and whatever the exact value of $x(t, n_{min}(t))$:

$$E_t(f(t+1, n_{min}(t))) < f(t, n_{min}(t)) \tag{10}$$

If we suppose now, moreover, that we have pure SBS, i.e. $x(t,n) > x(t, n-1)$ for $n = n_{min}(t)+1, \ldots, n_{max}(t)$, then (9) yields that:

$$E_t(f(t+1,n)) < f(t,n) \tag{11}$$

By induction on t inequalities (10) and (11) also yield that, for $t' > t$:

$$E_t(f(t',n)) < f(t,n) \tag{12}$$

for $n = n_{min}(t), .., n_{max}(t)$, (where we have used that $E_{t-1} E_t = E_{t-1}$, by the definition of a conditional expectation). The meaning of inequality (12) is that for all $n = n_{min}(t), .., n_{max}(t)$, and $t' > t$, averaged over our knowledge at time t, we expect to have fewer and fewer sources with n items and this for all possible n, i.e. n between $n_{min}(t)$ and $n_{max}(t)$. This result is in agreement with (Baayen, 2001, p. 52).

Note, however, that equation (10), derived without the extra pure SBS assumption, already suggests that f(t, n_{min}(t)) decreases, and hence is expected to become zero. This leads – intuitively – to a window n_{max}(t) - n_{min}(t), moving to the right.

The use of the SBS formalism yields a first, partial, explanation for the occurrence of size-frequency functions f(n) (we have dropped the symbol t) having low values for small n (in the limit this might lead to f(n) = 0 for n = 1, …, n_{min} -1). We do not think that it is possible to derive more relevant facts from the SBS - conditional expectation - model. In the following sections we will derive more concrete results.

An argument explaining the frequent occurrence of alpha exponents between 1 and 2 for the size-frequency function of IPPs without low-productive sources, based on the Naranan-Egghe formalism

Assuming that IPPs without low-productive sources can be described by a size-frequency function of the form

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

we will show in this section that the exponent α is expected to have a low value, i.e. lies between 1 and 2.

The following theorem is essentially due to Naranan (1970), see also (Egghe, 2005a,c).

Naranan's Theorem. Consider an arbitrary IPP, and assume that
(i) the number of sources grows exponentially in time:

$$N(t) = c_1 a_1^t \tag{13}$$

where N(t) denotes the total number of sources at the time t, $c_1 > 0$, $a_1 > 1$;

(ii) the number of items in each source grows also exponentially in time, where the growth rate is the same for each source:

$$j(t) = c_2 a_2^t \tag{14}$$

where j(t) denotes the number of items in each source at time t, $c_2 > 0$, $a_2 > 1$.

Then the size-frequency function of this IPP has the form

$$f(j) = \frac{C}{j^{\alpha}} \tag{15}$$

where $j \geq c_2$ [this corrects a small mistake in Egghe (2005 a,c) where it is written that $j \geq 0$. The argument given in Egghe (2005 a,c) is, however, still valid] and where

$$\alpha = 1 + \frac{\ln(a_1)}{\ln(a_2)} \tag{16}$$

In Egghe (2005 a,c) Naranan's theorem has been extended and reinterpreted as follows:

Theorem (a fractal interpretation of Naranan's theorem). Under the conditions of Naranan's theorem this IPP is a self-similar fractal with fractal dimension D, given by

$$D = \alpha - 1 = \frac{\ln(a_1)}{\ln(a_2)} \tag{17}$$

These results will next be used in the framework of IPPs without low-productive sources. Assume that we have such an IPP, then $a_1 \approx 1$ (but larger than 1). Hence the size-frequency function of this IPP is

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

where $j \geq c_2$, and $\alpha \approx 1$ (but larger than 1). Since the Naranan model is rather general (it does not go into specific details) and does not aim at providing a fitting method for real situations, we just conclude from the previous observations that for IPPs without low-productive sources, the alpha-value is expected to be low. Classically (Lotka, 1926) α is close to two, hence alpha-values between 1 and 2, can be considered to be 'low'. Moreover, the value two divides the set of Bradford curves, i.e cumulative rank-frequency representations on a semi-log plot, into two distinct classes: for $\alpha \geq 2$ this curve is convex, for $\alpha < 2$ is shows a so-called Groos droop at the end, i.e. the Bradford curve is convex for small r-values, but then has an articulation point where the curve becomes concave (Rousseau, 1988).

