A remarkable example in three-dimensional informetrics. The geometric law: Distribution of use or distribution of structure?

Abdelatif Agouzal, Thierry Lafouge

A remarkable Example in Three-dimensional Informetrics. The Geometric Law: Distribution of Use or Distribution of Structure?

Abdellatif, AGOUZAL;
Thierry LAFOUGE

1. Introduction and Background

To describe an informetric process, we use a general framework that consists in studying a population of sources which randomly produces items over time. Many examples have been studied in the past decades. We mention two historic examples: the production of articles by researchers (Lotka, 1926), and the contributions of research articles in a particular field in scientific journals (Bradford, 1934). A framework, referred to as IPP (Information Production Process) was defined (Egghe, 1990) to study such informetric processes in which the population is itself a set of sources[1] (researchers, articles, web pages…) producing items (articles, citations, links…). These historic examples do not explicitly consider time dependence in their mathematical formulation of the processes. They are stationary distributions in the form of counts (number of articles, number of citations…) observed over a predefined time period. Probabilistic laws are fitted to the empirical distributions. Their properties have been known for a long time in the field of scientometrics (Haitun, 1982). Almost all of the distributions are aggregative (above-average variance), decreasing, and are generally long-tailed. Certain researchers have implemented a stochastic model to take the time dependence explicitly into account and explain such phenomena more clearly. We can, for example, cite the works of Burrell on bibliometric processes (Burrell, 1988). It is also important to mention the works of Price in which the time unit (Price, 1976) is taken into account. The cumulative advantage model (a variant being Success-Breeds-Success (SBS)) describes the evolution of an IPP in time. In (Egghe and Rousseau, 1995), a model generalising Price's model (general SBS) is defined.

To the best of our knowledge, there are no historic laws – such as Lotka's or Bradford's – that deal with the use of documents. However, the statistical regularities in the distributions of book loans in libraries were observed early on. Such studies were undertaken to improve library management. Predictive tools, which originated from operational research (Morse, 1968) were set up. In his book, Morse chose the Poisson process when implementing Markovian matrices to regulate the circulation of books in libraries. The negative binomial (Bagust, 1983) law was chosen to model stationary distributions of library book loans (Leemans et al., 1992). Stochastic models, which depend on several parameters and which take explicit consideration of time dependence, were used (Burrell, 1990) (Burrell and Fenton, 1993).

In the digital era, borrowing the physical copy of a document is not as crucial as it used to be. Uses are now more of interest. However, statistical regularities linked to uses are still relevant, and mathematical models are still effective. Recently, studies (Ajiferuke and Famoye, 2016) used very different data sets that represented four broad informetric subfields where different counting models were tested.

These authors created statistics based on variables that are used in altmetrics (Priem et al., 2012) such as: statistics for the number of views, statistics for the number of readers.

Distributions of use related to citations are still largely studied and constitute invaluable indicators – notably citations in the evaluations of research. A multitude of types of distributions (Exponential, Weibul, Log-normal, Yule distribution…) have been tested to adjust these distributions. A state of the art of these various studies can be found in (Bertoli-Barsotti and Tommaso, 2015). This study leads the authors to suggest a new formula to calculate the h-index, based on the geometric law, to model the distributions of citations. Burrell (2014) noticed that the distribution of citations of three renowned researchers in informetrics followed a geometric law. The body of work is admittedly small, but the result is nevertheless surprising.

[1] It would be more precise to talk about generalized bibliographical sources.

The "No Use" sources, which are non-producing sources, are not *a priori* excluded from the framework of an IPP, but they are not studied as a stand-alone element. Taking into account these "non-producers" often modifies the type of model used for the adjustment. For example, in the study of loans (Burrel, 1980), this factor was taken into account. Egghe and Rousseau (2012) described a Lotka distribution which includes non-producers (the so-called shifted Lotka function). We must note that, in practice, the definition of an IPP generally implies that the entire set of sources is defined with the help of the produced items. Taking into account the affiliated time factor and "No Use" sources is, of course, crucial in the analysis of citations and in many other cases. In longitudinal studies, over time, the set of sources varies (some entering, some leaving). To the best of our knowledge, there are few theoretical and/or applied studies that take into account the time factor.

For a long time, some studies noticed that two IPPs could be linked in a natural way. These IPPs are called Three-dimensional Information Production Processes (Egghe, 2005, chapter 3). Rousseau (1992) considered the case in which researchers who published articles then also received citations. Burrell (1992) defined a stochastic model to study this type of problem: a population of researchers who publish articles over time and who subsequently receive citations. Burrell suggested a model in which he counted the number of published papers of a publishing author during [0, t], with a time-dependent geometric distribution.
Looking at data from the journal provider, we studied the demand for scientific articles (Lafouge, 1998) by researchers.
This experimentation led us to define a three-dimensional IPP with the following linear framework: journals produce scientific articles by including them in volumes, and the scientific articles are then requested by researchers.
We then conducted several theoretical studies (Lafouge and Lainé Cruzel, 1997) (Lafouge and Guinet, 1999) (Lafouge, 2001). The results from these articles are re-explained in the present paper using a more general method. To the best of your knowledge, few theoretical or practical studies on three-dimensional informetrics exist.

The aim of this article is to revisit the three results mentioned above. We had shown, under certain conditions, that if the distribution of article production in journals was of a certain type (Poisson, geometric…), then the distribution of use was of the same type. Here, we are building a generic model that is in line with the three dimensional IPP. It uses the properties of the probability generating function. We introduce time dependence in a non-explicit way by having the proportion of sources that no longer produce items (No-use) – those that are no longer used – tending towards 1. Such a reality illustrates what is known as information obsolescence.

Our article is organized in three parts:
- Defining the problem: we describe the informational process in which our study operates (see Section 2).
- Defining the general theory (see Section 3).
This section is divided in two subsections. In the first part (see Subsection 3.1), we study a stationary problem in which the geometric law is highlighted (see Theorem 3.6). In a second section (see Subsection 3.2), we indirectly introduce the time dependence (see Theorem 3.7).
- Discussion and conclusion (see Section 4).

2. Statement of the Problem

Preliminary Thoughts

The statistical regularities of the Information Production Processes are one of the most fascinating aspects of informetrics. This article focuses, from a theoretical point of view, on the regularities of distributions of use in a body of scientific articles over a predetermined period. This leads us to define what we call a distribution of "structure". To the best of our knowledge, such a concept has not yet been used in informetrics, or, at least, not under that name. The distribution of structure quantifies the number of articles published in the journals' volumes. Various hypotheses have been introduced to explain regularities in distributions of use. We do not, strictly speaking, seek to create a new explanatory model. The distribution of use and the distribution of structure are linked. We therefore make the hypothesis that, under certain conditions, these regularities can be seen as a consequence of other regularities. The distribution of use is necessarily geometric because the distribution of structure is geometric. The question we ask now is: why is the distribution of structure geometric?

