Proof of a conjecture of Moed and Garfield on authoritative references and extension to non-authoritative references

LEO EGGHE,[a,b,c] I. K. RAVICHANDRA RAO,[a] BIBHUTI BHUSAN SAHOO [a]

[a] *Documentation Research and Training Centre (DRTC), Indian Statistical Institute (ISI), Bangalore (India)*
[b] *Universiteit Hasselt (UHasselt), Diepenbeek (Belgium)*
[c] *Universiteit Antwerpen (UA), Campus Drie Eiken, Wilrijk (Belgium)*

In a recent paper [H. F. MOED, E. GARFIELD: In basic science the percentage of "authoritative" references decreases as bibliographies become shorter. *Scientometrics* 60 (3) (2004) 295–303] the authors show, experimentally, the validity of the statement in the title of their paper. In this paper we give a general informetric proof of it, under certain natural conditions. The proof is given both in the discrete and the continuous setting.

An easy corollary of this result is that the fraction of non-authoritative references increases as bibliographies become shorter. This finding is supported by a set of data of the journal Information Processing and Management (2002 + 2003) with respect to the fraction of conference proceedings articles in reference lists.

I. Introduction

Why do scientists cite? Many arguments can be given (cf. WEINSTOCK (1971) and EGGHE & ROUSSEAU (1990)). Of a higher level of explanation one can make distinction between the normative theory of citation and the social theory of citation, according to MERTON (1988) (see also MOED & GARFIELD (2004)). The normative theory of citation states that scientists cite the "necessary" references, i.e. the ones on which they base their article on; otherwise said: they cite to give credit where credit is due (MOED & GARFIELD (2004)). The social theory of citation focusses on social benefits that one can gain from citing: persuasion, citing important background sources, improving ones position in a scientific community.

Of course, in general, reference lists will show a mixture of both theories of citation and the question raised by Moed and Garfield is: how does the length of a reference list influences the nature of the references (as described above). Experimental data in Physics and Astronomy and also in Molecular Biology and Biochemistry reveal that there is an increasing relationship between the length of a reference list and the *fraction* of authoritative references. The latter is in MOED & GARFIELD (2004) defined as the

ones belonging to the 10% most frequently cited documents in the field. Of course, other percentages could be used. In other words, if the length of the reference list decreases, the fraction (or percentage) of authoritative sources decreases.

Rephrased in terms of the above described normative and social theories of citation, one could hence conclude that the normative theory applies more to the shorter reference lists and the social theory applies more to the longer reference lists.

In the next section we will describe how this problem can be rephrased in terms of the dual theory of informetrics, i.e. in terms of sources and items, including the use of a general decreasing size frequency functions f, such as the one (but not exclusively) of Lotka (cf. the theories developed in EGGHE (1990, 2005), EGGHE & ROUSSEAU (1990). In this model we generally define "authoritative" sources (extending the notion to sources above a certain threshold m>1 of items they contain) and then we prove that the fraction of authoritative sources is an increasing function of the total number of sources as expressed by the increase of the maximal number of items per source (if reference lists increase (in the $\subset$ sense) the maximal number of items in a source cannot decrease). This proves the Statement of Moed and Garfield. We give the proof, both in the discrete and the continuous setting.

As an obvious corollary of the above theory we also prove (in Section III) that, when reference lists gets longer, the fraction of *non*-authoritative sources (expressed as sources below a certain threshold m>1 of items they contain) becomes shorter.

This finding is then illustrated by considering the journal Information Processing and Management in the years 2002 and 2003 combined. Non-authoritative sources are (here) "defined" to be conference proceedings articles. A graph of the fraction of these conference proceedings articles versus the length of the reference list shows a cloud of points filling the lower triangle which could be considered as a "semi-decreasing" relationship in the sense that, the longer the reference list, the shorter *and* lower the range of the possible fractions of conference proceedings articles.

A similar (but weaker) finding (outside the theory of "(non)-authoritative sources" – as we think) is, experimentally, given for the fractions of references to books in a reference list.

