Information Processing & Management Vol. 28, No. 1, pp. 19–33, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + .00

THE GINI INDEX AND THE LEIMKUHLER CURVE FOR BIBLIOMETRIC PROCESSES

QUENTIN L. BURRELL
Statistical Laboratory, Dept. of Mathematics, University of Manchester,
Oxford Road, Manchester M13 9PL, U.K.

Abstract—It has recently been emphasized that the Leimkuhler curve and the Gini index are valuable in giving respectively graphical and numerical summaries of the concentration of bibliometric distributions. In this paper these tools are further investigated from a probabilistic viewpoint. In particular, the importance of the time parameter and the special nature of the "nonproducers" in bibliometric studies are highlighted.

Keywords: Concentration measures; Leimkuhler curve; Gini index; Pareto distribution; Gamma-Poisson process; Generalized Waring process.

1. INTRODUCTION

Carpenter (1979) was the first to point out that the measure of class concentration proposed by Pratt (1979) is essentially the same as the Gini index already well known in the field of econometrics and introduced by Gini (1909, 1912) over 60 years previously. This and many other possible measures of concentration were included in the survey presented by Ravichandra Rao (1988) at the *First International Conference on Bibliometrics* in 1987. Subsequently Egghe and Rousseau (1990, in press), Bonckaert and Egghe (1991), Egghe (1991), and Rousseau (1991) have investigated various "desirable properties" of such measures and have found that the Gini index performs notably well.

In all of these studies, however, the concentration measures have been applied to bibliometric distributions and data sets collected over a fixed period of time. By contrast, in a series of papers Burrell (1980,1987,1988,1990a,1991a) has argued in favour of bibliometric processes that develop in time, and hence that bibliometric models should explicitly incorporate a time parameter. In this paper we consider the consequent time-dependent form of the Gini index and point out the crucial role of the nonproducers.

2. THE GINI INDEX AND THE LEIMKUHLER CURVE

2.1 *Terminology, notation, and definition*

We shall adopt the "source-item" terminology familiar in bibliometrics and informetrics. Thus we have a population of sources (e.g., journals) producing items (e.g., articles on a particular subject) in some random fashion over time. For the moment we assume that the period of time is fixed and we are interested in the probability distribution of the random variable X, the number of items produced by a source during the period, and more particularly in the concentration of this distribution. Note that although, by its nature, X is non-negative and integer-valued, it is convenient to include continuous random variables in the discussion. Hence for the rest of this section we shall speak in terms of a nonspecific random variable Y. The following definition is adapted from Stuart and Ord (1987).

Definition. For a random variable Y with finite mean $\mu_Y = E[Y]$, the *Gini index* or *coefficient of concentration* is denoted γ_Y and defined by

$$\gamma_Y = \frac{E[|Y_1 - Y_2|]}{2E[Y]} \tag{1}$$

$$= \frac{\iint |y_1 - y_2| \, dF_Y(y_1) \, dF_Y(y_2)}{2 \int y \, dF_Y(y)} \tag{2}$$

where Y_1, Y_2 are independent copies of Y, and F_Y is the cumulative distribution function of Y.

Note that the Gini index is independent of the scale of measurement (i.e., $\gamma_{cY} = \gamma_Y$ for any positive constant c), but has the disadvantage that it is dependent on choice of origin (i.e., $\gamma_{Y+c} \neq \gamma_Y$ for a non-zero constant c). This dependence on origin will cause no real concern in what follows, since in the contexts of interest there is a natural origin at zero.

The numerator in (1) is called the (coefficient of) mean difference and its calculation (e.g., using the numerator in (2)) may not be straightforward even when the probability distribution of Y is known. A useful alternative is provided by the following:

Theorem I

For a non-negative random variable Y,

(i) *if Y is integer-valued then*

$$\gamma_Y = \frac{\sum_{j \geq 0} \mathrm{F}_Y(j)\Phi_Y(j+1)}{\mu_Y} = 1 - \frac{\sum_{j \geq 1} \Phi_Y(j)^2}{\mu_Y}. \tag{3}$$

(ii) *If Y is continuous then*

$$\gamma_Y = \frac{\int_0^\infty \mathrm{F}_Y(y)\Phi_Y(y) \, dy}{\mu_Y} = 1 - \frac{\int_0^\infty \Phi_Y(y)^2 \, dy}{\mu_Y} \tag{4}$$

where $\Phi_Y(x) = P(Y \geq x)$ is the tail distribution function of Y.

Proof. See Appendix.

Note. In the case of an empirical data set in which the distinct observed values are $x_0, x_1, \ldots, x_m$ with corresponding frequencies $f(0), f(1), \ldots, f(m)$, if we put

$$N = \sum f(j) = \text{total number of “sources”}$$

and

$$M = \sum x_j f(j) = \text{total number of “items”}$$

then

$$P(Y = x_j) = \frac{f(j)}{N}, \qquad \mu_Y = \frac{M}{N}$$

and the mean difference is $(1/N^2)\sum_{i,j} |x_i - x_j| f(i) f(j)$. Hence the empirical form of the Gini index is

$$\gamma_Y = \frac{\sum_{i,j} |x_i - x_j| f(i) f(j)}{2MN}.$$

(Note that in the expression for the empirical mean difference $1/N^2$ is sometimes replaced by $1/N(N-1)$, giving the so-called “mean difference without repetition” (see Stuart & Ord, 1987, p. 47) and a modified Gini index of $[N/(N-1)]\gamma_Y$.)

