András Schubert—
a world of models and metrics

a festschrift on the occasion of his

70th

birthday

András Schubert—
A World of Models and Metrics

A Festschrift
on the Occasion of his 70th Birthday

This special volume of the ISSI Newsletter is sponsored by

- Springer
- :: schlemmer photo & media :: | s-press.hu
- Centre for Research & Development Monitoring (ECOOM), KU Leuven
- International Society for Scientometrics and Informetrics (ISSI)

András Schubert—
A World of Models and Metrics

A Festschrift
on the Occasion of His 70th Birthday

special volume of the e-zine of the
international society of scientometrics and informetrics

vol. 12–S, March 2016

Published by **ISSI**

András Schubert—A World of Models and Metrics.
A Festschrift on the Occasion of His 70th Birthday

Published as special volume of the
ISSI e-Newsletter, vol. 12–S March 2016
with the financial support of Springer

Cover, typesetting and graphical works:
:: schlemmer photo & media :: | s-press.hu

Printed by HVG Press Kft. (Budapest, Hungary)
March 2016

Contents

foreword

A World of Models and Metrics —A Festschrift on the Occasion of András Schubert's 70th Birthday

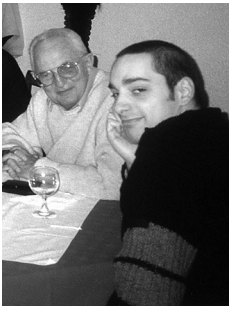
WOLFGANG GLÄNZEL & BALÁZS SCHLEMMER

This special issue of the ISSI periodical presents a collection of papers to honour a great personality in science on the occasion of his 70th birthday. About forty authors, colleagues and friends have submitted academic contributions and homages, letters, short communications and articles to portray way and work of a unique scholar in the broad area of quantitative science studies. However, mirroring the professional careers of other outstanding scientometricians of his generation, András Schubert, too, started his career in one of the established science fields. He is a skilled chemist and despite of his new commitments in scientometrics, he has remained true to his roots and succeeded in combining the two areas of activity. In both fields he has two big passions, models and measures. And as we know, good passions keep us young and zestful – as Judit Bar-Ilan tellingly expresses with a twinkle in her eye. Thus András still follows up the newest trends and challenges in scientometrics and responds to those with own models, methods and indicators. The about 25 pieces collected in this volume portray his work and impact on the community from various perspectives. We have organised the festschrift according to these perspectives. Jointly with all contributors we wish András health and energy to continue his outstanding work.

letters

The Chemistry between Ursidae and a Mellifluous Part-Time Clarinetist

ANDRÁS BRAUN & TIBOR BRAUN



Edward Bear, known to his friends as Winnie-the-Pooh, or Pooh for short, was walking through the Second District in Budapest one day, citing proudly to himself. What he was citing was a song from the Beatles and went like this:

> '*Honey pie, you are making me crazy*
> *I'm in love but I'm lazy*
> *So won't you please come home*?'

Well, he was citing this cite to himself, hoping that by citing this over and over he could increase his h-index, which, as far as Pooh was concerned, was his Honey index. Of course, if the citing failed, he could go to the Impact Factory for more.

Anyway, as he was walking, suddenly he heard something. It was a sound, a sound so mellifluous, that he thought of honey right away.

It *was* honey, he tasted honey with his ears.

He had read it somewhere that most bees buzz in the key of A, unless they are tired, when they buzz in the key of E. But it was not in the key of E, so it was not a tired bee. But not in the key of A either.

'In that case it's just an *al*most bee,' reasoned Pooh out loud.

'Strangely enough it buzzes in the key of Sch,' said Pooh, because he had absolute pitch and some more. 'And if it's in the key of Sch', he continued, 'that means my friend, Schubaa, which means, unfortunately, *not* honey, but it means, which is as good as honey (if not better),' and he cheered up very much, 'that I'll have a Parsnip Edibility Index, or,' he added, because didn't like parsnip that much, 'PHOOEY (φ-ey).' But the chance of a joint jam session with Schubaa always delighted him.

'A whole jar of jam!,' said Pooh excitedly. 'And don't forget the Erdős number for me! If I collabro…ate…if I ate a collarbone with Schubaa! An Erdős number of 4!'

And because 'erdős' means 'wooded' in english and Pooh was by this time quite lost in the Second District and just as tired as a bee buzzing in the key of E and longed very much *for* his forest, he followed the sound.

And, as he was following the sound, he made up a little song to the mellifluous melody and sang it in the key of Sch:

An actor is
said to have a Parsnip
Edibility Index φ-ey, if with φ-ey
of his/her n parsnips had at least φ-ey
joint actions each,
and with the other (n – φ-ey)
parsnips had no more than φ-ey
joint actions each.

And he traipsed through the whole Second District, and still has not found the source of the sound, because Schubaa was in Australia at the moment and Pooh heard the sound from over there, and it was soooo mellifluous that he wasn't sad at all, and just gone home, humming to himself this limerick:

There was a man from Budapest
Who could tame bears with clarinet.
He asked: 'Is it a crime
To be an arctophile?'
Not if you're septuagenarian.

Solos, Duos, Trios...

QUENTIN BURRELL
Ballabeg, Isle of Man, quentinburrell@manx.net

Dear András

Who else could have come up with a paper entitled "Jazz Discometrics"? And the idea behind it—an analogy between collaboration in scientific work and that in the world of jazz recordings—is absolutely appropriate. Thinking along the same lines, one of the great joys of listening to any musical group—whether chamber music, a concert band, a percussion ensemble, or a jazz combo—is to hear the dynamic interplay between the various individual musicians. We get striking solos, playful duos, interweaving trios and other groupings bringing in other subsets of the ensemble, each illustrating the special collaborative contribution that every individual makes although never losing sight of the whole.

The analogy with your own published work is immediate. When I looked up your list of 85 publications in scientometrics—or Information and Library Science according to Web of Science—I was struck not just by the number of classics that were there but the high degree of collaboration that was evident. There are 34 duos, 29 trios, 11 quartets and one quintet! Truly a collegial philosophy at play and at work!

I mentioned Jazz Discometrics also because it brought to mind fond memories of the 2006 STI Conference in the beautiful city of Leuven where with my wife I enjoyed some fine beer, some very good company and some excellent music! András, I look forward with pleasure to raising a glass to wish you very good health or, in Manx Gaelic, "Shoh Slaynt" on the occasion of your seventieth birthday! And welcome to the club—it is a fine age to be. Let us look ahead to a fine healthy future. And may the collaborations roll on, in whatever genre, whether musical or scientific!

András Schubert, a Life of Bibliometric Thought and Action. A Brief "Fest" Note.

KOENRAAD DEBACKERE
KU Leuven, MSI & ECOOM, Faculty of Economics and Business

András Schubert celebrates his 70th anniversary. This is a moment to reflect, to look back in order to better understand the opportunities ahead of us. A "zero order", quick analysis via Google Scholar Citations learns the following. As of February 14th 2016, the reader will find 224 publications co-authored by András listed in the Google database. Those 224 publications have attracted 9,144 citations (of which 3,979 since 2011). They lead to an h-index of 46. The Google database lists 16 frequent core co-authors (but of course, in total, there are many more). Cumulatively, those 16 core co-authors stand for 259,881 citations received (excluding the ones received by András himself). András thus is a visible member of a highly visible, well-connected community.

In short, András is one of the highly productive bibliometric researchers of our time. He is an esteemed author and co-author to many of us. He has applied bibliometrics techniques and insights to study a variety of scientific disciplines, ranging from social sciences, over agrifood, astronomy, chemistry and neurosciences, to jazz discometrics. In doing so, he has often been a pioneer. For instance, already in 1997, András studied the emerging field of nanoscience and nanotechnology and published his findings in the journal Scientometrics, together with Tibor Braun and Sándor Zsindely.

In this process, the design of subject classification schemes and the development of robust indicators (e.g. to measure and to map scientific excellence) became his trademark, often together with his long-time friend and colleague Wolfgang Glänzel. András' work offers a rich and consistent reading and insight into the various scientific and policy foundations, applications and implications of bibliometric research. It points the way to future research challenges and policy questions, advocating the need for a well-founded, multidimensional toolbox of bibliometric data and indicators in order to

capture ever better the subtleties of current scientific endeavours, their global nature, their varied outcomes and their multiple contributions to both knowledge and practice.

Current scientific activity is marked by increasing degrees of collaboration and co-authorship. András leads the way in mapping and measuring the nature and effects of collaborative research. Many of his (recent) publications offer valuable insights such as the one that methodologies developed in the context of co-authorship networks are useful for a systematic study of other complex evolving networks, such as the world wide web, Internet, or other social networks. Those insights have attracted many thousands of citations by now.

They show that András not only delves deep into the nature of scientific networks, but that he also acts as an efficient and effective scientific networker and gatekeeper. And this bodes well for the many years that are ahead. But first, enjoy your 70th birthday András, and then continue the writer's journey, both on the scientific and the literary frontier. Congratulations! All the very best!

Dear András,

BLUMA PERITZ
Professor Emerita of the Hebrew University of Jerusalem

Dear András

For 38 years now you have been one of the editors and pillars of Scientometrics along side with your own interesting research.

According to the Oxford English Dictionary an editor is one who "prepares" the work of others for publication.

András, until 120 there is a long way to go and you will have much more to "prepare".

Let me wish you on your anniversary, good health, stay alert and keep your intellectual freshness for many many years to come.

Fondly,
Bluma Peritz

articles I: models

A Researcher in the Mirror of his Models

WOLFGANG GLÄNZEL, KOENRAAD DEBACKERE, BART THIJS, SARAH HEEFFER
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)

Abstract: András Schubert is known as a versatile scientometrician who has been active in many topics within our field. His activities associated with the journal Scientometrics, his commitments for the field of quantitative science studies, for sound methodology and the correct use of the results, his methodological and applied research and exploration of data sources and measurable phenomena in scientific communication processes have yielded worldwide visibility and recognition that reaches out far beyond the scientometric community. One of his most important scholarly merits is undoubtedly his contribution to the development of models the significance of which are not limited to quantitative science studies. In the present piece we intend to discuss two of these models in the context of his work and check whether he stands the test of his models.

Introduction

Scientometricians of the first and second generation are truly multitalented scientists; among the pioneers of the field we find even polyhistors. Derek de Solla Price and Vasily V. Nalimov might just be mentioned as pars pro toto for the all-round nature of the first-generation scientometricians. Such comprehensive knowledge was necessary to found and establish a new interdisciplinary field, even more to create a new paradigm in the very focal point of philosophy, sociology and history of science, information science and economics. Their successors, skilled scientists with background in physics, chemistry, medical sciences, mathematics, sociology or other fields, still followed the tradition of acquiring the indispensable versatility in knowledge and competencies. Also András Schubert is a member of this "second" generation.

In the early 1970s András started up his career as a chemical engineer at Gödöllő University of Agricultural Sciences close to Budapest. Later he moved to the field of scientomet-

rics to join Tibor Braun's team in Budapest but he has never forgotten his scientific roots. In the early 1990s he assumed the function of the Editor in Chief of the international journal ACH–Models in Chemistry. The journal's title tellingly reflects his passion for models in science. Hence it does not surprise his detours to the world of mathematics and physics in the 1980s and around 2000.

The arts brought further enrichment to his life—and so did András to the world of music and literature. It the 1990s András founded the Medvecukor[1] Jazz Band where he performs as clarinettist. Most recently he detected his talents as a writer—not as a scientific one, since he publishes scientific literature already for decades, but as novelist. Two volumes of his children's book on the adventures of "Fuzzy Cardigan" appeared so far and the first one was already translated into German.—These are certainly new models in a multifaceted career of a person with versatile skills.

Chapter 1. András Schubert and scientometric models

In the early years of his work at the Library of the Hungarian Academy of Sciences András often said that he was an enthusiast of mathematical models, notably of those describing dynamics and evolution of population and he would be delighted to apply such models to the processes of scientific communication. The Lotka–Volterra rules based on a simple pair of first-order differential equations and used to describe the population dynamics of biological systems was one of his preferred example; the other one was related to a system of deterministic and probabilistic models that describe birth and death processes with interaction of the environment such as immigration and emigration. His dream came true and jointly with his colleagues András Telcs and Wolfgang Glänzel he elaborated and implemented some mathematical models that were suited to describe publication and citation processes in scientometrics. Here we will mention two of the most important models that are related to distribution theory (a third one jointly constructed with Albert L. Barabási and collaborators is dealing with complex evolving networks).

The first model is based on a simple semi-probabilistic birth model that assumes interaction with the environment. This model, which aims at describing the changing distribution of publications by authors in time, assumes an open population, that is, new members can join and "retired" members might leave the system (Schubert and Glänzel, 1984). This model is furthermore designed to provide a realistic picture of publication processes, where authors stepwise move from one publication status to the next one and the solution of the underlying system of first-order, linear differential equations is a stochastic process with real-valued time parameter. The process has, under certain conditions, a stationary limiting distribution, if time tends to infinity, and this solution is, in particular, a *Waring distribution*. At the same time, the model describes the dynamics of the population and its publication output. Interestingly, the

[1] Medvecukor (Hungarian) = liquorice

model was not only used in scientometrics but also applied in other research fields (Boxenbaum et al., 1987), and was recognised even in mathematics (Panaretos and Xekalaki, 1986; Xekalaki et al., 1987). Having now a model was fine, but the question arose whether the Waring distribution and related models do also occur in real life of scientometrics. Moving straight to mathematics was therefore a logical consequence: A characterisation theorem for the family of generalised Waring distributions proven by Glänzel et al. (1984) provided a statistical test and a plethora of application possibilities in scientometrics but also beyond the field. The method could, among others, be applied to the estimation of censored or unknown data (Schubert and Telcs, 1986), and the statistical analysis of literary vocabulary, that is, to word-frequency statistics in poetry, fiction and essayistic literature (Telcs et al., 1985).

The second model refers to citation frequency. The underlying idea was to create a reduction of citations distributions to a limited set of standard reference classes for the purpose of comparison, benchmarking and normalisation. Although the proposed method provides a parameter-free solution, it has remarkable properties for Paretian distributions, that is, distributions that asymptotically follow a power law, as could be derived from the above mentioned characterisation theorem. That is the reason why the method was called "Characteristic Scales and Scores" (CSS).

In what follows we will apply these models to characterise András Schubert's work and check in how far he "obeys his own rules".

Chapter 2. András Schubert in the mirror of his models

2.1 Schubert's scientific vocabulary

The first model will be applied to András Schubert's scientific vocabulary. Unlike in the study by Telcs et al. mentioned in the previous chapter, this time scientific text corpora are studied. Of course, we expect differences between vocabularies and word usage in scientific and literary texts. In scientific text technical terms are among the most frequent terms and, on the other hand, fillers, embolalia are assumed to be less common. The reason lies in the efforts for efficiency, particularly, in the intension to compress as much information as possible into possibly short texts. This has two causes, one is part of the nature of scholarly communication, the other one might be a consequence of the often experienced space limitation in scientific journals. We just mention in passing that the second cause will probably gradually disappear due to the electronic communication that is currently becoming prevalent. Furthermore, vocabulary of the scientific text comprises technical abbreviations, acronyms, formulas and other artificial language constructs as well as specific terms that are, otherwise, not commonly used in poetry and fiction. In the sciences we might therefore expect more skewed and polarised word-frequency distributions than in literary text along with specific vocabularies where technical terms play a determinative part.

Vocabulary characteristics vs. lexical similarity

The statistical analysis of the text is based on the determination of the vocabulary. This is done in a semi-automated way. The algorithms are similar to those used in text mining for cognitive document clustering. In both tasks all terms of a document are first extracted and stemmed. For this purpose, versions of the Porter-stemmer (Porter, 1980) are commonly used. Once a first raw vocabulary has been compiled, subsets of this vocabulary are treated in a different manner depending on the actual application, namely the statistical analysis of text corpora and cognitive clustering based on lexical document similarity, respectively. In this context we have mainly to distinguish between the partially different functions of stop-words, homonyms, synonyms, acronyms and various types of names. While common words, which are frequently used, form an essential part of a vocabulary and the use of certain common words can even be considered characteristic for a person's style, these high-frequency common terms are, on the other hand, considered noise in calculating document similarity since, from the cognitive viewpoint, those do not bear relevant information. Therefore we have not removed frequent common words, except for the articles 'the' and 'a(n)', which form a unit with the subsequent noun and are consequently not considered independent words. Homonyms are always problematic and need to be resolved manually. In our case, 'small' (adjective) and "Small" (Henry) is a typical example. This applies to both lexical applications. Synonyms are different again. While synonyms cover the same information, they need to be resolved in similarity-based clustering. In statistical vocabulary analysis they are, however, regarded as enrichment and their use as typical of an individual's style. Consequently, we left synonyms unresolved. The treatment of acronyms and person names might be subject to discussion. We decided to remove most acronyms and person names, except for eponymic use. Thus, for instance, Zipfian and Waring have been kept, Braun and Glänzel were removed.

Data and methods

For this exercise we have collected 18 articles by András Schubert, which have been published in three periods of András' career: the early 1980s, the mid-nineties and the most recent period around 2010. All selected papers were research articles mostly dealing with new methods and their application; case studies were, however, not taken into consideration. Earlier papers had to be OCR processed, spell-checking has been applied and remaining typos have been removed manually. Also bibliographic information and all references have been removed from all documents as being foreign sources. The 18 articles published between 1983–1985, 1993–1998 and 2010–2013, respectively, are listed in the Appendix.

The methodology of statistical text analysis is described in Telcs et al. (1985). Similarly to Herdan (1964), a Waring model is used but in the more recent paper a weighted regression analysis is applied. The challenge of modelling word-frequency distributions is that the distribution is truncated at point 1. The reason is that the complete

vocabulary of a writer and thus the share of unused words, which is otherwise part of the writer's vocabulary, is unknown. In the case of the Waring distribution this issue can readily be resolved. This is straightforward from the definition of the distribution as can be seen in the following.