IPPs without low-productive sources have low alpha-values implying that such systems have a low fractal dimension. Indeed, $D = \alpha - 1$ takes here a value between zero and one.

Note that $j \geq c_2$ includes that $0 = f(1) = f(2) = \ldots$ Yet, this is only part of what we hope to explain. Indeed, we think that n_{min} increases during the growth process of the IPP, while the number c_2 is a constant. Such a growing lower bound, combined with small exponents α will be obtained when using a positive reinforcement argument.

The fact that IPPs without low-productive sources usually have low alpha-exponents is confirmed by Rosen & Resnick (1980). They present a table of so-called Pareto exponents (our βs in equation (3)) for city sizes in 44 countries. The average Pareto exponent in their list is 1.136, with a minimum of 0.81 and a maximum of 1.96. These values correspond to an average α-value of 1.88 and a minimum and maximum value of 1.51 and 2.23.

A positive reinforcement approach to systems without low-productive sources

Definition. An IPP is said to be mature if growth in the number of sources has stopped. It may only grow through an increase in the number of produced items.

Suppose that a mature IPP without low-productive sources is given. Tracking its further growth we assume that the number of sources is T (constant) and the total number of produced items is A. Its rank-frequency function is denoted as g_A. Assume now that g_{A+1}, the rank-frequency function when a new item has been produced, is given as:

$$g_{A+1}(r) = B_A \left(g_A(r) \right)^{\gamma_A} \tag{19}$$

for $r \in [0, T]$, $B_A > 1$, $\gamma_A > 1$. Equation (19) expresses the fact that sources with the higher number of items grow relatively faster than the ones with a lower number of items. This is another way of expressing success-breeds-success (namely without conditional expectations). As

$$A + 1 = \int_0^T g_{A+1}(s)\, ds \tag{20}$$

the parameters in equation (19) are related through the equality:

$$A + 1 = B_A \int_0^T \left(g_A(s) \right)^{\gamma_A} ds \tag{21}$$

We take $\gamma_A > 1$ such that

$$A < \int_0^T \left(g_A(s) \right)^{\gamma_A} ds < A + 1 \tag{22}$$

Clearly, as $A=\int_0^T g_A(s)\,ds$, γ_A must be taken larger than one. Because of the right-hand inequality in (22) also B_A must be larger than 1. Yet, equation (22) leaves some freedom in the choice of the parameters. Note also that parameters B_A and γ_A are assumed to depend on A. These assumptions lead to the following theorem, explaining the size-frequency structure of IPPs without low-productive sources.

Theorem

Assume that we have a mature IPP growing from A to A+1 items such that the relation between successive rank-frequency functions is given as:

$$g_{A+1}(r)=B_A\,(g_A(r))^{\gamma_A} \tag{19}$$

for $r\in[0,T]$, $B_A > 1$, $\gamma_A > 1$. If the size-frequency function f_A is given as:

$$f_A(j_A)=\frac{C_A}{j_A^{\alpha_A}} \tag{23}$$

with $j_A \geq B_A \geq 1$, $\alpha_A > 1$, then

$$f_{A+1}(j_{A+1})=\frac{\dfrac{C_A B_A^{\frac{\alpha_A-1}{\gamma_A}}}{\gamma_A}}{\left(j_{A+1}\right)^{\frac{\alpha_A+\gamma_A-1}{\gamma_A}}}=\frac{C_{A+1}}{j_{A+1}^{\alpha_{A+1}}} \tag{24}$$

with,

$$C_{A+1}=\frac{C_A B_A^{\frac{\alpha_A-1}{\gamma_A}}}{\gamma_A}\ and\ \ \alpha_{A+1}=\frac{\alpha_A+\gamma_A-1}{\gamma_A} \tag{25}$$

Consequently: $\alpha_{A+1}<\alpha_A$ and $j_{A+1}=B_A j_A^{\gamma_A}$. Finally, $B_{A+1}=B_A.B_A^{\gamma_A}>B_A$, showing that the lower bound for the argument of f_A is strictly increasing in A.