In most practical situations the possible observed values are non-negative integers, so we may take $x_j = j$, $j = 0, 1, \ldots, m$ and Theorem I gives

$$\gamma_Y = 1 - \frac{\sum_{j=1}^{m} r(j)^2}{MN}$$

where

$$\begin{aligned} r(j) &= f(j) + f(j+1) + \ldots + f(m) \\ &= \text{number of sources with productivity } \geq j \end{aligned}$$

(see Burrell, 1991b).

2.2 *The Leimkuhler curve*

For the non-negative random variable Y with finite mean μ_Y we define its *tail moment function* Ψ_Y by

(i) if Y is integer-valued then

$$\Psi_Y(j) = \frac{1}{\mu_Y} \sum_{k \geq j} kP(Y = k), \quad j = 0,1,2,\ldots.$$

(Note that $\Psi_Y(0) = \Psi_Y(1) = 1$.)

(ii) if Y is continuous with probability density function f_Y then

$$\Psi_Y(x) = \frac{1}{\mu_Y} \int_x^\infty y f_Y(y)\, dy, \quad x \geq 0.$$

Then a plot of Ψ_Y as ordinate against Φ_Y as abscissa gives the so-called *Leimkuhler curve*, a variant of the Lorenz curve of concentration (see Burrell, 1991b). The Leimkuhler curve passes through the origin and (1,1) and is concave to the Φ_Y axis. Its direct connection with the Gini index is given graphically by

$$\begin{aligned} \gamma_Y &= 2 \text{ (area beneath Leimkuhler curve)} - 1 \\ &= 2 \int_0^1 \Psi_Y \, d\Phi_Y - 1 \qquad (5) \end{aligned}$$

which gives an alternative method of calculation. This form also makes it clear that $0 \leq \gamma_Y \leq 1$.

2.3 *Some examples*

2.3.1 *The Pareto distribution.* The Pareto distribution with index $\alpha > 0$ is specified by the probability density function

$$f_Y(x) = \frac{\alpha}{x^{1+\alpha}}, \quad x > 1,$$

giving the tail distribution function as

$$\Phi_Y(x) = \begin{cases} 1 & \text{if } x \leq 1, \\ \dfrac{1}{x^\alpha} & \text{if } x > 1. \end{cases} \qquad (6)$$

The mean is finite only if $\alpha > 1$, in which case $\mu_Y = \alpha/(\alpha - 1)$ and

$$\begin{aligned}\Psi_Y(x) &= (\alpha - 1)\int_x^\infty \frac{1}{y^\alpha}\,dy \\ &= \frac{1}{x^{\alpha-1}}.\end{aligned}$$

Using (4) we find

$$\begin{aligned}\gamma_Y &= 1 - \frac{\alpha - 1}{\alpha}\int_0^\infty \Phi_Y(y)^2\,dy \\ &= 1 - \frac{\alpha - 1}{\alpha}\left[1 + \int_1^\infty \frac{1}{y^{2\alpha}}\,dy\right] \quad \text{from (6)} \\ &= 1 - \frac{\alpha - 1}{\alpha}\left[1 + \frac{1}{2\alpha - 1}\right] \\ &= \frac{1}{2\alpha - 1}.\end{aligned}$$

Alternatively, note that the Leimkuhler curve is given by

$$\Psi_Y = (\Phi_Y)^{(\alpha-1)/\alpha}, \quad 0 \le \Phi_Y \le 1,$$

so using (5) we have

$$\begin{aligned}\gamma_Y &= 2\int_0^1 x^{(\alpha-1)/\alpha}\,dx - 1 \\ &= 2\left[\frac{1}{\dfrac{\alpha - 1}{\alpha} + 1}\right] - 1 \\ &= \frac{2\alpha}{2\alpha - 1} - 1 \\ &= \frac{1}{2\alpha - 1}\end{aligned}$$

as before.

2.3.2 *The Bradford distribution.* If the Pareto distribution is arbitrarily truncated at some value $1 + \beta > 1$, then the resulting density function has finite mean for all $\alpha > 0$. The particular case $\alpha = 1$ corresponds to

$$f_Y(x) = \frac{1 + \beta}{\beta x^2}, \quad 1 < x < 1 + \beta \tag{7}$$

which is a truncated continuous version of Lotka's inverse-square distribution. It is shown (e.g., by Burrell, 1991b) that

$$\gamma_Y = 1 - 2[(\ln(1 + \beta))^{-1} - \beta^{-1}]$$

while the Leimkuhler curve is given by

$$\Psi_Y = \frac{\ln(1 + \beta\Phi_Y)}{\ln(1 + \beta)}, \quad 0 \le \Phi_Y \le 1. \tag{8}$$

(8) above, rather than (7), is termed the Bradford distribution by Leimkuhler (1967). See also Burrell (1990b,1991b). For other cases with $\alpha \neq 1$, see Egghe (1991).

