Scientometrics (2014) 98:725–742
DOI 10.1007/s11192-013-1018-4

The individual author's publication–citation process: theory and practice

Quentin L. Burrell

Received: 15 February 2013 / Published online: 30 April 2013

Abstract The model proposed by Burrell (Information Processing and Management 28:637–645, 1992, Journal of Informetrics 1:16–25, 2007a) to describe the way that an individual author's publication/citation career develops in time is investigated further, the aim being to describe in more detail the form of the citation distribution and the way it evolves over time. Both relative and actual frequency distributions are considered. Theoretical aspects are developed analytically and graphically and then illustrated using small empirical data sets relating to some well-known informetrics scholars. Perhaps surprisingly, it is found that the distribution may well be approximated in some cases by a simple geometric distribution.

Keywords Informetric process · Citation distribution · Mixed Poisson process · Lotkaian informetrics · Pareto distribution · Geometric distribution · Price medallists

Introduction

In an attempt to understand the development of an author's h-index over time and to investigate factors affecting it, Burrell (2007a) proposed a stochastic model for what was termed the (stochastic) publication/citation process. However, the original paper gave little indication of the general form of the publication/citation distribution at any point in time. On the other hand several authors, most notably Egghe (2006, 2010) and Egghe and Rousseau (2006, 2012a, b), have explicitly assumed that the distribution is of Pareto/Lotka form although, like Burrell (2007a), without providing any empirical evidence in support. (Recently Burrell (2013) has presented some empirical examples that cast doubt on the Lotka assumption in this context.) In this note the aim is to present further results concerning the stochastic model. The paper is in three general parts. The first (technical) part

Q. L. Burrell
Centre for R&D Monitoring (ECOOM), KU Leuven, Waaistraat 6, Leuven, Belgium

Q. L. Burrell (✉)
119 Friary Park, Ballabeg, Isle of Man IM9 4EX, UK
e-mail: quentinburrell@manx.net

concerns the purely theoretical/mathematical properties of the stochastic publication/citation model. (Even here, some of the mathematical detail is deferred to an Appendix.) The general reader might well prefer to skip this section and move directly to the second part, which seeks to illustrate the general properties of the model and the development over time of an author's citation distribution via graphical representations. The third presents some empirical data sets and seeks to demonstrate that, at least in general terms, they conform to the theoretical model.

The publication/citation process distribution

The rudiments of the stochastic model were described in Burrell (1992) and more fully developed in Burrell (2007a). The essential features are that (a) an author publishes papers in some random fashion over time; (b) these papers subsequently attract citations, again in some random fashion over time; and (c) the citation rate varies over the published papers.

Notation Write Y_t = number of papers published by time t and, for a randomly chosen paper, X_t = total number of citations acquired by time t. Then $\{Y_t; t \geq 0\}$ and $\{X_t; t \geq 0\}$ are continuous time stochastic counting processes.
For the model, we assume that

(a) From the start of his/her publishing career (at time zero), an author publishes papers according to a Poisson process of rate θ, giving the mean number of publications per unit time and called the *publication rate*. Thus

$$P(Y_t = r) = e^{-\theta t}\frac{(\theta t)^r}{r!} \quad \text{for } r = 0, 1, 2, \ldots \text{ and note that } E[Y_t] = \theta t \text{ for } t \geq 0 \qquad (1)$$

(b) Any particular publication acquires citations according to a Poisson process of rate Λ, where Λ varies from paper to paper. Here Λ denotes the mean number of citations to the paper per unit time following publication, called the *citation rate*.

(c) The citation rate Λ varies over the set of publications according to a gamma distribution of index $\nu > 1$ and scale parameter $\alpha > 0$. Thus the probability density function (pdf) of Λ is given by

$$f_\Lambda(x) = \frac{\alpha^\nu}{\Gamma(\nu)}x^{\nu-1}e^{-\alpha x}, \quad x > 0 \qquad (2)$$

Remarks

(i) The Poisson assumption is the obvious one for a general counting process while the variable citation rate accords with intuition and experience. The assumption of a gamma distribution is more contentious. Its main virtues are that it is intuitively reasonable on account of its shape and, just as important, it is analytically convenient in that it allows the explicit calculations required to find the closed form expression in the Theorem.

(ii) In general, the gamma pdf is defined for any $\nu > 0$. The restriction to values greater than one is to ensure the convergence of various integrals leading to the formula given in the following Theorem. (Note that there is a slight error in Burrell (2007a), repeated in Burrell (2007b, c, d, e), where the restriction is stated as $\nu \geq 1$. In fact the case $\nu = 1$ must be excluded also.)

The crucial result is the following:

Theorem (Burrell 2007a) *Under the assumptions of the model, the probability mass function (pmf) of X_t, the number of citations to a randomly chosen paper by time t, is given by*

$$P(X_t = r) = \frac{\alpha}{(\nu-1)t}\frac{\Gamma(r+\nu)}{r!\Gamma(\nu-1)}\int_0^{t/(\alpha+t)} y^r(1-y)^{\nu-2}dy \tag{3}$$

$$= \frac{\alpha}{(\nu-1)t}B\left(\frac{t}{\alpha+t};r+1,\nu-1\right) \quad \text{for } r=0,1,2,\ldots, \tag{4}$$

where $B(x;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x y^{a-1}(1-y)^{b-1}dy$ is the cumulative distribution function (cdf) of a beta distribution (of the first kind) with parameters a and b.

Remarks

(i) Although the full model specification involves four parameters θ, t, ν and α, each of which can be chosen freely, (3) does not involve the publication rate θ since this only affects the *actual frequencies* rather than the probabilities or *relative frequencies*. The role of θ will become apparent in the next section. Also note that evaluation of the RHS of (3) only requires specification of ν and t/α, i.e. it is in effect a two-parameter distribution. The first of these is the index of the gamma distribution of citation rate and is essentially a shape parameter. Upon reflection, the dependence on the ratio t/α, rather than the two parameters separately, is not surprising since $1/\alpha$ is a scale parameter for the citation *rate* distribution so that t/α can be thought of as a sort of *standardised* time parameter. Hence it is reasonable to simplify notation and in the following write $s = t/\alpha$ for "standardised time". It is also useful to introduce $q = q(t,\alpha) = \frac{t}{\alpha+t} = \frac{s}{1+s}$. Thus (4) can be rewritten as

$$P(X_t = r) = \frac{1}{(\nu-1)s}B\left(\frac{s}{1+s};r+1,\nu-1\right) = \frac{1}{(\nu-1)s}B(q;r+1,\nu-1) \tag{5}$$

Note that, in general, for calculation of the probabilities in (3), numerical methods are required. This is not a great drawback in practice since many spreadsheet packages, such as MS Excel®, include the cdf of the beta distribution as standard. Such numerical methods are relied upon to a large extent on in what follows.

