An Empirical Study of the Measurement of Similarity of Concentration between Different Informetric Distributions

Quentin L. Burrell

q.burrell@ibs.ac.im
Isle of Man International Business School, The Nunnery, Old Castletown Road, Douglas,
Isle of Man IM2 1QB, via United Kingdom

Abstract

There is a well-established literature on the use of concentration measures in informetrics. However, these works have usually been devoted to measures of concentration *within* a productivity distribution. In a recent paper Burrell (2005a) introduced two new measures, both based on the Gini ratio, for measuring the similarity of concentration of productivity *between* two different informetric distributions. The first was derived from Dagum's (1987) notion of relative economic affluence; the second – in some ways analogous to the correlation coefficient – is a completely new approach. This was extended with further theoretical examples in Burrell (2005b). The purpose of this study is to develop a purely empirical approach to comparative studies of concentration between informetric data sets using both *within* and *between* measures thereby greatly extending the original study where Burrell (2005a) considered just two data sets for purposes of illustration of the methods of calculation of the measures.

Introduction

One of the fundamental descriptions of an informetric process is that of a population of sources producing items over a period of time. This covers, for instance, studies of library circulations where the sources are the books in a library (the population) and the items are borrowings; citation analysis where one considers a body of published work (the population) with the sources being the individual papers/articles and the items being the received citations; author productivity where the sources are the researchers (in a particular field) and the items are the authored papers. Note that all of these are essentially dynamic environments in that they involve production processes that develop over time. Indeed, there have been many attempts to propose stochastic models to describe the evolution of such processes. This paper is concerned with neither the dynamics nor the modelling, merely with the interpretation – and only so far as concentration aspects are concerned – of the distributions.

We begin with the graphical illustration of inequality via the Leimkuhler curve; then we consider the measurement of inequality using the Gini index, emphasising the simplest way of calculating it in practice; finally we investigate the use of the new measures of comparative concentration.

The data sets

All of the data sets used are ones that have previously appeared and been analysed in the literature, sometimes extensively and, in several cases, many times over. As we are only interested in the data themselves and not their context, let us very briefly describe the sets and their origins together with references that the reader can explore for further details.

(i) Applied Geophysics (hereafter referred to as AG).

This is one of the original data sets studied by Bradford (1934) that helped instigate the ongoing industry of proving, reformulating or illustrating the "Bradford law". We will not follow that well-worn path here.

(ii) Lubrication (Lub).

Bradford's other data set.

(iii) ORSA

This is a well-known bibliography on operational research introduced and analysed by Kendall (1960) as part of an attempt to rationalise mathematically the regularities observed by Bradford (1934).

Proceedings of ISSI 2005, pages 129–139.
Edited by P. Ingwersen and B. Larsen.

(iv) "Mast Cell" (Mast)

This is a bibliography covering a period of over 80 years, compiled by Selye (1968), and analysed by Goffman & Warren (1969, 1980).

(v) Schistosomiasis (Schi).

A bibliography covering a 110-year period, also presented by Goffman & Warren (1969, 1980).

(vi) Information Science (Pope).

This is an extensive bibliography on information science presented by Pope (1985).

(vii) Statistical Methods (Sachs).

These data have been taken from Egghe (1990) who derived them from Sachs (1986).

(viii) Wishart Library (Wish).

This is a very small data set relating to the number of loans of books from a University departmental library over a three-year period. The background is given in Burrell (1980).

(ix) Sussex University (Suss).

This is another library circulations data set reported by Burrell (1980), this time a large university library.

Note that all the sets (i) – (vii) are discussed in detail by Egghe (1990) where they are subjected to a Bradford-type analysis.

Graphical analysis – The Leimkuhler curve

Since in practical informetric applications both the number of sources and the number of items will be finite, let us write $g(j)$ for the number of sources producing j items, j = 0, 1, 2, …, n, where n is the largest observed productivity, N for the total number of sources, and M for the total number of items produced.

Then, $N = \sum_j g(j)$, $M = \sum_j jg(j)$

and mean number of items produced per source = M/N.

Definition/Notation

Tail distribution function = proportion of sources producing at least j items

$$= \Phi(j) = \left(\sum_{k\geq j} g(k)\right) / N,$$

Tail-moment distribution function = proportion of items accounted for by those sources producing at least j items

$$= \Psi(j) = \left(\sum_{k\geq j} kg(k)\right) / M$$

Of course, since we only have a finite number of sources, and a maximum observed productivity, n, the above sums are in fact from j to n and note that $\Phi(j) = \Psi(j) = 0$ for j > n.

Note: As has been pointed out before, by e.g. Burrell (1991, 1992b, 2005b), using the standard convention for ranking sources in decreasing order of production,

$N\Phi(j) = \sum_{k\geq j} g(k) =$ number of sources producing at least j items

= rank of a source producing j items
= $r(j)$, say, for $j = 0, 1, 2, \ldots , n$,

so that the tail distribution function can be thought of as the normalised form of the rank, r, as used in Egghe (1990).