Proof. The proof is an extension of a result shown in Egghe (2004a, 2005a). Using the relations between rank and size-frequency functions (Egghe, 2005a) we have for all $r\in[0,T]$, T constant:

$$r=g_{A+1}^{-1}(j_{A+1})=\int_{j_{A+1}}^{\rho_{m,A+1}} f_{A+1}(k)dk \tag{26}$$

and $$r = g_A^{-1}(j_A) = \int_{j_A}^{\rho_{m,A}} f_A(k)dk \tag{27}$$

where $\rho_{m,A}$ and $\rho_{m,A+1}$ denote the maximum item densities in a source. Hence, using the relation $j_{A+1} = g_{A+1}(r)$ (see equation (26)) and $j_A = g_A(r)$ (equation (27)), we have, by (19):

$$j_{A+1} = B_A\, j_A^{\gamma_A} \tag{28}$$

Inserting equation (28) in (26) and (27) yields:

$$\int_{B_A j_A^{\gamma}}^{\rho_{m,A+1}} f_{A+1}(k)dk = \int_{j_A}^{\rho_{m,A}} f_A(k)dk \tag{29}$$

for all $j_A \geq B_A$. Taking in equation (29) the derivative with respect to j_A yields:

$$-f_{A+1}(B_A\, j_A^{\gamma_A})\, B_A \gamma_A\, j_A^{\gamma_A - 1} = -f_A(j_A) \tag{30}$$

for all $j_A \geq B_A$. Using equations (28) and (23) leads to:

$$f_{A+1}(j_{A+1}) = \frac{f_A(j_A)}{B_A \gamma_A\, j_A^{\gamma_A - 1}} = \frac{C_A}{B_A \gamma_A\, j_A^{\alpha_A + \gamma_A - 1}} \tag{31}$$

Again using equation (28) yields:

$$f_{A+1}(j_{A+1}) = \frac{\frac{C_A B_A^{\frac{\alpha_A - 1}{\gamma_A}}}{\gamma_A}}{\left(j_{A+1}\right)^{\frac{\alpha_A + \gamma_A - 1}{\gamma_A}}}$$

which proves equations (24) and (25). As j_{A+1} is defined for values larger than or equal to B_{A+1} en j_A is defined for values larger than or equal to B_A, equation (28) shows that

$$B_{A+1} = B_A\, B_A^{\gamma_A} > B_A \geq 1 \tag{32}$$

Equation (32) shows that the lower bound of the argument of f_A is strictly increasing in A. Finally, we show that $\alpha_{A+1} < \alpha_A$.

Indeed, as $\alpha_{A+1} = \frac{\alpha_A + \gamma_A - 1}{\gamma_A}$ we have to show that

$$\frac{\alpha_A+\gamma_A-1}{\gamma_A}<\alpha_A \quad or$$
$$\alpha_A+\gamma_A-1<\alpha_A.\gamma_A \quad or$$
$$0<(\alpha_A-1)(\gamma_A-1).$$

This last inequality is true, as α_A and γ_A are assumed to be strictly larger than 1. This proves the theorem.

Remarks

We took $B_A > 1$ in the model discussed above. If we take $B_A = 1$ as a limiting case, then equation (20) becomes:

$$\int_0^T \left(g_A(s)\right)^{\gamma_A} ds = A+1 \tag{33}$$

while (19) becomes: $g_{A+1}(r)=\left(g_A(r)\right)^{\gamma_A}$. Since $\gamma_A > 1$ this situation is an example of a positively reinforced IPP. This means that there exists a strictly increasing function φ such that φ(1) = 1, φ(x) ≥ x and

$$g_{A+1}(x)=\left(\varphi \circ g_A\right)(x) \tag{34}$$

for all $r\in[0,T]$. Such positively reinforced IPPs were studied in (Egghe 2004a, 2005a). There the relation between successive Lotka-exponents (equation (25)) was already proved. The difference is that in the earlier approach j_A always (i.e. for all A) starts in the point 1. The extension given in this article to an increasing lower bound for values for the variable j_A is new. It explains the occurrence of IPPs without low-productive sources, but, clearly, the relation between successive Lotka-exponents is not influenced.

Decreasing Lotka-exponents as occurring in the previous theorem lead to the following consequence.

Corollary. If L(h) denotes the Lorenz curve of a function h, then:

$$L(f_A) > L(f_{A+1})$$

$$L(g_A) < L(g_{A+1})$$

These inequalities show that the size-frequency inequality decreases, while the rank-frequency inequality increases.

Proof. We refer the reader to (Egghe 2005a,b) for the definition of the Lorenz curve of a continuous function. In these references (see e.g. Egghe, 2005a, p. 204-205) it is shown that the Lorenz curve of the size-frequency function increases strictly in α (the Lotka-exponent) while the Lorenz curve of the rank-frequency function decreases in α. This proves this corollary.

