

An interpretation of the effort function through the mathematical formalism of Exponential Informetric Process

Thierry Lafouge, Agnieszka Smolczewska

► **To cite this version:**

Thierry Lafouge, Agnieszka Smolczewska. An interpretation of the effort function through the mathematical formalism of Exponential Informetric Process. Information Processing and Management, Elsevier, 2006, 42. sic_00112695

HAL Id: sic_00112695

https://archivesic.ccsd.cnrs.fr/sic_00112695

Submitted on 9 Nov 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An interpretation of the effort function through the mathematical formalism of Exponential Informetric Process

Thierry Lafouge

Laboratoire Ursidoc Université Claude Bernard Lyon 1

43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex, France

Lafouge@univ-lyon1.fr

Agnieszka Smolczewka

Laboratoire Ursidoc Université Claude Bernard Lyon 1

43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex, France

Agnieszka.Smolczewska@univ-lyon1.fr

Abstract

Statistical distributions in the production or utilization of information are most often studied in the framework of Lotkaian informetrics. In this article, we show that an Information Production Process (IPP), traditionally characterized by Lotkaian distributions, can be fruitfully studied using the effort function, a concept introduced in an earlier article to define an Exponential Informetric Process. We thus propose replacing the concept of Lotkaian distribution by the logarithmic effort function. In particular, we show that an effort function defines an Exponential Informetric process if its asymptotic behavior is equivalent to the logarithmic function $\beta \cdot \text{Log}(x)$ with $\beta > 1$, which is the effort function of a Lotkaian distribution.

1. Introduction

Statistical regularities observed in the production or use of information have been studied for a long time in informetric processes. Today, they are again very topical, as is testified by the many articles. They are characterized by phenomena of invariance of scale during research into the traffic on Internet (Aby & al., 2004), (Barabasi & al., 2000). They are also observed when the topology of the Web is studied (Bilke & al., 2001) or when counting the frequencies of the number of pages or the number of degrees entering or leaving the Web pages in a collection of sites (Prime Claverie & al., 2005). Their most current mathematical formulation is that of an inverse power function, usually called a Lotkaian informetric

distribution. A recent book from Egghe (Egghe, 2005) proposes a mathematical approach to the framework of Lotkaian informetrics, illustrated by several examples.

In this article, we continue a study begun previously (Lafouge & Prime Claverie, 2005) where we defined the Exponential Informetric Process by introducing the concept of effort function. We are studying here informetric processes while drawing on traditional mathematical formulation in continuous mode. Mathematically, we represent the effort function by the logarithmic function, which is related to the effort function that appears in the law of Lotka (Lotka, 1926).

2. Information Production Process and effort function

Statistical distributions in the production or utilization of information, such as the law of Lotka (Lotka, 1926) - production of articles by researchers in a scientific community - generally fit into simple unidimensional models. These models can be represented by the diagram of Figure 1, introduced into informetric systems by Leo Egghe (Egghe, 1990) and called "Information Production Process" (IPP). An IPP is a triplet made up of a bibliographical source, a production function, and all the elements (items) produced.

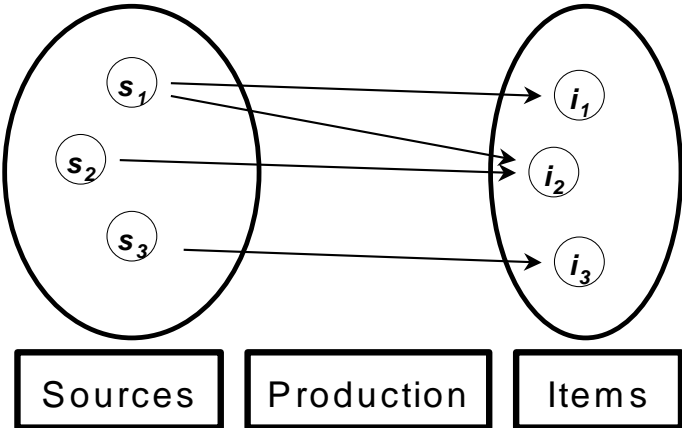


Figure 1: Schematic representation of an Information Production Process

In (Lafouge & Prime Claverie, 2005) we assume that an item produced requires a certain amount of effort and therefore we define the informetric process by introducing the effort