(ii) In the mathematical as opposed to the statistical literature, the above beta cdf is often referred to as the incomplete beta function ratio or the regularized (incomplete) beta function with a different standard notation as $I_x(a,b) = B(x;a,b)$. Thus the result in the Theorem may alternatively be written as

$$P(X_t = r) = \frac{1}{(\nu-1)s}I_q(r+1,\nu-1)$$

(iii) At the heart of the model is the assumption that sources (publications) produce/receive items (citations) according to independent Poisson processes of variable rates, leading to a consideration of mixed Poisson processes. Mixtures of Poisson distributions and processes have been used widely within informetrics, dating back at least to Morse (1976). The most all-embracing version of a mixed Poisson process is the generalized inverse Gaussian Poisson (GIGP) of Sichel (1985), see also Burrell

and Fenton (1993) for informetric applications. A more tractable—though slightly less flexible—version that has been used with considerable success particularly in the context of circulations of library materials is the gamma mixture, leading to the gamma-Poisson process (GPP), see Burrell (1980, 1990) and Burrell and Cane (1982) among others. The GPP has also been used in the context of citation studies but in situations where there is a clearly specified set of publications at the outset, see for instance Burrell (2001, 2002a, b, 2003, 2005a, b), whereas here new publications can be added at any time.

(iv) A Referee has pointed out that a legitimate alternative to the GPP in many situations is the so-called generalised Waring process (GWP), see for instance Zografi and Xekalaki (2001), Burrell (2005) and Xekalaki and Zografi (2008). However, the same objection as above, regarding a fixed set of publications rather than a growing set, applies to its appropriateness for the current context.

Rather than the integral formulation in (3), elementary closed expressions for (4) can be given when v takes an integer value as in the following.

Proposition 1 *If $v \geq 2$ is an integer, then*

$$\begin{aligned} P(X_t = r) &= \frac{1}{(v-1)s} B(q; r+1, v-1) = \frac{1}{(v-1)s} I_q(r+1, v-1) \\ &= \frac{1}{(v-1)s} \sum_{j=r+1}^{r+v-1} \frac{(r+v-1)!}{j!(r+v-1-j)!} q^j (1-q)^{r+v-1-j} \end{aligned} \tag{6}$$

with q and s as before.

Proof This is a standard result for the incomplete beta function, easily demonstrated by repeated integration by parts.

Other closed forms when $v \geq 2$ is an integer can easily be found by first expanding the second term in the integrand of (3) using the binomial theorem and then integrating term by term.

A particularly striking simple case, and one that will be of especial interest in what follows is provided by taking $v = 2$. □

Corollary *If $v = 2$, then*

$$P(X_t = r) = \left(\frac{1}{1+s}\right)\left(\frac{s}{1+s}\right)^r = (1-q)q^r, \quad r = 0, 1, 2, \ldots$$

Thus X_t follows a geometric distribution with parameter $p = 1 - q = \frac{1}{1+s}$.

Proof This follows simply by putting $v = 2$ in (6) but alternatively substituting into (3) gives

$$\begin{aligned} P(X_t = r) &= \frac{1}{s} B(q; r+1, 1) = \frac{1}{s} \frac{\Gamma(r+2)}{\Gamma(r+1)\Gamma(1)} \int_0^q y^r dy \\ &= \frac{r+1}{s} \left[\frac{y^{r+1}}{r+1}\right]_0^q = \frac{1}{s}\left(\frac{s}{1+s}\right)^{r+1} = \left(\frac{1}{1+s}\right)\left(\frac{s}{1+s}\right)^r = pq^r \quad \text{for } r = 0, 1, 2, \ldots \end{aligned}$$

□

Remark The simplest version of the standard GPP model involves an exponential distribution, i.e. a gamma distribution of index one, as the mixing distribution and leads to a geometric distribution, as in Burrell (1980). It is important to stress that this model is not equivalent to the one leading to the Corollary above where it is a gamma distribution of index *two* that leads to the geometric form. As mentioned before, the crucial difference between the two models is that in the simple GPP model it is assumed that all sources are present from time zero whereas here new sources are appearing at random from time zero onwards to the present.

For the moments of the distribution, note that it was shown in Burrell (2007a) that the mean of the distribution in (3) is given by $E[X_t] = \frac{\nu t}{2\alpha} = \frac{\nu s}{2}$. This result is intuitively clear since, at time t, a typical paper has been in circulation for, on average, time $t/2$ during which time it has been acquiring citations at an average rate of ν/α per unit time.

For higher moments it is convenient to work with the probability generating function as follows.

Definition/Notation The probability generating function (pgf) of the non-negative integer-valued random variable X_t is given by

$$G_t(z) = E\left[z^{X_t}\right] = \sum_{r=0}^{\infty} z^r P(X_t = r)$$

Proposition 2 *With X_t as in the Theorem, the pgf is given by*

$$\begin{aligned} G_t &= \frac{1}{(\nu-1)s(z-1)}\left[\left(\frac{p}{1-qz}\right)^{\nu-1} - 1\right] \\ &= 1 + \frac{\nu}{2!}(s(z-1)) + \frac{\nu(\nu+1)}{3!}(s(z-1))^2 + \cdots \\ &= \sum_{r=0}^{\infty} \frac{\Gamma(\nu+r)}{\Gamma(r)(r+1)!}(s(z-1))^r \end{aligned} \tag{7}$$

with s as before.

Proof See the Appendix. □

From Proposition 2 it is easy to determine the factorial moments as in the following.

Proposition 3

$$E[X_t(X_t-1)\ldots(X_t-r+1)] = \frac{\nu(\nu+1)\ldots(\nu+r-1)}{r+1}s^r \quad \text{for r} = 1, 2, \ldots \tag{8}$$

Proof See the Appendix. □

Corollary

$$E[X_t] = \frac{\nu s}{2}$$

$$V(X_t) = \frac{\nu s}{12}(6 + s(\nu+4))$$

Proof Making use of (8), the mean is simply the case $r = 1$ while, for the variance, putting $r = 2$ gives

$$E[X_t(X_t - 1)] = E\left[X_t^2\right] - E[X_t]\frac{v(v+1))}{3}s^2$$

and then

$$\begin{aligned} V(X_t) &= E\left[X_t^2\right] - (E[X_t])^2 = E[X_t(X_t - 1)] + E[X_t] - (E[X_t])^2 \\ &= \frac{v(v+1)s^2}{3} + \frac{vs}{2} - \frac{v^2s^2}{4} = \frac{vs}{12}(6 + s(v+4)) \end{aligned}$$

□

Remark That the mean is directly proportional to time should be self-evident. It is interesting to note that the variance is a quadratic function of time, necessarily passing through the origin.

Note The index of dispersion ID_t of a stochastic counting process is given by the ratio of the variance to the mean so that here we have $ID_t = \frac{V(X_t)}{E[X_t]} = 1 + s\left(\frac{v+4}{6}\right)$ and hence the publication/citation process is over-dispersed compared to a simple Poisson process, for which the index is always equal to one. Indeed, here it increases linearly over time, although not directly proportional to time. As an aside, it is worth mentioning that for a GPP the index of dispersion is also a linear function of time but is independent of the gamma parameter v.

