Mathl. Comput. Modelling Vol. 18, No. 9, pp. 63–77, 1993
Printed in Great Britain. All rights reserved

0895-7177/93 $6.00 + 0.00
Copyright© 1993 Pergamon Press Ltd

Consequences of Lotka's Law in the Case of Fractional Counting of Authorship and of First Author Counts

L. EGGHE
Limburgs Universitair Centrum
Universitaire Campus, B-3590 Diepenbeek, Belgium*
and
Universitaire Instellin Antwerpen
Universiteitsplein 1, B-2610 Wilrijk, Belgium

(Received August 1992; revised and accepted March 1993)

Abstract—In a recent paper, Rousseau [1] notes the fact that if we give weights of $1/m$ to each author in an m-authored paper, Lotka's law does not apply. However, he also notes that the function modeling the number of authors with weights j, $j \geq 0$, starts increasing from zero to about one and then decreases. In that paper, the present author is quoted as stating that this is not a breakdown of Lotka's law but merely a composition of two Lotka laws: one for $j \geq 1$ modeling papers per author, and one for $0 \leq j \leq 1$ modeling authors per paper. The stochastic problem of how these two laws fit into each other was not solved in Rousseau's paper, however. This is done in this paper, where we will show that the weight-distribution function has indeed a maximum for the weight equal to one. We then study the same problem in the case where only the first author gets weight one and the others weight zero. We solve this case completely providing a formula for the probability of the weights. Also this function has a maximum for the weight equal to one. The main tool in these models is the technique of repeated convolution of continuous or discrete distribution functions.

1. INTRODUCTION

The problem studied in this paper goes back to the year 1926, when Alfred Lotka (see [2]) introduced his celebrated frequency "law" on the fraction of the authors that publish x papers, ($x \in \mathbf{R}$) (on a fixed topic and in a fixed time period, e.g., one year):

$$\varphi(x) = \frac{C}{x^{\alpha}}, \tag{1}$$

where α is usually ≥ 1. Most classically, Lotka's law was expressed for $\alpha = 2$, C then being $6/\pi^2$ (this follows from the requirement that $\sum_{x=1}^{\infty} \varphi(x) = 1$ and the fact that $\sum_{x=1}^{\infty}(1/x^2) = \frac{\pi^2}{6}$, as is well-known, see, e.g., [3]); hence $C \approx 0.61$, meaning that about 60% of the authors publish only one paper. But already in [2], one acknowledges the need for a more general law as in (1), with $\alpha \geq 1$.

This celebrated law of Lotka was then followed by the apparent different laws of Bradford and the law of Mandelbrot (linked with Zipf's law), see [4–6], formulated differently using, e.g., the formalism of articles in journals (as in a bibliography) or of occurrences of words in texts

The author is grateful to R. Rousseau for mentioning the problem of modeling fractional author counts and for correcting a mistake in an earlier version of this paper.
*Permanent address.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

(cf. the terminology of Herdan: a word is the "type" and the occurrence of this word in a text is the "token"—see [7]).

Only later the proof of the equivalence of most of these laws was given (see [8] for an extensive discussion on this matter and for more references), using the generalized "dual" (cf. [9,10]) framework of sources (the "type," i.e., the objects that produce) and items (the "token," i.e., the objects that are produced). Examples are given above: authors "produce" articles, journals "produce" articles too, word types "produce" the occurrences and so on. In fact, this dual mechanism is also encountered outside this field, e.g., employees "produce" their salary, cities "have" inhabitants and so on.

In studying Lotka's law, in the framework of authors who publish articles, something special is going on. Unlike the other examples of dual situations, sources (authors) and items (articles) can be interchanged to yield: articles are the sources and they "produce" (i.e., have) authors (i.e., the items in this case). For example, an article could be written by 3 authors but an article cannot be published by 3 journals!

The calculation of author weights in such situations is not uniquely determined. One method is the one of "fractional counting," i.e., an author receives a weight $1/3$ in a 3-authored paper (one of them being this author). Another method consists of only giving the first named author a weight 1 and the others a weight 0. This method is called "straight counting." Finally, one could also give every author a weight 1, called "total counting" or "normal counting."

Let us focus on the fractional counting method. In (1), we assumed a total counting procedure (i.e., every article (co-)authored by an author counts as one publication). One can ask if this law (possibly with another α) is true in the case of fractional counting. This problem goes back to Bookstein [11] who proved that under certain conditions Lotka's law is stable for the applied method of authors counts. If we can speak of a version of the law with the form C/x^{α} to describe the productivity when we give full weight to every author of a paper, this will also be the case, but possibly with another α, if we give fractional weights of authorship.

This "property," however, is easily seen to be false. Indeed, as Rousseau points out [1] only a few papers are written by, say 8 authors, so that the weight $1/8$ will occur only a few times. Probably more papers are written by 4 authors, so that the weight $1/4$ will occur more often. Even more frequent should be two-authored papers so that the weight $1/2$ occurs even more frequently. Also, $1/2$ might be reached in the case where an author participates in two 4-authored papers. It is also intuitively clear that such an increase continues until about one, after which the classical decreasing Lotka law applies.

Rousseau investigates this idea and finds an initial increase until weights of about one, after which the weight distribution function decreases. Rousseau does not give an explanation of this fact but suggests a lognormal curve, although without any statistical fitting. In this paper, Rousseau acknowledges a remark of the present author as follows: the fact that there is an initial increase followed by a decrease does not imply the breakdown of Lotka's law as suggested by Rousseau but is merely a consequence of it. This can be seen as follows. Let:

$$\varphi(x) = \frac{C}{x^{\alpha}} \tag{2}$$

be the "classical" distribution of Lotka with $\alpha > 1$ for $x \geq 1$, φ measures the density of authors with x publications. For $x \leq 1$, we note that we are counting fractional authorships which are a consequence of multi-authored papers. So, if $x \leq 1$, the dual Lotka law (dual in the sense explained above) $\psi(y)$ could be used, where $y \geq 1$. Here $\psi(y)$ is the density of papers with y authors. As for φ, we can suppose ψ to decrease. Hence, for each such paper, each author receives a weight $x = 1/y$, and this corresponds to a function ψ^*

$$\psi^*(x) = \psi\left(\frac{1}{y}\right), \tag{3}$$

$0 \leq x \leq 1$. Note that ψ^* is an increasing function of x.