We say that a random variable X has a Waring distribution, if

$$P(X=k)=\frac{\alpha}{N+\alpha}\cdot\frac{N\ldots(N+k-1)}{(N+\alpha)\ldots(N+\alpha+k)} \tag{1}$$

where N and α are positive real parameters. Consequently, we have

$$P(X=k\,|\,k>0)=\frac{P(X=k)}{1-P(X=0)}=\frac{\alpha}{N+\alpha+1}\cdot\frac{(N+1)\ldots(N+k-1)}{(N+\alpha+2)\ldots(N+\alpha+k)} \tag{1*}$$

and, if we shift this distribution back to 0, we obtain a Waring distribution again, this time with parameters (N+1) and α.

Telcs et al. analysed four subsamples of an essays by Thomas Babington on Francis Bacon and the word frequency in Alexander Pushkin's story *The Captain's Daughter* using the newly elaborated test. Previously, Herdan (1964) has used a more conservative method to estimate the parameters of the word frequency in the Pushkin text according to the Waring model. The α values of both samples considerably differed: For the essay Telcs et al. obtained about 2.65, while for the Pushkin text the estimate was 1.33. Interestingly, the word-frequency distribution of Pushkin's text does not belong to the domain of the attraction of the normal distribution. Yet, another issue emerged in this context: The tail of the distribution behaved irregularly so that the hypothesis of the Waring model had to be rejected for this sample. There was also a further difference between the two vocabularies: While the essay data were based on nouns only, the Pushkin sample referred to the complete text. The authors came to the conclusion that this might be one source of the difference in the nature of the two samples.

For the following exercise, we decided to use all words (except the cases discussed above) and also to apply a conservative method to estimate the parameters of the Waring distribution. This method uses the first moment and the frequency of words that were used only once because these are the most robust statistics of a Waring distribution that is truncated at point 0. Although the fit was expectedly not perfect, we will discuss the results more in detail in the following subsection.

The evolution

The application of the three periods offers two important options. On one hand, this allows to monitor the evolution of the characteristics of Schubert's vocabulary and, secondly, it also permits to trace the major changes in his vocabulary. To begin with, we found parameters that are surprisingly similar to the Pushkin data. From the robust estimation formula

$$\alpha=\frac{f(x-1)}{fx-1}\quad\text{and}\quad N=\frac{x(1-f)-f(x-1)}{fx-1} \tag{2}$$

we obtain the estimated parameter pairs that are presented in Table 1. Both empirical and estimated data are plotted in Figure 1. Observations greater than 25 are aggregated.

Table 1 Waring parameters estimated from 18 Schubert papers published in three different periods

Parameters	1983–1985	1993–1998	2010–2013
α	1.36	1.28	1.38
N	1.19	0.91	1.07
Word count	1444	1524	1329

Also Schubert's data reveal the non-Gaussian nature of his vocabulary. α ranges between 1.3 and 1.4. Furthermore, *N* is about 1.0, that is, the distribution is close to a Yule distribution. And both parameters proved quite stable over time. This implies that the word-frequency distribution in Schubert's work has practically not changed. The closeness of the parameters to those of the Pushkin data is particularly striking because the nature of the two text corpora was assumed to be completely different: On one hand we have a very focussed scientific text and, on the other hand, a literary, poetic text with all the richness that language can provide. The average word use amounts to roughly 7 in both cases.

After we have seen that the statistical characteristics of Schubert's vocabulary has not changed during three decades, independently of the topics that were dealt with, we would also like to see if the vocabulary as such has changed. In order to do so, we have looked at structural changes. In particular, we have calculated the occurrence of the words, that were most frequently used, for each period separately. We have applied the coefficient of variation (CV) to their distribution over the three periods. The thresholds were 25 for the total frequency and 1.25 for the CV. We have only removed "Inorganica Chimica Acta" because this journal was the subject of the analysis (see article [11] in the Appendix). The results are shown in Table 2. Data are grouped by trends and ranked in decreasing order by their CV values. In principle, four scenarios are possible, 1) a decrease of frequency (D), 2) the opposite trend (E), 3) a maximum or peak in the second period (P) and 4) and the opposite, a minimum or valley (V) in the second period. The fourth case was not observed.

Of course, the selection of papers—and thus of the topics these papers are dealing with—was arbitrary, even if we aimed at balance. The topic choice has clear and measurable effect on the vocabulary as well. Nevertheless the patterns we have found, are in line with our expectations. The first period was the time András devoted to a great extent to the elaboration and testing of models (but partially also to the development of scientometric indicators). The corresponding terms can be found in the first section of Table 2; distribution-related issues (paper [2], [4] and [6]) and matrix decomposition (paper [3]) are here in the foreground. These topics have become less relevant for András' work in later phases of his academic career. The second period, in the 1990s, is characterised by bibliometric profiling, indicator research and the evolution of the field of scientometrics (paper [8] and [9]). These topics, notably the latter one, are not adequately reflected

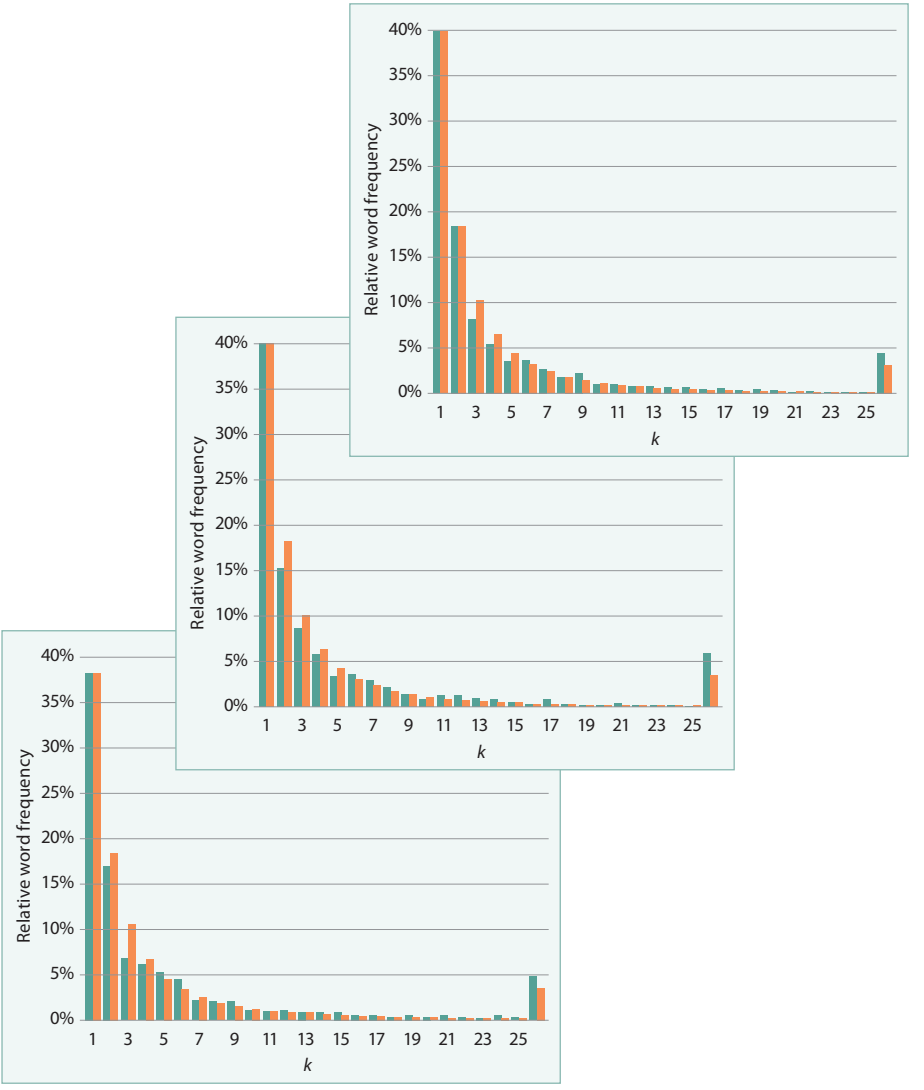


Figure 1 Empirical (teal) and estimated (orange) word frequency of selected Schubert texts, arranged in three periods: 1983–1985 (left), 1993–1998 (centre), 2010–2013 (top)

by a distinguished vocabulary that is, otherwise, not used in other contexts. Some of the terms, that are frequently used in this period only, are listed in group (P) in Table 2. Most recently, András has increasingly focused on network-related models and indicators. This is reflected by the corresponding (and partially specific) terms in the second group (E). To sum up, we can conclude that the statistical characteristics of the use of his vocabulary has not changed over time, while the vocabulary as such did.

Table 2 Changes in Schubert's vocabulary (D: decrease, E: increase, P: peak in period 2)

Term	Occurrence	Trend	CV
estim	28	D	1.732
meet	41	D	1.669
book	63	D	1.610
random	27	D	1.544
clinic	34	D	1.509
waring	33	D	1.502
medicin	29	D	1.389
scientist	36	D	1.387
neg	28	D	1.378
matrix	25	D	1.338
partnership	33	E	1.732
cluster	62	E	1.690
network	33	E	1.577
core	25	E	1.529
categor	82	E	1.370
similar	131	E	1.317
assess	41	P	1.480
subfield	35	P	1.445
period	72	P	1.381

2.2 Schubert's work and the community

The aim of this section is to apply the models introduced in the previous chapter to capture András' impact on the scientific community. Before we discuss the second model in this context, we shortly apply the already mentioned characterisation theorem by Glänzel et al. (1984) to model the citation distribution of his research work. The underlying distribution family builds upon Irwin's distribution family (cf., Irwin, 1975), which is also referred to as the Generalised Waring distribution. We say that a random variable X has an Irwin distribution with positive real parameters a, b and c, if

$$P(X=k)=P(X=k-1)\cdot\frac{(a-1)k^2+k(b+c-(a-1))-c}{((a-1)k+a+b+c)} \tag{3}$$

In order to estimate the parameters of Schubert's citation distribution, we first collected the empirical sample. The sample comprises all 'citable papers' indexed in Thomson Reuters Web of Science Core Collection (WoS) beginning with his Gödöllő era till the present (1972–2015). Citations have been counted from the date of publication till the present. The corresponding genesis of the model could be considered to be similar to the dynamic Waring process since an open population could be assumed in this case

as well. New papers are published and thus entering the system, obsolete ones are "retired" and leave the system. Yet, publication and citation processes are subject to different mechanisms. One of the most important limitations to publications processes are of physical nature, writing new papers is a function of the author's natural capacity, while there is almost no limit for receiving citations. Sheldrick's paper on the history of SHELX with almost 46,000 citation since its publication eight years ago might just serve as an example of the open-ended citation scale. Consequently, the Waring model is assumed to fail in the case of dynamic citation processes. We just mention in passing that for closed populations and static citation windows the negative binomial process proved a good model (Glänzel and Schubert, 1995; Mingers and Burrell, 2006). Therefore, we do not attempt to fit a negative binomial or Waring distribution but will apply the characterisation theorem for Irwin-type distributions to search the solution within this broad distribution family. According to this theorem, a non-negative integer-valued random variable X has a distribution belonging to Irwin's system if and only if the following equations holds for all $k \geq 0$.

$$\mathrm{E}\left(X \mid X \geq k\right) = a \cdot k + b + c \cdot k \cdot \mathrm{E}\left(\left(X+1\right)^{-1} \mid X \geq k\right) \tag{4}$$

where $\mathrm{E}(A|B)$ denotes the expectation of A under the condition B and a, b and c are the parameters as displayed in Eq (4). If we denote the corresponding empirical values by $e(k)$ for the series of k-truncated first moments on the left-hand side and $d(k)$ for the series of k-truncated first negative moments, respectively, we can rewrite the above theoretical equation (4) as the following empirical characterisation formula.

$$\left(e\left(k\right) - b\right) / k = a + c \cdot d\left(k\right) \tag{5}$$

Furthermore, this formula can be used to estimate the parameters of the distribution. Note that b is the mean value and as such an unbiased estimator for the expected value. If the sample is large enough. b can be substituted into Eq (5) and the exercise reduces to a simple linear regression analysis in order to determine the parameters a and c. The distributions of Irwin's family can be best visualised on the (a, c) parameter plane for any arbitrary but fixed parameter b. This is very convenient because b is the estimator for the expectation and can in most cases be regarded as known. The domains are then bounded by graphs of linear and parabolic functions. The horizontal and vertical axis at $c = 0$ and $a = 1$ mark the domain of the Waring and the negative binomial/binomial/Poisson distribution, respectively (see Figure 3).

Out of the 180 papers we could retrieve from the database, we selected the 144 "citable" items (i.e., articles, proceedings papers, letters, notes and reviews). The earliest paper was published in 1972, the most recent one appeared in 2015. The most cited paper co-authored by András was published in 2002 and received 801 citations in total. Figure 2 shows the plot of $(e(k)—b)/k$ vs. $c \cdot d(k)$ for $k < 25$. Beyond this value, frequencies are low and the sizes of the underlying truncated samples are too small to provide reliable statistics. The correlation coefficient and the constant in the linear regression model then immediately provide the two missing parameters. In particular, we obtain the following three parameters for the Irwin or Generalised Waring dis-

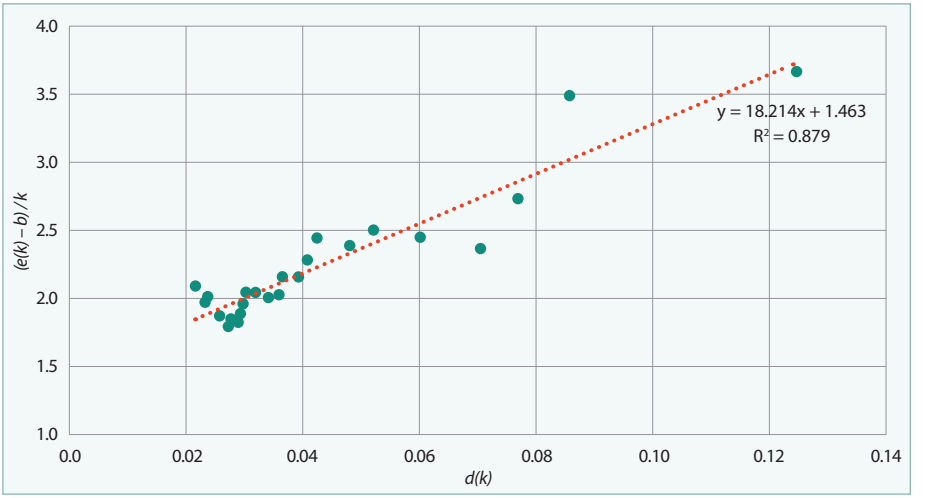


Figure 2 Plot of truncated moments for characterisation and parameter estimation of Schubert's citation impact

tribution: a = 1.463, b = 29.375 and c = 18.214. These parameters suggest an inverse Pólya-Eggenberger model and show that the particular distribution is far from the Waring model because of the large c parameter value (see Figure 3).

András Schubert's citation distribution is very flat, not very skewed and heavy tailed (see Figure 4). It is quite impossible to test any fit to such data. Note that one third of his publications is cited at least 25 times each and the extremely long tail is a sequence of 0 and 1 frequencies. Under these circumstances, the fit of the inverse Pólya-Eggenberger distribution is quite acceptable. Despite of the heavy tail, the tail parameter $\alpha = a/(a-1) = 3.16$ indicates that the distribution is of Gaussian type.

Now we can take the next step and apply the second model, the method of Characteristic Scores and Scales, to his citation distribution. Characteristic scores are obtained from iteratively truncating samples at their mean value and recalculating the mean of the truncated sample. The procedure is repeated until a given number k of scores is reached. Usually three scores are sufficient, where the first one is identical with the mean value of the original sample. The resulting four classes are obtained by the intervals defined by adjoining scores (b_i with b_0:=0 by definition). This way we obtain the following four citation classes ($CC_i = [b_{i-1}, b_i)$): $CC_1 = [b_0, b_1)$ is the class of 'poorly cited' papers, $CC_2 = [b_1, b_2)$ contains 'fairly cited' papers, $CC_3 = [b_2, b_3)$ contains 'remarkably cited' papers and $CC_4 = [b_3, \infty)$, putting $b_4 = \infty$, is the class of 'outstandingly cited' papers. The values k=2 and k=3 together are also used to identify *highly cited* papers. The model has four important advantages. 1. CSS is not biased by ties in the underlying citation ranking, 2. CSS scores are self-adjusting and thus not defined on arbitrary pre-set values, 3. In comparative analysis CSS can be calculated for each sample or even subpopulation separately, and 4. CSS provides robust classes in terms of their insensitivity to publication year, citation windows and subject. Although CSS is not directly

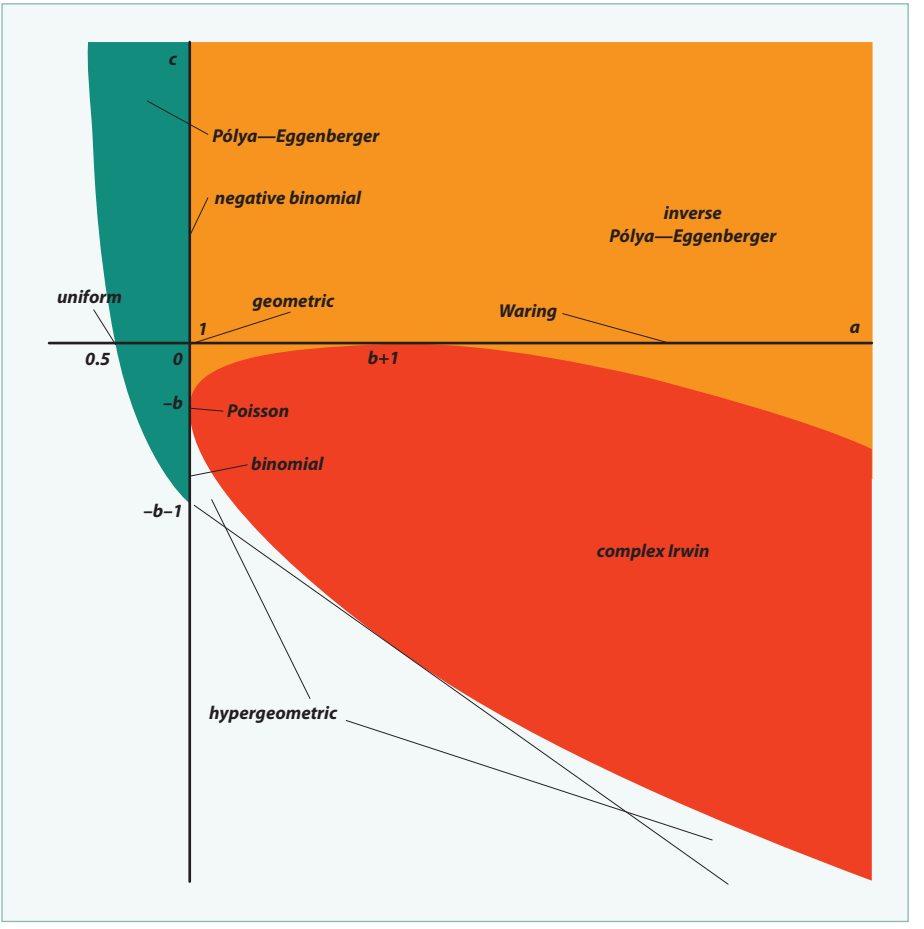


Figure 3 The domains of Irwin's family on the (a, c) parameter plane at fixed b

linked to percentiles, the distribution of papers over classes in general is about 70% (1), 21% (2), 6%–7% (3) and 2%–3% (4). This is a consequence of the Paretian nature of citation distributions (cf. Glänzel, 2007). The deviation of the profile under study from a given reference standard provides a multifaceted picture of citation impact.

