Information Processing & Management Vol. 29, No. 4, pp. 499–513, 1993
Printed in Great Britain.

0306-4573/93 $6.00 + .00

EVOLUTION OF INFORMATION PRODUCTION PROCESSES AND ITS RELATION TO THE LORENZ DOMINANCE ORDER

LEO EGGHE[1] and RONALD ROUSSEAU[2]
[1]LUC, Universitaire Campus, B-3590, Diepenbeek, Belgium
and UIA, Speciale Licentie Informatie en Bibliotheekwetenschap,
Universiteitsplein 1, B-2610, Wilrijk, Belgium
[2]KIHWV, Zeedijk 101, B-8400 Oostende, Belgium
and UIA, Speciale Licentie Informatie en Bibliotheekwetenschap,
Universiteitsplein 1, B-2610, Wilrijk, Belgium

(*Received* 27 *March* 1992; *accepted in final form* 19 *June* 1992)

Abstract—We investigate the evolution and growth of information production processes (in short IPPs). An important role in this investigation is played by the Lorenz curve of concentration and the majorization poset. This leads to a number of relations between the total number of sources, the total number of items, the average production, and the maximum production of an IPP, when one or several of these quantities show an evolution in time. We also establish a relation between the 80/20 rule and the Lorenz dominance order. We conclude that Lorenz curves and the majorization poset are important tools in the investigation of evolution and growth.

1. INTRODUCTION

Traditionally, informetrics (or bibliometrics) studies bibliographies or publication lists; yet, more generally, most theoretical work in informetrics can be described using the framework of information production processes (in short IPPs), introduced by Egghe (1989) and studied in Egghe (1990a,b). In the discrete case an IPP is defined as follows:

1.1 Definition: discrete IPPs. A discrete IPP is a triple of the form (S, I, i), where S and I are countable sets and where, for every $s \in S$, $i(s) \subset I$. The elements of the set S are called sources, and those of I are called items. The function i assigns to the source s the set of all items produced by s.

1.2 *Two examples*

(a) Lotka (1926) studied a group of researchers (the sources) and the number of publications (the items) they had produced. This gave rise to the well known Lotka distribution:

$$f(y) = \frac{D}{y^{\beta}} \tag{1}$$

where $f(y)$ denotes the number of sources (researchers) producing y items (articles), and D and β are strictly positive parameters. In this particular case the function i assigns to every researcher the set of papers of which he or she is the senior author.

(b) In database design (Casas & Sevcik, 1990) sources can be database blocks, and items are accesses to these blocks. In Casas & Sevcik (1990), it is assumed that block accesses are distributed according to a Bradford distribution, a distribution closely related to Lotka's, which has been the subject of intensive research in the field of informetrics (Egghe & Rousseau, 1990).

For other examples we refer the reader to Egghe (1989,1990a) and Rousseau (1991).

This article will study the evolution of IPPs. In particular, we will investigate the relations between the total number of sources, the total number of items, the average production, and the maximum production. We will show that the Lorenz curve (for a definition we refer the reader to section 2.1) and the Lorenz dominance order (see section 2.2) can play an important role in these investigations.

General IPPs will be characterized by the following parameters. A: the total number of items; T: the total number of sources; m: the maximum production, that is, the number of items in the source containing the largest number of items (when sources are ordered in decreasing order (according to production), this is the production of the source ranked first); and μ: the average number of items per source. When comparing two different IPPs, the respective characteristic numbers will be denoted: $A(i)$, $T(i)$, $m(i)$, and $\mu(i)$; $i = 1,2$. We note the following relations between these characteristic values.

$$\mu = \frac{A}{T} \tag{2}$$

and if there are no empty sources:

$$T \leq A \leq mT. \tag{3}$$

When the total number of sources (T) stays fixed, then

$$\mu(1) < \mu(2) \Leftrightarrow A(1) < A(2). \tag{4}$$

This means that if the number of sources is fixed, average growth is equivalent to growth in the number of items. Further down, less obvious relations will be established; for example, if we have two Lotka distributions without gaps and the Lorenz curve of the first lies strictly under the Lorenz curve of the second, then we were able to show the following inequality (cf. Theorem 3.2), involving the maximum productions ($m(i)$, $i = 1,2$) and the means ($\mu(i)$, $i = 1,2$)):

$$m(1).(2\mu(2) - 1) < m(2).(2\mu(1) - 1).$$

1.3 Definition: Rank-frequency vectors and pure frequency vectors. In our investigations we will use two different kinds of vectors. The first are simple rank-frequency vectors (also called **R**-vectors), denoted as:

$$\mathbf{R} = (z_1, z_2, \ldots, z_T). \tag{5}$$

Here z_i is the production of the source ranked at the ith position; $z_1 = m \geq z_2 \geq \ldots z_T > 0$ (in this article we will always assume that all sources are non-empty). Of course, it often happens that there is a functional relation between the z_i, that is, $z_i = g(i)$, as in Zipf's law, where $g(i) = C/i$ and C is a positive constant (cf. Egghe & Rousseau, 1990, Section 4.1.3.3).

The second kind are pure frequency vectors (**F**-vectors), denoted as:

$$\mathbf{F} = (l_1, l_2, \ldots, l_m). \tag{6}$$

Here l_i denotes the number of sources with production i. Usually $l_m = 1$ (i.e., there is only one source with the maximum production). If all productions between 1 and m actually occur, then **F** is an m-tuple, but this is not necessarily the case. In general, we will consider **F** as a p-tuple, with $1 \leq p \leq m$. We will often assume a functional relation between the l_is: $l_i = f(i)$, as in Lotka's square law, where $f(i) = D/i^2$ and D is a constant (cf. Egghe & Rousseau, 1990; Section 4.1.3.1). Because of the relation with Lotka's law, we have used the symbols l_i for the components of this frequency vector.

