TOP-Curves

Leo Egghe
Hasselt University, Agoralaan, B-3590 Diepenbeek, Belgium and Antwerp University, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium. E-mail: leo.egghe@uhasselt.be

Ronald Rousseau
KHBO (Association K.U. Leuven), Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium; Hasselt University, Agoralaan, B-3590 Diepenbeek, Belgium; and Antwerp University, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium. E-mail: ronald.rousseau@khbo.be

Sandra Rousseau
K.U. Leuven, Centre for Economic Studies, Naamsestraat 69, B-3000 Leuven, Belgium. E-mail: sandra.rousseau@econ.kuleuven.be

Several characteristics of classical Lorenz curves make them unsuitable for the study of a group of top-performers. TOP-curves, defined as a kind of mirror image of TIP-curves used in poverty studies, are shown to possess the properties necessary for adequate empirical ranking of various data arrays, based on the properties of the highest performers (i.e., the core). TOP-curves and essential TOP-curves, also introduced in this article, simultaneously represent the incidence, intensity, and inequality among the top. It is shown that TOP-dominance partial order, introduced in this article, is stronger than Lorenz dominance order. In this way, this article contributes to the study of cores, a central issue in applied informetrics.

Introduction

Lorenz curves were introduced in 1905 as a graphical device to show intrinsic inequality among a set of sources (Lorenz, 1905). These sources can be persons (as in the original use of the Lorenz curve), actors (a terminology often used in social network analysis), performers, authors, articles, and so on (as a case in point see, Egghe, 2005). Since its introduction, it has become clear that this is a very powerful device that can be used in many fields and for many applications. Examples include income distributions (Kleiber & Kotz, 2003; Lambert, 2001), plant-size inequality (Weiner, 1985), evenness studies in ecology (Nijssen et al., 1998), vegetation studies based on satellite images (Bogaert, Zhou, Tucker, Myneni, & Ceulemans, 2002), and research evaluation (Rousseau, 1998). Besides the original form, variations of the Lorenz curve have been proposed to study poverty (Jenkins & Lambert, 1997; Zheng, 2000), ecological diversity (Patil & Taillie, 1979), hierarchies (Egghe, 2002), own-group preference (Egghe & Rousseau, 2004), and overlap (Egghe & Rousseau, 2006). Note that the Lorenz curve is a real part of classical bibliometrics because as shown by Burrell (1991, 1993), the Bradford distribution as derived by Leimkuhler (1967) is actually a theoretical form of the Lorenz curve. In the context of Lotkaian informetrics, the further elaboration of this observation occupies the larger part of chapter 4 in Egghe's (2005) book on power laws in the information production process.

If $X = (x_1, x_2, \ldots, x_N)$ denotes an ordered array of the productions of N sources, then its classical Lorenz curve is denoted as $L(X)$ or L_X. Note that we prefer the term "array" for this type of n-tuple as it is not a vector in the strict sense; that is, it cannot be multiplied by an arbitrary real number (including negative numbers) and still be of the same type. The Lorenz curve of X starts in the origin, and connects consecutive points of the form

$$\left(\frac{j}{N}, \frac{\sum_{k=1}^{j} x_k}{\sum_{k=1}^{N} x_k} \right),$$

where $x_1 \geq x_2 \geq \cdots \geq x_N \geq 0$, and $j = 1, \ldots, N$. It always ends in the point with coordinates $(1, 1)$. In this context, arrays X and Y, with components $(x_j)_{j=1,\ldots,N}$ and $(y_j)_{j=1,\ldots,M}$—not necessarily of the same length—are said to be equivalent if they have the same Lorenz curves. This happens if the

Received December 14, 2005; revised June 1, 2006; accepted June 1, 2006

 • Published online 1 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20539

components of X and Y are permutations of one another; if there exists a positive value c such that $x_j = cy_j$, $j = 1, \ldots, N$; if $Y = \text{REPEAT}_r(X)$, for some $r \in N_0$, where $\text{REPEAT}_r(X)$ denotes the array (Rousseau, 1992, p. 112):

$$\text{REPEAT}_r(X) = \left(\underbrace{x_1, \ldots, x_1}_{r\ times}, \underbrace{x_2, \ldots, x_2}_{r\ times}, \ldots, \underbrace{x_N, \ldots, x_N}_{r\ times}\right)$$

or if X and Y are related through a finite sequence of these operations. The property that Y and $X = cY$, $c > 0$, have the same Lorenz curve is called scale invariance while the property that $Y = \text{REPEAT}_r(X)$ and X have identical Lorenz curves is referred to as replication invariance (Dalton, 1920).

The Lorenz curve is the basis of a partial order among arrays of finite length, referred to as Lorenz dominance order. In this partial order, X Lorenz dominates Y, denoted as $X \geq Y$, if the Lorenz curve of X is situated above or on the Lorenz curve of Y: $L_X(t) \geq L_Y(t)$, with strict inequality in at least one point $t \in]\,0, 1[$, and hence in infinitely many. It is well known (Dalton, 1920) that if $x_k \leq x_j$ and $x_k \geq h > 0$ then $L_X < L_{X'}$, where $X = (x_1, \ldots, x_j, \ldots, x_k, \ldots, x_N)$ and $X' = (x_1, \ldots, x_j + h, \ldots, x_k - h, \ldots, x_N)$. Lorenz dominance order is therefore said to satisfy the transfer principle.