We have calculated CSS classes on the basis of Schubert's citation distribution and, as kind of benchmark, those derived from the citation distribution of all 4104 citable papers published in the journal *Scientometrics* from the first volume in 1978 till the most recent one. Note that Schubert's oeuvre does not form a subsample of the journal's paper set although there is a considerable overlap between the two samples: Although András has also published in other fields than bibliometrics, above all in chemistry and mathematics, the journal might be considered a proper baseline since more than 60% of his papers was published in Scientometrics. Table 3 gives the scores and class distributions

as well as the cross-benchmarking of the two paper sets. All properties of the model become immediately apparent. The b_i threshold values for the Schubert sample displayed in the second column and the corresponding values for the journal in the fourth column impressively illustrates that András plays in a completely different league. The fact that the distribution of papers over citation classes is similar in the two samples again militates for the robustness of the method. The most revealing results are obtained, if both scales are gauged against each other: Column 6 in Table 3 gives the share of András Schubert's according to the Scientometrics standard, for instance, only about half of his papers are poorly cited according to the SCIM standard, while ¾ of his work are poorly cited according to his own rules. Similarly, almost one fifth of this papers can qualify as highly cited (classes 3 and 4) but according to his own scale only 7% can be considered as being of this type. Conversely, 90% of Scientometrics papers are poorly cited in the Schubert scale and less 2% of the journal's publications are highly cited here (cf. column 7). This is admittedly a statistical exercise that assumes similar publication dynamics in both samples: the flexible citation window prevents us from identifying individual papers on the basis of these scales; most of the poorly cited papers will be found among the most recent ones and highly cited papers are probably older. From the statistical viewpoint this does, of course, not affect the validity of the above discussion. Thus we can conclude that also this model can successfully be applied to András Schubert's published work and that his own rules are in line with the assumptions and properties of the models—but do, perhaps, represent a somewhat different standard.

Table 3 András Schubert's (AS) citation impact in the light of the CSS model and gauged against the Scientometrics (SCIM) standard

i	András Schubert		Scientometrics		AS vs. SCIM	SCIM vs. AS
	b_i	CC_i	b_i	CC_i		
1	0.0	74.3%	0.0	72.5%	50.7%	90.4%
2	29.4	18.8%	12.5	19.9%	30.6%	7.9%
3	88.4	4.9%	35.1	5.4%	9.7%	1.4%
4	205.6	2.1%	74.0	2.2%	9.0%	0.2%

Conclusion

There is one important conclusion that can be drawn from this admittedly somewhat irregular study: Do not create scientometric models if you do not wish that those will be applied to you. But the journey is not yet finished. Recently Glänzel and Schubert (2016) have published a book chapter on the statistical models of bibliometric distributions elaborated by the authors in the 1980s, and on their various relations and perspectives. The chapter bears the subtitle "A Success-Breeds-Success Story". Indeed, this is what we are talking about.

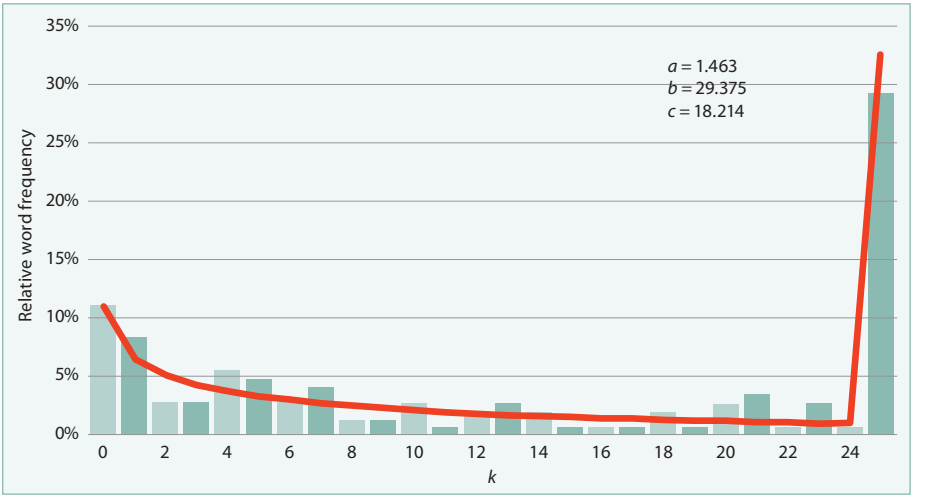


Figure 4 Empirical (bars) and estimated (line) citation frequencies of András Schubert's 144 citable papers (1972–2015) indexed in the WoS according to the inverse Pólya-Eggenberger model

Acknowledgement

Figure 3 is recreated from Glänzel (2009) with permission of the publisher.

References

Boxenbaum, H., Pivinski, F., Ruberg, S.J. (1987), Publication rates of pharmaceutical scientists—application of the Waring distribution. *Drug Metabolism Reviews*, 18 (4), 553–571.

Mingers, J., Burrell, Q.L. (2006), Modeling citation behavior in Management Science journals. *Information Processing & Management*, 42 (6), 1451–1464

Glänzel, W., Telcs, A., Schubert, A. (1984), Characterization by truncated moments and its application to Pearson-type distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66, 173–183. (Correction: *Probability Theory and Related Fields*, 74, 1987, 317.)

Glänzel, W., Schubert, A. (1988), Characteristic Scores and Scales in assessing citation impact. *Journal of Information Science*, 14, 123–127.

Glänzel, W. (1994), *IrWin—A characterization tool for discrete distributions under Windows*. In: R. Dutter, W. Grossmann (Eds.): Short Communications in Computational Statistics, Proceedings of COMPSTAT '94, Vienna, 199–200.

Glänzel, W., Schubert, A. (1995), Predictive aspects of a stochastic model for citation processes. *Information Processing & Management*, 31 (1), 69–80.

Glänzel, W. (2007), Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, 1 (1), 92–102.

Glänzel, W. (2009), The Multi-Dimensionality of Journal Impact. *Scientometrics*, 78 (2), 355–374.
Glänzel, W., Schubert, A., Thijs, B., Debackere, K. (2009), Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, 78 (1), 165–188.
Glänzel, W., Schubert, A. (2016), *From Matthew to Hirsch: A Success-Breeds-Success Story*. In: C.R. Sugimoto (Ed.), Theories of Informetrics and Scholarly Communication, De Gruyter Mouton, forthcoming.
Herdan, G. (1964), *Quantitative linguistics*. London, Butterworths
Irwin, J.O. (1975), The generalized Waring distribution. Part I, II, III. *Journal of the Royal Statistical Society A*, 138, 18–31, 204–227, 374–384.
Panaretos, J., Xekalaki, E. (1986), The Stuttering Generalized Waring Distribution, *Statistics & Probability Letters*, 4 (6), 313–318.
Porter, M.F. (1980), An algorithm for suffix stripping. *Program*, 14(3), 130-137.
Schubert, A., Glänzel, W. (1984), A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, 6 (3), 149–167.
Schubert, A., Telcs, A. (1986), Publication potential—an indicator of scientific strength for cross-national comparisons. *Scientometrics*, 9 (5-6), 231–238.
Telcs, A., Glänzel, W., Schubert, A. (1985), Characterization and statistical test using truncated expectations for a class of skew distributions. *Mathematical Social Sciences*, 10, 169–178.
Xekalaki, E., Panaretos, J., Philippou, A. (1987), On some mixtures of distributions of order *k*. *Fibonacci Quarterly*, 25 (2), 151–160.

Appendix

1) Articles published in 1983–1985

Schubert, A., Zsindely, S., Telcs, A. et al. Quantitative-analysis of a visible tip of the peer-review iceberg—Book Reviews in chemistry. *Scientometrics*, 6 (6), 1984, 433-443.
Glänzel, W., Schubert, A., Price distribution. An exact formulation of Price's Square Root Law. *Scientometrics*, 7 (3-6), 1985, 211–219.
Schubert, A., Zsindely, S., Braun, T., Scientometric analysis of attendance at international scientific meetings. *Scientometrics*, 5 (3), 1983, 177-187.
Schubert, A., Glänzel, W., Statistical reliability of comparisons based 1983 on the citation impact of scientific publications. *Scientometrics*, 5 (1), 1983, 59–74.
Schubert, A., Zsindely, S., Braun, T., Scientometric indicators for evaluating medical-research output of mid-size countries. *Scientometrics*, 7 (3-6), 1985, 155-163.
Schubert, A., Glänzel, W., A dynamic look at a class of skew distributions—a model with scientometric applications. *Scientometrics*, 6 (3), 1984, 149-167.

2) Articles published in 1993–1998

Schubert, A., The profile of the Chemical Engineering Journal and Biochemical Engineering Journal as reflected in its publications, references and citations, 1983-1996. *Chemical Engineering Journal*, 69 (3), 1998, 151-156.

Schubert, A., Maczelka, H., Cognitive changes in scientometrics during the 1980s, as reflected by the reference patterns of its core journal. *Social Studies of Science*, 23 (3), 1993, 571-581.

Schubert, A., Little Scientometrics, Big Scientometrics—and beyond. *Scientometrics*, 30 (2-3), 1994, 411-413.

Schubert, A., Braun, T., Cross-field normalization of scientometric indicators. *Scientometrics*, 36 (3), 1996, 311-324.

Schubert, A. P., Schubert, G. A., Inorganica Chimica Acta: its publications, references and citations. An update for 1995-1996. *Inorganica Chimica Acta*, 266 (2), 1997, 125-133.

Schubert, A., Braun, T., Reference-standards for citation based assessments. *Scientometrics*, 26 (1), 1993, 21-35.

3) Articles published in 2010–2013

Schubert, A., Jazz discometrics—A network approach. *Journal of Informetrics*, 6 (4), 2012, 480-484.

Schubert, A., A Hirsch-type index of co-author partnership ability. *Scientometrics*, 91 (1), 2012, 303-308.

Schubert, A., X-centage: a Hirsch-inspired indicator for distributions of percentage-valued variables and its use for measuring heterodisciplinarity. *Scientometrics*, 102 (1), 2015, 307-332.

Schubert, A., A reference-based Hirschian similarity measure for journals. *Scientometrics*, 84 (1), 2010, 133-147.

Schubert, A., Soos, S., Mapping of science journals based on h-similarity. *Scientometrics*, 83 (2), 2010, 589-600.

Schubert, A., Measuring the similarity between the reference and citation distributions of journals. *Scientometrics*, 96 (1), 2013, 305-313.

The Matthew Effect of Science, the Bible and András Schubert

ANDREA SCHARNHORST
Royal Netherlands Academy of Arts and Sciences,
Data Archiving and Networked Services (DANS), Amsterdam, Netherlands

When I wrote down the title a thought crossed my mind. Could it be that information scientists—just for a little—might have an inclination to feel as being gods?

Richard Smiraglia, editor in chief of the journal *Knowledge Organisation*, another family tribe in information science, not always still known to the scientometricians, once explained to me that it was not by accident what is in class 0 of the UDC. Paul Otlet started the Universal Decimal Classification with SCIENCE AND KNOWLEDGE. ORGANIZATION. COMPUTER SCIENCE. INFORMATION. DOCUMENTATION. LIBRARIANSHIP. INSTITUTIONS. PUBLICATIONS (class 0), before turning to PHILOSOPHY. PSYCHOLOGY (class 1), RELIGION. THEOLOGY (class 2), and so down towards the very tangible concreteness of GEOGRAPHY. BIOGRAPHY. HISTORY (class 9). Indeed there is an ordering principle behind this sequence. Olet started with the fundamentals, if not given by a higher authority. This is the place where information science is situated—class 0. András Schubert, as I know him, is far too modest to claim a place among the gods of a scientific Olympus himself, and still all the other phrases in the title are there with a perfect good right. And even if András is known for his modesty, for me, when I first tipped my toe into the deep waters of scientometrics, guided by Jan Vlachý, Hildrun Kretschmer and last but not least Manfred Bonitz, András was among those 'gods' which traces I aspired to follow.

András Schubert—as part of the iconic *Budapest group*—has written about many different topics relevant for the quantitative studies of science. Among them, when looking at (ScholarGoogle!) citations, one stands out. It is the paper from 2002 on "Evolution of the social network of scientific collaborations", with Barabási and others. One of the most important functions of this paper in my view was that it alert-

ed the statistical physicists which just had collected themselves under the new flag of *complex networks* and were now foraging for data for the new Network Science to the existence of a body of knowledge they might want also to take into account—namely scientometrics. With this paper András was among those building a bridge between new concepts in the natural sciences and information science.

But for me it did not need the 2002 paper to learn this. I had long recognized András as one of those linking natural sciences and information science. In my encounter with his work three things stand out: *models, maps* and the *Matthew Effect of Science*. Let me explain why and how.

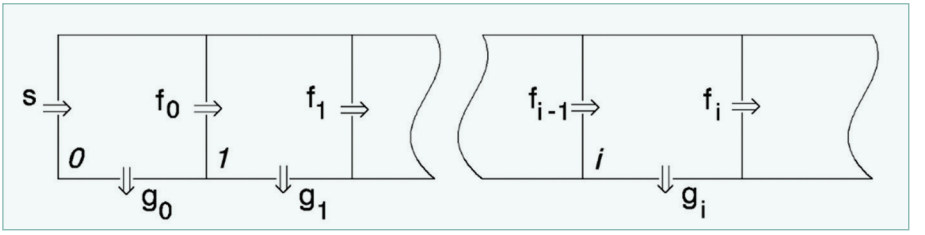


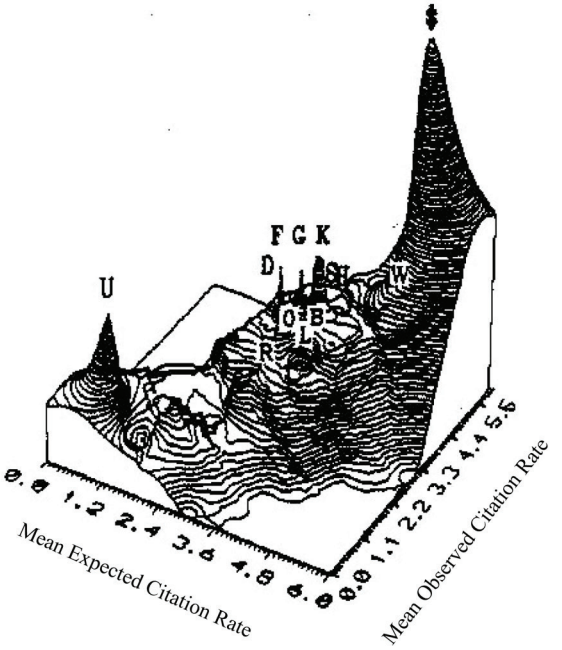
Figure 1: Scheme of substance flow according to Schubert and Glänzel (1984)

The position of mathematical *models* inside of scientometrics is still an ambivalent one. Some pioneers followed by parts of the community have embrace theories of stochastic processes which can be used to explore and understand the statistical laws we know as bibliometric laws—Lotka's law of productivity, Bradford's law of scattering and the Cumulative advantage distribution of Price. Few have made the step to predictive models. András Schubert is among those, and the early model of Wolfgang Glänzel and András Schubert about skew distributions caught my eye for various reasons when I started reading *Scientometrics*. First, it looked at scientists as unit of modelling while most of the other studies dealt with citations. And our own model was about the growth of fields measured in terms of scientists too. Second, it spoke of *self-reproduction* and non-linear processes, and I just had fledged the nest of physics of self-organisation. The paper proposes a model to explain the growth of publications in a scientific community. At its core we find a flow between cells, an urn model (see Figure 1). When entering the science system with a first publication, a scientist finds her/himself in cell 0 and which each new publication moves one cell up to the right. While this model has been taken up in later publications, its original presentation 1984 is still a good read! Some papers are like wine—they only get better!

Having recognized András as *Bruder im Geist*, the next piece of his work which has accompanied me for a long time, and still is, is actually a visualization. It is not so much as *map* as a 3D *landscape*. I wrote at this time about complex mathematical models describing the dynamics of systems in an attribute space. All very abstract. But, the core metaphor was that of a landscape in which for instance researchers, whole scientific communities, institutions, or even countries would position themselves and develop their research profile according to an unknown fitness landscape.

