Information Processing & Management Vol. 28, No. 1, pp. 35–44, 1992
Printed in Great Britain.

0306-4573/92 $5.00 + .00

DUALITY ASPECTS OF THE GINI INDEX FOR GENERAL INFORMATION PRODUCTION PROCESSES

L. EGGHE
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium*
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

Abstract—This paper studies information production processes (IPP) (e.g., bibliographies) from the point of view of concentration theory. More specifically, the Gini index is studied for an IPP as well as for its dual IPP. We prove that both Gini indices are the same. We also remark that such a result is not true for other well-known concentration measures.

1. INTRODUCTION

1.1 *Concentration theory and the Gini index*

Concentration theory studies the degree of inequality in a set of numbers $x_1, x_2, \ldots, x_n$ (usually, n is very high). It originates from econometrics, where one wants to study the degree of income inequality in a certain population. Among the many references, we mention Gini (1909), a historically important paper. More generally, in all areas of sociology, concentration problems occur; see, for example, Allison (1977,1978,1980), Cole (1983), Chapman and Farina (1982), Johnson (1979), Ray and Singer (1973), Theil (1967).

More specifically (and also more recently), concentration has also been studied in informetrics; see, for example Egghe and Rousseau (1990a,1991), Egghe (1987), Bonckaert and Egghe (1991), Heine (1978), Burrell (1990), Pratt (1977), and Carpenter (1979).

Measuring inequality between numbers is, evidently, linked with the standard deviation of these numbers. However, this measure has some undesirable properties. In general, we are looking for a function:

$$f: (x_1, x_2, \ldots, x_n) \to f(x_1, x_2, \ldots, x_n)$$

satisfying a number of "good" concentration properties (e.g., scale invariance and the transfer principle). We do not go into detail here since this has been worked out in, for example, Egghe and Rousseau (1990a,1991). From these studies—and that is what is important here—it follows that Gini's index is a very good concentration measure. It is defined as follows.

1.1.1 *Gini's index in the discrete case.* Let $x_1, x_2, \ldots, x_n$ be our numbers. We suppose they are arranged in increasing order. We construct the following graph (called the Lorenz-curve): Divide the interval [0,1] into n equal parts. With the abscis j/n ($j = 1, \ldots, n$), we link the value $\sum_{i=1}^{j} x_i / \sum_{i=1}^{n} x_i$ (and we take 0 in 0). Then we obtain a graph λ as in Fig. 1. The Lorenz curve is always below the first bissectrice, since the x_i are increasing. Consider the area D between the Lorenz curve and this first bissectrice. Then, the Gini index G is defined as

$$G = 2D. \tag{1}$$

Otherwise stated, G is equal to

$$G = 1 - 2 \text{ (area beneath the Lorenz curve).} \tag{2}$$

*Permanent address.

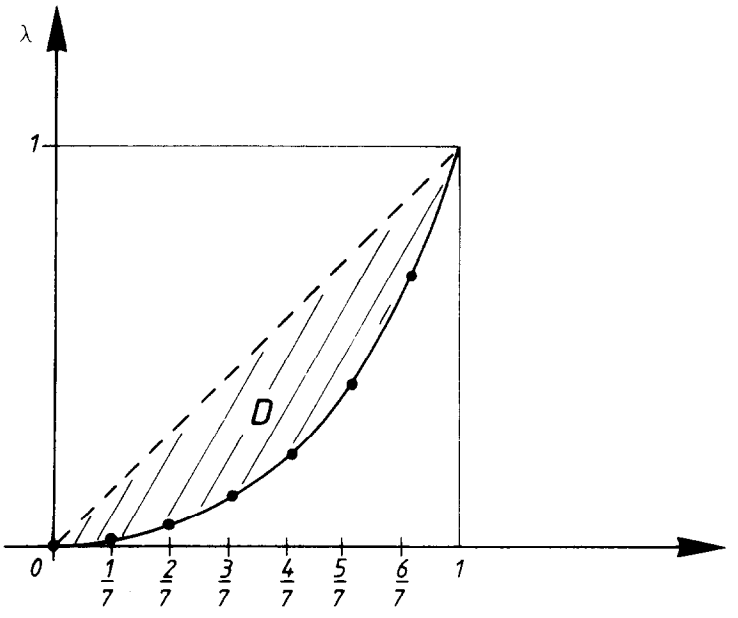

Fig. 1. Example of a Lorenz-curve with $n = 7$.