II. Informetrics theory of reference lists and (non-) authoritative references

II.1 Dual framework

In dual informetrics (also called two-dimensional informetrics, cf. EGGHE (1990, 2005), EGGHE & ROUSSEAU (1990)) sources are considered to be producing items. Classical examples are journals (as sources) and journal articles (as items). But articles can also be sources "producing" references or citations.

In order to interpret the situation, described in MERTON (1988) and MOED & GARFIELD (2004), in the framework of sources and items we will consider here (we think for the first time) references as *sources* and the items they generate are the received citations. This will enable us to also interpret the notion of "authoritative" source in the source-item framework. This can be done in two equivalent ways. Firstly we define an authoritative source as one belonging to the a% most productive sources (within a given field Ω, say a reference is an authoritative reference if it is among the a% most frequently cited references). In Moed & Garfield one takes $a = 10$ but this is just an example.

If we fix the field (represented by Ω) then we can also define an authoritative source in a different but equivalent way (that is more suitable in this paper): in a fixed field (and interpreting sources and items as above) there is a one-to-one correspondence between the percentage $\underline{a}$ (as above) and m = the minimum number of citations that an element (a reference) x in this part of Ω must have: indeed, if we rank the elements in Ω decreasingly according to the number of citations they receive, then limiting ourselves to the top a% of these references determines m being the number of citations to this last reference that appears in this truncated list (threshold). Conversely, limiting ourselves to the references with m or more citations determines the top references and their fraction of the total number of references ($\#\Omega$) equals $a/100$, hence $\underline{a}$ is determined.

So, instead of a fixed percentage (as in MOED & GARFIELD (2004)), we will define an authoritative source as a reference receiving m (>1) or more citations and m will be fixed (if the field Ω is fixed). We will put, in the sequel, no further requirements on m so that we considerable extend the classical notion of authoritative sources to a group of sources with a minimum number $m>1$ of items. A more general name for this could be an "upper class group of sources" since it is more general than "authoritative sources". Since the latter are contained in the former and since this paper is devoted to the study of the Statement of Moed and Garfield, we will continue to use the notion of "authoritative sources" as generally defined above.

II.2 Explanation of the Statement of Moed and Garfield: Discrete setting

In the sequel we will use the term source for a reference in a reference list and item for the number of citations it receives (in a certain period). Let $f(n)$ denote the number of references receiving n citations, where n ranges (discretely) in $\{1,2,\ldots,n_{max}\}$, n_{max} being the highest number of citations (items) to a reference (source) in a first (long) reference list. For the second (shorter) reference list we will assume (for reasons of neutrality with respect to the studied problem) that we have, up to a positive constant, the same size-frequency function

$$f^*(n) = Df(n) \qquad (1)$$

$n = 1,\ldots, n^*_{max}$. The second list is assumed to be a subset of the first list, expressing the "more selectiveness" or "more selective" construction of the second reference list with respect to the first one – see MOED & GARFIELD (2004). Hence we have, necessarily

$$n^*_{max} < n_{max} \tag{2}$$

Theoretically $n_{max} = n^{\sim}_{max}$ is also possible but then we have $\leq$ in (2) and we will only be able to prove the Statement of Moed and Garfield in the non-strict sense. We henceforth will use (2).

Discussion on assumptions (1) and (2). As pointed out by one of the referees assumptions (1) and (2) are not always true. The authors certainly agree with this. Assumption (1) is, however, not controversial. It "limits" the problem to this "neutral" special case, also expressing that we make the comparison in the same field of research. Condition (2) is less evident. If we consider two reference lists (a long one and a short one) it is of course not so (even in the same field) that the shorter one is a subset of the longer one and even the weaker assumption (2) need not be true. Condition (2) needs to be interpreted – as indicated by one of the referees – in the universe of all citing papers in a field. In the sense, increasing selectivity is expressed by considering a genuine subset of the larger one, which implies (2). Another interpretation of (2) is as follows. An author of a paper with a (long) reference list can be asked to be more selective by shortening the reference list. Then we are certainly in case (2). Finally, inequality (2) can also be interpreted, in a fixed field, but considering two types of papers e.g. short communications versus "regular" papers or "regular" papers versus e.g. review papers. In each of these examples, one paper of the first set can be "matched" to a paper of the second set where (2) is valid (and where we even have – approximately – that the shorter reference list is a subset of the longer one, an assumption that is not used in this paper: only the much weaker (2) is used).