Similarly $M\Psi(j) = \sum_{k\geq j} kg(k)$ = cumulative number of items produced by sources producing at least j items each
= R(j), say, for $j = 0, 1, 2, \ldots, n$

so that the tail-moment distribution is the normalised form of R in Egghe (1990).

Note. (i) It is important to note both here and later that, in the above, "rank" can be defined for each productivity j, whether or not there is a source with this productivity.

(ii) The original Bradford (1934) graphical analysis (see also e.g. Egghe (1990)) is essentially a plot of $\Psi(j)$ against $\log\Phi(j)$, see Burrell (1992b), compared with the plot of $\Psi(j)$ against $\Phi(j)$ which we will consider here and gives what we term the Leimkuhler curve. This is a well-known method of representing informetric data sets graphically. It is closely connected with the well-known Lorenz curve from econometrics; see Burrell (1991, 1992b,c, 2005c).

The construction of the Leimkuhler curve is straightforward, the required calculations being illustrated in Table 1 for the Wishart library data – this being the smallest set considered. In this particular context, "items" correspond to "loans" and "sources" to "books". The r(j) and R(j) columns are found by cumulating from the bottom the g(j) and jg(j) columns respectively.

Table 1. Calculation of Φ and Ψ functions for the Wishart data.

No. items, j	**No. sources, g(j)**	**r(j)**	**Φ(j)**	**jg(j)**	**R(j)**	**Ψ(j)**
1	65	N=122	1.000	65	M=243	1.000
2	26	57	00.467	52	178	0.733
3	12	31	0.254	36	126	0.519
4	10	19	0.156	40	90	0.370
5	5	9	0.074	25	50	0.206
6	3	4	0.033	18	25	0.103
7	1	1	0.008	7	7	0.029
>7	0	0	0.000	0	0	0.000

Notice that what we really construct is a plot of distinct points, as in Figure 1.

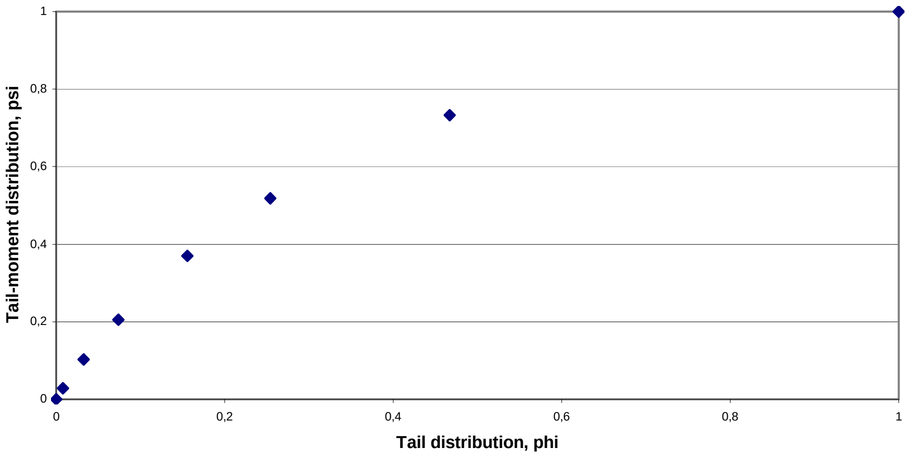

Figure 1. Basic Leimkuhler plot

The standard presentation of the Leimkuhler curve would then convert this into a simple polygonal plot by joining the plotted points with straight lines. Instead, since we are only concerned with the overall visual presentation, and comparisons between different plots, in Figure 2 (a) – (i) we have suppressed the original data points and merely give "smooth curve" approximations of the polygonal plots.