The proportion of articles that are never used increases over time. This obsolescence concept is well-known, and it is linked to the exponential growth of the amount of information: an article ceases to be cited after a certain period following its publication date. This does not necessarily mean that the oldest articles loose in scientific value, but that the more recent articles receive a surplus of citations or orders. An important result showed that exponential growth and Lotkaian informetrics are linked (Egghe, 2004). In Section 3, we use the geometric law which is the

discrete version of the exponential law. We indirectly introduce time dependence in order to take the concept of obsolescence into account.

In this section, we define the process of use of a body of scientific articles by highlighting the link between the distribution of use and the distribution of structure.

Implementing the Process

Before defining the informational process in which we operate, we shall first rapidly recall the three dimensional IPPs (Egghe, 2005, chapter 3). If the mathematical formulation of our process (see equation [1], Section 3) is different than the one used by Egghe, our model is in line with three-dimensional informetrics.

a) Reminder

The most well-known informetrics theory is two-dimensional informetrics. A framework, referred to as IPP (Information Production Process) is defined. An IPP is a triplet (S, f, I) in which S is all of the sources, and I is all of the items produced by these sources. The f function is the size frequency function[2]: for every $n \in \mathbb{N}, n = 1,2, \ldots \rho_S$, $f(n)$ is the number of sources with n items and ρ_S is the maximum number of items that a source can produce. Our theoretical model (see Theorem 3.6, Subsection 3.1) favours the case in which f is geometric.

In three-dimensional informetrics, we focus here on the case in which there are two source sets and one item set. Two patterns are possible:

- Linear pattern

Let there be two IPPs (S_1, f, I_1), and (S_2, g, I_2). In the linear pattern, the first IPP's item set I_1 is the second IPP's source set, $I_1 = S_2$ Rousseau (1992) considered the case in which researchers published articles in which these articles received citations. This linear pattern is illustrated in Figure 1.

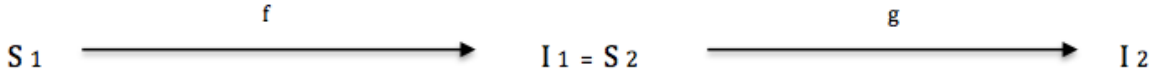

Figure 1: Linear three-dimensional informetrics

- Triangular pattern

Let there be, for example, two IPPs (S_1, f, I) and (S_2, g, I), in which S_1 is a group of researchers producing articles and in which S_2 is a group of journals publishing articles. In this example we consider the researchers and journals as producers of articles, f and g are independent. This situation is illustrated in Figure 2.

[2] Other types of mathematical formulas exist to describe the production of items by the sources.

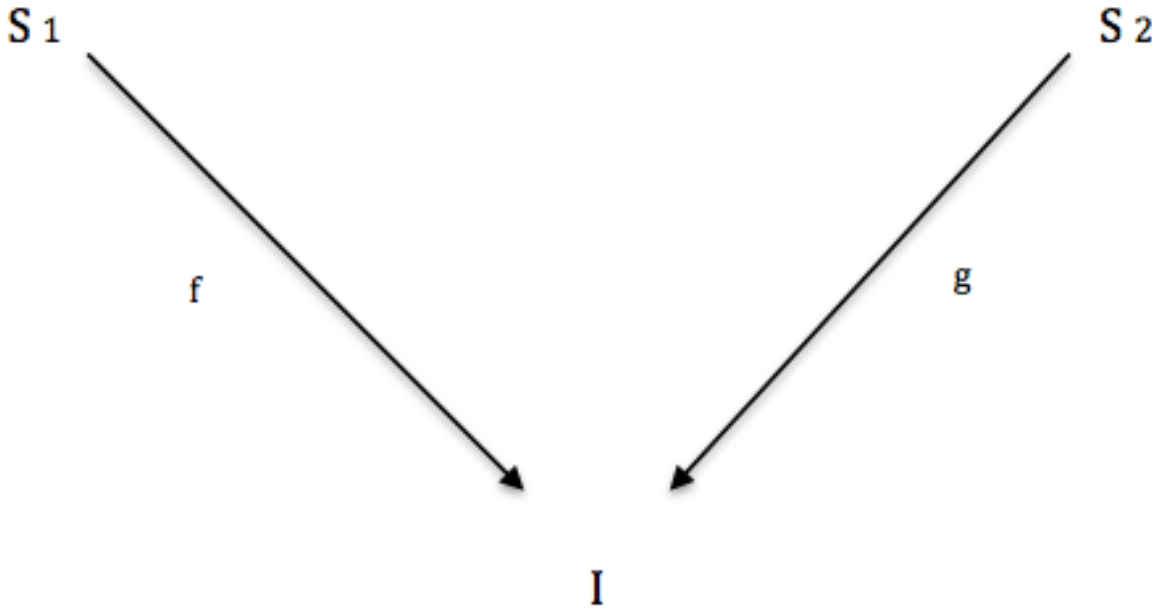

Figure 2: Triangular three-dimensional informetrics

We must comment that the choice of a pattern is arbitrary and that an informational process can be represented by different patterns.

b) Link between use and structure

Let there be a body of scientific journals (known as J). Let there be the articles published in these journals (known as A). During a given period of time, we observe the use (known as I) of the body of journals. When we say use we take into account downloads when consulting electronic journals (Boukacem-Zeghmouri et al., 2016), citations of articles, or "viewed" and "read" tags on platforms that use altmetrics. Two patterns are possible:

- Linear pattern

Journals produce scientific articles by including them in volumes, and the scientific articles are then requested by researchers. In this case, the first IPP's (J, f, A) number of sources is necessarily inferior to the second IPP's (A, g, I) number of sources. This situation is illustrated in Figure 3.

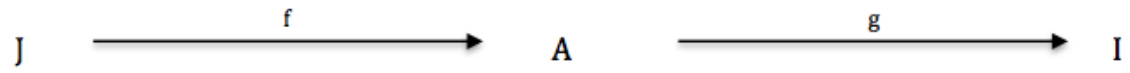

Figure 3: Linear view of the process

$f(n)$ is the number of journals that produced n articles. f is the distribution's size frequency function which we shall from now on call a distribution of structure.

- Triangular pattern

- Each time an article is requested by a researcher, a journal volume is necessarily sought. We write m as the maximum number of uses of a volume .
Figure 4 sheds light on the distribution of structure's role :
Where
- E designates the state space: $E = [1,2..m] \supset [1,2..\rho_A]$ (ρ_A is the maximum number of uses of an article).
- U is a function from A into E quantifying the number of times an article is used over a given period.

- V is a function from J into E quantifying the number of times a journal is used over a given period.
- F is the relation of $(J \times A)$ which expresses the fact that all article belongs to a volume and that each volume contains at least one article.

