Journal of Informetrics 13 (2019) 387–396

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

How mean rank and mean size may determine the generalised Lorenz curve: With application to citation analysis

Lucio Bertoli-Barsotti [a,*], Tommaso Lando [a,b]

[a] Department of Management, Economics and Quantitative Methods, University of Bergamo, via dei Caniana 2, 24127 Bergamo, Italy
[b] Department of Finance, VŠB -TU Ostrava, Sokolskà 33, 70121 Ostrava, Czech Republic

ARTICLE INFO

Article history:
Received 21 November 2018
Received in revised form 4 February 2019
Accepted 4 February 2019

Keywords:
Lorenz curve
Stochastic dominance
Journal ranking
Journal impact factor
Geometric distribution

ABSTRACT

Within the wide framework of information production processes, we present a conversion formula that expresses the generalised Lorenz (*GL*) curve of a size-frequency distribution as a function of the corresponding rank-size distribution using a fully discrete modelling approach. Based on this conversion formula, we introduce a somewhat universal model for the *GL* curve of the empirical size-frequency distribution. This study's approach to determining the *GL* curve is indirect, as we obtain our model for the size-frequency framework by modelling the rank-size distribution and not by directly modelling the size distribution or the *GL* curve itself, as is usually done. Our *GL* curve model is particularly appealing because it provides a simple and economical description of the distribution that depends on only three quantities: the (i) mean size, (ii) mean rank, and (iii) maximal rank. The model's performance in predicting the shape of the empirical *GL* curve is illustrated through a case study involving citation analysis.

1. Introduction

In informetrics, size-frequency (SF) and rank-size (RS) distributions can be viewed as 'dual' approaches within the more general framework of information production processes (IPPs). As Egghe (1990, 2005a, p.8) points out, an IPP is a formal mechanism describing 'sources that produce items'. Examples of IPPs include analyses of the frequency with which words occur in a text; of the number of employees in a company; of the number of articles produced by certain researchers; of the number of people living in certain cities at a given time; of the number of visits to a certain webpage; of the number of research articles published in a particular field; of the number of articles published by journals on a given subject; and of the number of citations received by a journal within a specific time window. In a broad sense, the 'sizes' are the 'production quantities of the different sources' (Egghe, 2005a, p. vii). The size, also called 'production' (Rousseau, 1990), of a source, may be simply defined as the number of items produced by that source. For this reason, the function which describes the observed size of a given source, depending on its rank, is usually called rank-frequency function, in the IPPs terminology (Egghe, 2005a). Indeed, in this special case the rank-frequency function is a genuine counting function (if the frequency is not zero). Nevertheless, the size may also be conceived as a point of the image set of a more general non-negative real-valued function, defined on the set of the items, as it is, for example, in the case that the journals are the sources and the journal impact factors play the

* Corresponding author.
E-mail address: lucio.bertoli-barsotti@unibg.it (L. Bertoli-Barsotti).

https://doi.org/10.1016/j.joi.2019.02.003

role of sizes (see Egghe & Waltman, 2011). Therefore, to be more precise about terminology, in what follows we shall use the more general term rank-*size*, instead of rank-*frequency*. In general, the SF function yields the number of sources, giving a specific size x, while the RS function represents the size x_i produced by the source on rank $i = 1, 2, 3, \ldots$, where the sources are ranked in decreasing order of their sizes. Obviously, SF and RS functions, as broadly defined here, can be equivalently obtained through their corresponding cumulative distribution functions (CDFs), and through their corresponding generalised Lorenz (*GL*) curves (Shorrocks, 1983).

As part of the same process, SF and RS distributions, which can be summarised through their mean values, say mean size and mean rank, respectively, are mathematically interrelated. Indeed, by considering the size x as a continuous variable (adopting a continuous approach as opposed to a discrete one; Rousseau, 1990), the relationship between SF and RS distributions can be described, writing an integral sign instead of a sum, by saying that 'the rank-size curve is essentially the integral of the size-frequency curve' (Rapaport, 1978). Each functional model for the RS distribution theoretically corresponds to a specific functional model for the SF distribution, and vice versa, even if their relationship is not simple to derive in practice (Li, Miramontes, & Cocho, 2010). Partial results concerning the relationship between the shapes of SF and RS distributions have been obtained by, for example, Egghe and Rousseau (2006), Egghe & Rousseau (2012) and Egghe and Waltman (2011). These works and other studies concerning the modelisation of SF distributions typically use the continuous approach 'since this is mathematically more convenient' (Egghe & Rousseau, 2006; see also Egghe, 2005b).

By contrast, we consider the natural framework of the empirical data, even if mathematically less convenient, that entails a discrete setting where i) the supports of both the size and rank variables are finite (or countably infinite) sets and ii) ties (i.e. cases in which different sources produce the same size) are allowed. Pointing toward citation analysis applications, this study pays special attention to an empirical case in which articles (sources) produce citations (items). We recall that the *GL* curve has been proven effective as a tool for characterising citation patterns (Lando & Bertoli-Barsotti, 2017). Specifically, the *GL* curve can be used to distinguish between journals with similar 'impact factors' expressed as a mean size. Indeed, the convexity of the *GL* curve reflects the degree of dispersion in citation data (e.g. due to the presence of a few highly cited papers) and may be measured consistently by a linear function of the mean rank parameter, that is, by a linear function of the Gini index.

The aim of this study is twofold. 1) First, it seeks to provide the exact relation between SF and RS distributions in a discrete setting through a conversion formula that expresses the *GL* curve of the SF distribution as a function of the rank distribution. This formula shows that the first-order stochastic dominance between RS CDFs corresponds to the second-order stochastic dominance between SF CDFs (under the condition of equal means). 2) Second, the study obtains a simple parametric model for the *GL* curve of the SF distribution starting from the above conversion formula. This *GL* curve model relies on the assumption of a geometric law for the RS distribution and is essentially governed by only two parameters, the mean size and the mean rank, the two simplest moments characterising the two distributions at hand. Despite its simple structure, our *GL* curve model can provide a very good fit, as shown in the case study presented below.