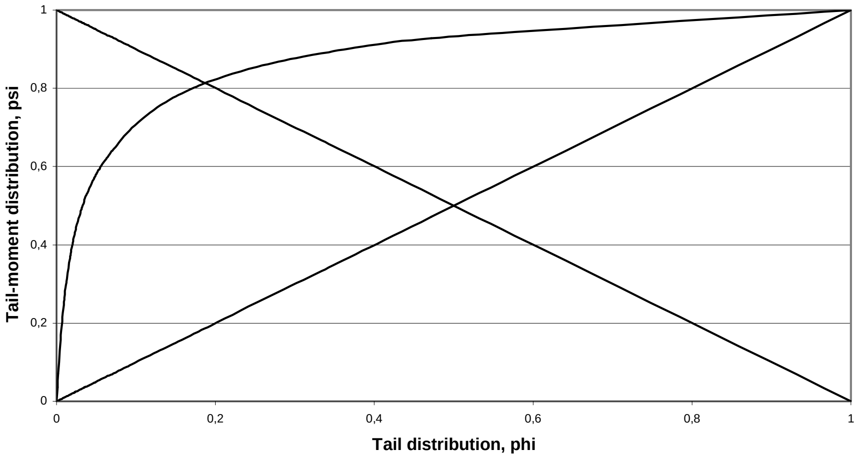

Figure 2(a). Leimkuhler curve, Pope

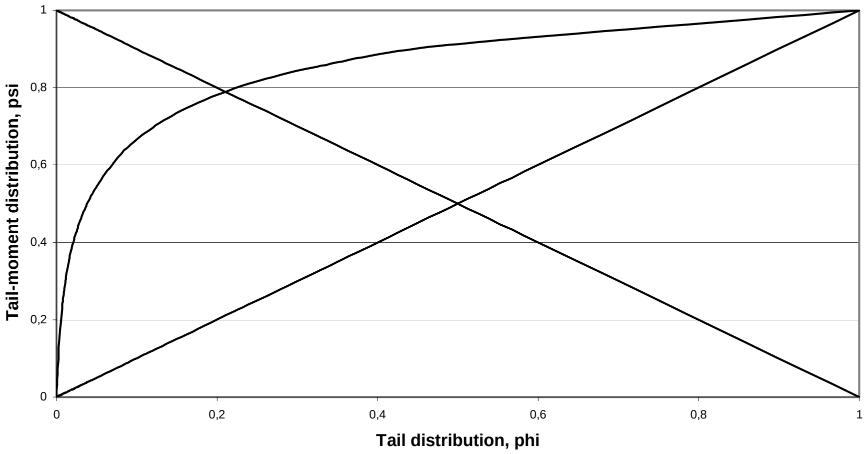

Figure 2(b). Leimkuhler curve, Schi

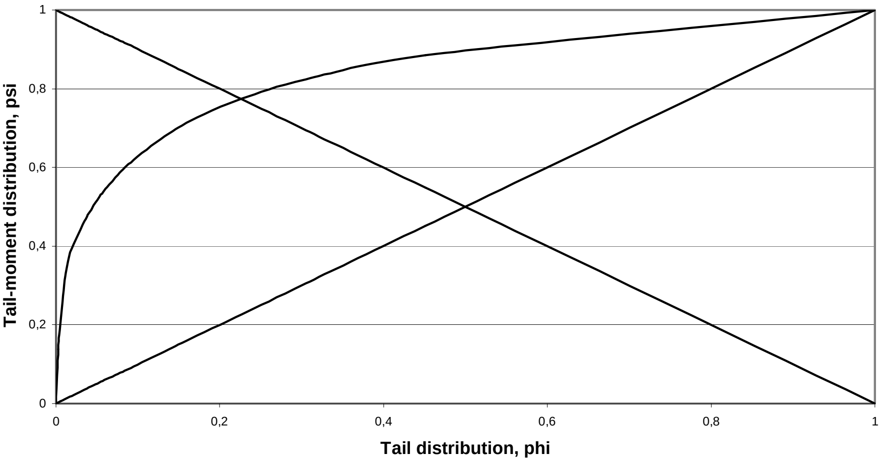

Figure 2(c). Leimkuhler curve, ORSA

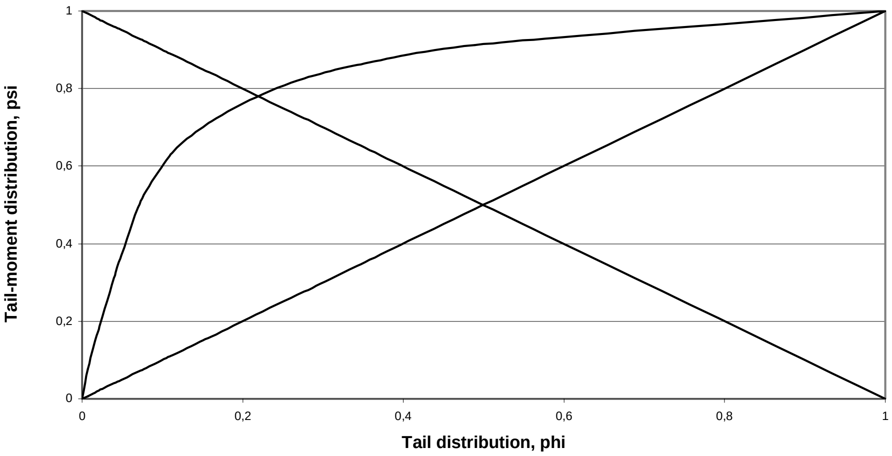

Figure 2(d). Leimkuhler curve, Sachs

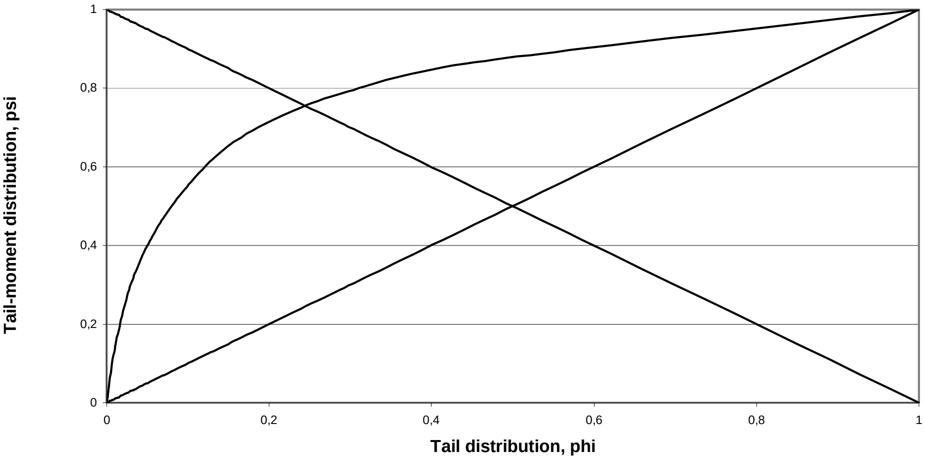

Figure 2(e). Leimkuhler curve, Mast

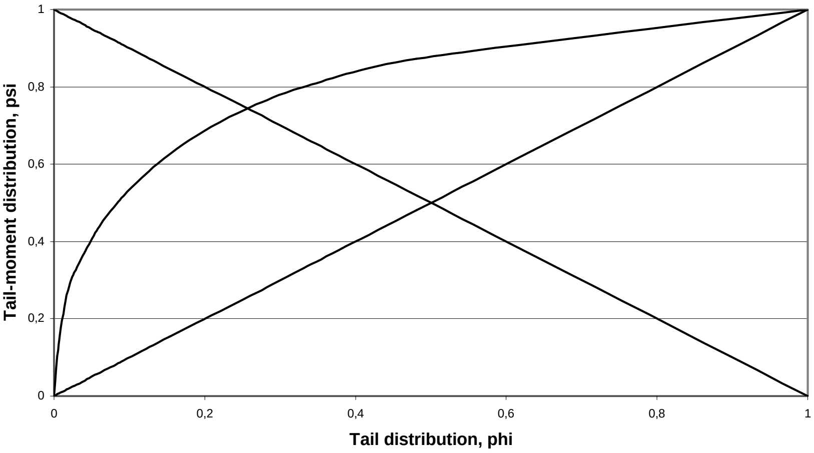

Figure 2(f). Leimkuhler curve, AG

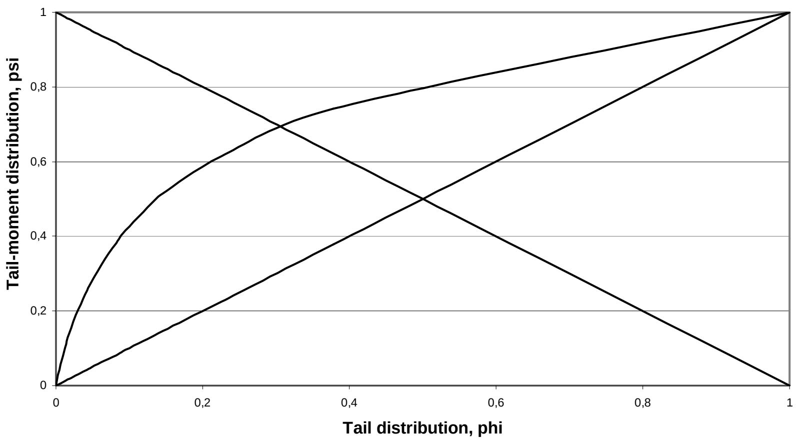

Figure 2(g). Leimkuhler curve, Lub

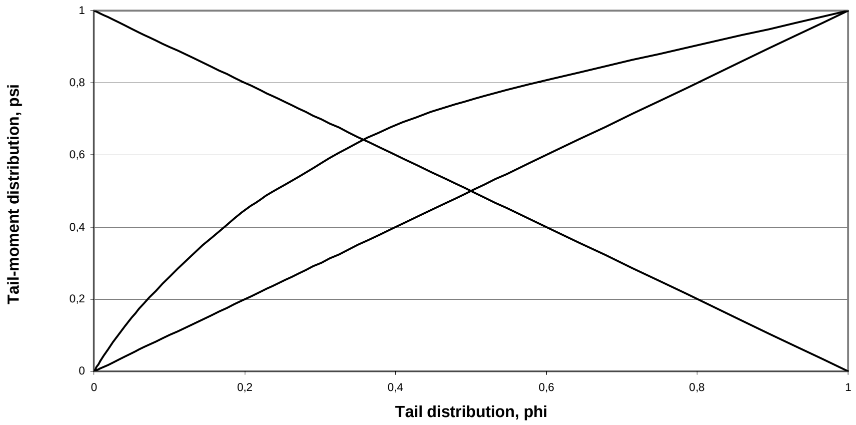

Figure 2(h). Leimkuhler curve, Wish

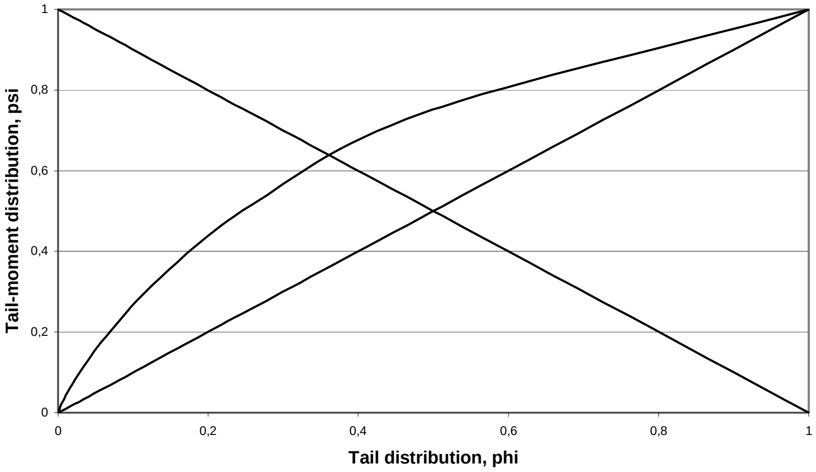

Figure 2(i). Leimkuhler curve, Suss

Note that we have ordered these graphs not as in our original listing of the data sets but according to their general geometric form. Recall that in a situation where all sources are equally productive, the Leimkuhler curve would be of the form Ψ = Φ, corresponding to the main diagonal in these graphs, whereas at the other extreme where all productivity is within a single source, the graph would consist of the vertical line through Φ = 0 and the horizontal through Ψ = 1. Indeed, it is well known (Burrell, 1991) that in general the standard Gini coefficient of concentration (see next section) is given geometrically as twice the area between the Leimkuhler curve and the main diagonal Φ = Ψ. Hence