1.1.2 *Gini's index in the continuous case.* This formula (2) is also used when we deal with a continuous situation: instead of $x_1, \ldots, x_n$, we have now a function $f(x)$, $x \in [0,1]$ such that f is integrable and increasing. The construction of the Lorenz curve implies a graph of the function:

$$\lambda : x \to \lambda(x) = \frac{\int_0^x f(x')\,dx'}{\int_0^1 f(x')\,dx'} \tag{3}$$

and hence, according to (2):

$$G = 1 - 2\int_0^1 \lambda(x)\,dx$$

or

$$G = 1 - 2\int_0^1 \frac{\int_0^x f(x')\,dx'}{\int_0^1 f(x')\,dx'}\,dx. \tag{4}$$

It is this Gini index that we will study in the dual framework of information production processes (IPP).

1.2 *Information production processes (IPP)*

The information production processes (IPP), as introduced in Egghe (1989), are continuous mathematical models of production processes such as bibliographies, economic production processes, social usages (e.g., of words in a text) and so on. (See also Egghe 1990a–c, Egghe & Rousseau, 1990b, and Rousseau, 1991a.) We briefly repeat the notation.

An IPP is a triple of the form

$$(S, I, V) \tag{5}$$

where $S = [0, T]$, $I = [0, A]$ and

$$V: S \to I$$

is a function such that V is strictly increasing and differentiable, such that V' is strictly increasing and such that $V(0) = 0$ and $V(T) = A$. To be clearer, we think of S as the source set (sources being the objects that produce) and of I as the item set (items being produced by the sources). In the sequel we will consider $V(r)$ (for every $r \in S \setminus \{0\}$) to be the cumulative number of items in all sources $s \in [T - r, T]$. Hence V is an integral of a function, called ρ (hence $V' = \rho$). Since $V(0) = 0$, we have:

$$V(r) = \int_0^r \rho(r')\, dr'. \tag{6}$$

The dual IPP of (S, I, V) is the IPP

$$(I, S, U), \tag{7}$$

where

$$U(i) = T - V^{-1}(A - i) \tag{8}$$

and where V^{-1} denotes the inverse function of V (note that V is injective since it is strictly increasing). As for (S, I, V), we define (U is differentiable by (8)): $U' = \sigma$; hence

$$U(i) = \int_0^i \sigma(i')\, di' \tag{9}$$

(since $U(0) = 0$, as follows from (8)). When expressed as a function of i in the IPP (S, I, V) (hence $i = V(r)$), $\rho(r)$ becomes

$$\rho(i) = V'(V^{-1}(i)). \tag{10}$$

Otherwise stated: If $i = V(r)$, we define

$$\rho(i) = \rho(r), \tag{11}$$

for every $r \in [0, T]$ (or $i \in [0, A]$). We also repeat the following (easy) result (Egghe, 1989,1990a):

LEMMA

For every IPP (S, I, V),

$$\sigma(i) = \frac{1}{\rho(A - i)}, \tag{12}$$

for every $i \in I$.

This mathematical model has turned out to be very efficient at proving and understanding certain regularities in, for instance, informetrics; see Egghe (1989,1990c).

In this paper, we will try to study Gini's index in the dual framework of IPPs. Note that dual situations are often encountered in informetrics. Two examples:

- Citation bibliographies: If we change "cited" to "citing," we change the IPP into its dual.
- If the IPP studies (as in Lotka's law) "authors with a certain number of articles," the dual IPP studies "articles with a certain number of authors" (meaning the number of co-authors per article—an important issue in, e.g., science policy studies).