The use of a volume is determined by the cumulation of the use of the articles that belong to this volume. Therefore, the most natural way to define V would be as follows:

$$j \in J\,, \qquad V(j) = \sum_{a} U(a)\,, \qquad where\ a = F(j)$$

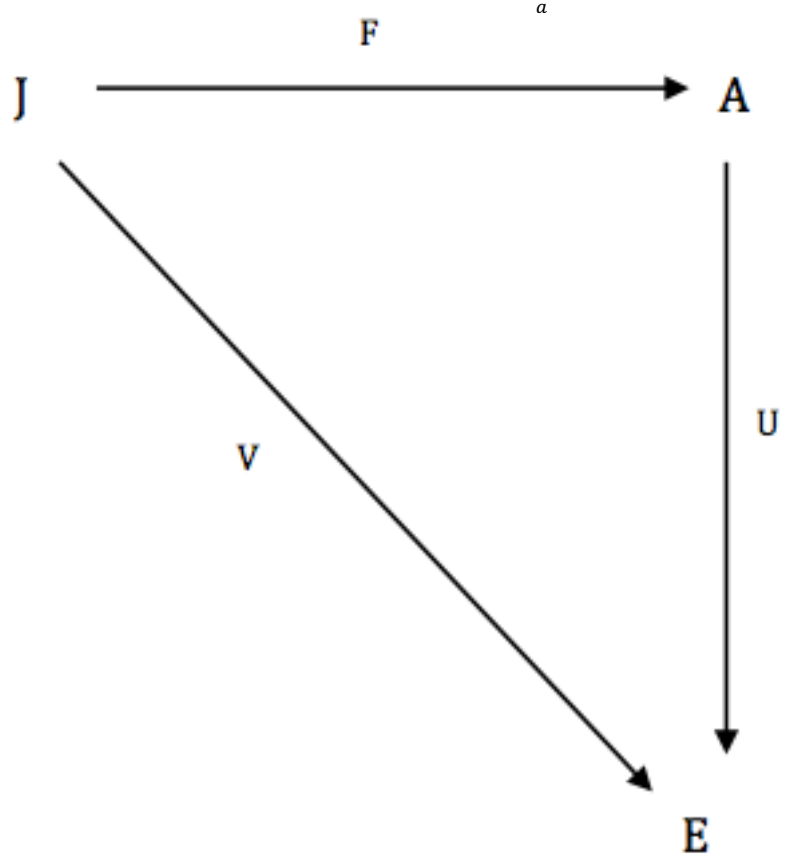

Figure 4: Triangular view of the process

This symbolic formula links the distribution of use (J, V, E) to the distribution of use (A, U, E). The F relation allows us to count the number of articles in each volume. Each article belongs to only one scientific journal volume, for which, in turn, we know the overall number of articles.

When we talk about the "number of articles in the volume" in this study, several meanings are implied:
- The most common meaning (Lafouge 1995) is the number of articles published in a volume: traditionally, this number does not vary much from one volume to another, when the articles come from the same journal. However, mega-journals (such as Plos One) are new types of journals that publish incrementally and that are modifying the world of publishing.

Other meanings are possible:
- Number of research articles in the volume,
- Number of articles in a research institution,
- Number of articles stemming from a research theme,
- Number of leading articles in the volume. The term "leading articles" can cover different meanings: articles from high-ranking authors, articles on sensitive or controversial news topics.

This last count is different in nature compared to the others.

All of these examples lead us to postulate the existence of another distribution, which we shall call "distribution of structure".

Objective

In reality, we can observe the IPP's distribution of use $J \to E$. The same does not always apply for the distribution of use $A \to E$.
We have postulated the existence of the IPP's distribution of structure (J,A). We have seen that this distribution of structure can be interpreted in different ways. Our objective is to build a theoretical model that links these three distributions. The triangular pattern (see Figure 4, Figure 5) highlights the link between the two sets of sources and the set of items. The letters T, S, X will be used in the mathematical formula (See equation [1]) in the following section and will designate the random variables that correspond to the distribution of structure and to the distributions of use.

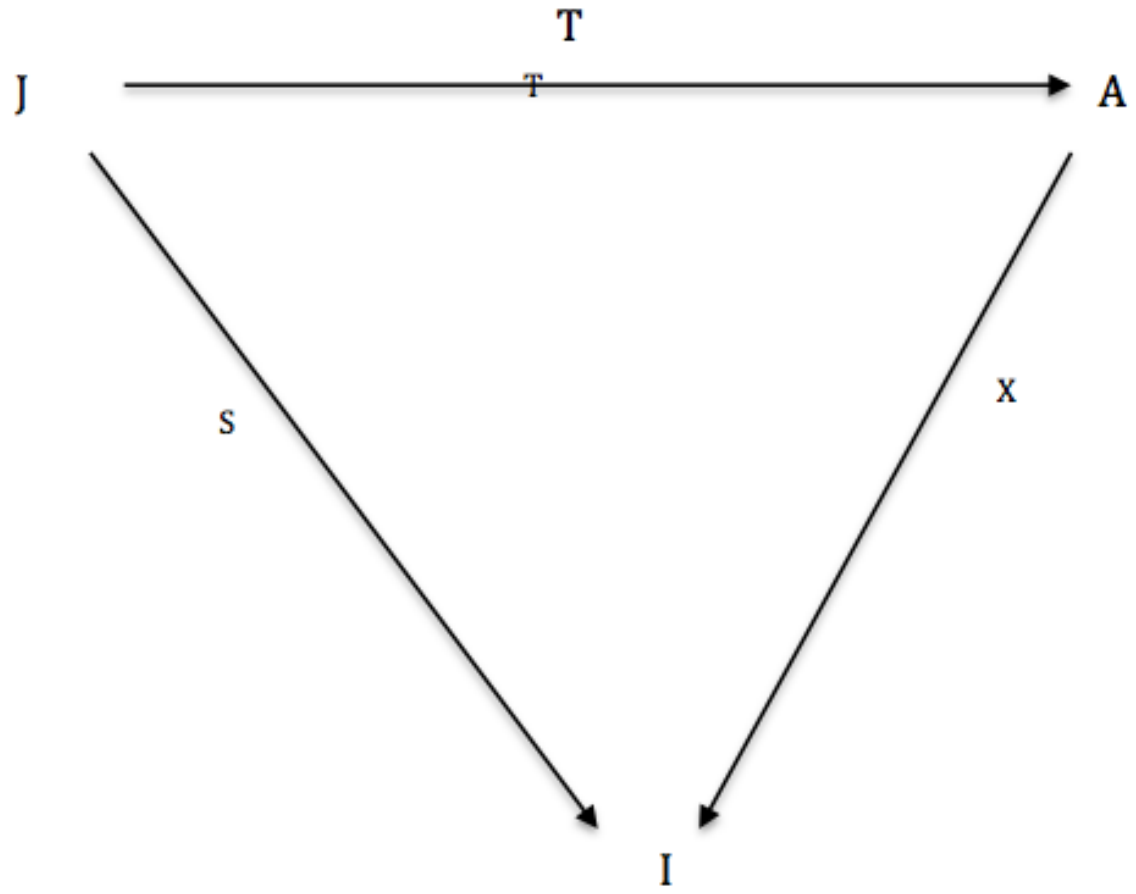

Figure 5: Link between use and structure

3. General Theory

Notation
Let X be a non-negative discrete random variable. We define X^* the discrete random variable that takes on values 1, 2, 3… $X^*(\omega) = X(\omega) + 1$.
If we write $\mathcal{L}$ as the probability law of X, $\mathcal{L}^*$ designates the probability law of X^*.
We write:

$G_X(\lambda)$: The probability generating function of X (see Appendix).
$G_X^{-1}(\lambda)$: The inverse function of $G_X(\lambda)$.
$G_X'(\lambda)$: The derivative function of $G_X(\lambda)$.
p_0 : The proportion of No-use: $0 \leq p_0 \leq 1$.
$Ge(\beta)$: The Geometric law of parameter $\beta < 1$.
$\mathcal{P}(\alpha)$: The Poisson law of mean α.
$Ber(p)$: The Bernoulli law of parameter $p < 1$.
$\mathcal{B}(n,p)$: The Binomial law of parameters $p < 1, n$ naturel number.
$\mathcal{B}n(\beta,r)$: The Negative Binomial law of parameters $\beta < 1$, r real number.