The rest of this paper proceeds as follows. In Section 2 we establish the notation used in our analysis. In Section 3, we present the conversion formula that relates the RS distribution to the (*GL* curve of the) SF distribution. In Section 4, we introduce our parametric model for the *GL* curve of SF distribution. In Section 5, we illustrate the proposed approach with a case study. Finally, in Section 6, we present our conclusions.

2. Rank and size distributions

2.1. Rank distributions

We consider the case of T objects, entities we will call 'sources', each of which has/produces a set of items. In turn, this set of items corresponds to a quantity/measure that we will call 'size' (of the source). As mentioned above, the size may be defined, for example, as the number of items produced by the source. The RS function is defined as $\tilde{n}(i)$ and represents the size of the source on rank i, $i = 1, 2, 3, \ldots, T$, where

$$\tilde{n}(1) \geq \tilde{n}(2) \geq \ldots \geq \tilde{n}(T) \geq 0$$

In empirical IPPs, the number of sources, although possibly large, is finite; thus, $T < \infty$. T may also be interpreted as the maximal rank. To simplify the notation, we also write $x_i = \tilde{n}(i)$, $i = 1, 2, 3, \ldots, T$, where $x_1 \geq x_2 \geq \ldots \geq x_T \geq 0$. By definition, x_i need not be an integer number (Egghe & Waltman, 2011). This allows us to cover the general case in which the size represents, for example, a weighted counting (Glänzel & Moed, 2002), a relative frequency, a proportion, a percentage, a quantile, a generic degree, or a value of a statistical indicator such as an arithmetic mean. For example, Egghe and Waltman (2011) (see also Sarabia, Prieto, & Trueba, 2012 and Mansilla, Köppen, Cocho, & Miramontes, 2007) studied the rank-order behaviour of journal impact factors, which are clearly not necessarily integer numbers. We also assume $0 < x_1 < \infty$, to avoid a degenerate case, and we define the total size as $\sum_{i=1}^{T} x_i = C < \infty$.

The assumption of no ties, made by several authors assuming a continuous approximation model (e.g. Egghe & Waltman, 2011; Burrell, 2005), oversimplifies a more complex real situation. We thus prefer to deal with the most general case in which ties are allowed. In this case, $\tilde{n}$ is not an invertible function. Thus, for every i, $i = 1, 2, \ldots, T$, $r(\tilde{n}(i)) \geq i$, where $r(x) =$

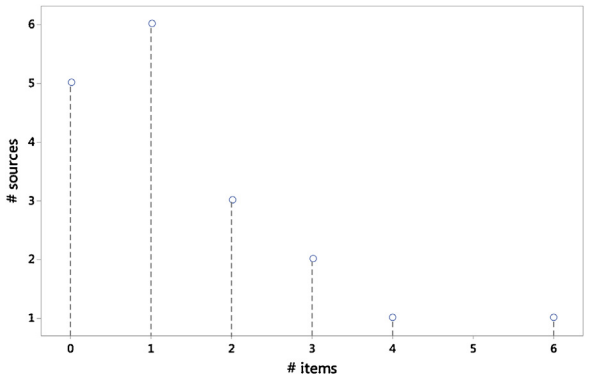

Fig. 1. SF function. Set of pairs (*size, #sources*): (0,5), (1,6), (2,3), (3,2), (4,1), (6,1).

$\max\left\{i : \tilde{n}(i) = x,\ i \in \left\{1, 2, \ldots, T\right\}\right\}$ is the generalised inverse function of $\tilde{n}$. Let $\tilde{N}(t) = \sum_{i \le t} \tilde{n}(i)$. In particular, $\tilde{N}(t) = C$ for every $t \ge T$. The normalised version of $\tilde{N}(t)$ is the following function:

$$\tilde{F}(t) = \sum_{i \le t} \frac{\tilde{n}(i)}{C} = \sum_{i \le t} \tilde{f}(i) = C^{-1} \sum_{i \le t} x_i,\ -\infty < t < \infty$$

where $\tilde{f}(t) = \frac{\tilde{n}(t)}{C}$ for every t, $t = 1, 2, \ldots, T$ and 0 otherwise denotes a function that is non-negative and that sums to 1. Then, $\tilde{f}$ behaves as a probability mass function, while $\tilde{F}$ behaves as a discrete probability distribution function on the real line ($\lim_{t \to -\infty} \tilde{F}(t) = 0$; $\lim_{t \to \infty} \tilde{F}(t) = 1$; $\tilde{F}(0) = 0$; $\tilde{F}(1) = x_1 C^{-1} > 0$; $\tilde{F}(T) = 1$), with a support set $\mathcal{N}$ that is an *initial segment* of the set of positive integers, $\mathbb{N}$. Recall that a set $\left\{m \in \mathbb{N} | m < k\right\}$ for some $k \in \mathbb{N}$ is said to be an initial segment of $\mathbb{N}$ (Rubin, 1967, p. 161).

We will refer to functions $\tilde{f}$ and $\tilde{F}$ as, respectively, 'RS density function' ('density' with respect to the counting measure) and 'RS CDF', or 'rank density' and 'rank distribution' for short. The mean of the ranks is thus defined as

$$\tilde{\mu} = \frac{\sum_{i=1}^{T} i x_i}{C} = \sum_{i=1}^{T} i \cdot \tilde{f}(i) = \sum_{s=0}^{T-1} \left(1 - \tilde{F}(s)\right). \tag{1}$$

Let us refer to $\tilde{\mu}$ as the mean rank.

2.2. Size distributions

The SF function, $n(x)$, is defined as the number of sources producing the size x, for every $x \in \mathcal{M}$, where $\mathcal{M} = \left\{x | x = \tilde{n}(i),\ i = 1, 2, \ldots, T\right\}$ is the image set of the function $\tilde{n}$, $\tilde{n} : \left\{1, 2, \ldots, T\right\} \to \mathcal{M}$, and 0 otherwise. By definition, $\sum_{x \in \mathcal{M}} n(x) = T$. Then, the function $F(x) = \sum_{z \le x} f(z)$, where $f(x) = \frac{n(x)}{T}$, $x \in \Re$, behaves as a discrete probability distribution function on the real line (with $\lim_{x \to -\infty} F(x) = 0$; $\lim_{x \to \infty} F(x) = 1$; $F(x_1) = 1$). Let us refer to f and F as the 'SF density function' and the 'SF CDF' respectively, or 'size density' and 'size distribution' for short. The corresponding mean of the sizes is defined as

$$\mu = \sum_{x \in \mathcal{M}} x f(x) = \sum_{x \in \mathcal{M}} \frac{x n(x)}{T} = \frac{\sum_{i=1}^{T} x_i}{T} = \frac{C}{T}.$$

Let us refer to μ as the 'mean size'.