In this framework it is important to investigate the dual properties of IPPs. This has already commenced in Egghe (1989,1990a-c) and continues here. Now we emphasize the dual properties of the Gini index for IPPs.

2. DUAL THEORY OF THE GINI INDEX

2.1 *Preliminaries*

Let (S, I, V) be as in section I, where $I = [0, A]$, $S = [0, T]$, and let (I, S, U) be its dual. Since $V(r)$ represents the cumulative number of items in all sources $s \in [T - r, T]$, for every $r \in [0, T]$, it follows from (8) that $U(i)$ is the cumulative number of sources that produce the items in $[0, i]$. Since $\sigma = U'$, we then have that σ is the density of the number of sources in the item-coordinate i. Hence (since $\rho = V'$ increases strictly and by (12)) σ is strictly increasing. So we conclude that σ represents the generalized continuous version of Bradford's law. This function can be used for the calculation of Gini's index, by putting $f = \sigma$ in (3) and (4). To use σ for concentration calculations is even more natural, since the classical law of Bradford has often been considered as an "expression" of degree of concentration of a bibliography (even in the historical paper of Bradford himself (1934) one talks about three groups of journals with a different status: the ones that have to be bought, the ones that one buys if one has the money, and the ones that should not be bought!). The original law of Bradford (but here in a continuous, group-free setting, which is equivalent with the historical law of Bradford—see Egghe, 1989,1990a-c) states that

$$\sigma(i) = M \cdot K^i \tag{13}$$

for every $i \in I$, where M and K are parameters. Parameter K is larger than 1 (since σ increases strictly) and is called the group-free Bradford factor (or multiplicator). This law corresponds to classical informetric laws, such as Lotka's law (with exponent 2) and the law of Leimkuhler: Let $R(r)$ denote the cumulative number of items in the sources $s \in [0, r]$, for every $r \in [0, T]$ in the IPP (I, S, U). Then

$$R(r) = a \ln(1 + br) \tag{14}$$

where a and b are parameters. Note that $R = U^{-1}$. In Egghe (1989,1990c) and Rousseau (1988), generalizations of the above functions (for Lotka-exponents $\alpha \neq 2$) are constructed. These generalizations as well as (13) and (14) will be used in the next section, where we will calculate Gini's index for these cases, based on the function σ.

However, the main topic of this paper is to investigate the relation between the Gini index G for the IPP (S, I, V) and the Gini index G_d of its dual (I, S, U). It is surprising that we can prove that $G = G_d$, for general IPPs (hence not supposing a certain fixed function as, e.g., (13)).

2.2 *The Gini indices G and G_d*

Let (S, I, V) be any IPP and (I, S, U) its dual. We define the Gini index of (S, I, V) to be (cf. (4)):

$$G = 1 - \frac{2T}{A} \int_0^1 \int_0^x \rho(Tx')\, dx'\, dx. \tag{15}$$

Note that $r = Tx$, $r' = Tx'$, in the notation of section 1. The use of A and T in (15) is to make sure that our function

$$\lambda : x \to \int_0^x \frac{T\rho(Tx')}{A}\, dx'$$

is a function for which $\varphi(1) = 1$. Indeed:

$$\frac{T}{A}\int_0^1 \rho(Tx')\, dx' = \frac{1}{A}\int_0^T \rho(x'')\, dx''$$

$$= \frac{1}{A}\, V(T) = 1,$$

as requested by (3). Logically, we define the Gini index of the dual IPP (I, S, U) to be

$$G_d = 1 - \frac{2A}{T}\int_0^1 \int_0^y \sigma(Ay')\, dy'\, dy. \tag{16}$$

Note that $i = Ay$, $i' = Ay'$, in the notation of section I.

2.3 *Relationship between G and G_d*

We will prove a relationship between G and G_d. First, some preliminary results.

