Information Processing & Management Vol. 28, No. 5, pp. 637–645, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + 00

A SIMPLE MODEL FOR LINKED INFORMETRIC PROCESSES

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

(*Received* 2 *January* 1992; *accepted in final form* 6 *January* 1992)

Abstract—The notion of linked informetric processes is considered and a simple mathematical formulation is presented. Analysis of the model gives some insight on the differences in concentration between such processes which has been noted in empirical work.

Keywords: Informetric process, Concentration measure, Coefficient of variation, Gini index, Leimkuhler curve.

1. INTRODUCTION

The general setting for the description of an informetric process is that of a population of sources producing items at random over time. Standard examples include authors writing papers, monographs accumulating loans, journals providing references for a bibliography, papers receiving citations. Much of the mathematical work in informetrics over the past decade or so has been concerned with the manner in which the items are distributed over the sources in such situations although the origins can be traced back to Lotka (1926) for the case of scientists publishing research papers and Bradford (1934) for the contributions of different journals to a subject bibliography. However, as Rousseau (in press) has remarked, two such source-item pairings might well be linked in a natural way. For example, scientists write papers which then receive citations so that the items from the first pairing become the sources for the second. In this situation, we can consider the secondary items as arising from the original as well as from the secondary sources, that is, we could consider the citations as relating to the original authors as well as to the published papers. [This is an example of what Egghe & Rousseau (1990, p. 378) call a three-dimensional informetric study.]

In this case of a linked pair of informetric processes, it is of some interest to compare the distributions of, for example, "papers over scientists" with "citations over scientists." Rousseau (in press) quoted just such an example using data taken from Allison (1980), and noted that the inequality, as measured by the coefficient of variation, was greater in the latter case than in the former. He showed, by consideration of the Leimkuhler curve (equivalent to the standard Lorenz curve), that this inequality could be explained in the case where the number of citations is a simple function (having certain monotonicity properties) of the number of papers. This result turns out to be a special case of a more general theorem of Fellman (1976), see also Burrell (in press).

In this article, we present a slightly more realistic setting in which the original production process is stochastic, developing in time, and in which the number of citations received by a published paper is random.

2. THE STOCHASTIC MODEL

To simplify and at the same time motivate the discussion, we shall talk in terms of the particular context alluded to above, viz. we imagine a population of scientists publishing papers over time and that each paper then receives citations, where both the number of papers published and the number of citations to a published paper may be zero.

A citation study typically takes the form "citations acquired during 1985–1990 by papers published during 1985". More generally, consider a publication period of length t and

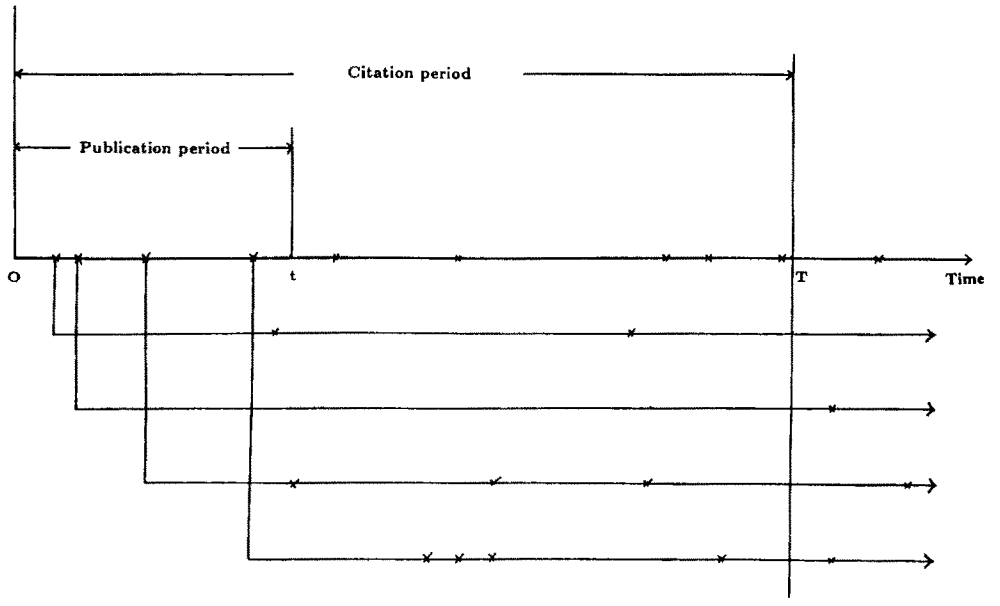

Fig. 1. The publication/citation process.