Below, we describe the general equation for the three-dimensional information product process of Figure 5.

Definition
Let X_i be a sequence of non-negative discrete random and independent variables, identically distributed T is a non-negative discrete random variable: S is defined as the following random variable:

$$S(\omega) = \sum_{i=1}^{T^*(\omega)} X_i(\omega) \quad [1]$$

Meaning

The ω random events are the uses of journals.
$T^*(\omega)$ is the number of articles in the volume. T is the random variable that corresponds to the distribution of structure which has been defined above.
We have previously stated that, in this study, we postulate its existence. In other words, we suppose that it has a statistical regularity that can be modelled.
$X_i(\omega)$ is the number of articles used in a journal's volume comprised of i articles, i varies between 1 and $T^*(\omega)$.
$S(\omega)$ is the total number of articles used in a journal. S is the distribution of use that is usually observed.
From a mathematical perspective, Equation [1] is made up of three unknowns. One of our study's aims is to calculate one unknown according to the two others when possible. The bibliometric aim consists in studying the properties of the distribution of structure (represented by T) in relation to the distribution of use (represented by S).

Assumption

A tenable hypothesis consists in supposing that the X_i variables are independent. This seems to work naturally in many examples.
The X_i variables are supposed to be of the same law. We were not able to find arguments to justify this hypothesis. We admit that this counter is similar in type to the success-breeds-success philosophy (Burrell, 1992, p. 638).

3.1 Stationary Mathematical Model

Equation [1] is a random sum of random variables. Before formulating this article's main result, we shall first study an elementary and realistic case in which the sum is finite, namely where T is a constant random variable. To do this, we assume that the number of articles per volume is constant.

Theorem 3.1 If T is a constant random variable that verifies $P(T = n) = 1, n \geq 1$, then the two trivial results are as follows:
(i) if X_i is Bernouilli $Ber(1 - p_0)$ then equation [1] is Binomial $\mathcal{B}(n, 1 - p_0)$.
(ii) if X_i is Geometric $Ge(\beta)$ then equation [1] is Negative Binomial $\mathcal{B}n(\beta, n)$.

The results are well-known. In the case of (i), S is a finite sum of Bernoulli independent variables, *i.e* a binomial variable. The second case is a finite sum of independent geometric variables, *i.e* a negative binomial variable. Given the usage phenomena that interest us, case (ii) is the only realistic one, as we know that the negative binomial law often gives satisfactory results in many bibliometric processes (Ajiferuke and Famoye, 2016) - We cite the authors: "It was found that due to over-dispersion in most response variables, the negative binomial regression model often seems to be more appropriate for informetric datasets".

Remark 1
The previous theorems are not new, but without them, the following results would not mean anything. We now suppose that T is a non-constant random variable governed by a known probability distribution.

The geometric law, which is the discrete version of the exponential law, is often present in distributions of use, and is of particular interest.
It is used to model the different distributions of formula [1]:
- the distribution of structure T ,
- X_i the number of articles used in a journal or a volume, comprised of i articles (i varies between 1 and $T^*(\omega)$).

More precisely, formula [1] has three unknowns (meaning that the laws are believed to be unknown): we fixed two unknowns in order to calculate the third one. We suppose that both unknown distributions are geometric. Three cases are possible.

All of the theorem demonstrations below use the *Uniqueness Theorem*, which characterizes a PGF (see Appendix), and Theorem 1 given in the Appendix. The demonstration of the latter can be found in the (Burrell, 1992) appendix, and in several online courses (see Appendix).

a. X_i *and* T^* are fixed

Theorem 3.2

(i) If X_i^* is Geometric $Ge^*(\alpha)$ and T^* is Geometric $Ge^*(\beta)$ then S :

$$S(\omega) = \sum_{i=1}^{T^*(\omega)} X_i^*(\omega)$$

is Geometric $Ge^*(\alpha\beta)$.

(ii) If X_i is Geometric $Ge(\alpha)$ and T^* is Geometric $Ge^*(\beta)$ then S :

$$S(\omega) = \sum_{i=1}^{T^*(\omega)} X_i(\omega)$$

is Geometric $Ge(h)$ with $h = \frac{\alpha.\beta}{1-\alpha(1-\beta)}$.

Proof (i)
According to Theorem 1 and Proposition 2, (see Appendix), we have:

$G_S(\lambda) = G_{T^*}(G_{X_i^*}(\lambda)) = \frac{\frac{\lambda.\alpha.\beta}{1-(1-\alpha).\lambda}}{1-\frac{(1-\beta)\lambda\alpha}{1-(1-\alpha).\lambda}} = \frac{\lambda.\alpha.\beta}{1-(1-\alpha.\beta).\lambda}$

hence S is Geometric $Ge^*(\alpha.\beta)$. □

Proof (ii)
According to Theorem 1 and Proposition 2, (see Appendix), we have:

$G_S(\lambda) = G_{T^*}(G_{X_i}(\lambda) = \frac{\frac{\beta\alpha}{1-(1-\alpha)\lambda}}{1-(1-\beta).\frac{\alpha}{1-(1-\alpha)\lambda}} = \frac{h}{1-(1-h)\lambda}$ where $h = \frac{\alpha.\beta}{1-\alpha(1-\beta)}$

hence S is Geometric $Ge(h)$. □

Remark 2
This theorem exists in another form in Egghe (1994) in which he studies the distribution of the production of articles (in the line of Lotka) with multiple authors, using two geometric laws.

$$\varphi(i) = \sum_{j=1}^{\infty} \varphi_j(i).\psi(j)$$

$\varphi_j(i)$ = fraction of the authors with *i* publications with the condition that all papers have exactly *j* authors,
$\psi(j)$ = fraction of the papers that have *j* authors,
$\varphi(i)$ = fraction of the authors with *i* publications.

φ_j is calculated thanks to discrete convolutions where we suppose that φ_1 is geometric. If we also suppose that ψ is equally geometric, φ is a geometric distribution.