function (see Figure 2). We use the size frequency form, and denote as F the frequency distribution where $F(i)$ represents the number of sources that have produced i items $i=1,2,\dots,i_{\max}$ (maximum number of items produced). The effort function $f(i)$ denotes the amount of effort from a source needed to produce i items $i=1,2,\dots,i_{\max}$. The amount of effort, denoted as AF , produced by an IPP is:

$$AF = \sum_{i=1}^{i_{\max}} f(i) \cdot F(i)$$

If f is the identity function $f(i) = i$, the amount of effort produced by the process is simply equal to the number of items produced. Since the production and the effort (function) appear to be logically connected, we use both to define the Exponential Informetric Process.

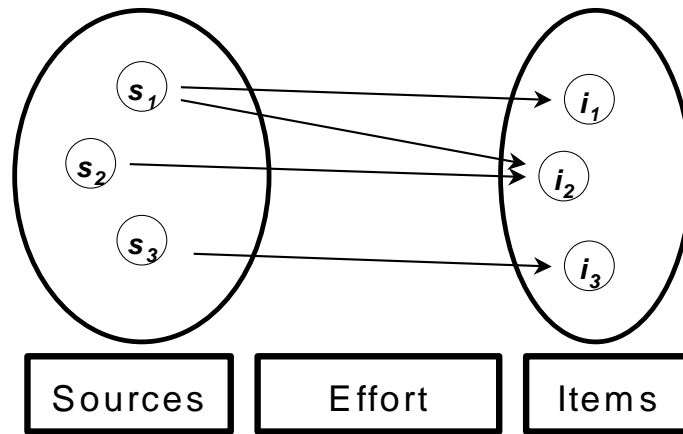


Figure 2: Schematic representation of an informetric process using the effort function

3. Exponential Informetric Process

In the article (Lafouge & Prime Claverie, 2005) we define an Exponential Informetric Process in terms of an exponential density and an effort function where the average quantity supplied by the sources to produce all the items is finite. More precisely, we define a set of functions denoted EF :

$$EF = \{ f: [1..\infty[\rightarrow \mathfrak{R}^+, \exists A \geq 1, f \text{ increasing over } [A..\infty[\text{ and not majorized } \}$$

We call effort function an element of EF . Let a be a number greater than 1, $f \in EF$, we call Exponential Informetric Process the following density function $v(f, a)$:

$$v(f, a)(x) = k \cdot a^{-f(x)} \quad (k \text{ constant of normalization}) [1],$$

where

$$\overline{F} = \int_1^{\infty} \nu(f, a)(x) \cdot f(x) dx \text{ is finite [2].}$$

\overline{F} corresponds to the average of effort produced by the density process $\nu(f, a)$. The results (b) and (c) that follow, explain the relationship between average of effort and entropy.

3.1. Entropy and effort

The Maximum Entropy Principle (MEP) maximizes the entropy subject to the constraint that the effort remains constant, whereas the Principle of Least Effort (PLE) minimizes the effort subject to the constraint that the entropy remains constant (Egghe, 2005).

Assuming f an effort function and a a number greater than 1, we show in (Lafouge & Prime Claverie, 2005) that these conditions imply the following properties:

- (a) $\nu(f, a)$ is a function with decreasing density over the interval $[A.. \infty[$.
- (b) The two principles, maximum entropy and least effort are verified simultaneously.
- (c) If \overline{H} and \overline{F} describe the average information content and effort produced by the process, we have the following proportional relationship: $\overline{H} = -\text{Log}(k) + \text{Log}(a) \cdot \overline{F}$.

Note: Here we will use $a = e$, where e is the Euler number, and $\text{Log} = \text{Log}_e$. All the results are valid for a logarithmic function in any base.

Also, in the following, for $f \in EF$ we denote as $\nu(f) = e^{-f}$ the associated density function (we suppose $k=1$). $\nu(f)$ is an Exponential Informetric Process if condition [2] is verified,

that is if the average effort, $\int_1^{\infty} f(x) \cdot e^{-f(x)}$ is finite.