a citation period of length T with $T \gg t$. Consider first an individual scientist for whom the publication/citation process may be pictured as in Figure 1. Here, the marked points on the main time axis denote the publication events, while those on the branches denote the subsequent citation events for each of those papers. Thus, this author published four papers during the publication period for which the first paper published received two citations during the period of study, the second received none, the third three, and so on. In general, the total number of citations to the author's work is given by the sum of the numbers of citations to each of the individual papers. Hence, if we write $X = X_t$ = number of papers published during $[0, t]$ by a randomly chosen author; Y_j = number of citations to the jth paper of this author; and N = total number of citations during $[0, T]$ received by these papers then we have

$$N = Y_1 + Y_2 + \ldots + Y_X \tag{1}$$

So that N is a random sum of random variables. In order to simplify matters we make the following:

ASSUMPTION

$Y_1, Y_2, \ldots$ are independent and identically distributed random variables.

Some comments on this assumption are in order. While the independence assumption may be tenable, there are several objections to the Y_js all having the same distribution. It is clear, for example, that papers appearing late in the publication period have less opportunity to attract citations than those published early on. It is on this point that the requirement $T \gg t$ is important because it is then reasonable to suppose that slight differences in publication date will have little real effect on the resulting citation opportunity and the identical distribution assumption might then be justified.

[Aside: In the limit as $T \to \infty$, $N = N_t$ gives the total number of future citations to papers publishing during $[0, t]$ and the assumption is perfectly justified. If $\{X_t; t \geq 0\}$ were assumed to be a Poisson process, then $\{N_t; t \geq 0\}$ would be called a compound Poisson process (see e.g. Parzen (1962, p. 128)).]

A further objection is that it is assumed that the distribution of citations to individual papers is the same for all authors. We have no direct evidence to support this assumption and admit that it is counter to the success-breeds-success philosophy. However, given

these admittedly rather severe assumptions, the resulting model allows simple yet suggestive analysis of the difference in concentration between the X and N distributions.

3. CONCENTRATION ASPECTS

3.1 *The coefficient of variation*

In keeping with the context under discussion, assume that X takes values $1,2,\ldots$, that is, only authors who have published during the period are included, while for any of these the number of citations received, N, may be $0,1,2,\ldots$. Of course a citation index only lists papers which have actually been cited, that is, authors for whom $N = 0$ would not be listed. However, knowing the number of publishing authors and the number of cited authors allows us to fill in the number of uncited authors. (Similar remarks apply to citations to individual papers so that we know the number of uncited papers, i.e. those for which $Y = 0$, even though these are not listed.)

According to their classification, Egghe and Rousseau (1991) have shown that the best concentration measures are those equivalent to what they term the "generalized Pratt measures". Of these, the most widely known is that of order 2, otherwise the *coefficient of variation* (CV). The CV of a random variable, or its probability/frequency distribution is just the ratio of its standard deviation to its mean. Actually, in the following it is more convenient to work with the square of the CV which we denote G as it is sometimes referred to as Gastwirth's index. Thus, for a random variable W we have

$$G_W = \frac{\mathrm{Var}(W)}{(E[W])^2} = (CV_W)^2$$

For the model of §2 the result is:

THEOREM 1

$$G_N = G_X + \frac{G_Y}{E[X]}$$

Proof. See Appendix.

From this we have immediately:

COROLLARY 1

$$CV_N > CV_X$$

This result therefore supports the empirical findings of Rousseau (in press) mentioned in §1, although the context is not exactly equivalent to the one modelled here.

3.2 *The uncited papers*

The assumption that it is the *full* distribution of citations over authors, including those from whom $N = 0$, is important in the above result. If instead we work with the zero-truncated form N^* giving the number of citations to an author receiving at least one citation, so that N^* has possible values $1,2,\ldots$ and its distribution may be determined directly from the citation index, then we have:

COROLLARY 2

$$CV_{N^*} \gtrless CV_X \text{ according as } \frac{P(N \neq 0)}{P(N = 0)} \gtrless \frac{(G_X + 1)E[X]}{G_Y} \tag{2}$$

(Again, the proof is given in the Appendix). Note that the left-hand side of the above is just the odds against a published paper being uncited.

The importance of being specific about whether or not the nonproducers are included in informetric studies has been raised many times before. In the context of library circu-

lation models, see the discussion following Burrell and Cane (1982), the pointed remarks of Bagust (1983) and the response by Burrell (1984). The effect on concentration measures, in particular the Gini index, is graphically illustrated in Burrell (1992).