Practical examples and discussion

The determination of Lotka's α for systems without low-productive items is not straightforward. In informetric studies Lotka's size-frequency function usually begins with many sources producing one item. Indeed, the mode of Lotka's distribution is always situated at the point 1. This is not the case here. Hence, a program such as LOTKA (Rousseau & Rousseau, 2000) cannot be used.

Consequently, we will determine Lotka's α from the corresponding β in Mandelbrot's function. Recall that Mandelbrot's rank-frequency function, denoted as g(r), and the relation between the exponents α and β are given as (Egghe, 2005a, p. 128):

$$g(r) = \frac{E}{(1+Fr)^{\beta}}\ , r \in [0,T] \tag{2}$$

$$and\ \beta = \frac{1}{\alpha - 1} \tag{3}$$

When we determine α based on formulae (2) and (3) we refer to this approach as method M. As an alternative, Lotka's α can also be determined from the cumulative rank-frequency form, denoted as G(r) (Egghe, 2005a, p. 128; Rousseau, 1988):

$$G(r) = \frac{C}{2-\alpha}\left(\rho_m^{2-\alpha} - \left(\rho_m^{1-\alpha} - \frac{1-\alpha}{C}r\right)^{\frac{2-\alpha}{1-\alpha}}\right),\ r \in [0,T] \tag{35}$$

where ρ_m denotes the highest density of items in a source. Determining α from equation (35) will be called method R.

In the examples that follow we will apply these two methods where curve fitting is performed using the Marquardt algorithm for nonlinear regression as implemented in Statgraphics ©.

Note that when it comes to fitting there is no difference between 'natural' IPPs without low-productive sources and other IPPs for which the size-frequency function has been left-truncated (or equivalently: the rank-frequency function is

right-truncated). For the truncated IPPs there is, however, no special reason to expect low values of the exponent α.

Test

In order to check how well the proposed algorithms work we first test them on a truncated version of Bradford's *Applied Geophysics* data. More precisely, we only consider the first 45 sources (beginning with the most productive one), restricted moreover to one source with a specific production. More precisely, there are, e.g., four sources with production 16, occupying ranks 10, 11, 12 and 13. We only kept the one at rank 13.

The alpha-value for the complete data obtained using method R is 2.07 [R^2 = 0.9974] (Rousseau, 1994b), while using method M it is 1.81 [R^2 = 0.9746]. The truncated data yield 2.13 [R^2 = 0.9981], using method R, and 1.75 using method M [R^2 = 0.9699]. Differences between the two methods seem larger than between complete and truncated versions. Close inspection of the data represented on a semi-log scale shows a small 'droop' at the end. This indicates that the 'real' alpha-value is somewhere near 1.9. Hence, in this test, method M seems to give the better result. We postpone final judgment as to which method is the better of the two till the end of this section.

Example 1. Country sizes

Countries and their number of inhabitants are good examples of IPPs without low-productive sources. There clearly are no countries with 1 or two inhabitants. Using data from the website: www.gazetteer.de on the number of inhabitants of 237 countries (visited on July 10, 2005) we obtained an α-value of 1.69 with method R [R^2 = 0.9876], and an α-value of 1.64 with method M [R^2 = 0.9369]. Fig. 1 illustrates the data and the best-fitting curve according method R. These α-values are, as predicted, 'small', i.e. smaller than two.

