Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

A bibliometric study in crystallography

Heinrich Behrens[a] **and Peter Luksch**[b]*

[a]Kolbenäckerweg 12, 76144 Stutensee, Germany, and [b]FIZ Karlsruhe, 76344 Eggenstein-Leopoldshafen, Germany

Correspondence e-mail: peter.luksch@fiz-karlsruhe.de

Received 24 May 2006
Accepted 2 August 2006

This is an application of the mathematical and statistical techniques of bibliometrics to the field of crystallography. This study is, however, restricted to inorganic compounds. The data were taken from the Inorganic Crystal Structure Database, which is a well defined and evaluated body of literature and data published from 1913 to date. The data were loaded in a relational database system, which allows a widespread analysis. The following results were obtained: The cumulative growth rate of the number of experimentally determined crystal structures is best described by a third-degree polynomial function. Except for the upper end of the curve, Bradford's plot can be described well by the analytical Leimkuhler function. The publication process is dominated by a small number of periodicals. The probability of the author productivity in terms of publications follows an inverse power law of the Lotka form and in terms of database entries an inverse power law in the Mandelbrot form. In both cases the exponent is about 1.7. For the lower tail of the data an exponential correction factor has to be applied. Multiple authorship has increased from 1.4 authors per publication to about four within the past eight decades. The author distribution itself is represented by a lognormal distribution.

1. Introduction

Bibliometrics can be defined as the application of methods of mathematical and statistical analysis to study and quantify the process of publication. Bibliometrics may comprise publication counts such as the growth rate of the number of publications in a discipline. It may be applied to collections of articles and periodicals for the investigation of the productivity of authors, or the use of citation analysis methods in a wider sense, to mention just the most important fields of application. Scientometrics is defined as the application of bibliometric techniques in science. Scientometrics usually goes beyond a simple application of such techniques by also examining the social and political background of scientific developments (see Diodato, 1994).

This paper presents a bibliometric study in the specific field of crystallography, but in this case is restricted to inorganic compounds. The study is based on the data published from 1913 to the present, which were taken from the Inorganic Crystal Structure Database (ICSD). The following topics are considered:

(i) the magnitude and growth rate of a number of experimentally determined specific crystal structures (entries in the ICSD) and the community of authors who published the corresponding measurements;

Table 1
Basis data in the ICSD..

Information entity	Occurrence
Time coverage	1913 to present
Entries	86 306
Primary references	44 886
Periodicals	1132
Authors	33 478
Unique structures	66 172

(ii) the application of Bradford's law to periodicals publishing crystal structure data and the top journals ranked by percentage of crystallographic papers published in the area under consideration;

(iii) the repeated determination of a specific crystal structure;

(iv) the probability that a publication contains more than one entry;

(v) the productivity of authors (Lotka's law);

(vi) the question of multiple authorship.

Two earlier bibliometric studies in the field of crystallography should be mentioned in this context. The first study (Hawkins, 1980) dealt with the whole crystallographic literature (organic and inorganic) available at the time, including citation analyses, but was restricted to fewer topics and to the literature for the relatively small time range 1972–1976. The second study (Redman *et al.*, 2001) covers the Cambridge Structural Database and is a citation analysis, which is, however, not the focus of the present investigation.

2. Content of the Inorganic Crystal Structure Database (ICSD)

The ICSD is a comprehensive numerical database containing fully determined crystal structures of inorganic compounds (for more details see Bergerhoff & Brown, 1987; Behrens, 1996; Fluck, 1996; Belsky *et al.*, 2002). Since the application of X-ray diffraction to the determination of crystal structures was first demonstrated by Bragg in 1913, this method has become a powerful tool for the acquisition of crystallographic data, and the first papers referenced in the ICSD are from 1913. It is the goal of the ICSD to record all crystal structures published from 1913 to the present. The common aim of crystallographic databases is not only to record the relevant bibliographic information but also to provide the primary numerical results of crystal analysis experiments.

The ICSD has been loaded in a relational database system and offers unique capabilities for retrieving any desired information and for investigating relationships between various entities. It is therefore an ideal data pool for a bibliometric study of the publication behavior in a well defined body of literature and data over a time period of almost one century.

The database is organized by entries. One database entry describes the determination of one structure and is based on one main article, called the 'primary reference'. Besides this main reference the entry may also refer to other articles, called 'secondary references'. Most of the figures presented here are based on primary references because authors are given only for these primary reference articles.

Some structures were investigated several times by different authors under different experimental conditions. Therefore, the total number of entries in the database is larger than the total number of unique structures. In order to define unique structures the comparison of unit-cell parameters is the best tool, but this approach was not practicable for this study. Instead, a unique structure was defined by its formula, space group, Pearson symbol and Wyckoff sequence.

The definition of an inorganic compound was not applied in a very strict way, and over the past few years, an increasing number of metallic compounds have been added to the database. However, one can assume that the publication behavior of scientists working on inorganics does not differ from those working on metals. The ranking of journals, however, could be influenced by the types of compounds.

For a bibliometric study it is important to have reliable bibliographic data. In the current implementation of the ICSD in a relational database system, the author names are stored in a table with unique names that were checked manually in order to have a unique spelling for each name. All authors of a publication are recorded even if their number is unusually high. The author sequence is the same as in the published article.

The consistent nomenclature for journal names is another advantage of the ICSD database. A reference table for journal names ensures that each publication is assigned to the correct journal or periodical. The additional recording of issue, volume and page numbers allows the unambiguous identification of articles.

Data (as summarized in Table 1) were extracted for this study by running SQL queries in the MySQL version of the database.

3. Bibliometric laws, distributions and fits

Bradford's law and Lotka's law are major laws which are applied in bibliometrics. Bradford's law is applied to the distribution of publications (articles) in a set of periodicals in a particular discipline. Lotka's law describes the publication frequency of authors in a given field. Additional laws and distributions are explained in the following.

3.1. Growth phenomena

The growth in science was first discussed extensively by de Solla Price (see, for example, de Solla Price, 1963). Often simple models were used for the analytical description of growth rates. The following models are applied in the present paper: linear growth, quadratic growth, cubic growth and exponential growth.