As an example, ψ could be a Lotka function as in (2):

$$\psi(y) = \frac{D}{y^{\beta}}, \tag{4}$$

($\beta \geq 1$) and in this case, $\psi^*(x) = D\,x^{\beta}$, indeed an increasing function of x.

From this first idea, we end up with a function ξ illustrated in Figure 1. Note that $\psi^{**}(x)$ is proportional to $\psi^*(x)/x$, since there are $1/x$ times as much authors in papers with $1/x$ authors than there are papers with $1/x$ authors (see further on for the exact proof). Here the function ξ can be written as:

$$\xi(x) = \psi^{**}(x)\,\chi_{[0,1]}(x) + \varphi(x)\,\chi_{]1,\infty[}(x), \tag{5}$$

where χ_A denotes the characteristic function of a certain set A, i.e., $\chi_A(x) = 1$ iff $x \in A$, and $\chi_A(x) = 0$ iff $x \notin A$.

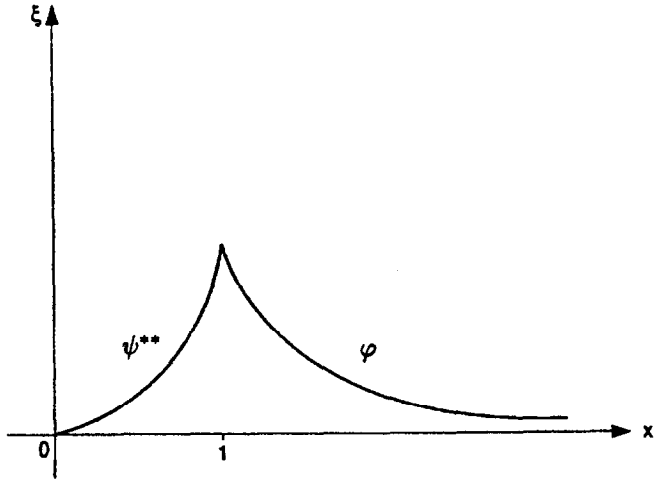

Figure 1. Lotka's law and its dual: primitive model of the fraction of the authors with x papers.

Of course, the above argument is not totally correct, since it assumes that every weight j comes from a publication with $1/j$ authors. This is clearly not true since, for example, a weight of 3/2 could come from one publication with one author and the fractional weight from a 2-authored paper. This weight could also come from three 2-authored papers or six 4-authored papers, or other combinations. Intuitively, we need a stochastic argument to "mix" all these possibilities. This will be done in this paper. The tool that will be used hereby will be convolution theory as explained in, e.g., [12,13]. The reason for using this theory will be explained in the sequel.

We will investigate the problem in the next sections. Section 2 deals with the problem in its full generality, using repeated convolutions of functions of continuous variables. It is indeed so that, since we deal here with the case of fractional counting, discrete convolutions involving rational numbers are needed. It is not clear how to implement these in this context. Therefore, the continuous approach is followed. Note that, theoretically, any rational number can occur as the total fractional weight of an author, and that the rational numbers are dense (in the mathematical sense) in the set of the real numbers. The validity of Rousseau's conjecture will be proved for a class of functions ψ, including the function $\psi(x) = 1/x^3$.

There is also a general interest in this result: from the dual frequency laws φ and ψ (based on counting entire numbers of items) we can deduce the fractional frequency law of the distribution of the authors with certain "fractions" of authorships.

In Section 3, the case of $\psi(x) = 1/x^2$ (which was not covered in the previous section) is investigated, by directly calculating the repeated convolutions. Here, for computational reasons, we had to restrict the calculations to the case of one, two or three papers per author. Already in

these simpler cases, the observation of Rousseau is proved, and we conjecture that this continues for productivities higher than 3.

We will then investigate the same problem in the case of "straight counts" (i.e., only the first author receives weight 1 and all other authors receive weight 0). We now are able to prove, for any number of papers per author, that the weight distribution function attains its absolute maximum at 1.

Only in the case "normal counts," which give weights 1 to all authors do we find that the weight distribution function decreases and in fact it is the Lotka distribution itself.

2. MODELING FRACTIONAL COUNTS—GENERAL THEORY

As explained above, since fractional author counts involve, in principle, all rational numbers, we adapt a continuous approach, since discrete convolutions over the rational numbers are very difficult to calculate.

2.1. DEFINITIONS AND NOTATIONS. *Let $\psi \to \psi(y) : [1, \infty[\to \mathbf{R}^+$ be the density of papers with y authors, $y \geq 1$ and $\varphi \to \varphi(x) : [1, \infty[\to \mathbf{R}^+$ be the density of authors with x papers. For φ, there is no complication to consider $\varphi : \mathbf{N} \to \mathbf{R}^+$ and hence this approach will be followed. Hence, for every $i \in$: $\mathbf{N}$, $\varphi(i)$ denotes the probability that an author has i papers. Fix $i \in \mathbf{N}$ and suppose we consider only those authors with i publications. Then we denote by $f_i(z)$ the density of these authors with a fractional count z. Let $f(z)$ denote the density of the authors with a fractional count z, when all author productions are allowed. We will also assume that the maximal number of authors per paper is finite; we will denote it by N. Hence, it is trivial to see that $f_i(z) = 0$, for $z \in [0, i/N[$ (for all $i \in \mathbf{N}$). Of course, also $f_i(z) = 0$, for $z > i$.*

PROPOSITION 2.1. *For every $z \in [1/N, 1]$,*

$$f_1(z) = \frac{\psi\left(\frac{1}{z}\right)}{z^3 \mu}, \tag{6}$$

where μ is the average number of authors per paper.