our ordering has been chosen so that the data sets are given in decreasing order of concentration (as measured by the Gini coefficient) as we move through from (a) to (i).
In Figure 2 (a) – (i) we have also given the other diagonal corresponding to Φ + Ψ = 1 to illustrate that, at least in the cases (a) – (g), most of the contribution to the Gini coefficient comes from the lower end of the curve, which of course corresponds to the most productive sources. In other words, in these examples, productivity is concentrated in the most productive sources.

As an aside, it is interesting to note that all of the curves (a) – (f) pass fairly close to the intersection of the line Ψ = 0.8 and the line Φ + Ψ =1, which intersection indicates the ubiquitous 80/20 rule, see Trueswell (1969) and Burrell (1985). (Some of the later cases indicate something more like a 75/25 rule, an accepted variant.)

Remark. It might appear that what we have is a sequence of curves in which each one "dominates" – i.e. always lies above – the next; in fact there are cases where the curves intersect. In such cases a graphical approach is not totally adequate in itself. We shall not discuss the case of intersecting Leimkuhler curves further but refer the reader to Fellman (1976), Rousseau (1992), Burrell(1992a, 2005c) and Lambert(2001) for more details.

Numerical analysis – The Gini coefficient

The Gini coefficient is usually held to be one of the, if not *the*, best inequality measures in that it obeys all seven of the "desirable" properties proposed by Dalton (1920) for such a measure, see Dagum (1983). One of these properties is that it is invariant under scale, or is independent of the unit of measurement. (Note that the Leimkuhler curve also has this scale-invariance property.) This is clearly almost a necessary property in measuring inequality within a population; for instance in measuring the inequality of incomes within a population it should clearly be immaterial whether income is measured in £stg or Euros. There are many different ways of formulating and calculating the Gini coefficient (Yitzhaki, 1998) of concentration. The one that seems easiest to use for the purposes of calculation with the types of distributions we are considering is given by:

$$\text{Gini coefficient} = \gamma = 1 - \frac{\sum_{j=1}^{n} r(j)^2}{NM} = 1 - \frac{\sum_{j=1}^{n} \Phi(j)^2}{M/N} \qquad (1)$$

See Arnold & Laguna (1977), Dorfman (1979), Burrell (1991, 1992a, 2005a,b), Yitzhaki (1998) and Kleiber & Kotz (2003, p30).

We find it useful to also consider also

$$\text{Coefficient of equality} = \theta = 1 - \gamma = \frac{\sum_{j=1}^{n} r(j)^2}{NM} = \frac{\sum_{j=1}^{n} \Phi(j)^2}{M/N} \qquad (2)$$

(See Burrell, 2005a,b.)

It is important to stress that in both (1) and (2) above, the summation is over all values of j from 1 to n – i.e. the maximal observed production - including those j for which g(j) = 0.

