SHAPING THE CITATION-PAPER RANK DISTRIBUTIONS: BEYOND HIRSCH'S MODEL

Vladimir Atanassov, Ekaterina Detcheva

Abstract*: It is well known that the h-index has been originally introduced by analyzing a simple deterministic model of individual's scientific activity. The basic features of Hirsch's model are listed as follows: i) constant publication rate in time; ii) constant number of citations gained by each paper per unit time. One of its predictions is the formation of (approximately) linear negative slope citation-paper rank distribution. However, the actual citation-paper rank distributions obtained by analyzing real life scientometric data are, apart from some rare exceptions, convex and far from linear.*

In this paper we consider two possible reasons that might explain the deviation from linear negative-slope (Hirsch) citation-paper rank distribution. The first one is an increasing in time publication rate, reflecting the fact of growing productivity in the course of a scientific career, as well as the increase of publication rates on global scale. The second one is an increasing in time number of citations gained by individual papers per unit time, associated, e.g. with improving publication quality or growing popularity. Both factors lead to a shaping of convex citation-paper rank distribution from the linear one, in a way that generalizes Hirsch's results.

Keywords*: citation-paper rank distributions, Hirsch's model, linear negative-slope distribution, convex distributions, time dependence, empirical relationships, scientometric data analysis*

ACM Classification Keywords*: H. Information Systems, H.2. Database Management, H.2.8. Database applications, subject: Scientific databases; I. Computing methodologies, I.6 Simulation and Modeling, I.6.4. Model Validation and Analysis*

Introduction

Since the beginning of scientometric studies citation dynamics has been established as one of their most fascinating parts. This might be due to the fact that it gives more answers to questions 'why?' rather than 'who?' the latter being of main interest for the application-oriented citation analysis. Both areas complement each other by exchanging data and theoretical models. Citation dynamics itself has miscellaneous aspects, *e.g.* the psycho-socio-economical issues of the publication-citation process ([Asknes, 2005; Bar-Ilan, 2008; Markov et al, 2013]), including dynamics of visibility and quality, as well as citation motivation and tactics. The main efforts, however, have been concentrated on studying the citation distributions and the (stochastic) processes behind them. A bouquet of references, not necessarily going on the same line, follows further on: [de Solla Price, 1976], [Bookstein, 1990 a, b], [Sen, 1996], [van Raan, 2001 a, b], [Egghe, 2005], [Burrell 2007 a, b]. Studies on a global scale (time spans sometimes exceeding a century, worldwide range, various scientific fields, huge amount of citation data) have been performed ([Redner, 2005], [Wallace et al, 2008]) to find out what kind of distribution function fits best a given data set, to explain deviations, to refine details as distribution tails *etc*. The variety of distributions starts with the one parameter beta function resulting from a Cumulative Advantage (or Success Breeds Success, Preferential Attachment) process ([de Solla Price, 1976]). The latter reproduces in a limiting case Bradford and Lotka laws, as well as Pareto and Zipf distributions (comments on all this could be found in

[Bookstein, 1990 a, b] and [Watts and Gilbert., 2011]). It is worth to note the two-step (first to publish and second, to get cited) competition process [van Raan, 2001 a, b], where a paper-citation distribution represented by a modified Bessel function (K_0) has been obtained rather than the expected power-law one. A rather comprehensive model [Burrell 2007 a, b] assumes Poisson publication process of rate equal to the publication rate and gamma-distributed citation rate for an individual paper, which eventually leads to beta-distributed citations to individual publications. Further on, [Wallace et al, 2008] prefer the stretched exponential function to fit their data. The problems of Zipf's law to explain empirical citation distributions motivated [Gupta et al, 2005] and [Peterson et al, 2010] to suggest a 'truncated' power-law distribution, specified by cumulative advantage processes or random choice (exponential decay), for large and small citation counts, respectively. Discrete generalized beta distribution has been implemented by [Petersen et al, 2011] to analyze rank citation data of 300 distinguished physicists. In a recent study on modeling citation dynamics ([Eom and Fortunato, 2011]) it has been found that citation distributions are best described by a shifted power-law, the citation dynamics is determined by citation bursts appearing few years after publication and the microscopic mechanism, responsible for the evolution of the citation networks is the preferential attachment with time-decaying initial attractiveness (see also [Redner, 2005]). Preferential attachment together with aging (loss of novelty in the course of time) and fitness (overall community response) are the basic mechanisms responsible for the long-term citation behavior of individual papers for [Wang et al, 2013]. It has been also shown [Golosovsky and Solomon, 2012] that in addition to the preferential attachment, papers exceeding a certain threshold of overall citation count (50-70, meets about 10 percent of all papers) continue to be cited, due to an epidemic-like self-excited cascade process.

One may conclude from this somewhat hectic development that citation dynamics will remain on the top of scientometric research (probably) for a long time. All these theoretical and empirical studies have additional motivation in practical applications, like improvement of scientific assessment – from establishing the milestones of excellence in science up to introducing more and more scientometric indices that better represent citation paper rank distributions.

The scope of this paper is studying some possible reasons for deviation of citation-paper rank distributions from the linear negative slope one, obtained in [Hirsch, 2005] by assuming time-independent publication and citation rates. Hirsch's model, sometimes commented as elementary, base and averaged one [Liu and Rousseau, 2008], inspired us to treat the problem of publication-citation processes from pure dynamical point of view, rather than entering into much detail by following the modern trends in statistical physics. For the purpose of our analysis we consider (as usual in similar studies) continuous real valued distributions of real argument, without further discussion on integer-real and discrete-continuous aspects of citation-paper rank distributions ([Bookstein, 1990b], [Atanassov and Detcheva, 2013]).

The paper is organized as follows. In the first section we briefly consider Hirsch's model and its assumptions and consequences. A set of equations that govern the publication-citation process for time-dependent sources (publication rate and citation rate of each paper – not to be mistaken with the 'items being produced by sources' approach in [Egghe, 2005]) is derived in the second section and the conditions for existence of rank convex, linear and concave distributions are obtained. The third section contains analysis of the basic equations for the case of stationary time-dependent papers' citation rate. Next three sections illustrate the theory with examples of citation-paper rank distributions (obtained in explicit form or numerically) for several combinations of more or less realistic time-dependent sources.