Putting this in perspective with Egghe's result is interesting since we are in two very distinct situations: while an article is published in a journal's single volume, an article can have multiple authors. Egghe is more specifically interested in the informetric processes in which items can have multiple sources (Egghe, 2005, chapter 7).

Theorem 3.3
If X_i is Bernoulli $Ber\ (1-p_0\ )$ and T^* is Geometric $Ge^*(\beta)$, then S is the sum of Geometric $Ge^*(q)$ and Bernoulli $Ber(1-p_0)$ where $q=\frac{\beta}{1-(1-\beta).p_0}$.

Proof
We have $G_{T^*}(\lambda)=\frac{\beta\lambda}{1-(1-\beta)\lambda}$, $G_{X_i}(\lambda)=(1-p_0)\lambda+p_0$ hence, according to Theorem 1 (see Appendix):
$G_S(\lambda)=\frac{\beta}{1-(1-\beta)((1-p_0)\lambda+\ p_O)}(1-p_0)\lambda+p_0\ )$

We put $q=\frac{\beta}{1-(1-\beta).p_0}$
We then have: $G_S(\lambda)=f(\lambda).((1-p_0)\lambda+p_0)$ with $f(\lambda)=\frac{\beta}{1-(1-\beta)((1-p_0)\lambda+\ p_0\ )}$
We can write: $f(\lambda)=\frac{q}{1-(1-q)\lambda}$ where $q=\frac{\beta}{1-(1-\beta).p_0}$
According to proposition 1 (see Appendix) we deduce that:
$S=G^*+Y$
where Y is Bernouilli $Ber(1-p_0)$ and G^* is Geometric $Ge^*(q)$ $q=\frac{\beta}{1-(1-\beta).p_0}$. □

Remark 3
Thus S converges in law when $p_0\to 0$ towards a geometric distribution $Ge^*(\beta)$.
This theorem means that, if $p_0\to 0$, the distribution of use and the distribution of structure are identical. This theorem sheds light on the significance of the notion of structure. When all sources are producing, the distribution of structure is the distribution of use. Therefore, it does not seem strange to define the distribution of structure by counting the number of leading articles in a volume (see Section 2). In this case the leading articles are those that are requested or cited.

b. $X_i\ and\ S$ are fixed

Theorem 3.4
If X_i is Geometric $Ge^*(q)$ and S is Geometric $Ge^*(\alpha)$ and $\alpha<q$, hence T^* is Geometric $Ge^*(h)$ where $h=\frac{\alpha}{q}$

Proof
According to Theorem 1 (see Appendix) $G_S(\lambda)=G_{T^*}(\frac{q.\lambda}{1-(1-q).\lambda})$ we put $z=\frac{q.\lambda}{1-(1-q).\lambda}$
hence $\lambda=\frac{z}{q+z.(1-q)}$. We know that $G_S(\lambda)=\frac{\alpha.\lambda}{1-(1-\alpha).\lambda}$, thus $G_{T^*}(z)=\frac{\alpha.z}{q+z(1-q)-(1-\alpha).z}$
$G_{T^*}(z)=\frac{\alpha.z}{q-(q-\alpha).z}=\frac{\frac{\alpha}{q}.z}{1-\left(1-\frac{\alpha}{q}\right).z}$ hence $T^*\ is\ Geometric\ Ge^*(h)\ \ wih\ h=\frac{\alpha}{q}$. □

Remark 4
This theorem gives meaning to the distribution of structure: if both distributions of use are geometric, then the distribution of structure is necessarily geometric. The condition $\alpha<q$ is natural, given that: $P(S=1)\le P(X_i=1)\Longrightarrow\alpha\le q$. The percentage of articles requested once is necessarily higher than the percentage of volumes requested once. Such a study on the uses of several collections of journals has been conducted in (Lafouge, 1998) where the distributions of use were observed at several levels of granularity.

c. $T^*\ and\ S$ are fixed

Theorem 3.5
If T^* is Geometric $Ge^*(\beta)$ and S is Geometric $Ge(\alpha)$, then X_i is Geometric $Ge\ (h)$ where

$h = \frac{\alpha}{\alpha+\beta-\alpha\beta}$.

Proof
$G_{T^*}(\lambda) = \frac{\beta.\lambda}{1-(1-\beta).\lambda}$ thus $G_{T^*}^{-1}(\lambda) = \frac{\lambda}{\beta+\lambda(1-\beta)}$, Theorem 1 (see Appendix) allows us to write:

$G_{X_i}(\lambda) = G_{T^*}^{-1}(G_S(\lambda))$
However $G_S(\lambda) = \frac{\alpha}{1-(1-\alpha).\lambda}$ thus $G_{X_i}(\lambda) = \frac{\alpha}{\alpha(1-\beta)+\beta-(1-\alpha).\beta.\lambda} = \frac{h}{1-(1-h).\lambda}$ where $h = \frac{\alpha}{\alpha+\beta-\alpha.\beta}$
thus X_i *is Geometric* $Ge(h)$. □

The following theorem summarizes the properties of the geometric law in the IPP defined by Equation [1]:

Theorem 3.6
Let X_i the number of articles used in a journal's volume comprised of i articles (X_i are identically distributed, independent variables) T the distribution of structure, S the total number of articles used in a journal is defined as the following random variable:

$$S(\omega) = \sum_{i=1}^{T^*(\omega)} X_i(\omega) \quad [1]$$

(i) If X_i is Geometric $Ge(\alpha)$ and T^* is Geometric $Ge^*(\beta)$, then S is also Geometric $Ge\,(h)$.

where $h = \frac{\alpha.\beta}{1-\alpha(1-\beta)}$
(ii) If X_i is Geometric $Ge^*(q)$ and S is Geometric $Ge^*(\alpha)$ with $< q$, then T is also Geometric $Ge\,(h)$ where $h = \frac{\alpha}{q}$.

(iii) If T^* is $Ge^*(\beta)$ and S is Geometric (α) , then X_i is Geometric $Ge\,(h)$ where $h = \frac{\alpha}{\alpha+\beta-\alpha\beta}$.

If at least two distributions are geometric, then the third one is necessarily geometric. This result is note-worthy.

In reality, it is difficult to observe distribution X_i. Therefore, in the following paragraphs, we do not make assumptions on this particular distribution.

3.2 Pseudo-stationary Mathematical Model

Our aim in this section is to take into account time dependence without introducing it explicitly in our equations. We do not formulate a hypothesis on the nature of X_i . We suppose that X_i all have the same law and depend upon at least one p_0 parameter, $0 \leq p_0 < 1$, where p_0 is the proportion of No-use. We shall write them as $X_i^{p_0}$. We also suppose that the moment of order one, written $E(X_i^{p_0})$, exists.
Equation [1] is written in this way:

$$S_{p_0}(\omega) = \sum_{i=1}^{T^*(\omega)} X_i^{p_0}(\omega) \quad [1a]$$

We look for the law of distribution S_{p_0} (use), supposing that we know the law of distribution T (structure). To do this, we make the following hypotheses:

We suppose $p_0 \to 1$ [a]:
[a] expresses a certain type of temporal dependency in the model, which explains the term "pseudo-stationary" in this section's title.
Let $M > 0$ we suppose that the following three conditions are verified:

$$\lim_{p_0 \to 1} E(X_i^{p_0}) = 0 \quad [b]$$

$$\lim_{p_0 \to 1} E(T) = \infty \qquad [c]$$
$$\lim_{p_0 \to 1} E(T).E(X_i^{p_0}) = M \qquad [d]$$

Meaning of the boundary conditions:
- (b) translates the fact that, throughout the studied period, use (citations, downloads) decreases: this is known as the obsolescence of information: obsolescence is usually expressed by the decline in time of the use of a document,
- (c) translates the fact that the number of published articles increases throughout the studied period,
- (d) translates the fact that a stationary state has been reached: there is a balance between the obsolescence of information on one hand, and the increase in the number of published articles on the other.

