A model for the size-frequency function of co-author pairs

EGGHE, Leo (2008) A model for the size-frequency function of co-author pairs. In: JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 59(13). p. 2133-2137.

DOI: 10.1002/asi.20900
Handle: http://hdl.handle.net/1942/8488

A model for the size-frequency function of co-author pairs

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium
leo.egghe@uhasselt.be

ABSTRACT

Lotka's law was formulated to describe the number of authors with a certain number of publications. Empirical results [S.A. Morris and M.L. Goldstein. JASIST 58(12), 1764-1782, 2007] indicate that Lotka's law is also valid if one counts the number of publications of co-author pairs.

This paper gives a simple model proving this to be true with the same Lotka exponent if the number of co-authored papers is proportional to the number of papers of the individual co-authors. Under the assumption that this number of co-authored papers is more than proportional to the number of papers of the individual authors (to be explained in the paper), we can prove that the size-frequency function of co-author pairs is Lotkaian with an exponent that is higher than the one of the Lotka function of individual authors, a fact that is confirmed in experimental results.

---

[1] Permanent address

Key words and phrases: size-frequency function, co-author pairs, Lotka.
Acknowledgement: The author is grateful to Profs. Dr. C. Borgman, J. Furner, W. Glänzel, H. Kretschmer and R. Rousseau for interesting discussions on the topic of this paper.

I. Introduction

The law of Lotka is the celebrated decreasing power law introduced in Lotka (1926) to describe the number of authors (in a certain field) with $n = 1,2,3,...,n_{max}$ publications (n_{max} is the highest number of publications of an author). The simple function is

$$f(n) = \frac{C}{n^{\alpha}} \qquad (1)$$

where $C > 0$, $\alpha > 1$. The exponent α is the most important one (the parameter C is only used to make sure that the summation of (1) for all $n = 1,2,3,...$ gives the total number T of authors). Usually $1 < \alpha < 3$ with $\alpha = 2$ as a classical midpoint. It is well-known that $\alpha = 2$ is a turning point in informetrics in the sense that derived functions change forms at $\alpha = 2$, e.g. in case of the semi-logarithmic Leimkuhler curve (Groos droop or not) or the cumulative first-citation distribution, cf. Egghe (2005), Rousseau (1988), Groos (1967), Egghe (2000).

The simple function (1) was found to be valid by Lotka in case of senior author counts (i.e. where in co-authored papers only one author (the senior) receives a credit of one and the other authors receive a credit of zero). This way, Lotka circumvented the problem of fractional counting (i.e. where in co-authored papers, each author receives a credit of one divided by the total number of authors), cf. Egghe, Rousseau and Van Hooydonk (2000), in which case (1) is not valid – cf. Egghe (1993), Kretschmer and Rousseau (2001) and Egghe and Rao (2002).

Yet, if total counting is applied (i.e. where in co-authored papers, each author receives a credit of one, cf. Egghe, Rousseau and Van Hooydonk (2000)) one still has the validity of (1) for the size-frequency function of author production, cf. Egghe (1994).

Another way of looking at author production is by studying co-author pairs and the number of their (joint) publications. This topic is becoming more and more important since collaboration increases in time, cf. Lipetz (1999), Schubert (2002).

A bivariate distribution, based on (1), is studied in Kretschmer and Kretschmer (2007), hence producing three-dimensional graphs. In Morris and Goldstein (2007), another approach is

followed. Here one studies the (univariate) distribution $\varphi(n)$ denoting the number of co-author pairs with $n = 1,2,3,...$ (joint) publications. Hence, a classical framework is studied here but where "authors" are replaced by co-author pairs. In fact, in Morris and Goldstein (2007), one studies both functions $f(n)$ as in (1) for authors and $\varphi(n)$ (described above) for co-author pairs.

In one draws the graph of (1) on a log-log scale it is clear that we obtain a decreasing straight line with slope $-\alpha$. The data in Morris and Goldstein (2007) clearly show this form (cf. Figs. 7, 8 and 9, graphs (b)). In the same Figs., graphs (e), graphs in a log-log scale are found for the number $\varphi(n)$ of co-author pairs with n common papers. They clearly show the same linear trend with slopes smaller than or equal to $-\alpha$ (the ones for the author size-frequency), hence the validity of Lotka's law with Lotka exponents larger than or equal to α.