Examples of mapping science and technology:
III Landscapes of scientometric output indicators

Landscape of observed vs. expected citation rates, 1990-1994



Data source are the bibliometric datafiles based on the Science Citation Index. The third dimension is given by the cubic root of the total number of publications in all fields combined.
$: USA, B: BEL, C: CAN, D: DNK, F: FRA, G: DEU, H: NDL, K: UKD, L: ISR, O: FIN, R: IRL, S: SWE, U: SUN, W: CHE
Graphical representation after *Braun, Schubert, Scientometrics 38(1998)175*

Figure 2: Landscape of observed vs. expected citation rates (1990–1994) according to Schubert and Braun (1997)

I was sourcing the literature of bibliometrics maps, only to find networks or other 2D representation, until I found the paper of Braun and Schubert in 1998. The graphics is nothing more than an extension of the so-called relational charts! The coordinates of the countries are given by their MECR and MOCR, and the third dimension is given by their size in terms of total output of publications (Figure 2). The former Soviet Union and the United States seem to form the antipodes in a landscape which co-ordinates represent values of publication venues (expected citation rates) and actual received reward (observed citation rate). The introduction of *Expected Citations* was

as simple as ingenious and opened a whole new way to look at the distribution of citations across countries in the whole Web of Science database, in specific fields, and in certain journals. It enabled to empirical test Merton's postulated *Matthew effect in science*. Priority for the bibliometrical empirics around the Matthew effect in science belongs to András and his colleagues. But for many years my old friend and colleague Manfred Bonitz enjoyed himself, delving into investigations around the Matthew effect in science, all based on the concept to compare Expected and Observed Citations. Manfred's work after 1989 (and my own collaboration with him on this matter) would not have been possible, had not the Budapest colleagues so generously shared data with us. In 1989 they published an aggregated data set about countries and journals in a large volume, which Manfred only called 'the bible' (Schubert et al., 1989). Data is seen as the New Gold of research nowadays, and data reuse is healed as the new imperative in science. Data for sure is also at the heart of bibliometrics, and still data sharing is not a culture maintained in the community. What András and his colleagues did in 1989/1997 was not only unprecedented, it has also not been succeeded. I think it was both revolutionary as well as visionary. As webometrics, altmetrics and other new kids on the block of scientometrics show—the future of bibliometrics is in open data. I don't remember if I ever asked András, Wolfgang or Tibor what triggered this publication—but doing this is in line with curiosity, pushing boundaries and generosity. Those features seem to be so characteristic in both András work as personality.

Having said this what remains is a toast with a glass of good wine to the sound of good music to the Jubilee!

Ackowledgement

Figures 1 and 2 are reproduced from Schubert and Glänzel (1984) and Braun and Schubert (1997), respectively, with permission of the publisher. Figure 2 is redrawn by the author as it was reproduced in Scharnhorst (2000).

References

Schubert, A., Glänzel, W. (1984). A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, 6(3), 149–167. doi:10.1007/BF02016759

Schubert, A., Glänzel, 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. doi:10.1007/BF02093234

Braun, T. Schubert, A. (1997). Dimensions of scientometric indicator datafiles—World science in 1990-1994. *Scientometrics*, 38(1), 175–204. doi:10.1007/bf02461130

Scharnhorst, A. (2000). *Evolution in Adaptive Landscapes — Examples of Science and Technology Development*. WZB Discussion Paper FS II 00-302.

From Metrics to Modeling

ANDRÁS TELCS[1,2,3]
[1] *Wigner Institute for Physics, Konkoly Thege M. ut 29--33,1112 Budapest Hungary,*
[2] *Budapest University of Technology and Economics*
[3] *University of Pannonia*
telcs.szit.bme@gmail.com

Introduction

The best boss and the best friend, that's András for me. He was the boss of the little scientometric unit at the Library Hungarian Academy of Sciences when I applied to the junior position opened there. At that time scientometric was an almost new field. Two Hungarian scientists András and Tibor Braun were among the founders and pioneers. They were the scientometers and we, the juniors, Wolfgang Glänzel and me the scientomilimeters helping them with our mathematical, statistical background. We had to learn a scientific approach very different from mathematics, learn data acquisition, handling, cleaning and making meaningful statistics and in parallel learn to use computers (mainframe at that time) as well as writing papers. András was a brilliant mentor, guide and teacher in all of that.

We were in the era when scientometrics was a kind of quantitative sociological field which describes publication activity of scientists. Data were difficult to collect. The field started to grow out from the early descriptive phase. That was the challenge and inspiration for the chemical engineer, András, to step on a higher level. His ambition was to build mathematical models which properly describe the macroscopic picture of scientific communication based on the microscopic patterns of publication, reference and citation usage. The first and memorable piece of this kind was the paper by András and Wolfgang on the modeling of publication activity of scientists based on the cumulative advantage effect [1]. In some respect this paper is the forrunner of Barabasi's celebrated works [2],[3]. The preferential attachment and the cumulative advantage both has a linear contribution to the acquisition rate let it be a new connection or a new publication. András' idea was ground breaking. And several further unique ideas came, hard to enumerate them. The author left the field of scientometics, hence he is not expert of it, so he does not

attempt to survey András's numerous scientific results. After a longer pause we did some work together, we combined the network view and novel scientometric idea, the Hirsh index. The collaboration was fruitful (c.f. [6],[7]) and delightful as in the good old times.

While András concentrated on modeling, the mainstream followed a less demanding approach. The indicator creation industry started. András was skeptical, critical and developed rigorous principles, criteria for indicators and their usage. We came up with bold ideas and András patiently repeated the clear principles and prevented us from cheap but not lasting solutions. He was the master and still he is. Shapes and influences the study of scientific communication. Today he is one of the leading figures not only in scientometrics but in social science methodology [4],[5]. Here it is proper to note that his mentoring resulted that one of his students, Wolfgang Glänzel, gained similar exceptional position on.

Today scientometrics similarly to many other disciplines enjoys the power of computers, sheer size of storage capacity and abundance of data. The new possibilities bring to life two paradigms, Data Mining and Big Data, and new topics to investigate which were either infeasible or even below the horizon. In what follows we bring up some uncombed ideas of this kind. The hidden intention again to call András' interest and initiate joint thinking and work. We all know András's bound to music, to jazz, that's why the rest of this paper is arranged into three themes and free improvisation around them.

Theme one—Academic inbreeding [10]

This first topic is going to serve as an example to the new, the data mining approach. In the publication arena there are different types of competing actors. Scientists compete for more recognition through more a publication, higher visibility and more citations. Of course sole scientific merit attracts acknowledgement, but there are several known tricks which can increase such kind of reward of authors. But this topics is not about such tricks. We also witness the competition of scientific publishers and even countries. Our focus is on the intermediate level, on journals. Journals also compete for authors, for subscribers, for readers and last but not least for governmental or independent sponsors. Journals serve the smaller or wider scientific community as their founders/publishers/owners define their mission. Their mission statement and practice might not in line. Here we formulate a framework by which journals might be caught which deviating from the service of interest of the global scientific community. In particular a simple tool is suggested to find journals with biased acceptance policy, papers published coming from a very closed circle of institutions or from other kind small groups.

Before we proceed one intention of our master, András should be recalled. The statistical method we use is not an oracle. The pinpointed journals might belong to the majority of the fair ones regardless of their limited scope of institutions, others after a second closer look may truly fall into the gray zone of science. Without that second careful investigation, no judgment should be made.

Now let us consider a field of science e.g. economy. Using Thomson's WOS database we collect data on affiliation of authors publishing in journals of a given field (in

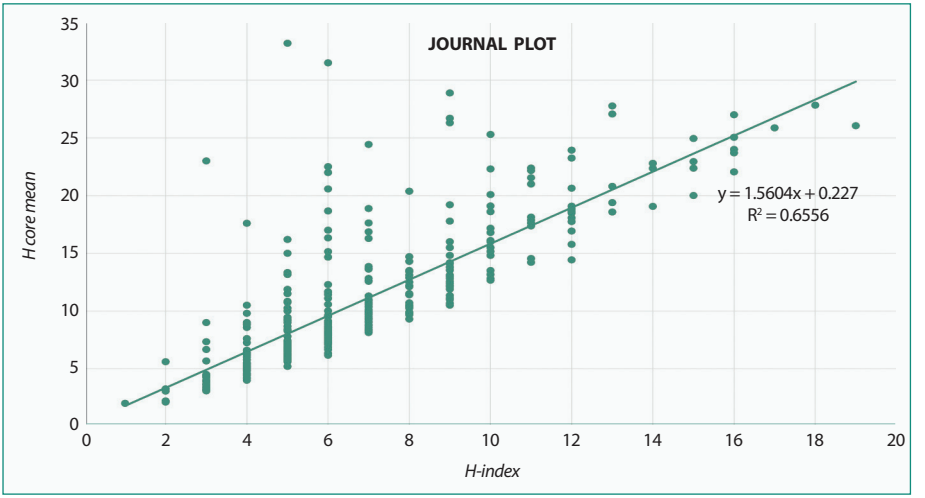


Figure 1. H-core mean vs H-ndex

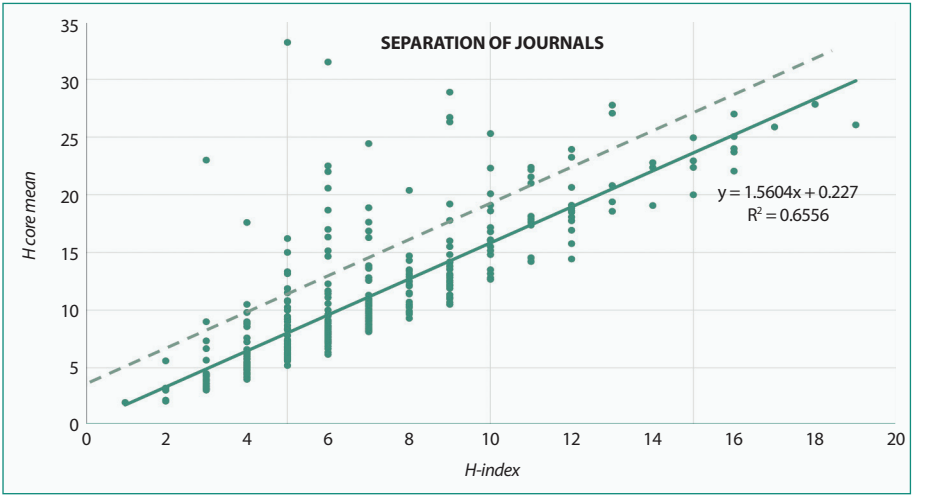


Figure 2. The separation of journals

our case economics). For each journal let us develop the distribution of entries for institutes *DI*, how many times a paper is accepted by the journal from a given institute. We claim, that distribution itself contains enough information about the fairness or bias of the journal acceptance policy. After collection of data, several indicators for the *DI* can be developed hoping it reflects well the biased policy. Among others the Gini index, the H-index, the average count in the H-core is calculated and completed with the size of the journal, field adjusted impact factor of the journal and country of origin if it can be identified.

Our assumption is that the biased policy reflected in a very skewed, tail heavy distribution. If this is the case that is reflected in the H-core of the institutes for the given journal. The average count in the H-core is much higher than it is expected given the size of the core, i.e. the H-index. The conjecture is supported by the trends visible on Figure F1 (two items, VALUE HEALTH and AM ECON REV have been removed for better scaling of the figure).

One can find a nice linear connection between the average of count in the core and the H-index. But there is a small bunch of journals which deviate from the overall trend. One can separate them with the shift of the found trend line as Figure F2 shows.

The dashed line represents the upper confidence interval boundary for the population trend line with the width of 2.5 times the standard deviation. There are 29 journals above that level. On Figure F3 and F4 the trend difference is presented for the two group of journals.

The two groups can be further investigated and one can find significant differences in other characteristics. Among them the mean Gini indexes for the selected 0.5069 and majority 0.3567 are significantly different. Similarly the field corrected Impact factors 0.4184, 0.4721 show significant difference, the majority of journals performs much better. One by one investigation of the suspicious items reveal that the AM ECON REV is not a representative element of that group but a very selective excellent journal. Others are pheriferic journals, mostly publishing works from their own less developed home country.

Theme two—Citation inflation [11]

It is already commonsense that the number of citations as measure of scientific quality is heavily biased by the activity of a minority of authors who artificially increase the number of their received citation. There are many tricks to do so, among others the mutual favour citations, "scratch each other's backs" is our point of interest. Can we sort out those cheating authors by analyzing only the citation graph? If not what further data and technique might be useful?

We started a little pilot study in the field of Computer Science, Cybernetics of WOS. A little literature research revealed that similar phenomena are known and cause similar ethical and business problems with respect of website link. Link or citation farms and other tricks are used to increase the visibility, promoting to the top in Google hit list the given web site. There are several methods which help to identify such websites and activity, but those use very sophisticated web specific data, not only the graph structure.

Our starting point was the citation graphs (Cit) built by using Grauwin's algorithm [9]. The second equally important graph is the coauthor graph (CA). We performed two basic procedure on both, exploration of cycles and strongly connected components. Our main tool was the Louvain algorithm [8] and an additional simple indicator the distribution with multiplicity of the referred authors by a fixe author. Short cycles, small strongly connected components and extreme reference values call for closer investigation. Of course publication time, affiliation of authors, personal links are all factors may contribute to the detection of cheating, but not reflected by the Cit and CA graph.

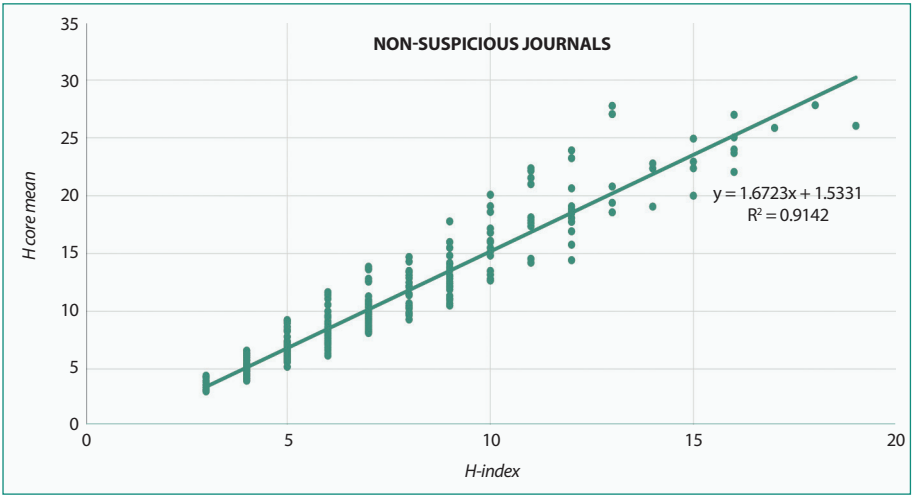


Figure 3.Trend line for the regularly behaving journals

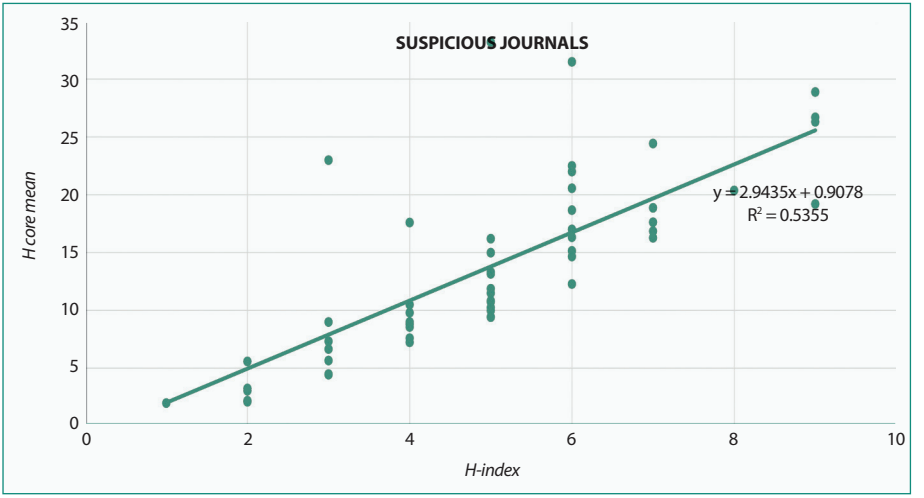


Figure 4. The trendline for suspicious journals

Our investigation resulted a few suspicious authors. The cursory investigation found that some basically different groups can be identified.

1. Small, almost full graphs, where the group of mutually citing authors are coauthors as well.
2. Two authors cite each other, and it can be assumed for good reason that one is supervisor the other is student.
3. Small number of authors with mutual citation dating back to one key paper, when and where their research interest met and their activity ran in parallel.
4. Other citation patterns.

Practically the identification of the first type can be automated. Without incorporating further data it is very difficult to distinguish between the others. Inclusion of temporal information, affiliation data, and reveal of joint conference visits may help to improve the analysis. Consequently the proper identification of favour citations needs further, more advanced data mining techniques. The most promising approach might be text mining, in particular topics analyses [12],[13]. If the topics of the suspicious cited and citing paper have no or very little overlap, if that is the practice in the investigated group, the likelihood of cheating increased, and the case is worth for a human check.

Theme 3 Normalized indicators [16]

The struggle for good normalization of indicators with respect to scientific fields/subfields is as old as the indicators themselves. For a long time it was impossible to build enough big databases and perform tedious computations to build the proper normalizing factors [14]. That is why the author's proposal to use reference groups to build normalization factors failed in the 80's. Today such calculation is feasible thanks to the waste storage and computing power. In this Theme we present the idea applied to reference group based journal impact factor normalization [16]. Our actual work uses the dataset of the Journal Citation Report 2006. The cited and citing records are both given for the 2006 year. We prepare the graph of bibliographic coupling [17],[18] of the journals given a similarity treshold *w*. As a result for each journal we have the set of neighbors *N(x)* and let

$$V(x) = N(x) \cup \{x\},$$

the journals which linked to it, and *x* itself. That set is considered the reference group for the journal *x* and the reference group based normalized impact factor (in short corrected IF, *CIF*) defined using

$$\overline{IF}(x) = \frac{1}{|V(x)|} \sum_{y \in V(x)} IF(y)$$

as

$$C[IF](x) = \frac{IF(x) - \overline{IF}(x)}{\overline{IF}(x)},$$

where *IF(y)* denotes the impact factor of a journal *y*. It is clear that this definition of $C[F] = C[F]_w$ depends on *w* but here we do not discuss how *w* should be chosen. One can see that the corrected impact factor oscillates around 0.

Here we may refer back to Theme 1. If a journal, in order to increase the impact factor, applies great pressure to the authors to cite papers from the same journal, than the higher of the concentration is the closer the *CIF* is to one, regardless of the real merit of the journal. Consequently those journals, which have low impact factor but *CIF* is close to 0 might belong to the gray zone of scientific communication.

The proposed scheme generalizes to other indicators. If $I(x)$ is such one, we use

$$\bar{I}(x)\frac{1}{|V(x)|}\sum_{y\in V(x)} I(y)$$

and define the corrected indicator as

$$C[I](x)=\frac{I(x)-\bar{I}(x)}{|\bar{I}(x)|},$$

provided

$$\sum_{y\in V(x)} I(y)\neq 0$$

Let us note that in case of signed indicators, the normalized indicator not restricted to [–1, 1] It can be very large if $\bar{I}(x)$ is close to zero. In such case it might be that other normalization is more appropriate or at least careful interpretation needed.