LEMMA II.3.1

$$\int_0^1 \int_0^y \sigma(Ay')\, dy' = \int_0^1 (1 - y)\sigma(Ay)\, dy \tag{17}$$

Proof. We apply the theorem of Fubini (Apostol, 1974). Hence

$$\int_0^1 \int_0^y \sigma(Ay')\, dy' = \int_0^1 \int_{y'}^1 \sigma(Ay')\, dy\, dy'$$

$$= \int_0^1 [y\sigma(Ay')]_{y=y'}^{y=1}\, dy'$$

$$= \int_0^1 (1 - y')\sigma(Ay')\, dy'$$

$$= \int_0^1 (1 - y)\sigma(Ay)\, dy. \qquad \square$$

We now express G in terms of σ, instead of ρ.

LEMMA II.3.2

$$G = 1 - \frac{2A}{T}\int_0^1 y\sigma(A - Ay)\, dy \tag{18}$$

where $V(Tx) = V(r) = i = Ay$.

Proof. By (15)

$$G = 1 - \frac{2T}{A}\int_0^1 \int_0^x \rho(Tx')\, dx'\, dx. \tag{15}$$

Since $Ay = V(Tx)$, we have $Tx = V^{-1}(Ay)$ and hence

$$dx = \frac{A}{T} \frac{dy}{V'(V^{-1}(Ay))}.$$

By (10) and (11), we have $V'(V^{-1}(Ay)) = \rho(Ay) = \rho(Tx)$, and the same for dx':

$$dx' = \frac{A}{T} \frac{dy'}{\rho(Tx')}.$$

Putting this in (15) yields

$$G = 1 - \frac{2A}{T} \int_0^1 \int_0^y \frac{dy'\, dy}{\rho(Ay)}.$$

We now use (12) to obtain:

$$G = 1 - \frac{2A}{T} \int_0^1 \int_0^y \sigma(A - Ay)\, dy'\, dy$$

$$G = 1 - \frac{2A}{T} \int_0^1 y\sigma(A - Ay)\, dy. \qquad \square$$

This is an expression of G in terms of the function σ of the dual IPP. We can now prove:

THEOREM II.3.3

For any IPP (S, I, V) with dual (I, S, U) we have that

$$G = G_d. \tag{19}$$

Proof. Note that, for every integrable function:

$$\int_0^1 \psi(y)\, dy = \int_0^1 \psi(1 - y)\, dy. \tag{20}$$

Indeed, substituting $y' = 1 - y$ in the left-hand side of (20), we find

$$\int_0^1 \psi(y)\, dy = -\int_1^0 \psi(1 - y')\, dy'$$

$$= \int_0^1 \psi(1 - y')\, dy'.$$

From (20) we deduce:

$$\int_0^1 (1 - y)\sigma(Ay)\, dy = \int_0^1 y\sigma(A - yA)\, dy.$$

Hence, by (18):

$$1 - G = \frac{2A}{T} \int_0^1 y\sigma(A - Ay)\, dy$$

$$= \frac{2A}{T} \int_0^1 (1 - y)\sigma(Ay)\, dy.$$

So,

$$\frac{T}{A}(1-G) = 2\int_0^1 (1-y)\sigma(Ay)\,dy.$$

Combining (16) and (17) yields:

$$\frac{T}{A}(1-G) = \frac{T}{A}(1-G_d).$$

Hence:

$$G = G_d. \qquad \square$$

3. THE GINI INDEX FOR INFORMETRIC IPPs

We will now calculate the formulae:

$$G_d = 1 - \frac{2A}{T}\int_0^1 \int_0^y \sigma(Ay')\,dy'\,dy \tag{16}$$

or

$$G_d = 1 - \frac{2A}{T}\int_0^1 (1-y)\sigma(Ay)\,dy \tag{17}$$

for some informetric laws. From (19), we then have the expression for G automatically. We start with the classical Bradford law.