3.3 *Another example as a special case*

A citation index often includes also an author index, that is, a listing of cited authors and their cited papers. One might then consider the distribution of the number of cited papers over this population of cited authors and compare this with the original distribution of publications as another example of a linked pair of informetric processes. This can be modelled exactly as before simply by redefining

$$Y_j = \begin{cases} 1 & \text{if the author's } j\text{th published paper is cited,} \\ 0 & \text{otherwise} \end{cases}$$

M = Number of the author's published papers which are cited

$$= Y_1 + Y_2 + \ldots + Y_X$$

and as a special case of the theorem we have

COROLLARY 3

$$G_M = G_X + \frac{1-\theta}{\theta E[X]}$$

where θ is the probability that a published paper is cited.

In particular note, therefore, that $CV_M > CV_X$. Of course, the author index will only include authors having at least one cited paper so that we should consider the zero-truncated form M^* in which case we have

COROLLARY 4

$$CV_{M^*} \gtrless CV_X \text{ according as } \frac{P(M \neq 0)}{P(M=0)} \gtrless (G_X + 1)E[X]\,\frac{\theta}{1-\theta} \tag{3}$$

Note that $M = 0 \Leftrightarrow N = 0$ and the left-hand side of the above is again the odds against a published paper being uncited.

4. A SIMPLE TIME-DEPENDENT EXAMPLE

Suppose that each author in the population publishes papers as a Poisson process where the rates of these processes vary over the population as an exponential distribution. The resulting publication process for the population is then the well-known exponential mixture of Poisson processes which has been employed in many informetric models, see, for example, Burrell (1980) and Burrell and Cane (1982) in the context of library circulations. The resulting distribution of X_t, the number of published papers of a publishing author during $[0,t]$, is then a time-dependent geometric distribution:

$$P(X_t = k) = pq^{k-1}, \quad k = 1,2,\ldots$$

where the parameter is

$$p = p(t) = \frac{1}{1+\theta t} = 1 - q \tag{4}$$

and $\theta > 0$ is a (time) scale parameter giving the mean number of publications per unit time for the entire population of potentially publishing authors (whether or not they happen to have published during the particular period of observation).

Let us further assume that the common distribution of the number of citations to a published paper is given by

$$P(Y = k) = \alpha\beta^k, \quad k = 0,1,2\ldots$$

where $\alpha + \beta = 1$, so that the Y_js have a geometric distribution on 0, 1, 2.... Note that there is no time dependence in the distribution of the Y_js so that essentially we are assuming that $T \to \infty$ or equivalently that each published paper effectively receives all its citations immediately upon publication.

Using well-known properties of the geometric distribution we have

$$E[X] = 1/p, \qquad E[Y] = \beta/\alpha$$

$$\mathrm{Var}(X) = q/p^2, \qquad \mathrm{Var}(Y) = \beta/\alpha^2$$

so that

$$G_X = q, \qquad G_Y = 1/\beta$$

and hence from the theorem

$$G_N = q + \frac{p}{\beta}$$

For the case where the uncited papers are not counted note that $P(N = 0) = p\alpha/(1 - q\alpha)$ (see Appendix) so that from Corollary 2

$$\begin{aligned} CV_{N^*} > CV_X &\Leftrightarrow \frac{P(N \neq 0)}{P(N = 0)} > \frac{(G_X + 1)E[X]}{G_Y} \\ &\Leftrightarrow \frac{\beta}{p\alpha} > (1 + q)\frac{\beta}{p} \\ &\Leftrightarrow \frac{1}{\alpha} > 1 + q \qquad (5) \\ &\Leftrightarrow \frac{\beta}{\alpha} > q = 1 - p \end{aligned}$$

In particular, note that if $\alpha \leq \frac{1}{2}$ (so that $E[Y] \geq 1$) then necessarily

$$CV_{N^*} > CV_X$$

that is, if the mean number of citations per published paper exceeds 1, then the coefficient of variation of the distribution of citations over (cited) authors exceeds that of the distribution of publications over (published) authors.

If on the other hand $\alpha > \frac{1}{2}$ (so that $E[Y] < 1$) then (4) and (5) together yield

$$CV_{N^*} > CV_X \Leftrightarrow \theta t < \frac{\beta}{2\alpha - 1}$$

which inequality will hold initially but fail for larger t.