Hirsch's model outline

In his well known paper [Hirsch, 2005] where the h-index has been introduced, Hirsch considered a simple deterministic publication-citation process, briefly described as follows. An author publishes p papers per unit time (*e.g.* per year) and each published paper gains c citations per unit time (year). Assuming constant p and c we arrive at the following equations for the paper P published at time τ ($P+1$ represents the total number of author's papers at this time) and for the citations C gained by this paper within a time interval $t-\tau \geq 0$:

$$P = p\tau\ ,\ C = c(t-\tau)\,. \tag{1}$$

At time moment t this model produces a *linear negative slope* citation-paper rank distribution (or *Hirsch distribution*)

$$C = C_{\max} - sP\ , \tag{2}$$

where the slope $s = c\,/\,p = C_{\max}\,/\,P_{\max}$, $C_{\max} = ct$, $P_{\max} = pt$. Since s is time-independent, the distribution does not change its shape in the course of the publication-citation process.

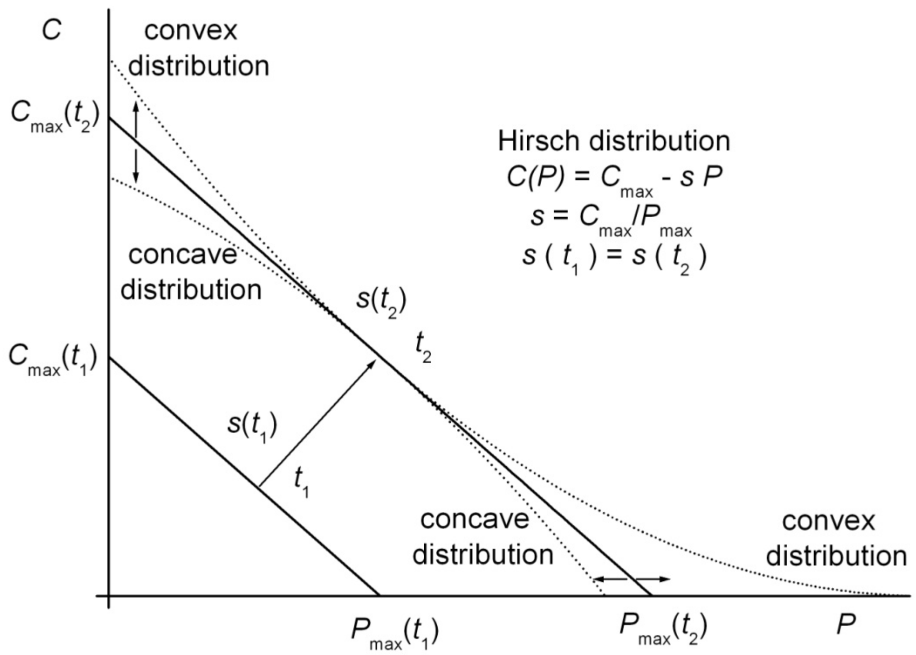

Figure 1. Time evolution of a linear negative slope (Hirsch) citation-paper rank distribution with corrections for surplus or deficit of citations C and/or papers P, respectively

Two questions appear in connection with the distribution (2): i) why the real life citation-paper rank distributions are usually convex and seldom linear or concave; ii) what happens if the publication and citation rates p and c change in time and cannot be considered as constants. Some intuition leads us to a conclusion (illustrated by Fig. 1), that *increasing* in time publication rate p would result in a surplus of papers in the distribution tail containing least cited papers; in the same way, *increasing* in time citation rate c would result in a surplus of citations in the core of most cited papers. Both factors could lead to formation of a *convex-shaped* citation-paper rank distribution. Analogously, *decreasing* in time publication rate p and/or citation rate c could produce a *concave* citation-paper rank distribution. A combination of increasing and decreasing in time p and c could even form distributions with at least one *inflection* point.

One may find some support of these considerations in the results of a case study on time evolution of scientific activity (see [Atanassov, 2012] for more detail). On Fig. 2a we observe a sharp increase in publication rate (productivity). Fig. 2b demonstrates, apart from some statistical variability, a slight positive trend in *average* citation rate per paper (impact). The corresponding extremely convex citation-paper rank histogram is shown on Fig. 2c. Similar increase in *average* citation rate per paper could be deduced also from the results of a global study on citation distributions [Wallace et al, 2008]. This makes clear the need of generalizing Hirsch's model for time-dependent publication and citation rates. This could be done by deriving equations that govern the publication-citation process and eventually form the citation-paper rank distribution.

Equations governing the publication-citation process

Let us assume that a scientist publishes $p > 0$ papers per unit time. At a moment τ the paper count $P(\tau)$ is