Note 1

More generally, we can easily show that if f defines an Exponential Informetric Process and that g is a limited positive function, then $h = f + g$ also defines an Exponential Informetric Process, provided that h is an increasing function on the interval $[A.. \infty[$, where $A \geq 1$. This note enables us to envisage building a multitude of effort functions starting from an Exponential Informetric Process (see *Example c*) below).

Note 2

It is important to note that there exist effort functions for which $\nu(f)$ is a density function but the condition [2] is not satisfied.

The reader will see that, for the following effort function:

$$x \in [1 \dots \infty[, f(x) = \text{Log}(x+1) + \text{Log}(\text{Log}^2(x+1))$$

we have

$$\int_1^{\infty} \nu(f)(x) dx = \frac{1}{\text{Log}(2)} \text{ and that } \int_1^{\infty} f(x) \cdot \nu(f)(x) dx \text{ is infinite.}$$

Proposition 1

Assume f and g two effort functions, so that g is greater or equal than f on the interval $[B \dots \infty[$ where $B \geq 1$. If $\nu(f)$ is an Exponential Informetric Process the same holds for $\nu(g)$.

Proof

Given that f and g are increasing functions not majorized such that $\forall x \geq B, f(x) \leq g(x)$, there is a positive number A greater than 1 such that $\forall x > A, 1 < f(x) \leq g(x)$. Knowing that the function $x \cdot e^{-x}$ is decreasing on the interval $[1 \dots \infty[$ we can write:

$$\forall x > A, \int_1^{\infty} e^{-g(x)} dx \leq \int_1^{\infty} e^{-f(x)} \cdot f(x) dx$$

We can therefore conclude that if $\nu(f)$ is an Exponential Informetric Process the same holds for $\nu(g)$.

□

Considering *Note 2* and the result of *Proposition 1*, it would be interesting to define an Exponential Informetric Process with a different condition from condition [2]. This condition and related results obtained by means of *Proposition 1*, will be presented in the following.

4. Characterization of an Exponential Informetric Process

As mentioned before, we seek to define an adequate condition that will allow us to say if any effort function defines an Exponential Informetric Process. To do this, we will compare the asymptotic behavior of an effort function with the effort functions of a Lotkaian distribution.

4.1. Lotkaian distribution

We saw in the introduction that the most current formulation for modeling many phenomena related to the use or production of information is that of Lotka, given here in continuous form:

$$v(x) = \frac{k}{x^{\alpha+1}} \quad \alpha > 0 \quad x \in [1.. \infty[\quad [3]$$

The properties of these hyperbolic distributions or inverse power laws have been widely studied (Egghe, 2005). It seems essential to us to be able to compare these functions with an exponential density, defined by an effort function.

We easily go from [3] to [1] taking as effort function $f(x) = (\alpha + 1) \cdot \text{Log}(x)$, where $a = e, k = \alpha$. The effort function of an inverse power function is a function proportional to the logarithmic function. This characteristic clarifies the principle of the least effort. This result is implicit in the article of Yablonsky (Yablonsky, 1981), where he shows the relationship between the principle of the maximum entropy and an inverse power function. Also, it is known that such a distribution (Lafouge & Michel, 2001) has an entropy equal to $-\text{Log}(\alpha) + \frac{1}{\alpha} + 1$ and thus a quantity of effort equal to $\frac{1}{\alpha} + 1$. An inverse power function thus defines an Exponential Informetric Process (noted $v(\alpha)$) if α is strictly positive. For α less than or equal to 0, $v(\alpha)$ is not an Exponential Informetric Process and neither is a density function.

4.2. Effort function and Exponential Informetric Process

4.2.1. Characterization by logarithmic effort function

In the following, we will characterize the Exponential Informetric Process by the effort function of a Lotkaian distribution.

Theorem 1

Let g be an effort function so that $\text{Limit}_{x \rightarrow \infty} \frac{g(x)}{\text{Log}(x)}$ exists¹, and verifies the condition

$$\text{Limit}_{x \rightarrow \infty} \frac{g(x)}{\text{Log}(x)} = C \quad [4],$$

where C is a finite number greater than one or infinite. Then g defines an Exponential Informetric Process.