2.3.3 *The exponential distribution.* Here we have a one-parameter family of distributions with probability density function

$$f_Y(x) = \lambda e^{-\lambda x}, \quad x > 0,$$

where $\lambda > 0$. As λ is a scale parameter, we can make use of the scale invariance of the Gini index and without loss of generality take $\lambda = 1$. Then $\mu_Y = 1$ and

$$\Phi_Y(x) = e^{-x}, \quad x > 0;$$

so from (4)

$$\begin{aligned}\gamma_Y &= 1 - \int_0^\infty e^{-2x}\,dx \\ &= \tfrac{1}{2}.\end{aligned}$$

Note that the tail moment function is given by

$$\begin{aligned}\Psi_Y(x) &= \int_x^\infty y e^{-y}\,dy \\ &= e^{-x}(1 + x);\end{aligned}$$

so that the equation of the Leimkuhler curve is

$$\Psi_Y = \Phi_Y[1 - \ln \Phi_Y], \quad 0 \leq \Phi_Y \leq 1. \tag{9}$$

2.3.4 *The geometric distribution.* For this discrete distribution, Y has probability mass function

$$P(Y = j) = pq^j, \quad j = 0,1,2,\ldots$$

where $0 < p < 1$ and $q = 1 - p$. Note that $\mu_Y = q/p$ and

$$\Phi_Y(j) = q^j, \quad j = 0,1,2,\ldots. \tag{10}$$

According to (3), then, the Gini index is given by

$$\begin{aligned}\gamma_Y &= 1 - \frac{p}{q}\sum_{j\geq 1} q^{2j} \\ &= 1 - \frac{p}{q}\cdot\frac{q^2}{1-q^2} \\ &= 1 - \frac{q}{1+q} \\ &= \frac{1}{1+q}.\end{aligned} \tag{11}$$

For the Leimkuhler curve, note first that

$$\begin{aligned}\mu_Y \Psi_Y(j) &= \sum_{k\geq j} kpq^k \\ &= pq \sum_{k\geq j} kq^{k-1} \\ &= pq \sum_{k\geq j} \frac{d}{dq} q^k \\ &= pq \frac{d}{dq}\left(\sum_{k\geq j} q^k\right) \\ &= pq \frac{d}{dq}\left(\frac{q^j}{1-q}\right) \\ &= pq\cdot q^{j-1} \frac{(pj+q)}{(1-q)^2} \\ &= \frac{q^j}{p}(pj+q).\end{aligned}$$

Thus

$$\Psi_Y(j) = q^j\left(1 + \frac{p}{q} j\right), \quad j = 0,1,2,\ldots$$

and hence from (10) the points $(\Phi_Y(j), \Psi_Y(j))$ lie on the curve

$$\Psi_Y = \Phi_Y\left[1 + \frac{p}{q \ln q} \ln \Phi_Y\right] \quad 0 \leq \Phi_Y \leq \Phi_Y(1) = q. \tag{12}$$

Remark. The reader might note that the delightfully simple formula for the Gini index for the geometric distribution given by (11) bears no relation to that given by Egghe (1987). The basic reason for this is that we have given the index for the distribution of "*items over sources*," whereas Egghe calculates it for the distribution of "*sources over productivities.*" While Egghe's viewpoint may be valid in certain situations, it is not really consistent with the one of interest here. (We would similarly dispute the relevance of Egghe's calculations for the truncated Lotka inverse-square distribution.)

3. BIBLIOMETRIC PROCESSES

3.1 *Standard models*

When we start to consider the production of items by sources as a system evolving over time, we denote by X_t the number of items produced by a source during $[0, t]$ and then $\{X_t; t \geq 0\}$ is a *stochastic counting process* (i.e., X_t is non-negative integer-valued and is non-decreasing with t). Let us write

$$\begin{aligned}p_t(j) &= P(X_t = j), \quad j = 0,1,2,\ldots \\ \Phi_t(j) &= P(X_t \geq j) \\ &= \sum_{k\geq j} p_t(j).\end{aligned}$$

If the mean exists (and this will not necessarily be the case in the models to be considered), write

$$\begin{aligned}\mu_t &= E[X_t] \\ &= \sum_{k\geq 1} kp_t(k)\end{aligned}$$

and

$$\Psi_t(j) = \frac{\sum_{k\geq j} kp_t(k)}{\mu_t}$$

as well as γ_t for the Gini index.

A number of models have been proposed for such processes, based in the main on either stochastic birth and death processes (e.g., Schubert & Glänzel, 1984, Glänzel & Schubert, 1991) or mixtures of simple counting processes (e.g., Burrell, 1980,1987,1988, 1990a,1991a, Sichel, 1985). We mention in particular:

3.1.1 *The gamma-Poisson process (GPP).* This counting process arises as a gamma mixture of Poisson processes (see, e.g., Burrell, 1987,1988,1990a, for further details) and

$$p_t(j) = \binom{j+\nu-1}{j}\left(\frac{1}{1+\beta t}\right)^{\nu}\left(\frac{\beta t}{1+\beta t}\right)^{j}, \quad j = 0,1,2,\ldots.$$

Hence X_t has a negative binomial distribution of index $\nu > 0$ and parameter $(1+\beta t)^{-1}$. Here β is a time-scale parameter. For this process

$$\mu_t = \nu\beta t.$$

In general, $\Phi_t(j)$ and $\Psi_t(j)$ can both be written in terms of incomplete beta functions, but there is no simple closed expression for Ψ_t as a function of Φ_t except in the special case where $\nu = 1$.