Of course, the Theorem, together with Proposition 1 and its Corollary merely give mathematical expressions for the probability distribution of citations at a given moment (the time t since the commencement of the author's publication career); it gives no obvious clue to what the distribution looks like at any particular time, nor how it changes with increasing time. These aspects are best demonstrated by graphical methods. (As the aim is to illustrate the general form of the distribution we just give some examples—the reader can easily generate other cases starting from the basic formula in (3).)

Although the pmf in (3) involves the citation rate parameters $\alpha > 0$ and $v > 1$, together with current time, t, we have shown that it is actually a function of just the gamma index v and the standardised time $s = t/\alpha$. In what follows we will fix the gamma index, since this governs the shape of the distribution, and the natural time since this shows how the distribution develops over time. (The parameter α is not so important for purposes of illustration since it is just a scaling of the citation rate and, as has been shown, can essentially be absorbed into the time parameter.) To aid comparisons we will take $\alpha = v$ throughout this section, so that the assumed citation rate has mean equal to one per unit time.

It is not difficult to show that the pmf in (3) is a strictly decreasing function of r. However, the rate of decrease can be quite slow as can be seen from Fig. 1a–c which gives examples of the pmf for times $t = 10$, 20 and 30 with $v = 1.01$, 2 and 3. (As the main interest is the evolving shape of the distributions, these are plotted as continuous functions for visual clarity.)

(Note that the vertical scales are different in the three figures.)

It is interesting to observe that in the examples with $v = 1.01$ and $v = 2$, the graphs are reverse J-shaped, whereas for $v = 3$ they are an elongated reverse S-shape. (We conjecture that these properties are true in general for distinguishing between cases where $v \leq 2$ and

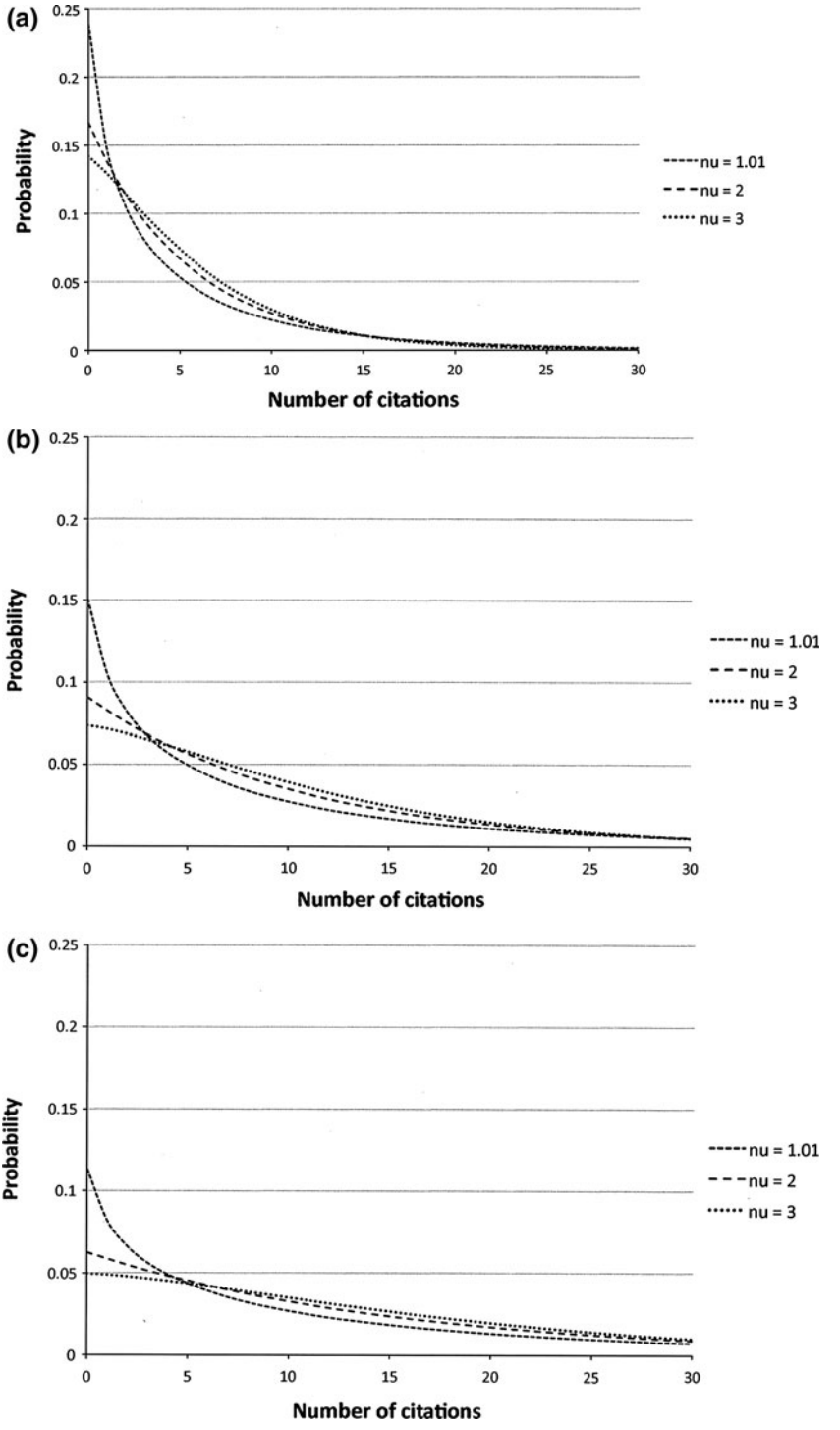

Fig. 1 **a** Probability mass function, $t = 10$. **b** Probability mass function, $t = 20$. **b** Probability mass function, $t = 30$

$v > 2$ but have not been able to prove it.) In all cases, though, the probabilities are small and decline very slowly with a long tail. Furthermore, the graphs become ever flatter as t increases. This is to be expected since from (3) it is clear that for each r, $P(X_t = r)$ decreases to zero as $t \to \infty$. In particular $P(X_t = 0) \to 0$ as $t \to \infty$ so that the expected proportion of uncited items tends to zero with increasing time. Note that this is not the same as saying that the actual *number* of uncited items necessarily declines. For this we need to consider the behaviour of the *expected* frequency distribution, which will be our next concern.

The expected frequency distribution

The important difference when considering the expected frequency distribution is that it takes account of the increasing number of contributing publications as time proceeds.

Notation Write $Z_t(r)$ = number of papers with exactly r citations at time t.

Given the number of papers that have been published by time t, say $Y_t = n$, then the conditional distribution of $Z_t(r)$ is binomial corresponding to n (the number of papers) "trials" with probability $P(X_t = r)$ of "success" at each trial. It follows that

$$E[Z_t(r)] = E[Y_t P(X_r = r)] = E[Y_t]P(X_t = r) = \theta t P(X_t = r)$$

since Y_t has a Poisson distribution with mean θt, see (1).