It is the purpose of this paper to present a rationale for this, under some simple assumptions. Under the assumption that the number of co-authored papers is proportional to the number of papers per author we prove that $\varphi(n)$, the size-frequency distribution for the number of papers of co-authored pairs, is Lotkaian with the same exponent as in (1). Under the assumption that the number of co-authored papers is more than proportional to the number of papers of the individual authors (in a way to be expressed exactly in the sequel), we can even prove that $\varphi(n)$ is Lotkaian but with a Lotka exponent which is larger than α in (1), a fact that is confirmed in the graphs in Morris and Goldstein (2007). This will be executed in the next section, both in the discrete and continuous setting.

The closing section draws conclusions and presents open problems.

II. The size-frequency function of co-author pairs

Throughout this paper we will suppose that (1) is the size-frequency function for the number of authors with $n = 1,2,...,n_{max}$ publications (papers). Then

$$T = \sum_{n=1}^{n_{max}} f(n) \tag{2}$$

is the total number of authors.

Let $\varphi(k)$ denote the size-frequency distribution of the number of co-author pairs with k (joint) publications. Then we have the following theorem.

Theorem II.1 :
Let f and φ be as above. Suppose that the number of joint papers of 2 co-authors is proportional to these authors' fraction of papers in the paper set. Then we have that, for every k (to be specified further),

$$\varphi(k) = \frac{D}{k^{\alpha}} \tag{3}$$

where α is the same as in (1) and where $D > 0$ is a constant.

Proof:
Let A denote the number of papers in the paper set. By assumption we have that, if the first author has m papers in total and the second author has n papers in total, then the probability for a joint paper is $\frac{m}{A}.\frac{n}{A}$. Indeed, $\frac{m}{A}$ is the probability (in the paper set) (or fraction) to have a paper of the first author and $\frac{n}{A}$ is the same for the second author. The proportionality assumption for a joint paper by these two authors is expressed by the independence rule: probability for a joint paper equals $\frac{mn}{A^2}$. Otherwise said, $\frac{mn}{A^2}$ is the probability to pick the

same paper in two independent trials, one with probability $\frac{m}{A}$ (to pick m papers) and one with probability $\frac{n}{A}$ (to pick n papers), the same paper then being a paper written by these two co-authors. Since there are A papers in total, these two co-authors hence have a number k of joint papers equalling

$$k = \frac{mn}{A} \tag{4}$$

(note that k is not necessarily an entire number but even if $k \in \mathbb{N}$, the distribution $\varphi(k)$ will describe the properties inherited from $f(n)$).

Note that the same k can be found using different pairs (m,n) (e.g. $(m,n)=(1,12)$ or $(m,n)=(2,6)$ or $(m,n)=(3,4)$, …or m and n reversed). For a fixed pair (m,n) we have that the probability to have an author with m papers in total and to have an author with n papers in total equals

$$\frac{f(m)}{T}\cdot\frac{f(n)}{T} \tag{5}$$

by (2). This contributes to the probability to have a co-author pair with production k. We only have to sum up for all possible pairs (m,n) that yield (4).

Hence, by definition of $\varphi(k)$ we have

$$\varphi(k)=\sum_{n=1}^{n_{max}} \frac{f\left(\frac{Ak}{n}\right)f(n)}{T^2} \tag{6}$$

$$=\frac{1}{T^2}\sum_{n=1}^{n_{max}} \frac{C}{\left(\frac{Ak}{n}\right)^{\alpha}}\frac{C}{n^{\alpha}}$$

$$= \frac{C^2}{T^2}\frac{n_{max}}{A^\alpha}\frac{1}{k^\alpha}$$

which is (3) for

$$D = \frac{C^2 n_{max}}{T^2 A^\alpha}.$$ □

Note:

If we allow all $n = 1,2,...,n_{max}$ as in (6), we can end up with several cases where $\frac{Ak}{n}$ Ï ¥ . Since, in the discrete case (1) we want to restrict ourselves to arguments in ¥ we can approximate $\frac{Ak}{n}$ by the natural number which is closest to $\frac{Ak}{n}$, in which case the calculation above is approximate. We still think that the above heuristic argument sheds some light on why the size-frequency distribution of co-author pairs is of the form (3). This problem will not be encountered later on where we will use continuous variables for the arguments of f and φ.