$$P(\tau) = \int_0^\tau dt' p(t') . \qquad (3)$$

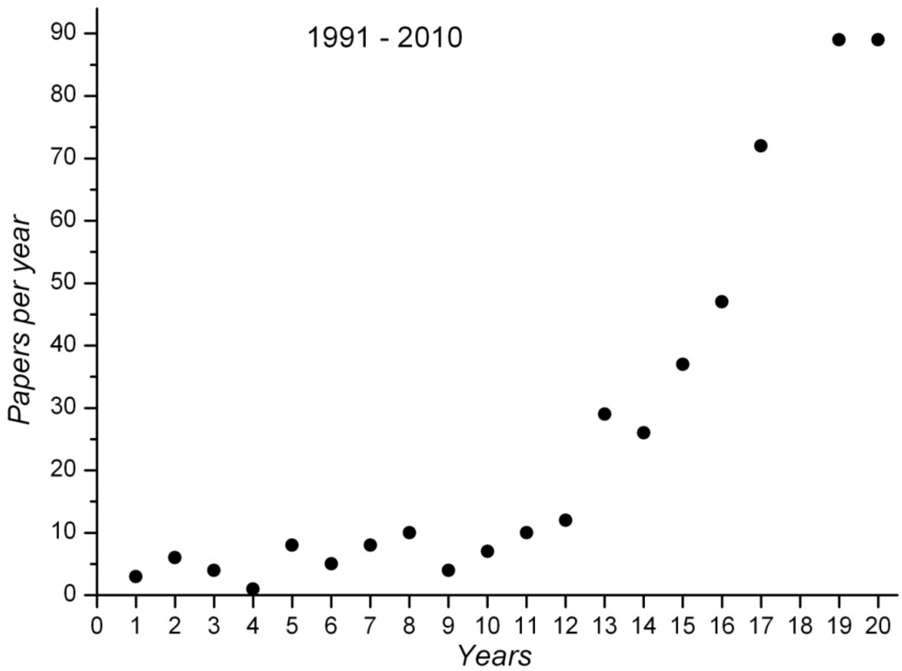

Figure 2a. Example for time evolution of *average* publication rate

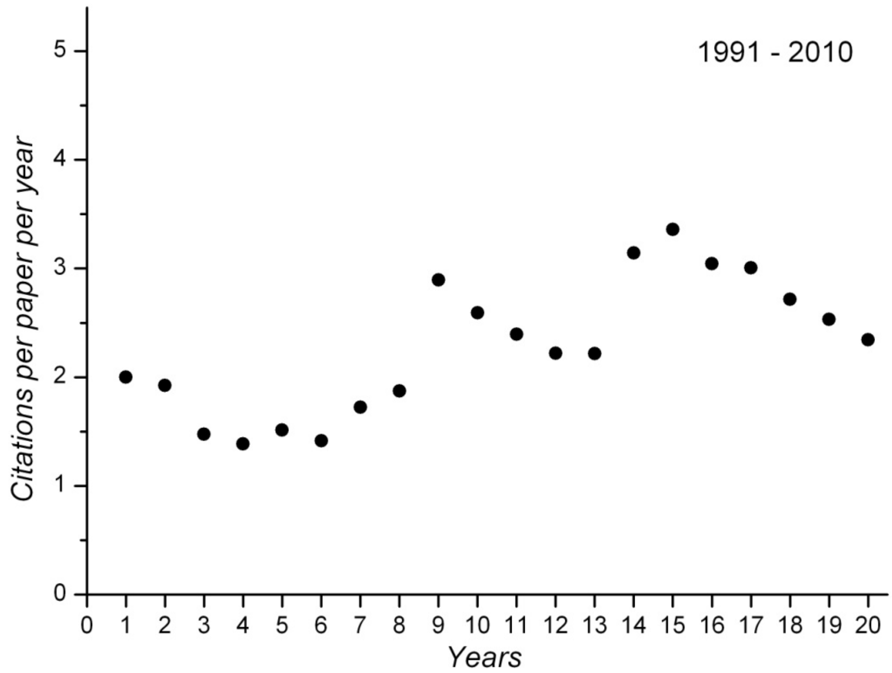

Figure 2b. Example for time evolution of *average* citation rate per paper

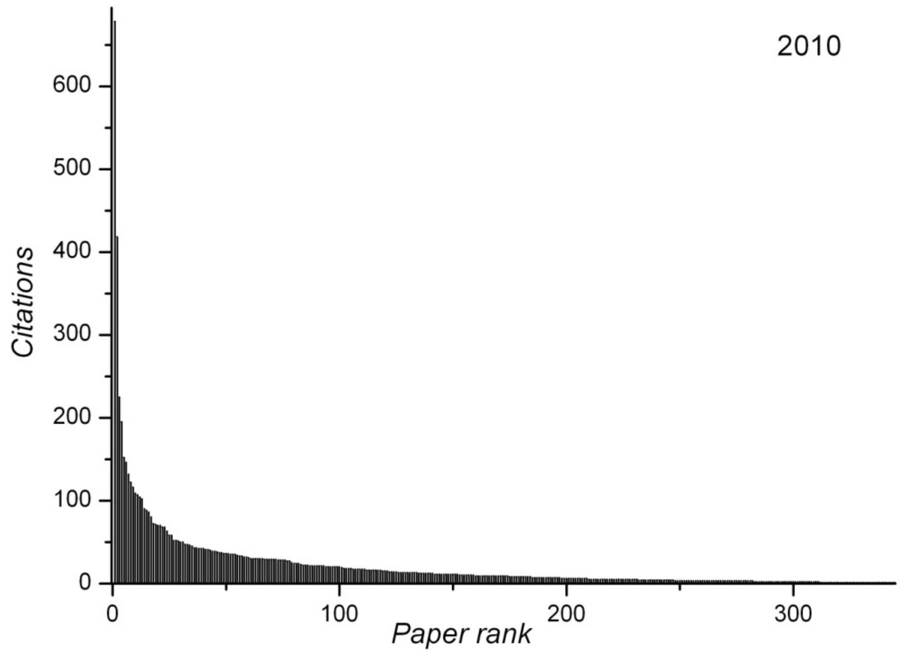

Figure 2c. Citation-paper rank histogram for the case study presented on Figure 2a, b

Equation (3) establishes one-to-one correspondence between a paper P and the time of its publication τ and

$$\frac{d\tau(P)}{dP} = \frac{1}{p\left[\tau(P)\right]}.$$

Each paper $P(\tau)$ starts earning $c(\tau,t) \geq 0$ citations per unit time, provided that $t \geq \tau$ (causality). Hence, for the number of citations gained by a paper $P(\tau)$ within a time interval $[\tau, t]$ we have

$$\underset{\sim}{C}(\tau,t) = \int_{\tau}^{t} dt'\, c(\tau,t'),\ \tau \leq t. \tag{4}$$

The set of equations (3)-(4) govern the time evolution of the publication-citation events. The citation-paper distribution at time t is

$$C(P,t) \equiv \underset{\sim}{C}\left[\tau(P),t\right] = \int_{\tau(P)}^{t} dt'\, c\left[\tau(P),t'\right]. \tag{5}$$

It is worth to emphasize the difference between τ (time of publication of paper P) and t (time of 'measurement', when all citations for all published papers have been taken into account to construct the histogram/distribution).

It should be noted that Eq. (5) alone does not necessarily specify a citation-paper *rank* distribution. The latter is obtained (when processing real life scientometric data) by rearranging the citation-paper sequence in descending order of the number of citations gained, *i.e.* most cited paper placed first. The model considered here cannot account for such procedure; it can only rely on the specifics of the publication-citation process that automatically produce rank distribution. Therefore, an additional condition must be imposed on the publication and citation sources p and c . By definition, a rank distribution $C(P,t)$ satisfies $\partial C(P,t) / \partial P \leq 0$. We have

$$\frac{\partial C(P,t)}{\partial P} = \frac{1}{p\left[\tau(P)\right]}\left\{\int_{\tau(P)}^{t} dt'\left[\frac{\partial c(\tau,t')}{\partial \tau}\right]_{[\tau(P),t']} - c\left[\tau(P),\tau(P)\right]\right\}. \tag{6}$$

Hence the condition for existence of rank distribution is the inequality

$$\int_{\tau(P)}^{t} dt'\left[\frac{\partial c(\tau,t')}{\partial \tau}\right]_{[\tau(P),t']} \leq c\left[\tau(P),\tau(P)\right] \tag{7}$$

that must be fulfilled throughout the publication-citation process.