Hence the distribution of expected citation frequencies is just the underlying probability distribution scaled by the expected number of publications, which is in turn proportional to t. The publication rate θ is just a (vertical axis) scaling so, to illustrate the changing form, there is no harm in taking $\theta = 1$. The expected frequency distributions are illustrated in Fig. 2a–c for the same parameter values as before but now with the purpose of showing the change over time for each of these separately.

Remarks on the general shapes of the distributions are of course the same as those made after Fig. 1a–c but we can say more. Firstly, it is clear that the entire distribution increases as t increases, i.e. $E[Z_t(r)]$ increases with t for any value of r. In particular, the expected number of uncited items increases, which might seem counter-intuitive. It implies that new uncited sources are appearing faster than existing uncited sources are becoming cited for the first time. This aspect was considered in detail by Burrell (2012), in response to Egghe et al. (2011), and it was shown that, according to this model, the expected number of uncited items tends to the value $\theta\alpha/(v - 1)$. For the parameter values in Figs. 2a–c the limits are 101, 2 and 1.5 respectively. The manner of convergence is clear.

When considering real data sets, as we will do in the next section, they can actually be quite small. How many papers will a scientist publish in the first 10 years of his/her career? Obviously this will depend on the subject area as well as the particular scientist's talents but we would hazard that it is, in general, no more than about 50. (In our examples, it is about 25.) With such small data sets, it is difficult to determine the shape of the distribution just by consideration of the frequency distribution and it is often more informative to consider instead the tail distribution function.

Definition For any non-negative integer-valued random variable W, its tail distribution function is given by

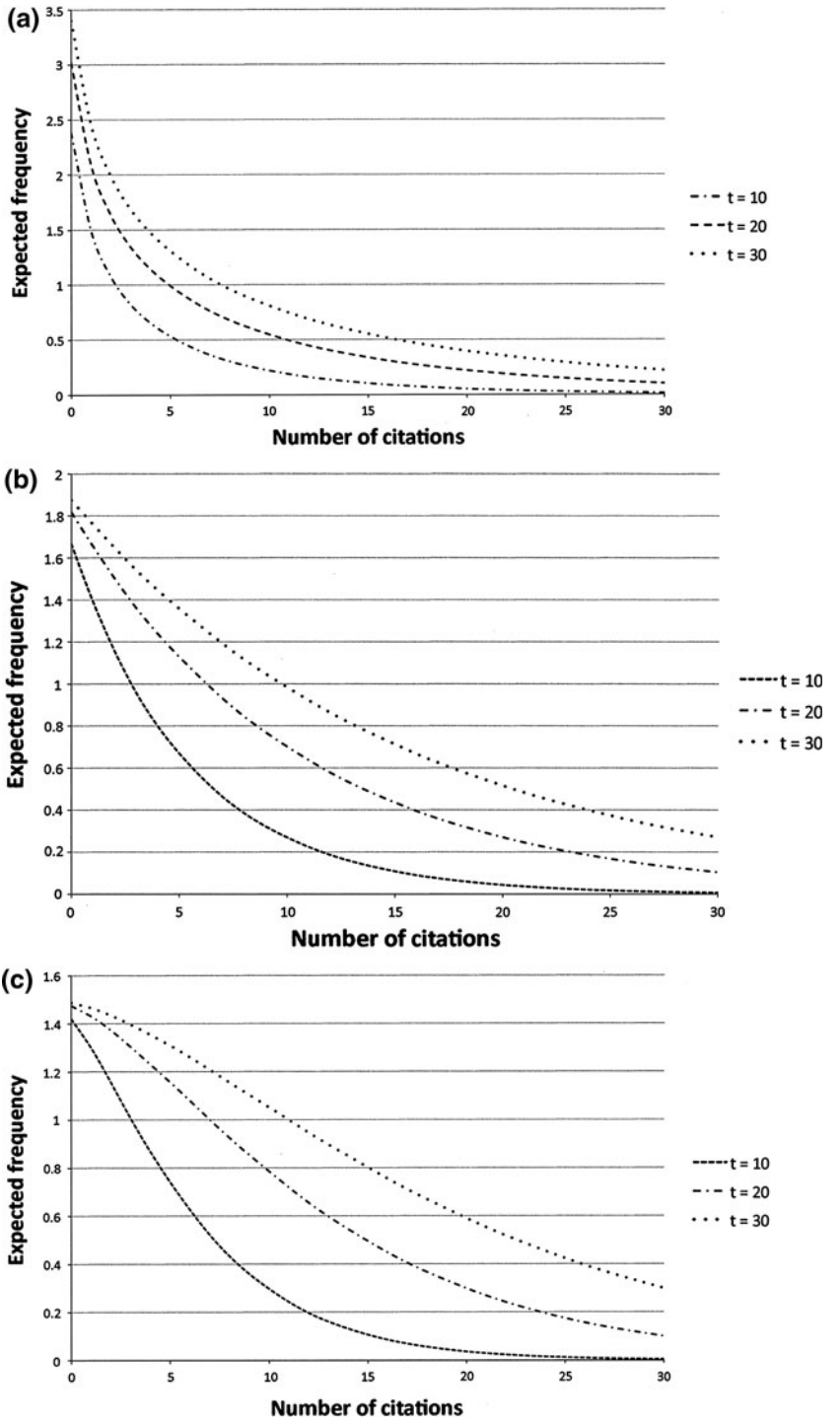

Fig. 2 **a** Expected frequency distribution, $\nu = 1.01$. **b** Expected frequency distribution, $\nu = 2$. **c** Expected frequency distribution, $\nu = 3$

$$\phi(k) = P(W \geq k) = \sum_{j=k}^{\infty} P(W = j)$$

In particular $\phi(k) = \phi_t(k) = P(X_t \geq k)$ is just the probability that by time t a paper has gathered at least k citations

Except in the case where $v = 2$, there do not seem to be simple closed form expressions for the tail distribution of the publication/citation distribution in (3) and so we again rely upon numerical evaluations and graphical presentation. In the special case where $v = 2$, corresponding to a simple geometric distribution, the tail distribution function is given by

$$\phi(k) = P(X_t \geq k) = \sum_{j=k}^{\infty} P(X_t = j) = \sum_{j=k}^{\infty} pq^j = q^k = \left(\frac{s}{1+s}\right)^k = \left(\frac{t}{\alpha+t}\right)^k \quad (9)$$

for k $= 0, 1, 2, \ldots$

An important feature of this case is revealed when we consider instead the log tail distribution, which is a linear function of k. Indeed, from (9) we have that

$$\log \phi(k) = k \log q = k \log\left(\frac{t}{\alpha+t}\right) \quad (10)$$

(Here and throughout logs are taken as natural logarithms.)