PROOF. For every $z \in [1/N, 1]$:

$$\begin{aligned} f_1(z)\,dz &= P(\text{weight of an author} \in [z, z + dz]) \\ &= \frac{\#\text{ authors with weight} \in [z, z + dz]}{\text{total } \#\text{ authors}}. \end{aligned}$$

Since authors have exactly 1 publication, we see that:

$$f_1(z)\,dz = \frac{\frac{1}{z}\left(\#\text{ papers with authors between } \frac{1}{z+dz} \text{ and } \frac{1}{z}\right)}{\text{total } \#\text{ authors}}.$$

Now, since dz is small

$$\frac{1}{z + dz} = \frac{1}{z}\left(\frac{1}{1 + \frac{dz}{z}}\right) \approx \left(1 - \frac{dz}{z}\right);$$

we have:

$$\begin{aligned} f_1(z)\,dz &= \frac{\frac{1}{z}\left(\#\text{ papers with weight} \in \left[\frac{1}{z} - \frac{dz}{z^2}, \frac{1}{z}\right]\right)}{\text{total } \#\text{ authors}} \\ &= \frac{\frac{1}{z}\,\psi\left(\frac{1}{z}\right)\frac{dz}{z^2}\,(\text{total } \#\text{ papers})}{\text{total } \#\text{ authors}} \\ &= \frac{\psi\left(\frac{1}{z}\right)}{z^3 \mu}\,dz. \end{aligned}$$

∎

Note that $x \to f_1(x)$ is indeed a density on $[1, N]$, since $\int_{1/N}^{1} f_1(x)\,dx = 1$, by definition of ψ.

EXAMPLES.

1. For $\psi(x) = \frac{C}{x}$, we see that $C = \frac{1}{\ln N}$ (since ψ is a density) and then, by (6),

$$f_1(z) = \frac{1}{(N-1)\,z^2}.$$

2. For $\psi(x) = \frac{C}{x^2}$, we have $C = \frac{N}{N-1}$ and then, by (6),

$$f_1(z) = \frac{1}{z \ln N}.$$

3. For $\psi(x) = \frac{C}{x^3}$, we have $C = \frac{2N^2}{N^2-1}$ and then, by (6),

$$f_1(z) = \frac{N}{N-1},$$

a constant.

4. For $\psi(x) = \frac{C}{x^4}$, we have $C = \frac{3N^3}{N^3-1}$ and then, by (6),

$$f_1(z) = \frac{2N^2 z}{N^2-1}.$$

We now invoke the following well-known results from probability theory (cf. [12, pp. 144–146]).

THEOREM 2.1. *Let X_1 and X_2 be independent random variables with distribution functions F_1 and F_2, respectively. Then $X_1 + X_2$ has the distribution function $F_1 * F_2$, where*

$$(F_1 * F_2)(x) = \int_{-\infty}^{\infty} F_1(x-y)\,dF_2(y), \tag{7}$$

is the convolution of the two distribution functions.

When the distribution functions have densities, we have the following theorem.

THEOREM 2.2. *The convolution of two distribution functions with densities g_1 and g_2 is a distribution function with density $g_1 * g_2$, where $(g_1 * g_2)(x) = \int_{-\infty}^{\infty} g_1(x-y)\,g_2(y)\,dy$.*

We can now continue with our main theory.

PROPOSITION 2.2. *For every $i \in \mathbf{N}$, $i \geq 2$, and every $z \in [i/N, i]$*

$$f_i(z) = (\underbrace{f_1 * \cdots * f_1}_{i \text{ times}})(z). \tag{8}$$

PROOF. Let $i = 2$. A weight $z \in [2/N, 2)$, in case we only consider authors with two publications, comes from a weight $y \in [1/N, i]$ in the first publication and a weight $z - y \in [1/N, 1]$ in the second one. Here y is arbitrary. Hence, using Theorems 2.1 and 2.2 above:

$$\begin{aligned} f_2(z) &= (f_1 * f_1)(z) \\ &= \int_{D_2} f_1(y)\,f_1(z-y)\,dy, \end{aligned} \tag{9}$$

where the integration is over this region such that y and $z - y \in [1/N, 1]$, the domain of definition of f_1. In general, a weight $z \in [i/N, i]$, in case we only consider authors with i publications, comes

from a weight $y \in [(i-1)/N, i-1]$ in the first $i-1$ publications and a weight $z-y \in [1/N, 1]$ in the i^{th} one. Here y is arbitrary. Hence

$$f_i(z) = \int_{D_i} f_{i-1}(y)\, f_1(z-y)\, dy, \tag{10}$$

where the integration is over this region such that $y \in [(i-1)/N, i-1]$ and $z-y \in [1/N, 1]$, the domains of definition of f_{i-1}, respectively, f_1. Consequently, by the associative property of convolutions,

$$f_i(z) = (\underbrace{f_1 * \cdots * f_1}_{i \text{ times}})(z) \qquad \blacksquare$$

PROPOSITION 2.3. *For all* $z \in [1/N, \infty[$,

$$f(z) = \sum_{i=1}^{\infty} f_i(z)\, \varphi(i), \tag{11}$$

where we put $f_i(z) = 0$ *for* $z \notin [i/N, i]$.

PROOF. This follows from the above and Bayes' rule. $\blacksquare$

NOTE. Formula (11) above gives a direct relationship between the total frequency counts φ and ψ and the fractional frequency counts: for all $z \in [1/N, \infty[$,

$$f(z) = \sum_{i=1}^{\infty} (\underbrace{f_1 * \cdots * f_1}_{i \text{ times}})(z)\, \varphi(i), \tag{12}$$

where

$$f_1(x) = \frac{\psi\left(\frac{1}{x}\right)}{x^3\, \mu}.$$

It is however clear that the concrete calculation of f, based on φ and ψ is far from trivial. We therefore continue our qualitative study of formulae (10) and (11).