2.3. An example

To illustrate the notions and notations introduced above, let us consider an artificial dataset constituted by $T = 18$ papers, where five papers have zero citations each; six papers have one citation each; three papers have two citations each; three papers have three citations each; one paper has four citations, and one paper has six citations. Here, the papers represent the sources and the citations represent the items. The size is the number (possibly zero) of items. From these data, we determine the following: the total number of citations, $C = 28$; the support of the SF distribution, $\mathcal{M} = \{0, 1, 2, 3, 4, 6\}$; the SF density function, defined by $f(0) = 0.278$, $f(1) = 0.333$, $f(2) = 0.167$, $f(3) = 0.111$, $f(4) = 0.056$ and $f(6) = 0.056$ (see Fig. 1); and the mean size, $\mu = \sum_{z \in \mathcal{M}} z f(z) = C/T = 28/18 = 1.556$. The numbers of citations per article, ranked from the

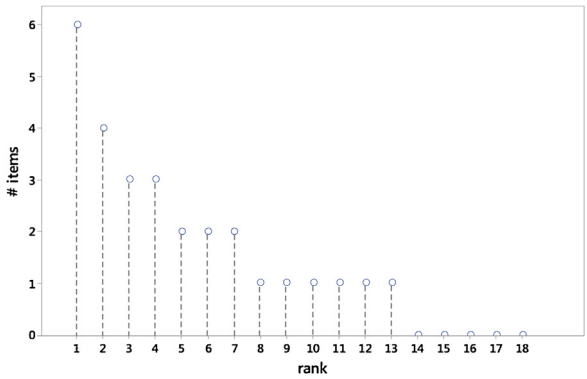

Fig. 2. RS function. Set of pairs (*size*, *#sources*): (0, 5) , (1, 6) , (2, 3) , (3, 2) , (4, 1) , (6, 1) . The domain is an initial segment of $\mathbb{N}$.

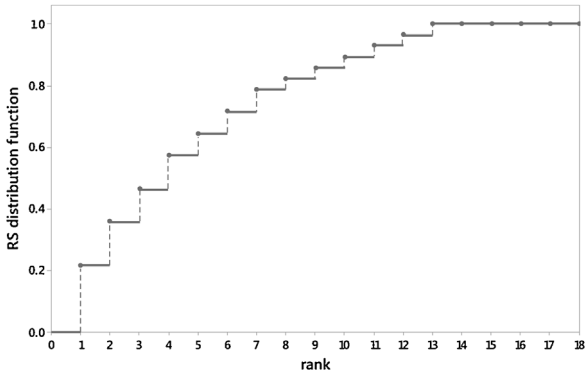

Fig. 3. RS CDF. Set of pairs (*size, # sources*): (0,5), (1,6), (2,3), (3,2), (4,1), (6,1). RS CDF.

most cited to the least cited paper, are $x_1 = 6, x_2 = 4; x_3 = x_4 = 3, x_5 = x_6 = x_7 = 2, x_8 = x_9 = x_{10} = x_{11} = x_{12} = x_{13} = 1, x_{14} = x_{15} = x_{16} = x_{17} = x_{18} = 0$. The RS density function is defined by $\tilde{f}_1 = 0.214, \tilde{f}_2 = 0.143, \tilde{f}_3 = \tilde{f}_4 = 0.107; \tilde{f}_5 = \tilde{f}_6 = \tilde{f}_7 = 0.071, \tilde{f}_8 = \tilde{f}_9 = \tilde{f}_{10} = \tilde{f}_{11} = \tilde{f}_{12} = \tilde{f}_{13} = 0.036, \tilde{f}_{14} = \tilde{f}_{15} = \tilde{f}_{16} = \tilde{f}_{17} = \tilde{f}_{18} = 0$ (see Fig. 2). The RS distribution function (see Fig. 3) is defined by $\tilde{F}_1 = 0.214, \tilde{F}_2 = 0.357, \tilde{F}_3 = 0.464, \tilde{F}_4 = 0.571, \tilde{F}_5 = 0.643, \tilde{F}_6 = 0.714, \tilde{F}_7 = 0.786, \tilde{F}_8 = 0.821, \tilde{F}_9 = 0.857, \tilde{F}_{10} = 0.893, \tilde{F}_{11} = 0.929, \tilde{F}_{12} = 0.964, \tilde{F}_{13} = \tilde{F}_{14} = \tilde{F}_{15} = \tilde{F}_{16} = \tilde{F}_{17} = \tilde{F}_{18} = 1$. The mean rank is $\widetilde{\mu} = \sum_{i=1}^{18} i \cdot \widetilde{f}(i) = 4.786$.

3. Conversion formula between the RS and SF domains

3.1. SF generalised Lorenz curve as a function of the rank distribution

As mentioned before, SF and RS distributions are related to each other (Egghe & Waltman, 2011) and can be uniquely characterized through their (respective) *GL* curves (Sarabia et al., 2012). In the present section, we derive a conversion formula that expresses the *GL* curve of the SF distribution as a function of the RS CDF. The importance of the *GL* curve for practical applications will be discussed in Section 5.

As is well-known, the *GL* curve of a T -point sample of non-negative numbers, $x_1, x_2, \ldots, x_T$, is defined as

$$GL\left(\frac{i}{T}\right) = \mu \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{T} x_{(j)}} = \frac{i}{T} \frac{\sum_{j=1}^{i} x_{(j)}}{i} \tag{2}$$

for every $i = 1, 2, \ldots, T$, and $GL(0) = 0$, where $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(T)}$, $\sum_{i=1}^{T} x_i = C$, and $\mu = C/T$. Note that, as a special case, one obtains $GL(1) = \mu$. The function *GL* extends uniquely to a function, still denoted by *GL*, defined by linear interpolation for every point in the set [0, 1]. Fig. 4 illustrates the *GL* curve derived from data in the above example.