Condition (2) is in this sense logical and certainly, the opposite relation ($n^*_{max} > n_{max}$) would be counter-intuitive. In the sequel we will be able to prove the exact statement of Moed and Garfield, only using (1) and (2), indicating a logical special case in which this conjecture is proved. This should then give evidence for the validity of the conjecture of Moed and Garfield (in general) over an entire field.

To give an example of (1) and its "neutrality" meaning we can use

$$f(n) = \frac{C}{n^{\alpha}} \tag{3}$$

$n = 1,\ldots, n_{max}$ and

$$f^*(n) = \frac{DC}{n^\alpha} = \frac{E}{n^\alpha} \tag{4}$$

$n = 1,\ldots, n^*_{max}$, where both functions are laws of Lotka *with the same exponent* α, expressing that the field Ω remains the same and also expressing that Lotka's exponent in f^* is neither larger nor smaller than the one in f, expressing neutrality with respect to the problem under study. But we underline that, in the sequel, we only need a general positive f (hence also f^* is positive).

As described above, we have a fixed number $m>1$ as minimal number of received citations in the fixed field Ω, for a reference to be defined (generally) "authoritative". The fraction of authoritative references in the first list is then

$$\frac{\sum_{n=m}^{n_{max}} f(n)}{\sum_{n=1}^{n_{max}} f(n)} \tag{5}$$

The fraction of authoritative references in the second list is then (same m since authoritativeness depends on the field Ω, which is fixed) and not on the reference list:

$$\frac{\sum_{n=m}^{n^*_{max}} f^*(n)}{\sum_{n=1}^{n^*_{max}} f^*(n)} = \frac{\sum_{n=m}^{n^*_{max}} f(n)}{\sum_{n=1}^{n^*_{max}} f(n)}, \tag{6}$$

using (1). If we define

$$\varphi(p) = \frac{\sum_{n=m}^{p} f(n)}{\sum_{n=1}^{p} f(n)} \tag{7}$$

we hence have proved the Statement of Moed and Garfield if we can show that φ is strictly increasing in p. This is done now.

Theorem II.2.1: φ strictly increases.

Proof:

$$\varphi(p_1) < \varphi(p_2)$$

$$\Leftrightarrow$$

$$\frac{\sum_{n=m}^{p_1} f(n)}{\sum_{n=1}^{p_1} f(n)} \quad \frac{\sum_{n=m}^{p_2} f(n)}{\sum_{n=1}^{p_2} f(n)}$$

$$\left(\sum_{n=m}^{p_1} f(n)\right)\left(\sum_{n=1}^{p_2} f(n)\right) \quad \left(\sum_{n=1}^{p_1} f(n)\right)\left(\sum_{n=m}^{p_2} f(n)\right)$$

$$\left(\sum_{i=m}^{p_1} f(i)\right)\left(\sum_{j=m}^{p_2} f(j) \quad \sum_{j=1}^{m-1} f(j)\right) \quad \left(\sum_{i=m}^{p_1} f(i) \quad \sum_{i=1}^{m-1} f(i)\right)\left(\sum_{j=m}^{p_2} f(j)\right)$$

$$\sum_{i=m}^{p_1} f(i) \sum_{j=1}^{m-1} f(j) \quad \sum_{i=1}^{m-1} f(i) \sum_{j=m}^{p_2} f(j)$$

$$p_1 < p_2$$

since $f > 0$. □

We now show that also the average number of citations per reference strictly increases in n_{max}, hence if the reference lists become shorter, the average number of citations per reference decreases. This can be regarded as a variant of the Statement of Moed and Garfield, because of Theorem II.2.1 and the next

Theorem II.2.2:

$$\frac{\sum_{n=1}^{n_{max}} nf(n)}{\sum_{n=1}^{n_{max}} f(n)} \tag{8}$$

strictly increases in n_{max}.