Special case: $\nu = 1$, $\beta = 1$. Here

$$p_t(j) = \left(\frac{1}{1+t}\right)\left(\frac{t}{1+t}\right)^{j}, \quad j = 0,1,2,\ldots$$

so that X_t has a geometric distribution with parameter $p = (1+t)^{-1}$. Hence we can make direct use of (11) to write the Gini index as

$$\gamma_t = \frac{1+t}{1+2t}. \tag{13}$$

Notice in particular that γ_t is strictly decreasing with t and that $\lim_{t\to 0}\gamma_t = 1$, $\lim_{t\to\infty} = \frac{1}{2}$. Turning to the equation of the Leimkuhler curve we find from (12) that

$$\Psi_t = \Phi_t\left[1 + \frac{1}{t\ln\left(\frac{t}{1+t}\right)}\ln\Phi_t\right], \quad 0 \leq \Phi_t \leq \Phi_t(1) = \frac{t}{1+t}. \tag{14}$$

3.1.2 *The generalized Waring process (GWP).* This arises (see e.g., Burrell, 1988, 1991a) as a mixture of negative binomial processes and leads to

$$p_t(j) = \frac{\Gamma(\alpha+\nu t)}{B(\alpha,\beta)\Gamma(\nu t)}\cdot\frac{\Gamma(\nu t+j)\Gamma(\beta+j)}{\Gamma(\nu t+\alpha+\beta+j)j!}, \quad j = 0,1,2,\ldots.$$

Here ν is the index of the underlying negative binomial process, α and β are the parameters of the mixing beta distribution, and $\Gamma(\cdot), B(\cdot,\cdot)$ are the gamma and beta functions, respectively. The parameter α governs the behaviour of the tail of the distribution. Indeed

$$p_t(j) \sim \frac{c(t)}{j^{\alpha+1}}$$

so that the GWP can be thought of as a process whose tails are asymptotically of Lotka form. In particular, only moments of order $< \alpha$ exist and

$$\mu_t = \frac{\nu\beta t}{\alpha - 1}\text{, provided } \alpha > 1. \tag{15}$$

Again there seem to be no manageable expressions for Φ_t and Ψ_t although the following provides an interesting example.

Special case: $\nu = 1$, $\alpha = 2$, $\beta = 1$. With these choices the probability mass function simplifies considerably to give

$$\begin{aligned} p_t(j) &= \frac{\Gamma(t+2)}{B(2,1)\Gamma(t)} \cdot \frac{\Gamma(t+j)\Gamma(j+1)}{\Gamma(t+j+3)j!} \\ &= \frac{2t(t+1)}{(t+j)(t+j+1)(t+j+2)}, \quad j = 0,1,2\ldots \\ &= t(t+1)\left[\frac{1}{t+j} - \frac{2}{t+j+1} + \frac{1}{t+j+2}\right]. \end{aligned}$$

Note that although from (15) $\mu_t = t < \infty$, the variance is infinite so measures of concentration derived from variance are not appropriate in this case. Turning to the tail distribution function we have

$$\begin{aligned} \Phi_t(j) &= \sum_{k=j}^{\infty} p_t(k) \\ &= t(t+1) \sum_{k=j}^{\infty} \left[\frac{1}{t+k} - \frac{2}{t+k+1} + \frac{1}{t+k+2}\right] \\ &= t(t+1)\left[\frac{1}{t+j} - \frac{1}{t+j+1}\right] \end{aligned}$$

and hence

$$\begin{aligned} \sum_{j=1}^{\infty} \Phi_t(j)^2 &= [t(t+1)]^2 \sum_{j=1}^{\infty} \left[\frac{1}{(t+j)^2} - \frac{2}{(t+j)(t+j+1)} + \frac{1}{(t+j+1)^2}\right] \\ &= [t(t+1)]^2 \left[\frac{1}{(t+1)^2} + 2\sum_{j=2}^{\infty} \frac{1}{(t+j)^2} - \frac{2}{t+1}\right] \\ &= [t(t+1)]^2 \left[2\sum_{j=1}^{\infty} \frac{1}{(t+j)^2} - \frac{2t+3}{(t+1)^2}\right]. \end{aligned}$$

Now substituting into (3) we find for the Gini index

$$\begin{aligned} \gamma_t &= 1 - t(t+1)^2 \left[2\sum_{j=1}^{\infty} \frac{1}{(t+j)^2} - \frac{2t+3}{(t+1)^2}\right] \\ &= (t+1)(2t+1) - 2t(t+1)^2 \sum_{j=1}^{\infty} \frac{1}{(t+j)^2}. \end{aligned} \tag{16}$$