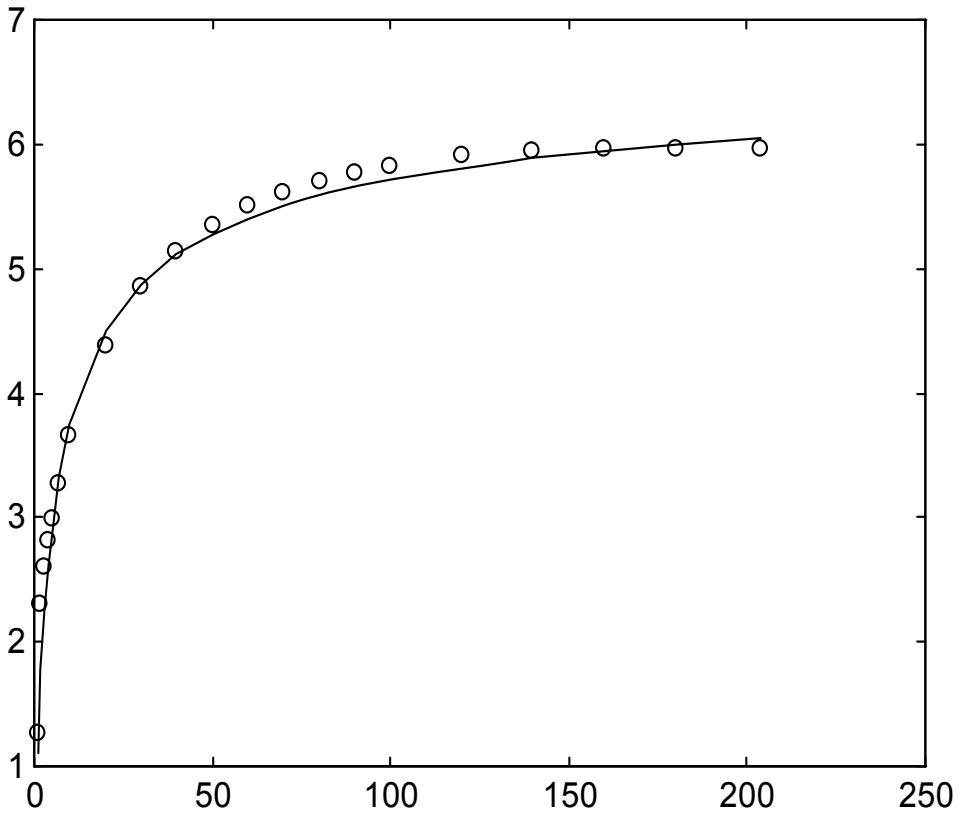

Fig. 1 Countries of the world and best-fitting cumulative rank-frequency function. Ranks on the abscissa, cumulative population (in 10^9 persons) on the ordinate axis. Only selected data are shown.

Example 2. Municipalities in Belgium (July 1, 2004 data)

The number of inhabitants of each municipality in Belgium can be found at: www.statbel.fgov.be/figures/download_nl.asp (visited on June 5, 2005)

We obtained an α-value of 2.31 with method R [R^2 = 0.9978], and an α-value of 2.54 with method M [R^2 = 0.9851]. Fig. 2 illustrates the data and the best-fitting curve. These exponent values are not 'small', suggesting that these data are of a different nature. This is, indeed, the case. In 1831 Belgium consisted of 2739 cities and villages. On January 1977 (1983 for the city of Antwerp) the existing cities and villages of Belgium were merged, such that only 596 municipalities remained. Merging has been done in such a way that municipalities are approximately of the same size. More equality in size means that due to this human intervention, the Lorenz curve of the rank-frequency function g, denoted as L(g), approached the diagonal line of total equality. As explained in (Egghe, 2005a,b) such an operation implies an increase in the corresponding α-exponent

for the size-frequency function f. Hence it is not surprising to find a best-fitting α-value which is somewhat larger than expected.