Proof:

$$\frac{\sum_{n=1}^{p_1} nf(n)}{\sum_{n=1}^{p_1} f(n)} \quad \frac{\sum_{n=1}^{p_2} nf(n)}{\sum_{n=1}^{p_2} f(n)}$$

$$\sum_{i=1}^{p_1} if(i) \sum_{j=1}^{p_2} f(j) \quad \sum_{i=1}^{p_1} f(i) \sum_{j=1}^{p_2} jf(j)$$

$$\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} if(i)f(j) \quad \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} jf(i)f(j)$$

$$p_1 < p_2$$

since $f > 0$. □

Note that in the above proofs we only used that $f > 0$. It is, however, clear that only decreasing size-frequency functions are needed in informetrics.

II.3 Explanation of the Statement of Moed and Garfield: Continuous setting

We refer to EGGHE (2005) for general results on continuous models (and Lotkaian models in particular, but we do not need this here) for size-frequency functions f(j), $j \in [1, \rho_m]$: j = item density, ρ_m = maximal item density. We now have

$$\frac{\int_1^{\rho_m} jf(j)dj}{\int_1^{\rho_m} f(j)dj} \tag{9}$$

Theorem II.3.1:

μ strictly increases in ρ_m

Proof:

The proof follows the lines of the proof of the discrete case (Theorem II.2.2):

$$\frac{\int_1^{\rho_1} jf(j)dj}{\int_1^{\rho_1} f(j)dj} < \frac{\int_1^{\rho_2} jf(j)dj}{\int_1^{\rho_2} f(j)dj}$$

$$\Leftrightarrow$$

$$\int_{i=1}^{i=\rho_1}\int_{j=1}^{j=\rho_2} if(i)f(j)didj < \int_{i=1}^{i=\rho_1}\int_{j=1}^{j=\rho_2} jf(i)f(j)didj$$

$$\Leftrightarrow$$

$$\rho_1 < \rho_2$$

since $f > 0$. □

Now let, as in the discrete case,

$$\varphi(\rho) =: \frac{\int_1^{\rho} f(j)dj}{\int_1^{\rho} f(j)dj}, \qquad (10)$$

i.e. the fraction of authoritative sources.

Theorem II.3.2:

φ strictly increases.

Proof:

Also this proof follows the lines of the one of Theorem II.2.1:

$$\varphi(\rho_1) < \varphi(\rho_2)$$

$$\Leftrightarrow$$

$$\frac{\int_m^{\rho_1} f(j)dj \quad \int_m^{\rho_2} f(j)dj}{\int_1^{\rho_1} f(j)dj \quad \int_1^{\rho_2} f(j)dj}$$

$$\int_m^{\rho_1} f(i)di \int_1^{\rho_2} f(j)dj \quad \int_1^{\rho_1} f(i)di \int_m^{\rho_2} f(j)dj$$

$$\left(\int_m^{\rho_1} f(i)di\right)\left(\int_m^{\rho_2} f(j)dj \quad \int_1^{m} f(j)dj\right) \quad \left(\int_m^{\rho_1} f(i)di \quad \int_1^{m} f(i)di\right)\left(\int_m^{\rho_2} f(j)dj\right)$$

$$\int_{i\ m}^{i\ \rho_1} \int_{j\ 1}^{j\ m} f(i)f(j)didj \quad \int_{i\ 1}^{i\ m} \int_{j\ m}^{j\ \rho_2} f(i)f(j)didj \quad \int_{i\ m}^{i\ \rho_2} \int_{j\ 1}^{j\ m} f(i)f(j)didj$$

$$p_1 < p_2$$

since $f > 0$. □

Note that, even in the case that the size frequency function f is Lotkaian:

$$f(j) = \frac{C}{j^{\alpha}} \tag{11}$$

$j \in [1, \ _m]$, $C > 0$, we can have that a reference list can be shortened (i.e. a lower number of sources), keeping the same exponent , such that (hence $_m$ by Theorem II.3.1) increases: following the theory in EGGHE (2005) (Theorem II.2.1.2.1, pp. 116–117) we have, if $1 < 2$ that for every $A > T > 0$ given (T = total number of sources, A = total number of items) there exist $_m > 1$ and $C > 0$ such that (11) gives the values A and T (via (II.20) and (II.21) in EGGHE (2005)).