When denoting the *GL* dominance relation using $\leq_{GL}$ (i.e. $G \leq_{GL} F$), we mean that the *GL* curve of the CDF F is nowhere below the *GL* curve of the CDF G. It is well-known that *GL* dominance is equivalent to second-order stochastic dominance (see e.g. Thistle, 1989), defined by the condition

$$\int_0^t F(z)\,dz \leq \int_0^t G(z)\,dz \text{ for all } t > 0.$$

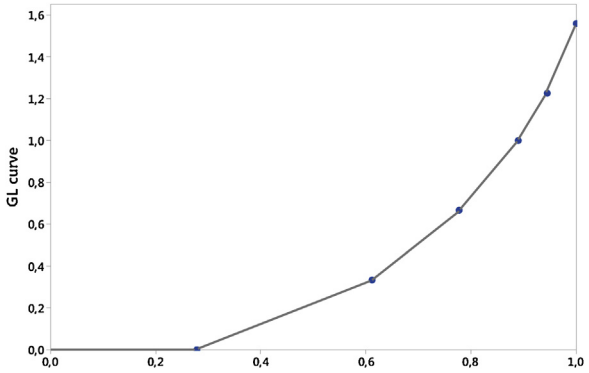

Fig. 4. *GL* curve. Set of pairs (*size, # sources*): (0,5), (1,6), (2,3), (3,2), (4,1), (6,1).

We identify a spurious mean size dominance relationship between two size distributions F and G whenever their *GL* curves (say GL_F and GL_G) intersect; thus, $GL_F(1) > GL_G(1)$ and $GL_F(t) < GL_G(t)$ (or $GL_F(1) < GL_G(1)$ and $GL_F(t) > GL_G(t)$) for at least one $t \in (0, 1)$. Alternatively, we identify full mean size dominance when the condition of *GL* ordering is fulfilled; thus, $GL_F(t) \geq GL_G(t)$ (or $GL_F(t) \leq GL_G(t)$) for every $t \in [0, 1]$. The condition $G \leq_{GL} F$ can also be expressed as follows: for every fixed p, $0 \leq p \leq 1$, the arithmetic mean of the first $100p$ percent of the ranked observations is greater for distribution F than for distribution G. This is why the *GL* curve provides a natural and reliable tool for ranking sources: the idea is to compare different sources by comparing their whole *GL* curves.

Thus, by virtue of Eq. (2) and considering that $\sum_{j=1}^{i} x_{(j)} = C - \sum_{j=1}^{T-i} x_j = C\left(1 - \tilde{F}(T-i)\right)$, the following conversion formula (by 'conversion' we mean, generally, how SF and RS distributions are related to each other) may be derived simply as follows:

$$GL\left(\frac{i}{T}\right) = \frac{\sum_{j=1}^{i} x_{(j)}}{T} = \frac{C\left(1 - \tilde{F}(T-i)\right)}{T} = \mu\left(1 - \tilde{F}(T-i)\right) \tag{3}$$

for every $i = 0,\ 1,\ 2, \ldots,\ T$. Note that $0 \notin \mathcal{N}$, then $\tilde{F}(T-i) = 0$ when $i = T$. This formula can be interpreted by saying that, for any pair of size distributions F and G with equal mean μ, the condition $G \leq_{GL} F$ between size distributions is equivalent to the condition $\tilde{G} \leq_{SD} \tilde{F}$ between their corresponding rank distributions $\tilde{G}$ and $\tilde{F}$, where $\leq_{SD}$ represents the first-order stochastic dominance relation. Recall that $\tilde{G} \leq_{SD} \tilde{F}$ if and only if $\tilde{G}(t) \geq \tilde{F}(t)$ for all t. Formula (3) can also be used to express the RS distribution function as a function of the SF *GL* curve, as follows: $\tilde{F}(j) = 1 - \mu^{-1} GL\left(1 - \frac{j}{T}\right), j = 0,\ 1,\ 2, \ldots,\ T$.

3.2. Second example

Two distributions with equal means can have significantly different *GL* curves. To compare the effect of the first-order stochastic dominance relation between rank distributions on the second-order stochastic dominance relation between the size distributions under the condition of equal means, let us consider a second citation distribution defined by a set of $T = 9$ articles. Let us suppose that there are four papers with zero citations; two papers with one citation; two papers with two citations, and one paper with eight citations.

To distinguish the distributions from those in the above example, let us denote by $\tilde{G}$ and G the RS and SF CDFs of these data. The total number of citations is now $C = 14$, but the mean size is the same as in the above example, in that $\mu = C/T = 14/9 = 1.556$ (equal mean sizes condition). Nevertheless, $\tilde{F}$ first-order stochastically dominates $\tilde{G}$ because $\tilde{G}(t) \geq \tilde{F}(t)$ for all t. Indeed, $\tilde{G}_1 = 0.571 > \tilde{F}_1$; $\tilde{G}_2 = 0.714 > \tilde{F}_2$; $\tilde{G}_3 = 0.857 > \tilde{F}_3$; $\tilde{G}_4 = 0.929 > \tilde{F}_4$; $\tilde{G}_5 = \tilde{G}_6 = \tilde{G}_7 = \tilde{G}_8 = \tilde{G}_9 = 1 > \tilde{F}_9$. Then, a second-order stochastic dominance relation between the size distributions F and G holds, as $G \leq_{GL} F$ (see Fig. 5). This represents a case of full mean size dominance, in that $GL_F(p) \geq GL_G(p)$ for every $p \in [0, 1]$.