Every **F**-vector can be interpreted in two ways. The first way is to consider this vector a p-tuple as such. The second way is to interpret it as a different representation of an **R**-vector, as above. When $\mathbf{F} = (l_1, l_2, \ldots, l_m)$, then the associated **R**-vector, denoted as $R(F)$, is:

$$(m, \ldots, \underbrace{2, \ldots, 2}_{l_2 \text{ times}}, \underbrace{1, \ldots, 1}_{l_1 \text{ times}}),$$

where we have assumed that there is only one source producing the maximum production m. Of course, if necessary, we can as well convert an **R**-vector into an **F**-vector.

We note that, for **R(F)** and **F**:

$$T = \sum_{i=1}^{m} l_i = \sum_{i=1}^{m} f(i) \tag{7}$$

and

$$\mu = \sum_{j=1}^{m} jf(j)/T. \tag{8}$$

When referring to the average, μ, of an **F**-vector, it will always be in the second interpretation (8).

1.4 Definition. We will say that two IPPs with frequency vectors $\mathbf{F}_1 = (f_1(1),\ldots, f_1(m(1)))$ and $\mathbf{F}_2 = (f_2(1),\ldots,f_2(m(2)))$ are t-related ('t' from 'by truncation') if there exists a function f, defined on the natural numbers and numbers K_1 and K_2 such that

$$f_1(n) = K_1 f(n), \quad \text{for every } n = 1,\ldots,m(1) \tag{9}$$

and

$$f_2(n) = K_2 f(n), \quad \text{for every } n = 1,\ldots,m(2). \tag{10}$$

In the case that $\sum_{n=1}^{\infty} f(n) = 1$ and when $m(i)$ is large, K_i is an approximation for $T(i)$, $i = 1,2$.

2. THE LORENZ CURVE IN CONCENTRATION THEORY

2.1 *Construction of a discrete Lorenz curve (cf. Rousseau, in press)*

In 1905 Lorenz presented a method to visualize inequality by using a curve, which since then is known as the Lorenz curve. In the discrete case, this curve is constructed as follows. Let $X = (x_1, x_2,\ldots,x_N)$ be a general vector, $x_1 \geq \ldots \geq x_N$; then we put for $i = 1,\ldots,N$,

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j} \tag{11}$$

$$= \frac{x_i}{\mu N}. \tag{12}$$

In this article the notation a_i will always be used for relative frequencies. In the applications X will be a rank-frequency vector **R** (or **R(F)**), or a frequency vector **F**.

The discrete Lorenz curve is then the polygonal curve joining the origin (0,0) to the points with coordinates

$$\left(\frac{i}{N}, \sum_{j=1}^{i} a_j\right), \quad i = 1,\ldots,N. \tag{13}$$

Note that

$$\sum_{j=1}^{N} a_j = 1, \tag{14}$$

so that the Lorenz curve ends in the point (1,1).

The Lorenz curve of an N-tuple X is also the graph of a function. This function will be denoted as $\mathcal{L}_X(t)$, $0 \le t \le 1$. As the construction of the Lorenz curve uses relative numbers, all **X**-vectors that differ only by a multiplicative constant will have the same Lorenz curve.

The equation of the function $\mathcal{L}_x(t)$ is given as:

$$\text{for } \frac{i}{N} \le t \le \frac{i+1}{N}, \quad i = 0, \ldots, N-1$$

$$\mathcal{L}_X(t) = \sum_{k=1}^{i+1} a_k + N a_{i+1}\left(j - \frac{i+1}{N}\right). \tag{15}$$

Obviously, a Lorenz curve is an increasing curve. As the x_i, hence also the a_i, are placed in decreasing order, the Lorenz curve is concave. We note here that, especially in econometrical investigations, it is customary to place the x_i in increasing order. In that case the Lorenz curve is convex.

2.2 *The Lorenz dominance order and the majorization poset*

When $\mathbf{X}_1$ and $\mathbf{X}_2$ are two vectors, we say that $\mathbf{X}_1$ is dominated by $\mathbf{X}_2$, denoted as $\mathbf{X}_1 \prec \mathbf{X}_2$ if for all $t \in [0,1]$, $\mathcal{L}_{\mathbf{X}_1}(t) \le \mathcal{L}_{\mathbf{X}_2}(t)$. Lorenz dominance is only a partial order. Two distributions are not comparable whenever their Lorenz curves intersect. The Lorenz dominance order $\prec$ is sometimes also referred to as the majorization poset (Hardy *et al.* 1952; Marshall & Olkin, 1979). These notions will be applied to **R**-vectors as well as to **F**-vectors.

Finally, we note that when two vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ have the same Lorenz curve, we will say that they are equivalent, denoted as $\mathbf{X} \equiv \mathbf{Y}$ (cf. Rousseau, 1992a). Here we note the following (trivial) result.

2.3 *Proposition*

If two vectors $\mathbf{F}_1$ and $\mathbf{F}_2$ are t-related and $m(1) = m(2)$, then $\mathbf{F}_1 \equiv \mathbf{F}_2$.

3. EVOLUTION OF IPPs

In this section we begin our investigations on the evolution of IPPs as shown by the influence that the change of parameters has on the Lorenz curve and on the dominance order.

3.1 *Proposition*

Let $\mathbf{F} = (a_1, \ldots, a_N)$ be an N-tuple, ranked in decreasing order, and let $\mathbf{F}'$ be the Nk-tuple REPEAT$_k$(**F**) constructed by replacing every a_i by k times a_i/k; hence $\mathbf{F} \equiv \mathbf{F}' =$ REPEAT$_k$(**F**). When $\mu = \sum i a_i$ then $\mu' = k\mu - (k-1)/2$.