Actually this example can be analyzed further because the distribution of N_t^* can be found explicitly as

$$P(N_t^* = k) = \left(\frac{p\alpha}{1 - q\alpha}\right)\left(\frac{\beta}{1 - q\alpha}\right)^{k-1}, \quad k = 1,2,\ldots$$

(see Appendix). Hence, N_t^* also has a geometric distribution, with parameter $p\alpha/(1 - q\alpha) = \alpha/[1 - (1 + \alpha)\theta t]$.

In general, it is shown in Burrell (1992) that if Z has a geometric distribution on $1,2,\ldots$ with parameter ρ, that is,

$$P(Z = k) = \rho(1 - \rho)^{k-1}, \quad k = 1,2,\ldots$$

then the Gini index or coefficient of concentration (equivalent to the generalized Pratt measure of order 1) of Z is given by

$$\gamma_Z = \frac{1 - \rho}{2 - \rho}$$

Hence,

$$\gamma_{N^*} > \gamma_X \Leftrightarrow \frac{\beta/(1 - q\alpha)}{1 + \beta/(1 - q\alpha)} > \frac{q}{1 + q}$$

This last inequality reduces, after a little algebra, to

$$\frac{1}{1 + q} > \alpha$$

exactly as at (5). Hence, the conditions for inequality of concentration as measured by the Gini index are exactly the same as those for when measured by the coefficient of variation. This is to be expected because, following Egghe and Rousseau (1991), we have that "good concentration measures respect the Lorenz (or Leimkuhler) dominance ordering". For the case of a geometric distribution on $1,2,\ldots$ with parameter ρ, the Leimhkuhler curve of concentration is given by Burrell (1992) as

$$\Psi = \Phi\left(1 + \frac{\rho}{(1 - \rho)\ln(1 - \rho)} \ln \Phi\right)$$

from which it easily follows that if Z_1, Z_2 have such distributions with parameters ρ_1, ρ_2 then the Leimkuhler curve of Z_1 dominates that of Z_2 if and only if $\rho_1 < \rho_2$. Hence, we have that the concentration measure (however defined) of N_t^* exceeds that of X_t

$$\Leftrightarrow \frac{\alpha p}{1 - q\alpha} < p$$

$$\Leftrightarrow \alpha < \frac{1}{1 + q}$$

5. CONCLUDING REMARKS

The proposed model is perhaps oversimplified and the assumptions unrealistic so that the results are suggestive rather than definitive. The model already allows for variation in publication rates across the population of authors, the next step would be to allow for a variation in citation rates also, if such a variation does indeed exist.

REFERENCES

Allison, P.D. (1980). Inequality and scientific productivity. *Social Studies of Science, 10*, 163–179.
Bagust, A. (1983). A circulation model for busy public libraries. *Journal of Documentation, 39*, 24–37.
Bradford, S.C. (1934). Sources of information on specific subjects. *Engineering, 137*, 85–86.
Burrell, Q.L. (1980). A simple stochastic model for library loans. *Journal of Documentation, 36*, 115–132.
Burrell, Q.L. (1984). Library circulation models. *Journal of Documentation, 40*, 68–71.
Burrell, Q.L. (in press). A note on a result of Rousseau for concentration measures. Research report, University of Manchester.
Burrell, Q.L. (1992). The Gini index and the Leimkuhler curve for bibliometric processes. *Information Processing & Management, 28*, 19–33.
Burrell, Q.L., & Cane, V.R. (1982). The analysis of library data (with discussion). *Journal of the Royal Statistical Society, Series A, 145*, 439–471.
Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics*. Elsevier: Amsterdam.
Egghe, L., & Rousseau, R. (1991). Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science, 42*, 479–489.
Feller, W. (1968). *An introduction to probability theory and its applications*, vol. 1 (3rd Edition). Wiley: New York.
Fellman, J. (1976). The effect of transformations on Lorenz curves. *Econometrica, 44*, 823–824.
Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*, 317–323.
Parzen, E. (1962). *Stochastic processes*. Holden-Day: San Francisco.
Rousseau, R. (in press). Concentration and diversity of availability and use in information systems: a positive reinforcement model.

APPENDIX

All of the random variables X, N and $Y_j, j = 1,2,\ldots$ as defined in §2 are non-negative and integer-valued. For such a random variable W, it is convenient to work with its probability generating function (*pgf*) $\Pi_W(z)$ where

$$\Pi_W(z) = \sum_{k=0}^{\infty} z^k P(W = k)$$
$$= E[z^W]$$

Here, E denotes expectation/expected value. The following result is well-known, see e.g. Feller (1968, p. 287).