The summary statistics are given in Table 2 for all the data sets, in decreasing order of γ. The ordering of the sets reflects that given in the graphical presentations in Figure 2. Note that there is a weak relationship between the mean productivity and the Gini coefficient of the data sets, illustrating that these address related but different features.

Table 2. Summary statistics for the individual data sets.

Source	n	N	M	Mean	∑r²	γ	θ
Pope	261	1011	7368	7.288	1794480	0.7591	0.2409
Schi	325	1738	9914	5.704	4864540	0.7177	0.2823
ORSA	242	370	1763	4.765	205023	0.6857	0.3143
Sachs	64	143	850	5.790	38402	0.6841	0.3159
Mast	66	587	2378	4.051	519586	0.6277	0.3723
AG	93	326	1332	4.086	166066	0.6176	0.3824
Lub	22	164	395	2.409	33933	0.4762	0.5238
Wish	7	122	243	1.992	19553	0.3404	0.6596
Suss	14	18854	37877	2.009	471802611	0.3393	0.6607

It is interesting to note that while construction of the Leimkuhler curve requires both Φ and Ψ, or r and R, calculation of the Gini coefficient γ only requires Φ, or r.

The co-concentration coefficient

The two measures introduced by Burrell (2005a) are both based upon the Gini ratio, defined by Dagum (1987). The idea behind the construction of the Gini ratio between two populations is exactly analogous to that of the Gini coefficient for a single population, namely we look at pairs of sources, but now one from each population, find the absolute difference between their productivities and average this difference over all possible pairs. Details of the construction, and methods of calculation are given in Burrell (2005a). Here we just repeat the formulae required for empirical studies.

When comparing two populations or data sets, let us distinguish between them by the use of suffices X and Y.

Definition. Assuming, without loss of generality, that $n_X \le n_Y$, the *empirical Gini ratio* is given by

$$G(X,Y) = 1 - \frac{2\sum_{j=1}^{n_X} r_X(j) r_Y(j)}{N_X M_Y + M_X N_Y} \tag{3}$$

The summation terminates at $j = n_X$ since $r_X(j) = 0$ for all larger j, but again we stress that the summation is over all j from 1 to n_X. Note that for evaluation of this ratio, we need to determine the sum of rank products. An illustration of the calculation is given by Burrell (2005a).

Although the Gini ratio gives some sort of measure of the degree of similarity/dissimilarity between two productivity distributions so far as their concentration/inequality is concerned, it is not very informative on its own. One problem is that the ratio is minimised when the two distributions are the same whereas we would like a comparative measure to be *maximised* in this situation. This is easily resolved if we instead of (3) we use:

$$H(X,Y) = 1 - G(X,Y) = \frac{2\sum_{j=1}^{n_X} r_X(j) r_Y(j)}{N_X M_Y + M_X N_Y} \tag{4}$$

The other major problem for comparative studies is that H is not normalised. To address this, Burrell (2005a) proposes the following:

Definition. The *coefficient of co-concentration* or *co-concentration coefficient* is given by

$$Q(X,Y) = \frac{H(X,Y)}{\sqrt{\theta_X \theta_Y}} = \frac{(1 - G(X,Y))}{\sqrt{(1-\gamma_X)(1-\gamma_Y)}} \tag{5}$$

Then we have
Theorem (Burrell, 2005a, Theorem 2)
Q(X,Y) is a normalised measure in that 0 < Q(X,Y) < 1 and we get Q(X,Y) = 1 if and only if the (probability) distributions of the two populations are the same.

Aside. If we combine the expression (4) for H(X,Y) above with, from (2)

$$\theta_X\theta_Y = (1-\gamma_X)(1-\gamma_Y) = \frac{\left(\sum_1^{n_X} r_X(j)^2\right)\left(\sum_1^{n_Y} r_Y(j)^2\right)}{N_X M_X N_Y M_Y} \tag{6}$$

we find the empirical Q-measure as

$$Q(X,Y) = \frac{2\sqrt{N_X M_X N_Y M_Y}}{N_X M_Y + M_X N_Y}\frac{\sum r_X r_Y}{\sqrt{\left(\sum r_X^2\right)\left(\sum r_Y^2\right)}} = \frac{2\sqrt{\overline{X}\overline{Y}}}{(\overline{X}+\overline{Y})}\frac{\sum r_X r_Y}{\sqrt{\left(\sum r_X^2\right)\left(\sum r_Y^2\right)}} \tag{7}$$

Although (7) gives a single formula for Q, for purposes of practical computation – particularly if carried out "by hand" – it is more convenient to calculate H(X,Y), θ_X and θ_Y separately and then substitute from (4) and (6) into the expression (5) for Q(X,Y) as given in the definition. Note that in the calculation of Q(X,Y), the only new quantity that needs to be computed is the sum of the rank cross-product terms, i.e. the $\sum_{j=1}^{n_X} r_X(j)r_Y(j)$, where in each case the summation is from j = 1 to the smaller of n_X and n_Y.