Although scale invariance, resulting from the fact that Lorenz curves are constructed using relative values, is a basic property of these curves, it is for some studies better to work with absolute values, such as when studying citations received by persons or journals, or income distributions among countries of a totally different nature, such as between highly developed and extremely poor countries. The latter observation led Shorrocks (1983) to introduce so-called generalized Lorenz curves. These generalized Lorenz curves are constructed by multiplying each ordinate of the classical Lorenz curve of X by the average production of the array: $\mu_X = \dfrac{\sum_{j=1}^{N} x_j}{N}$. In addition, the generalized Lorenz curve starts in the origin; it further connects consecutive points of the form $\left(\dfrac{j}{N}, \dfrac{\sum_{k=1}^{j} x_k}{N}\right)$, where $x_1 \geq x_2 \geq \cdots \geq x_N \geq 0$, and $j = 1, \ldots, N$. It always ends in the point with coordinates $(1, \mu_X)$. This curve will be denoted as GL_X.

In socioeconomic studies, scientists have often focused their attention on people with the lowest income. They introduced the notion of poverty line, a threshold line or value such that if someone's income falls below this threshold income, this person is considered to live in poverty. Inequality among the poor with respect to the whole situation (e.g., the whole country) is then studied by an adaptation of Shorrocks' generalized Lorenz curves: the so-called TIP-curves (Jenkins & Lambert, 1997) or absolute rotated Lorenz curves (Spencer & Fisher, 1992).

In information science and, in particular, in research evaluation studies, however, evaluators and decision makers are usually more interested in the most productive sources rather than in the low producers. For this reason, we introduce the opposite of TIP-curves, called TOP-curves, as they are designed to study the most productive sources. Similarly to the poverty line, we introduce the notion of top line or top value. This is a threshold, denoted as "t," separating the top from the rest.

TOP-Curves

Important notions when studying the most productive sources are *incidence*, *intensity*, and *inequality among the top*. These notions will be explained further.

Incidence

Given a top line, the notion of incidence is defined as the percentage of the population that belongs to the top group.

Preliminary Constructions

TOP-curves can be considered as generalized Lorenz curves (Shorrocks, 1983), specially designed to study the most productive sources in an information production process (IPP). We will show that TOP-curves portray simultaneously the incidence, intensity, and inequality among the top (see Figure 1).

Let $X = (x_1, x_2, \ldots, x_N)$ denote an array of productions of N sources, where $x_j \geq 0$, for $j = 1, \ldots, N$, denotes the (generalized) production of the j-th source. We assume that sources are ranked in decreasing order. Let $t > 0$ be the top line. The TOP array of X, given t, is then defined as

$$T_X = (\max(x_j, t))_{j=1}^{N} \tag{1}$$

This definition implies that if, for all j, $x_j \leq t$, then $T_X = T = (t, \ldots, t)$ and if, for all j, $x_j \geq t$, then $T_X = X$. These two cases will be referred to as trivial cases. Otherwise, as X is ordered in decreasing order, there exists an index $j_0 \in \{1, \ldots, N-1\}$

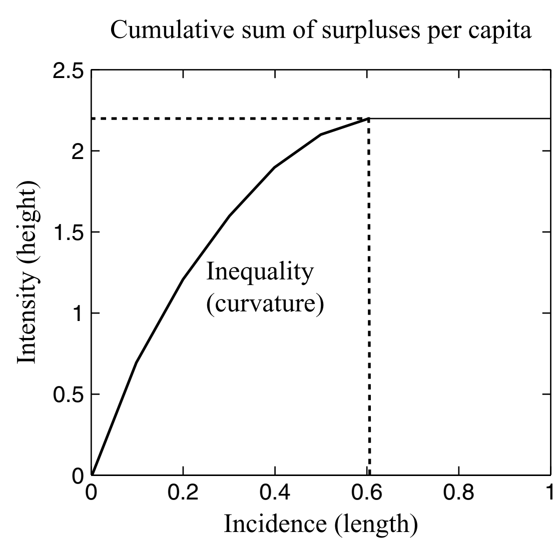

FIG. 1. General TOP-curve: in this example incidence $= j_0/N = 0.6$.

DOI: 10.1002/asi

such that $(T_X)_j = x_j > t$ for $1 \le j \le j_0$ and $(T_X)_j = t$ for $j_0 < j \le N$. From now on, we will always assume that we do not have a trivial case and, hence, that such an index j_0 exists.

Next, we construct the surplus array, denoted as S_X, by taking the difference $T_X - T$. Observe that this array is automatically ranked in decreasing order:

$$\begin{aligned} &S_X(j) = x_j - t > 0 \ \text{ for } 1 \le j \le j_0 \ \text{ and} \\ &S_X(j) = 0 \ \text{ for } j_0 + 1 \le j \le N \end{aligned} \tag{2}$$

The last $N - j_0$ components of this array are zeros. The array S_X will sometimes be denoted as $S_X[j_0]$, particularly when it is necessary to stress the fact that for $1 \le j \le j_0$, components of $S_X[j_0]$ are non-zero.

Surplus arrays are essentially "partial vectors" as studied by Egghe (2002) and Egghe and Rousseau (2005). More precisely, they are "partial vectors" on Level j_0. In socioeconomics and statistics, such arrays are often referred to as censored arrays (Zheng, 2000).

We are now able to define the notion of intensity.

Intensity

The intensity of the top sources of Array X is equal to the total surplus sum divided by N:

$$\frac{\sum_{k=1}^{N} S_X(k)}{N} = \frac{\sum_{k=1}^{j_0} (x_k - t)}{N} \tag{3}$$

The intensity, being the average surplus, is clearly a measure characterizing the "power" of the top sources.