The same dataset allows us to develop an other indicator, which reflect a kind of role of the journal in the scientific communication. The inspiration came again from András's papers [15],[23] and partly from the work of neuroscientists who investigated the network of brain areas[19-22]. In particular we propose the application of the local node convergence degree as a new indicator for nodes, journals of scientific communication. The definition is relatively simple. Let

$C(x) = \{y : y \text{ cites } x\}$,
$R(x) = \{y : x \text{ cites } y\}$,

and $cite(x) = |C(x)|$, $ref(x) = |R(x)|$. For non-isolated nodes

$$CLD(x)=\frac{cite(x)-ref(x)}{cite(x)+ref(x)}.$$

This indicator shows, as spelled out in the papers [19-22], the information dissemination or "absorption" role of a node. If we seek for an indicator for the qualitative contribution of the journal one may consider any measure q of quality of journals, let it be absolute or relative. Analogously we define

$$Qin(x)=\sum_{y\in C(x)} q(y)$$

$$Qout(x)=\sum_{y\in R(x)} q(y)$$

$$Q(x)=\sum_{y\in C(x)\cup R(x)} q(y)$$

and for x-s when Q(x) ≠ 0 let

$$LCD_q(x)=\frac{Qin(x)-Qout(x)}{Q(x)}.$$

As an example one can consider $q(x) = IF(x)$. It is clear that both convergence indicators are in [−1, 1] and by definition normalized and can be used on graphs covering different scientific fields as well. As it is indicated by András in [15] the journals where the cited papers appeared typically has higher impact factor then the citing journals. This means that this indicator will be typically negative. As we introduced above, we can normalize the local convergence degree and the qualitative version as well:

$$C\left[LCD\right](x) = \frac{LCD(x) - \overline{LCD}(x)}{\left|\overline{LCD}(x)\right|}$$

$$C\left[LCD_q\right](x) = \frac{LCD_q(x) - \overline{LCD_q}(x)}{\left|\overline{LCD_q}(x)\right|},$$

if $\overline{LCD}(x) \neq 0$, $\overline{LCD}_q(x) \neq 0$. .

Bibliography

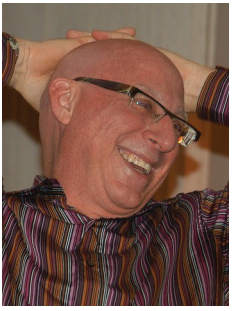
1. Schubert, A. and Glänzel, W., 1984. A dynamic look at a class of skew distributions. A model with scientometric applications. Scientometrics, 6(3), pp.149-167.
2. Albert, R., Jeong, H. and Barabási, A.L., 1999. Internet: Diameter of the world-wide web. Nature, 401(6749), pp.130-131.
3. Barabási, A.L. and Albert, R., 1999. Emergence of scaling in random networks. science, 286(5439), pp.509-512.
4. http://academic.research.microsoft.com/?SearchDomain=22&SubDomain=7&entitytype=2
5. https://scholar.google.hu/citations?user=DTNpAiwAAAAJ&hl=en
6. Korn, A., Schubert, A. and Telcs, A., 2009. Lobby index in networks. Physica A: Statistical Mechanics and its Applications, 388(11), pp.2221-2226.
7. Schubert, A., Korn, A. and Telcs, A., 2008. Hirsch-type indices for characterizing networks. Scientometrics, 78(2), pp.375-382.
8. Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), p.P10008.
9. Grauwin, S., Beslon, G., Fleury, E., Franceschelli, S., Robardet, C., Rouquier, J.B. and Jensen, P., 2012. Complex systems science: dreams of universality, interdisciplinarity reality. Journal of the American Society for Information Science and Technology, 63(7), pp.1327-1338.
10. Nagy, A.M.,Soos, S., Telcs, A., Török, A., Vida, Zs., Academic inbreeding (working title), in preparation
11. Danis, M., Konka,B., Török, Á., Telcs,A., Examination of mutual favors in the citation networks (working title), in preparation
12. Aggarwal, C.C. and Zhai, C., 2012. Mining text data. Springer Science & Business Media.

13. Berry, M.W. and Kogan, J., 2010. Text Mining. Applications and Theory. West Sussex, PO19 8SQ, UK: John Wiley & Sons.
14. Schubert, A. and Braun, T., 1993. Reference standards for citation based assessments. Scientometrics, 26(1), pp.21-35.
15. Schubert, A., 2013. Measuring the similarity between the reference and citation distributions of journals. Scientometrics, 96(1), pp.305-313.
16. Fischer, H., Telcs, A., Some notes on scientometric indicators (working title), in preparation
17. Glänzel, W. and Czerwon, H.J., 1996. A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. Scientometrics, 37(2), pp.195-221.
18. Glänzel, W., 2012. The role of core documents in bibliometric network analysis and their relation with h-type indices. Scientometrics, 93(1), pp.113-123.
19. Négyessy, L., Nepusz, T., Zalányi, L. and Bazsó, F., 2008. Convergence and divergence are mostly reciprocated properties of the connections in the network of cortical areas. Proceedings of the Royal Society of London B: Biological Sciences, 275(1649), pp.2403-2410.
20. Bányai, M., Nepusz, T., Négyessy, L. and Bazso, F., 2009, September. Convergence properties of some random networks. In Intelligent Systems and Informatics, 2009. SISY'09. 7th International Symposium on (pp. 241-245). IEEE.
21. Bányai, M., 2013. Functional modelling of cortical macro-networks: a dissertation submitted for the degree of Doctor of Philosophy.
22. Bányai, M., Négyessy, L. and Bazsó, F., 2011. Organization of signal flow in directed networks. Journal of Statistical Mechanics: Theory and Experiment, 2011(06), p.P06001.
23. Schubert, A. and Telcs, A., 2014. A note on the Jaccardized Czekanowski similarity index. Scientometrics, 98(2), pp.1397-1399.

articles II: networks

Tracking a Lifetime of Contributions at the Topic Level: Nodes and Edges Associated with the Work of András Schubert

KEVIN W. BOYACK[a] & RICHARD KLAVANS[b]
[a] *SciTech Strategies, Inc., Albuquerque, NM 87122 (USA) (kboyack@mapofscience.com)*
[b] *SciTech Strategies, Inc., Wayne, PA 19087 (USA) (rklavans@mapofscience.com)*



Introduction

As is the case for most researchers, we have far more ideas of experiments we wish to perform than time to perform them. Some ideas will succeed and others will fail, and we typically don't know ahead of time which will be which. How, then, do we choose which ideas to try? Sometimes outside influences intervene and provide an excuse or opportunity to try one of these ideas. Such is the case for this exploratory study.

We have an ongoing research interest in being able to model the structure and evolution of the science system in a way that will allow better decisions to be made by agencies, institutions, and individuals. This not only requires an ability to create accurate models of science (or as accurate a model as possible), but it also requires that we understand the interplay of the various actors and agents in the science system. One question that is continually asked by most stakeholders is how to identify the hottest, most emergent, or most innovative topics. Similar questions are asked about researchers—which researchers do the most innovative work? Of course, to ask and answer these questions, one must first define "innovative" and "emerging", and determine how they can be measured. This is, as we all know, a difficult thing to do (Rotolo, Hicks, & Martin, 2015). For instance, one widely held assumption in bibliometrics is that the most highly cited papers are also the most innovative (cf., Uzzi, Mukherjee, Stringer, & Jones, 2013). However, a recent survey of influential researchers with many highly cited papers found that these researchers consider their highly cited "synthesis" and "incremental advance" papers to be just as important as (and more prevalent than) their "innovative" papers (Ioannidis, Boyack, Small, Sorensen, & Klavans, 2014). Moreover, most highly cited papers are not associated with emerging topics (Small, Boyack, & Klavans, 2014).

In recent years, multiple research teams have begun to explore if edges in the scientific network (e.g., links between papers, references, chemicals, etc.) can be used to signal innovation. For example, Uzzi et al. (2013) used references from papers to create a journal network, and defined edges between journals (or co-cited journals) in that network as either 'typical' or 'atypical' (novel) edges based on comparison to their expected values. Individual papers were then classified based on the combination of typical and atypical edges associated with their references. More recently, Foster, Rzhetsky & Evans (2015) created a time-dependent network of chemicals based on their co-occurrence in PubMed records. Edges were defined as links between chemicals, and new edges were considered to be the most innovative. Both of these methods show promise, and we believe that exploring the relationship between edges and innovation is a potentially fruitful path to follow.

This brings us back to the subject of the current study. Can this new emphasis on an analysis of edges help us better characterize the innovativeness of an individual researcher? It is within this context that the 70th birthday festschrift for András Schubert provides us with an opportunity. As our first exploration into edge structures that are potentially related to innovation, this paper presents an analysis of András' publication history. The paper proceeds as follows. We first give a brief overview of our model and the method for defining edges. We then identify the topics (clusters) associated with András' work, show his entry into different topics over time, and compare the edges existing between those topics at the time of entry and in 2013. We then close with a discussion of the results and how they may help us to understand how to better characterize innovative publication strategies.

The STS global model of science

The SciTech Strategies (STS) global model of science was created from 24,615,844 indexed source documents from Scopus (1996-2012), and also contains 23,917,457 non-source documents that were each cited at least twice by the set of source documents. Thus, the model contains over 48.5 million documents in total. The set of 582 million direct citation links between these 48.5 million documents was used to create the model. Clustering was done using the recently created CWTS method and algorithm (Waltman & van Eck, 2013). Recent work suggests that this methodology may produce more accurate clusters than competing methods (Emmons, Kobourov, Gallant, & Börner, 2016). The CWTS algorithm can be tuned to produce different numbers of clusters using minimum cluster size and resolution parameters. Using settings of a minimum cluster size of 50 papers and a resolution of 3×10^{-5}, the resulting cluster solution contained 91,726 clusters, which was close to our desired number of approximately 100,000 clusters. Each cluster represents a topic, and is comprised of the papers on that topic and the community of researchers working on that topic. Our experience is that at the 100k cluster level, a) experts can easily differentiate between topics (Boyack, Klavans, Small, & Ungar, 2014), and b) funding can be assigned to topics and is correlated with innovation metrics (Boyack & Klavans, 2015).

A map of the 91,726 topics (or research communities) has been created to provide a visual depiction of the structure of science (Figure 1). This map was created using the following process. First, the similarity between pairs of topics was calculated from the titles and abstracts of the documents in each topic using the BM25 similarity measure (Sparck Jones, Walker, & Robertson, 2000). Second, the resulting similarity list was filtered to keep only the top-n (between 5 and 15) similarities per topic. Finally, a layout of the topics was created using the DrL (OpenOrd) algorithm (Martin, Brown, Klavans, & Boyack, 2011), which gives each topic an x,y position based on the similarity graph. Each of the 91,726 topics in the map has been designated as belonging primarily to one of twelve high-level fields using journal-to-field assignments from the UCSD journal schema (Börner et al., 2012), and each is colored correspondingly in Figure 1.

At a high level, the field structure in the STS map of science is similar to that of many other global science maps, including the consensus map of science (Klavans & Boyack, 2009). Physics, chemistry and engineering are highly related, and are adjacent to each other. The medical areas (disease, medicine, health sciences, brain sciences) also are adjacent to each other. Biology is adjacent to chemistry and medicine, earth sciences is primarily adjacent to engineering, and social sciences are adjacent to health sciences, while computer science lies between physics (which includes mathematics) and the social sciences.

The accuracy of this model and map of science was recently established by comparing large-scale models of Scopus data created using direct citation, co-citation, and bibliographic coupling (Klavans & Boyack, 2016). Using papers with at least 100 references as gold standards, direct citation was shown to concentrate references at a higher level (indicating more accurate clusters) than co-citation or bibliographic coupling.

Defining edges

Our unit of analysis is topics (clusters of papers from our model of science) rather than papers. Thus, we define edges as links between topics. These edges can be specified in a number of ways. For instance, to create the map of Figure 1 we specified edges as textual links between topics; this was an appropriate specification for creation of a static map. We could specify edges using the references in papers. Papers with large numbers of references to multiple topics are inherently signaling an ***intent*** to link those topics. Intentions, however, are often not realized. If a paper is published that proposes an edge between two topics, and those two topics are not cited together by subsequent papers, can we really say that the edge exists? We suggest that the answer to that question is no.

We thus prefer to specify edges that are based on citations. If, for example, a single paper is cited by a large number of papers in each of two topics, this suggests that both of those topics consider the paper to be important. The resulting ***outcome*** is that these two topics have linked themselves together by virtue of citing the same paper. We suggest that edges based on outcomes are more meaningful and more robust than edges based on intentions. We thus choose to specify edges as links between topics based on

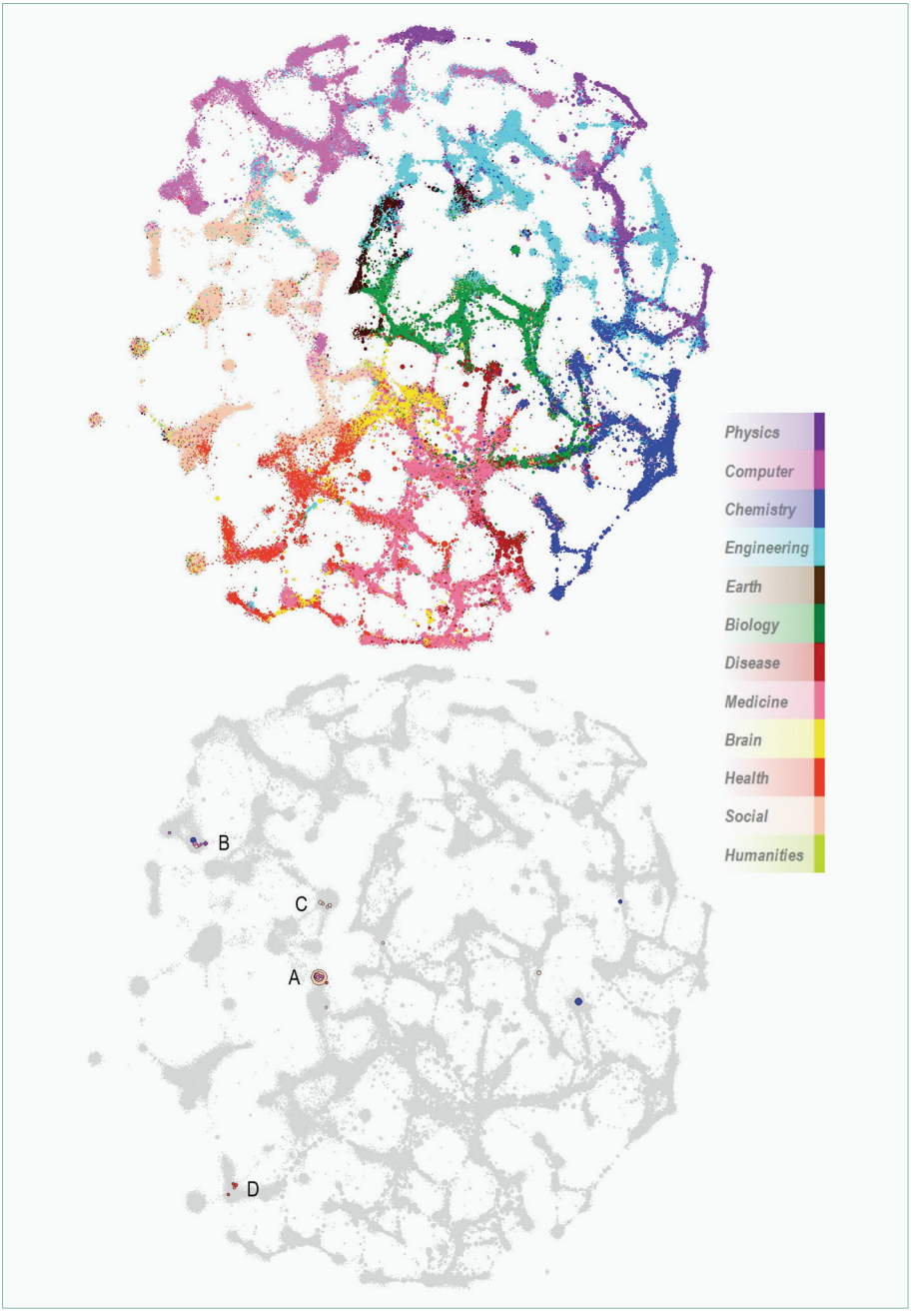


Figure 1. Visual map of the STS model of science (top) with an overlay of the papers published by András Schubert (bottom). Groupings of clusters are labeled A—D.

citations to highly cited papers. Why limit edge specification to highly cited papers? There are several reasons for this. First, the signal associated with citations to a paper that is cited only a few times is too small to definitively say that two or more topics are associating themselves with that paper. Secondly, if a paper is not highly cited, it is not part of the core of scientific network. This doesn't mean that the paper is not valuable, or that it does not contain important information. Rather, it simply indicates that others do not consider it to be part of the small core of science. Finally, outcomes may be very different than intentions. The communities that end up citing a paper, and their reasons for citing, may be different than what was intended by the authors. However, many citations must accrue for these unintended outcomes to become clear.

We define edges using a moving five-year window. For each five-year window from 1996-2000 through 2008-2012, we identify the top 1% highly cited papers from our model of science. Since the model contains so many documents, the numbers of top 1% papers are not small, ranging from 247,797 (those cited > 45 times) from the first time period to 514,130 (those cited > 62 times) from the final time period. Edges are specified between pairs of topics where each of the top two topics cites one of these top 1% papers at least 5 times during the time period.

Since multiple time periods are used, edges are born and then grow wider over time. This is analogous to a new footpath that is formed between villages, which can then grow into a bike trail, a road, and finally a superhighway as the villages grow into towns and cities and more travelers traverse the path. We suggest that innovative edges are those that are born and then quickly grow—these can be assumed to remain innovative until they are large enough to be considered part of the normal fabric of science, after which time they are no longer considered innovative. Using this framework, one might consider the most innovative researchers to be those who either 1) author papers associated with innovative edges, or 2) concurrently work in multiple topics that are not connected by an edge, but for which an edge is created shortly thereafter.