Proof

a) C is finite

In order to define an Exponential Informetric Process it suffices to show that condition [2] is verified. Since $C > 1$, then it exists a finite number α such that $C > \alpha > 1$. Let $f = \alpha \cdot \text{Log}(x)$ and $\beta = \frac{C}{\alpha}$, then $\beta > 1$.

According to the preceding, we know that f defines an Exponential Informetric Process.

Condition [4] implies:

$$\forall \varepsilon > 0, \exists A > 0 \text{ so that } \forall x > A, \left| \frac{\text{Log}(x)}{g(x)} - \frac{1}{C} \right| < \varepsilon$$

$$\forall \varepsilon > 0, \exists A > 0 \text{ so that } \forall x > A, f(x) < \left(\varepsilon + \frac{1}{\beta} \right) g(x).$$

Choosing $\varepsilon = \frac{\beta - 1}{\beta}$, we then have $\forall x > A, f(x) \leq g(x)$. Since f is an increasing function not majorized we can choose A so that $\forall x > A, 1 < f(x) \leq g(x)$.

$\nu(f)$ being an Exponential Informetric Process, the preceding proposition allows us to say that g defines an Exponential Informetric Process.

¹ There are functions for which this limit does not exist.

b) C is infinite

The relationship [4] allows us to write:

$$\forall B > 0, \exists A > 0 \text{ so that } \forall x > A, \frac{g(x)}{\text{Log}(x)} > B.$$

Proof follows, as for the finite case.

□

In other words, an effort function defines an informetric process if its asymptotic behavior is equivalent to the logarithmic function $\beta \cdot \text{Log}(x)$ with $\beta > 1$.

Examples

a) *Exponential case*

This case corresponds to a linear effort function:

$$f(x) = \beta \cdot (x-1), \quad \beta > 0, \quad x \in [1.. \infty[$$

where:

$$\text{Limit}_{x \rightarrow \infty} \frac{f(x)}{\text{Log}(x)} = +\infty$$

We find the well-known case of an exponential distribution : $v(f) = e^{-\beta \cdot (x-1)}$

b) *Mixed case*

This case corresponds to an effort function that is the sum of a linear type function and a logarithmic type function. The effort function is:

$$x \in [1.. \infty[, f(x) = \alpha \cdot (x-1) - (j-1) \cdot \text{Log}(x-1), \quad \alpha > 0, \quad j = 1, ..m, \quad m \in \mathbb{N}$$

In this case $\text{Limit}_{x \rightarrow \infty} \frac{f(x)}{\text{Log}(x)} = +\infty$.

The exponential process corresponding to the density function is :

$$v(x) = \beta^j \cdot \frac{(x-1)^{j-1}}{(j-1)!} \cdot e^{-\beta(x-1)},$$

which corresponds to the distribution of Erlang. If j is real and non-integer, we recognize the gamma distribution.

c) General inverse power law

The following example is an inverse power law modified for low frequencies. This case corresponds to an effort function of type:

$$x \in [1.. \infty[, f(x) = \frac{\mu}{x} + \beta \cdot \text{Log}(x), \quad \beta > 1$$

This function is increasing for $x > \frac{\mu}{\beta}$.

In this case, we have $\text{Limit}_{x \rightarrow \infty} \frac{f(x)}{\text{Log}(x)} = \beta$. We can then conclude.

The associated density function is:

$$v(f) = C \cdot e^{-\frac{\mu}{x}} \cdot x^{-\beta},$$

where C is a standard constant. This distribution, used for example in (Nararan, 1992), allows us to adjust statistical distributions on the vocabulary of various language corpora. In general, calculations give us β close to 2 and μ less than 1. This distribution is one of many possible generalizations of the law of Zipf (Shan, 2005). According to the sign of the constant μ , the effort function will have a different effect for low frequency values, which are known to have an important influence. In this case, the adjustment will be done via a multi-linear regression after transforming the data on a logarithmic scale.