We will now assume, more generally, that the number of joint papers of 2 co-authors is higher than proportional to these authors' fractions of papers in the paper set (cf. Borgman and Furner (2002): "higher RATES of collaboration are usually associated with higher productivity"). Borgman and Furner (2008) base themselves on earlier work by Pao (1992), Bordons and Gomez (2000), Subramanyam (1983), Beaver and Rosen (1979), Price and Beaver (1966), Pao (1981, 1982), Zuckerman (1967).

For a discussion of the assertions around the relation between collaboration and production, we refer the reader to the "Conclusions and open problems" section, where we formulate an open problem around this theme, which possible validity would imply some of the earlier assertions mentioned above.

Proportionality was expressed by independence in the proof of the above theorem: the probability for a joint paper of two authors with m and n papers in total was $\frac{mn}{A^2}$. Since $m,n < A$ obviously we can express a higher probability for a joint paper by

$$p = \left(\frac{mn}{A^2}\right)^{\beta} > \frac{mn}{A^2} \tag{7}$$

where $0 < \beta < 1$ (since $m, n < A^2$). Under this assumption we can prove the next theorem.

Theorem II.2:

Under the same notation as above and supposing (7) for the probability for a joint paper of 2 co-authors we have that, for every k,

$$\varphi(k) = \frac{E}{k^{\frac{\alpha}{\beta}}} \tag{8}$$

hence a Lotka law with exponent

$$\delta = \frac{\alpha}{\beta} > \alpha \tag{9}$$

Proof:

The proof follows the lines of the one of Theorem II.1. Now, if (by (7))

$$p = \left(\frac{mn}{A^2}\right)^{\beta}$$

is the probability for a joint paper then, since there are A papers, we have

$$k = A\left(\frac{mn}{A^2}\right)^{\beta} \tag{10}$$

joint papers of these two authors. Again, as in (6), we have to account for all possible combinations of m and n, yielding k. Therefore we now have

$$\varphi(k) = \sum_{n=1}^{n_{max}} \frac{f\left(\frac{A^2}{n}\left(\frac{k}{A}\right)^{\frac{1}{\beta}}\right) f(n)}{T^2} \qquad (11)$$

$$= \frac{1}{T^2} \sum_{n=1}^{n_{max}} \frac{C}{\left(\frac{A^2}{n}\left(\frac{k}{A}\right)^{\frac{1}{\beta}}\right)^{\alpha}} \frac{C}{n^{\alpha}}$$

$$= \frac{C^2 n_{max}}{T^2 A^{2\alpha - \frac{\alpha}{\beta}}} \frac{1}{k^{\frac{\alpha}{\beta}}}$$

which is (8) for

$$E = \frac{C^2 n_{max}}{T^2 A^{2\alpha - \frac{\alpha}{\beta}}}$$

Note that (8) is the law of Lotka for co-author pairs and that now the exponent

$$\delta = \frac{\alpha}{\beta} > \alpha$$

since $0 < \beta < 1$. □

Note:

The same remark as in the note following Theorem II.1 applies to the argument given in Theorem II.2 (and more specifically equation (11)) above.

We will now present the arguments, given in Theorems II.1 and II.2, in a continuous setting. Now equation (1) is replaced by

$$f(j) = \frac{C}{j^{\alpha}} \qquad (12)$$

where $f(j)$ denotes the density of the number of authors with publication density $j \in [1,\rho]$ and where $C > 0$ and $\alpha > 1$ - cf. Egghe (2005), Chapter II. Note that now (2) is replaced by

$$T = \int_1^{\rho} f(j)dj \tag{13}$$

for the total number T of authors.

We have the following results, being the continuous analogues of Theorems II.1 and II.2.

Theorem II.3:

Under the notation as above we have

(i) If the publication density k of co-author pairs is given by (or is proportional with)

$$k = \frac{jj'}{A} \tag{14}$$

(cf. (4)), where j and $j' \in [1,\rho]$ are the publication densities of the individual authors, we have that the size-frequency distribution of co-author pairs $\varphi(k)$ is given by

$$\varphi(k) = \frac{D}{k^{\alpha}} \tag{15}$$

where α is the same as in (12) and where $D > 0$ is a constant.