Proof. By definition the mean μ' is equal to

$$\begin{aligned}
&(1 + \ldots + k)a_1/k \\
&\quad + ((k+1) + \ldots + 2k)a_2/k \\
&\quad + \ldots \\
&\quad + (((i-1)k+1) + \ldots + ik)a_i/k \\
&\quad + \ldots \\
&\quad + ((N(k-1)+1) + \ldots + Nk)a_N/k.
\end{aligned}$$

$$= \frac{k(k+1)}{2} a_1/k$$

$$+ \left(\frac{2k(2k+1)}{2} - \frac{k(k+1)}{2}\right) a_2/k$$

$$+ \dots$$

$$+ \left(\frac{ik(ik+1)}{2} - \frac{(i-1)k[(i-1)k+1]}{2}\right) a_i/k$$

$$+ \dots$$

$$+ \left(\frac{kN(kN+1)}{2} - \frac{k(N-1)[k(N-1)+1]}{2}\right) a_N/k$$

$$= \sum_{i=1}^{N} \frac{2ik-k+1}{2} a_i \tag{16}$$

$$= k\mu - \frac{k-1}{2}. \tag{17}$$

□

THEOREM 3.2

Let f_i, $i = 1,2$, be decreasing functions, defined in all points k, $k = 1, \dots, m(i)$. The numerical values $f_i(k)$ are interpreted as the number of sources with k items. Considering now the **F**-*vectors*

$$\mathbf{F}_1 = (f_1(1), f_1(2), \dots, f_1(m(1)))$$

and

$$\mathbf{F}_2 = (f_2(1), f_2(2), \dots, f_2(m(2)))$$

then $\mathbf{F}_1 \prec \mathbf{F}_2$ *implies*

$$m(1)(2\mu(2) - 1) \leq m(2)(2\mu(1) - 1) \tag{18}$$

where, as stated before, the averages $\mu(i)$ refer to vectors $\mathbf{R}(\mathbf{F}_i)$. Moreover, equality in (18) *only occurs if* $\mathbf{F}_1 \equiv \mathbf{F}_2$.

Proof. If $\mathbf{F}_1 \prec \mathbf{F}_2$, and if $\mathbf{F}_1$ and $\mathbf{F}_2$ are not equivalent, the Gini index of $\mathbf{F}_1$, denoted as $G(\mathbf{F}_1)$ is strictly smaller than the Gini index of $\mathbf{F}_2$, denoted as $G(\mathbf{F}_2)$. If all productions occur, as assumed, then, by Rousseau (1992a, II.39)

$$G(\mathbf{F}_i) = \frac{m(i) + 1 - 2\mu(i)}{m(i)}, \quad i = 1,2 \tag{19}$$

Consequently, $\mathbf{F}_1 \prec \mathbf{F}_2$ implies:

$$\frac{m(1) + 1 - 2\mu(1)}{m(1)} \leq \frac{m(2) + 1 - 2\mu(2)}{m(2)} \tag{20}$$

with equality only in the case of equivalence. After some calculations, this becomes:

$$2m(1)\mu(2) + m(2) \leq 2m(2)\mu(1) + m(1) \tag{21}$$

or

$$m(1).(2\mu(2) - 1) \leq m(2).(2\mu(1) - 1). \tag{22}$$

□

This result leads to a number of important special cases.

3.3 Corollaries

(a) *Under the assumptions of the previous theorem, and if $m(1) = m(2)$, then $\mu(2) \leq \mu(1)$ with equality only if* $\mathbf{F}_1$ *and* $\mathbf{F}_2$ *are equivalent.*

Indeed, this result follows immediately from (18).

(b) *Under the same assumptions and if $m(1) > m(2)$, then $\mu(1) > \mu(2)$.*

This result too is a simple consequence of (18).

(c) *Under the same assumptions, and if $\mu(1) = \mu(2)$ then $m(1) \leq m(2)$ with equality only if* $\mathbf{F}_1 \equiv \mathbf{F}_2$.

(d) *Finally, under the same assumptions and if, moreover, $\mu(2) > \mu(1)$, this implies $m(1) < m(2)$.*

3.4 *Note*

We will show that a number of plausible converses of the results of 3.3 are actually false.

Indeed, take for instance $\mathbf{F}_1 = (6,3,2,2,2)$ and $\mathbf{F}_2 = (5,4,3,2,1)$. Then $m(1) = m(2) = 5$, $\mu(1) = 36/15 = 2.4$ and $\mu(2) = 35/15 = 2.333$. Yet, $\mathbf{F}_1 \equiv (5/15, 4/15, 3/15, 2/15, 1/15)$ and $\mathbf{F}_2 \equiv (6/15, 3/15, 2/15, 2/15, 2/15)$. From this we conclude that $\mathbf{F}_1$ and $\mathbf{F}_2$ are not comparable in the Lorenz dominance order. This shows that the converse of 3.3a is false.

The converse of 3.3b is false too. To see this, consider $\mathbf{F}_1 = (5,3,3,2,1,1)$ and $\mathbf{F}_2 = (5,4,3,2,1)$ as above. Then $m(1) = 6 > m(2) = 5$; $\mu(1) = 2.6 > \mu(2) = 2.333$. Yet, $\mathbf{F}_1$ and $\mathbf{F}_2$ are incomparable for the Lorenz dominance order.

Lemma 3.5

When two IPPs are t-related then

$$m(1) < m(2) \Leftrightarrow \mu(1) < \mu(2) \tag{23}$$

and

$$m(1) = m(2) \Leftrightarrow \mu(1) = \mu(2). \tag{24}$$

Proof. The inequality $\mu_1 < \mu_2$ is equivalent to the inequality

$$\sum_{i=1}^{m(1)} \sum_{j=1}^{m(2)} if(i)f(j) < \sum_{i=1}^{m(1)} \sum_{j=1}^{m(2)} jf(i)f(j). \tag{25}$$

Now, $m(1) < m(2)$ implies (25), while $m(1) \geq m(2)$ implies (25), but with the inequality sign $<$ replaced by $\geq$. This proves that $\mu(1) < \mu(2)$ is equivalent to $m(1) < m(2)$. Similarly, $m(1) = m(2)$ is equivalent to $\mu(1) = \mu(2)$. □

The following example shows that the requirement to be *t*-related cannot be omitted from Lemma 3.5.