3.2 *The "nonproducers"*

One of the major problems, as well as a distinctive feature, of bibliometric processes is that the nonproducers are in general not observed. For instance, how many (potentially contributing) journals did not publish a paper on bibliometrics in 1990? How many scientists did not have any paper published in 1990? In such examples the underlying population of "potentially productive sources" is ill defined and so the *number* of nonproducers in any period cannot be known with any degree of precision. Even in situations where the population is supposedly well defined there can be genuine uncertainty over this "zero class." For instance, in library circulation models it is often unclear whether a particular book that has not circulated just happens not to have been borrowed or is not borrowable through having been lost or stolen. We may thus be obliged, or prefer, to work only with the actually productive sources. We write:

$$X_t^* = \text{number of items produced by a productive source during } [0, t]$$

$$\begin{aligned} p_t^*(j) &= P(X_t^* = j) \\ &= P(X_t = j \mid X_t \neq 0) \\ &= \frac{p_t(j)}{1 - p_t(0)}, \quad j = 1,2,\ldots. \end{aligned} \tag{17}$$

Thus the observed process $\{X_t^*; t \geq 0\}$ is the zero-truncated version of the X_t process. (If $X_t = 0$ then X_t^* is not defined.) From (17) it is immediate that for the observed process we have

$$\mu_t^* = E[X_t^*] = \frac{\mu_t}{1 - p_t(0)} \quad \text{(if the mean exists)}$$

$$\Phi_t^*(j) = P(X_t^* \geq j) = \frac{\Phi_t(j)}{1 - p_t(0)}, \quad j = 1,2,\ldots \tag{18}$$

while

$$\begin{aligned} \Psi_t^*(j) &= \frac{\sum_{k \geq j} k p_t^*(k)}{\mu_t^*} \\ &= \frac{\sum_{k \geq j} k p_t(k)}{\mu_t} = \Psi_t(j), \quad j = 1,2,\ldots. \end{aligned} \tag{19}$$

It is clear from (18) and (19) that the Leimkuhler curve for the X_t^* process sits under that for the X_t process and hence, from the graphical derivation of the Gini index in (5), that $\gamma_t^* < \gamma_t$. More exactly we have:

Theorem II

(i)

$$\gamma_t^* = \frac{\gamma_t - p_t(0)}{1 - p_t(0)} \tag{20}$$

(ii)

$$\gamma_t^* < \gamma_t \quad \text{provided } p_t(0) > 0$$

(iii)

$$\lim_{t\to\infty} \gamma_t^* = \lim_{t\to\infty} \gamma_t \quad \text{if } \lim_{t\to\infty} p_t(0) = 0.$$

Proof. (i)

$$\gamma_t^* = 1 - \frac{\sum_{j\ge 1} \Phi_t^*(j)^2}{\mu_t^*}$$

$$= 1 - \frac{\sum_{j\ge 1} \Phi_t(j)^2}{(1 - p_t(0))\mu_t} \quad \text{from (18)}$$

$$= 1 - \frac{1 - \gamma_t}{1 - p_t(0)} = \frac{\gamma_t - p_t(0)}{1 - p_t(0)}.$$

(ii) and (iii) are immediate corollaries. □

Note that the insistence that $p_t(0) > 0$ merely says that by (finite) time t there are still some (potentially) productive sources that have not by then actually produced an item. In general, $p_t(0)$ will be decreasing with t; the requirement that the limit is zero says that there are no "never-producers," so that the population does indeed comprise "*potential* producers."

For purposes of illustration, let us return to the examples of 3.1.

Example 1: The GPP with $\nu = 1$, $\beta = 1$. In this case

$$p_t^*(j) = \left(\frac{1}{1+t}\right)\left(\frac{t}{1+t}\right)^{j-1}, \quad j = 1,2,3,\ldots.$$

Using Theorem II(i) and (13), the Gini index is

$$\gamma_t^* = \frac{\gamma_t - \dfrac{1}{1+t}}{1 - \dfrac{1}{1+t}} = \frac{\dfrac{(1+t)^2}{1+2t} - 1}{t}$$

$$= \frac{t}{1+2t}.$$

Hence γ_t^* is strictly increasing with t; indeed note that $\gamma_t + \gamma_t^* = 1$ in this example.

For the Leimkuhler curve we can make use of (18) and (19) together with the eqn (14) already derived for the non-truncated version to write

$$\Psi_t^* = \Phi_t^*\left[1 + \frac{1}{(1+t)\ln\left(\dfrac{t}{1+t}\right)} \ln \Phi_t^*\right]$$

Compare this with (14). See Fig. 1 for a graphical presentation.

A simple application of l'Hôpital's rule shows that

$$\lim_{t\to\infty} \frac{1}{t \ln\left(\dfrac{t}{1+t}\right)} = \lim_{t\to\infty} \frac{1}{(1+t)\ln\left(\dfrac{t}{1+t}\right)} = -1$$

so that the limiting form of the Leimkuhler curve in both cases is

$$\Psi = \Phi[1 - \ln \Phi], \quad 1 \le \Phi \le 1,$$

which is the Leimkuhler curve of the exponential distribution (9). (See Fig. 1.)