Remark It is not difficult to show that the converse to this result is true, namely that, if the plot of the log tail distribution function is linear, then the pmf is geometric.

Returning to the expected frequency distribution let us write

$$N_t(k) = \sum_{j=k}^{\infty} Z_t(j) = \text{Number of papers with at least k citations by time t}$$

(This corresponds to the notation $N(k; t)$ in Burrell (2007a, b, c, d, e).)

Then we have for its tail form:

$$\Phi(k) = \Phi_t(k) = E[N_t(k)] = \sum_{j=k}^{\infty} E[Z_t(j)] = \sum_{j=k}^{\infty} \theta t P(X_t = j) = \theta t \sum_{j=k}^{\infty} P(X_t = j) = \theta t \phi(k)$$

The log tail distribution is given as

$$\log \Phi(k) = \log \phi(k) + \log t + \log \theta$$

and in particular when $v = 2$ and $\theta = 1$ we have, from (10)

$$\log \Phi(k) = k \log\left(\frac{t}{\alpha+t}\right) + \log t$$

This linear form in k, for fixed t, allows a simple benchmark against which other forms can be compared. In Fig. 3 we give illustrations of the plot of the log tail distribution for the same set of parameter values as before.

Two aspects are worthy of comment. Firstly there is the expected general upward shift of the distributions as t increases as well as the horizontal stretching of the plots.

Also, these examples show that the graphs are convex for $v < 2$, linear for $v = 2$ and concave for $v > 2$ and again we conjecture that this is true in general. Taken together, these

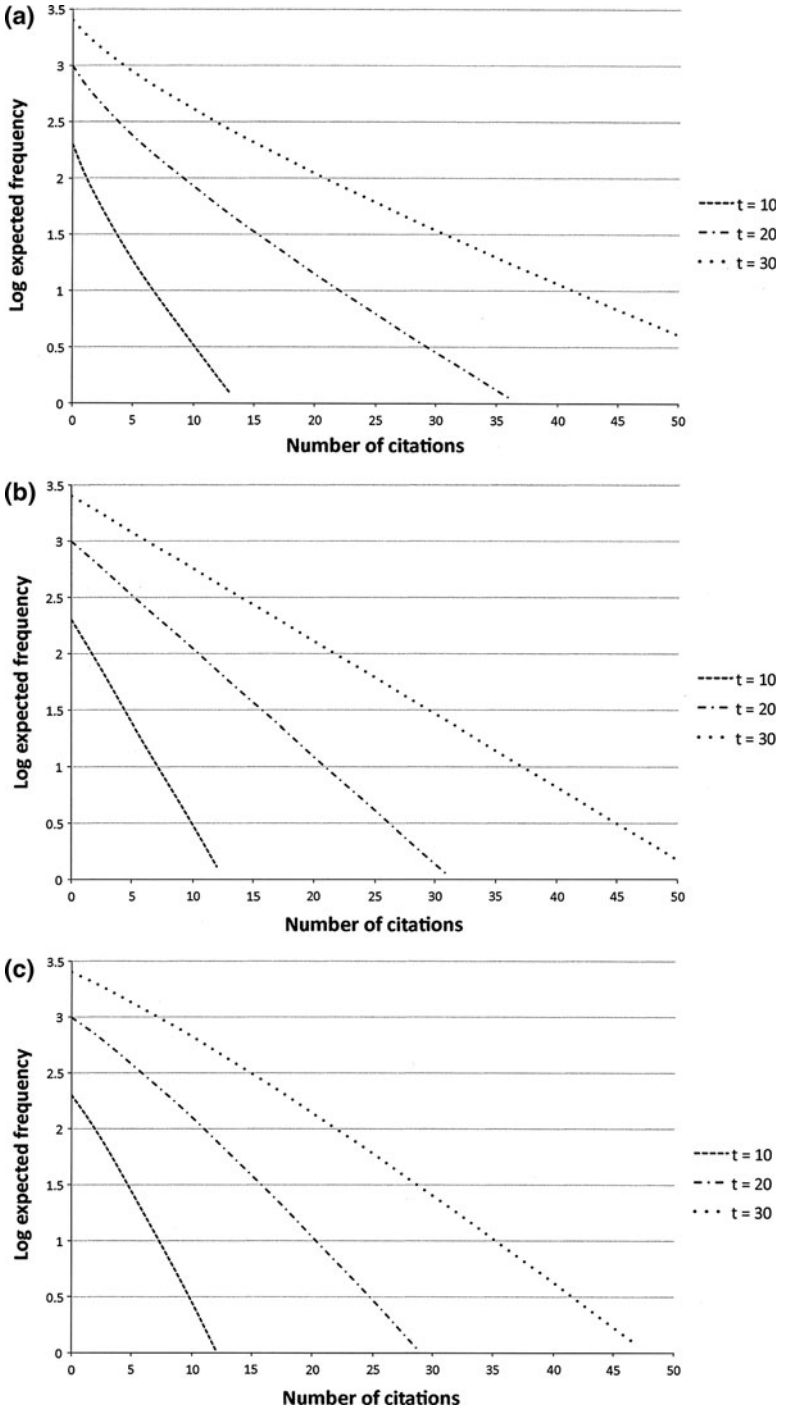

Fig. 3 **a** Log expected tail frequency distribution, $\nu = 1.01$. **b** Log expected tail frequency distribution, $\nu = 2$. **c** Log expected tail frequency distribution, $\nu = 3$

aspects suggest that this sort of plot that will most easily distinguish between the various forms of the distribution.

Some empirical examples

In Burrell (2009) the author has looked at certain features of the empirical development of an author's publication/citation career in the light of the stochastic model, in particular regarding the increases of publications and citations over time. Here we wish to look at the development over time of the entire frequency distribution of citations. Purely for purposes of illustration, we have analyzed the citation records of four of the most eminent currently active scholars in (mainly) mathematical approaches to problems in informetrics. All are of the same generation and are recipients of the Price Medal. In alphabetical order, they are Leo Egghe, first publication in 1978, Price Medal awarded in 2001; Wolfgang Glanzel, first publication 1983, Price Medal 1999; Loet Leydesdorff, first publication 1981, Price Medal 2003; Ronald Rousseau, first publication 1976, Price Medal 2001. (Year of first publication is according to Web of Science (WoS®).)