TOP-Curve

The TOP-curve of $X = (x_1, x_2, \ldots, x_N)$, defined after Jenkins and Lambert's (1997) TIP-curve, plots against $p = k/N$, $1 \le k \le N$, the sum of the first $k/N\ S_X$ values, divided by N. This TOP-curve is denoted as $\mathrm{TOP}_X(p)$, $0 \le p \le 1$. More precisely:

$$\mathrm{TOP}_X(0) = 0;\ \ \mathrm{TOP}_X(k/N) = \frac{\sum_{j=1}^{k} S_X(j)}{N}, \ \text{ for } \ k = 1, \ldots, N;$$

at intermediate points $\mathrm{TOP}_X(p)$ is determined by linear interpolation. Figure 1 illustrates this concept. Clearly, $\mathrm{TOP}_X(p)$ is, by definition, a concavely increasing curve. The part before the horizontal one is similar to a (generalized) Lorenz curve and reflects the inequality among the most productive sources.

Essential TOP-Curve

In many practical examples, the group of top sources constitutes a small minority. This would lead to a TOP-curve consisting for a large part of a horizontal line. For this reason, we propose the *essential* TOP-curve, which is identical to the TOP-curve, but shows only the part before this horizontal line. All essential characteristics can still be seen on this essential TOP-curve, as shown in Figure 2.

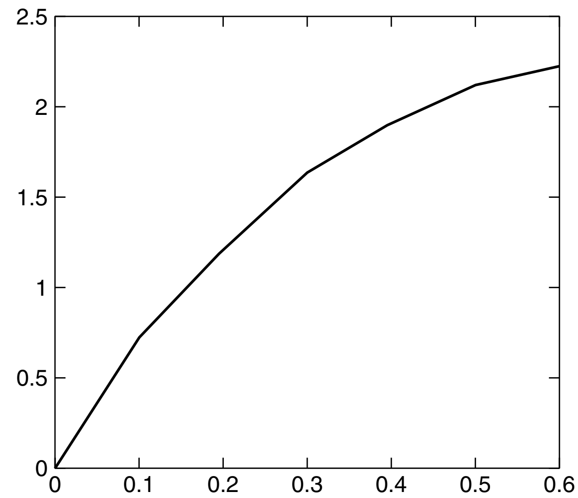

FIG. 2. Essential TOP-curve derived from Figure 1.

TOP-Equivalence

Arrays, not necessarily with the same number of sources but with the same TOP-curve, are said to be TOP-equivalent. If the context is clear, we will simply say "equivalent." If two N-arrays differ only in the components with values lower than or equal to the threshold line, then they are equivalent. Two N-arrays which consist of the same components (i.e., they are permutations of one another) also are equivalent, as TOP-curves are based on components ranked in decreasing order. By its definition, it immediately follows that the TOP-curves of X and $Y = \mathrm{REPEAT}_r(X)$ are identical, or stated otherwise, that X and $Y = \mathrm{REPEAT}_r(X)$ are TOP-equivalent. Hence, TOP-curves are replication invariant.

Based on the notions of Lorenz dominance and TIP-dominance, we introduce the notion of TOP-dominance.

TOP-Dominance

Array X TOP-dominates Array Y if $\mathrm{TOP}_X(p) \ge \mathrm{TOP}_Y(p)$ for all $p \in [0, 1]$.

Strict TOP-Dominance

Array X strict TOP-dominates Array Y if $\mathrm{TOP}_X(p) \ge \mathrm{TOP}_Y(p)$ for all $p \in [0, 1]$, where this inequality is strict for at least one p, and hence for infinitely many.

TOP-dominance determines a partial order in the space of finite arrays, given a Threshold Line t. The smallest element in this partially ordered set corresponds to the equivalence class of arrays where all components are smaller than or equal to the Threshold t.

TOP-curves can be interpreted as reflecting the general sense of dominance—the higher the TOP-curve, the higher the sense of dominance among the highest producers.

DOI: 10.1002/asi

Indeed, if intensity and incidence are given, the higher situated TOP-curve has more inequality among the top sources. Hence, intuitively, there is a higher sense of dominance. Further, a curve with a smaller intensity (and fixed incidence) or higher incidence value (and fixed intensity) can never TOP-dominate one with a higher intensity or smaller incidence value. So, intuitively, when there is only a small group of top sources (small incidence), they tend to dominate more. In addition, when the incidence is the same, the group with a larger intensity tends to dominate more.

TOP-Curves and Standard Lorenz Curves

The standard Lorenz curve of $S_X[j_0]$, denoted as $L_{S_X[j_0]}$ or for short L_{S_X}, is obtained from TOP_X by dividing each ordinate value by the intensity; that is, the total surplus sum divided by N. In particular:

$$L_{S_X}(k/N) = \frac{\sum_{j=1}^{k} S_j}{N} \cdot \frac{N}{\sum_{j=1}^{N} S_j} = \frac{\sum_{j=1}^{k} S_j}{\sum_{j=1}^{N} S_j}.$$

Lemma

If X and Y are N-arrays, if $L_{S_X} \geq L_{S_Y}$ for the classical Lorenz dominance relation, and if the total surplus of X is larger than the total surplus of Y, then X TOP-dominates Y. This is a trivial consequence of the definitions.

Before continuing with a theoretical study of TOP-curves showing, among other aspects, their relation with corresponding Lorenz curves, we first present a concrete example of a TOP-curve.

TOP-curve example. In this section, an example of a TOP-curve is presented. We consider all articles published in Volume 51 (Year 2000) of the *Journal of the American Society for Information Science* (*JASIS*). We do not take editorials (i.e., "In this issue"), letters to the editor, book reviews, "In memoriam," and introductions to special issues into account. In this way, we retained 105 articles. According to the *Journal Citation Reports* (*JCR*), *JASIS* published 106 articles in 2000, so the numbers correspond within expected margins. We ranked these articles according to the number of citations received in the *ISI Web of Knowledge* on October 16, 2005. This is the dataset we will describe using a TOP-curve.