The main points that have to be proved (to confirm Rousseau's observation) are

(i) f increases on $[1/N, 1]$;
(ii) f decreases on $]1, 2]$;
(iii) $\lim_{z \to 1} f(z) \geq \lim_{z \to 1} f(z)$.

These conditions will be investigated now. For computational reasons, we will replace $1/N$ by zero in the next theorem. This is never a problem for functions ψ such that $\frac{\psi\left(\frac{1}{z}\right)}{z^3}$ is bounded around zero: then $f_1(z)$ is bounded and the integrals that have to be calculated (to find f_i, $i \in \mathbf{N}$), approach the ones we are calculating if N is sufficiently high. $\frac{\psi\left(\frac{1}{z}\right)}{z^3}$ is bounded for, e.g., all functions ψ of the form

$$\psi(z) = \frac{C}{z^\alpha},$$

$\alpha \geq 3$.

Under this assumption, the domain of integration in (10), for each $z \in [0, i]$ is shown in Figure 2.

We have the following result (proving assertion (i)).

THEOREM 2.3. *The function* f *increases on* $[0, 1]$ *for all functions* ψ *such that the function* $\theta(x) = x^3\, \psi(x)$ *decreases on* $[1, \infty[$ *(e.g., for all functions* $\psi(x) = C/x^\alpha$, $\alpha \geq 3$*).*

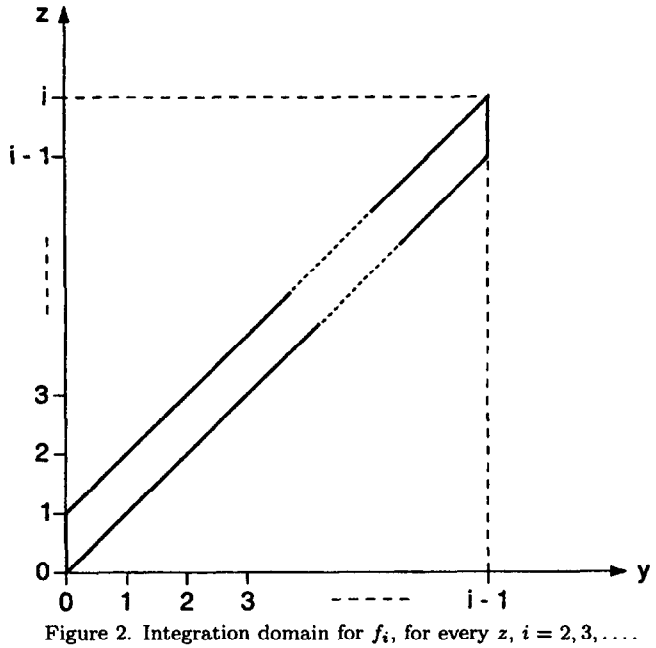

Figure 2. Integration domain for f_i, for every z, $i = 2, 3, \ldots$.

PROOF. It is clear that $f_1(z) = \left(\psi\left(\frac{1}{z}\right)/z^3\mu\right)$ increases on $[0, 1]$, since $x^3\,\psi(x)$ decreases on $[1, \infty[$. Consequently, for all $i \in \mathbf{N}$, $i \geq 2$ and by Figure 2.

$$f_i(z) = \int_0^z f_{i-1}(y)\, f_1(z-y)\, dy,$$

increases on $[0, 1]$ since f_1 does, and since $f_{i-1} \geq 0$ for all i. Finally, f increases on $[0, 1]$, since $\varphi \geq 0$ and

$$f(z) = \sum_{i=1}^{\infty} f_i(z)\, \varphi(i). \qquad \blacksquare$$

At the point $z = 1$, we have the following situation:

$$f(1) = \sum_{i=1}^{\infty} f_i(1)\, \varphi(i), \qquad \text{and}$$

$$\lim_{z \to 1} f(z) = \sum_{i=2}^{\infty} f_i(1)\, \varphi(i).$$

Hence,

$$f(1) - \lim_{z \to 1} f(z) = f_1(1)\, \varphi(1) > 0. \tag{13}$$

This proves (iii). Finally, we prove (ii) and even on the interval $[1, \infty[$.

THEOREM 2.4. *The function f decreases on $[1, \infty[$ for all functions ψ, such that $\theta(x) = x^3\,\psi(x)$ increases on $[1, \infty[$ and $\lim_{x \to \infty} x^3\,\psi(x) \leq \psi(1)$ (e.g., the function $\psi(x) = C/x^3$).*

PROOF.

(A) Let $z \in]1, 2]$. By Figure 2, note that, for $z \in]1, 2]$,

$$f_2(z) = \int_{z-1}^{1} f_1(y)\, f_1(z-y)\, dy, \tag{14}$$

and for $i = 3, 4, \ldots,$

$$f_i(z) = \int_{z-1}^{z} f_{i-1}(y)\, f_1(z-y)\, dy. \tag{15}$$

We invoke Lemma A in the Appendix, yielding

$$f_2'(z) = \int_{z-1}^{1} f_1(y)\, f_1'(z-y)\, dy - f_1(z-1)\, f_1(1)$$
$$= \int_{z-1}^{1} \frac{\psi\left(\frac{1}{y}\right)}{y^3\,\mu^2} \left(-\frac{\psi'\left(\frac{1}{z-y}\right)}{(z-y)^5} - \frac{3\psi\left(\frac{1}{z-y}\right)}{(z-y)^4} \right) - \frac{\psi\left(\frac{1}{z-1}\right)}{(z-1)^3\,\mu^2}\,\psi(1).$$