Schubert's contributions

Scopus has record of 129 documents authored by András Schubert from 1981-2012; 108 of these are available in our model of science. Of the other 21 papers that are not present in the model, most are excluded because Scopus lacked references for these pre-1996 papers. The 108 papers included in the model are located in 29 separate topics. Figure 1 shows that these 29 topics are located in several areas of the map of science, and that the majority of these topics are in four major groups (labeled A—D). It is interesting to note that topics from the social sciences, computer science, and the medical sciences are all co-located in group A, while topics from computer science, chemistry, physics, and the social sciences are present in group B. The fact that topics within the same group are assigned to different fields of science does not mean that they are topically dissimilar. To the contrary, it shows that similar topics are found in multiple fields, as evidenced by the representative terms shown for each topic in Figure 2.

It is also interesting that the papers in group D are all in topics related to health sciences even though these papers were published in the journal *Scientometrics*. Although one might think that these papers should have been located in clusters with large numbers of papers from *Scientometrics*, their references and the papers that cite them are primarily from health sciences fields.

Figure 2 shows the topics containing papers published by András Schubert. Each paper is not shown individually. Rather, the entry year into each topic is shown by the circles in the figure, and the size of the circle reflects the total number of Schubert's papers in that topic from the entry year through 2012. Horizontal lines show how long Schubert was actively publishing in each topic. Topic #432 contains the largest number (33) of his papers.

The largest portion of Schubert's work (72 papers) is found in group A, which is also the first group he enters, in 1981. As mentioned above, this is a multidisciplinary area that, overall, deals with the measurement of scientific output. Group A is his home—not only does it contain the majority of his papers, but it houses the topics in which he resided the longest. Over the past 30 years, Schubert has made visits to groups B, C, D, along with some isolated topics. These are not distant visits from a cognitive perspective. As already mentioned, group D papers are related to scientific output, but are linked through references and citations to the health sciences. The visits to groups B and C are to topics related to complex networks, community detection, and knowledge flows, all of which rely on methodologies that have been used in conjunction with citation analysis. These forays into related topics are like being a visiting professor—Figure 2 shows that these were relatively short-term visits that broadened Schubert's intellectual domain. Ultimately, however, his long-term research commitment was to the topics in group A.

Figure 2 also shows edges (as defined above) linking some of the topics in which Schubert was active. Edges were defined as of the year that the newer of the two topics was entered for topics after 2000, while edges were defined using 1996-2000 data for topics entered before 2000 (e.g., the edge between #34085 and #432). Unfortunately, we could not identify edges for dates earlier than this because we do not have references for papers published prior to 1996. We note that very few edges existed between Schubert's portfolio of topics prior to 2000. In this sense, the fact that he was active in these topics—thus inherently linking those topics through his work—and that the few realized edges between these topics were weak, suggest that he was a pioneer trying to link work in these topics. If we look at this same set of topics in 2012, we find that only one strong edge has developed—the edge between topic #11438 (scientific networks) and #432 (bibliometric indicators). Thus, some of the efforts by Schubert to link topics appears to have been successful.

The strongest edges between topics entered by Schubert (at the time of his entry) were to topics that he only visited. The link from topic #3770 (community detection) to #104 (complex networks) was already very strong in 2010, and continued to get stronger through 2012. The link from topic #9524 (citation networks) to #14807 (search engines) was moderately strong in 2010, but had decayed severely by 2012.

In this analysis we have shown the topical distribution of András Schubert's published works, how those topics fit within a larger context, and have identified links between those topics that might be seen to be innovative. We note that this is an ex-

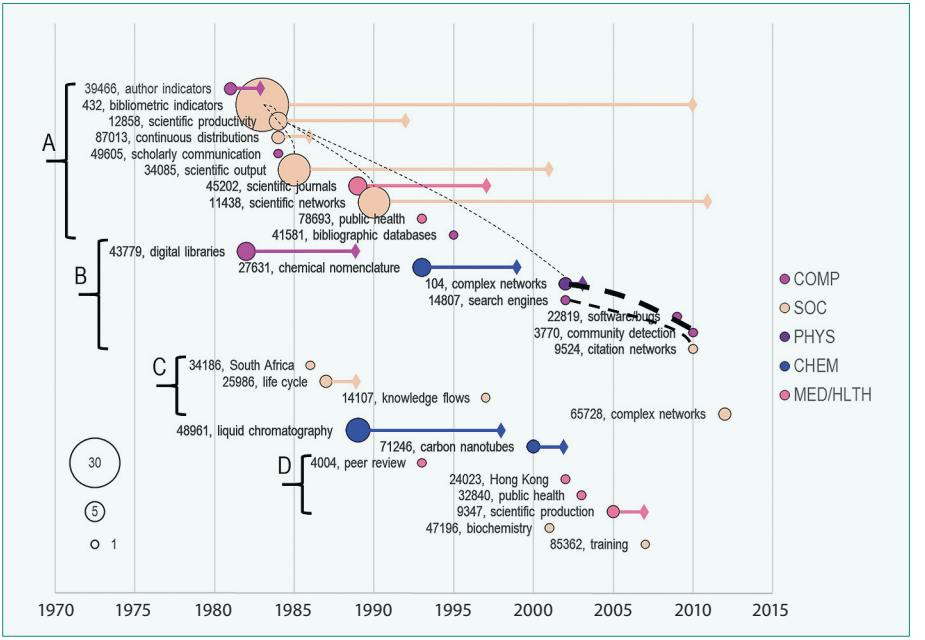


Figure 2. Topics and edges associated with the papers published by András Schubert. Groups A—D are the same as those shown in Figure 1. Node size indicates the number of papers by A.S. in the topic, while the horizontal lines indicate how long A.S. was active in the topic. Dashed lines indicate edges existing at the time A.S. entered the topic.

ploratory study, and that using links between topics as a proxy for innovation is not well understood. Much more work needs to be done to understand how bibliometric data can best be used to understand innovation.

Perhaps the most salient lesson we can learn from these data is the following. The history of Figure 2 clearly shows that 1) András Schubert has been relentless in his pursuit of focused knowledge as evidenced by the length of time that he has published in a number of topics, and 2) he was very willing to broaden his horizons by visiting a large number of topics. This can serve as an example to us all. We close with our best wishes to Dr. Schubert on his 70th birthday, and that he will have many more productive and joyful years to come. May we all be so productive!

References

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464.

Boyack, K. W., & Klavans, R. (2015). Is the most innovative research being funded? *20th International Conference on Science and Technology Indicators*.

Boyack, K. W., Klavans, R., Small, H., & Ungar, L. (2014). Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. *Journal of Engineering and Technology Management, 32*, 147-159.

Emmons, S., Kobourov, S., Gallant, M., & Börner, K. (2016). Analysis of network clustering algorithms and cluster quality metrics at scale. *PLoS ONE, submitted*.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review, 80*(5), 875-908.

Ioannidis, J. P. A., Boyack, K. W., Small, H., Sorensen, A. A., & Klavans, R. (2014). Is your most cited work your best? *Nature, 514*, 561-562.

Klavans, R., & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology, 60*(3), 455-476.

Klavans, R., & Boyack, K. W. (2016). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, forthcoming*.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: An open-source toolbox for large graph layout. *Proceedings of SPIE—The International Society for Optical Engineering, 7868*, 786806.

Rotolo, D., Hicks, D., & Martin, B. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827-1843.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*, 1450-1467.

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. Part 1. *Information Processing & Management, 36*(6), 779-808.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*, 468-472.

Waltman, L., & van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B, 86*, 471.

Mapping the Œuvre of András Schubert with Advanced Bibliometric Instruments

ANTHONY F. J. VAN RAAN
Centre for Science and Technology Studies, Leiden University
Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands
vanraan@cwts.leidenuniv.nl

Abstract: In this contribution we investigate important properties of the András Schubert papers. We present an approach in which the 'cognitive environment' of the Schubert papers is analyzed, based on the mapping of these papers and their citing papers using citation links and conceptual relations.

1 Introduction: Bibliometric Instruments to Map Scientific Work

There are two major bibliometric approaches to map scientific work, particularly in terms of its 'cognitive environment'. The first one is citation-based, the second is concept-based[1].

First the citation-based approach. As any other publication, the András Schubert papers have links with other (earlier) publications by their references and it is interesting to find out whether there are Schubert papers that have references in common. This might reveal clusters, in bibliometric terms these are bibliographically coupled Schubert papers. And the other way around, these common references are co-cited by the Schubert papers. Clusters of co-cited papers can be seen as thematic basic work for the Schubert papers.

In the second approach we use natural language processing (text mining) to extract the important, publication-specific concepts (terms such as keywords or noun phrases) from the titles and abstracts of the Schubert papers. By measuring all co-occurrences of any possible pair of concepts, co-word maps can be created in which the conceptual structure of the research represented by the set of the Schubert papers is visualized.

For both approaches we used the recently developed CWTS bibliometric instruments *CitNetExplorer* and the *VOS-viewer*.

[1] The text to describe these bibliometric approaches and their results is mainly drawn from Van Raan (2015) as this recent paper provides a general textual framework.

The CitNetExplorer (van Eck and Waltman 2014) is a software instrument specifically designed for analyzing and visualizing citation networks of scientific literature. It can be uploaded with sets of publication records directly from the Web of Science (WoS) or Scopus. Citation networks can then be explored interactively, for instance by drilling down into a network and by identifying clusters of closely related publications[2].

The VOS-viewer (van Eck and Waltman 2010) is a software instrument for constructing and visualizing (mapping) a broad range of bibliometric networks. These networks may for instance include journals, researchers, or individual publications, and they can be constructed with co-citation, bibliographic coupling, co-author or co-affiliation relations. In particular, the VOSviewer also offers a text mining functionality that can be used to construct and visualize conceptual (co-word based) networks of terms extracted from a body of scientific literature, particularly titles and abstracts of publications. The VOS viewer can be uploaded with any type of relational information and particularly with publications records of the WoS as well as of Scopus[3].

2 Analysis of the Schubert papers

2.1 Citation links

We analyze the citation network for the Schubert papers by creating a full WoS record (title, abstract, authors, institutions, references) set of these papers and uploading this set into the *CitNetExplorer*. Thus, the Schubert papers are the source publications and their references define the citation links. This procedure renders a citation network based on these references if sufficient citations links are available. In the visualization of the citation network each circle represents a publication. Publications are labeled by the last name of the first author. To avoid overlapping labels, some labels may not be displayed. The horizontal location of a publication is determined by its citation relations with other publications. The vertical location of a publication is determined by its publication year. The lines represent citation relations, citations point in upward direction: the cited publication is always located above the citing publication. Publications are clustered based on their citation relations. The identified clusters have different colors.

The CitNetExplorer algorithm applies threshold values of important parameters for the construction of the citation network, particularly for the minimum number of citation links and also for the minimum cluster size. In this sense, the CitNetExplorer operates as a community detection tool. We refer to the methodology section in the CitNetExplorer website for details (see footnote 2). A high value for the minimum number of citation links (e.g., 10) results in a sparse network and a low value (e.g., 2) gives an overcrowded picture. The interactive character of the CitNetExplorer enables the user to experiment with the

[2] More about CitNetExplorer: http://www.citnetexplorer.nl/Home

[3] More about VOSviewer: http://www.vosviewer.com/Home

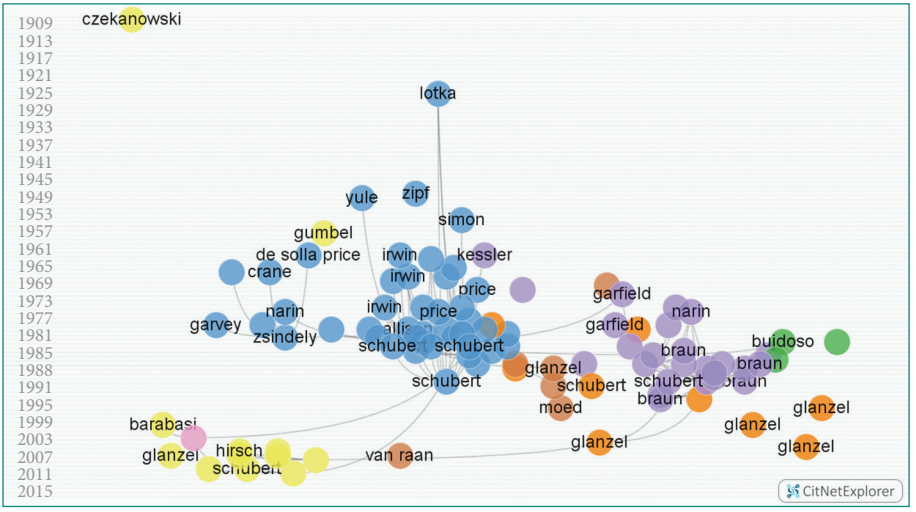


Figure 1: References map of the Schubert papers.

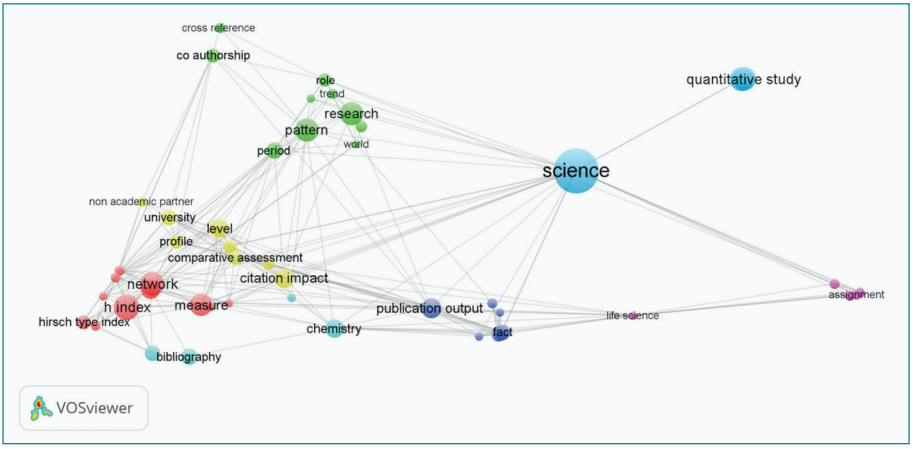


Figure 2: Concept map of the Schubert papers.

data and to find out the differences in the created networks imposed by the adjustable parameters. Thus, the CitNetExplorer allows to find an optimal network configuration.

By trying out several parameter values, we find a sensible representation of the overall citation network analysis with value 2 for the minimum number of citation links and value 2 for the minimum cluster size. This minimum number of 2 citation links means that in the set of the Schubert papers only references that occur at least in 2 different papers are included in the construction of the network. This provides us with a general overview shown in Fig. 1. Several clusters are detected, indicated by colors. Most of these clusters are small, mainly because of the threshold for the minimum number of

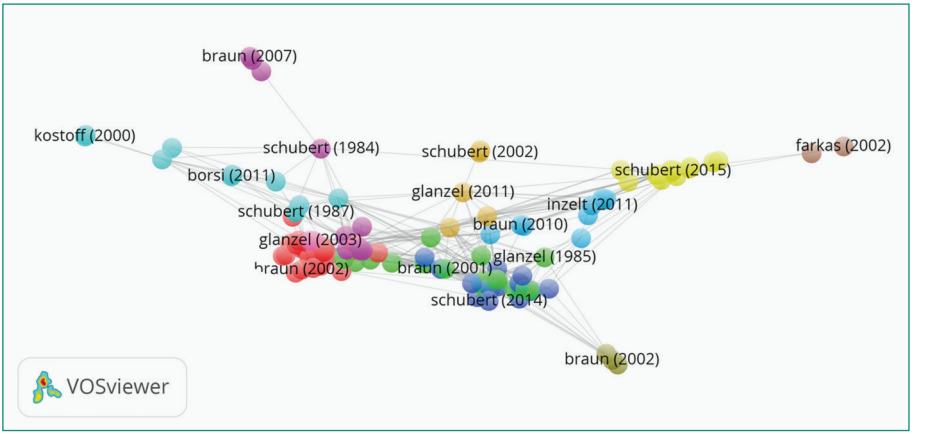


Figure 3: Bibliographic coupling of the Schubert papers (first author is indicated).

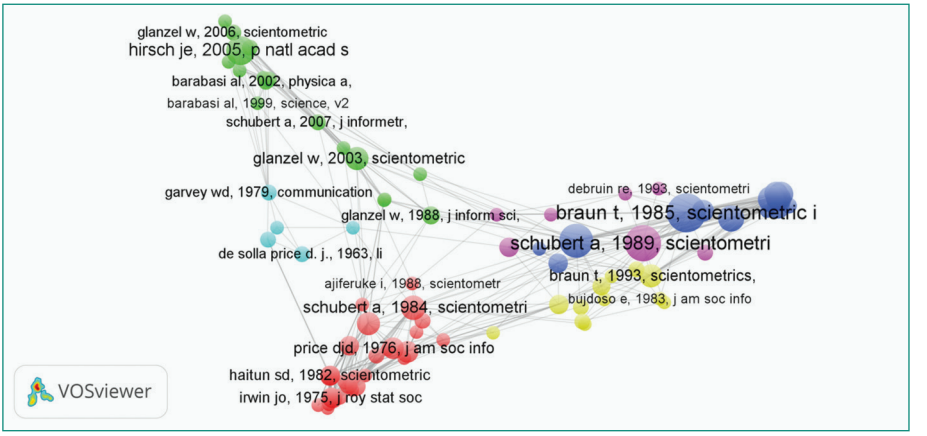


Figure 4: Co-citation map of the Schubert papers.