The choice of a threshold is always arbitrary. We choose the number of citations corresponding to the h-index as threshold line. The h-index was recently introduced as a number characterizing the scientific output of a researcher (Hirsch, 2005) and can easily be applied to other situations (here, one volume of a scientific journal). When a ranked list is given, the corresponding h-index is m if the first m items have a value larger than or equal to m while the item at rank $m + 1$ has a value strictly smaller than $m + 1$. For *JASIS* Volume 51, its citation h-index is equal to 15 because the article at rank 15 is cited 15 times while the article at rank 16 is cited less (i.e., 14 times). So we use 14.5 as the threshold. Recall that this number is completely arbitrary and used only as an illustration.

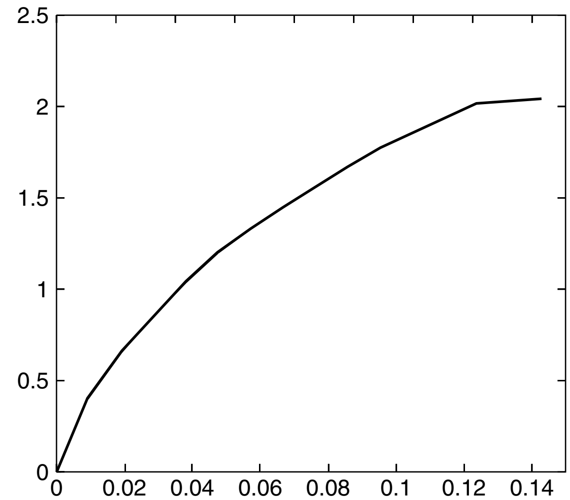

FIG. 3. Essential TOP-curve of the *JASIS* data.

As 15 of 105 articles belong to the top, the incidence is $15/105 = 0.143$. Table 1 shows these top articles. Based on Table 1 and the threshold value $t = 14.5$, we obtain an intensity value of $214.5/105 = 2.043$. This and other essential features are illustrated in Figure 3.

A referee asked if using another—acceptable—threshold (top line) would give a totally different result. Only more experience with TOP-curves based on other top lines can lead to a complete answer to this question. Just as an experiment, we consider a group of top sources based on another, historical threshold: the core group in the sense of Bradford (1934). Bradford divided a bibliography (IPP) into three groups with equal total production. The Bradford core consists of those most-productive sources producing one third of all items. In the case of *JASIS* Volume 51, this is the group of most-cited articles receiving one third of all citations. As the total number of citations on October 16, 2005 was 919, one third is 306.3, corresponding to the nine most-cited articles. These articles received at least 26 citations. So taking $t = 25.5$, the incidence is $9/105 = 0.086$ and the intensity is $76.5/109 = 0.702$. Clearly, using this threshold yields a totally different result. This is, of course, no surprise. Even after many years of using poverty lines, their exact definition (depending on family size and adapted to the cost of living) still leads to heated political debates in many countries (e.g., Ravallion, 1996; Sen, 1983).

Properties of TOP-Curves and Related TOP-Dominance Measures

First, we study a property of TOP-curves, based on the "dropping-out" transformation (defined later). This behavior will be contrasted with the behavior of the Lorenz curves of the surplus array under this same transformation. In this article, we always keep the top line fixed.

DOI: 10.1002/asi

TABLE 1. Top articles in *JASIS*, Volume 51.

Rank	Authors/Title	Citations received
1	Kling, R.; McKim, G. Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication	57
2	Bilal, D. Children's use of the Yahooligans! Web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks	41
3	Palmquist, R.A.; Kim, K.S. Cognitive style and on-line database search experience as predictors of Web search performance	35
4	Lazonder, A.W.; Biemans, H.J.A.; Wopereis, I.G.J.H. Differences between novice and experienced users in searching information on the World Wide Web	34
5	Harter, S.P.; Ford, C.E. Web-based analyses of e-journal impact: Approaches, problems, and issues	32
6	Xie, H. Shifts of interactive intentions and information-seeking strategies in interactive information retrieval	28
7	Zhang, Y. Using the Internet for survey research: A case study	27
8	Large, A.; Beheshti, J. The Web as a classroom resource: Reactions from the users	26
	Kim, H.J. Motivations for hyperlinking in scholarly electronic articles: A qualitative study	26
10	Tolle, K.M.; Chen, H.C. Comparing noun phrasing techniques for use with medical digital library tools	25
11	Sutcliffe, A.G.; Ennis, M.; Watkinson, S.J. Empirical studies of end-user information searching	23
	Case, D.O.; Higgins, G.M. How can we investigate citation behavior? A study of reasons for citing literature in communication	23
	Haas, S.W.; Grams, E.S. Readers, authors, and page structure: A discussion of four questions arising from a content analysis of Web pages	23
14	Dillon, A.; Gushrowski, B.A. Genres and the web: Is the personal home page the first uniquely digital genre?	17
15	Ross, N.C.M.; Wolfram, D. End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine	15

The "Dropping-Out" Transformation

Replace in X the component $x_{j_0} > t$ by $x'_{j_0} = t$, such that $x'_{j_0} = t \geq x_{j_0+1}$. Observe that this is always possible as $j_0 + 1$ is the first index j for which $x_j \leq t$. The new array obtained through this transformation is denoted as X':

$$X' = (x_1, \ldots, x_{j_0-1}, x'_{j_0} = t, x_{j_0+1}, \ldots, x_N) \tag{4}$$

This transformation will be referred to as a "dropping-out" transformation (DO-transformation), as source j_0 drops out of the group of top sources.