The data to be analysed for each author relate to the cumulated publications and citations acquired up to the end of 1990, end of 2000 and end of 2010. (These were all downloaded from WoS® in December 2012.) Only material that could be classified as research was included, so that editorial material, book reviews, etc. were excluded. The data have been cleaned to the best of our ability but we do not claim them to be absolutely definitive—only to be adequate for purpose! Note that, because of different years of first

Table 1 Citation frequency distribution giving the number publications with r citations for $r = 0, 1, 2, \ldots, 20$

No. of citations r	Egghe $N = 171$	Glanzel $N = 139$	Leydesdorff $N = 153$	Rousseau $N = 166$
0	32	24	24	32
1	15	8	9	15
2	18	1	8	12
3	17	2	8	13
4	8	6	9	14
5	9	7	7	5
6	6	6	6	5
7	8	5	9	6
8	7	5	5	8
9	4	4	6	5
10	5	1	7	3
11	2	4	4	0
12	3	1	2	8
13	2	4	4	1
14	6	3	5	1
15	2	3	1	3
16	1	0	4	2
17	3	5	1	3
18	2	6	6	3
19	3	1	1	3
20	3	1	0	1

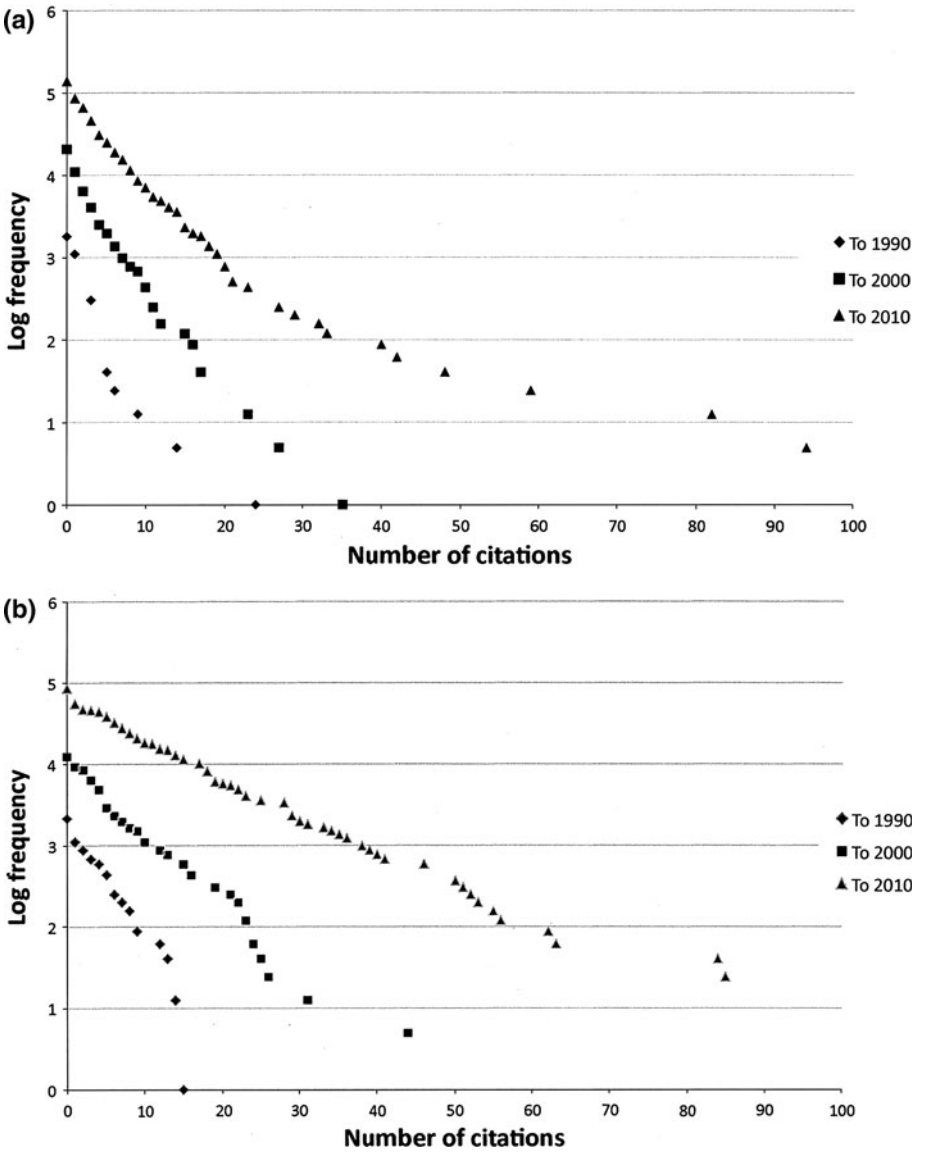

Fig. 4 **a** Log tail frequency distribution: Egghe. **b** Log tail frequency distribution: Glanzel. **c** Log tail frequency distribution: Leydesdorff. **d** Log tail frequency distribution: Rousseau

publication for the four authors, the lengths of time period covered are slightly different but this does not really affect the main objective of the exercise, which is illustrative.

The initial parts of the basic citation frequency distributions up to the end of 2010 for the four authors, together with their total numbers of publications, *N*, are given in Table 1. Presented in this way, perhaps the best we can say is that in each case there is a general decline in the frequencies as the number of citations increases but that it is, to say the least, a very irregular decline. As is often the case for such relatively small data sets, a graphical presentation of the raw frequency distribution adds little to this crude description. A better