Before demonstrating the theorem, the limit of S_{p_0} when $p_0 \to 1$, we must first demonstrate intermediate results [3] and [4]:
According to Lemma 1 (see Appendix) the PGF of [1a] is:
$G_{S_{p_0}}(\lambda) = G_T(G_{X_i^{p_0}}(\lambda)).(G_{X_i^{p_0}}(\lambda)$, according to Lemma 2 (see Appendix) we have:
$\lim_{p_0 \to 1} G_{S_{p_0}}(\lambda) = \lim_{p_0 \to 1} G_T(G_{X_i^{p_0}}(\lambda))$ [3]

Furthermore, when $G_{X_i^{p_0}}(\lambda)$ is the PGF of $X_i^{p_0}$; we know (see Appendix, Proposition 1) that $G_{X_i^{p_0}}(1) = 1$ and $G'_{X_i^{p_0}} = E(X_i^{p_0})$. We can therefore write, according to Lemma 3 (see Appendix, we put $f_\alpha(\lambda) = G_{X_i^{p_0}}(\lambda)$), that when $p_0 \to 1$

$$G_{X_i^{p_0}}(\lambda) \sim G_{X_i^{p_0}}(1) + G'_{X_i^{p_0}}(1).(\lambda - 1) \sim 1 + G'_{X_i^{p_0}}(1).(\lambda - 1) \quad [4]$$

Theorem 3.7
S_{p_0} being the random variable defined by equation [1a], supposing that the boundary conditions [b], [c] and [d] are verified, we have:

(i) If T is Poisson $\mathcal{P}(\beta)$, then S_{p_0} converges in law when $p_0 \to 1$ towards a Poisson $\mathcal{P}(M)$.

(ii) If T is Geometric $Ge(\beta)$, then S_{p_0} converges in law when $p_0 \to 1$ towards a Geometric $Ge(\frac{1}{1+M})$.

(iii) If T is Negative Binomial $Bn(\beta, r)$, then S_{p_0} converges in law when $p_0 \to 1$ in law towards a Negative Binomial $Bn(\frac{1}{1+M}, r)$.

Proof (i)
According to [3], we have: $\lim_{p_0 \to 1} G_{S_{p_0}}(\lambda) = \lim_{p_0 \to 1} G_T(G_{X_i^{p_0}}(\lambda))$.
According to our hypothesis and [4], we have:

$$G_T\left(G_{X_i^{p_0}}\right)(\lambda) \sim \exp\left(\beta.(1 + G'_{X_i^{p_0}}(1)(\lambda - 1) - 1)\right) \sim \exp\left(\beta.G'_{X_i^{p_0}}(1).(\lambda - 1)\right)$$

We have $E(T) = \beta$, according to boundary condition [b], we have: $\beta \to \infty$ thus we have :
$\lim_{p_0 \to 1} E(T).E(X_i^{p_0}) = \lim_{p_0 \to 1} \beta.G'_{X_i^{p_0}}(1) = M$.
We obtain the result $G_{S_{p_0}}(\lambda) \sim \exp(M(\lambda - 1))$, thus, according to proposition 1 (see Appendix) :
S_{p_0} *converges in law towards* $\mathcal{P}(M)$ when $p_0 \to 1$ □

Proof (ii)
We have $E(T) = \frac{1-\beta}{\beta}$, according to boundary condition [b], we have: $\beta \to 0$ thus we have :
$\lim_{p_0 \to 0} E(T).E(X_i^{p_0}) = \lim_{p_0 \to 0} \frac{1}{\beta} G'_{X_i^{p_0}}(1) = M$
for the same reasons as above, we can write:

$$G_T\left(G_{X_i^{p_0}}\right)(\lambda) \sim \frac{\beta}{\beta - (1-\beta)(G'_{X_i^{p_0}}(\lambda)(\lambda - 1)} \sim \frac{1}{1 - (\frac{1}{\beta} - 1)G'_{X_i^{p_0}}(1)(\lambda - 1)} \sim \frac{1}{1 - M(\lambda - 1)}$$

Let $G_T\left(G_{X_i^{p_0}}\right)(\lambda) \sim \frac{\frac{1}{1+M}}{1-(1-\frac{1}{1+M})\lambda}$, thus, according to proposition 1 (c) (see Appendix) :
S_{p_0} *converges in law towards* $Ge\ (\frac{1}{1+M})$ when $p_0 \to 1$. □

Proof (iii)

This demonstration follows the same reasoning as the previous ones.

Remark 5

With this theorem, we wanted to consider a general case. We do not hypothesize on the X_i distributions except to say that these distributions have the same law and have a moment of order 1. This last hypothesis eliminates the very important case of power distributions, which are inevitable in informetrics.
This theorem seems to say:
If the distribution of structure follows a law (Poisson, geometric…), then the distribution of use follows a law of the same nature.

The question that arises then is: does a counter example exist? Theorem 3.8 answers this question: when a binomial distribution is chosen, then the distribution of use is not a binomial distribution but a Poisson distribution.

Theorem 3.8
S_{p_0} being the random variable defined by equation [1a], let us suppose that boundary conditions [b], [c] and [d] are verified we have :
If T is Binomial $\mathcal{B}(n,p)$, then S_{p_0} converges in law when $p_0 \to 0$ towards a Poisson $\mathcal{P}(M)$.

Proof
We first recall the following analytical result :
$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = \exp(a)$ [5]
According to our hypothesis, we have $G_T(\lambda) = \left(q.\lambda + (1-q)\right)^n \quad n\ positive\ integer, q < 1.$
According to [3] we have: $\lim_{p_0\to 1} G_{S_{p_0}}(\lambda) = \lim_{p_0\to 1} G_T(G_{X_i^{p_0}}(\lambda))$.
According to [4] we have:
$G_T\left(G_{X_i^{p_0}}\right)(\lambda) = \sim(\, q\left(\ 1 + G'_{X_i^{p_0}}(1).(\lambda-1)\ \right) + 1 - q\,)^n$
However $E(T) = nq$, we then have according to [c] $\lim_{p_0\to 1} E(T) = \infty$, thus $\lim_{p_0\to 0}\ n = \infty$, as $q < 1$.
However, according to [d], $\lim_{p_0\to 1} E(T).E(X_i^{p_0}) = \lim_{p_0\to 1} nq.G'_{X_i^{p_0}}(1) = M$.
Let $G_T\left(G_{X_i^{p_0}}\right)(\lambda) \sim (1\ + \mathrm{q}.G'_{X_i^{p_0}}(1).(\lambda-1)\,)^n \sim (1+\frac{M}{n}(\lambda-1))^n$
According to [5], $G_T\left(G_{X_i^{p_0}}\right)(\lambda) \sim \exp(M.(\lambda-1))$.
Thus S_{p_0} converges in law when $p_0 \to 1$ towards a Poisson distribution $\mathcal{P}(M)$.