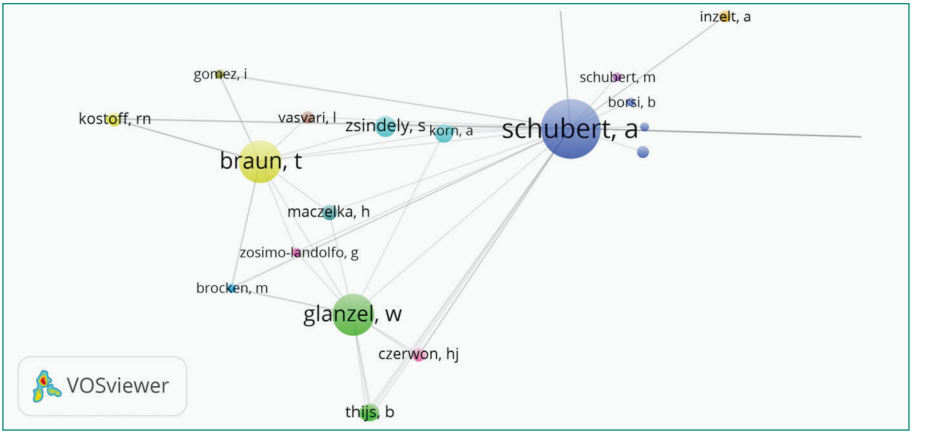


Figure 5: Co-author map (main part) of the Schubert papers.

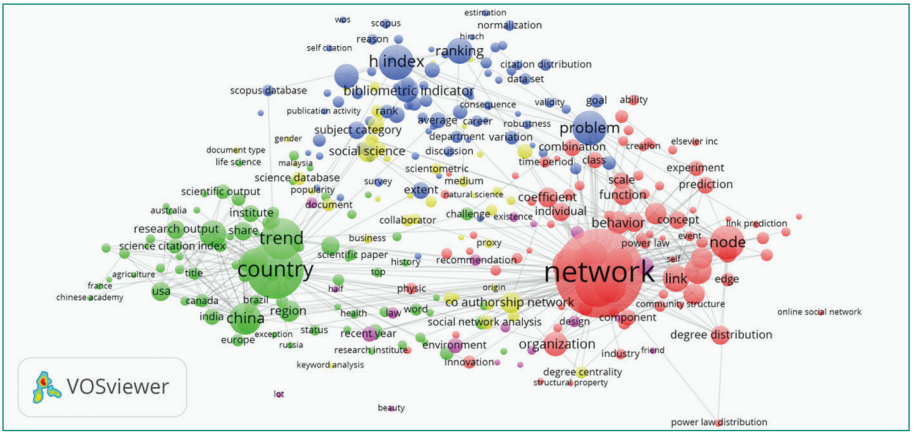


Figure 6: Concept map of the 500 most recent publications citing the Schubert papers.

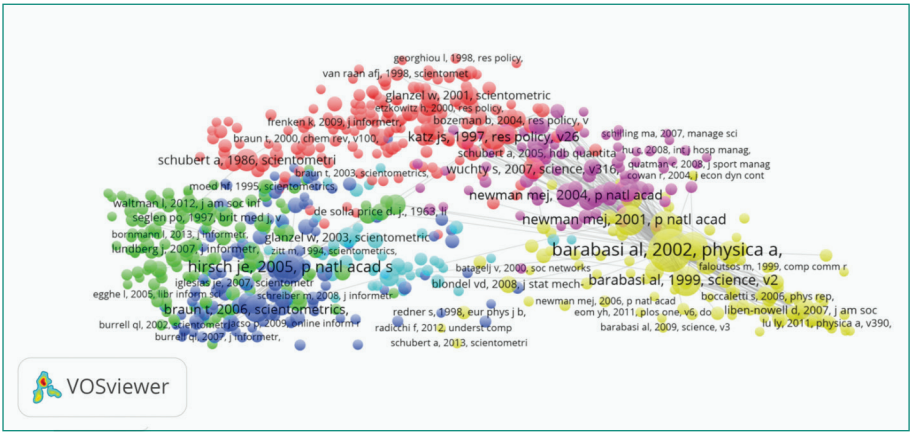


Figure 7: Co-citation map of the 500 most recent publications citing the Schubert papers.

citations links. Major clusters are for instance on the h-index and network analysis (yellow cluster, bottom left) and on statistical methods (blue cluster in the middle). Notice the 'old' references of Czekanowski (1909) on similarity measures, and the famous Lotka paper (1926) on the frequency distribution of the number of publications per author.

2.2 Concept Maps

The same WoS-based full record set of the Schubert papers used as source publications for uploading in the CitNetExplorer can also be used for uploading in the VOSviewer. Several choices can be made with the VOSviewer: co-citation, bibliographic coupling, co-authorship, and term co-occurrence (co-word) networks. We first apply the term co-occurrence facility to create concept maps. After uploading the set of full records, the VOSviewer applies a natural language text processing technique to collect terms (mainly noun phrases) from the titles and the abstracts of the Schubert publication records. In

a next step the VOSviewer calculates -after choosing a specific occurrence threshold- all term co-occurrences, i.e., in how many publications of the set any possible pair of terms co-occurs. This provides the data for the construction of the co-occurrence matrix.

A major challenge in the construction of concept maps is the selection of terms. Although many irrelevant (the, and, of, between, etc.) terms are automatically removed by the VOSviewer natural language text processing algorithm, still the algorithm selects terms such as 'theory', 'approximation', 'dependence', 'correlation', 'possibility', 'calculation', 'comparison', 'assumption', 'level', etc. The problem is that these terms may be relevant in some sets of publications, whereas in other sets they are not. Therefore it is sometimes unavoidable to remove specific terms manually. The VOSviewer provides this facility of manual term selection. This, however, is a tricky matter. If the set of publications is in the order of magnitude of 100, like in our case, the removal of just one term, for instance 'level' may quite drastically change the structure of the map. On the other hand, if rather general terms such as 'level' or 'principle' are not removed, it is possible that two clusters representing quite distant parts of an oeuvre are 'linked together' because in both subfields the term 'principle' may have a high occurrence (and thus co-occurrence with other terms). Often one has to find the best solution by trial and error, and particularly in the case of relatively small sets of publications (100, as in this case) full counting (FC) instead of binary counting (BC)[4] is a better choice.

The text processing of the Schubert papers rendered 1,142 terms, of which 83 meet the occurrence threshold-value 4 in the case of full counting. The results are shown in Fig. 2. We see a landscape with a variety of topics. These topics form clusters which are indicated by colors. Very clearly (and no surprise) András' work is about science, connected with a multitude of research themes such as citation impact, publication output, h-index, assessment, chemistry. A special position is for 'quantitative study' due to several general review papers on quantitative studies of science. The size of the circles represents the number of publications in which the term occurs in title or abstract. Lines indicate relatively strong connections between terms.

By using different occurrence thresholds, the main structure of the map will remain stable, but the details may differ considerably. Like the CitNetExplorer, the VOSviewer too is an interactive tool. Uploading the dataset into the VOSviewer enables the user to investigate carefully the influence of different occurrence thresholds, of binary versus full counting, and also of the removal of specific terms.

2.3 Citation- and Author-Based Maps

As discussed in the foregoing section, the VOSviewer offers the possibility to perform citation-based analyses, in particular bibliographic coupling (BCpl) and co-citation (CCit)

4 The VOSviewer offers the possibility to choose for full counting (FC), i.e., all occurrences of a term in a publication are counted; or binary counting (BC), i.e., only presence or absence of a term in a publication is counted, thus the actual number of occurrences of a term in a publication does not matter. Clearly, in FC those terms that are mentioned more than once in the abstract of a publication, for instance because these terms are central to the research discussed in the publication, get a heavier weight in the co-occurrence matrix calculations (we refer to http://www.vosviewer.com/Home)

analysis, and also co-author (CA) analysis. In the BCpl analysis the Schubert papers are analyzed in their role as *citing* papers and thus the BCpl-map shows how the Schubert papers are related to each other on the basis of commonly *cited* papers. We show the results in Fig. 3. The papers are indicated with their first author and year of publication. The circles have the same size because in bibliographic coupling papers are similar entities. We observe several clusters of Schubert papers, for instance the red cluster on classification of scientific fields and the light green cluster on the h-index

In Fig. 4 we present the results of the CCit analysis. Here the relation of the references (*cited* papers) of the Schubert papers (as *citing* papers) are mapped. References can be cited together (co-cited) in a reference list of a paper, and the more this occurs, the stronger their relation. In order to avoid overloading the map, the threshold for the minimum number of cited references is 3. Now we see that papers have different size, depending on the extent to which they are co-cited. The CCit map of the Schubert papers shows more clearly defined clusters than BCpl map. This is understandable because the number of references is an order of magnitude larger than the number of papers in which these references are mentioned. We observe that the clusters with different colors clearly mark different themes: for instance, the green cluster in the upper left corner on the Hirsch-index and on network analysis, the red cluster on field classification issues statistical issues and the dark blue cluster on the more general bibliometric issues.

We present the CA analysis in Fig. 5. We see the close relation with co-authors Tibor Braun and Wolfgang Glänzel. The size of the circles depends on the number of co-authored papers.

We remind that the above discussed maps are based on the entire oeuvre of András which covers a long time period of 35 years, from 1981 until today. It is possible to divide the oeuvre into 'slices' of, say, 5 years and perform similar analyses which then reveal developments over time. However, because the number of papers per slice of 5 years is, on average, 20, the maps will be quite sparse. In the next section we present an approach to investigate the Schubert oeuvre from a more recent perspective.

3 Analysis of the Papers Citing the Schubert Papers

3.1 Concept Maps

In this part of our contribution we will focus on the cognitive environment of the Schubert papers by creating a concept map and a co-citation map of the publications *citing the Schubert papers*. The Web of Science Core Collection covers in total 2,495 publications (February 16, 2016) that cite the Schubert papers. In order to create a more recent perspective, we select the 500 *most recent* citing papers, from 2013 until January 2016.

The text processing of the 500 most recent citing papers rendered 9,925 terms, of which 534 meet the occurrence threshold value 5 in the binary counting method. Obviously, to avoid an overcrowded map, we have to take a higher threshold as compared to mapping of the Schubert papers given the larger number of publications involved. The resulting map is shown in Fig.6. In this map we can clearly observe the same

research themes as those of the majority of the (cited) Schubert papers: the blue cluster on the h-index, bibliometric indicators and ranking issues; the green cluster on country-related bibliometric studies; and the red cluster on network analysis.

3.2 Citation-Based Maps

In Fig. 7 the co-citation (CCit) map of the 500 most recent citing papers is presented. Thus, a clustering of the references of the citing papers is constructed, in which many of the Schubert papers show up, as all citing papers cite by definition at least one of the Schubert papers. There is a considerable similarity between the clusters in the CCit map and those in the above discussed concept map of the 500 most recent citing papers: the light green cluster on network analysis; the blue cluster on the h-index; bibliometric indicators and ranking issues; the light green cluster on country-related bibliometric studies; and the red cluster on network analysis.

4 Concluding Remarks

In this contribution we investigated important properties of the Schubert papers. We showed an approach in which the 'cognitive environment' of the Schubert papers is analyzed, based on the mapping of these papers and their citing papers using citation links and conceptual relations. The interactive facilities of the CitNetExplorer and the VOSviewer enable a detailed analysis of scientific oeuvres by using different mapping modalities.

References

van Raan AFJ (2015). Dormitory of Physical and Engineering Sciences: Sleeping Beauties May Be Sleeping Innovations. *PLoS ONE* 10(10): e0139786.

van Eck NJ, Waltman L (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics 8*(4): 802-823.

van Eck NJ, Waltman L (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics 84*(2): 523-538.

András Schubert: The Scholar Who Does Not Take Himself Too Seriously

GUILLAUME CABANAC
Department of Computer Science, IRIT UMR 5505 CNRS, University of Toulouse, France

Picture yourself skimming through the latest table of contents of a major journal in your field. One single-authored paper stands out. It's not only because it's a contribution by a prominent researcher in the field. You're intrigued by the title: "A Hirsch-type index of co-author partnership ability" (Schubert, 2012). It suggests a *tour de force*: the shaping of the partnership φ-index, an indicator assessing the vitality of an academic's social ties *via* sustained co-authorships … based on the h-index (Hirsch, 2005), a flagship indicator of individual performance.

This was the first time I came across András's powerful, concise, elegant, and sometimes tongue-in-cheek writing. A concluding remark in his paper invited others to "upscale the present 'test tube' study to larger samples" (p. 308)—a playful nod to his background as a doctor of chemistry. András's scientific record is impressive: not only because of his contributions to various areas of science, but also for his ability to sustain successful partnerships with his "φ-core" closest partners: Wolfgang Glänzel, Tibor Braun, Sándor Zsindely, András Telcs, and Bart Thijs.

How can one represent the lifelong touch of a scholar? Well, let's try to shape it by looking at the words he has used to convey ideas throughout his career! For technical reasons, we should restrain our search to his published materials—his clarinet gigs are out of the scope of this study. I have not been able to retrieve the full texts of all of his 200+ publications[1] dating back from 1972—to do so I would need to subscribe to dozens of publishers and it would raise the cost of this modest study. I also purposely overlooked his contributions to youth literature (Schubert 2010, 2012) not to include words like 'cardigan' and 'knitting' in the corpus. Eventually, I gathered a non-random sample of 72 research papers András (co)-authored in English from 1981 onwards.

[1] See http://j.mp/AndrásSchubertBibliography

Iramuteq (Ratinaud, 2009) and Gephi (Bastian, Heymann & Jacomy, 2009) were used to analyse and visualise their textual contents once extracted from PDFs and manually cleaned to remove the headers and the bibliographic sections. The selected 72 papers were published in *Scientometrics*, the *Journal of Information Science*, the *Journal of Informetrics*, *Physica A*, the *Journal of Radioanalytical*

Figure 1 Word cloud of a non-random sample of 72 papers by András and colleagues (1981-2015).

and Nuclear Chemistry, *Trends in Analytical Chemistry*, and the *Journal of Chemical Information and Computing Sciences*, just to name a few.

András's main interests are illustrated in Figure 1, as reflected by word frequencies. It is difficult to distinguish this "word cloud" from one we would produce with the 100+ volumes of *Scientometrics*, a journal he has been serving as Editor for decades. The astute reader will spot, however, some specificities related to (but not restricted to) *chemistry* and *fullerenes* (see, e.g., Braun, Schubert, & Kostoff, 2002) or the *Waring* process modelling the distribution of authors by publication productivity (see, e.g., Schubert & Glänzel, 1984).

Classifying the co-occurring keywords of his publications shows the five topics depicted in Figure 2. András addressed research issues related to the communication of science and the peer review process (class 1). The SCI and ESI provided by the ISI[2] were core resources used in his contributions in scientometrics (class 2). András and his Budapest colleagues were pioneers in the area of geographic evaluative scientometrics and they assessed the relative productivity of countries (class 3). As a trained chemist, he contributed to the study of the research front of fullerene molecules with Tibor Braun (class 4). Overall, most of his papers showcased mathematical methods and formal models of the problems under study (class 5).

In a joint collaboration with his oldest son Gábor, the Schuberts sought 'bisociations' in the titles of *Inorganica Chimica Acta* to detect emergent areas in inorganic chemistry. These reveal unprecedented associations of words, such as "Dinuclear—Stereochemis-

[2] The *Science Citation Index* and the *Essential Science Indicators* were developed at the *Institute for Scientific Information*—now distributed by Thomson-Reuters.

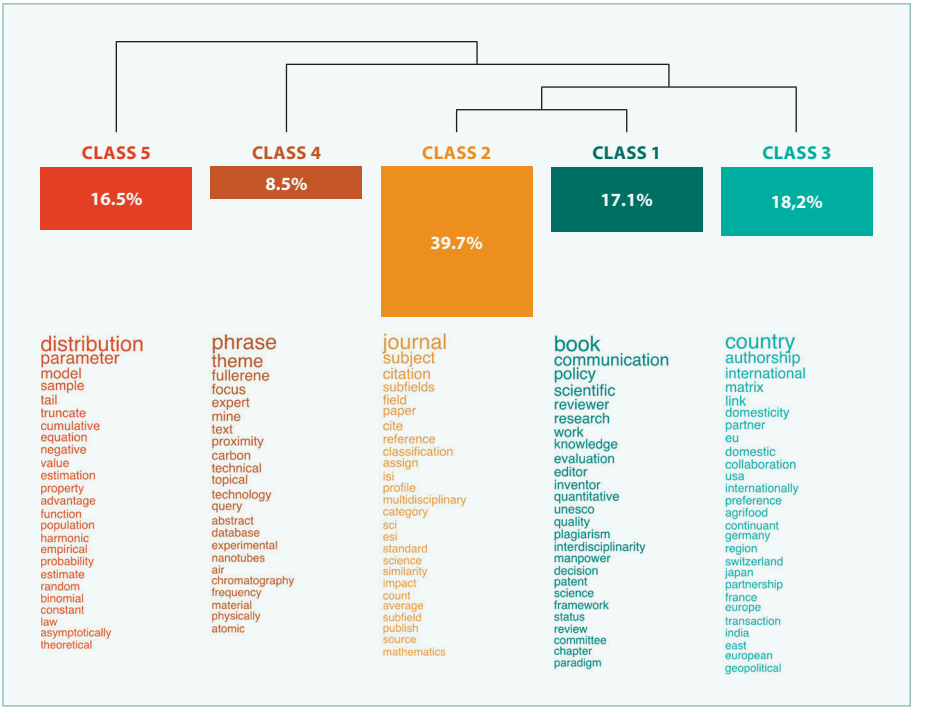


Figure 2 Descending classification of the corpus of András's 72 selected papers (1981-2015).

try" (Schubert & Schubert, 1997). Figure 3 shows a graph of the most frequently co-occurring words in András's papers, with detected communities emphasised in colour. Expanding this graph would reveal the bisociation "partnership—index" coined by András. How many other such creative bisociations are waiting for a prince to awake them by reusing these concepts, in line with van Raan's (2004) sleeping beauties?

Now back to András's invitation to upscale the "test-tube study" of the partnership φ-index (Schubert, 2012). This compelling allusion—at least from a computer scientist's perspective—led me to work on a validation of his model of φ with a database of a million researchers' bibliographic records. Here I should thank Tibor Braun, my 50-year older (to the day!) mentor who suggested that I liaise with András in the first place. This I did, and that's how I began corresponding with a remarkably creative scientist, dedicated gatekeeper, and witty gentleman. I remember the frightened face of my colleagues as he revealed the title of his talk before the Department of Sociology in Toulouse in 2013: "20 years, 20 papers in 20 minutes." Only to find he was joking and winking at his wife Zsuzsi who kept knitting while smiling at the back of the auditorium! I should have remembered the late Manfred Bonitz's comment as András received the 1993 Derek de Solla Price Award: "He thinks of his scientific activities as fun, a creative game played according to definite rules" (Bonitz, 1993).