By the previous definitions,

$$S_{X'} = S_{X'}[j_0 - 1] = \left(\underbrace{x_1 - t, \ldots, x_{j_0-1} - t,}_{j_0-1\ components} \quad \underbrace{0, \ldots, 0}_{N-j_0+1\ times} \right) \tag{5}$$

where the first $j_0 - 1$ components are strictly larger than zero. The new array $S_{X'}[j_0 - 1]$ given by (5) is the surplus array of the DO-transformed situation. Clearly, the total surplus of X' is smaller than that of X. Array $S_{X'}[j_0 - 1]$ is a partial vector at Level $j_0 - 1$.

Proposition

Let $L_{S[j_0]}$ and $L_{S[j_0-1]}$ denote the Lorenz curves of the arrays $S_X[j_0]$ and $S_{X'}[j_0 - 1]$ as defined earlier. Then

$$L_{S[j_0]} < L_{S[j_0-1]} \tag{6}$$

For TOP-dominance, however, we obtain the opposite result. If X' is the DO-transformed array of X, then X TOP-dominates X':

$$\mathrm{TOP}_{X'} < \mathrm{TOP}_X \tag{7}$$

DOI: 10.1002/asi

Proof. If $1 \leq k \leq j_0 - 1$, then

$$L_{S[j_0]}\left(\frac{k}{N}\right) = \frac{\sum_{j=1}^{k} S_j}{\sum_{j=1}^{N} S_j} = \frac{\sum_{j=1}^{k} S_j}{\sum_{j=1}^{j_0} S_j} \text{ while } L_{S[j_0-1]}\left(\frac{k}{N}\right) = \frac{\sum_{j=1}^{k} S_j}{\sum_{j=1}^{j_0-1} S_j}$$

This shows that in this segment, the transformed Lorenz curve is situated above the original one. From $k = j_0$ on, both curves always take the value 1. This proves inequality (6). This result also can be found in Egghe (2002) and in Egghe and Rousseau (2005).

For the corresponding TOP-curves we have for $1 \leq k \leq j_0 - 1$:

$$\mathrm{TOP}_X\left(\frac{k}{N}\right) = \frac{\sum_{j=1}^{k} S_j}{N} = \mathrm{TOP}_{X'}\left(\frac{k}{N}\right)$$

From $k = j_0$ on, $\mathrm{TOP}_X\left(\frac{k}{N}\right) = \frac{\sum_{j=1}^{j_0} S_j}{N}$ while $\mathrm{TOP}_{X'}\left(\frac{k}{N}\right) = \frac{\sum_{j=1}^{j_0-1} S_j}{N}$ This shows that $\mathrm{TOP}_{X'}$ is situated below TOP_X. This proves the proposition.

These results are interpreted as follows: If the weakest of the TOP-group is removed from this group (all other things being the same), the inequality of the surplus array increases. This seems to be an acceptable conclusion. Yet, the general feeling of dominance decreases. The reason for this is that the total surplus has decreased. Lorenz curves use relative values on the ordinate axis while TOP-curves use absolute values. Hence, the two constructions shine a different light on the same operation.

Remarks

1. Because the total surplus of X is always larger than that of X', we could not deduce inequality (7) directly form inequality (6), using the previous lemma.
2. Note that the earlier result also is correct if the component $X_{j_0} > t$ in X is replaced by $X'_{j_0+1} \leq t$ (not necessarily equal to t) such that $X_{j_0+1} \leq X'_{j_0} \leq t$.

Corollary

If a given N-array X (not equivalent to the null TOP-curve) is transformed step by step through a series of DO-transformations to the null TOP-curve, then at each step we obtain a Lorenz curve for the corresponding surplus array which is strictly more concentrated than the previous one. Hence, applying DO-transformations is a stepwise procedure increasing at each step the inequality in the corresponding surplus arrays. Similarly, the general feeling of dominance decreases in this stepwise procedure.[1]

Definition

An (acceptable) strict TOP-dominance measure is a function f respecting the TOP-dominance order of finite arrays, given the top line t. In other words, if X and Y are arrays and if $\mathrm{TOP}_X > \mathrm{TOP}_Y$ then $f(X) > f(Y)$ (strict inequality).

Examples. The area under the TOP-curve is an acceptable strict TOP-dominance measure. Similarly, the length of the TOP-curve is an acceptable strict TOP-dominance measure.

This area can be calculated as:

$$\frac{\sum_{k=1}^{j_0} (2N - 2k + 1)S_X(k)}{2N^2},$$

while the length of the TOP-curve is:

$$\frac{\sum_{k=1}^{j_0}\left(\sqrt{1 + S_X(k)^2}\right) + (N - j_0)}{N}$$

Calculations can be found in Appendix A.

Thon (1979, p. 433) formulated a series of desirable properties for poverty measures. Based on his ideas, we will study the corresponding properties of TOP-curves and, hence, of dominance measures respecting TOP-dominance. The complete set of desirable properties (according to Thon) is covered, explicitly or implicitly.

Proposition A: Restricted Form of the Transfer Principle

If two top sources (i.e., sources with a production above the top line) are ranked one after the other but have a different production, then replacing each source's production by their average production yields a TOP-curve which is situated strictly below the original one. This means that such an operation decreases the feeling of dominance.

Proof. By the general transfer principle, this property is true for Lorenz curves (Dalton, 1920). As the total surplus and the incidence are not changed by this transformation, the property also holds for TOP-curves.

Proposition B: Monotonicity (Sen, 1976)

If the production of one top source decreases, the new TOP-curve is situated strictly under the old one. This loss of production may be caused by external reasons or by a transfer to a source or sources under the top line, as long as the production of the source(s) stays under the top line.