Cumulative data are obtained by integration of per annum data over the time period considered. Consequently, if an exponential growth model for the per annum data is applied, the cumulative data will also grow exponentially. On the other

hand, a quadratic model for the per annum data will result in a cubic model for the cumulative data. Very often in the literature, *a priori* exponential growth is assumed. This assumption was originally suggested by de Solla Price (1963). However, as explicitly discussed by Behrens & Lankenau (2006), there are many cases in which it is not justified.

3.2. Bradford's Law

Bradford's (1934) law describes the distribution of the literature for a given subject in periodicals (journals).

The graphical representation of Bradford's law is a plot of the cumulative number of articles (vertical coordinate) *versus* the logarithm of the cumulative number of periodicals (horizontal coordinate), with the latter arranged in decreasing order of productivity. The plot usually has the form of an 'S' shape with a central straight section following Bradford's log-linear law. The upward curving bottom of the curve represents the nuclear zone of the most relevant periodicals. The upper end of the curve, usually termed the Groos droop, represents the peripheral zone showing saturation effects where relevant articles are widely scattered among a large number of periodicals. In many cases Bradford's law shows that a few very productive periodicals, as far as numbers of articles are concerned, are dominating the whole publication process in a specific area.

A detailed discussion and a derivation of the analytical Leimkuhler formula are, for example, given by Bookstein (1990). An analytical description is given by

$$Y = A\ln(1 + BN), \tag{1}$$

the form given by Leimkuhler (1967), where A and B are constants. The parameter B determines the curvature in the nuclear zone of the most relevant periodicals and the beginning of the log-linear zone, *i.e.* also the extent of the dominance of a few periodicals. For example, the yield ratio of 10 periodicals to 1000 periodicals has a value of 4% for $B = 0.01$, a value of 15% for $B = 0.1$ and a value of 35% for $B = 1$.

The upper end of the Bradford curve is not described by Leimkuhler's formula given above. Thus, the graphical representation is normally in closer agreement with the existing data as far as the saturation tail of the curve is concerned.

3.3. Lotka's Law

Lotka's (1926) law deals with the scientific productivity of authors in a particular area. In its strict form, it is an inverse-square law claiming that the number of authors of n papers in a population is about $1/n^2$ of the number of authors of one paper.

However, Lotka's law is nowadays defined in a more generalized form (see for example Bookstein, 1990), which reads as

$$g(n) = a/n^b, \tag{2}$$

where $g(n)$ represents the probability of an author making n published contributions to a subject area, and where a and b (b being the Lotka exponent) are characteristic parameters.

Such inverse power law distributions are quite natural models for complex systems. They are scale invariant, *i.e.* if n is multiplied by a factor c the scale is changed but not the form of the law. It is reproduced with the exception of a different value for the parameter a. Inverse power law distributions are known to be closely related to the problem of self-organization, which is often explained by the dynamic properties of a complex system characterized by self-organized criticality. Inverse power laws are observed in many fields. Examples are the frequencies of events of varying size in self-organized critical systems, *e.g.* the Gutenberg–Richter law of earthquake magnitudes or the occurrence of avalanches in sand piles (granular matter).

Mandelbrot (1982) formulated an even more generalized version of these laws, which can be written as

$$g(n) = a/(c + n)^b, \tag{3}$$

with the Mandelbrot parameter c being a further constant.

3.4. Lognormal distribution

Many data distributions in physics and biology show a more or less skewed behavior. Such skewed distributions can often be fitted by a lognormal distribution. The lognormal distribution is described by the following formula

$$y = \frac{1}{ax(2\pi)^{1/2}}\exp\left[-(\ln x - b)^2/2a^2\right]. \tag{4}$$

The mean is

$$\mu = \exp(b + a^2/2) \tag{5}$$

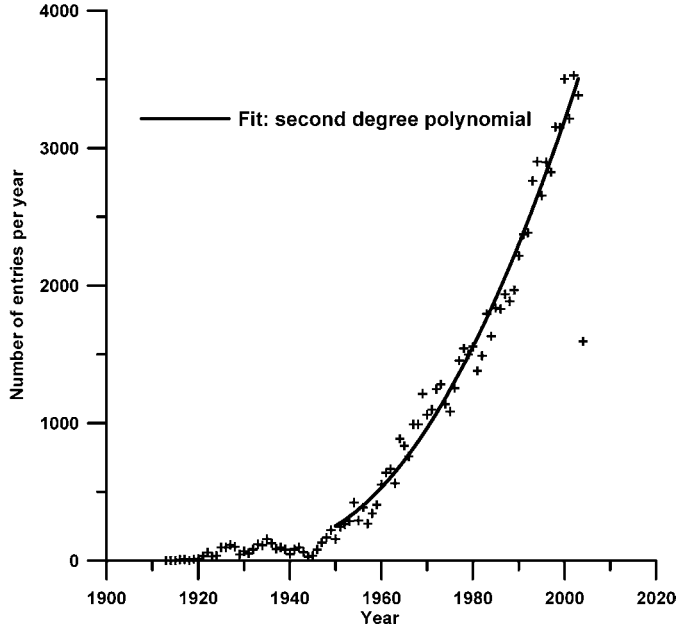

Figure 1
Number of entries per year.

and the variance

$$\sigma^2 = \mu^2\eta^2, \tag{6}$$

with

$$\eta^2 = \exp(a^2) - 1. \tag{7}$$

The lognormal distribution is determined by two parameters, *i.e.* the dispersion parameter a and location parameter b. The coefficient of variation σ/μ depends only on the dispersion parameter a. The relative positions of the median and mode are $x = \exp(b)$ and $x = \exp(b - a^2)$, respectively. The coefficient of skewness has the value $\gamma_1 = \eta^3 + 3\eta$.

3.5. Fits

Usually linear or nonlinear regression fits[1] a model to the data (the method of least squares). In this context a model is a mathematical description of a process. The goodness of fit is quantified by parameters like the sum-of-squares of the residuals or the related coefficient of determination[2] R^2. The goal is to obtain the best-fit values of the parameters of the model chosen. However, for the data treated in this paper there is no such basic theory. Therefore, the parameters of these models do not correspond to known basic processes. Normally, a model is not unique, and various other models could also be applied. It is therefore crucial to apply a sensible model. When comparing different models, one can only state that one specific model explains the data better than another, but not which model is correct. It is obvious that models with a larger number of parameters will yield better fits to the data. Therefore, it makes sense to compare only models with an equal or similar number of parameters. In the case of models with very different numbers of parameters, other statistical approaches should be used.