Remark 6
The hypothesis of a binomial distribution is realistic if the distribution of structure consists in counting the number of articles per volume.

Remark 7
Theorems 3.7 and 3.8 have been demonstrated in previously cited publications (Lafouge and Lainé Cruzel S. 1997) (Lafouge and Guinet, 1999) (lafouge, 2001) using a more traditional method. These demonstrations are in need of a hypothesis for the X_i law. However, this is not the case here. We can use Theorem 3.8 with any type of distribution. The proposed method in this study is more general than in previous publications.
We point out a mistake in Lafouge and Lainé Cruzel (1997): Theorem 1 on page 525 is incorrect. The distribution is not geometric but is rather the sum of a geometric distribution and a Bernoulli distribution. Theorem 3.3 demonstrates this result.

Remark 8
A connection can be made between Theorem 3.8 and a well-known result in probabilities, which is the convergence in law of the binomial law $\mathcal{B}(n\ p)$ towards the Poisson law $\mathcal{P}(M)$ when $n \to \infty, p \to 0\ and\ n.p \to M$.
In this case, the binomial distribution can model book loans in a library, in which:
- n is the number of publications,
- p is the probability of borrowing a publication,
- M is the average number of loans for a publication.
The boundary conditions are the same ones as in Theorem 3.8 and the result is identical. The probability law which models the book loans or the uses of the articles is a Poisson distribution.

4. Discussion and Conclusion

The three dimensional IPP, shown on Figure 1, can describe many informetric processes. We have not yet confronted the theoretical results with real data, at least to calculate the distribution of structure. Therefore the reader may remain sceptical, since one might ask what the value of a result is without experimentation. For now, our aim is theoretical: to find characteristics in the geometric law – not in the mathematical sense, but in distributional terms – in which the size frequency function of a three-dimensional IPP is geometric. In the preface to his book on Lotkaian Informetrics, Egghe (2005) said: "The only axiom used in this book is that the size-frequency function is Lotkaian, *ie* a power function". Of course, the author could establish such an axiom given the amount of articles he published on the subject and given the many theoretical and practical studies done in the field of Lotkaian Informetrics.

We more modestly postulate the existence of a distribution of structure or a distribution of use that is geometric. The results summarized in Theorem 3.6 are unexpected. Along with such a result, a question arises and remains, for now, an open problem: is the geometric law alone in verifying Theorem 3.6? If this is indeed the case, it could be pertinent to create three-dimensional informetrics, which would be based on this theorem. The results obtained in Section 3.2 are interesting, since we do not hypothesize on X_i .

Finally, another open problem concerns the Lotkaian Informetrics. The problem is not easy to solve since no analytical expression exists for the power law's PGF. Furthermore, in the pseudo-stationary model, laws must have moments, which, as we know, is not always true for power laws.

Burrel (2008) extends Informetrics by defining Pareto's law, known as type II. He suggests that: "It would be interesting to see to what extent Egghe's development of Lotkaian informetrics can be replicated using the Pareto type II family, including the right truncated version." Therefore, the results obtained in the present article with the geometric distribution lead us to develop our theory in a similar way as what we had started to explore in Lafouge (2007) as in families of exponential distributions that are, in fact, the continuous version of geometric distributions.

Appendix

In this appendix, we recall the properties of the probability generating functions (PGF) that are used in this article. Many bibliographical references exist online:

such as the course "Applied Mathematics § Theoretical Physics" from Queen's University in Belfast: http://www.am.qub.ac.uk/users/g.gribakin/sor/Chap3.pdf

We chose to demonstrate only Lemma 2 in this appendix.

Definition and Uniqueness Theorem

Let X be a discrete non-negative random variable:

we write: $p_k = P(X = k),\ \ k = 0,1,2 \ldots$

The probability generating function (PGF) of X, written G_X , is defined as:

$$G_X(\lambda) = \sum_{k=0}^{\infty} p_k \,.\, \lambda^k$$

Let X and Y be two discrete non-negative random variables.

If X and Y have PGF G_X and G_Y respectively, then (i) and (ii) are equivalent:

(i) $G_X(\lambda) = G_Y(\lambda)$ for all λ .

(ii) $P(X = k) = P(Y = k), \quad k = 0, 1, 2 \ldots$

Theorem 1

Let X_i be a sequence of non-negative discrete random and independent variables identically distributed, each with PGF G_X. T is a non-negative discrete random variable, we define S as the random variable:

$$S(\omega) = \sum_{i=1}^{T^*(\omega)} X_i\,(\omega) \quad [1]$$

The probability generating function G_S of S is defined as:

$$G_S(\lambda) = G_{T^*}(G_X(\lambda))$$

If X_i and T have moments $E(X_i)$ and $E(T)$ respectively we infer the following corollary:

Corollary 1

$$E(S) = E(T^*).E(X)$$

Proposition 1

(i) $G_X(0) = 1$; $G'_X(0) = E(X)$.

(ii) Let X and Y be independent and let $Z = X + Y$ then $G_Z(\lambda) = G_X(\lambda).G_Y(\lambda)$.

(iii) If $G_{X_\alpha} \to G_{X_\beta}$ when $\alpha \to \beta$ then X_α converges in law towards X_β.

Lemma 1

Let X_i be a sequence of non-negative discrete random and independent variables identically distributed, each with PGF G_X. We suppose that the moments of order one written $E(X)$ exists. T is a non-negative discrete random variable with moment $E(T)$. We define S as the random variable:

$$S(\omega) = \sum_{i=1}^{T^*(\omega)} X_i\,(\omega)$$

(i) $G_S(\lambda) = G_T\big(G_X(\lambda)\big).G_X(\lambda)$

(ii) $E(S) = E(T).E(X) + E(X)$

Proposition 2

Let X be a non-negative discrete random variable and X^* defined as $X^*(\omega) = X(\omega) + 1$ we have:

$$G_{X^*}(\lambda) = \lambda.G_X(\lambda)$$

For the common distribution, the PGFs are :

1. If X is Constant: $p_n = 1,\ p_k = 0, k \neq n$
   $G_X(\lambda) = \lambda^n$

2. If X is Bernoulli $Ber(1 - p_0)$: $0 \leq p_0 < 1$.
   $G_X(\lambda) = (1 - p_0)\lambda + p_0$

3. If X is Binomial $\mathcal{B}(n, 1 - p_0)$: $0 \leq p_0 < 1, n\ positive\ integer$
   $G_X(\lambda) = ((1 - p_0)\lambda + p_0)^n$