[1]One can show that the Lorenz curves of T_X and $T_{X'}$ [as defined by (1)] always intersect in a point in]0, 1[(A proof can be obtained from the authors.) Hence, this proposition is not valid for T_X.

DOI: 10.1002/asi

Proof. Let $X = (x_1, x_2, \ldots, x_N)$ be a given N-array, and let $S_X[j_0]$ be its surplus array:

$$S_X = \left(x_1 - t, x_2 - t, \ldots, x_{j_0} - t, \quad \underbrace{0, \ldots, 0}_{N - j_0 \, times} \right).$$

The transformation described in this proposition leads to a new array, denoted as X'. It implies that there exists a (unique) index $j_1 \in \{1, \ldots, j_0\}$ such that $x_{j_1} - t$ is replaced by a, with $0 \le a < x_{j_1} - t$. Here, $a = 0$ if this loss brings this source's production under or on the top line, otherwise $a > 0$. If $a = 0$, then

$$S_{X'} = \left(x_1 - t, x_2 - t, \ldots, x_{j_1 - 1} - t, x_{j_1 + 1} - t, \ldots, x_{j_0} - t, \ \underbrace{0, \ldots, 0}_{N - j_{0+1} \, times} \right)$$

If $a > 0$ then there exists an index $j_2 \in \{j_0 + 1, \ldots, N\}$ such that $S_{X'} = \Big(x_1 - t, x_2 - t, \ldots, x_{j_1 - 1} - t, x_{j_1 + 1} - t, \ldots, x_{j_2 - 1} - t,$

$$a, x_{j_2} - t, \ldots, x_{j_0} - t, \underbrace{0, \ldots, 0}_{N - j_0 \, times} \Big)$$

The first components of S_X and $S_{X'}$ (from $x_1 - t$ to $x_{j_1 - 1} - t$) are the same. From that point on until the component with value a, the TOP-curve of X' certainly stays under the TOP-curve of X as, at each component, the surplus values for X' are smaller than or equal to those for X. If $a = 0$, this already proves this proposition. Otherwise, at the component with value a, the ordinate value for TOP_X is $\frac{\sum_{k=1}^{j_2 - 1} (x_k - t)}{N}$ while the ordinate value of $\mathrm{TOP}_{X'}$ is $\frac{\sum_{k=1, k \neq j_1}^{j_2 - 1} (x_k - t) + a}{N}$. As $a < x_{j_1} - t$ this shows that also here the TOP-curve of X' is strictly under the TOP-curve of X. From this point on, the components of S_X and $S_{X'}$ are again the same. This ends the proof of this proposition.

From a "domination" point of view, this result is as one might expect. Yet, this result is not true for concentration curves (see examples). This clearly illustrates why classical Lorenz curves are not suited for the study of top sources.

Examples. Let $X = (8, 6, 4, 2)$, let $t = 4$ and let $X'_1 = (8, 5, 4, 2)$ and $X'_2 = (8, 5, 4, 3)$. In the second case, there has been a transfer from a top source to a less-performing source; in the first case, there is simply a decrease in performance for the second source. In any case, we have: $S_X = (4, 2, 0, 0)$ and $S_{X'} = (4, 1, 0, 0)$, and hence, the Lorenz curve of S_X is situated under the Lorenz curve of $S_{X'}$. Yet, considering another transformation of the same type, namely X to $X' = (7, 6, 4, 2)$ leads to $S_{X'} = (3, 2, 0, 0)$. Now, the Lorenz curve of S_X is situated above the one for $S_{X'}$.

By the way TOP-curves are constructed, it immediately follows that TOP-dominance values are not affected by changes in the performance of non-top sources. This property is sometimes referred to as "focus" (Zheng, 2000). TOP-dominance values also are "anonymous;" that is, they depend only on the actual data, not on the source that has contributed a particular component of the studied array.

Relation Between Generalized Lorenz Curves of Finite Arrays and TOP-Curves

Proposition C

Let $X = (x_1, x_2, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_M)$ be finite arrays, and let t be a given top line, then

$$\mathrm{GL}_X < \mathrm{GL}_Y \text{ implies } \mathrm{TOP}_X < \mathrm{TOP}_Y.$$

The rather technical proof can be found in Appendix B.

Corollary. Let $X = (x_1, x_2, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_N)$ be N-arrays such that $\sum_{j=1}^{N} x_j = \sum_{j=1}^{N} y_j$ and let t be a given top line, then

$$L_X < L_Y \text{ implies } \mathrm{TOP}_X < \mathrm{TOP}_Y$$

This follows from the fact that $\sum_{j=1}^{N} x_j = \sum_{j=1}^{N} y_j$ implies that $L_X < L_Y$ is equivalent to $GL_X < GL_Y$.

Remark. If $L_X = L_Y$ and $\sum_{j=1}^{N} x_j = \sum_{j=1}^{N} y_j$, then clearly $L(S_X) = L(S_Y)$.

Note that the opposite implication of the corollary [i.e., if $\sum_{j=1}^{N} x_j = \sum_{j=1}^{N} y_j$, then $\mathrm{TOP}_X \le \mathrm{TOP}_Y$ implies $\mathrm{L}_X < \mathrm{L}_Y$ ($\Leftrightarrow \mathrm{GL}_X < \mathrm{GL}_Y$)], is in general not true (see counterexample described next). This proves that the TOP-dominance partial order (given t) is stronger than the Lorenz dominance order. Recall that given a set U and two partial orders $<$ and $\ll$ on U, then $<$ is said to be stronger than $\ll$ (and therefore $\ll$ is weaker than $<$) if for every x and y in U: $x \ll y$ implies $x < y$. This relation between these two partial orders is another reason why TOP-curves are better suited to study the top sources than are classical Lorenz curves. This conclusion is, of course, also true for the corresponding measures of inequality or TOP-dominance.