In the context of the inverse power laws it is also essential not to use truncated data sets for the fits as Perline (2005) recently pointed out. Otherwise, a power law could be mimicked. Perline makes a distinction between strong, weak and false inverse power laws. The term 'weak power law' (Perline, 2005) refers to the case where only the upper portion of the distribution is following an approximate inverse power law and the term 'false inverse power law' refers to the case where the observations of samples drawn from other distributions can mimic an inverse power law.

In all figures the probabilities are given in percent.

[1] In this paper the representation and fitting of the data has been carried out by the Software *Grapher 3* (Golden Software Inc, Golden, Colorado, USA).
[2] The value R^2 quantifies the goodness of fit and is a fraction between 0.0 and 1.0. The closer R^2 is to 1, the better the fit describes the data. The definition of R^2 is given as shown in the following. The method of least squares says that the parameters of the fitted equation are adjusted such that the sum of the squares of the residuals $S_2 = \sum(\hat{y}_i - y_i)^2$ is a minimum. y_i and x_i are the data values of the ordinate and of the abscissa, respectively. $\hat{y}_i = f(x_i)$ are the fitted ordinate values for the mathematical function $f(x)$. The smaller S_2 is at the minimum, the better the fitted function describes the data. Usually, the square sum S_2 of the residuals at the minimum is, however, not taken as a measure for the goodness of fit but a related quantity, *i.e.* $R_2 = S_1/(S_1 + S_2)$, with $S_1 = \sum(\hat{y}_i - \bar{y})^2$. In principle, R^2 is the amount of the variance $\sum(y_i - \bar{y})^2/(n - 1)$, which is explained by the fit.

4. Results

4.1. Growth rates

First the growth rate of the number of experimentally determined specific crystal structures is considered. Fig. 1 shows the increase of the number of entries in the ICSD published per year as a function of time (publication date) in a linear representation. During the Second World War a collapse of the growth rate occurred. After the war, the best fit is a quadratic function $n(t) = 0.78\ (t - t_0)^2 + 120$ in the time range 1950–2003 with $t_0 = 1937.0$ years (shown in Fig. 1). The coefficient of determination here is $R^2 = 0.981$. By using this fit an extrapolation into the future is possible. Thus, for the year 2010 a yearly publication rate of 4300 entries can be forecast, for the year 2020 a rate of 5500 and for 2030 a rate of 6800.

In addition, the corresponding cumulative number of entries is presented in Fig. 2 in semi-logarithmic representation. The cumulative numbers correspond to an integration of the yearly entries shown above. Thus, the curve is smoothed and the effects of the Second World War are no longer very explicitly visible. The data in Fig. 2 can be fitted by two models. The usual method is by a fit of an exponential function in the time range 1935–2004 (shown here) with $R^2 = 0.989$ and a doubling time $T = 10.6$ years. This doubling time is in agreement with earlier results (Behrens, 1996) and approximately in agreement with the doubling time of the growth of the whole chemical literature, which has a value of 14 years (Behrens & Lankenau, 2006). The second method uses a fit of a third-degree polynomial with a coefficient of determination $R^2 = 0.9997$, also in the time range 1935–2004 (shown in Fig. 3). This latter fit is much better than that by the exponential model where we have systematic deviations between fit and data. Therefore, only the cubic growth model should be used for forecasting the future. By carrying out such an extrapolation

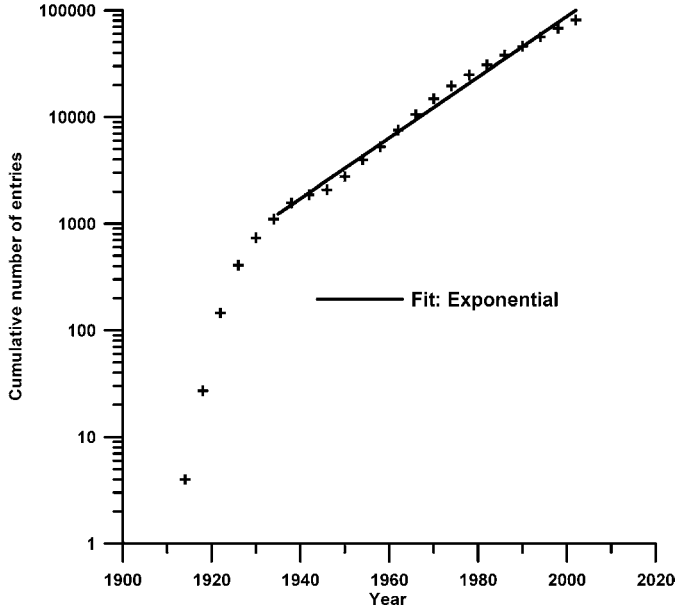

Figure 2
Cumulative number of entries, exponential fit. Here, only every second data point is shown in the figure.

we obtain 110 000 cumulative entries for the year 2010, 160 000 cumulative entries for the year 2020 and 210 000 cumulative entries for 2030.

In the next step, the magnitude and growth rate of the community of authors who published the structure determinations is evaluated. The corresponding representation is given in Fig. 4. In this representation the community of authors is defined as the number of individual authors (each name is counted only once) who published structures within a time span of 5 years. After the Second World War we have a continual increase which is best described by a linear growth with a rate of 203 authors per year in the time period 1960–1998 ($R^2 = 0.994$). This fact is in agreement with the results for other scientific communities (see Behrens & Lankenau, 2006).