Fig.2

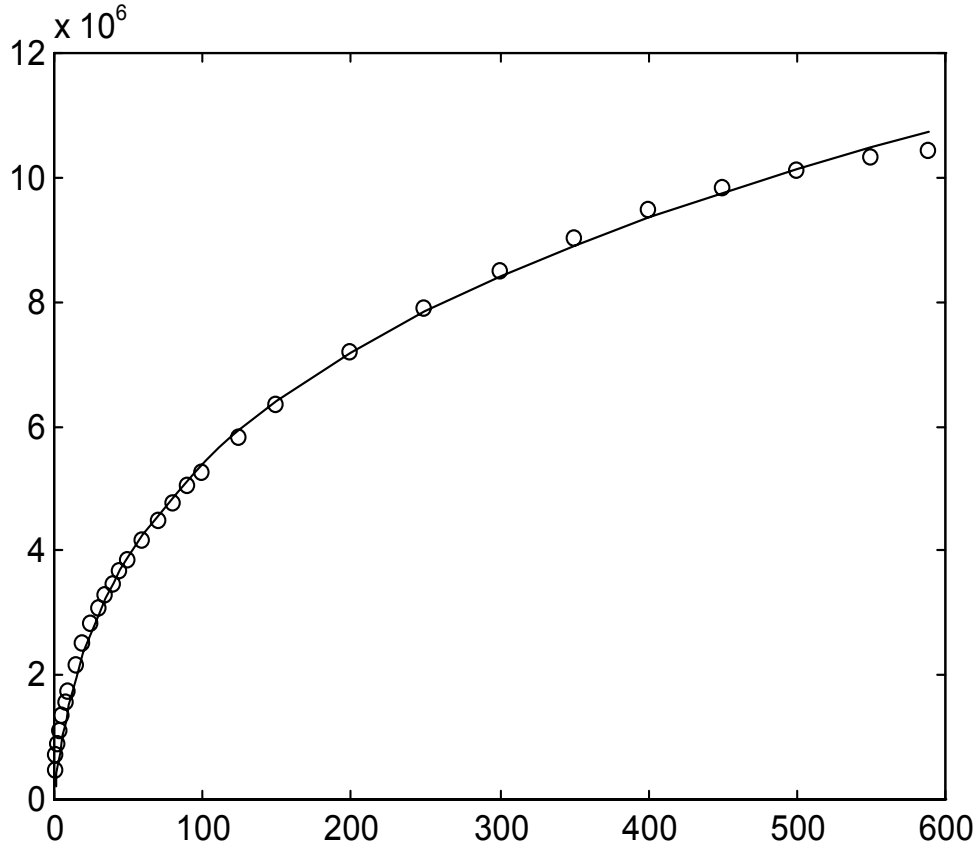

Fig. 2 Belgian population and best-fitting cumulative rank-frequency function. Ranks on the abscissa, cumulative population on the ordinate axis. Only selected data are shown.

Example 3. Local councils in Malta (1997 data)

Malta is only a small country where a merging operation such as in Belgium did not take place. As we had access (from the local Yellow Pages) to the complete data for the 68 local councils we checked if these data were more according to our predictions. (Actually data are only provided for 67 local councils, as the population of Mtarfa is included in that of Rabat (Malta)).

For the Malta data we found an α-value of 1.12 with method R [$R^2 = 0.9995$], and an α-value of 1.31 with method M [$R^2 = 0.9785$]. Fig. 3 illustrates the data and the best-fitting curve. These exponent values are 'small', corresponding to the predictions of our theory.

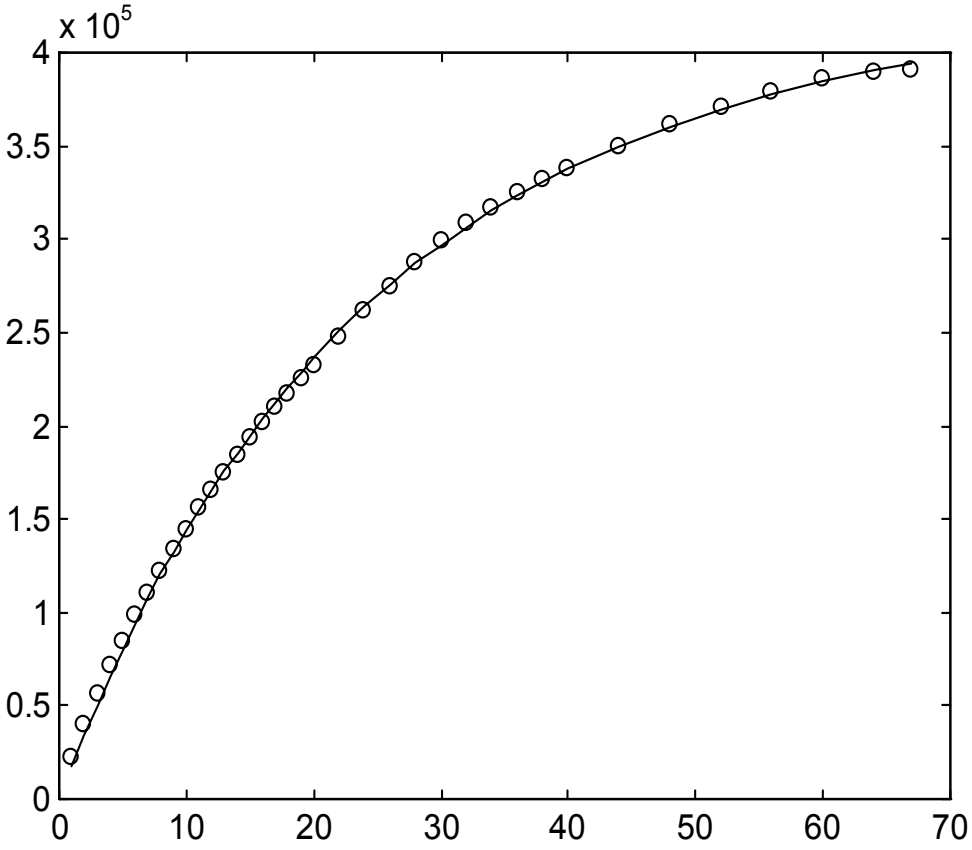

Figure 3. Population of Maltese local councils and best-fitting cumulative rank-frequency function. Ranks on the abscissa, cumulative population on the ordinate axis. Only selected data are shown.