Hence, $f_2'(z) < 0$ if

$$\psi'\left(\frac{1}{z}\right) \geq -3z\,\psi\left(\frac{1}{z}\right), \tag{16}$$

for all $z \in [0,1]$. This is so since $x^3\,\psi(x)$ increases. Analogously, formula (15) and Lemma A yield, for all $i \geq 3$,

$$f_i'(z) = \int_{z-1}^{z} f_{i-1}(y)\,\frac{1}{\mu} \left(-\frac{\psi'\left(\frac{1}{z-y}\right)}{(z-y)^5} - \frac{3\psi\left(\frac{1}{z-y}\right)}{(z-y)^4} \right) dy + f_{i-1}(z)\,\frac{A}{\mu} - f_{i-1}(z-1)\,\frac{\psi(1)}{\mu}.$$

Applying complete induction, we see that formula (16), together with $f_{i-1}'(z) < 0$ implies $f_i'(z) < 0$, since we assumed that

$$A = \lim_{x\to\infty} x^3\,\psi(x) \leq \psi(1).$$

Since we already proved that $f_2'(z) < 0$, we have shown that, for all $i = 2, 3, \ldots,$ f_i decreases on $]1, 2]$. Since, for $z \in]1, 2]$,

$$f(z) = \sum_{i=2}^{\infty} f_i(z)\,\varphi(i),$$

this is also the case for f. By (13), f decreases on $[1, 2]$.

(B) Let $z \in]i, i+1]$, $i = 1, 2, \ldots$. More generally but analogously with the above arguments, we have, for all $z \in]i, i+1]$,

$$f_{i+1}(z) = \int_{z-1}^{i} f_i(y)\, f_1(z-y)\, dy,$$

and, for all $j \geq i+2$,

$$f_j(z) = \int_{z-1}^{z} f_{j-1}(y)\, f_1(z-y)\, dy;$$

and

$$f(z) = \sum_{j=i+1}^{\infty} f_j(z)\,\varphi(j).$$

It is now clear that the same conditions on ψ imply that f decreases on $]i, i+1]$.

(C) Furthermore, in the connecting points, we have (analogously to (13)), for all $i = 1, 2, \ldots,$

$$f(i) - \lim_{z\to i} f(z) = f_i(i)\,\varphi(i) > 0.$$

This proves that f decreases on $]1, \infty[$, and even on $[1, \infty[$. ∎

COROLLARY 2.1. *For $\psi(x) = \dfrac{2N^2}{(N^2-1)\,x^3}$ (cf. Example 3), we have that the fractional counting function f increases on $[0,1]$ and decreases on $[1, \infty[$ and has negative jumps in every $i \in \mathbf{N}$.*

PROOF. This follows readily from Theorems 2.3 and 2.4. ∎

COROLLARY 2.2. *For $\psi(x) = C/x^\alpha$ ($\alpha \geq 3$), we have that the fractional counting function f increases on $[0,1]$ and has negative jumps in every $i \in \mathbf{N}$.*

PROOF. This follows from Theorem 2.3 and the proof (part C) of Theorem 2.4. ∎

NOTE 1. The estimates in Theorems 2.3 and 2.4 are rough (but we do not know how to refine them), so that, most probably, the results can be extended to a much larger class of α's.

NOTE 2. In [14], these findings are confirmed in practice by using extensive computer simulations. Also, for the power functions $\psi(x) = C/x^\alpha$ the Rousseau observation is tested by simulation and shown to be true. These cases escape, my general theory, however. Nevertheless, in the next section, we can calculate f_1, f_2, f_3, for $\psi(x) = C/x^4$ and we can show that the approximation

$$\sum_{i=1}^{3} f_i(z)\,\varphi(i)$$

of f is indeed increasing on $[0,1]$ and decreasing on $[1,2]$.

3. MODELING FRACTIONAL COUNTS— APPROXIMATIVE THEORY

In this section, we restrict ourselves to the case (cf. Example 2 in the previous section), for $x \in [1, N]$,

$$\psi(x) = \frac{C}{x^4}. \tag{17}$$

Here $C = \dfrac{3N^3}{N^3 - 1}$. Now

$$f_1(z) = \frac{2N^2 z}{N^2 - 1}, \tag{18}$$

for $z \in [1/N,\ 1]$.

Direct calculations now yield (putting $1/N \approx 0$ in the integration interval), for $z \in [0,2]$,

$$f_2(z) = \frac{4N^4}{(N^2-1)^2}\left(\frac{z^3}{3!}\,\chi_{[0,1]}(z) + \left(\frac{z^3}{3} - \frac{3z^2}{2} + 2z - \frac{2}{3}\right)\chi_{]1,2]}(z)\right). \tag{19}$$

The second order approximation g_2 of f is now (using only $\varphi(1)$ and $\varphi(2)$ now), for $z \in [0,2]$,

$$g_2(z) = \frac{2N^2}{N^2-1}\left[\left(z\,\varphi(1) + \frac{2N^2}{N^2-1}\,\frac{z^3}{3!}\,\varphi(2)\right)\chi_{[0,1]}(z) + \frac{2N^2}{N^2-1}\left(\frac{z^3}{3} - \frac{3z^2}{2} + 2z - \frac{2}{3}\right)\varphi(2)\,\chi_{]1,2]}(z)\right]. \tag{20}$$

It is clear that g_2 increases on $[0,1]$ (cf., also the proof of Theorem 2.3). On $]1,2]$, we have that $g_2'(z)$ is proportional to

$$z^2 - 3z + 2,$$

which has $z = 1$ and $z = 2$ as its roots; hence $g_2'(z) < 0$ on $]1,2[$ and so, g_2 decreases on $]1,2]$. Furthermore,

$$g_2(1) - g_2(1+) = \frac{2N^2}{N^2-1}\,\varphi(1) > 0.$$

For f_3, we find for $z \in [0,3]$

$$f_3(z) = \frac{8N^6}{(N^2-1)^3}\left(\frac{x^5}{5!}\,\chi_{[0,1]}(z) + \left(\frac{z^5}{40} - \frac{z^4}{8} + \frac{5z^3}{12} - \frac{z^2}{2} - \frac{5z}{4} - \frac{13}{120}\right)\chi_{]1,2]}(z) + \left(-\frac{z^5}{60} + \frac{z^4}{8} - \frac{z^3}{6} - \frac{13}{12}\,z^2 - \frac{17}{12}\,z - \frac{63}{40}\right)\chi_{]2,3]}(z)\right). \tag{21}$$