Note 1 allows us to consider many other functions of this type.

d) Other examples

Among many other possible examples we have, for instance, the normal Log law with an effort function of type:

$$f(x) = \frac{(\text{Log}(x-1) - m)^2}{\sigma}$$

Every effort function of type:

$$f(x) = kx^n, k > 0$$

also defines an Exponential Informetric Process.

We can show that for these functions, the quantity of effort is finite and is equal to:

$$\int_1^{\infty} x^n \cdot e^{-x^n} dx = \frac{1}{n} \cdot \Gamma\left(\frac{1}{n}, 1\right)$$

We can also quote the distribution of Weibul here. This distribution, like the normal Log distribution, can be increasing then decreasing. This is why we require an effort function to increase only on the interval $[A \dots \infty[$ where $A \geq 1$ and not necessarily on the interval $[1 \dots \infty[$.

□

Regrettably, the preceding theorem is not a necessary and sufficient condition. In fact, when C is equal to 1, the two following cases are possible:

- if $g(x) = \text{Log}(x)$ we know that $\nu(g)$ is not an informetric process

- if $g(x) = \text{Log}(x) + \sqrt{\text{Log}(x)}$ the reader can show that $\int_1^{\infty} \nu(g)(x) \cdot g(x) dx < \infty$

Nevertheless, we can provide a necessary condition, presented in *Theorem 2*.

Theorem 2

Assume g is an effort function so that the $\text{Limit}_{x \rightarrow \infty} \frac{g(x)}{\text{Log}x}$ exists and verifies:

$$\text{Limit}_{x \rightarrow \infty} \frac{g(x)}{\text{Log}(x)} = C \quad [5]$$

where C is a positive number less than 1, then g is not recognized as an Exponential Informetric Process since its average of produced effort is infinite.

Proof

Since $1 > C$, then it exists a finite number α verifying $1 > \alpha > C$. Let $f = \alpha \cdot \text{Log}(x)$ and

$\beta = \frac{C}{\alpha}$, then $\beta < 1$.

Since $\alpha < 1$, we know that f is not an Exponential Informetric Process.

Condition [5] implies:

$$\forall \varepsilon > 0, \exists A > 0 \text{ so that } \forall x > A, \frac{g(x)}{f(x)} < \beta + \varepsilon. \text{ Therefore } \varepsilon = 1 - \beta.$$

Since g is an increasing function not majorized we can choose A so that:

$$\forall x > A, 1 \leq g(x) \leq f(x).$$

Since the quantity of effort $\int_1^{\infty} e^{-f(x)} \cdot f(x) dx$ is infinite, we can conclude using the same argument as in *Proposition 1*.

□

4.2.2. Characterization by generic effort function

The two previous theorems can be generalized replacing the logarithmic function by an effort function having a finite (*Theorem 3*) or infinite (*Theorem 4*) quantity of effort.

Theorem 3

Assume g is an effort function and $v(f)$ an Exponential Informetric Process so that

Limit $\lim_{x \rightarrow \infty} \frac{g(x)}{f(x)}$ exists and verifies the condition:

$$\lim_{x \rightarrow \infty} \frac{g(x)}{f(x)} = C,$$

with $C > 1$, then g is also an Exponential Informetric Process.

Proof

The same as in the *Theorem 1*, using the argument:

$$\int_1^{\infty} e^{-f(x)} \cdot f(x) dx < \infty.$$

□

Like *Theorem 3*, which generalizes the result of *Theorem 1*, the following theorem generalizes the result of *Theorem 2*.

Theorem 4

Assuming g is an effort function and f an effort function with an infinite quantity of effort and verifying the condition:

$$\text{Limit}_{x \rightarrow \infty} \frac{g(x)}{f(x)} \text{ exists and verifies } \text{Limit}_{x \rightarrow \infty} \frac{g(x)}{f(x)} = C \text{ with } C < 1,$$

then g is not an Exponential Informetric Process, since its quantity of effort is infinite.