(ii) If the publication density k of co-author pairs is given by (or is proportional with)

$$k = A\left(\frac{jj'}{A^2}\right)^{\beta} \tag{16}$$

(cf. (10)), where $0 < \beta < 1$, then we have that the size-frequency distribution of co-author pairs $\varphi(k)$ is given by

$$\varphi(k) = \frac{E}{k^{\frac{\alpha}{\beta}}} \tag{17}$$

where $E > 0$ is a constant and where the Lotka exponent is

$$\delta = \frac{\alpha}{\beta} > \alpha \tag{18}$$

Proof:

The proof follows the lines of the proofs of Theorems II.1 and II.2. Now (6) is replaced by

$$\varphi(k) = \int_1^{\rho} \frac{f\left(\frac{Ak}{j}\right) f(j)}{T^2} dj \tag{19}$$

using (14) (deleting the eventual proportionality factor). Note that there are no problems with the argument $\frac{Ak}{j}$ of f now: by (16) and since $j, j' \in [1, \rho]$ we have that $Ak \in [1, \rho^2]$ and hence $\frac{Ak}{j} \in [1, \rho]$, which is the domain of f.

Equation (19) now gives

$$\varphi(k) = \frac{1}{T^2} \int_1^{\rho} \frac{C}{\left(\frac{Ak}{j}\right)^{\alpha}} \frac{C}{j^{\alpha}} dj$$

$$= \frac{C^2(\rho - 1)}{T^2 A^{\alpha}} \frac{1}{k^{\alpha}}$$

which is of the form (15).

(iii) Similarly, with (16),

$$\varphi(k) = \int_1^{\rho} \frac{f\left(\frac{A^2}{j}\left(\frac{k}{A}\right)^{\frac{1}{\beta}}\right) f(j)}{T^2}\,dj$$

$$= \frac{C^2}{T^2}\int_1^{\rho} \frac{1}{\left(\frac{A^2}{j}\left(\frac{k}{A}\right)^{\frac{1}{\beta}}\right)^{\alpha}} \cdot \frac{1}{j^{\alpha}}\,dj$$

$$= \frac{C^2(\rho - 1)}{T^2 A^{2\alpha - \frac{\alpha}{\beta}}} \frac{1}{k^{\frac{\alpha}{\beta}}}$$

and where $\delta = \frac{\alpha}{\beta} > 1$. □

Note:

Although the continuous setting in Theorem II.3 solves the problems of the argument of f in the discrete setting, we still have a problem in case (ii) of Theorem II.3. Now, since $j, j' \in [1, \rho]$ are densities, it is not always so that $j, j' < A$ (this is so in the discrete case). But since $j, j' \in [1, \rho]$ and since $A > T$? 1 we will have that $\frac{jj'}{A^2} < 1$ in most cases. So we can conjecture that (16) still is a good continuous interpretation of the (always) correct equation (10). Note also that this problem does not occur in Theorem II.3 (i) since no β is applied to (14). Hence the result in Theorem II.3 (i) can be considered (within the used model) as correct.

Experimental evidence for these results are found in Morris and Goldstein (2007). There, three datasets are presented. By way of example we reproduce their Figs. 8(b) and (e), with permission, see Figs. 1 and 2 here. The data deal with a collection of papers on distance education. Fig. 1 presents the size-frequency function of the authors with a certain number of papers in a log-log scale. The linear trend is clear, conforming with Lotka's law (1). Fig. 2 presents the size-frequency function of co-author pairs with a certain number of papers in a log-log scale. Also here is there a linear trend, conforming with Lotka's law (1). However it is

clear that the linear trend in Fig. 2 has a lower negative slope $-\delta < -\alpha$, the slope of the linear trend in Fig. 1. Hence $\delta > \alpha$ conforming with Theorem II.2. This is also found in their Fig. 9 (b) and (e). On the other hand, their Fig. 7 (b) and (e) shows also decreasing linear trends, again with $\delta > \alpha$ but $\delta \gg \alpha$, conforming with Theorem II.1. In general we can say that $\delta \geq \alpha$ where δ is the Lotka exponent of the size-frequency function of co-author pairs and where α is the Lotka exponent of the size-frequency function of the authors themselves.