3.1 *The Gini index for Bradfordian IPPs (i.e., Lotka's law with exponent $\alpha = 2$)*

In this case we suppose

$$\sigma(Ay) = MK^{Ay} \tag{20}$$

for $y \in [0,1]$ (hence $i = Ay \in [0,A]$ as in section 1). Since (17) is easier to calculate than (16), we use the former:

$$G_d = 1 - \frac{2A}{T}\int_0^1 (1-y)MK^{Ay}\,dy.$$

Partial integration yields:

$$G_d = 1 - \frac{2M}{T\ln K}\left(\frac{K^A - 1}{A\ln K} - 1\right). \tag{21}$$

So, we also have, by (19)

$$G = 1 - \frac{2M}{T\ln K}\left(\frac{K^A - 1}{A\ln K} - 1\right). \tag{22}$$

Note that, intuitively, the Bradford factor (the historical group-dependent one (k) or the one above (K)) was considered as a "measure of concentration" but, as is well known, this simple parameter does not have all the good concentration properties as described, for

example, in Egghe and Rousseau (1990a,1991). Formulae (21) and (22) provide an answer to this problem, since they incorporate K.

We can also express (21) and (22) in terms of the parameters a and b occurring in Leimkuhler's law:

$$R(r) = a \ln (1 + br) \tag{14}$$

for $r \in [0, T]$. Here $R(r)$ denotes the cumulative number of items in the sources $s \in [0, r]$. We have the formulae (Egghe, 1989):

$$a = \frac{1}{\ln K} \tag{23}$$

$$b = \frac{\ln K}{M}. \tag{24}$$

Hence, (21) and (22) yield:

$$G = G_d = 1 - \frac{2}{bT}\left(\frac{(e^{A/a} - 1)a}{A} - 1\right). \tag{25}$$

Finally, in terms of the classical Bradford parameters $p, r_o = r_o(p)$, $k = k(p)$, where p is the number of groups, r_o is the number of sources in the first group (the nucleus) and k is the p-dependent Bradford factor, we have the formulae (Egghe 1989) (with $y_o = A/p$):

$$a = \frac{y_o}{\ln k} \tag{26}$$

$$b = \frac{k - 1}{r_o} \tag{27}$$

Now (25) yields:

$$G = G_d = 1 - \frac{2r_o}{T(k-1)}\left(\frac{k^p - 1}{p \ln k} - 1\right). \tag{28}$$

Note that formula (25) corresponds with the analogous calculation of Burrell (1990). However, Burrell uses a different notation:

$$\frac{1}{A} R(x) = \psi(x) = \frac{\ln (1 + \beta x)}{\ln (1 + \beta)} \tag{29}$$

for $x \in [0,1]$. In comparison with our notation, we thus have $a = A/\ln(1 + \beta)$ and $bT = \beta$ (Tx being r). Hence we refind Burrell's formula from (25):

$$G_d = 1 - 2\left(\frac{1}{\ln (1 + \beta)} - \frac{1}{\beta}\right). \tag{30}$$

3.2 *The Gini index for generalized Bradfordian IPPs (i.e., Lotka's law with general exponent α)*

It is shown in Egghe (1989,1990a,c) that if we have the well-known law of Lotka

$$f(y) = \frac{C}{y^\alpha},$$

where $\alpha > 1$ and $C > 0$ are parameters and where $f(y)$ denotes the density of the number of sources with y items, then we have (equivalently)

$$\sigma(i) = \left(\left(\frac{A(2-\alpha)}{C} + 1\right) - i\frac{2-\alpha}{C}\right)^{-1/(2-\alpha)} \tag{31}$$

for every $i \in I$. Now we will calculate $G = G_d$ for σ as above. To simplify the notation, put

$$B = \frac{2-\alpha}{C} \tag{32}$$

So,

$$\sigma(i) = ((AB + 1) - iB)^{-(1/BC)}. \tag{33}$$

In this setting, it is easiest to calculate G by using formula (18). This gives:

$$G_d = G = 1 - \frac{2C}{TAB}\left[\frac{(1+AB)^{-(1/BC)+2}}{2BC-1} - \frac{(1+AB)^{-(1/BC)+1}}{BC-1} + \frac{1}{(BC-1)(2BC-1)}\right]. \tag{34}$$

This is analogous to a result obtained by Rousseau (1991b). In this reference, together with Burrell (1990) one can find numerical calculations of G, based on these formulae.