LEMMA 1

For N defined via (1),

$$\Pi_N(z) = \Pi_X(\Pi_Y(z)). \tag{A1}$$

where Y has the same distribution as $Y_1, Y_2, \ldots$

Proof of Theorem 1. Recall that for any *pgf* $\Pi_W(z)$ we have

$$\Pi_W(1) = 1$$

$$\Pi'_W(1) = E[W]$$

$$\Pi''_W(1) = E[W(W-1)]$$

Differentiating (A1) twice then gives, on putting $z = 1$,

$$E[N] = E[X]E[Y]$$

$$E[N(N-1)] = E[X^2]E[Y]^2 + E[X]\mathrm{Var}(Y) - E[X]E[Y]$$

so that

$$E[N^2] = E[X^2]E[Y]^2 + E[X]\mathrm{Var}(Y)$$

Writing

$$G_W = \frac{\text{Var}(W)}{E[W]^2} = \frac{E[W^2]}{E[W]^2} - 1$$

so that G_W is the square of the coefficient of variation of W

$$\begin{aligned} G_N &= \frac{E[X^2]E[Y]^2 + E[X]\text{Var}(Y)}{E[X]^2E[Y]^2} - 1 \\ &= \frac{E[X^2]}{E[X]^2} + \frac{\text{Var}(Y)}{E[X]E[Y]^2} - 1 \qquad \text{(A2)} \\ &= G_X + \frac{G_Y}{E[X]} \end{aligned}$$ □

For Corollary 1 note that $G_Y > 0, E[X] > 0$ so that $G_N > G_X$ and hence

$$CV_N > CV_X$$ □

For the case of zero-truncated data,

$$P(N^* = k) = P(N = k | N \neq 0) = \frac{P(N = k)}{P(N \neq 0)} \qquad \text{(A3)}$$

and then

$$E[(N^*)^r] = \frac{E[N^r]}{P(N \neq 0)}$$

so that

$$\begin{aligned} G_{N^*} &= \frac{E[(N^*)^2]}{E[N^*]^2} - 1 \\ &= P(N \neq 0)\frac{E[N^2]}{E[N]^2} - 1 \qquad \text{(A4)} \\ &= P(N \neq 0)(G_N + 1) - 1 \end{aligned}$$

Now $CV_{N^*} > CV_X \Leftrightarrow G_{N^*} + 1 > G_X + 1$

$$\Leftrightarrow P(N \neq 0)(G_N + 1) > G_X + 1 \quad \text{from (A4)}$$

$$\Leftrightarrow P(N \neq 0)\left(G_X + 1 + \frac{G_Y}{E[X]}\right) > G_X + 1 \quad \text{from (A2)}$$

$$\Leftrightarrow P(N \neq 0)\frac{G_Y}{E[X]} > P(N = 0)(G_X + 1)$$

$$\Leftrightarrow \frac{P(N \neq 0)}{P(N = 0)} > (G_X + 1)\frac{E[X]}{G_Y} = \frac{E[X^2]}{E[X]}\frac{1}{G_Y}$$

For the special case mentioned in §3.3, note that $E[Y] = \theta$, and $\text{Var}(Y) = \theta(1 - \theta)$ so that $G_Y = (1 - \theta)/\theta$.

EXAMPLE.

$$P(X = k) = pq^{k-1}, \quad k = 1,2,\ldots$$

$$P(Y = k) = \alpha\beta^k, \quad k = 0,1,2,\ldots$$

Here

$$\Pi_X(z) = \frac{pz}{1 - qz}, \qquad \Pi_Y(z) = \frac{\alpha}{1 - \beta z}$$

so from (A1)

$$\begin{aligned}\Pi_N(z) &= \frac{p\alpha/(1 - \beta z)}{1 - q\alpha/(1 - \beta z)} \\ &= \frac{A}{1 - Bz} \\ &= A \sum_{k=0}^{\infty} B^k z^k\end{aligned}$$

where

$$A = \frac{p\alpha}{1 - q\alpha}, \qquad B = \frac{\beta}{1 - q\alpha}$$

Thus

$$\begin{aligned}P(N = k) &= \text{coeff. } z^k \\ &= AB^k, \quad k = 0,1,2,\ldots\end{aligned}$$

and from (A2)

$$\begin{aligned}P(N^* = k) &= \frac{AB^k}{1 - A} \\ &= \left(\frac{p\alpha}{1 - \alpha}\right)\left(\frac{\beta}{1 - q\alpha}\right)^k \\ &= \left(\frac{p\alpha}{1 - q\alpha}\right)\left(\frac{\beta}{1 - q\alpha}\right)^{k-1}, \quad k = 1,2,\ldots.\end{aligned}$$