If Eq. (7) takes place it makes sense to define a maximum citation count $C_{\max}$, representing the number of citations to the first, most cited paper:

$$C_{\max}(t) \equiv C(0,t) = \int_{0}^{t} dt'\, c(0,t'). \tag{8}$$

Further on, bearing in mind that τ cannot exceed t one may introduce a maximum paper number (equal to the maximum number of papers) of *nonzero* citation count:

$$P_{\max}(t) \equiv P(t) = \int_{0}^{t} dt' p(t'), \tag{9}$$

where, obviously, $C\left[P_{\max}(t),t\right] = 0$. One should bear in mind that, although t , $C_{\max}$ and $P_{\max}$ may tend to infinity, in order to have a *distribution* the total number of citations

$$C_{total}(t) \equiv \int_0^{P_{\max}} C(P,t)dP = \int_0^t d\tau p(\tau) \underset{\sim}{C}(\tau,t) = \int_0^t d\tau p(\tau) \int_\tau^t dt' c(\tau,t') \quad (10)$$

must remain finite.

The citation-paper rank distribution shape is determined by the its second derivative

$$\frac{\partial^2 C(P,t)}{\partial P^2} = \left\{ \frac{1}{p^2(\tau)} \left[\int_\tau^t \frac{\partial^2 c(\tau,t')}{\partial \tau^2} dt' - 2\frac{\partial c(\tau,t'')}{\partial \tau} - \frac{\partial c(\tau,t'')}{\partial t''} - \frac{dp(\tau)}{d\tau} \frac{\partial C(P,t)}{\partial P} \right] \right\}_{\tau=\tau(P),t''=\tau(P)} . \quad (11)$$

Thus we have a convex, (negative-slope) linear or a concave distribution depending on whether $\partial^2 C(P,t) / \partial P^2$ is positive, zero or negative, respectively.

Analysis for a stationary citation rate

We start our analysis of the basic set of equations derived in the previous chapter by considering publication-citation processes with a *stationary* paper citation rate

$$c(\tau,t) \equiv c_s(t-\tau) \geq 0, \ \tau \leq t . \quad (12)$$

The meaning of this approach is that all published papers have the same citation behavior, however shifted in time due to the different moments of publication τ. The dependence of individual paper's citation rate on its age $t-\tau$ could be quite complicated (see, *e.g.* [Wang et al, 2013]). The empiric results ([Asknes, 2003]) demonstrate that a typical behavior (of rather universal character) consists of an approximately linear growth, reaching a maximum followed by an exponential-like decay. Another type of citation behavior can be observed for the so called *sleeping beauties* ([van Raan, 2004]) or *revived classics* [Redner, 2005]. Obviously, Eq. (12) cannot hold simultaneously for both types. On the other side, the *average* citation rate per paper on a global scale could be considered as slowly varying (increasing) within the hundred years time interval under investigation [Wallace et al, 2008]. Stationarity approach might require reducing this time interval in order to make such variations small enough.

Now the number of citations gained by a paper $P(\tau)$ within a time interval $[\tau,t]$ is

$$\underset{\sim}{C}(\tau,t) \equiv \underset{\sim}{C}_s(t-\tau) = \int_0^{t-\tau} dt' c_s(t'), \ \tau \leq t . \quad (13)$$

The citation-paper distribution at time t is

$$C(P,t) \equiv \underset{\sim}{C}_s\left[t-\tau(P)\right] = \int_0^{t-\tau(P)} dt' c_s(t') . \quad (14)$$

Further on, from Eq. (6) we obtain

$$\frac{\partial C(P,t)}{\partial P} = -\frac{c_s\left[t-\tau(P)\right]}{p\left[\tau(P)\right]} \leq 0 . \quad (15)$$

Hence *stationary citation rates always produce rank distributions*.

The expression for $\partial^2 C(P,t) / \partial P^2$ (Eq. (11)) is reduced to

$$\frac{\partial^2 C(P,t)}{\partial P^2} = \frac{1}{p^3\left[\tau(P)\right]} \left\{ c_s\left[t-\tau(P)\right] \left[\frac{dp(\tau)}{d\tau}\right]_{\tau=\tau(P)} + p\left[\tau(P)\right] \frac{dc_s\left[t-\tau(P)\right]}{dt} \right\} . \quad (16)$$

It follows from Eq. (16) that:

- At least one of the sources p, c_s must *increase/decrease* in time in order to have a *convex/concave* citation-paper rank distribution;
- If *both* p and c_S are *increasing/decreasing* functions of time, the sources produce *convex/concave* distributions;
- For a constant citation rate c_s the distribution is *convex, linear or concave*, depending on whether the publication 'acceleration' $dp / d\tau$ is *positive, zero or negative*;
- For a constant publication rate p the distribution is *convex, linear or concave*, depending on whether the citation 'acceleration' of each paper dc_S / dt is *positive, zero or negative*;
- A citation-paper rank distribution is (negative slope) linear one if and only if the ratio c_s / p does not depend on the time of publication τ (or, which is equivalent; on the paper count P);
- A citation-paper rank distribution has no inflexion points if and only if the ratio c_s / p is a monotonously increasing/decreasing function of the time of publication τ (or, equivalently; on the paper count P);

These conclusions are drawn for quite general case of sources: strictly positive publication rate p, a nonnegative stationary papers' citation rate c and finite total number of citations (Eq. (10)).

Examples I (constant paper citation rate)

The simplest example of a publication-citation process is the one considered by Hirsch [Hirsch, 2005]. It is characterized with a constant publication rate $p > 0$ and a constant citation rate for each paper $c \geq 0$. The citation-paper rank distribution (Eq. (2)) as well as the relations for $P_{\max}$ and $C_{\max}$ are easily obtained from Eq. (5) (or Eq. (14)) and Eqs. (8), (9), respectively.