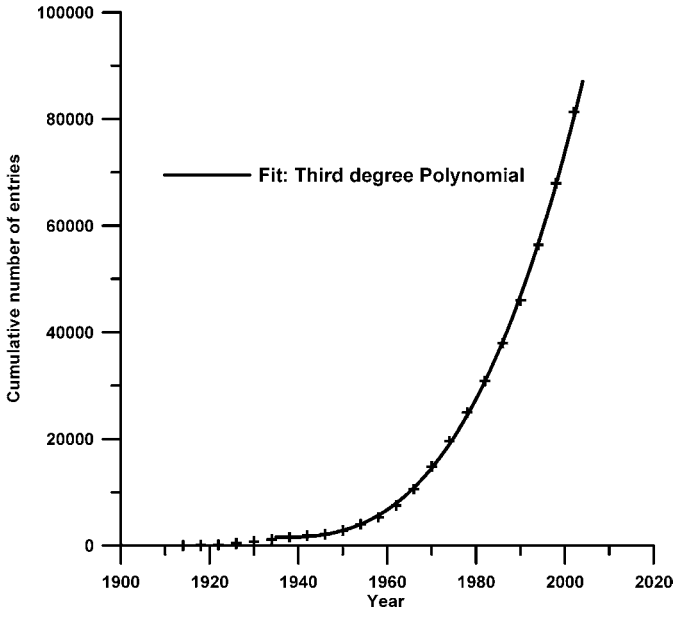

Figure 3
Cumulative number of entries, cubic fit. Here, only every second data point is shown in the figure.

4.2. Bradford's plot

In the following the application of Bradford's law to periodicals publishing crystal structure data will be treated. Fig. 5 shows the cumulative number of entries as a function of the number of periodicals (most of them journals) ranked in decreasing order of productivity. The scale is semi-logarithmic. By considering this figure we recognize that the nuclear zone covers the first eight periodicals, then we have the log-linear range from nine to 70 periodicals and finally the Groos droop area from 71 to 1056 periodicals. For the fit the Leimkuhler function $y = A\ln(1 + Bx)$ as discussed in §3.2 [see equation (1)] was used. By fitting the range of the first 70 periodicals the parameters $A = 20\,490$ and $B = 0.51$ ($R^2 = 0.998$) are obtained. The value of the B parameter is an indication of the dominance of few periodicals. In fact, 20 periodicals contain 59% of all entries, 70 periodicals contain 83%, 100 periodicals contain 88% and 150 periodicals 92%. Thus, about 120 periodicals practically cover the field of crystallography for inorganic compounds.

In addition, in Fig. 6 the percentage of total entries originating from the nine top journals is represented as a bar chart in order to give an impression of which journals were the most important ones. However, the fraction of these top journals was not constant over the time period investigated. Therefore, the share of the total entries for four top journals as a function of time for the past decades is given as a bar diagram in Fig. 7.

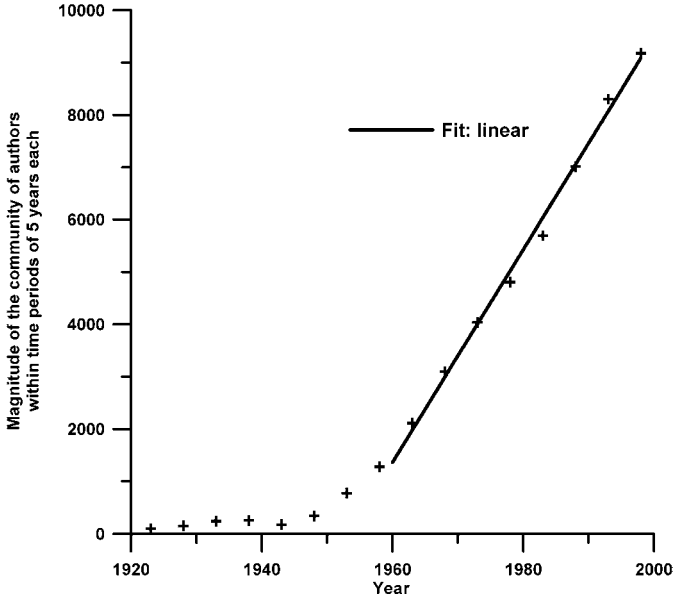

Figure 4
Magnitude of the community of authors evaluated in time periods of five years.

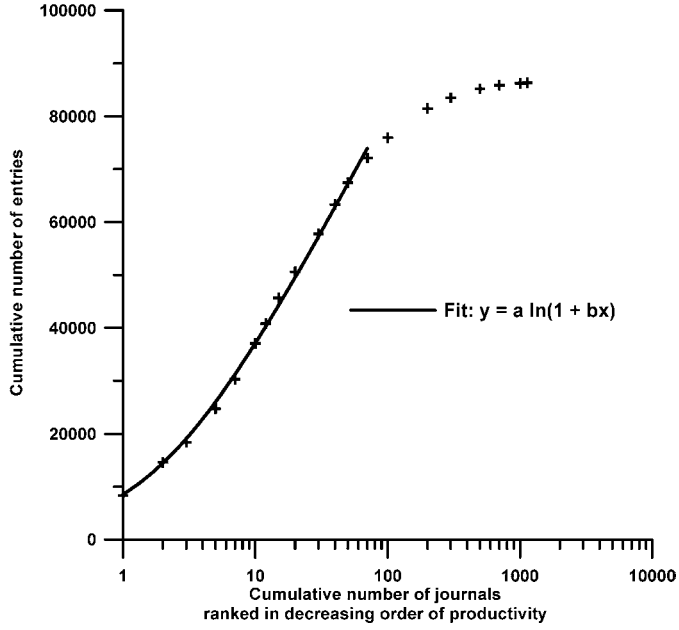

Figure 5
Bradford plot for the ICSD.

4.3. Repeated determinations of a specific crystal structure

A certain number of crystal structures have been experimentally investigated more than once, either to improve the accuracy of the measurement or to investigate different experimental conditions. The probability for such a repeated measurement is shown in Fig. 8 in double-logarithmic scale. The data can only be suitably described by two functions where the probability follows an inverse power law behavior, *i.e.* $y = a/x^b$, but with different sets of parameters a and b. For the range from one to seven investigations the parameters are $a = 85.2$ and $b = 3.13$ with $R^2 = 0.99995$, and for the range from seven to 164 investigations $a = 15.1$ and $b = 2.36$ with $R^2 = 0.988$. There is a sharp bend in the curve at seven investigations. The average number of investigations of a structure is 1.30.