Example 3.6. We choose $m(1) = 3$. Let further be $c_i = f_1(i)/T(1)$, then we choose

$$c_1 = 0.4$$

$$c_2 = 0.35$$

$$c_3 = 0.25.$$

Then

$$\mu(1) = \sum_{k=1}^{3} kc_k = 1.85.$$

Next, we choose $m(2) = 4$ and $d_i = f_2(i)/T(2)$, then we choose

$$d_1 = 0.7$$

$$d_2 = 0.1$$

$$d_3 = 0.1$$

$$d_4 = 0.1.$$

Then

$$\mu(2) = \sum_{k=1}^{4} kd_k = 1.6 < 1.85 = \mu(1),$$

although $m(2) > m(1)$.

THEOREM 3.7

If $\mathbf{R}(\mathbf{F}_1) \prec \mathbf{R}(\mathbf{F}_2)$ *and both IPPs have sources with production one then*

(a) $\mu(1) \le \mu(2)$
(b) $m(1) \le m(2)$
(c) $m(1)\mu(2) \le m(2)\mu(1)$ (*this result does not depend on the assumption of having sources with production 1*).

Proof. When $\mathbf{R}(\mathbf{F}_1) \prec \mathbf{R}(\mathbf{F}_2)$ then the slope of the first segment of the Lorenz curve of $\mathbf{R}(\mathbf{F}_1)$ is smaller than the slope of the first segment of $\mathbf{R}(\mathbf{F}_2)$. This yields:

$$\frac{\dfrac{f_1(m(1)).m(1)}{A(1)}}{\dfrac{f_1(m(1))}{T(1)}} \le \frac{\dfrac{f_2(m(2)).m(2)}{A(2)}}{\dfrac{f_2(m(2))}{T(2)}} \tag{26}$$

or

$$\frac{m(1)T(1)}{A(1)} \le \frac{m(2)T(2)}{A(2)} \tag{27}$$

or

$$\frac{m(1)}{\mu(1)} \le \frac{m(2)}{\mu(2)}. \tag{28}$$

Similarly, the slope of the last segment of the Lorenz curve of $\mathbf{R}(\mathbf{F}_1)$ is larger than the slope of the Lorenz curve of $\mathbf{R}(\mathbf{F}_2)$. This yields:

$$\frac{\dfrac{f_1(1)}{A(1)}}{\dfrac{f_1(1)}{T(1)}} \ge \frac{\dfrac{f_2(1)}{A(2)}}{\dfrac{f_2(1)}{T(2)}} \tag{29}$$

$$\Rightarrow \mu(1) \le \mu(2). \tag{30}$$

Combining (28) and (30) yields $m(1) \le m(2)$. □

COROLLARY 3.8

If $\mathbf{R}(\mathbf{F}_1) \prec \mathbf{R}(\mathbf{F}_2)$, *both IPPs have sources with production 1 and* $m(1) = m(2)$ *then* $\mu(1) = \mu(2)$.

3.9 *Note*

The result of the previous theorem can be slightly generalized when using rank-frequency vectors. The point is that we have assumed that there are always sources with production 1. If this is not the case, the general notation for rank-frequency vectors (5) is somewhat easier to work with. Absence of sources with production 1 does not influence (28), but (29) now becomes:

$$\frac{\dfrac{z_{T(1)}}{A(1)}}{\dfrac{1}{T(1)}} \ge \frac{\dfrac{z_{T(2)}}{A(2)}}{\dfrac{1}{T(2)}}. \tag{31}$$

When $z_{T(1)} = z_{T(2)}$, (31) = (29), showing that the assumptions that the lowest production is 1 is not necessary, we only have to assume that they are equal. On the other hand, when, for example, $z_{T(1)} = 1$ and $z_{T(2)} = 2$, we find

$$2\mu(1) \le \mu(2)$$

and hence also $2m(2) \le m(1)$.

In general, eqn (31) yields $z_{T(2)}\mu(1) \le z_{T(1)}\mu(2)$ and hence:

$$z_{T(2)}m(1) \le z_{T(1)}m(2). \tag{32}$$

Note also that this approach clearly shows that, for this result, we do not make the assumption that all productions between 1 and m occur.

3.10 *A negative result*

One could expect that when two IPPs are t-related, $\mu(1) < \mu(2)$ and $m(1) < m(2)$, or even $m(1)\mu(2) < m(2)\mu(1)$ (cf. 3.7), then $\mathbf{F}_1 \prec \mathbf{F}_2$. That this is not true is shown by the following counterexample.