3.3. SF Gini index as a function of the mean rank

As a by-product of the above formula (3), we can provide a simple expression for the Gini index of size distribution in terms of the mean rank. On the basis of Eq. (3), the SF Lorenz curve L can also be expressed as a function of the rank distribution $\tilde{F}$ as

$$L\left(\frac{i}{T}\right) = \mu^{-1} GL\left(\frac{i}{T}\right) = 1 - \tilde{\tilde{F}}(T-i),\ i = 0,\ 1,\ 2, \ldots,\ T,$$

with a continuous version that is given by the piecewise linear curve $L(p) = \mu^{-1} GL(p)$, $0 \leq p \leq 1$. For SF distributions with equal means, the *GL* ordering coincides with that of Lorenz. Since, within this class, the Gini index is consistent with the *GL* ordering, we would expect to find a relationship, via Eq. (3), through the mean rank, an index necessarily consistent with the first-order stochastic dominance. The SF Gini index, which is equal to one minus twice the area under the Lorenz curve, can

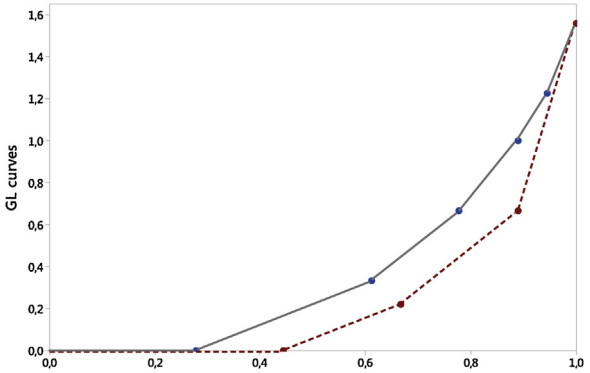

Fig. 5. *GL* curves of SF distributions F (solid line) and G (dashed line). The SF distribution F is defined by the pairs (*size*, *#sources*): (0, 5), (1, 6), (2, 3), (3, 2), (4, 1), (6, 1). The SF distribution G is defined by the pairs (*score*, *#sources*): (0,4), (1,2), (2,2), (8,1). Both distributions have the same mean, $\mu_F = \mu_G = 1.556$, but F second-order stochastically dominates G because $\tilde{F}$ first-order stochastically dominates $\tilde{G}$. Here we find $\tilde{\mu}_F = 4.786 > \tilde{\mu}_G = 1.929$. This demonstrates the utility of the mean rank statistics, in conjunction with the mean size, for adequately distinguishing between citation profiles.

then be obtained by integrating the function $2(p - L(p))$ resulting in $G = \int_0^1 2(p - L(p))\,dp = 1 - 2\int_0^1 L(p)\,dp$. Now, since $\int_0^1 L(p)\,dp = (2T)^{-1}\sum_{i=1}^{T}\left(L\left(\frac{i}{T}\right) + L\left(\frac{i-1}{T}\right)\right)$ (the area under the Lorenz curve is easily computed by a summation of the area of T trapezoids) and since, by virtue of the above Eq. (1), we have

$$\sum_{i=1}^{T} L\left(\frac{i}{T}\right) = \sum_{i=1}^{T}\left(1 - \tilde{F}(T-i)\right) = \sum_{s=0}^{T-1}\left(1 - \tilde{F}(s)\right) = \tilde{\mu}$$

and

$$\sum_{i=1}^{T} L\left(\frac{i-1}{T}\right) = \sum_{i=0}^{T-1}\left(1 - \tilde{F}(T-i)\right) = \sum_{s=1}^{T}\left(1 - \tilde{F}(s)\right) = \tilde{\mu} - 1,$$

we finally conclude that

$$G = 1 - 2(2T)^{-1}\left(2\tilde{\mu} - 1\right) = 1 - \frac{2\tilde{\mu} - 1}{T}.$$

As can be seen, this formula expresses the Gini index as a linear function of the mean rank. The Gini index of size distribution decreases as the mean rank increases. Note that the inequality $\tilde{\mu} \leq \frac{T+1}{2}$ is a consequence of the monotonicity of the rank density (as a function of the rank).

4. Modelling the *GL* curve

Statistical regularities in IPPs have been studied extensively. They generally depend on the context. In the citation analysis literature, several probabilistic models have been proposed to describe empirical SF and RS distributions. The former case includes, for example, the two-parameter Pareto distribution of the second kind (Schubert & Glänzel, 2007), Tsallis distribution (Bletsas & Sahalos, 2009), Weibull distribution (Bertoli-Barsotti & Lando, 2017b), and geometric distribution (Burrell, 2014; Bertoli-Barsotti & Lando, 2017a; Burrell, 2013). For the latter case, we may recall the Zipf law (Egghe & Rousseau, 2006; Rousseau, 2002), the stretched exponential model (Hirsch, 2005; Iglesias & Pecharroman, 2007; Laherrère & Sornette, 1998), and the discrete generalised beta distribution (Campanario, 2010; Mansilla et al., 2007; Martínez-Mekler et al., 2009; Naumis & Cocho, 2008).

A simple interpretative parametric representation model for the *GL* curve is now possible using the conversion formula (3). We assume that the rank distribution follows a simple one-parameter model: the shifted-geometric distribution with probability function $\tilde{f}^*(i) = \theta(1-\theta)^{i-1}$, $i = 1, 2, 3, \ldots$ and zero elsewhere, where θ is a real parameter, $0 < \theta < 1$. As is well-known, the CDF of this random variable is available in closed form as

$$\tilde{F}^*(i;\theta) = \sum_{j=1}^{i} \tilde{f}^*(j) = 1 - (1-\theta)^i,\ i = 1, 2, 3, \ldots$$

and the corresponding expectation produces $\mu^* = \sum_{i=1}^{\infty} i \cdot \tilde{f}^*(i) = \theta^{-1}$. Our goal here is not to achieve the best fit possible for the rank distribution (in general, empirical rank distributions may be very well fitted with beta-like function distributions; Naumis & Cocho, 2008), but rather to obtain an effective and easy-to-interpret model for the *GL* curve of SF distribution.