4.4. Number of entries per publication

A publication may describe the determination of several structures. This yields several entries in the ICSD. The probability that a publication covers more than one entry has therefore been investigated. The result is given in Fig. 9 in double-logarithmic scale. For a good description of the data it is necessary to make use of a strongly modified inverse power law behavior, in the form of

$$y = a/x^b \exp(-c/x^2). \quad (8)$$

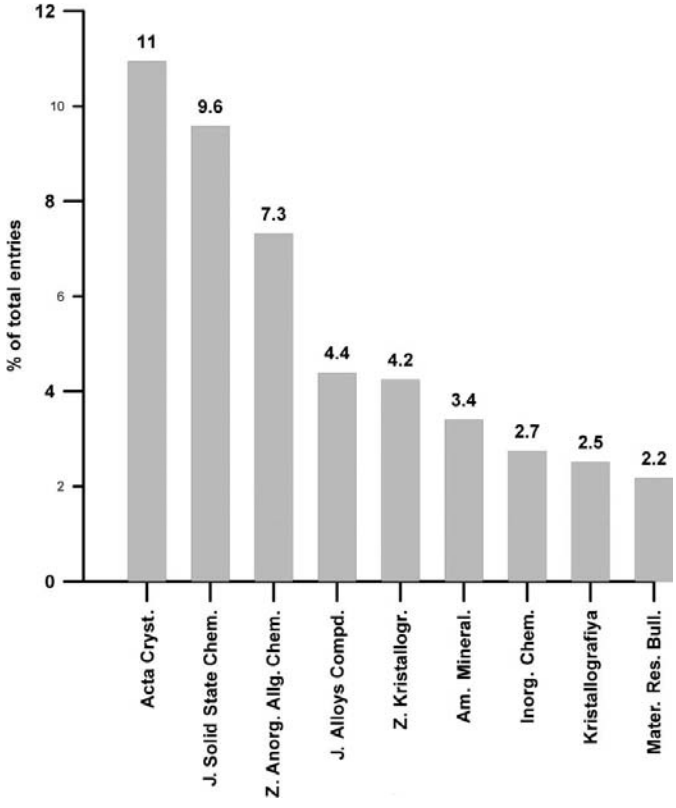

Figure 6
Percentage of total entries originating from the nine top journals. The terms 'Acta Cryst.' and 'Z. Kristallogr.' refer to all different journal sections in the series.

The parameters are $a = 143$, $b = 2.69$ and $c = 0.787$ with $R^2 = 0.99993$.

In this case we have an inverse power law for the lower tail of the data distribution, *i.e.* for the range from six to 87 entries per publication. It is therefore not possible to describe the entire range of data by an inverse power law. This can only be

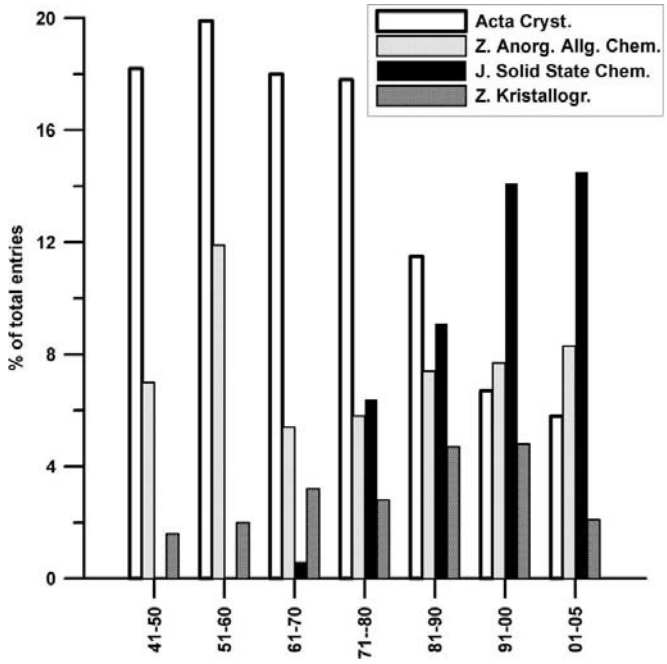

Figure 7
Percentage of total entries of some top journals as a function of time. The terms 'Acta Cryst.' and 'Z. Kristallogr.' refer to all different journal sections in the series.

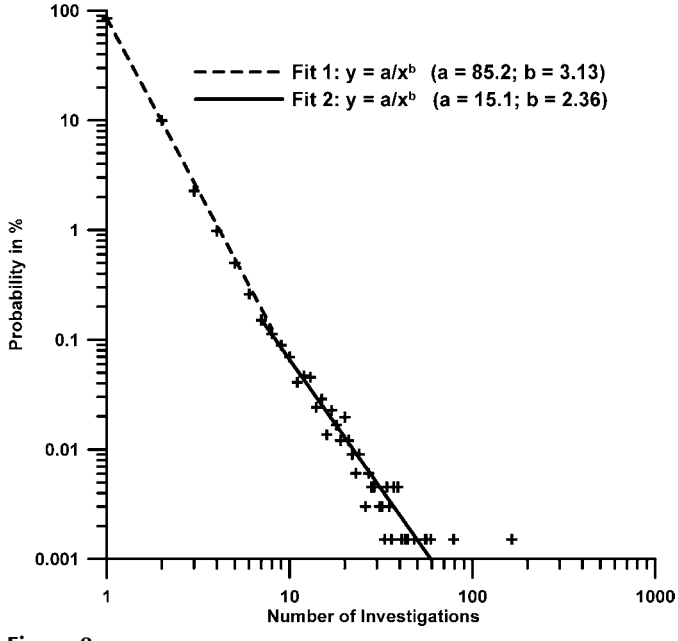

Figure 8
Probability for repeated investigations of a compound.

achieved by a different distribution function with a correction factor as shown above.

The average number of entries per reference is 1.92.