4. CLOSING REMARKS

We have studied the Gini index in the dual framework of IPPs and we proved that the values of the Gini index in the original IPP are the same as the ones in the dual IPP.

We furthermore calculated explicit formulae for the Gini index in case the IPP satisfies Lotka's laws ($\alpha = 2$, respectively $\alpha \neq 2$).

We finally note that we tried to prove analogous results for other good concentration measures (such as the coefficient of variation and Theil's measure—see Egghe & Rousseau 1990a, 1991—but a simple relationship between the values of these measures in an IPP and the ones of the dual IPP does not seem to exist. Hence the present paper reveals a special property of the Gini index.

REFERENCES

Allison, A. (1977). The reliability of variables measured as the number of events in an interval of time. In K.F. Schuessler (Ed.), *Sociological methodology 1978* (pp. 238–253). San Francisco: Jossey-Bass.

Allison, P.D. (1978). Measures of inequality. *American Sociological Review, 43*, 865–880.

Allison, P.D. (1980). Inequality and scientific productivity. *Social Studies of Science, 10*, 163–179.

Apostol, T.M. (1974). *Mathematical analysis.* Reading, MA: Addison-Wesley.

Bonckaert, P., & Egghe, L. (1991). Rational normalization of concentration measures. Unpublished manuscript.

Bradford, S.C. (1934). Sources of information on specific subjects. *Engineering, 137*, 85–86. Reprinted in *Collection Management, 1*, 95–103, 1976–1977; also reprinted in *Journal of Information Science, 10*, 148 and 176–180, 1985.

Burrell, Q.L. (1990). *The Bradford distribution and the Gini index.* Preprint.

Carpenter, M.P. (1979). Similarity of Pratt's measures of class concentration to the Gini index. *Journal of the American Society for Information Science, 30*, 108–110.

Chapman, I.D., & Farina, C. (1982). Concentration of resources: the National Research Council's (Canada) grants in aid of research: 1964–1974. *Scientometrics, 4*, 105–117.

Cole, S. (1983). The hierarchy of the sciences? *American Journal of Sociology, 89*, 111-139.

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

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

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

Egghe, L. (1990b). A note on different Bradford multipliers. *Journal of the American Society for Information Science*, *41*(3), 204–209.

Egghe, L. (1990). New Bradfordian laws, evolving from a source-item duality argument. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90*, Proceedings of the Second International Conference on Bibliometrics, Scientometrics and Informetrics, London (Canada) 1989. Amsterdam: Elsevier, 79–96.

Egghe, L., & Rousseau, R. (1990a). Elements of concentration theory. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90*, Proceedings of the Second International Conference on Bibliometrics, Scientometrics and Informetrics, London (Canada) 1989. Amsterdam: Elsevier, 97–137.

Egghe, L., & Rousseau, R. (1991). Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science*, *42*(7), 479–489.

Egghe, L., & Rousseau, R. (1990b). Introduction to Informetrics. *Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

Gini, C. (1909). Il diverso accrescimento delle classi sociali e la concentrazione della ricchezza. *Giornale degli Economisti*, serie *11*, 37.

Heine, M.H. (1978). Indices of literature dispersion based on qualitative attributes. *Journal of Documentation*, *34*, 175–188.

Johnson, R.L. (1979). Measures of vocabulary diversity. In D.E. Ager, F.E. Knowles, & J. Smith (Eds.), *Advances in computer-aided literary and linguistic research* (pp. 213–227). Birmingham: AMLC.

Pratt, A.D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, *28*, 285–292.

Ray, J.L., & Singer, J.D. (1973). Measuring the concentration of power in the international system. *Sociological Methods and Research*, *1*, 403–437.

Rousseau, R. (1988). Lotka's law and its Leimkuhler representation. *Library Science with a Slant to Documentation and Information Studies*, *25*, 150–178.

Rousseau, R. (1991a). Category theory and informetrics: Information production processes. *Scientometrics* (in press).

Rousseau, R. (1991b). Written communication.

Theil, H. (1967). *Economics and Information Theory*. Amsterdam: North-Holland.