Take $\mathbf{F}_1 = (4,4,1)$ and $\mathbf{F}_2 = (4,4,1,1)$. Then $\mu(1) = 1.6667$ and $m(1) = 3$; $\mu(2) = 1.9$ and $m(2) = 4$ (hence $m(1)\mu(2) = 5.7 < m(2)\mu(1) = 6.6667$). Now

$$\mathbf{F}_1 \equiv (4,4,4,4,4,4,4,4,1,1,1,1) \equiv \left(\tfrac{4}{36}, \tfrac{4}{36}, \ldots, \tfrac{4}{36}, \tfrac{1}{36}, \tfrac{1}{36}, \tfrac{1}{36}, \tfrac{1}{36}\right)$$

and

$$\mathbf{F}_2 \equiv (4,4,4,4,4,4,1,1,1,1,1,1) \equiv \left(\tfrac{4}{30}, \ldots, \tfrac{4}{30}, \tfrac{1}{30}, \ldots \tfrac{1}{30}\right).$$

The ordinates of the vertices of the Lorenz curves are:

$$\text{for } \mathbf{F}_1\text{:} \quad \left(\tfrac{20}{180}, \tfrac{40}{180}, \ldots, \tfrac{140}{180}, \tfrac{160}{180}, \tfrac{165}{180}, \tfrac{170}{180}, \tfrac{175}{180}, \tfrac{180}{180}\right)$$

and

$$\text{for } \mathbf{F}_2\text{:} \quad \left(\tfrac{24}{180}, \tfrac{48}{180}, \ldots, \tfrac{144}{180}, \tfrac{150}{180}, \tfrac{156}{180}, \tfrac{162}{180}, \tfrac{168}{180}, \tfrac{174}{180}, \tfrac{180}{180}\right).$$

This clearly shows that in the first part the Lorenz curve of $\mathbf{F}_2$ is situated above the Lorenz curve of $\mathbf{F}_1$, but that at the end the situation is reversed. Hence, both Lorenz curves cross and $\mathbf{F}_1$ and $\mathbf{F}_2$ are incomparable in the Lorenz dominance order, although 3.7 a, b, and c are satisfied. The reason for this phenomenon is that μ is a numerical value associated with numbers, as such, whereas the Lorenz dominance order and the Lorenz curve depend only on relative numbers.

The next theorem shows a relation between $T(i)$s, the total number of sources, and $A(i)$s, the total number of items. Note that we already know that for IPPs that are t-related, growth in the maximum production is equivalent to growth in mean (Lemma 3.5).

THEOREM 3.11

If two IPPs are t-related with the same constants (*i.e., the K_is in eqns* (9) *and* (10) *coincide*), *then*

$$T(1) < T(2) \Leftrightarrow A(1) < A(2) \Leftrightarrow m(1) < m(2).$$

Proof.

$$T(1) < T(1)$$

$$\Leftrightarrow$$

$$K \sum_{j=1}^{m(1)} f(j) < K \sum_{j=1}^{m(2)} f(j)$$

$$\Leftrightarrow$$

$$m(1) < m(2)$$

$$\Leftrightarrow$$

$$K \sum_{j=1}^{m(1)} jf(j) < K \sum_{j=2}^{m(2)} jf(j)$$

$$\Leftrightarrow$$

$$A(1) < A(2)$$ □

3.12 *Note*

When IPPs are t-related but with different constants, then it is possible that $T(1) < T(2)$, $A(1) > A(2)$, $m(1) > m(2)$, and $\mu(1) > \mu(2)$.

An example: Take $\mathbf{F}_1 = (1000, 900, 190)$, and $\mathbf{F}_2 = (1110, 999)$. Then $T(1) = 2090 < T(2) = 2109$, yet $m(1) = 3 > m(2) = 2$, $A(1) = 3370 > A(2) = 3108$, and $\mu(1) = 1.612 > \mu(2) = 1.473$; yet both IPPs are t-related.

From the requirements '$\mathbf{F}_1$ and $\mathbf{F}_2$ are t-related' and $T(1) < T(2)$, we can conclude nothing concerning the Lorenz order between $\mathbf{R}(\mathbf{F}_1)$ and $\mathbf{R}(\mathbf{F}_2)$. It is possible that $\mathbf{R}(\mathbf{F}_1) \prec \mathbf{R}(\mathbf{F}_2)$, that $\mathbf{R}(\mathbf{F}_1) \succ \mathbf{R}(\mathbf{F}_2)$, or that both are incomparable. Examples are:

- Take $\mathbf{F}_1 = (2,1)$ and $\mathbf{F}_2 = (4)$. These two vectors are t-related and $T(1) = 3 < T(2) = 4$. As $\mathbf{R}(\mathbf{F}_1) = (2,1,1)$ and $\mathbf{R}(\mathbf{F}_2) = (1,1,1,1)$, we see that $\mathbf{R}(\mathbf{F}_2) \prec \mathbf{R}(\mathbf{F}_1)$.
- Take $\mathbf{F}_1 = (2,1)$ and $\mathbf{F}_2 = (2,1,1)$. Also these frequency vectors are t-related and $T(1) = 3 < T(2) = 4$.

As $\mathbf{R}(\mathbf{F}_1) = (2,1,1) \equiv (2,2,2,2,1,1,1,1,1,1,1,1) \equiv (\frac{42}{336}, \frac{42}{336}, \frac{42}{336}, \frac{42}{336}, \frac{21}{336}, \frac{21}{336}, \ldots, \frac{21}{336})$

and

$\mathbf{R}(\mathbf{F}_2) = (3,2,1,1) \equiv (3,3,3,2,2,2,1,1,1,1,1,1) \equiv (\frac{48}{336}, \frac{48}{336}, \frac{48}{336}, \frac{32}{336}, \frac{32}{336}, \frac{32}{336}, \frac{16}{336}, \ldots, \frac{16}{336})$.

Here: $\mathbf{R}(\mathbf{F}_1) \prec \mathbf{R}(\mathbf{F}_2)$.

- In example 3.10, $\mathbf{F}_1$ and $\mathbf{F}_2$ are t-related, $T(1) = 9 < T(2) = 10$, and both vectors are incomparable in the Lorenz dominance order.

4. A RELATION WITH THE 80/20 RULE

In its classical formulation, the 80/20 rule states that 20% of the sources (the most productive ones) are responsible for 80% of the total number of items. In general, one can formulate a $100\,\Theta/100\,x$-rule, $0 < x$, $\Theta \leq 1$ (see, e.g., Egghe, 1986).