Note. In practical cases it has been argued in Burrell (2005b) that, because when comparing distributions of similar general form – such as the long right-tailed distributions frequently encountered in informetrics – the Q value can be very close to one, it is more informative to instead consider Q^2. This is analogous to using the coefficient of determination rather than the standard correlation coefficient in regression analysis and this is the way we proceed in the following. The Q^2 matrix is given for all the data sets in Table 3.

The calculated values of Q^2 in Table 3 vary from 0.549 to 0.999 – of course Q^2(X,X) = 1 in all cases, but some general patterns and features can be discerned. The tendency is for values being higher the closer are the values of the Gini coefficient, smaller values corresponding to pairs whose Gini coefficients are further apart. Note also that for a pair of data sets having similar Gini coefficients – look in particular at Mast & AG and at Wish & Suss – the rows of Q^2 values are very similar, in other words each member of the pair has the same sort of co-concentration with each of the other data sets.

Table 3. The Co-Concentration or Q^2 matrix.

Source	Pope	Schi	ORSA	Sachs	Mast	AG	Lub	Wish	Suss	γ
Pope	1.000	0.975	0.963	0.963	0.899	0.883	0.657	0.549	0.552	0.7591
Schi	0.975	1.000	0.986	0.966	0.962	0.961	0.800	0.676	0.679	0.7177
ORSA	0.963	0.986	1.000	0.928	0.986	0.989	0.859	0.763	0.765	0.6857
Sachs	0.963	0.966	0.928	1.000	0.924	0.908	0.697	0.587	0.589	0.6841
Mast	0.889	0.962	0.986	0.924	1.000	0.992	0.895	0.795	0.797	0.6277
AG	0.883	0.961	0.989	0.908	0.992	1.000	0.896	0.808	0.810	0.6176
Lub	0.657	0.800	0.859	0.697	0.895	0.896	1.000	0.965	0.964	0.4762
Wish	0.549	0.676	0.763	0.587	0.795	0.808	0.965	1.000	0.999	0.3404
Suss	0.552	0.679	0.765	0.589	0.797	0.810	0.964	0.999	1.000	0.3393

The relative concentration coefficient
This measure, an adaptation of Dagum's relative economic affluence (Dagum, 1987), is also based on the Gini ratio and its empirical form is given by (Burrell, 2005a,b)

$$D(X,Y) = \frac{\overline{X} - \left(\sum r_X r_Y\right)/N_X N_Y}{\overline{Y} - \left(\sum r_X r_Y\right)/N_X N_Y} = \frac{M_X N_Y - \sum r_X r_Y}{N_X M_Y - \sum r_X r_Y} \tag{8}$$

where we have written $\overline{X} = \frac{M_X}{N_X}$ and $\overline{Y} = \frac{M_Y}{N_Y}$ for the two sample mean productivities and have assumed, wlog that $\overline{Y} \geq \overline{X}$. (The ranges of summation in the above are as before.)

It can be shown (Burrell, 2005a, Proposition 5, Corollary) that the relative concentration coefficient is also normalized in that it lies between 0 and 1 and the upper bound is achieved if and only if the two means are the same. In fact, as pointed out in Burrell (2005b), if the two means are not the same then the upper bound is given by the ratio of the smaller to the larger. To get around this objection, we instead consider the (modified) relative concentration coefficient proposed by Burrell (2005b).

$$D^*(X,Y) = \frac{\overline{Y}}{\overline{X}} D(X,Y) = \frac{1 - (\sum r_X r_Y)/N_Y M_X}{1 - (\sum r_X r_Y)/N_X M_Y} \tag{9}$$

where we are still assuming wlog that $\overline{Y} \geq \overline{X}$, so that $N_X M_Y \geq N_Y M_X$.

The D^* matrix of relative concentration coefficients is given in Table 4. Note that in the construction of this matrix we have ordered the data sets in decreasing order of mean productivity, reflecting the crucial role of the mean in the construction of the coefficient. Again we can note the general features of the matrix – in this case more clear cut than for the Q^2-matrix. The further apart are the means, the smaller is the D^*-value. Also, pairs of data sets with similar mean productivities, e.g. Sachs & Schi, AG & Mast and Suss & Wish, have very similar rows of coefficients, so very similar relative concentrations with each of the other data sets.