4.5. Productivity of authors

Two kinds of author productivities have been considered, namely productivity in terms of entries and productivity in terms of publications (the more common approach). First the number of publications produced by each author is discussed in order to try to verify Lotka's law. The result is given in Fig. 10 in double-logarithmic scale. By choosing the fit function

$$y = a/x^b \exp(-cx), \tag{9}$$

an excellent description of the data is obtained. This is an inverse power law, but modified by a correction factor $\exp(-cx)$. The parameters of the fit are given by $a = 54$, $b = 1.69$ and $c = 0.0194$ with $R^2 = 0.99995$. For $n \gtrsim 10$ the inverse power law is corrected in the sense that the productivity of highly productive authors is reduced. This is not unreasonable, as already shown by de Solla Price (1963) and especially by Perline (2005). Otherwise the contribution of authors with a very high productivity would be unreasonably high. As Bailón-Moreno *et al.* mentioned, Lotka's law presents good fits in the area of authors with low productivity in many other cases (Bailón-Moreno *et al.*, 2005). This finding is confirmed here.

Omission of the correction factor would also have the consequence that the calculation of the total number of publications in the ensemble by carrying out a summation of production probabilities[3] multiplied by publication frequencies of the authors would not converge.[4] Thus, the correction factor ensures correct behavior of the production function in the high productivity area. Another interesting point is that half the sum of all papers have been produced by authors with a productivity of more than nine publications. This point was also mentioned by de Solla Price (1963). The average number of publications per author is 4.19.

Next, the number of entries produced by each author is considered. The corresponding result is represented in Fig. 11 in double-logarithmic scale. Here, by choosing a somewhat different fit function,

$$y = a/(c + x)^b \exp(-dx), \tag{10}$$

an excellent description of the data is obtained. This is a modified inverse power law in the Mandelbrot form [see equation (3)], but again with a correction factor $\exp(-dx)$. The parameters of the fit are given by $a = 103$, $b = 1.73$, $c = 0.88$ and $d = 0.0063$ with $R^2 = 0.99992$. The same arguments as for the previous discussion on publications are valid here.

The average number of entries for each author is 8.36.

4.6. Multiple authorship

In many cases more than just one author contributed to the investigation of a specific crystal structure and therefore also to the publication. Multiple authorship will be discussed in the following. In Fig. 12 the probability distribution of the number

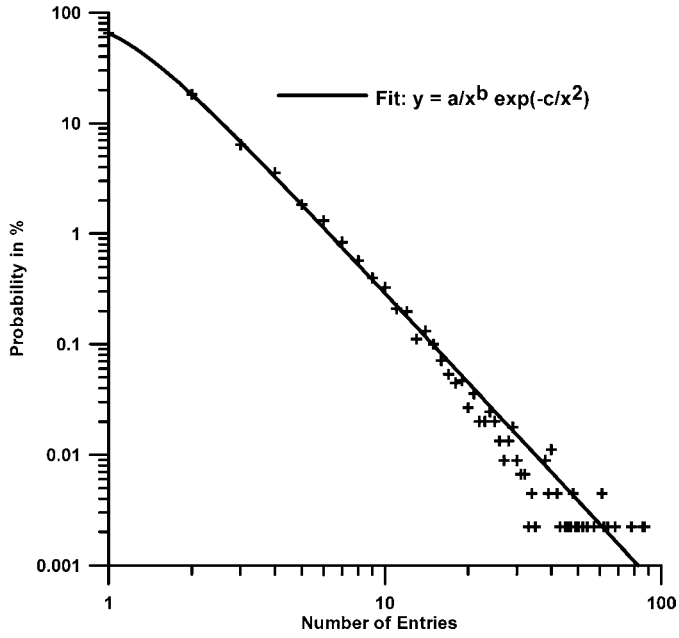

Figure 9
Probability for the number of entries per publication.

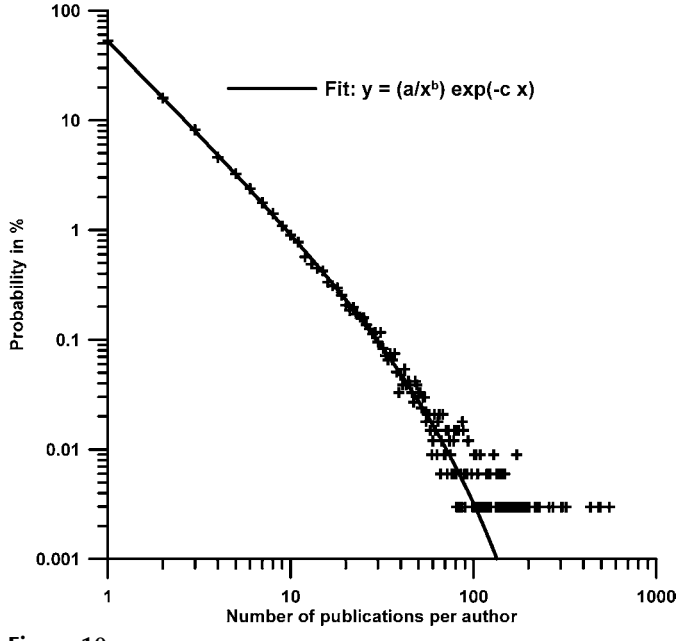

Figure 10
Probability for the number of publications per author.

[3] The probabilities have also been multiplied by the total number of authors.
[4] Another important point is the treatment of multiple authorship. Lotka in his original paper credited joint contributions to the senior author only (Lotka, 1926). But, in our case, we fully counted each author even when several authors published a paper together. Thus, the problem of multiple authorship and its influence on the law remains. On the other hand, as Bookstein showed, Lotka's law is not sensitive to how the authors are counted (Bookstein, 1990). However, the summation just mentioned does not give the correct total number of publications but a sum which is too large by a factor depending on the average number of authors per publication in the ensemble. In our case this average number is 3.09 (see the discussion in §4.6).