Example 4. Database sizes

Our next example deals with database size. Although any database begins in theory with one source and one item, in reality when considering a group of databases, there never are databases (sources) with a small number of items. The specific data we will use, however, are of the left-truncated type. Indeed, we use data presented in (Hood & Wilson, 2003) on database sizes, limited to the topic "fuzzy set theory". Since we only have data on the twenty largest databases, this example is indeed a left-truncated case. So, whether or not the complete data set follows a power law, we only try to model its tail behavior.

Table 1. Database sizes (number of documents present) on the topic "fuzzy set theory". Data taken from (Hood & Wilson, 2003).

Rank	Production
1	5,815
2	5,149
3	4,980
4	2,902
5	1,708
6	980
7	677
8	652
9	556
10	373
11	365
12	308
13	265
14	261
15	244
16	212
17	154
18	132
19	102
20	76

According to method M $\alpha = 1.001$ [$R^2 = 0.9524$], method R yields $\alpha = 1.09$ [$R^2 = 0.9919$]. The result using method M does not lead to a good fit (visually), its R^2-value is smaller than that for the R-method, and the resulting α basically yields a divergent curve. For this reason we conclude that $\alpha = 1.09$ is the best value. Although this value is small there was no a priori reason to expect a small α-value, as the data are truncated.

Example 5. Unique documents in databases

Hood and Wilson (2003) present another interesting table, namely the number of unique documents in each of the above twenty databases. This table is ranked in decreasing order, and hence ranking is not necessary the same as for table 2.

Table 2. Unique documents (taken from (Hood & Wilson, 2003))

Rank	Unique documents
1	2,456
2	1,650
3	1,618
4	681
5	431

6	336
7	316
8	266
9	234
10	151
11	138
12	126
13	120
14	99
15	86
16	82
17	62
18	57
19	56
20	56

According to method M $\alpha = 1.14$ [$R^2 = 0.9671$], method R yields $\alpha = 1.33$ [$R^2 = 0.9959$]. Although this value is small also here there was no a priori reason to expect a small α-value, as the data are truncated.

Conclusions and open problems

In this article we studied IPPs without low-productive sources. Although such systems must, by necessity, have evolved from systems in which some sources have low-productivity, we cannot say that such systems are necessarily old (this is certainly not the case for texts or web links), nor can we say that systems where most sources produce just one item are necessarily young (any bibliography of authors or journals).

We established a success-breeds-success mechanism in which, from some point in time on, no new sources are created. Such systems are called mature systems. When time increases further on in mature systems the expected number of sources with $n \in$ □ items strictly decreases, if such a source already existed. An adaptation of the Naranan-Egghe model indicates that IPPs without low-productive sources must have small alpha exponents ($\alpha < 2$). Yet this model does not provide an explanation why the minimum number of items in a source is increasing in time.

An extension of the positive reinforcement model explains all the essential properties. Using this approach we have shown that, when time increases in mature systems the alpha exponent of the Lotka size-frequency function decreases, while, moreover, the minimum source size increases. Another argument in favor of lower alpha-values is presented in the appendix.

Examples related to country sizes, city sizes and databases illustrate the concepts and results discussed in this article.

Recall that our positive reinforcement model assumes that the size-frequency function of an IPP can be modeled through a decreasing power density function, corresponding to a rank-frequency power (Zipf-Pareto) law (Gabaix, 1999). Other models may and do exist. William Reed (2001, 2002) for instance derives a double-Pareto-lognormal model for the size-frequency distribution of human settlement. Consequently, our derivations only apply to his tail distribution.

An intriguing open problem is to find an explanation why certain IPPs always stay within the confines of the classical framework where the majority of sources have only one item [a necessary condition seems to be that there should be a virtually infinite possibility for creating new sources, as in the author case], why some IPPs move to a situation where $f(1) \neq 0$, but the mode of f is reached for a value $n > 1$ (e.g. articles and their references, web sites and outlinks), and, finally, why some IPPs may reach a state where $f(n) = 0$ for $n = 1, \ldots, n_{min} - 1$, as for texts, databases sizes, countries or municipalities. This problem is related to that of the emergence of cities (Batty, 2003), the writing process of texts (Baayen, 2001) and simulations of the generation of bibliographies, e.g. through some form of success-breeds-success mechanism (Wilkinson, 1972; Brookes, 1988).