A Counterexample

Let $X = (6, 6, 4, 2)$, $Y = (7, 5, 3, 3)$, and $t = 4$. Clearly, X and Y are incomparable according to the Lorenz order, as the corresponding Lorenz curves intersect. $S_X = (2, 2, 0, 0)$ and $S_Y = (3, 1, 0, 0)$, and hence $\mathrm{TOP}_Y > \mathrm{TOP}_X$.

Summary and Conclusion

In this contribution, we have introduced TOP-curves because several characteristics of classical Lorenz curves make them unsuitable for the study of a group of top sources. For

DOI: 10.1002/asi

example, Lorenz curves do not reflect intensity or incidence of the top. They are, moreover, invariant under scale transformations.

TOP-curves, defined as a kind of mirror image of TIP-curves as used in poverty studies, are shown to possess properties necessary for an adequate empirical ranking of various data arrays and this based on the properties of the highest performers. They simultaneously represent the incidence, intensity, and inequality among the top. Moreover, the TOP-dominance partial order (given a top line t) is proven to be stronger than the Lorenz dominance order. We therefore advocate the use of TOP-curves when focusing on the top sources in information-production processes. In this way, this article contributes to the study of cores, a central issue in applied informetrics.

Besides rankings of TOP-curves with a fixed top line, one also may study the influence of varying top lines. This is similar to the difference between poverty-measure ordering and poverty-line ordering (Zheng, 2000). This type of study is, however, left to future research.

Acknowledgment

Thanks are due to two anonymous reviewers for insightful comments and suggestions.

References

Bogaert, J., Zhou, L., Tucker, C.J., Myneni, R.B., & Ceulemans, R. (2002). Evidence for a persistent and extensive greening trend in Eurasia inferred from satellite vegetation index data. Journal of Geophysical Research, 107(ACL 4–1), 4–14.

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

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

Burrell, Q.L. (1993). The Gini index and the Leimkuhler curve for bibliometric processes. Information Processing and Management, 29, 512–522.

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

Egghe, L. (2002). Development of hierarchy theory for digraphs using concentration theory based on a new type of Lorenz curve. Mathematical and Computer Modelling, 36, 587–602.

Egghe, L. (2005). Power laws in the information production process: Lotkaian informetrics. Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (2004). How to measure own-group preference? A novel approach to a sociometric problem. Scientometrics, 59, 233–252.

Egghe, L., & Rousseau, R. (2005). Comparing partial and truncated conglomerates from a concentration theoretic point of view. Mathematical and Computer Modelling, 41, 301–311.

Egghe, L., & Rousseau, R. (2006). Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. Information Processing and Management, 42, 106–120.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, USA, 102, 16569–16572.

Jenkins, S.P., & Lambert, P.J. (1997). Three "I"s of poverty curves, with an analysis of UK poverty trends. Oxford Economic Papers, 49, 317–327.

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

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

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

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

Nijssen, D., Rousseau, R., & Van Hecke, P. (1998). The Lorenz curve: A graphical representation of evenness. Coenoses, 13(1), 33–38.

Patil, G.P., & Taillie, C. (1979). An overview of diversity. In J.F. Grassle, G.P. Patil, W. Smith, & C. Taillie (Eds.), Ecological diversity in theory and practice (pp. 3–27). Fairland, MD: International Cooperative.

Ravallion, M. (1996). Issues in measuring and modelling poverty. Economic Journal, 106(483), 1328–1343.

Rousseau, R. (1992). Concentration and diversity measures: Dependence on the number of classes. Belgian Journal of Operations Research, Statistics and Computer Science, 32, 99–126.

Rousseau, R. (1998). Evenness as a descriptive parameter for department or faculty evaluation studies. In E. de Smet (Ed.), Informatiewetenschap 1998 (pp. 135–145). Antwerp, Belgium: Werkgemeenschap Informatiewetenschap.

Sen, A.K. (1976). Poverty: An ordinal approach to measurement. Econometrica, 44, 219–231.

Sen, A.K. (1983). Poor, relatively speaking. Oxford Economic Papers, 35, 153–169.

Shorrocks, A.F. (1983). Ranking income distributions. Economica, 50, 3–17.

Spencer, B., & Fisher, S. (1992). On comparing distributions of poverty gaps. Sankhya: The Indian Journal of Statistics, Series B, 54, 114–126.

Thon, D. (1979). On measuring poverty. Review of Income and Wealth, 25, 429–439.

Weiner, J. (1985). Size hierarchies in experimental populations of annual plants. Ecology, 66, 743–752.

Zheng, B. (2000). Poverty orderings. Journal of Economic Surveys, 14(4), 427–466.

Appendix A

Area Under a TOP-Curve and Length of a TOP-Curve

We consider the graph (see Figure 1) and calculate the area of each horizontal zone starting from below. Each zone consists of a triangle followed by a rectangle. This leads to the following sum:

$$\frac{1}{2N}\cdot\frac{S_1}{N}+\left(\frac{N-1}{N}\right)\cdot\frac{S_1}{N}+\frac{1}{2N}\cdot\frac{S_2}{N}+\left(\frac{N-2}{N}\right)\cdot\frac{S_2}{N}$$

$$+\cdots+\frac{1}{2N}\cdot\frac{S_{j_0}}{N}+\left(\frac{N-j_0}{N}\right)\cdot\frac{S_{j_0}}{N}$$

$$=\left(\frac{2N-1}{2N}\right)\cdot\frac{S_1}{N}+\left(\frac{2N-3}{2N}\right)\cdot\frac{S_2}{N}+\cdots$$

$$+\left(\frac{2N-(2j_0-1)}{2N}\right)\cdot\frac{S_{j_0}}{N}=\frac{\sum_{k=1}^{j_0}(2N-2k+1)S_X(k)}{2N^2}.$$

This is the area under a TOP-curve.