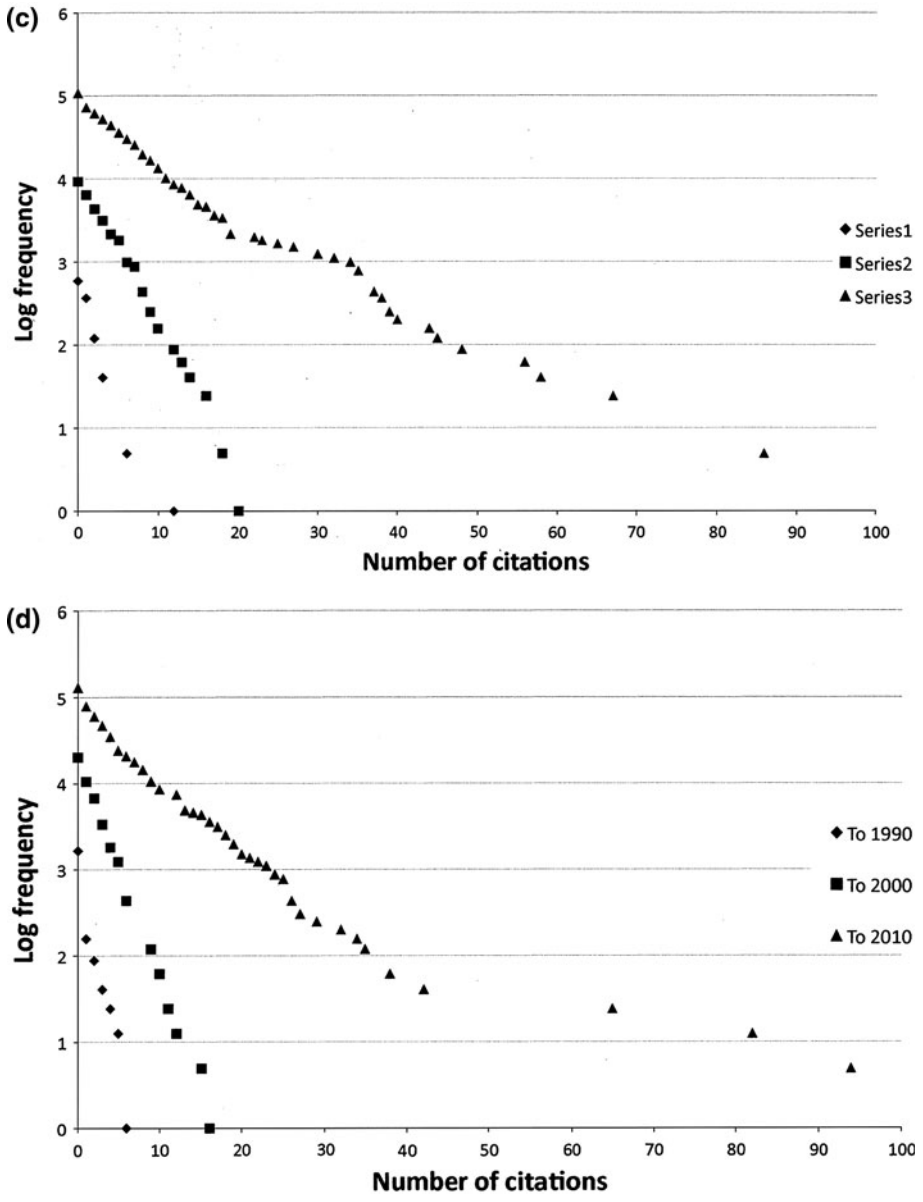

Fig. 4 continued

visual appreciation is gained when, as in the theoretical framework, we look at the tail frequency distribution as this has the effect of smoothing some of the irregularities in the individual frequencies. Again in view of our comments on the preceding graphical representations of the theoretical distributions, in Fig. 4 are plotted the log tail distributions of the empirical citation frequencies of the four authors.

Note that in order to facilitate visual comparison of the plots for the four authors, we have truncated the number of citations at 100. In the case of Egghe, this required omitting his most highly cited paper (Egghe 2006), which by the end of 2010 had attracted 162 citations; for Glanzel, this cut the top three papers, the most highly cited being (Schubert et al. 1989) with 157 citations; for Leydesdorff, the single paper (Etkowitz and Leydesdorff 2000) with 410 citations and for Rousseau just the paper (Ahlgren et al. 2003) with 103 citations.

It is striking that in all cases there is an evident degree of regularity in the early parts of the graph but with increasingly wild behaviour in the later parts, which are of course governed by the most highly cited papers. Indeed, in the earliest parts there seems to be evidence of linearity in all of the graphs, although it can be argued that in some cases there is a suggestion of convexity. Perhaps the best that can be said is that, at least in general terms, they conform to the typical theoretical examples in Fig. 3.

Concluding remarks

Mathematical modelling in the social sciences is much more problematic than in the physical sciences because there are so many external unforeseen, uncontrollable and unmeasurable sources of variability, even at the general level. When looking at behaviour at the individual level, matters are even worse. In this paper we have investigated a model that seeks to address the two main aspects of the publication–citation process for an individual scientist. The first, regarding the publication process, assumes a constant (average) rate of production, but even this is surely open to question. As a scientist gathers experience—and seniority—it could well be that his/her rate increases because of extra time for research or extra opportunities to collaborate. On the other hand there could be many external reasons for a disruption or interruption in the research career. (Some of these aspects were raised in Burrell (2007b).)

When it comes to the citation process, we would argue that the general ideas are valid but that in practice they are too prescriptive. Yes, the citation rate will vary over sources but, as remarked earlier, the gamma assumption is mainly for analytic convenience. And again, the assumption of a constant citation rate for any particular publication is open to question. What about, on the one hand, obsolescence and, on the other, sleeping beauties?

We happily acknowledge that the basic objection to the proposed model is that it assumes that everything flows smoothly in an unchanging external environment whereas real life is not like that. And that is why we should not expect such models to provide anything more than a general description of actual behaviour.

Our argument is that he graphical analysis in "Some empirical examples" section shows that the actual publication–citation process for individual authors does conform, at least in general terms, with that predicted by the theoretical stochastic model.

As a final remark, it should be stressed that the general form of the publication–citation distribution investigated here is very different from the (Type I) Pareto, or so-called (continuous) Lotka model, as in Egghe and Rousseau (2006), or the shifted version presented by Egghe and Rousseau (2012 a, b) or even the Pareto (Type II) analysed in detail by Burrell (2008). This is most clearly seen via graphical presentations of the different versions as in, e.g. Figs. 1, 2 and 3 of Burrell (2008).

Appendix

In the proofs we make free use of standard results regarding Poisson processes. The reader can refer to such standard references as Ross (1996) and Stirzaker (2005).

Proof of Proposition 2 The proof is constructed by conditioning on T, the time since publication, and Λ, the citation rate of a paper.

$$G_t(z) = E\left[z^{X_t}\right] = E_{T,\Lambda}E\left[z^{X_t}|T,\Lambda\right]$$

Given T and Λ, we have that citations following a Poisson distribution with mean ΛT so that the pgf is given by

$$E\left[z^{X_t}|T,\Lambda\right] = \exp[\Lambda T(z-1)]$$

Now averaging over T, which has a uniform distribution on $[0, t]$, we find

$$\begin{aligned} E\left[z^{X_t}|\Lambda\right] &= E_T E\left[z^{X_t}|T,\Lambda\right] = \int f_T(s)e^{\Lambda s(z-1)}ds = \int_0^t \frac{1}{t}e^{\Lambda s(z-1)} \\ &= \frac{1}{t\Lambda(z-1)}\left[e^{\Lambda s(z-1)} - 1\right] \end{aligned} \tag{A1}$$

Finally averaging over Λ, which has a gamma distribution as in (2), we get

$$\begin{aligned} G_t(z) = E\left[z^{X_t}\right] &= E_\Lambda E\left[z^{X_t}|\Lambda\right] = f_\Lambda(\lambda)\frac{1}{t\lambda(z-1)}\left[e^{\lambda t(z-1)} - 1\right]d\lambda \\ &= \frac{1}{t(z-1)}\int_0^\infty \frac{\alpha^v\lambda^{v-1}}{\Gamma(v)}e^{-\alpha\lambda}\frac{1}{\lambda}\left[e^{\lambda t(z-1)} - 1\right]d\lambda \\ &= \frac{\alpha^v}{t(z-1)\Gamma(v)}\int_0^\infty \left[\lambda^{v-2}e^{-\lambda(\alpha - t(z-1))} - \lambda^{v-2}e^{-\alpha\lambda}\right]d\lambda \\ &= \frac{\alpha^v}{t(z-1)\Gamma(v)}\left(\frac{\Gamma(v-1)}{(\alpha - t(z-1))^{v-1}} - \frac{\Gamma(v-1)}{\alpha^{v-1}}\right) \end{aligned}$$