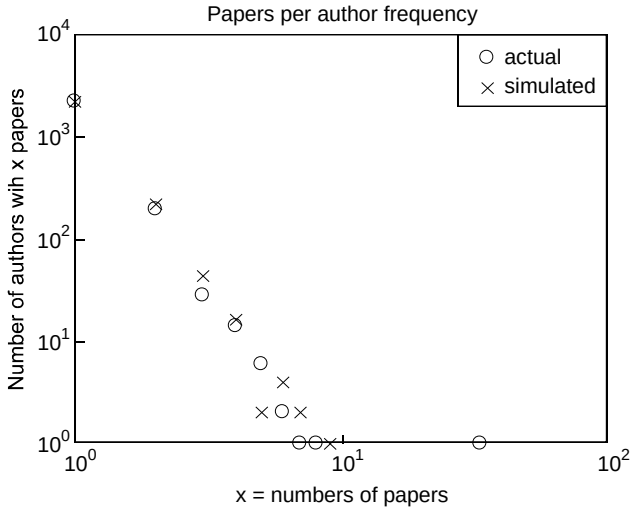

Fig. 1 Size-frequency function of authors.
Reprinted with permission of Wiley InterScience.

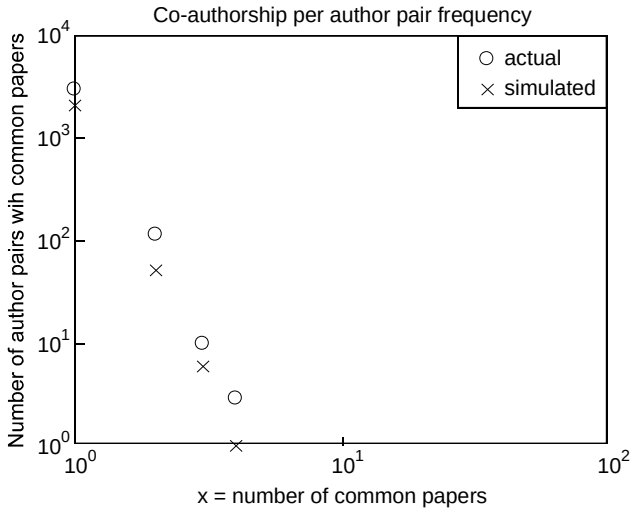

Fig. 2 Size-frequency function of co-author pairs
Reprinted with permission of Wiley InterScience.

Note: That Lotka's law also holds for co-author pairs has also been found in Kretschmer and Kretschmer (2008)

III. Conclusions and open problems

We showed that the size-frequency function of co-author pairs retains the Lotkaian form which is present in the size-frequency function of authors (vs. their papers). Experimental evidence of Morris and Goldstein (2007) that the Lotka exponent in the former case is larger than the one in the latter case is mathematically proved based on a principle that higher productivity leads to higher rates of collaboration, as advocated in e.g. Borgman and Furner (2002).

Although we have experimental evidence of Morris and Goldman, we did not find experimental results on the Borgman-Furner assertion, although we think it is a "logical" principle, certainly in its reverse orde: higher rates of collaboration should lead to higher production of the individual authors (where authorship should be counted in the total way – see Egghe, Rousseau and Van Hooydonk (2000)).

It is not easy to give experimental evidence for this assertion. To put things more clearly we state the following problem.

Problem:
Prove that, in general, the higher the number of papers of an author, the higher is his/her fraction of co-authored papers (with at least one co-author).

Of course, this will not be true for every author but we conjecture that, given a field, the cloud of points – each point representing the number of papers of an author (obscissa) versus his/her fraction of co-authored papers (ordinate) – has an increasing regression line.