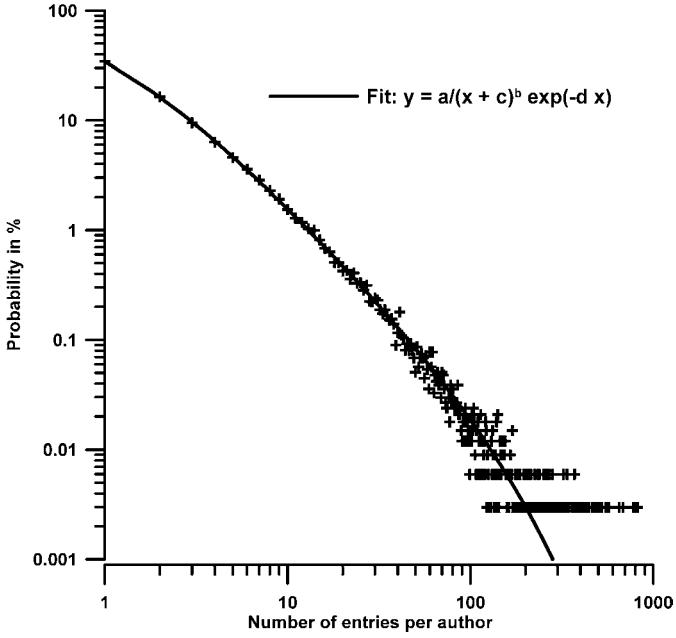

Figure 11
Probability for the number of entries per author.

of authors per entry is represented for the whole ensemble of data considered. The average number of authors per entry μ is 3.24. Possible statistical distributions in this context are the Erlang Distribution, the Inverse Gaussian (Wald) Distribution and the Lognormal Distribution. The best fit is obtained by the lognormal distribution[5] [see equation (4)], which is able to describe the data well with the parameters a = 0.55 and b = 1.07 with R^2 = 0.998. The distribution of the authors per publication is nearly identical and therefore not shown. The average number of authors per publication is 3.09.

However, the distribution of the number of authors per entry (and per publication) shows a strong time dependence, as illustrated in Fig. 13 for three different decades. It is evident that the number of authors per entry is increasing over time. However, the coefficient of variation (standard deviation/mean) of the distribution remains constant over the years and has a value of 0.47 $\pm$ 0.03.

Fig. 14 shows the average number of authors[6] per entry μ as a function of time for eight different decades. The fit has been carried out by an exponential equation [this follows from equation (5)] with the doubling time T = 47 years (coefficient of determination R^2 = 0.983). The corresponding curve (not shown here) for the average number of authors per publication is almost identical.

It is even more interesting to consider the time dependence of the parameters a and b. The location parameter b [in ln(years)] increases strictly linearly with the time t (in years) *i.e.* $b = 0.24\ (t - t_0)$ with t_0 = 1925.5. The fit has an R^2 of 0.986. On the other hand, the dispersion parameter a [in ln(years)] remains constant over time (a = 0.47). This latter result [see equations (5), (6) and (7)] is in agreement with the fact that the coefficient of variation σ/μ is constant over time, as mentioned above.

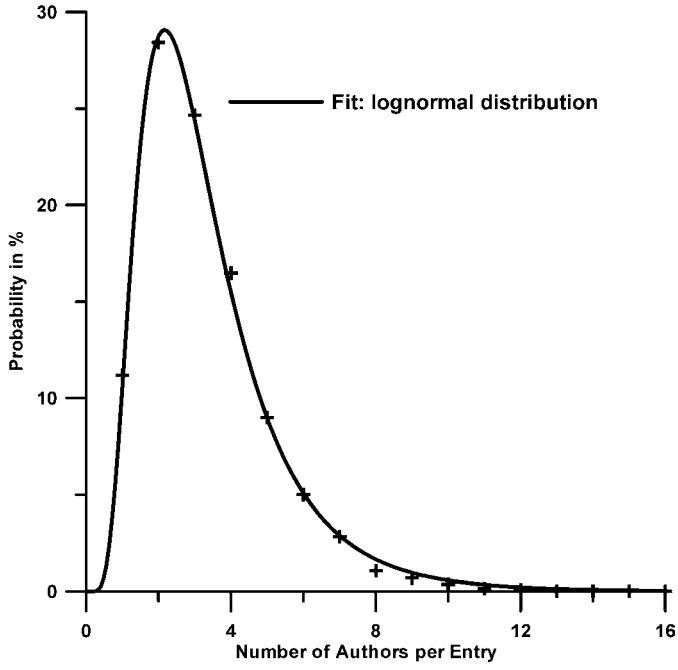

Figure 12
Distribution of the number of authors per entry for the total ensemble of data.

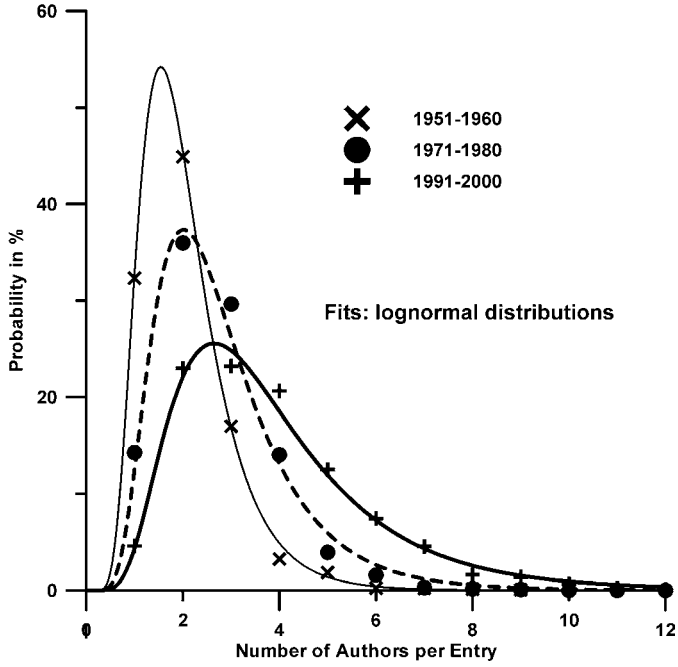

Figure 13
Time dependence of the distributions of the number of authors per entry.

[5] For the tail of the distribution (number of authors per entry > 8) an improved description can be obtained by introducing a correction factor of $\exp(-cx^5)$ to the lognormal distribution. For the whole ensemble of data the parameter c has a value of 4×10^{-6}. Then R^2 is improved from 0.998 to 0.999. It should be remarked that the parameter c of this correction factor is a function of time if the time dependence of the multiple authorship is considered.

[6] The average number of authors is calculated exactly in this study and not from the parameter sets of the fitted lognormal distributions.

5. Summary and discussion

The following conclusions can be drawn for the field of crystallography in the area of inorganic compounds.