Further on in this section we consider the effect of *time-dependent* publication rate $p(\tau)$ on the shape of citation-paper rank distributions by assuming constant papers' citation rate $c \equiv c_0 \geq 0$. This represents the simplest case of a stationary citation rate. Therefore, all distributions discussed later on are *rank* distributions. They are convex, linear or concave for increasing, constant or decreasing in time publication rates $p(\tau)$, respectively.

- **linear dependence** $p(\tau) = p_0 + p_1\tau$,

where $p_0 > 0$ and p_1 are constants and $p_0 > -p_1 t$ must be fulfilled to keep p positive. By solving the integral in Eq. (3) one arrives at

$$P(\tau) = p_0\tau + \frac{1}{2}p_1\tau^2,\ P_{\max}(t) \equiv P(t) = p_0 t + \frac{1}{2}p_1 t^2 \tag{17}$$

and

$$\tau(P) = \left(2P / p_0\right) / \left[1 + \sqrt{1 + \left(2p_1 P / p_0^2\right)}\right], \tag{18}$$

provided that $-2p_1 P_{\max} < p_0^2$. The citation-paper rank distribution

$$C(P,t) = C_{\max}(t) - c_0\tau(P),\ C_{\max}(t) = c_0 t \tag{19}$$

is a convex, linear or concave one for p_1 greater than, equal to or less than zero, respectively (Fig. 3a).

- **power law dependence** $p(\tau) = \alpha\tau^{\beta}$,

where $\alpha > 0$, $\beta > -1$, and $\tau > 0$. Now we have

$$P(\tau) = \alpha\tau^{1+\beta} / (1+\beta),\ P_{\max}(t) \equiv P(t) = \alpha t^{1+\beta} / (1+\beta) \tag{20}$$

and

$$\tau(P) = \left[(1+\beta)P / \alpha\right]^{\frac{1}{1+\beta}}. \tag{21}$$

The citation-paper rank distribution

$$C(P,t) = C_{\max}(t)\left[1 - (P / P_{\max})^{\frac{1}{1+\beta}}\right],\ C_{\max}(t) = c_0 t, \tag{22}$$

appears in [Atanassov and Detcheva, 2013] as the *three-parameter positive exponent power-law distribution*. It is convex, linear or concave for $\beta > 0$, $\beta = 0$ or $-1 < \beta < 0$, respectively.

- **exponential dependence** $p(\tau) = \alpha \exp(\beta\tau)$,

where $\alpha > 0$. It is easy to obtain

$$P(\tau) = (\alpha / \beta)\left[\exp(\beta\tau) - 1\right],\ P_{\max}(t) \equiv P(t) = (\alpha / \beta)\left[\exp(\beta t) - 1\right] \tag{23}$$

and

$$\tau(P) = (1/\beta)\ln\left[1 + (\beta / \alpha)P\right]. \tag{24}$$

The citation-paper rank distribution

$$C(P,t) = C_{\max}(t) - (c_0 / \beta)\ln\left[1 + (\beta / \alpha)P\right],\ C_{\max}(t) = c_0 t, \tag{25}$$

is convex, linear or concave for positive, zero or negative values of β, respectively.

Examples II (constant publication rate)

Further on in this section we consider the effect of *time-dependent* (however, *stationary*) paper citation rate $c(\tau,t) = c_s(t-\tau) \geq 0$ on the shape of citation-paper rank distributions by assuming constant publication rate $p \equiv p_0 > 0$. It follows then that $P(\tau) = p_0\tau$, $P_{\max}(t) \equiv P(t) = p_0 t$. As in the previous chapter, all distributions discussed later on are *rank* ones. They are convex, linear or concave for increasing, constant or decreasing in time (t) paper citation rates $c_s(t-\tau)$, respectively.

- **linear dependence** $c_s(t-\tau) = c_0 + c_1(t-\tau)$,

where $c_0 \geq 0$ and c_1 are constants and $c_0 \geq -c_1 t$ must be fulfilled to keep c non-negative. Eq. (14) yields

$$\frac{C(P,t)}{C_{\max}(t)} = 1 - \left(1 + \frac{1}{2}\rho\right)\frac{P}{P_{\max}} + \frac{1}{2}\rho\left(\frac{P}{P_{\max}}\right)^2, \tag{26}$$

where

$$C_{\max}(t) = c_0 t + \frac{1}{2}c_1 t^2,\ \rho = c_1 t^2 / C_{\max}(t).$$

This citation-paper rank distribution appears in [Atanassov and Detcheva, 2013] as the *three-parameter polynomial distribution*. It is convex, linear or concave for $c_1 > 0$, $c_1 = 0$, or $-c_0 / t \le c_1 < 0$, respectively (Fig.3b).

- **power law dependence** $c_s(t-\tau) = \alpha(t-\tau)^{\beta}$,

where $\alpha > 0$, $\beta > -1$, and $t - \tau > 0$. Now we have

$$C(P,t)/C_{\max}(t) = \left[1 - \left(P/P_{\max}\right)\right]^{1+\beta},\ C_{\max}(t) = \alpha t^{1+\beta}/(1+\beta). \tag{27}$$

As it could be expected, this citation-paper rank distribution is convex, linear or concave for $\beta > 0$, $\beta = 0$ or $-1 < \beta < 0$, respectively.

- **exponential dependence** $c_s(t-\tau) = \alpha \exp\left[\beta\left(t-\tau\right)\right]$,

where $\alpha > 0$. We obtain

$$\frac{C(P,t)}{C_{\max}(t)} = \frac{\exp\left[-(\beta/p_0)P\right] - \exp\left[-(\beta/p_0)P_{\max}\right]}{1 - \exp\left[-(\beta/p_0)P_{\max}\right]},\ C_{\max}(t) = \frac{\alpha}{\beta}\left[\exp\left(\beta t\right) - 1\right]. \tag{28}$$

The rank distribution is convex, linear or concave for positive, zero or negative β, respectively.

Examples III (time dependent publication and citation sources)

Let us consider situations, where both publication rate $p(\tau)$ and (stationary) paper citation rate $c_s(t-\tau)$ are *time-dependent*.