Formulae (18), (19), and (21) together yield, for $z \in [0,3]$,

$$g_3(z) = \frac{2N^2}{N^2-1}\left[\left(z\varphi(1) + \frac{2N^2}{N^2-1}\frac{z^3}{3!}\varphi(2) + \frac{4N^4}{(N^2-1)^2}\frac{z^5}{5!}\varphi(3)\right)\chi_{[0,1]}(z)\right.$$
$$+ \frac{2N^2}{N^2-1}\left(\left(\frac{z^3}{3} - \frac{3z^2}{2} + 2z - \frac{2}{3}\right)\varphi(2)\right.$$
$$+ \frac{2N^2}{N^2-1}\left(\frac{z^5}{40} - \frac{z^4}{8} + \frac{5z^3}{12} - \frac{z^2}{2} - \frac{5z}{4} - \frac{13}{120}\right)\varphi(3)\Big)\chi_{]1,2]}(z)$$
$$\left. + \frac{4N^4}{(N^2-1)^2}\left(-\frac{z^5}{60} + \frac{z^4}{8} - \frac{z^3}{6} - \frac{13}{12}z^2 - \frac{17}{12}z - \frac{63}{40}\right)\varphi(3)\chi_{]2,3]}(z)\right]. \tag{22}$$

Again it is clear that g_3 increases on $[0,1]$. Let us approximate by putting $\varphi(3) \ll \varphi(2)$ so that the sign of g_3' on $]1,2]$ equals the sign of $z^2-3z+2 < 0$ on $]1,2[$. Furthermore, using $\varphi(3) \ll \varphi(1)$, we have

$$g_3(1) - g_3(1+) \approx \frac{2N^2}{N^2-1}\varphi(1) > 0.$$

So this sequel shows that the second and third approximations of f satisfy

(i) increasing on $[0,1]$,
(ii) decreasing on $[1,2]$, and
(iii) negative jump in 1.

So this section and Corollaries 2.1 and 2.2 give substantial proof of the observation of Rousseau. Furthermore, in the next section we are also able to prove the conjecture of Rousseau in the case of straight author counts.

4. MODELING STRAIGHT AUTHOR COUNTS

In this case, only the first author receives a weight of 1 and, hence, the only possible weights belong to the set $\{0\} \cup \mathbf{N}$, the natural numbers extended with 0. For this reason, we will try to work with discrete distributions, although—in general—taking consecutive convolutions of discrete distributions is very difficult (cf. [13]). We will solve the problem completely and in an exact way (i.e., without any approximations).

Now we must find the discrete distribution

$$p_1(x) = P(\text{weight} = x \text{ in one article}). \tag{23}$$

Now,

$$P(\text{weight} = x \text{ in one article}) = \begin{cases} \dfrac{y-1}{y}, & \text{if there are } y \text{ authors and if } x = 0, \\ \dfrac{1}{y} & \text{if there are } y \text{ authors and if } x = 1. \end{cases}$$

Hence,

$$p_1(x) = \left(\int_1^\infty \frac{y-1}{y}\psi(y)\,dy\right)\chi_{\{0\}}(x) + \left(\int_1^\infty \frac{1}{y}\psi(y)\,dy\right)\chi_{\{1\}}(x), \tag{24}$$

where ψ is the density function of the number of authors per paper. Formula (24) is rewritten as:

$$p_1(x) = a\,\chi_{\{0\}}(x) + b\,\chi_{\{1\}}(x). \tag{25}$$

Let $p_n(x)$ be the probability to have a cumulative weight x over n articles ($n = 1,2,3,\dots$) for straight author counts. Hence $x \in \{0,1,2,\dots,n\}$, necessarily.

PROPOSITION 4.1. *For any function ψ as in the previous section*

$$p_n(x) = C_n^x\, b^x\, a^{n-x}, \tag{26}$$

for every $x = 0, 1, 2, \ldots, n$.

PROOF. We give the proof by complete induction. For $n = 1$, we have $p_1(x) = a$ if $x = 0$, and $p_1(x) = b$ if $x = 1$. This is in accordance with formula (26). Let us now assume (26) to be true for $n \in \mathbf{N}$ and we must prove it for $n + 1$. By [13], we see that

$$p_{n+1}(x) = \sum_{i=0}^{x} p_n(i)\, p_1(x - i), \tag{27}$$

for all $x = 0, 1, \ldots, n + 1$, where $i = 0, 1, \ldots, n$ and $x - i = 0, 1$ (the discrete convolution).

(a) Let $x \neq 0$ and $x \neq n + 1$. Then, in (27), $i = x$ or $i = x - 1$, (since $x - i = 0, 1$). Hence, by (27) and (26), (for n)

$$\begin{aligned} p_{n+1}(x) &= p_n(x)\, p_1(0) + p_n(x-1)\, p_1(1) \\ &= C_n^x\, b^x\, a^{n-x}\, a + C_n^{x-1}\, b^{x-1}\, a^{n-x+1}\, b = (C_n^x + C_n^{x-1})\, b^x\, a^{n-x+1}, \\ p_{n+1}(x) &= C_{n+1}^x\, b^x\, a^{n+1-x}, \end{aligned}$$

i.e., formula (26) for $n + 1$.

(b) Let $x = 0$. Now $i = 0$ is the only possible value. So

$$p_{n+1}(0) = p_n(0)\, p_1(0) = C_n^0\, b^0\, a^n\, a = C_{n+1}^0\, b^0\, a^{n+1},$$

i.e., formula (26) is true for $n + 1$.