(i) The growth of the cumulative number of entries after the Second World War is best described by a third-degree polynomial and not – as is often done – by an exponential growth. This result is in agreement with the growth in other disciplines of natural science, *i.e.* in astronomy/astrophysics, physics and chemistry (Behrens & Lankenau, 2006). The corresponding authors' community is growing linearly, which is also not unreasonable (Behrens & Lankenau, 2006).

(ii) Bradford's plot can be described relatively well in the nuclear and log-linear zone by the analytical Leimkuhler function with the parameter $B = 0.51$. The publication process itself is dominated by a small number of periodicals, *i.e.* 59% of all entries originated from only 20 periodicals. This situation is due to the fact that the subject field is very compact, as physics was in the 19th century. In the 20th century the subject field for physics is, on the other hand, not so compact ($b \simeq 0.09$) (see for example Bennion, 1986). Other fields, such as life science or sociology, are more diffuse and show a larger scattering in the journal distribution.

(iii) Complex systems often show a scale invariant behavior, *i.e.* the occurrence of inverse power laws. This is also the case here, but with certain modifications. For the probability of repeated determinations of a specific crystal structure, the data can be described by inverse power laws, but only by two equations with different parameters sets, one in the lower range of the abscissa values and one in the higher range of the abscissa values. In the lower range the exponent in the denominator is 3.1; in the higher range it is 2.4. Also, for the probability that a publication covers more than one entry, the entire range of data cannot be described by an inverse power law but only by a strongly modified distribution $y = a/x^b \exp(-c/x^2)$. Nevertheless, the lower tail of the data (number of entries $n > 5$) can be fitted well by an inverse power law with an exponent in the denominator of 2.7 because in this range the factor $\exp(-c/x^2)$ is negligible.

(iv) The productivity probabilities of authors (Lotka's law) could be described by inverse power laws [for entries in the Mandelbrot form of equation (3) and for publications in the generalized Lotka form of equation (2) with a Lotka exponent of about 1.7 in both cases]; however, both are multiplied by an exponential correction factor $\exp(-cx)$, which is important for the lower tail of the data (number of publications $n > 10$). The correction parameter c in this factor has a value of about 0.006 for entries and about 0.02 for publications. This is not improbable, as already shown by Bailón-Moreno *et al.* (2005). In the sense of Perline (2005) this behavior corresponds to a weak inverse power law.

(v) Today multiple authorship is usual. This was not the case 70 years ago, but the occurrence of multiple authorship has been increasing relatively strongly over the past few decades. In fact, the average number of authors per entry or per publication has been growing from a value of about 1.4 in the 1920s to a value of about four in the 1990s. The author distribution per entry can be described well by a lognormal distribution. A very interesting point is that the parameter b of the fitted lognormal distributions [see equation (4)] is growing strictly linearly with time. Another interesting point is that the coefficient of variation σ/μ is constant over time.

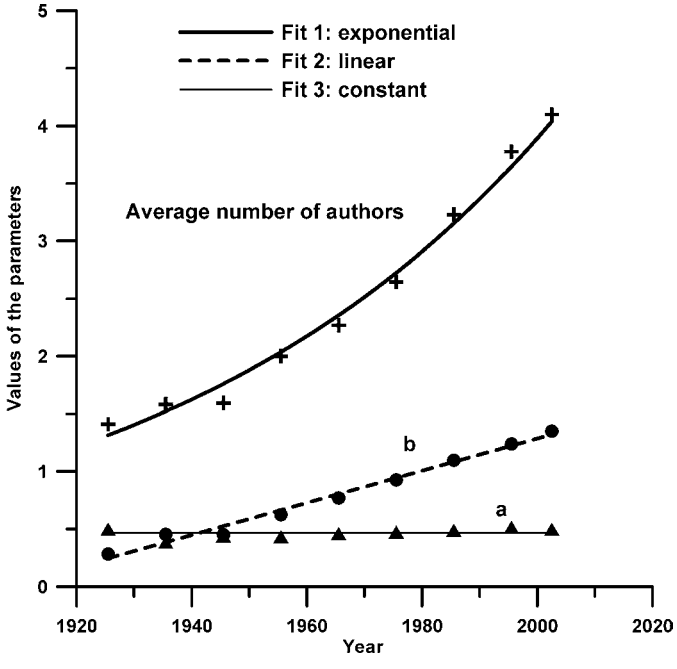

Figure 14
Average number of authors and parameters of the lognormal distribution as a function of time.

References

Bailón-Moreno, R., Jurado-Almeda, E., Ruiz-Baños, R. & Courtial, J. P. (2005). *Scientometrics*, **63**, 209–229.
Behrens, H. (1996). *J. Res. Natl Inst. Stand. Tech.* **101**, 365–373.
Behrens, H. & Lankenau, I. (2006). *Ber. Wissenschaftsgesch.* **29**, 89–108.
Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. (2002). *Acta Cryst.* B**58**, 364–369.
Bennion, B. C. (1986). *Czech. J. Phys. B*, **36**, 19–22.
Bergerhoff, G. & Brown, I. D. (1987). *Crystallographic Databases*, edited by F. H. Allen, G. Bergerhoff & R. Sievers, pp. 77–95. Chester: International Union of Crystallography.
Bookstein, A. (1990). *J. Am. Soc. Inf. Sci.* **41**, 368–375, 376–386.
Bradford, S. C. (1934). *Engineering*, **137**, 85–86.
Diodato, V. (1994). *Dictionary of Bibliometrics.* New York/London/Norwood: The Haworth Press.
Fluck, E. (1996). *J. Res. Nat. Inst. Stand. Technol.* **101**, 217–220.
Hawkins, D. T. (1980). *Acta Cryst.* A**36**, 475–482.
Leimkuhler, F. F. (1967). *J. Doc.* **23**, 197–207.
Lotka, A. J. (1926). *J. Wash. Acad. Sci.* **16**, 317–323.
Mandelbrot, B. (1982). *The Fractal Geometry of Nature.* New York: Freeman.
Perline, R. (2005). *Stat. Sci.* **20**, 68–88.
Redman, J., Willett, P., Allen, F. H. & Taylor, R. (2001). *J. Appl. Cryst.* **34**, 375–380.
Solla Price, D. J. de (1963). *Little Science, Big Science.* New York: Columbia University Press.