The proof is the same as in *Theorem 3*, using the argument that the quantity of effort

$$\int_1^{\infty} e^{-f(x)} \cdot f(x) dx \text{ is, in this case, infinite.}$$

4.3. Moment and Exponential Informetric Process

It is a known fact that a Lotkaian distribution has only moments of order n for n strictly less than β (where $\beta = \alpha + 1$ is the exponent of the inverse power law). An Exponential Informetric Process may have, or not have, moments of order n . We will give a sufficient condition so that an Exponential Informetric Process has a moment of order n . In order to do this, we will compare the asymptotic behavior of an effort function with an Exponential Informetric Process, whose effort function is the power function $f(x) = x^n$, where n is an integer greater than 1.

Theorem 5

Let n be an integer greater than 1 and g an effort function so that $\text{Limit}_{x \rightarrow \infty} \frac{g(x)}{x^n}$ exists and verifies the condition:

$$\text{Limit}_{x \rightarrow \infty} \frac{g(x)}{x^n} = C$$

where C is a finite number greater than one or infinite, then g defines an Exponential Informetric Process characterized by moments of order n .

Proof

According to *Theorem 1*, g is an Exponential Informetric Process. Let us show that it has moments of order n :

a) C is finite

$$\forall \varepsilon > 0, \exists A > 0 \text{ so that } \forall x > A, \text{ we have } \frac{x^n}{g(x)} < \frac{1}{C} + \varepsilon,$$

therefore for $\varepsilon = 1 - \frac{1}{C}$, we have $\forall x > A, x^n < g(x)$,

$$\text{therefore } \int_A^{\infty} x^n \cdot e^{-g(x)} < \int_A^{\infty} g(x) \cdot e^{-g(x)}.$$

We can then conclude because g is an Exponential Informetric Process.

b) C is infinite

The proof is identical.

□

5. Conclusion

An IPP (Information Production Process) is defined by sources, a production function and the elements produced. Given that each element produced requires a certain quantity of effort, it is justified to approach an IPP as much by its effort function as by its distribution. In an Exponential Informetric Process, both are closely linked.

In informetrics, the traditional cases of IPP are generally characterized by Lotkaian distributions. Given that it is justified to approach an IPP by the effort function, we here propose replacing the concept of Lotkaian distribution by the logarithmic effort function. The results presented in this article, in particular *Theorem 1* and *Example c*), show that if a distribution of effort is equivalent to the logarithmic function $\beta \cdot \text{Log}(x)$ with $\beta > 1$, we have an Exponential Informetric Process.

Acknowledgements

The authors wish to thank A. Agouzal, University of Lyon 1, for his advice in mathematics, and L. Egghe for his relevant remarks on Theorem 1.

References

Aby, P., Flandrin, P., Hohn, N., & Veitch, D. (2004). Invariance of scale in the Internet, published in "Mesures de l'Internet" under the direction of Eric Guichard, Les Canadiens en Europe, 96-111.

Barabasi, A.L., Albert, R. & Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide-web. *Physica A*, 311, 590-614.

Bilke, S. & Peterson, C. (2001). Topological properties of citation and metabolic networks. *Physical Reviews E*, 6403(3), 76-80.

Egghe, L. (1990). On the duality of informetric systems with application to the empirical law. *Journal of Information Science*, 16, 17-27.

Egghe, L. (2005). Power laws in the information production process: Lotkaian Informetrics. Elsevier.

Lafouge, T., & Prime Claverie, C. (2005). Production and use of information. Characterization of informetric distributions using effort function and density function. *Information Processing and Management*, Vol 41, 1387-1394.

Lafouge, T., & Michel, C. (2001). Links between information construction and information gain. Entropy and distributions. In *Journal of Information Science*, 27(1), pp 39-49.

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

Naranan, S. (1992). Statistical laws in information science, language and system of natural numbers: some striking similarities. *Journal of scientific and industrial Research* 51(8-9), 736-755.

Prime Claverie, C., Beigbeder, M., & Lafouge, T. (2005). Limits and feasibility of cositation method on the web an experiment on the French speaking Web; In *Proceedings of the 10th international conference of scientometrics and informetrics*, p78-86, July 2005, University of Stockholm, Sweden

Shan, S. (2005). On the generalized Zipf distribution. Part 1. *Information Processing and Management*, Vol 41, 1369-1386.

Yablonsky, A.L. (1981). On fundamental regularities of the distribution of scientific productivity. *Scientometrics* 2(1), 3-34.