- **linear dependence** $p(\tau) = p_0 + p_1\tau$, $c_s(t-\tau) = c_0 + c_1(t-\tau)$,

where $p_0 > 0$, p_1, $c_0 \ge 0$ and c_1 are constants and $p_0 > -p_1 t$, $c_0 \ge -c_1 t$ must hold to keep p positive and c non-negative. The citation-paper rank distribution is given by

$$C(P,t) = c_0\left[t - \tau(P)\right] + \frac{1}{2}c_1\left[t - \tau(P)\right]^2, \tag{29}$$

with $\tau(P)$ specified by Eq. (18). The shape of this distribution depends on the sign of

$$\partial^2 C(P,t)/\partial P^2 = (p_1 c_0 + p_0 c_1 + p_1 c_1 t)/p^3\left[\tau\left(P\right)\right]. \tag{30}$$

Although $\partial^2 C(P,t)/\partial P^2$ varies with P, it obviously cannot change its sign. Therefore, *rank distributions shaped by sources that depend linearly on time, have no inflection points*.

In order to illustrate the general idea (Fig. 1) and the analytics, we have solved numerically Eqs. (3) - (5) for sources linearly varying in time. The results are plotted on Fig. 3, as follows: (a) – for a constant papers' citation rate; (b) – for a constant publication rate; (c) – for varying publication and citation rates. In particular, we have demonstrated (Fig. 3c, dot curve) that *a linear citation-paper rank distribution may exist even for non-constant publication and citation rates*. This has been achieved by appropriately choosing p_1 and c_1 (obviously of opposite signs) to anneal the right part of Eq. (30).

- **power-law dependence** $p(\tau) = \alpha_P \tau^{\beta_P}$, $c_s(t-\tau) = \alpha_C (t-\tau)^{\beta_C}$,

where $\alpha_P > 0$, $\alpha_C > 0$, $\beta_P > -1$, $\beta_C > -1$, and $0 < \tau < t$. The citation-paper rank distribution is

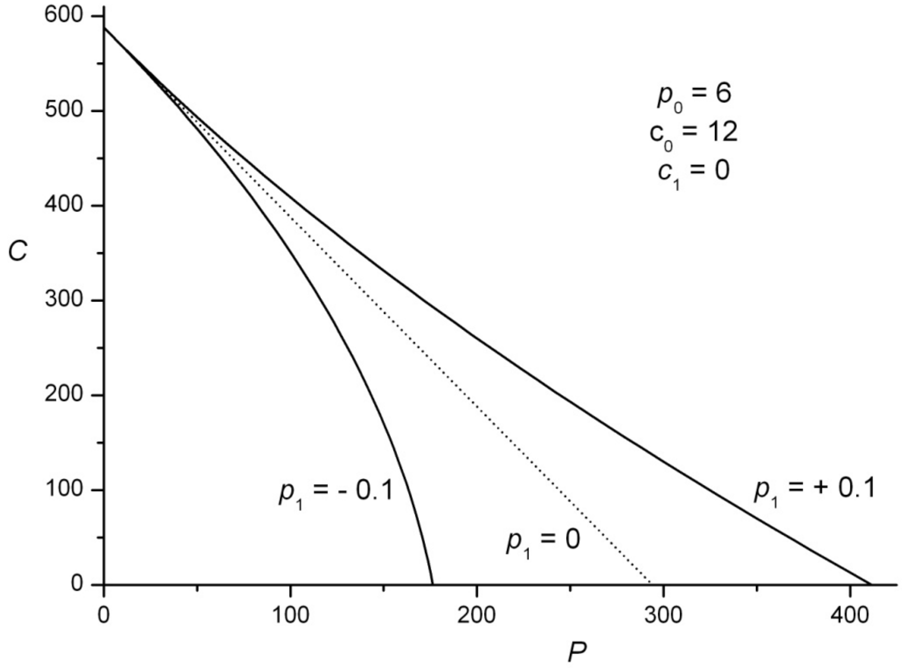

Figure 3a. Citation-paper rank distribution for linearly varying publication rate and constant citation rate

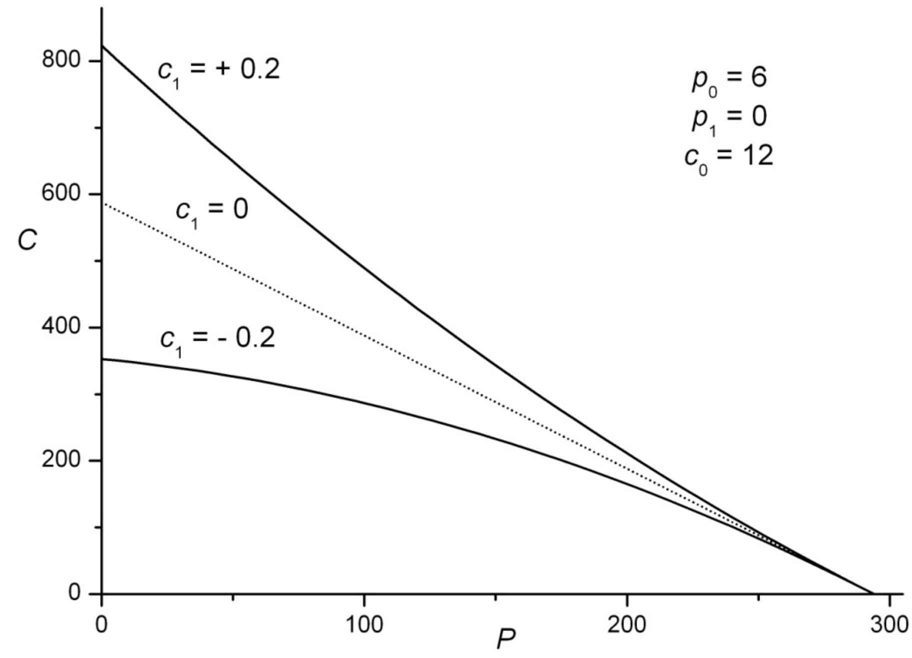

Figure 3b. Citation-paper rank distribution for constant publication rate and linearly varying citation rate