Taking the mean rank $\tilde{\mu}$ as a proxy for the expectation μ^*, we may now approximate the empirical rank distribution $\tilde{F}$ with the function $\tilde{F}^*\left(i;\tilde{\mu}^{-1}\right)/\tilde{F}^*\left(T;\tilde{\mu}^{-1}\right)$, $i = 1, 2, \ldots, T$, using, at the first instance, a truncated version of the geometric

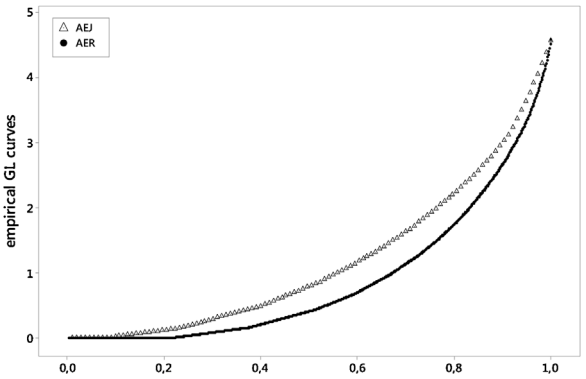

Fig. 6. Two journals with similar mean sizes may have significantly different citation distributions. Here, we show the empirical *GL* curves of the SF distributions for the *American Economic Journal: Applied Economics* (*AEJ*) and the *American Economic Review* (*AER*). Both journals have the same mean size (4.57), but the *AEJ* has a larger mean of the first 100p% of the ranked observations (sizes), for every fixed p, $0 \le p \le 1$.

distribution. To simplify further, if T is large enough, we can drop the constant factor $\tilde{F}^*\left(T; \tilde{\mu}^{-1}\right) \cong 1$, and take $\tilde{F}^*\left(i; \tilde{\mu}^{-1}\right)$ as an estimate of the rank distribution $\tilde{F}$ for $i = 1, 2, \ldots, T$. Then, by virtue of Eq. (3), we obtain the following general model for the *GL* curve of size distribution:

$$\hat{GL}\left(\frac{i}{T}\right) = \mu\left(1 - \tilde{\mu}^{-1}\right)^{T-i},\ i = 1,\ 2, \ldots,\ T \tag{4}$$

As seen above, this function $\hat{GL}$ extends uniquely to a function, still denoted by $\hat{GL}$, defined on the set $[0, 1]$ by linear interpolation. Formula (4) yields a parametric model of the *GL* curve generated according to the general formula (3) and the specific choice of the geometric model for the rank distribution. It should be stressed that this estimation approach is 'indirect' because it is obtained through a parameterisation of the rank distribution and not, as is usually done in other studies, through a direct parameterisation of the size distribution or the *GL* curve itself. What is remarkable about the estimating function (4) is that it depends only on three simple basic quantities, μ, $\tilde{\mu}$, and T: the mean size, mean rank, and maximal rank, respectively. Despite its simplicity, it does not significantly limit the generality or flexibility of the model. We illustrate its goodness-of-fit capabilities in the next section.

5. Application to citation data

We apply our model (4) to a case of empirical *GL* curves deriving from a set of documents (sources) published by a given journal that receive citations (items) for fixed specific citation and publication windows. We obtain a quasi-impact factor for a journal (JIF) by dividing the number of citations of it by the number of its articles for given citing and cited windows. By choosing a suitable pair of citing–cited windows and a suitable set of 'citable' documents, we may refer, without any loss of generality, to the mean size $\mu = C/T$ as a 'JIF' (or JIF-like measure).

Lando and Bertoli-Barsotti (2017) recommended the use of indicators consistent with the *GL* ordering as alternative or complementary tools for overcoming the weaknesses of the JIF. Since two journals with equal JIFs can have very different *GL* curves (see Fig. 6), the idea is to compare the journals by analysing their entire citation distributions. This can be done by comparing their *GL* curves. Lando and Bertoli-Barsotti (2017) also proposed the stabilised JIF (s-JIF) index, which is equal to twice the area under the *GL* curve, as a simple measure consistent with the *GL* ordering that takes into account both the JIF and deviations from it. The s-JIF obeys the following principle: given two citation distributions with equal means, the one with the greater dispersion should be preferred. The s-JIF, a sort of 'corrected impact factor', is more reliable than the JIF because it also contains information about deviations from this statistic. It considers that the 'reliability' of the JIF inherently depends on the dispersion of the size distribution. Investigations along these lines have been recently discussed by Cockriel and McDonald (2018), who suggested an entire new class of indices that consider the dispersion factor.

The theoretical model of the proposed *GL* curve (4) depends on the mean size μ (to be intended as the JIF or JIF-like measure for alternative choices of citation and publication windows), as well as on the maximal rank T as a scaling constant, but also on a third parameter, the mean rank $\tilde{\mu}$, which helps indicate whether the *GL* ordering is effective or not. This corroborates the utility of the mean rank statistics, in conjunction with the mean size, for enabling a more thorough evaluation of a journal's impact.

To evaluate the goodness-of-fit of model (4) to the empirical data, we exploited the same dataset considered by Bertoli-Barsotti and Lando (2017a), to which the reader can refer for further details. As reported there, the data were downloaded from Scopus during the last week of April 2016 and comprise more than 74,000 citations received in 2014 by the top 100 journals within the Scopus subject area 'Economics, Econometrics and Finance', ranked according to the Scopus 2014 impact factor (at that time called 'Impact per Publication'). For each journal (only journals with a minimum of 50 published documents during that period are considered in the list), citations received during 2014 to documents published in it were

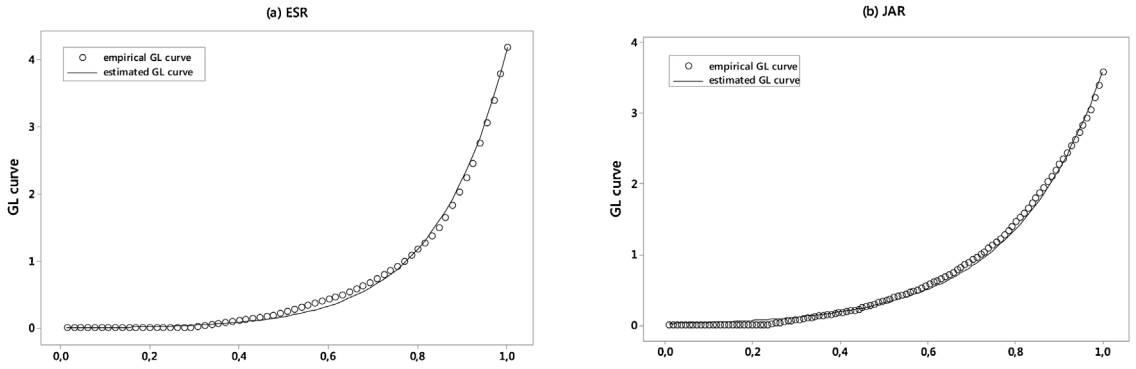

Fig. 7. *GL* curves for *Economic Systems Research* (*ESR*) and *Journal of Accounting Research* (*JAR*). Circles represent points of the empirical *GL* curve, while fits using formula (4) for the estimated *GL* curve are shown as solid lines.