Experimental evidence of this assertion is not available as is also confirmed by the colleages mentioned in the acknowledgement. Some "weaker" assertions or variants of the assertion

above are proved experimentally. Pao (1992) investigates global and local collaborators: global collaborators are authors that have co-authors from other laboratoria while local collaborators only have co-authors from their own lab. Pao proves that the global collaborators are much more productive than the local ones. Price and Beaver (1966) prove that researchers with many collaborators are far more productive than researchers with few collaborators. In Beaver and Rosen (1979) one finds that even in the period 1799-1830, the French scientific elite had a high average productivity in the group of scientists that collaborated. Zuckerman (1967) finds that Nobel laureates publish and collaborate more than a matched sample of scientists (a predictable fact). The study of Pao (1982) in computational musicology is a bit inconclusive with respect to collaboration and production, mainly due to the different sociological habits of the humanities (in comparison with the sciences) – see also Pao (1981).

The kind of problem as formulated above on collaboration are of increasing interest since, worldwide, collaboration in scientific publications is increasing (Lipetz (1999), Schubert (2002)).

References

Beaver D. deB. and Rosen R. (1979). Studies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830. Scientometrics 1(2), 133-149.

Bordons M. and Gómez I. (2000). Collaboration networks in science. Festschrift in honor of E. Garfield, ASIST, Chapter 10, 197-213.

Borgman C.L. and Furner J. (2002). Scholarly communication and bibliometrics. In: B. Cronin (ed.). Annual Review of Information Science and Technology 36. Medford, NJ: Information Today, 3-72.

Borgman C.L. and Furner J. (2008). Personal communication.

Egghe L. (1993). Consequences of Lotka's law in the case of fractional counting of authorship and of first author counts. Mathematical and Computer Modelling 18(9), 63-77.

Egghe L. (1994). Special features of the author-publication relationship and a new explanation of Lotka's law based on convolution theory. Journal of the American Society for Information Science 45(6), 422-427.

Egghe L. (2000). A heuristic study of the first-citation distribution. Scientometrics 48(3), 345-359.

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

Egghe L. and Ravichandra Rao I.K. (2002). Duality revisited: construction of fractional frequency distributions based on two dual Lotka laws. Journal of the American Society for Information Science and Technology 53(10), 789-801.

Egghe L., Rousseau R. and Van Hooydonk G. (2000). Methods for accrediting publications to authors or countries: consequences for evaluation studies. Journal of the American Society for Information Science 51(2), 145-157.

Groos O.V. (1967). Bradford's law and the Keenan-Atherton data. American Documentation 18, 46.

Kretschmer H. and Kretschmer T. (2007). Lotka's distribution and distribution of co-author pairs' frequencies. Journal of Informetrics 1(4), 308-337.

Kretschmer H. and Kretschmer T. (2008). Distribution of co-author pairs' frequencies: self-similarity and power laws. Preprint.

Kretschmer H. and Rousseau R. (2001). Author inflation leads to a breakdown of Lotka's law. Journal of the American Society for Information Science and Technology 52(8), 610-614.

Lipetz B.-A. (1999). Aspects of JASIS authorship through five decades. Journal of the American Society for Information Sience 50(11), 994-1003.

Lotka A.J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16(12), 317-324.

Morris S.A. and Goldstein M.L. (2007). Manifestation of research teams in journal literature: a growth model of papers, authors, collaboration, coauthorship, weak ties and Lotka's law. Journal of the American Society for Information Science and Technology 58(12), 1764-1782.

Pao M.L. (1981). Co-authorship as communication measure. Library Research 2, 327-338.

Pao M.L. (1982). Collaboration in computational musicology. Journal of the American Society for Information Science 33(1), 38-43.

Pao M.L. (1992). Global and local collaborators: s study of scientific collaboration. Information Processing and Management 28(1), 99-109.

Price D.J. De Solla and Beaver D. DeB (1966). Collaboration in an invisible college. American Psychologist 21, 1011-1018.

Rousseau R. (1988). Lotka's law and its Leimkuhler representation. Library Science with a Slant to Documentation and Information Studies 25(3), 150-178.

Schubert A. (2002). The Web of Scientometrics. A statistical overview of the first 50 volumes of the journal. Scientometrics 53(1), 3-20.

Subramanyam K. (1983). Bibliometric studies of research collaboration. Journal of Information Science 6, 33-38.

Zuckerman H. (1967). Nobel laureates in science: patterns of productivity, collaboration, and authorship. American Sociological Review 32(3), 391-403.