(c) Let $x = n + 1$. Now $i = n$ is the only possible value. Hence,

$$p_{n+1}(n+1) = p_n(n)\, p_n(1) = C_n^n\, b^n\, a^0\, b = C_{n+1}^{n+1}\, a^0\, b^{n+1},$$

i.e., formula (26) is true for $n + 1$. ∎

Let $p(x)$ denote the probability to have a weight of x in the general case. By Bayes' rule we have, for all $x = 0, 1, 2, 3, \ldots$,

$$p(x) = \sum_{j=x}^{\infty} C_j^x\, b^x\, a^{j-x}\, \varphi(j), \tag{28}$$

for any function φ as in the previous section. Here, we assume that $C_0^0 =: 0$. In this way, the formula (28) is also correct for $p(0)$.

THEOREM 4.1. *For any function ψ and φ as above,*

$$p(0) < p(1),$$

if $a \leq b$. This is true for all functions ψ, e.g., of the form $\psi(x) = C/x^\alpha$, $\alpha \geq 2$ (and any φ).

PROOF. By (28),

$$\begin{aligned} p(0) &= \sum_{j=1}^{\infty} a^j\, \varphi(j), \\ p(1) &= \sum_{j=1}^{\infty} j\, b\, a^{j-1}\, \varphi(j). \end{aligned}$$

Hence, $p(0) < p(1)$ is satisfied if (sufficient but not necessary condition) $j\,b \geq a$, for all $j \geq 1$ (and at least one "greater or equal" is a "greater"). This is true for $a \leq b$. This condition is equivalent to (see (24)):

$$\int_1^\infty \frac{y-1}{y}\,\psi(y)\,dy \leq \int_1^\infty \frac{1}{y}\,\psi(y)\,dy.$$

For $\psi(x) = C/x^\alpha$, we find the condition $\alpha \geq 2$. ∎

NOTE. Theorem 4.1 is true for all functions ψ such that:

$$\int_1^2 \frac{2-y}{y}\,\psi(y)\,dy \geq \int_2^\infty \frac{y-2}{2}\,\psi(y)\,dy.$$

It is possible to satisfy this inequality in the important case of densities ψ who are initially increasing and then decreasing, if only the maximum is attained for small y. As pointed out by Rousseau, a Poisson-type function ψ is natural in this context (cf. [14]).

The general behavior of p for $x > 1$ is left open in this very general case. However, we have the following result, valid for $\alpha = 2$ and for $\varphi(j) = \frac{6}{\pi^2 j^2}$, $j = 1, 2, \ldots$, (note that $\sum_{j=1}^\infty p(j) = 1$ since $\sum_{j=1}^\infty \frac{1}{j^2} = \frac{\pi^2}{6}$ — see [3]).

THEOREM 4.2. *For $\psi(x) = \frac{1}{x^2}$, $x \in [1, \infty[$, and $\varphi(j) = \frac{6}{\pi^2 j^2}$ $(j = 1, 2, \ldots)$, we have that*

$$p(1) > p(2).$$

PROOF. It follows from (28) that in this case ($a = b = \frac{1}{2}$):

$$p(1) = \frac{6}{\pi^2}\sum_{j=1}^\infty \frac{1}{j\,2^j} = \frac{6}{\pi^2}\ln 2 = 0.42138, \qquad \text{and}$$

$$p(2) = \frac{6}{\pi^2}\sum_{j=2}^\infty \frac{C_j^2}{j^2\,2^j} = \frac{6}{\pi^2}\left(\sum_{j=2}^\infty \frac{1}{2^{j+1}} - \sum_{j=2}^\infty \frac{1}{j\,2^{j+1}}\right);$$

$$p(2) = \frac{6}{\pi^2}\left(\frac{1}{4} - \frac{1}{2}\left(\ln 2 - \frac{1}{4}\right)\right) = 0.01728 \ll p(1).$$ ∎

NOTE.

(1) We can also verify directly here that $p(0) < p(1)$, since

$$p(0) = \frac{6}{\pi^2}\sum_{j=1}^\infty \frac{1}{j^2\,2^j} = \frac{1}{2} - \frac{3\,(\ln 2)^2}{\pi^2} = 0.35396$$

(see [15, p. 9]).

(2) After $x = 2, p$ does not continue to decrease, even in this simple case:

$$p(3) = \frac{6}{\pi^2}\sum_{j=3}^\infty \frac{C_j^3}{2^j\,j^2} = \frac{1}{\pi^2}\left(\sum_{j=3}^\infty \frac{j}{2^j} - 3\sum_{j=3}^\infty \frac{1}{2^j} + 2\sum_{j=3}^\infty \frac{1}{j\,2^j}\right);$$

$$p(3) = 0.03914 > p(2).$$

See Figure 3 for a partial graph of p in this simple case.

(3) After $x = 3$, the behavior of p is unclear, but it is also an unimportant issue. Note in any case the absolute maximum is $x = 1$. The problem of the weight distribution in the case of straight counts is hereby solved to a good extent.

We close the paper by adding a trivial section on total (also called normal) author counts.

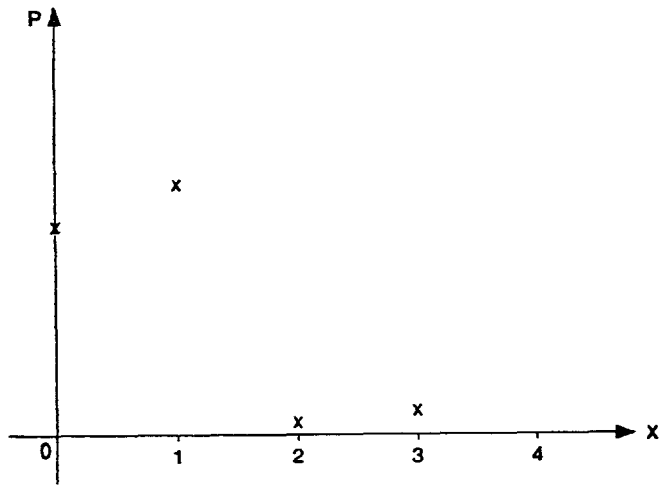

Figure 3. Partial graph of p.