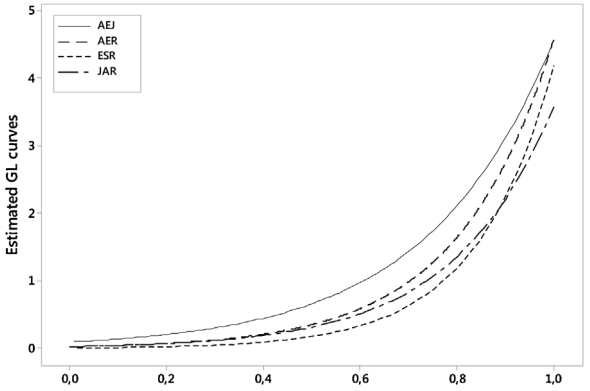

Fig. 8. Fitted *GL* curves for *American Economic Journal: Applied Economics* (*AEJ*), *American Economic Review* (*AER*), *Economic Systems Research* (*ESR*) and *Journal of Accounting Research* (*JAR*). The estimated *GL* curves confirm that *AEJ* dominates the remaining three journals; *AER* dominates both *ESR* and *JAR*; while there is a spurious mean size dominance relationship between (the SF distributions of) *ESR* and *JAR*. The simple scalar value of the mean size, $\mu = GL(1)$, is clearly not able to be distinguished at this level of comparison.

calculated during a three-year citation window (2011–2013). These journals contribute 19,889 documents. Non-citable documents (e.g. notes) were excluded to obtain sets of publications as close as possible to those considered by Scopus for the computation of the impact factors, even if differences were observed between the computed mean sizes C/T and the corresponding impact factors for 2014 published by Scopus. Fig. 7 shows two representative examples from this dataset. The empirical *GL* curves of *Economic Systems Research* (*ESR*; $\mu = 4.18$) and *Journal of Accounting Research* (*JAR*; $\mu = 3.59$) are compared with the estimated *GL* curves obtained by formula (4). Both the fits are excellent, with correlation coefficients (between observed and fitted values of *GL* curves) above 0.99. *ESR* has a higher JIF, but *JAR* is characterised by a lower level of dispersion. A spurious mean size dominance relationship is revealed between these (empirical) distributions. This kind of relationship is confirmed by the estimated *GL* curves, as can be seen in Fig. 8.

We have verified that formula (4) appears to be accurate for almost all journals. The correlation coefficients are all above 0.990, except for four cases with values between 0.965 and 0.990. Overall, the mean (and median) of the correlation coefficients obtained from the set of all journals under consideration is 0.995. The first quartile is equal to 0.994. A single outlier is present (*Experimental Economics*), whose correlation coefficient is 0.965 (the observed smaller value). However, this journal is characterised by a very skewed citation distribution, with a high concentration of citations (the highest by far of all the journals tested) among a small number of articles. If we define pillar articles as those belonging to the set of the most-cited 5% for a given journal (over the period considered), the proportion of citations obtained by the pillar articles, $\Pi = \frac{\sum_{i=1}^{p} x_i}{C}$, where $p = \lfloor 0.05\,T \rfloor$ ($where \lfloor z \rfloor denotes the floor function$), is a raw index of the concentration of the citations for the journal. For *Experimental Economics*, we find $\Pi = 0.54$, which means that the 5% most-cited articles received 54% of all the citations of this journal. This high percentage leads to a certain degree of misfit in both ending tails of its estimated *GL* curve. Overall, the fit is very good if the level of Π does not exceed a value around 0.36 (see Fig. 9).

6. Conclusion

As Egghe and Waltman (2011) point out, both the SF and RS (that they called rank-frequency) approaches should convey the same amount of information. Nevertheless, the latter approach offers advantages: (1) a more regular support set of the

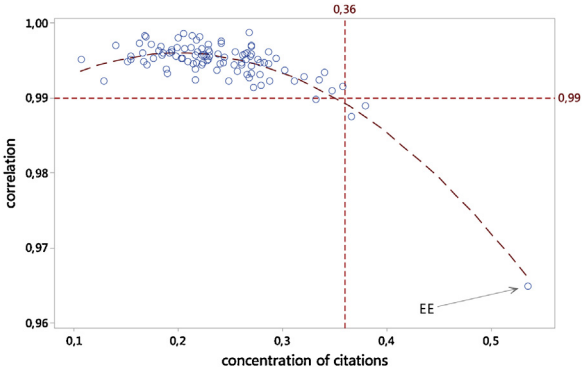

Fig. 9. Correlation between observed and fitted values of *GL* curves as a function of the concentration of citations, measured by the proportion Π of citations received by the top 5% of most-cited articles for all 100 journal considered in this study (each circle represents a journal). The figure illustrates that the fit begins to rapidly degrade only when this proportion exceeds 36%, a very high value. In the case of *Experimental Economics* (EE), we observed an anomalous level of concentration in that the 5% most-cited articles received 54% of all the citations of the journal. For this item, the correlation falls below 0.97.

distribution, with the support of the RF distribution being an *initial segment* of $\mathbb{N}$; (2) a more simple density function shape, as the rank density is, necessarily, a monotone non-increasing function (with a typical reverse J-shaped form); and (3) a more efficient data structure, in terms of Shannon's information measure, as argued by Brookes (1980, p. 212–213). This may explain why the estimation approach based on a modelisation of the RF distribution tends to be more effective than its counterpart based (directly) on the SF framework.