4.1 *Assertion: Generalized 80/20 rule*

If two IPPs $\mathbf{F}_1$ and $\mathbf{F}_2$ are t-related and $\mu(1) \leq \mu(2)$, then $x(1) \geq x(2)$, for every $\Theta \in [0,1[$, i.e., $\mathbf{R}(\mathbf{F}_1) \prec \mathbf{R}(\mathbf{F}_2)$. Furthermore, $\mu(1) < \mu(2)$ implies $x(1) > x(2)$.

This assertion has been proved (Egghe, 1991) for a number of distributions, including the geometric distribution and Lotka's distribution $f(y) = D/y^{\beta}$, with $\beta = 1$, 1.5, 2, or 3.

4.2 *Note*

The requirement to consider t-related functions cannot be dropped from the assumptions of 4.1. Indeed, consider, for example, $\mathbf{F}_1 = (5,4,3,2,1)$ and $\mathbf{F}_2 = (6,3,2,2,2)$, as in 3.4. Then $\mathbf{R}(\mathbf{F}_1) = (5,4,4,3,3,3,2,2,2,2,1,1,1,1,1)$ and $\mathbf{R}(\mathbf{F}_2) = (5,5,4,4,3,3,2,2,2,1,1,1,1,1,1)$. Then $\mu(1) = \frac{35}{15} = 2.333$ and $\mu(2) = \frac{36}{15} = 2.4$. Yet, $\mathbf{R}(\mathbf{F}_1)$ and $\mathbf{R}(\mathbf{F}_2)$ are incomparable for the Lorenz order.

4.3 Corollary

If we have two IPPs that are t-related, where the sources with lowest production have the same production, and such that the generalized 80/20 rule is valid, then

$$\mathbf{R}(\mathbf{F}_1) \prec \mathbf{R}(\mathbf{F}_2) \Leftrightarrow \mu(1) \leq \mu(2). \tag{33}$$

Proof. This follows readily from Theorem 3.7, Note 3.9, and 4.1. □

5. GROWTH: ADDING ONE SOURCE

This section investigates what happens when adding one source. The main result is that only in the case where the production of this source equals the average production, the original situation dominates the new one.

5.1 *Constructions*

Let $\mathbf{R} = (r_1, r_2, \ldots, r_T)$ and let $S = \mathrm{ADD}_I(\mathbf{R}) = (I, r_1, \ldots, r_T)$, where this new vector is not yet ordered. We assume that the new source, with production I is ranked at the kth place. Then

$$\begin{aligned} s_i &= r_i && \text{when } 1 \leq i < k \text{ (if such an } i \text{ exists)} \\ s_i &= I && \text{when } i = k \\ s_i &= r_{i-1} && \text{when } k < i \leq T + 1 \text{ (if such an } i \text{ exists).} \end{aligned} \tag{34}$$

Let us agree that in case of equality, the new source is placed at the highest possible rank. Of course we have the following equality:

$$\sum_{j=1}^{T+1} s_j = \left(\sum_{j=1}^{T} r_j\right) + I. \tag{35}$$

We assume from now on that the components of S are placed in decreasing order and put

$$a_i = \frac{r_i}{\sum_{j=1}^{T} r_j} \quad i = 1, \ldots, T;$$

and

$$b_i = \frac{s_i}{\sum_{j=1}^{T+1} s_j}, \quad i = 1, \ldots, T+1. \tag{36}$$

We further put

$$A_i = \sum_{j=1}^{i} a_j, \quad \text{and} \quad B_i = \sum_{j=1}^{i} b_j; \quad A_0 = B_0 = 0. \tag{37}$$

We note that $\sum r_j = T\mu_R$ and, finally we put

$$J = \frac{\sum r_j}{\sum r_j + I} = \frac{T\mu_R}{T\mu_R + I}. \tag{38}$$

The symbols a_i, A_i, b_i, and B_i are related as follows:

$$b_i = \frac{r_i}{\sum r_j + I} = a_i J \quad \text{when } 1 \le i < k \text{ (if such an } i \text{ exists)}$$

$$b_i = \frac{I}{\sum r_j + I} = \frac{IJ}{T\mu_R} \quad \text{when } i = k$$

$$b_i = \frac{r_{i-1}}{\sum r_j + I} = a_{i-1} J \quad \text{when } k < i \le T+1 \text{ (if such an } i \text{ exists);} \tag{39}$$

and further:

$$B_i = A_i J \quad \text{when } 1 \le i < k$$

$$B_i = A_{i-1} J + \frac{IJ}{T\mu_R} \quad \text{when } k \le i \le T+1. \tag{40}$$

Now, two remarks are in order. First, the abscissa of the vertices of the Lorenz curve of S are equal to $i/(T+1)$. These numbers fall between the abscissa of the vertices of the Lorenz curve of **R**, that is:

$$\forall i, i = 1, \ldots, T: \frac{i-1}{T+1} \le \frac{i-1}{T} \le \frac{i}{T+1} < \frac{i}{T} \le \frac{i+1}{T+1} \tag{41}$$

with equality on the left hand side only if $i = 1$, and on the right hand side only if $i = T$.

Second, to express that the Lorenz curve $\mathcal{L}_Y$ of a general vector **Y** is situated above the Lorenz curve $\mathcal{L}_X$ of a general vector **X**, it suffices to require this inequality in the vertices, both of $\mathcal{L}_X$ and of $\mathcal{L}_Y$. This yields the following requirements:

Theorem 5.2

R $\prec S = \mathrm{ADD}_I(\mathbf{R})$, *that is*: *the Lorenz curve of* **R** *is situated under the Lorenz curve of S*

$$\Leftrightarrow$$

$$\forall i = 1, \ldots, T-1 \quad A_i \le B_i + \frac{i}{T}\, b_{i+1} \tag{42}$$

$$\& \ \forall i = 0, \ldots, T-1 \quad A_{i+1} - a_{i+1}\, \frac{i+1}{T+1} \le B_{i+1}. \tag{43}$$

For $S \prec R$ the inequality signs in (42) and (43) must be reversed.