5. MODELING TOTAL AUTHOR COUNTS

For the sake of completeness, we add also the important (but simple) case of total author counts: each author receives a weight of one per paper. In the notation of the previous sections, this means that

$$p_1(x) = 1,$$

if and only if $x = 1$. Since, by (27),

$$p_2(x) = \sum_{i=0}^{x} p_1(i)\, p_1(x-i),$$

we have that

$$p_2(x) = 1,$$

if and only if $x = 2$. More generally,

$$p_n(x) = \chi_{\{n\}}(x), \qquad x = 0, 1, 2, \ldots, n. \tag{29}$$

Hence,

$$p(x) = \sum_{j=x}^{\infty} p_j(x)\, \varphi(j),$$

(where φ is a general distribution as above);

$$p(x) = \varphi(x), \qquad x = 1, 2, 3, \ldots, \tag{30}$$

i.e., the probability of having weight x is equal to $\varphi(x)$, being the probability of an author writing x papers. This case is hence the *only* case that gives a decreasing p but note that x only starts from one here.

6. SUMMARY

In this paper, we studied the distribution of the weights among authors if we count authorship fractionally, in a straight way, or totally. Apart from the trivial case of total author counts, we showed that this distribution increases on $[0, 1]$ and then starts decreasing from one onward. This result has been reached under fairly normal assumptions on the frequency functions φ and ψ (respectively, the densities of papers per author and authors per paper).

A general model, involving continuous convolutions, has been developed in the case of fractional counting and a general model, involving discrete convolutions, has been developed in the case of straight counting.

We leave open the study of the same problem for functions ψ that are initially increasing and then decreasing.

APPENDIX

LEMMA A.

1. Let h be an integrable function on $[a,b]$ and g a differentiable function on $[a,b]$. Then, for

$$f(x) = \int_0^x h(y)\, g(x-y)\, dy,$$

we have, for $x \in [a,b]$,

$$f'(x) = \int_0^x h(y)\, g'(x-y)\, dy + h(x)g(0), \qquad \text{where } g'(x) = \frac{dg}{dx}(x). \tag{A.1}$$

2. Under the same conditions, we also have for

$$f(x) = \int_{x-1}^1 h(y)\, g(x-y)\, dy,$$

that

$$f'(x) = \int_{x-1}^1 h(y)\, g'(x-y)\, dy - h(x-1)\, g(1). \tag{A.2}$$

3. Analogously, for

$$f(x) = \int_0^{x-1} h(y)\, g(x-y)\, dy,$$

we have that

$$f'(x) = \int_0^{x-1} h(y)\, g'(x-y)\, dy + h(x-1)\, g(1). \tag{A.3}$$

PROOF. We only show that (A.1) is valid; the proof of the other formulae is exactly the same.

$$\begin{aligned} f'(x) &= \lim_{p\to 0} \frac{1}{p}\left(\int_0^{x+p} h(y)\, g(x+p-y)\, dy - \int_0^x h(y)\, g(x-y)\, dy\right) \\ &= \lim_{p\to 0} \frac{1}{p}\left(\int_0^{x+p} h(y)\big(g(x+p-y)-g(x-y)\big)dy + \int_0^{x+p} h(y)g(x-y)dy - \int_0^x h(y)g(x-y)dy\right) \\ &= \int_0^x h(y)\, g'(x-y)\, dy + \frac{d}{dz}\left(\int_0^z h(y)\, g(x-y)\, dy\right)(x) \\ &= \int_0^x h(y)\, g'(x-y)\, dy + h(x)\, g(0) \end{aligned}$$

(cf. [3]). ∎

REFERENCES

1. R. Rousseau, Breakdown of the robustness property of Lotka's law: The case of adjusted counts for multi-authorship attribution, *Journal of the American Society for Information Science* **43** (10), 645–647 (1992).
2. A.J. Lotka, The frequency distribution of scientific productivity, *Journal of the Washington Academy of Sciences* **16**, 317–323 (1926).
3. T.M. Apostol, *Mathematical Analysis*, Addison-Wesley, Reading, Massachusetts, (1974).

4. S.C. Bradford, Sources of information on specific subjects, *Engineering* **137**, 85–86 (1934); Reprinted in *Collection Management* **1**, 95–103 (1976–1977); Also reprinted in *Journal of Information Science* **10** (148), (facsimile of the first page) and 176–180 (1985).
5. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York, (1977).
6. G.K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, (1949); Reprinted, Hafner, New York, (1965).
7. G. Herdan, *Type-Token Mathematics, A Textbook of Mathematical Linguistics*, Mouton's Gravenhage, (1960).
8. L. Egghe and R. Rousseau, *Introduction to Informetrics, Quantitative Methods in Library, Documentation and Information Science*, Elsevier, Amsterdam, (1990).
9. L. Egghe, The duality of informetric systems with applications to the empirical laws, Ph.D. Thesis, City University, London, U.K, (1989).
10. L. Egghe, The duality of informetric systems with applications to the empirical laws, *Journal of Information Science* **16** (1), 17–27 (1990).
11. A. Bookstein, Informetric distributions, Part II: Resilience to ambiguity, *Journal of the American Society for Informatin Science* **41** (5), 376–386 (1990).
12. K.L. Chung, *A Course in Probability Theory*, Academic Press, New York, (1974).
13. G. Blom, *Probability and Statistics, Theory and Applications*, Springer-Verlag, Berlin, (1989).
14. R. Rousseau, Fractional counts for authorship atrribution: A numerical study, *Preprint* (1993).
15. I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, New York, (1965).