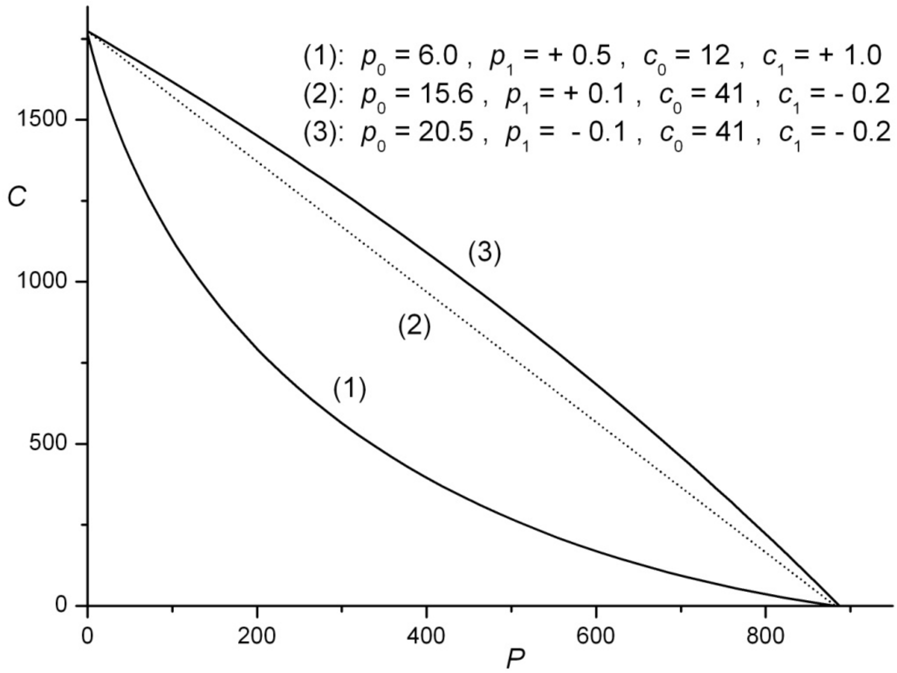

Figure 3c. Citation-paper rank distribution at linearly varying publication rate and citation rate

$$\frac{C(P,t)}{C_{\max}} = \left[1 - \left(P / P_{\max}\right)^{\frac{1}{1+\beta_P}}\right]^{1+\beta_C}, \tag{31}$$

and

$$P_{\max} = \alpha_P t^{1+\beta_P} / (1+\beta_P),\ C_{\max} = \alpha_C t^{1+\beta_C} / (1+\beta_C).$$

This distribution has one inflection point

$$P_{\inf} = P_{\max}\left[1 - \left(\beta_C / \beta_P\right)\right]^{-1/(1+\beta_P)}, \tag{32}$$

provided that $\beta_C / \beta_P < 0$. More precisely, it is convex-concave for $\beta_P < 0$, $\beta_C > 0$ and concave-convex one for $\beta_P > 0$, $\beta_C < 0$. If both β_P and β_C are *negative/positive*, the distribution is *convex/concave,* respectively.

- **exponential dependence** $p(\tau) = \alpha_P \exp(\beta_P \tau)$, $c_s(t-\tau) = \alpha_C \exp\left[\beta_C\left(t-\tau\right)\right]$,

where $\alpha_P > 0$, $\alpha_C > 0$. The citation-paper rank distribution is

$$\frac{C(P,t)}{C_{\max}} = \frac{\left[1 + \left(\beta_P P / \alpha_P\right)\right]^{-(\beta_C/\beta_P)} - \left[1 + \left(\beta_P P_{\max} / \alpha_P\right)\right]^{-(\beta_C/\beta_P)}}{1 - \left[1 + \left(\beta_P P_{\max} / \alpha_P\right)\right]^{-(\beta_C/\beta_P)}} \tag{33}$$

with

$$P_{\max} = \alpha_P\left[\exp\left(\beta_P t\right) - 1\right] / \beta_P,\ C_{\max} = \alpha_C\left[\exp\left(\beta_C t\right) - 1\right] / \beta_C.$$

In the limit $\beta_P \to 0$ (constant publication rate) Eq. (33) reproduces Eq. (28) (with $\alpha_P = p_0$). The distribution is convex, linear or concave depending on whether the exponents sum $\beta_P + \beta_C$ is positive, zero or negative. In particular, for $\beta_P \equiv \beta_C = \beta$ it follows that

$$C / C_{\max} = \left[1 - \left(P / P_{\max}\right)\right] / \left[1 + \left(\beta P / \alpha_P\right)\right]. \tag{34}$$

We note that for small citation count ($C / C_{\max} << 1$), where P approaches $P_{\max}$ the citation-paper rank distribution is dropping faster then expected from Zipf's law ([Bookstein, 1990a]). In the limit $P_{\max} \to \infty$ (this implies $\beta_P > 0$, $\beta_C > 0$), Eq. (33) could be written as

$$C / C_{\max} = 1 / \left[1 + \left(\beta_P P / \alpha_P\right)\right]^{\frac{\beta_C}{\beta_P}}. \tag{35}$$

Eq. (35) represents a Mandelbrot-Zipf distribution with power exponent equal to the ratio of citation and publication growth rates, respectively. This result may be compared with the one derived in [Egghe, 2005] within the Lotkaian informetrics approach. In particular, if both growth rates coincide, the Lotka exponent equals 2, which corresponds to the Zipf exponent of 1, obtained here.

Summary and conclusions

This paper considers some possible reasons for the deviation of citation-paper rank distribution from the linear negative-slope (Hirsch) one. We have derived equations that govern the publication-citation process for time-dependent sources (publication rate and paper citation rate) and obtained conditions for existence of rank convex, concave and linear distributions. The basic equations are analyzed for a stationary temporal dependence of the paper citation rate. A summary of our main conclusions for this particular case follows further on:

- *Stationary* citation rates always lead to *rank* distributions;
- At least one of the sources must *increase/decrease* in time in order to have a *convex/concave* citation-paper rank distribution; publication rates and paper citation rates that simultaneously *increase/decrease* in time produce *convex/concave* distributions;
- For a constant paper citation rate the distribution is *convex, linear or concave*, depending on whether the publication rate is *increasing, constant or decreasing* function of time; for a constant publication rate the distribution is *convex, linear or concave*, for an *increasing, constant or decreasing* paper citation rate;
- The citation-paper rank distribution is a linear one if and only if the ratio of (stationary) paper citation rate and publication rate does not depend on the time of publication (or equivalently, on the paper count); linear negative-slope citation-paper rank distributions may exist even for time-dependent publication and citation rates.