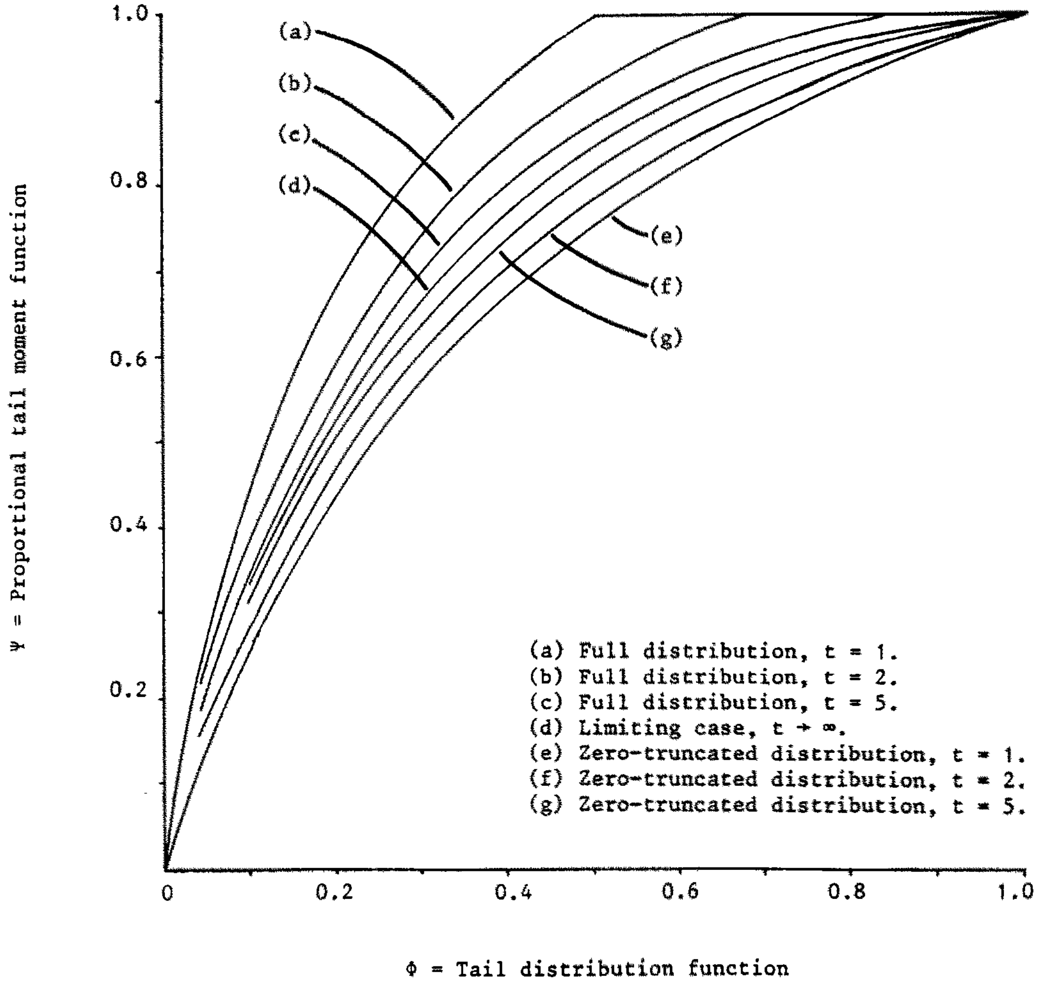

Fig. 1. Leimkuhler curves for the full and zero-truncated forms of the gamma-Poisson process with $\nu = 1, \beta = 1$.

Example 2: The GWP with $\nu = 1$, $\alpha = 2$, $\beta = 1$. Note that $p_t(0) = 2/(t+2)$ so using (20) we find

$$\gamma_t^* = \frac{\gamma_t - p_t(0)}{1 - p_t(0)}$$

$$= \frac{1}{t}\,[(t+2)\gamma_t - 2]\,.$$

Computation of γ_t, and hence of γ_t^*, is fairly straightforward if we just consider integer time points $n, n = 1,2,\ldots$, since then

$$\gamma_n = (n+1)(2n+1) - 2n(n+1)^2 \sum_{j=n+1}^{\infty} \frac{1}{j^2}$$

$$= (n+1)(2n+1) - 2n(n+1)^2\left[\frac{\pi^2}{6} - \sum_{j=1}^{n} \frac{1}{j^2}\right].$$

Calculated values are given in Table 1. Note that, as in Example 1, the Gini index decreases with time for the full distribution, but increases with time for the zero-truncated version.