To obtain inequalities (42) and (43) we have written the equations of the line segments connecting $(i/(T+1), B_i)$ to $((i+1)/(T+1), B_{i+1})$ and similarly for the segments connecting $(i/T, A_i)$ and $((i+1)/(T+1), A_{i+1})$. Next we have expressed that $(i/T, A_i)$ is situated under the first segment, and similarly that $((i+1)/(T+1), B_{i+1})$ is situated above the second (see Fig. 1).

For the case $S = \mathrm{ADD}_I(\mathbf{R})$ we will formulate simpler conditions to satisfy the requirements $\mathbf{R} \prec S$ or $S \prec \mathbf{R}$. We will first give necessary conditions and we will later investigate whether these conditions are also sufficient. If $\mathbf{R} \prec S$ the slope of the first segment of the Lorenz curve of **R** must be smaller than the slope of the first segment of S. Similarly, the slope of the last segment of the Lorenz curve of **R** must be larger than the slope of the last segment of the Lorenz curve of S. For the relation $S \prec \mathbf{R}$ these conditions must be reversed.

5.3 $S \prec \mathbf{R}$: *A first necessary condition*

The requirement for the first segment leads to:

$$Ta_1 \ge (T+1)b_1 \tag{44}$$

$$\Leftrightarrow Ta_1 \ge (T+1)a_1 \frac{T\mu_R}{T\mu_R + I} \quad \text{if } k > 1$$

$$\Leftrightarrow I \ge \mu_R \tag{45}$$

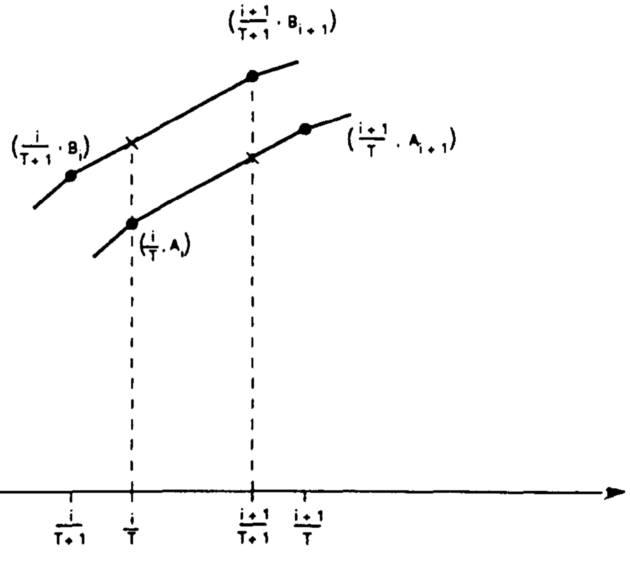

Fig. 1. The relation between the Lorenz curves $\mathcal{L}_R$ and $\mathcal{L}_S$.

and if $k = 1$:

$$Ta_1 \geq (T+1)\frac{I}{T\mu_R + I} \tag{46}$$

$$\Leftrightarrow I \geq \frac{T^2 a_1 \mu_R}{T+1-Ta_1}. \tag{47}$$

We finally note that when $k = 1$, the relation $I > r_1 = a_1 \mu_R T \geq \mu_R$ holds automatically.

5.4 $S \prec \mathbf{R}$: *A second necessary condition*

The condition on the last segment leads to:

$$Ta_T \leq (T+1)b_{T+1} \tag{48}$$

$$\Leftrightarrow Ta_T \leq (T+1)a_T \frac{\mu_R T}{\mu_R T + I} \quad \text{if } k < T+1 \tag{49}$$

$$\Leftrightarrow a_T I \leq a_T \mu_R. \tag{50}$$

As we assume that all cells are non-empty, this is equivalent to:

$$I \leq \mu_R. \tag{51}$$

If $k = T+1$ we find:

$$Ta_T \leq (T+1)\frac{I}{T\mu_R + I} \tag{52}$$

$$\Leftrightarrow I \geq \frac{T^2 a_T \mu_R}{T+1-Ta_T}. \tag{53}$$

Note that $k = T+1$ implies that $I \leq r_T \leq \mu_R$.

5.5 *Analysis of a necessary condition to ensure that* $S \prec \mathbf{R}$

From 5.3 and 5.4 we see that $S \prec \mathbf{R}$ if the following conditions are satisfied:

$$\left[(I \geq \mu_R \;\&\; I \leq r_1) \vee \left(I \leq \frac{T^2 a_1 \mu_R}{T+1-Ta_1} \;\&\; I > r_1\right)\right]$$

&

$$\left[(I \leq \mu_R \;\&\; I > r_T) \vee \left(I \geq \frac{T^2 a_T \mu_R}{T+1-Ta_T} \;\&\; I \leq r_T\right)\right]$$

$\Leftrightarrow$

$$I \geq \mu_R \;\&\; I \leq r_1 \;\&\; I \leq \mu_R \;\&\; I > r_T \tag{54}$$

(this means: $I = \mu_R$ and not the equality situation)

$\vee$

$$I \geq \mu_R \;\&\; I \leq r_1 \;\&\; I \geq \frac{T^2 a_T \mu_R}{T+1-Ta_T} \;\&\; I \leq x_T \tag{55}$$

(this means: $I = \mu_R$ in the equality situation).

Hence we conclude that $I = \mu_R$ is a necessary condition for $S \prec \mathbf{R}$. We will now show the main result of this section, namely, that this condition is also sufficient.