To calculate the length of a TOP-curve, we simply add lengths of hypotenuses of triangles and add the length of the

DOI: 10.1002/asi

horizontal line segment on the end. This yields:

$$\sum_{k=1}^{j_0}\sqrt{\frac{1}{N^2}+\frac{S_k^2}{N^2}}+\left(1-\frac{j_0}{N}\right)$$

$$=\frac{\sum_{k=1}^{j_0}\left(\sqrt{1+S_X(k)^2}\right)+(N-j_0)}{N}.$$

Appendix B: Proof of Proposition C

Proposition C

Let $X = (x_1, x_2, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_M)$ be finite arrays, ranked in decreasing order, and let t be a given top-line, then

$$\mathrm{GL}_X < \mathrm{GL}_Y \text{ implies } \mathrm{TOP}_X < \mathrm{TOP}_Y$$

Proof. We first transform X and Y such that they become arrays of the same length (NM) without changing their average, by using the REPEAT-operation. The transformed X is denoted as RX = REPEAT$_M$(X) and is defined as:

$$RX = \left(\underbrace{x_1, \ldots, x_1}_{M\ times}, \ldots, \underbrace{x_N, \ldots, x_N}_{M\ times}\right)$$

Similarly, Y is transformed into RY = REPEAT$_N$(Y) and is defined as:

$$RY = \left(\underbrace{y_1, \ldots, y_1}_{N\ times}, \ldots, \underbrace{y_M, \ldots, y_M}_{N\ times}\right)$$

The components of RX and RY will be denoted as $(\xi_k)_{k=1,\ldots,NM}$ and $(\eta_k)_{k=1,\ldots,NM}$. Clearly, this transformation leaves averages of arrays invariant. This implies that the generalized Lorenz curves of GL_X and GL_{RX} coincide, as do GL_Y and GL_{RY}. Consequently $\mathrm{GL}_X < \mathrm{GL}_Y$ implies $\mathrm{GL}_{RX} < \mathrm{GL}_{RY}$. From this inequality, we deduce that for all j, $1 \le j \le$ NM:

$$\sum_{m=1}^{j}\xi_m \le \sum_{m=1}^{j}\eta_m, \tag{8}$$

with at least one strict inequality. Then we also have:

$$\sum_{m=1}^{j}(\xi_m - t) \le \sum_{m=1}^{j}(\eta_m - t) \tag{9}$$

where some of the terms in these sums can be negative.

Let j_0 be defined as in the previous sections, then, $S_{RX} = \left(\boldsymbol{\xi}_1 - t, \ldots, \boldsymbol{\xi}_{j_0} - t, \underbrace{0, \ldots, 0}_{NM-j_0\ times}\right)$ and let k_0 play the same role for the array RY. Hence, $S_{RY} = \left(\eta_1 - t, \ldots, \eta_{k_0} - t, \underbrace{0, \ldots, 0}_{NM-k_0\ times}\right)$ Note that the corresponding indices for X and Y are j_0/M and k_0/N. As TOP-curves are replication invariant, it follows that proving that $\mathrm{TOP}_X < \mathrm{TOP}_Y$ is equivalent to proving $\mathrm{TOP}_{RX} < \mathrm{TOP}_{RY}$.

The proof considers two main cases: $j_0 < k_0$ and $j_0 \ge k_0$.

Case A: $j_0 < k_0$

This first case is subdivided into three parts: $j \le j_0$, $j_0 < j \le k_0$ (if $j_0 \ne k_0$), and $k_0 < j \le$ NM.

If $j \le j_0$ then, by inequality (9)

$$S_{RX}(j) = \frac{\sum_{m=1}^{j}(\xi_m - t)}{NM} \le \frac{\sum_{m=1}^{j}(\eta_m - t)}{NM} = S_{RY}(j)$$

If $j_0 < j \le k_0$ (if $j_0 \ne k_0$), then

$$S_{RX}(j) = \frac{\sum_{m=1}^{j_0}(\xi_m - t)}{NM} < \frac{\sum_{m=1}^{j_0}(\eta_m - t)}{NM}$$

$$\le \frac{\sum_{m=1}^{j}(\eta_m - t)}{NM} = S_{RY}(j)$$

Finally, if $k_0 < j \le$ MN, then

$$S_{RX}(j) = \frac{\sum_{m=1}^{j_0}(\xi_m - t)}{NM} < \frac{\sum_{m=1}^{k_0}(\eta_m - t)}{NM} = S_{RY}(j).$$

Case B: $j_0 \ge k_0$

This case also is subdivided into three parts: $j \le k_0$, $k_0 < j \le j_0$ (if $j_0 \ne k_0$), and $j_0 < j \le$ NM.

If $j \le k_0$, then the first part of Case A is still valid. If $k_0 < j \le j_0$ (if $j_0 \ne k_0$), then $S_{RX}(j) = \frac{\sum_{m=1}^{j}(\xi_m - t)}{NM} \le \frac{\sum_{m=1}^{j}(\eta_m - t)}{NM} < \frac{\sum_{m=1}^{k_0}(\eta_m - t)}{NM} + 0 = S_{RY}(j)$, and finally, the third case is again the same as for Case A. This proves Proposition C.

DOI: 10.1002/asi