Table 4. The relative concentration or D^* matrix

Source	Pope	Sachs	Schi	ORSA	AG	Mast	Lub	Suss	Wish	Mean
Pope	1.000	0.926	0.917	0.855	0.776	0.775	0.520	0.388	0.384	7.288
Sachs	0.926	1.000	0.983	0.907	0.824	0.821	0.540	0.400	0.396	5.790
Schi	0.917	0.983	1.000	0.927	0.849	0.848	0.565	0.417	0.413	5.704
ORSA	0.855	0.907	0.927	1.000	0.922	0.919	0.614	0.458	0.453	4.765
AG	0.776	0.824	0.849	0.922	1.000	0.995	0.647	0.474	0.469	4.086
Mast	0.775	0.821	0.848	0.919	0.995	1.000	0.659	0.497	0.492	4.051
Lub	0.520	0.540	0.565	0.614	0.647	0.659	1.000	0.776	0.766	2.409
Suss	0.388	0.400	0.417	0.458	0.474	0.497	0.776	1.000	0.983	2.009
Wish	0.384	0.396	0.413	0.453	0.469	0.492	0.766	0.983	1.000	1.992

Remark. Note that although both the co-concentration and relative concentration coefficients involve the rank products through $\sum r_X r_Y$, the other quantitative information required to calculate D^* is considerably less than that for Q^2. For instance the calculation of Q^2 requires the sum of squared ranks while D^* does not. This reflects the fact that D^* is primarily based upon the relative means of the distributions while Q^2 is based upon their relative concentrations.

Concluding remarks

We have tried to illustrate several different but related approaches to measuring the similarity, or otherwise, of concentration between different informetric data sets; the main thrust being the empirical analysis of comparative studies. The results reported here are certainly not definitive but we would hope to have encouraged others to make use of the methods to increase understanding of the appropriate interpretation of the measures through further empirical work. Two obvious areas of possible application would be the investigation of

(i) differences between different subject areas, possibly as a complement to consideration of differences of impact factors

(ii) differences over different time periods, either for year-by-year changes or for the evolutionary distributional change over extending periods of time.

As a final thought, although our examples have been in the familiar context of sources producing items, there is no reason why the same sort of approach could not be adopted in other informetric contexts such as citation age distributions. Again, investigation of differences between different subject areas should be of interest.

References

Arnold, B. C. (1987). *Majorization and the Lorenz order*. Lecture Notes in Statistics, 43. Berlin & New York: Springer.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, 137, 85-86.

Burrell, Q. L. (1985) The 80/20 rule: library lore or statistical law? *Journal of Documentation*, 41, 24-39.

Burrell, Q. L. (1991). The Bradford distribution and the Gini index. *Scientometrics*, 21, 181-194.

Burrell, Q . L. (1992a). A note on a result of Rousseau for concentration measures. *Journal of the American Society for Information Science*, 43, 452-454.

Burrell, Q. L. (1992b). The dynamic nature of bibliometric processes: a case study. In I. K. Ravichandra Rao (Ed.), *Informetrics – 91: selected papers from the Third International Conference on Informetrics* (pp. 97-129), Bangalore: Ranganathan Endowment.

Burrell, Q. L. (1992c). The Gini index and the Leimkuhler curve for bibliometric processes. *Information Processing and Management*, 28, 19-33.

Burrell, Q. L. (2005a). Measuring similarity of concentration between different informetric distributions: Two new approaches. *Journal of the American Society for Information Science and Technology*, 56, 704-714.

Burrell, Q. L. (2005b). Measuring relative equality of concentration between different income/wealth distributions. *International Conference in memory of two Social Scientists: C. Gini and M. O. Lorenz.* University of Siena, 23-26 May, 2005.

Burrell, Q. L. (2005c). Symmetry and other transformation features of Lorenz/Leimkuhler representations of informetric data. *Information Processing and Management*. (To appear.)

Dagum, C. (1983). Income inequality measures. In S. Kotz & N. S. Johnson (Eds.), *Encyclopaedia of Statistical Sciences, Volume 4* (pp. 34-40), New York: Wiley.

Dagum, C. (1987). Measuring the economic affluence between populations of income receivers. *Journal of Business and Economic Statistics*, 5, 5-11.

Dalton, H. (1920). The measurement of inequality of incomes. *Economic Journal*, 30, 348-361.

Dorfman, R. (1979). A formula for the Gini coefficient. *Review of Economics and Statistics*, 61, 146-149.

Egghe, L. (1990). Applications of the theory of Bradford's law to the calculation of Leimkuhler's law and to the completion of bibliographies. *Journal of the American Society for Information Science*, 41, 469-492.

Fellman, J. (1976). The effect of transformations on Lorenz curves. *Econometrica*, 44, 823-824.

Kleiber, C. & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. New Jersey: Wiley.

Lambert, P. J. (2001). *The distribution and redistribution of income. 3rd edition.* Manchester: Manchester University Press.

Rousseau, R. (1992). Concentration and diversity of availability and use in information systems. *Journal of the American Society for Information Science*, 43, 391-395.

Trueswell, R. W. (1969). Some behavioral patterns of library users: the 80/20 rule. *Wilson library bulletin*, 43, 458-461.

Yitzhaki, S. (1998). More than a dozen alternative ways of spelling *Gini*. *Research on Income Inequality*, 8, 13-30.