Examples:

1. $\alpha = 1.5$, $A = 15{,}000$, $T = 10{,}000$ hence $\mu = A/T = 1.5$. The equation for ρ_m is given by (II.37), p. 118 in EGGHE (2005):

$$-\frac{x^{0.5}}{1.5} - x^{-0.5} + \frac{2.5}{1.5} = 0$$

yielding $\rho_m = x = 2.251$.

2. $\alpha = 1.5$, $A = 10{,}000$, $T = 5{,}000$ hence $\mu = A/T = 2$. So A and T are smaller than the corresponding values in Example 1 but μ is larger, hence also ρ_m is larger (by Theorem II.3.1). ρ_m is given by the equation (using again (II.37), p. 118 in EGGHE (2005)):

$$-\frac{x^{0.5}}{2} - x^{-0.5} + \frac{3}{2} = 0$$

yielding $\rho_m = 4$ exactly.

This is excluded in our explanation of the Statement of Moed and Garfield: from the discrete model we have (2) hence μ decreases by Theorem II.2.2. By Theorem II.3.1 we hence have that ρ_m decreases and Theorem II.3.2 shows that the Statement of Moed and Garfield is also proved in the continuous case (as we did already in the discrete case).

III. Extension of the Statement of Moed and Garfield to the case of non-authoritative sources

If we express "non-authoritative sources" as sources with a number of items *below* a certain threshold $m > 1$, we can prove the opposite Statement of Moed and Garfield:

Statement for non-authoritative sources:

If the number of sources decreases, the fraction of non-authoritative sources increases.

In terms of reference lists: if reference lists become shorter, the percentage of non-authoritative references increases.

Proof:

We will give the proof for the continuous model; the one for the discrete model is similar.

Define

$$*(\ ) \quad \frac{\int_{1}^{m} f(j)dj}{\int_{1}^{\rho} f(j)dj} \tag{12}$$

given a size-frequency function f and a threshold $m > 1$ as in the previous section. Hence $*(\ )$ is the fraction of non-authoritative sources (more generally, the fraction of sources belonging to a "lower class group of sources").

Hence

$$*(\ ) \quad \frac{\int_{1}^{p} f(j)dj \quad \int_{m}^{p} f(j)dj}{\int_{1}^{\rho} f(j)dj}$$

$$*(\ ) \quad 1 \quad \frac{\int_{m}^{\rho} f(j)dj}{\int_{1}^{\rho} f(j)dj}$$

$$*(\ ) = 1 - \ (\ ) \tag{13}$$

The above Statement now follows from (13) and the fact that increases in (Section II).

Looking for confirmation of the above Statement we have considered conference proceedings articles which can be considered (or defined) as "non-authoritative" sources. We have analyzed the combined volumes 2002 + 2003 of the journal Information Processing and Management (IPM). Figure 1 shows the relation between the total number of references and the fraction (or percentage) of conference articles in each IPM paper. It is clear that the points of this scatter diagram fill a triangle situated in the lower parts of the abscissa and ordinate. We can call this a "semi-decreasing" relationship since, if the number of references increases the range for the fraction of conference articles becomes smaller and lower, showing the validity of this complementary Statement (complementary – but not in contradiction! – to the Statement of Moed and Garfield).

Outside the field of (non-) authoritative sources are books. Although the relation between the number of references and the fraction of books is weaker than in the case of conference articles (Figure 1), Figure 2 also shows a semi-decreasing relationship. We leave it to the reader for an explanation of this (weaker) phenomenon.