In this work, we first analysed the duality between SF and RS frameworks and presented a 'conversion formula' that relates rank distribution to size distribution in a fully discrete setting. The formula proves that the first-order stochastic dominance between RS CFDs corresponds to second-order stochastic dominance between SF CFDs under the condition of equal SF means.

Subsequently, as an application of the obtained formula, we introduced a simple parametric model of a *GL* curve (of the SF distribution), depending on only three quantities: the mean rank, the mean size and, as a mere scaling factor, the maximal rank. Our *GL* curve model proves that the first moment of both the RS and SF distributions are the only crucial parameters for determining the shape of the *GL* curve: the mean size captures the magnitude of the size variable, while the mean rank, as a linear transformation of the Gini index, captures its variability. Despite its simplicity, the formula shows an excellent agreement for all the journals of the citation dataset considered in this study, with the exception of specific cases of abnormal concentration of citations in a few articles. Of course, the degree of our model's universality remains to be further investigated.

Author contribution

Lucio Bertoli-Barsotti: Conceived and designed the analysis, Performed the analysis, Wrote the paper.
Tommaso Lando: Conceived and designed the analysis, Collected the data, Performed the analysis.

Funding

This research was supported by the Czech Science Foundation (GACR) under project 17-23411Y (to T.L.)

References

Bertoli-Barsotti, L., & Lando, T. (2017a). A theoretical model of the relationship between the h-index and other simple citation indicators. *Scientometrics*, *111*(3), 1415–1448.
Bertoli-Barsotti, L., & Lando, T. (2017b). The h-index as an almost-exact function of some basic statistics. *Scientometrics*, *113*(2), 1209–1228.
Bletsas, A., & Sahalos, J. N. (2009). Hirsch index rankings require scaling and higher moment. *Journal of the American Society for Information Science and Technology*, *60*, 2577–2586.
Brookes, B. C. (1980). The foundations of information science: Part II. Quantitative aspects: Classes of things and the challenge of human individuality. *Journal of Information Science*, *2*(5), 209–221.
Burrell, Q. L. (2005). Symmetry and other transformation features of Lorenz/Leimkuhler representations of informetric data. *Information Processing & Management*, *41*(6), 1317–1329.
Burrell, Q. L. (2013). The h-index: A case of the tail wagging the dog? *Journal of Informetrics*, *7*, 774–783.
Burrell, Q. L. (2014). The individual author's publication-citation process: Theory and practice. *Scientometrics*, *98*, 725–742.
Campanario, J. M. (2010). Distribution of ranks of articles and citations in journals. *Journal of the American Society for Information Science and Technology*, *61*(2), 419–423.
Cockriel, W. M., & McDonald, J. B. (2018). The influence of dispersion on journal impact measures. *Scientometrics*, *116*, 609–622.
Egghe, L. (1990). The duality of informetric systems with applications to the empirical laws. *Journal of Information Science*, *16*(1), 17–27.
Egghe, L. (2005a). *Power laws in the information production process: Lotkaian informetrics*. Oxford: Elsevier.
Egghe, L. (2005b). Relations between the continuous and the discrete Lotka power function. *Journal of the American Society for Information Science and Technology*, *56*(7), 664–668.

Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, *69*(1), 121–129.
Egghe, L., & Rousseau, R. (2012). The Hirsch index of a shifted Lotka function and its relation with the impact factor. *Journal of the American Society for Information Science and Technology*, *63*(5), 1048–1053.
Egghe, L., & Waltman, L. (2011). Relations between the shape of a size-frequency distribution and the shape of a rank-frequency distribution. *Information Processing & Management*, *47*(2), 238–245.
Glänzel, W., & Moed, H. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*(2), 171–193.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.
Iglesias, J., & Pecharroman, C. (2007). Scaling the *h*-index for different scientific ISI fields. *Scientometrics*, *73*, 303–320.
Laherrère, J., & Sornette, D. (1998). Stretched exponential distributions in Nature and Economy: Fat tails with characteristic scales. *The European Physical Journal B*, *2*, 525–539.
Lando, T., & Bertoli-Barsotti, L. (2017). Measuring the citation impact of journals with generalized Lorenz curves. *Journal of Informetrics*, *11*(3), 689–703.
Li, W., Miramontes, P., & Cocho, G. (2010). Fitting ranked linguistic data with two-parameter functions. *Entropy*, *12*(7), 1743–1764.
Mansilla, R., Köppen, E., Cocho, G., & Miramontes, P. (2007). On the behavior of journal impact factor rank-order distribution. *Journal of Informetrics*, *1*(2), 155–160.
Martínez-Mekler, G., Martínez, R. A., del Río, M. B., Mansilla, R., Miramontes, P., & Cocho, G. (2009). Universality of rank-ordering distributions in the arts and sciences. *PloS One*, *4*(3), e4791.
Naumis, G. G., & Cocho, G. (2008). Tail universalities in rank distributions as an algebraic problem: The beta-like function. *Physica A Statistical Mechanics and Its Applications*, *387*(1), 84–96.
Rapaport, A. (1978). Rank-size relations. *International Encyclopedia of Statistics*, *2*, 847–854.
Rousseau, R. (1990). Relations between continuous versions of bibliometric laws. *Journal of the American Society for Information Science*, *41*(3), 197–203.
Rousseau, R. (2002). George Kingsley Zipf: life, ideas, his law and informetrics. *Glottometrics*, *3*, 11–18.
Rubin, J. E. (1967). *Set theory for the mathematician*. New York: Holden-Day.
Sarabia, J. M., Prieto, F., & Trueba, C. (2012). Modeling the probabilistic distribution of the impact factor. *Journal of Informetrics*, *6*(1), 66–79.
Schubert, A., & Glänzel, W. (2007). A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, *1*(3), 179–184.
Shorrocks, A. F. (1983). Ranking income distributions. *Economica*, *50*(197), 3–17.
Thistle, P. D. (1989). Ranking distributions with generalized Lorenz curves. *Southern Economic Journal*, 1–12.