THEOREM 5.6

$$S = \mathrm{ADD}_I(\mathbf{R}) \prec \mathbf{R}$$

$$\Leftrightarrow$$

$$I = \mu_R$$

Proof. We already know that $S \prec X$ implies $I = \mu_R$. We only have to show the opposite implication $I = \mu_R \Rightarrow S \prec \mathbf{R}$. To do this, we have to consider four cases:

(a) $\forall i < k : A_i \geq B_i + (i/T)\, b_{i+1}$ (cf. (42))
(b) $\forall i < k : A_i - a_i\, (i/T + 1) \geq B_i$ (cf. (43))
(c) $\forall i \geq k : A_i \geq B_i + (i/T)\, b_{i+1}$ (cf. (42))
(d) $\forall i \geq k : A_i - a_i\, (i/T + 1) \geq B_i$ (cf. (43)).

Proof of (a). Let $i < k - 1$, then we have to show that

$$A_i \geq \frac{A_i \mu_R T}{\mu_R T + I} + \frac{i}{T}\, a_{i+1} \frac{\mu_R T}{\mu_R T + I}$$

$$\Leftrightarrow A_i\left(1 - \frac{\mu_R T}{\mu_R T + I}\right) \geq \frac{i a_{i+1} \mu_R}{\mu_R T + I}$$

$$\Leftrightarrow A_i \geq i a_{i+1}.$$

This inequality is always satisfied because the a_i are placed in decreasing order.

If $i = k - 1$, then we have to show that

$$A_i \geq \frac{A_i \mu_R T}{\mu_R T + \mu_R} + \frac{i \mu_R}{T(\mu_R T + \mu_R)}$$

$$\Leftrightarrow A_i \geq \frac{A_i T}{1 + T} + \frac{i}{T(T + 1)}$$

$$\Leftrightarrow A_i\left(1 - \frac{T}{1 + T}\right) \geq \frac{i}{T(T + 1)}$$

$$\Leftrightarrow A_i \geq \frac{i}{T}.$$

This inequality is also satisfied because the a_i are placed in decreasing order. □

Proof of (b). We have to show that

$$A_i - a_i \frac{i}{T + 1} \geq A_i \frac{T}{T + 1}$$

$$\Leftrightarrow$$

$$A_i\left(1 - \frac{T}{T + 1}\right) \geq a_i \frac{i}{T + 1}$$

$$\Leftrightarrow$$

$$A_i \geq i a_i,$$

which is true. □

Proof of (c). We have to show that

$$A_i \geq \frac{A_{i-1}T}{1+T} + \frac{T}{T(T+1)} + \frac{i}{T} a_i \frac{T}{1+T}$$

$$\Leftrightarrow A_i \geq \frac{TA_{i-1} + 1 + ia_i}{(T+1)}$$

$$\Leftrightarrow A_iT + A_i \geq TA_i - Ta_i + 1 - ia_i$$

$$\Leftrightarrow 1 - A_i \leq (T - i)a_i;$$

also this inequality is satisfied because the a_i are placed in decreasing order. □

Proof of (d). We have to show that

$$A_i - \frac{ia_i}{T+1} \geq \frac{TA_{i-1} + 1}{T+1}$$

$$\Leftrightarrow (T+1)(A_{i-1} + a_i) - ia_i \geq TA_{i-1} + 1$$

$$\Leftrightarrow Ta_i + A_{i-1} + a_i - ia_i \geq 1$$

$$\Leftrightarrow 1 - A_{i-1} \geq a_i(T + 1 - i),$$

which is also true. □

In Rousseau (1992a) we have also investigated when the curves $\mathcal{L}_R$ and $\mathcal{L}_S$ cross (hence **R** and S are incomparable) and when $R \prec S = \text{ADD}_T(\mathbf{R})$. For details we refer the reader to Rousseau (1992a).

6. CONCLUSION

We have investigated the relations between the total number of sources, the total number of items, the average production, and the maximum production in an IPP, when one or several of these quantities vary. It is shown that the Lorenz dominance order and the Lorenz curve can play an important role in these investigations. Hence, we propose that, in informetric studies, more use should be made of these tools.

REFERENCES

Casas, I.R., & Sevcik, K.C. (1990). A buffer management model for use in predicting overall database system performance. (Preprint 1990, Toronto.)

Egghe, L. (1986). On the 80/20 rule. *Scientometrics, 10,* 55-68.

Egghe, L. (1989). The duality of informetric systems with applications to the empirical laws. Ph.D. Thesis, The City University, London, U.K.

Egghe, L. (1990a). The duality of informetric systems with applications to the empirical laws. *Journal of Information Science, 16,* 17-27.

Egghe, L. (1990b). New Bradfordian laws equivalent with old Lotka laws, evolving from a source-item duality argument. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90, Proceedings of the 2nd International Conference on Bibliometrics, Scientometrics and Informetrics* (pp. 79-96). Amsterdam: Elsevier.

Egghe, L. (1991). Exact probabilistic and mathematical proofs of the relation between the mean μ and the generalized 80/20 rule. (Preprint.)

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics.* Amsterdam: Elsevier.

Hardy, H., Littlewood, J.E., & Polya, G. (1952). *Inequalities.* Cambridge: Cambridge University Press.

Lorenz, M.O. (1905). Methods of measuring concentration of wealth. *Journal of the American Statistical Association, 9,* 209-219.

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

Marshall, A.W., & Olkin, I. (1979). *Inequalities: Theory of majorization and its applications.* New York: Academic Press.

Rousseau, R. (1991). Information production processes as a general framework in informetric studies. In G. Kempen & P. de Vroomen (Eds.), *Informatiewetenschap 1991* (pp. 193-201). Nijmegen: Stinfon.

Rousseau, R. (1992a). Concentration and diversity in informetric research. Doctoral dissertation, Antwerp. U.I.A.

Rousseau, R. (in press). De Gini-index en de Lorenzkromme (The Gini index and the Lorenz curve; in Dutch). *Wiskunde en Onderwijs.*