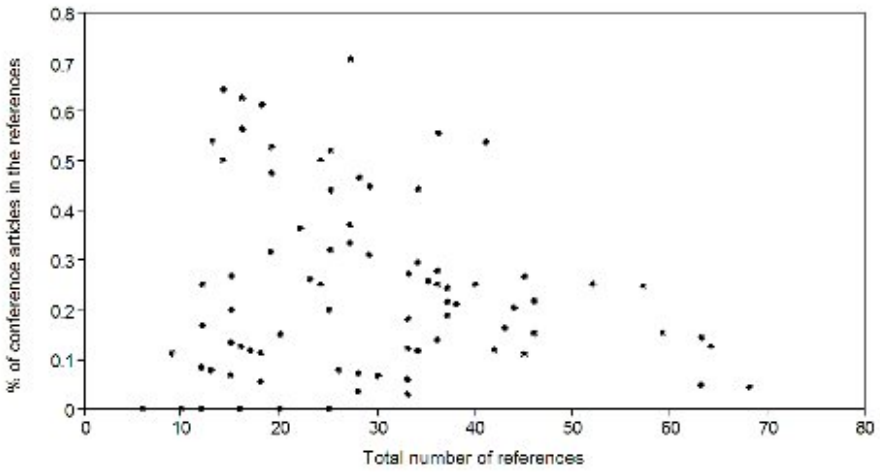

Figure 1. Scatter diagram of no. of references vs % of conference articles
(data from IPM 2002 & 2003)

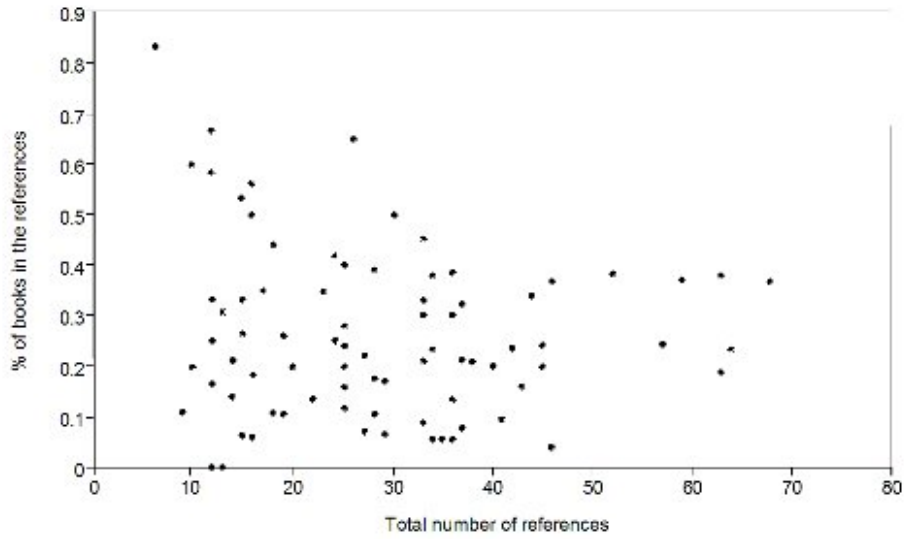

Figure 2. Scatter diagram of no. of references vs % of books
(data from IPM 2002 & 2003)

*

The first named author is grateful to the Indian Statistical Institute (ISI) for financial support during his stay at ISI as a visiting professor.

References

EGGHE, L. (1990), The duality of informetric systems with applications to the empirical laws. *Journal of Information Science*, 16 (1) : 17–27.

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

EGGHE, L., ROUSSEAU, R. (1990), *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science*. Elsevier, Amsterdam, the Netherlands.

MERTON, R. K. (1988), The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *ISIS*, 79 : 606–623.

MOED, H. F., GARFIELD, E. (2004), In basic science the percentage of "authoritative" references decreases as bibliographies become shorter. *Scientometrics*, 60 (3) : 295–303.

WEINSTOCK, M. (1971), Citation indexes. In: *Encyclopedia of Library and Information Science*, 5, pp. 16–40, Marcel Dekker, New York. Reprinted in: *Essays of an Information Scientist*, 1, pp. 188–216, 1977, ISI Press, Philadelphia, USA.