Table 1. Gini index for the full and the zero-truncated generalized Waring process with $\nu = 1$, $\alpha = 2$, $\beta = 1$

Time	Gini index	
t	γ_t	γ_t^*
1	0.8405	0.5216
2	0.7824	0.5647
3	0.7530	0.5883
4	0.7354	0.6031
5	0.7237	0.6132
6	0.7155	0.6206
7	0.7092	0.6262
8	0.7044	0.6305
9	0.7006	0.6341
10	0.6975	0.6370
11	0.6949	0.6394
12	0.6927	0.6414
13	0.6908	0.6432
14	0.6892	0.6448
15	0.6877	0.6461

4. CONCLUDING REMARKS

The main aim of this paper has been to extend the use of some established bibliometric techniques and to view them from a probabilistic standpoint. A consequence is that certain deficiencies in current practice have been brought out. Thus the paper raises several questions which may be the basis of useful projects for future research. To the author, some of the important questions are:

- Much interesting mathematical research concentrates on convenient continuous models of productivity (e.g., the Pareto and Bradford distributions of section 2). Where does time figure in such models? Again, the nonproducers are an important feature in both theoretical and applied work. How are these accommodated in continuous models?
- The work we have mentioned on general aspects of concentration measures is undoubtedly important. However, some of the "desirable properties" proposed require the nonproducers to be identifiable. How can this sort of axiomatic approach be modified if only the (non-zero) producers can be observed? Is there any way for time to be incorporated?
- The particular theoretical examples considered in section 3.2 suggest that the Gini index for the "full" population process *decreases* with time, while that for the observed process *increases* (both approaching the same limit). How general is this result?
- Bibliometrics is an applied subject and the author has consistently sought to stress the importance of the time parameter. Empirical studies recognizing this aspect are unfortunately few in number. Burrell (1991a) reports one such study for which inspection of the Leimkuhler curves supports the preceding suggestion that the Gini index for the producers increases with time. Are there any counterexamples? The increasing availability of bibliographic databases should allow many (comparative?) studies. Can we look forward to increasing reportage of these? (We would suggest that meaningful development of bibliometrics will not occur in their absence!)

REFERENCES

Bonckaert, P., & Egghe, L. (1991). Rational normalization of concentration measures. *JASIS* (accepted for publication).

Burrell, Q.L. (1980). A simple stochastic model for library loans. *Journal of Documentation*, *36*, 115–132.

Burrell, Q.L. (1987). A third note on ageing in a library circulation model; applications to future use and relegation. *Journal of Documentation*, *43*, 24–45.

Burrell, Q.L. (1988). Predictive aspects of some bibliometric processes. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 43–63). Amsterdam: Elsevier.

Burrell, Q.L. (1990a). Using the gamma-Poisson model to predict library circulations. *JASIS*, *41*, 164–170.

Burrell, Q.L. (1990b). *The non-equivalence of the Bradford, Leimkuhler and Lotka 'laws'.* Research Report, Statistical Laboratory, University of Manchester.

Burrell, Q.L. (1991a). *The dynamic nature of bibliometric processes: A case study.* To be presented at the *"Third International Conference on Informetrics."* Unpublished manuscript. Indian Statistical Institute, Bangalore, India.

Burrell, Q.L. (1991b). The Bradford distribution and the Gini index. *Scientometrics*, *21*, 117–130.

Carpenter, M.P. (1979). Similarity of Pratt's measure of class concentration to the Gini index. *JASIS*, *30*, 108–110.

Egghe, L. (1987). Pratt's measure for some bibliometric distributions and its relation with the 80/20 rule. *JASIS*, *38*, 288–297.

Egghe, L. (1991). Duality aspects of the Gini index for general information production processes. Preprint.

Egghe, L., & Rousseau, R. (1990). Elements of concentration theory. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90* (pp. 97–137). Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (1991). Transfer principles and a classification of concentration measures. *JASIS* (in press).

Gini, C. (1909). Il diverso accrescimento delle classi sociali e la concentrazione della ricchezzi. *Giornale degli Economisti*, *37*.

Gini, C. (1912). Variabilità e mutabilità, contributo allo studio delle distribuzioni e relazions statistiche. *Studi Economico-Giuridici dell'Univ. di Cagliari*, *3*, 1–158.

Glänzel, W., & Schubert, A. (1991). *Predictive aspects of a stochastic model for citation processes.* To be presented at the *"Third International Conference on Informetrics."* Unpublished manuscript. Indian Statistical Institute, Bangalore, India.

Leimkuhler, F.F. (1967). The Bradford distribution. *Journal of Documentation*, *23*, 197–207.

Pratt, A.D. (1979). A measure of class concentration in bibliometrics. *JASIS*, *28*, 285–292.

Ravichandra Rao, I.K. (1988). Probability distributions and inequality measures for analyses of circulation data. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 231–248). Amsterdam: Elsevier.

Rousseau, R. (1991). *Pielou's axiom and N-dependent concentration measures.* To be presented at the *"Third International Conference on Informetrics,"* Unpublished manuscript. Indian Statistical Institute, Bangalore, India.

Schubert, A., & Glänzel, W. (1984). A dynamic look at a class a skew distributions. A model with scientometric applications. *Scientometrics*, *6*, 149–167.

Sichel, H.S. (1985). A bibliometric distribution which really works. *JASIS*, *36*, 314–321.

Stuart, A., & Ord, J.K. (1987) *Kendall's advanced theory of statistics. Volume 1: Distribution theory* (5th Edition). London: Griffin.