4. If X is Geometric $Ge(\beta)$: $p_k = \beta.(1 - \beta)^k,\ k = 0,1 \ldots, 0 < \beta < 1$.
   $G_X(\lambda) = \dfrac{\beta}{1 - (1 - \beta)\lambda}$

5. If $X$is Negative Binomial $X\ \mathcal{B}n(\beta, r)$:
   $p_0 = \beta^r, p_k = r(r + 1) \ldots (r + k - 1).\frac{\beta^r}{k!}(1 - \beta)^k : k = 1,2 \ldots, r > 0,\ 0 < \beta < 1$.
   $G_X(\lambda) = \left(\frac{\beta}{1-(1-\beta)\lambda}\right)^r$

6. If $X$is Poisson $\mathcal{P}(\alpha) : p_k = Exp(-\alpha)\,.\alpha^k.\frac{1}{k!}$, $k = 0,1 \ldots, 0 < \alpha$.
   $G_X(\lambda) = Exp\,(\alpha.(\lambda - 1))$

Lemma 2

Let X^{p_0} be a sequence of random variables such that $\lim_{p_0 \to 1} E(X^{p_0}) = 0$. Then X^{p_0} converges in probability towards a certain variable $Z\ where\ P(Z = 0) = 1, P(Z = k) = 0,\ k \neq 0$. We also have $\lim_{p_0 \to 1} G_{X^{p_0}}(\lambda) = 1$.

Proof

Indeed, using Markov inequality $\forall\, \varepsilon > 0, P(X^{p_0} > \varepsilon) \leq \frac{E(X^{p_0})}{\varepsilon}$ thus $\lim_{p_0 \to 1} P(X^{p_0}) > \varepsilon) = 0$ thus X^{p_0} converges in probability towards a certain probability when $p_0 \to 1$. Therefore, according to proposition 1 (iii), we immediately deduce that $\lim_{p_0 \to 1} G_{X^{p_0}}(\lambda) = 1$. □

Lemma 3
We recall the following analytical result. Let there be the following function: $\lambda \to f_\alpha(\lambda)$ defined on $\mathbb{R}$, differentiable where α is a parameter and where we suppose that $\lim_{\alpha \to \alpha_0} f'_\alpha(1) = 0$. We have $f_\alpha(\lambda) \sim f_\alpha(1) + f'_\alpha(1)(\lambda - 1)$ when $\to \alpha_0$.

References

Ajiferuke, I. and Famoye, F. (2016). Modelling count response variables in Informetrics studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics*, Vol 9 N°3, 499-513.

Bagust, A., (1983). A circulation model for busy public libraries. *Journal of Documentation,* Vol 39 N°1, 24–37.

Bertoli-Barsotti, L. and Tommaso L. (2015). On a formula for the h-index. *Journal of Informetrics,* Vol 9 N°4, 762-776.

Boukacem-Zeghmouri, C. and Bador, P. and Lafouge, T. and Prost, H. (2016). Relationships between consumption, publication and impact in French universities in a value perspective: a bibliometric analysis. *Scientometics*, Vol 106, N°1, 88–105.

Bradford, S.C. (1934). Sources of information on specific subjects. *Engineering* 26 janvier 1934, 85-86.

Burrell, QL. (1980). A simple stochastic model for library loans. *Journal of Documentation*, Vol 36 N°2, 115-132.

Burrell, QL. (1988). Predictive aspects of some bibliometric processes. Informetrics 87/88: *Select proceedings of the first international conference on bibliometrics and theoretical aspects of information retrieval*. Elsevier, Amsterdam 1988.

Burrell, QL. (1990). Using the Gamma-Poisson Model to Predict library circulations. *Journal of the American Society for Information Science*, Vol 41 N°3, 164-170.

Burrell, QL. (1992). A simple model for linked informetrics processes. *Information Processing & Management*, Vol 38 N°1, 637-645.

Burrell, QL. and Fenton, M. (1993). Yes, the GIGP, really does work and is workable! *Journal of the American Society for Information Science and Technology*, Vol 44 N°2, 61–69.

Burrell, QL. (2008). Extending Lotkaian informetrics. *Information Processing and Management.* Vol 44 N°5, 1794-1807.

Burrell, QL. (2014). The individual author's publication-citation processes: theory and practice. *Scientometrics,*Vol 98 N°1, 725-742.

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

Egghe, L., (1994). Special Features of the Author-Publication Relationship and a New Explanation of Lotka's law based on convolution theory. *Journal of the American Society for Information Science and Technology*, Vol 45 N°6, 422-427.

Egghe, L., (2004). Solution of a problem of Buckland on the influence of obscolescence on scattering *Scientometrics,* Vol 59 N°2, 225-232.

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

Egghe, L. and Rousseau, R. (1995). Generalized success-breeds-success principle leading to time-dependent informetric distributions. *Journal of the American Society for Information Science.* Vol 46 N°6, 426-445.

Egghe, L. and Rousseau, R. (2012). Theory and practice of the shifted Lotka function. *Scientometrics* Vol 91 N°1, 295-301.

Haitun, D. (1982). Stationary scientometric distributions. *Scientometrics* n°4, Part I, 5-25, Part II, 89-104, Part III, 181-194.

Lafouge, T. (1995). Stochastic information field. *The international Journal of Scientometrics and Informetrics*, 1(2), 57-64.

Lafouge, T. (1998). Mathématiques du document et de l'information, Bibliométrie distributionnelle. Chapitre 4. *Habilitation à diriger des recherches*, Université Lyon3.

Lafouge, T. (2001). A mathematical model of documents circulation : use distribution, utility distribution, content distribution: example of scientific articles circulation in journals. *In Proceedings of the eight conference of the international Society for Scientometrics and Informetrics*, Sydney Australia 2001, 327–337.

Lafouge, T. (2007). The source-item coverage of the exponential function. *Journal of Informetrics*, Vol1 N°1, 59–67.

Lafouge, T. and Guinet, E. (1999). A new explanation of the negative binomial law and the Poisson law with regard to library journal circulation data. *Journal of Information Science*, Vol 25 N°1, 89–93.

Lafouge, T. and Lainé Cruzel, S. (1997). A new explanation of the geometric law in the case of library circulation data. *Information Processing and Management*, Vol 33 N°4, 523–527.

Leemans, M. and Maes, M. and Rousseau, R., and Ruts, C. (1992). The negative binomial distribution as a trend distribution for circulation data in flemish public libraries. *Scientometrics*, Vol 25 N°1, 47–57.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 16(2), 317–323.

Morse, P.M. (1968). Library effectiveness: A systems approach, Cambridge, Mass: M.I.T Press.

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

Priem, J. and Groth, P. (2012). The Altmetrics Collection. *Plos One* 7(11), e48753

Rousseau, R. (1992). Concentration and diversity of availability and use in information systems: a positive reinforcement model. *Journal of the American Society for Information science,* Vol 43 N°5, 391-395.