Thus the widely distributed convex shape of citation-paper rank distributions could be associated with increasing in time publication rate and/or paper citation rate. The factors for such *inflation* in productivity and impact might be growing productivity and popularity in the course of a scientific career as well as increase of publication and/or citation rates on global scale.

The equations that govern the publication-citation process are derived on the base of rather common (universal) principles, as *continuity* and c*ausality*. Therefore, we hope it would be possible to use them (perhaps, with appropriate modification) for treatment of more realistic (but difficult to handle) situations, *e.g.* the non-stationary citation rates, rearranging papers or stochastic sources. Moreover, since our analysis contains no presumptions for the active part of the publication-citation process, its results and conclusions could be extended to cover the scientific activity of arbitrary set of authors – a scientist, a research group *etc*.

Acknowledgments

This paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Asknes, 2003] D. W. Asknes. Characteristics of highly cited papers, Research Evaluation 12(3) 159-170 (2003)

[Asknes, 2005] D. W. Asknes. Citations and their use as indicators in science policy. Studies of validity and applicability issues with a particular focus on highly cited papers, Dissertation for the doctoral degree of the University of Twente (2005)

[Atanassov and Detcheva, 2013] V. Atanassov, E. Detcheva. Citation-paper rank distributions and associated scientometric indicators – a survey, Int. J. Information Models and Analyses 2(1) 46-61 (2013)

[Atanassov, 2012] V. Atanassov. Time evolution of scitation-paper rank distributions and its implications for scientometric models, presented at “Evaluating Science: Modern Scientometric Methods” Conference Sofia May 21-22, 2012, COST Action “Physics of Competition and Conflicts, https://sites.google.com/site/scientometrics2012sofia/presentations (2012)

[Bar-Ilan, 2008] J. Bar-Ilan. Informetrics at the beginning of the 21st century – A review, Journal of Informetrics 2 1-52 (2008)

[Bookstein, 1990a] A. Bookstein. Informetric distributions, Part I: Unified Overview, J. Am. Soc. Information Science 41(5) 368-375(1990)

[Bookstein, 1990b] A. Bookstein. Informetric distributions, Part II: Resilience to Ambiguity, J. Am. Soc. Information Science 41(5) 376-386(1990)

[Burrell 2007a] Q. L. Burrell. Hirsch’s h-index: A stochastic model, Journal of Informetrics 1 16-25 (2007)

[Burrell 2007b] Q. L. Burrell. On the h-index, the size of the Hirsch core and Jin’s A-index, Journal of Informetrics 1 170-177 (2007)

[de Solla Price, 1976] D. de Solla Price. A General Theory of Bibliometric and Other Cumulative Advantage Processes, J. Am. Soc. Information Science 27(5-6) 292-306 (1976)

[Egghe, 2005] L. Egghe. Power Laws in the Information Production Process: Lotkaian Informetrics, Elsevier (2005)

[Eom and Fortunato, 2011] Y.-H. Eom, S. Fortunato. Characterizing and Modeling Citation Dynamics, PLoS ONE 6(9), e24926 (2011)

[Golosovsky and Solomon, 2012] M. Golosovsky and S. Solomon. Stochastic Dynamical Model of a Growing Citation Network Based on a Self-Exciting Point Process, Phys. Rev. Lett. 109 098701 (2012)

[Gupta et al, 2005] H. M. Gupta, J. R. Campanha and R. A. G. Pesce. Power-Law Distributions for the Citation Index of Scientific Publications and Scientists, Braz. J. Phys. 35 (4A) 981-986 (2005)

[Hirsch, 2005] J.E. Hirsch. An index to quantify an individual’s scientific research output, Proc. Nat. Acad. Sci. 102 (46) 16569-16572 (2005)

[Liu and Rousseau, 2008] Y. Liu, R. Rousseau. Definitions of time series in citation analysis with special attention to the h-index, Journal of Informetrics 2 202-210 (2008)

[Markov et al, 2013] K. Markov, K. Ivanova, V. Velichko. Usefulness of scientific contributions, Int. J. Information Theories and Applications 20(1) 4-38 (2013)

[Petersen et al, 2011] A. M. Petersen, H. E. Stanley and S. Succi. Statistical regularities in the rank-citation profile of scientists, Sci. Rep. 1 181 (2011)

[Peterson et al, 2010] G. J. Peterson, S. Presse and K. A. Dill. Non-universal power-law scaling in the probability distribution of scientific citations, Proc. Nat. Acad. Sci. 107 (37) 16023-16027 (2010)

[Redner, 2005] S. Redner. Citation Statistics from 110 Years of Physical Review, Physics Today 49-54 (2005)

[Sen, 1996] S. K. Sen, Theoretical Issues in Citation Process: A Review, Int. J. Scientometrics and Informetrics 2(2-3) 159-198 (1996)

[van Raan, 2001a] A. F. J. Van Raan. Two-step competition process leads to quasi power law income distributions. Application to scientific publication and citation distributions, Physica A 298(3-4) 530-536 (2001)

[van Raan, 2001b] A. F. J. Van Raan. Competition amongst scientists for publication status: toward a model of scientific publication and citation distributions, Scientometrics 51 347-357 (2001)

[van Raan, 2004] A. F. J. Van Raan. Sleeping Beauties in science, Scientometrics 59(3) 461-466 (2004)

[Wallace et al, 2008] M. L. Wallace, V. Lariviere, Y. Gingras. Modeling a Century of Citation Distributions, Journal of Informetrics 3 296-303 (2009)

[Wang et al, 2013] D. Wang, C. Song, A.-L. Barabasi. Quantifying Long-Term Scientific Impact, Science 342 127-132 (2013)

[Watts and Gilbert., 2011] C. Watts, N. Gilbert. Does cumulative advantage effect affect collective learning in science? An agent-based simulation, Scientometrics 89(1) 437-463 (2011)

Authors' Information

Vladimir Atanassov *– Institute of Electronics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria; e-mail: v.atanassov@abv.bg*

Major Fields of Scientific Research: Plasma Physics and Gas Discharges, Radars and Ocean Waves, Nonlinearity and Chaos, Scientometric

Ekaterina Detcheva *– Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,1113 Sofia, Bulgaria; e-mail: detcheva@math.bas.bg*

Major Fields of Scientific Research: Web-based applications, Image processing, analysis and classification, Knowledge representation, Business applications, Applications in Medicine and Biology, Applications in Psychology and Special Education, Computer Algebra.