(recognising each of the integrands as being proportional to a gamma pdf)

$$\begin{aligned} &= \frac{\alpha}{(v-1)t(z-1)}\left[\frac{1}{(1-(t/\alpha)(z-1))^{v-1}} - 1\right] \\ &= \frac{1}{(v-1)s(z-1)}\left[\frac{1}{(1-s(z-1))^{v-1}} - 1\right] \end{aligned}$$

(where $s = t/\alpha$)

$$\begin{aligned} &= \frac{1}{(v-1)s(z-1)}\left[\left(\frac{1/(1+s)}{1-(s/(1+s)z)}\right)^{v-1} - 1\right] \\ &= \frac{1}{(v-1)s(z-1)}\left[\left(\frac{p}{(1-qz)}\right)^{v-1} - 1\right] \end{aligned} \tag{A2}$$

where, as before, $p = \frac{1}{1+s} = \frac{\alpha}{\alpha+t} = 1 - q$

For the power series representation, this follows from expanding the first term inside the square brackets of (A2) using the general binomial expansion for negative powers and then simplifying. Thus

$$\left(\frac{p}{(1-qz)}\right)^{\nu-1} - 1 = \left(\frac{1/(1+s)}{1-sz/(1+s))}\right)^{\nu-1} - 1 = 1 - (s/(z-1))^{-(\nu-1)} - 1$$
$$= \left(1 + (\nu-1)s(z-1) + \frac{(\nu-1)\nu}{2!}\left(s(z-1)^2\right) + \cdots\right) - 1$$
$$= (\nu-1)s(z-1) + \frac{(\nu-1)\nu}{2!}\left(s(z-1)^2\right) + \frac{(\nu-1)\nu(\nu+1)}{3!}\left(s(z-1)^3\right) + \cdots$$

Now dividing through by $(\nu - 1)s(z - 1)$ gives the result. □

Remark The power series representation could alternatively have been derived by expanding (A1) in standard power series form and then integrating term by term with respect to the gamma pdf of Λ.

Proof of Proposition 3 This is an application of a well-known result for probability generating functions. If the pgf $G(z)$ of a random variable X is differentiated wrt z r times then

$$G^{(r)}(z) = \frac{d^r G(z)}{dz^r} = E\left[X(X-1)\ldots(X-r+1)z^{X-r}\right]$$

so that $G^{(r)}(1) = E[X(X-1)\ldots(X-r+1)]$

Thus successively differentiating the power series expansion (7) and putting $z = 1$ each time gives the result. □

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology, 54*(6), 550–560.

Burrell, Q. L. (1980). A simple stochastic model for library loans. *Journal of Documentation, 36*(2), 115–132.

Burrell, Q. L. (1990). Using the gamma-Poisson model to predict library circulations. *Journal of the American Society for Information Science, 41*(3), 164–170.

Burrell, Q. L. (1992). A simple model for linked informetric processes. *Information Processing and Management, 28*(5), 637–645.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics, 52*, 3–12.

Burrell, Q. L. (2002a). On the nth-citation distribution and obsolescence. *Scientometrics, 53*, 309–323.

Burrell, Q. L. (2002b). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology, 53*, 232–235.

Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology, 54*(5), 372–378.

Burrell, Q. L. (2005a). The use of the generalised Waring process in modelling informetric data. *Scientometrics, 64*(3), 247–270.

Burrell, Q. L. (2005b). Are "Sleeping Beauties" to be expected? *Scientometrics, 65*(3), 381–389.

Burrell, Q. L. (2007a). Hirsch's h-index: a stochastic model. *Journal of Informetrics, 1*(1), 16–25.

Burrell, Q. L. (2007b). Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics, 73*(1), 19–28.

Burrell, Q. L. (2007c). On the h-index, the size of the Hirsch core and Jin's A-index. *Journal of Informetrics, 1*(2), 170–177.
Burrell, Q. L. (2007d). Hirsch's h-index and Egghe's g-index. In D. Torres-Salinas & H. F. Moed (Eds.), *Proceedings of ISSI 2007* (Vol. 1, pp. 162–169). Madrid: Centre for Scientific Information and Documentation (CINDOC).
Burrell, Q. L. (2007e). On Hirsch's h, Egghe's g and Kosmulski's h(2). *Scientometrics, 79*(1), 79–91.
Burrell, Q. L. (2008). Extending Lotkaian informetrics. *Information Processing and Management, 44*(5), 1794–1807.
Burrell, Q. L. (2009). The publication/citation process at the micro level: A case study. *Journal of Scientometrics and Information Management, 3*(1), 71–77.
Burrell, Q. L. (2012). Alternative thoughts on uncitedness. *Journal of the American Society for Information Science and Technology, 63*(7), 1466–1470.
Burrell, Q. L. (2013). Formulae for the h-index: A lack of robustness in Lotkaian informetrics? *Journal of the American Society for Information Science and Technology*, (Accepted for publication).
Burrell, Q. L., & Cane, V. R. (1982). The analysis of library data. (With discussion.). *Journal of the Royal Statistical Society (Series A), 145*(4), 439–471.
Burrell, Q. L., & Fenton, M. R. (1993). Yes, the GIGP really does work—and is workable! *Journal of the American Society for Information Science, 44*(2), 61–69.
Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics, 69*(1), 131–152.
Egghe, L. (2010). The Hirsch index and related impact measures. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 44, pp. 65–114). Medford, NJ: Information Today.
Egghe, L., Guns, R., & Rousseau, R. (2011). Thoughts on uncitedness: Nobel laureates and Fields Medalists as case studies. *Journal of the American Society for Information Science and Technology, 62*(8), 1637–1644.
Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics, 69*(1), 121–129.
Egghe, L., & Rousseau, R. (2012a). Theory and practice of the shifted Lotka function. *Scientometrics, 91*(1), 295–301.
Egghe, L., & Rousseau, R. (2012b). 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.
Etkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and Mode 2 to a Triple Helix of university-industry-government relations. *Research Policy, 29*(2), 109–123.
Morse, P. M. (1976). Demand for library materials: an exercise in probability analysis. *Collection Management, 1*, 47–78.
Ross, S. (1996). *Stochastic processes* (2nd ed.). New York: John Wiley.
Schubert, A., Glanzel, W., & Braun, T. (1989). Scientometric datafiles—A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981–1985. *Scientometrics, 16*(1–6), 3–478.
Sichel, H. S. (1985). A bibliometric distribution which really works. *Journal of the American Society for Information Science, 36*, 314–321.
Stirzaker, D. (2005). *Stochastic processes and models*. Oxford: Oxford University Press.
Xekalaki, E., & Zografi, M. (2008). The generalized Waring process and its applications. *Communications in Statistics - Theory and Methods, 37*, 1835–1854.
Zografi, M., & Xekalaki, E. (2001). The generalised Waring process. In E. A. Lypitakis (Ed.), *Proceedings of the 5th Hellenic-European Conference on Computer Mathematics and its Applications, Athens* (pp. 886–893). Athens: HERCMA.