APPENDIX

Proof of Theorem I. (i) For the discrete case, this is essentially the same as the Theorem in Burrell (1991b). Write $p_j = P(Y = j)$, $j = 0,1,2,\ldots$, let Y_1, Y_2 be independent copies of Y, and consider calculation of $E[|Y_1 - Y_2|]$. Note first that

$$P(|Y_1 - Y_2| = j) = \begin{cases} P(Y_1 = Y_2), & \text{if } j = 0, \\ 2P(Y_1 - Y_2 = j), & \text{if } j = 1,2,\ldots. \end{cases}$$

Thus

$$P(|Y_1 - Y_2| = 0) = \sum_{k=0}^{\infty} P(Y_1 = Y_2 = k) = \sum_{k=0}^{\infty} p_k^2$$

while if $j \neq 0$,

$$\{Y_1 - Y_2 = j\} = \bigcup_{k=j}^{\infty} \{Y_1 = k\} \cap \{Y_2 = k - j\}$$

so that

$$P(|Y_1 - Y_2| = j) = 2\sum_{k=j}^{\infty} p_k p_{k-j}, \quad \text{for } j = 1,2,\ldots.$$

Hence

$$\begin{aligned}
E[|Y_1 - Y_2|] &= \sum_{j=0}^{\infty} jP(|Y_1 - Y_2| = j) \\
&= 2\sum_{j=1}^{\infty} j \sum_{k=j}^{\infty} p_k p_{k-j} \\
&= 2\sum_{k=1}^{\infty} p_k \sum_{j=1}^{k} jp_{k-j} \\
&= 2\sum_{k=1}^{\infty} p_k \sum_{j=0}^{k-1} P(Y \le j) \\
&= 2\sum_{j=0}^{\infty} \sum_{k=j+1}^{\infty} p_k P(Y \le j) \\
&= 2\sum_{j=0}^{\infty} P(Y \ge j+1) P(Y \le j) \\
&= 2\sum_{j=0}^{\infty} F_Y(j)\Phi_Y(j+1) \qquad \text{(A1)} \\
&= 2\sum_{j=0}^{\infty} \Phi_Y(j+1)[1 - \Phi_Y(j+1)] \\
&= 2\left[\sum_{j=1}^{\infty} \Phi_Y(j) - \sum_{j=1}^{\infty} \Phi_Y(j)^2\right] \\
&= 2\left[\mu_Y - \sum_{j=1}^{\infty} \Phi_Y(j)^2\right]. \qquad \text{(A2)}
\end{aligned}$$

Then

$$\gamma_Y = \frac{E[|Y_1 - Y_2|]}{2\mu_Y}$$

gives

$$\begin{aligned}
\gamma_Y &= \frac{\sum_{j=0}^{\infty} F_Y(j)\Phi_Y(j+1)}{\mu_Y} \quad \text{from (A1)} \\
&= 1 - \frac{\sum_{j=1}^{\infty} \Phi_Y(j)^2}{\mu_Y} \quad \text{from (A2).}
\end{aligned}$$

(ii) For the continuous case, suppose that Y has probability density function $f_Y(x)$ on $x \ge 0$ and let Y_1, Y_2 be independent copies of Y. Then

$$\begin{aligned}
E[|Y_1 - Y_2|] &= \int_0^{\infty} \int_0^{\infty} |x - y| f_Y(x) f_Y(y)\, dx\, dy \\
&= 2 \iint_{x>y>0} (x - y) f_Y(x) f_Y(y)\, dx\, dy \\
&= 2 \int_0^{\infty} \left[\int_0^{x} (x - y) f_Y(y)\, dy\right] f_Y(x)\, dx
\end{aligned}$$

$$= 2\int_0^\infty \left[xF_Y(x) - \int_0^x yf_Y(y)\,dy\right] f_Y(x)\,dx$$

$$= 2\int_0^\infty \left[xF_Y(x) - yF_Y(y)|_0^x + \int_0^x F_Y(y)\,dy\right] f_Y(x)\,dx$$

$$= 2\int_0^\infty \left[\int_0^x F_Y(y)\,dy\right] f_Y(x)\,dx$$

$$= 2\int_0^\infty \left[\int_0^x f_Y(x)F_Y(y)\,dy\right] dx$$

$$= 2\int_0^\infty \left[\int_y^\infty f_Y(x)F_Y(y)\,dx\right] dy$$

$$= 2\int_0^\infty \Phi_Y(y)F_Y(y)\,dy \qquad \text{(A3)}$$

$$= 2\int_0^\infty \Phi_Y(y)\,[1 - \Phi_Y(y)]\,dy$$

$$= 2\int_0^\infty [\Phi_Y(y) - \Phi_Y(y)^2]\,dy$$

$$= 2\left[\mu_Y - \int_0^\infty \Phi_Y(y)^2\,dy\right]. \qquad \text{(A4)}$$

Then

$$\gamma_Y = \frac{E[|Y_1 - Y_2|]}{2\mu_Y}$$

$$= \frac{\int_0^\infty \Phi_Y(y)F_Y(y)\,dy}{\mu_Y} \quad \text{from (A3)}$$

$$= 1 - \frac{\int_0^\infty \Phi_Y(y)^2\,dy}{\mu_Y} \quad \text{from (A4).}$$

□