This article is only exploratory in nature. We do not claim to have found final or complete answers to the questions studied here. We, nevertheless hope to have shed some light on the intriguing phenomena of IPPs without low-productive sources, and on the evolution of IPPs. We hope that many colleagues will be interested in continuing this type of study, including fitting new data sets and explaining the results of this fitting exercise.

References

R.H. Baayen (2001). *Word frequency distributions*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

M. Batty (2003). The emergence of cities: complexity and urban dynamics. Preprint.

B.C. Brookes (1988). Comments on the scope of bibliometrics. In: *Informetrics 87/88* (Egghe & Rousseau, eds.). Amsterdam: Elsevier, p. 29-41.

L. Egghe (2004a). Positive reinforcement and 3-dimensional informetrics. *Scientometrics*, 60, 497-509. Correction: *Scientometrics*, 61, (2004) 283.

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

L. Egghe (2005a). *Power laws in the information production process: Lotkaian informetrics*. Elsevier, Oxford (UK).

L. Egghe (2005b). Relations between the continuous and the discrete Lotka power function. *Journal of the American Society for Information Science and Technology*, 56, 664-668.

L. Egghe and I.K.R. Rao (2002). Duality revised: construction of fractional frequency distributions based on two dual Lotka laws. *Journal of the American Society for Information Science and Technology*, 53, 789-801.

L. Egghe and R. Rousseau (1996). Stochastic processes determined by a general success-breeds-success principle. *Mathematical and Computer Modelling*, 23, 93-104.

X. Gabaix (1999). Zipf's law and the growth of cities. *American Economic Review*, 89, 129-132.

W. Hood and C.S. Wilson (2003). Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology*, 54, 1091-1103.

A. J. Lotka (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16, 317-324.

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

D. de Solla Price (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27, 292-306.

I.K.R. Rao (1995). A stochastic approach to analysis of distributions of papers in mathematics: Lotka's law revisited. In: *Proceedings of the fifth conference of the international society for scientometrics and informetrics* (A. Bookstein & M. Koenig, eds.), p. 455-464. Learned Information, Medford (NJ).

W.J. Reed (2001). The Pareto, Zipf and other power laws. *Economics Letters*, 74, 15-19.

W.J. Reed (2002). On the rank-size distribution for human settlements. *Journal of Regional Science*, 42, 1-17.

K.T. Rosen and M. Resnick (1980). The size distribution of cities: an examination of the Pareto law and primacy. *Journal of Urban Economics*, 8, 165-186.

B. Rousseau and R. Rousseau (2000). LOTKA: a program to fit a power law distribution to observed frequency data. *Cybermetrics*, 4(1), paper 4.

R. Rousseau (1988). Lotka's law and its Leimkuhler representation. *Library Science with a Slant to Documentation and Information Studies*, 25(3), 150-178.

R. Rousseau (1994a). The number of authors per article in library and information science can often be described by a simple probability distribution. *Journal of Documentation,* 50, 134-141.

R. Rousseau (1994b). Bradford curves. *Information Processing and Management*, 30, 267-277.

H.A. Simon (1955). On a class of skew distribution functions. *Biometrika*, 42, 425-440.

J. Tuldava (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics*, 3, 38-50.

E. Wilkinson (1972). The Bradford-Zipf distribution, a simulation study. OSTI Report 5172 (London).

Appendix 1

In this appendix we show that in the framework of so-called Lotkaian informetrics, decreasing alpha-values are quite natural when growth in the number of sources stagnates. This argument is, however, restricted to alpha > 2, while in general we expect lower α-values. For this reason we only mention this argument in an appendix.

Lotkaian informetrics and stagnation in the growth of sources

Assume that we work in the framework of Lotkaian informetrics, i.e. the size-frequency function of the system is given by a power law:

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

Then Egghe (2004b,2005a) has shown the following.

If the total number of items, denoted as A, is strictly larger than the total number of sources, denoted as T, and if ρ_m the maximum item per source density is equal to infinity, then the following relation is true:

$$\alpha = \frac{2A - T}{A - T} \tag{36}$$

Recall though that the theory leading to equation (36) implies that $\alpha > 2$.

If now the number of sources stagnates, say, becomes a constant, while the number of items still increases, then, clearly α decreases (by equation (36)). This finding implies that for IPPs without low-productive sources, very high alpha-values are not probable at all.