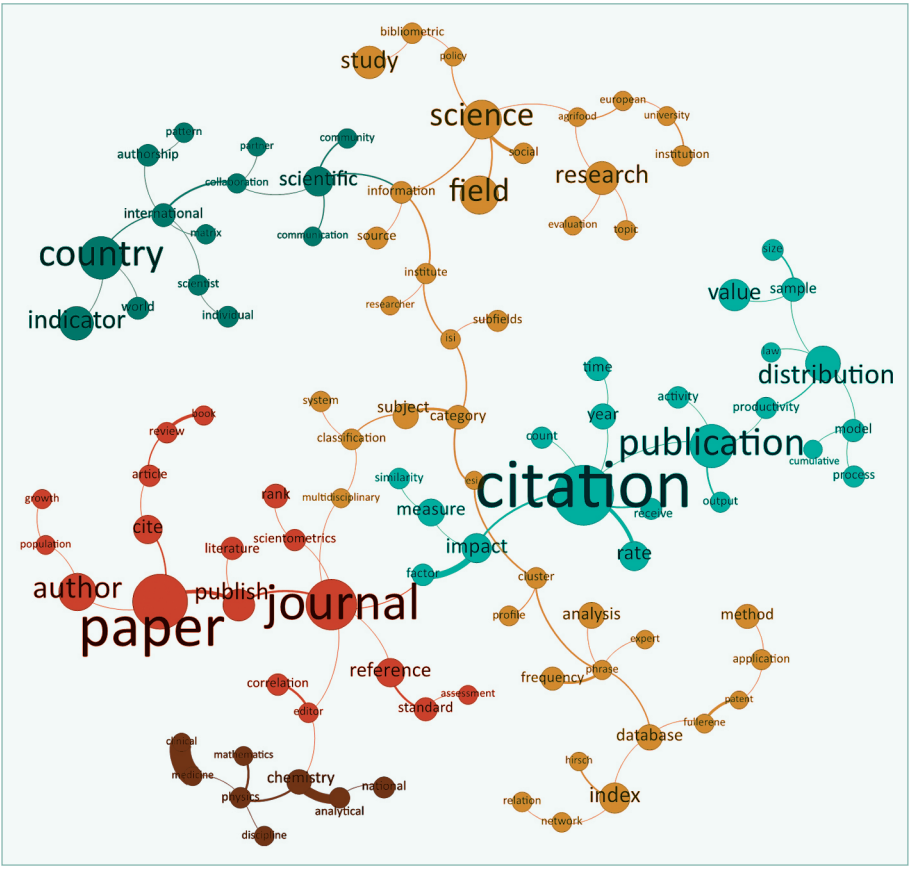


Figure 3 Graph of frequently co-occuring words in the corpus of András's 72 selected papers.

András has the habit of collecting sentences to remember from various sources, like plays, movies, jokes, and even *Scientometrics* (Schubert, 2014). I once asked him what had been the key to his success. "I don't take myself too seriously" he answered, almost instantly. What a wise reflection to remember!

« joyeux anniversaire »
„boldog születésnapot"
happy birthday
András!

References

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In *ICWSM'09: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

Bonitz, M. (1993). Comments on András Schubert, Recipient of the 1993 Derek de Solla Price Award. *Scientometrics*, *23*(3), 233–235. doi:10.1007/bf02026509

Braun, T., Schubert, A., Kostoff, R. N. (2002). A Chemistry Field in Search of Applications Statistical Analysis of U.S. Fullerene Patents. *Journal of Chemical Information and Computer Sciences*, *42*(5), 1011–1015. doi:10.1021/ci0200117

van Raan, A. F. J. (2004). Sleeping Beauties in Science. *Scientometrics*, *59*(3), 467–472. doi:10.1023/b:scie.0000018543.82441.f1

Ratinaud, P. (2009). IRaMuTeQ : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. Retrieved from http://www.iramuteq.org

Schubert, A. (2014). Sentences to remember from the first 100 volumes of the journal Scientometrics. *Scientometrics*, *100*(1), 1–13. doi:10.1007/s11192-014-1282-y

Schubert, A. (2013). *Fummelflausch Jacke und Ihre Freunde: Ein Märchenroman Für Kinder Und Erwachsene, Die Im Herzen Kinder Geblieben Sind*. Neckenmarkt: United P.C. [German translation of (Schubert, 2010)]

Schubert, A. (2012). A Hirsch-type index of co-author partnership ability. *Scientometrics*, *91*(1), 303–308. doi:10.1007/s11192-011-0559-7

Schubert, A. (2010). *Szöszmösz kardigán és barátai*. Budapest: Partvonal Kiadó.

Schubert, A., & Glänzel, W. (1984). A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, *6*(3), 149–167. doi:10.1007/bf02016759

Schubert, A. P., & Schubert, G. A. (1997). Inorganica Chimica Acta: Its publications, references and citations. An update for 1995–1996, *Inorganica Chimica Acta*, *266*(2), 125–133. doi:10.1016/S0020-1693(97)05910-0

Referenced Publication Year Spectroscopy (RPYS) and Algorithmic Historiography: The Bibliometric Reconstruction of András Schubert's Œuvre

LOET LEYDESDORFF[*a], LUTZ BORNMANN[b], JORDAN COMINS[c], WERNER MARX[d] & ANDREAS THOR[e]

[a*] *corresponding author; University of Amsterdam, Amsterdam School of Communication Research (ASCoR), PO Box 15793, 1001 NG Amsterdam, The Netherlands; loet@leydesdorff.net*

[b] *Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany; bornmann@gv.mpg.de*

[c] *Center for Applied Information Science, Virginia Tech Applied Research Corporation, Arlington, VA, United States; jcomins@gmail.com*

[d] *Max Planck Institute for Solid State Research, Information Service, Heisenbergstrasse 1, 70506 Stuttgart, Germany; w.marx@fkf.mpg.de*

[e] *University of Applied Sciences for Telecommunications Leipzig, Gustav-Freytag-Str. 43-45, 04277 Leipzig, Germany; thor@hft-leipzig.de*

Abstract: Referenced Publication Year Spectroscopy (RPYS) was recently introduced as a method to analyze the historical roots of research fields and groups or institutions. RPYS maps the distribution of the publication years of the cited references in a document set. In this study, we apply this methodology to the œuvre of an individual researcher on the occasion of a *Festschrift* for András Schubert's 70th birthday. We discuss the different options of RPYS in relation to one another (e.g. Multi-RPYS), and in relation to the longer-term research program of algorithmic historiography (e.g., *HistCite*™) based on Schubert's publications (n=172) and cited references therein as a bibliographic domain in scientometrics. Main path analysis and Multi-

RPYS of the citation network are used to show the changes and continuities in Schubert's intellectual career. Diachronic and static decomposition of a document set can lead to different results, while the analytically distinguishable lines of research may overlap and interact over time, and intermittently.
Keywords: RPYS, HistCite™, algorithmic historiography, main path, citation network

Introduction

In different compositions, the five of us have worked for the past two years on developing Referenced Publication Year Spectroscopy or—abbreviated—RPYS. RPYS is a bibliometric method which can be used to analyze the historical origins of research fields or researchers. This method analyzes the cited references (CR) and especially the referenced publication years of a publication set. The field CR in the Science Citation Index and the other databases at the Web of Science (WoS) contain a number of subfields separated by commas: the name of the first author, publication year, the abbreviated journal title, volume and page numbers, and increasingly also the doi (digital object identifier) of the cited document. In the online version (*SciSearch)* of the Science Citation Index at STN,[1] one can use these subfields for searching and retrieval (Marx, 2011; cf. Leydesdorff & Goldstone, 2014).

The first demonstration of RPYS as a method (Marx *et al*., 2014) was based on *SciSearch* at STN. In order to develop software for thus analyzing downloads from WoS, Lutz Bornmann linked up with Loet Leydesdorff, who extended his already existing software packages for bibliometric coupling[2] to this end (Leydesdorff *et al*., 2014). Andreas Thor further developed the program RPYS.exe (available at http://www.leydesdorff.net/software/rpys) into the Cited References Explorer (at http://crexplorer.net; Thor *et al*., in press). CRExplorer not only allows for RPYS, but also includes a tool for the disambiguation of misspelled references. Comins & Hussey (2015a and b; Comins & Leydesdorff, in press) further developed RPYS into a tool for Multi-RPYS (available at http://comins.leydesdorff.net). The occasion of a *Festschrift* for András Schubert's 70th birthday provides us with an opportunity to discuss the different options for RPYS in relation to the longer-term research program of algorithmic historiography—formulated by Garfield *et al.* (1964)—using Schubert's publications and citations as a bibliographic domain.

Garfield and Pudovkin further developed HistCite™ for the graphical user interfaces provided on both Windows and Apple computers in the late 1990s (Garfield *et al*., 2003; cf. Leydesdorff, 2010). The new version of HistCite™ (available at http://interest.science.thomsonreuters.com/forms/HistCite/) allows also for exporting the cita-

---

[1] STN (or Science and Technology Networks) is a fee-based host of databases maintained by the American Chemical Society.

[2] The program BibJourn.exe (available at http://www.leydesdorff.net/software/bibjourn) uses the subfield of the abbreviated journal name for mapping the knowledge bases of document sets (e.g., Leydesdorff & Goldstone, 2014).

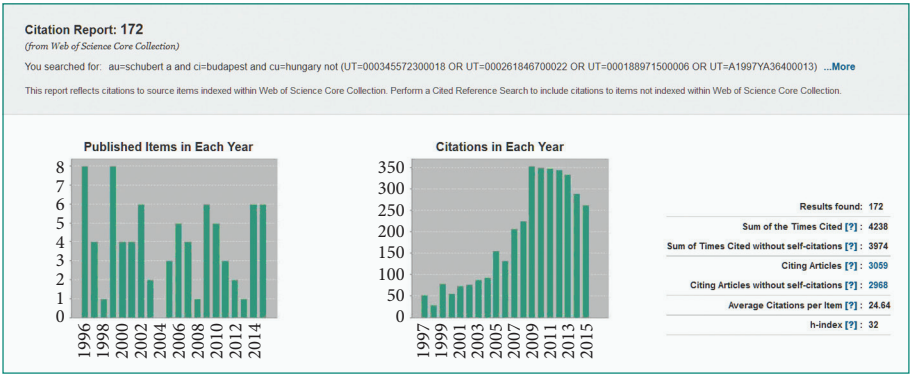


Figure 1: The Web of Science Citation Report for the 172 journal articles of András Schubert (January 18, 2016).

tion network into the Pajek format for social network analysis.[3] Hummon & Doreian (1989; Carley *et al*., 1993) first developed "main path analysis" that was integrated into Pajek in the 1990s. We will also pay attention to CitNetExplorer made available (at http://www.citnetexplorer.nl/) by researchers at the Center for Science and Technology Studies CWTS in Leiden (van Eck & Waltman, 2014) for citation network analysis.

Data

Searching for "AU = Schubert A and CI = Budapest", one retrieves 176 documents within the WoS domain of the Science and Social Science Citation Indices. Four of these documents are false positives (of Alfred Schubert).[4] We use the remaining 172 publications as our domain, downloaded on January 4, 2016.[5]

The WoS Citation Report in Figure 1 shows the publication and citation pattern of this set during the last twenty years. The legends show, among other things, that Schubert's papers are *on average* cited more than 24 times.

Figure 2 extends the graphs for the entire period 1972-2015. It shows the annual numbers of publications, citations, and cited references. As can be expected for a single author, publication and citation patterns fluctuate strongly over the entire period (if only for reasons of chance). Yet, both trends are upward as the dotted (regression) lines in Figure 2 reveal; there is a peak for publications in 1989 (n = 13) and for citations in 2002 (n = 892). Referencing is highest during the second half of the 1980s—the years

[3] Pajek is a program for network analysis and visualization, freely available for non-commercial purposes at http://mrvar.fdv.uni-lj.si/pajek/ .

[4] Four more papers can be added if conference proceedings are also taken into account; seven more documents were published during his doctorate period at the Physics Department of the University for Agricultural Sciences in Gödöllo. We are grateful to Wolfgang Glänzel for noting these corrections.

[5] Among these papers 20 are bibliographies and two meeting abstracts.

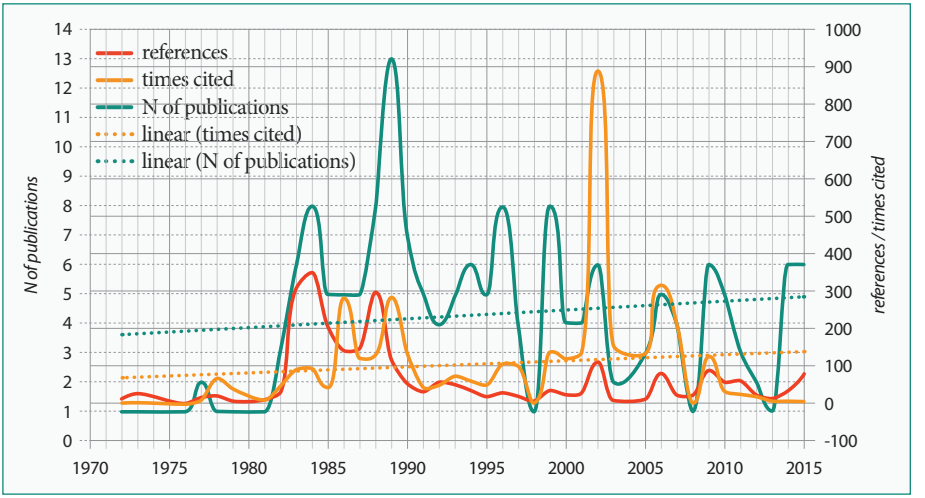


Figure 2: Publication, citation, and cited reference profiles of András Schubert, 1972-2015.

of the establishment of the Information Science and Scientometrics Research Unit (ISSRU) in Budapest in collaboration with Tibor Braun and Wolfgang Glänzel. The total number of references by the 172 publications is 2,715; that is, 15.8 references per publication on average. Citation peaks in 2002 with 892 citations in WoS during that year. This peak is largely due to Schubert's coauthorship of a single publication (Barabási *et al.*, 2002) that has been cited 784 times.

Algorithmic historiography

a. HistCite

As mentioned above, Eugene Garfield's original program for algorithmic historiography was revived and further elaborated by Alexander Pudovkin when graphical interfaces became available on Windows computers in the late 1990s (Garfield *et al.*, 2003). HistCite™ is nowadays available upon registration at http://interest.science.thomsonreuters.com/forms/HistCite/.[6]

Figure 3 shows the HistCite network based on the so-called "Local Citation Scores" *within* the publication set of András Schubert. An alternative representation can be obtained by using the Global Citation Scores which are based on the times-cited scores in the input file.

[6] Using HistCite, the header of an input file—downloaded from WoS—needs to be changed from "FN Thomson Reuters Web of Science™" (the current header) into "FN ISI Export Format" (the old format) before HistCite can read the file. Under Microsoft Windows, HistCite requires the presence of the Internet Explorer. The input has to be saved as ASCII/ANSI.

Since all input records are (co-)authored by Schubert, this figure shows the top-30 layer (n = 30) in his œuvre.[7] Self-citations to papers from the period 1983-1993 are prevalent in the set.

Table 1: Thirty papers selected by HistCite as the local citation network within Schubert's œuvre. (LCS: Local Citation Score within this network; GCS: Global Citation Score using times-cited values).

Nr in Fig. 2	Cited Reference	LCS	GCS
2	NOSZTICZ.Z, 1973, PERIOD POLYTECH CHEM, V17, P165	2	4
9	ZSINDELY S, 1982, SCIENTOMETRICS, V4, P57	4	26
10	ZSINDELY S, 1982, SCIENTOMETRICS, V4, P69	2	21
12	SCHUBERT A, 1983, SCIENTOMETRICS, V5, P59	6	62
18	SCHUBERT A, 1984, J RADIOANAL NUCL CH, V82, P215	7	9
20	SCHUBERT A, 1984, SCIENTOMETRICS, V6, P149	9	33
25	GLANZEL W, 1984, Z WAHRSCHEINLICHKEIT, V66, P173	8	33
26	TELCS A, 1985, MATH SOC SCI, V10, P169	4	10
31	SCHUBERT A, 1986, CZECH J PHYS, V36, P121	2	27
32	SCHUBERT A, 1986, CZECH J PHYS, V36, P126	4	21
33	SCHUBERT A, 1986, SCIENTOMETRICS, V9, P231	3	18
34	SCHUBERT A, 1986, SCIENTOMETRICS, V9, P281	16	215
36	BRAUN T, 1987, SCIENTOMETRICS, V11, P9	10	30
37	BRAUN T, 1987, SCIENTOMETRICS, V11, P127	9	24
38	BRAUN T, 1987, SCIENTOMETRICS, V12, P3	9	22
41	GLANZEL W, 1988, J INFORM SCI, V14, P123	5	37
45	BRAUN T, 1988, SCIENTOMETRICS, V13, P181	10	43
46	BRAUN T, 1988, SCIENTOMETRICS, V14, P3	9	28
47	BRAUN T, 1988, SCIENTOMETRICS, V14, P365	8	18
51	SCHUBERT A, 1989, J AM SOC INFORM SCI, V40, P291	4	12
52	BRAUN T, 1989, SCIENTOMETRICS, V15, P13	3	6
54	BRAUN T, 1989, SCIENTOMETRICS, V15, P325	4	21
55	SCHUBERT A, 1989, SCIENTOMETRICS, V16, P3	19	186
60	BRAUN T, 1989, TRAC-TREND ANAL CHEM, V8, P281	4	14
61	BRAUN T, 1989, TRAC-TREND ANAL CHEM, V8, P316	3	7
66	SCHUBERT A, 1990, SCIENTOMETRICS, V19, P3	6	108
69	BRAUN T, 1991, SCIENTOMETRICS, V20, P9	2	6
70	SCHUBERT A, 1991, SCIENTOMETRICS, V20, P317	2	12
74	SCHUBERT A, 1992, SCIENTOMETRICS, V23, P3	2	13
81	BRAUN T, 1993, SCIENTOMETRICS, V28, P137	8	20

[7] See for further explanation of the definitions in HistCite at http://garfield.library.upenn.edu/histcomp/guide.html.

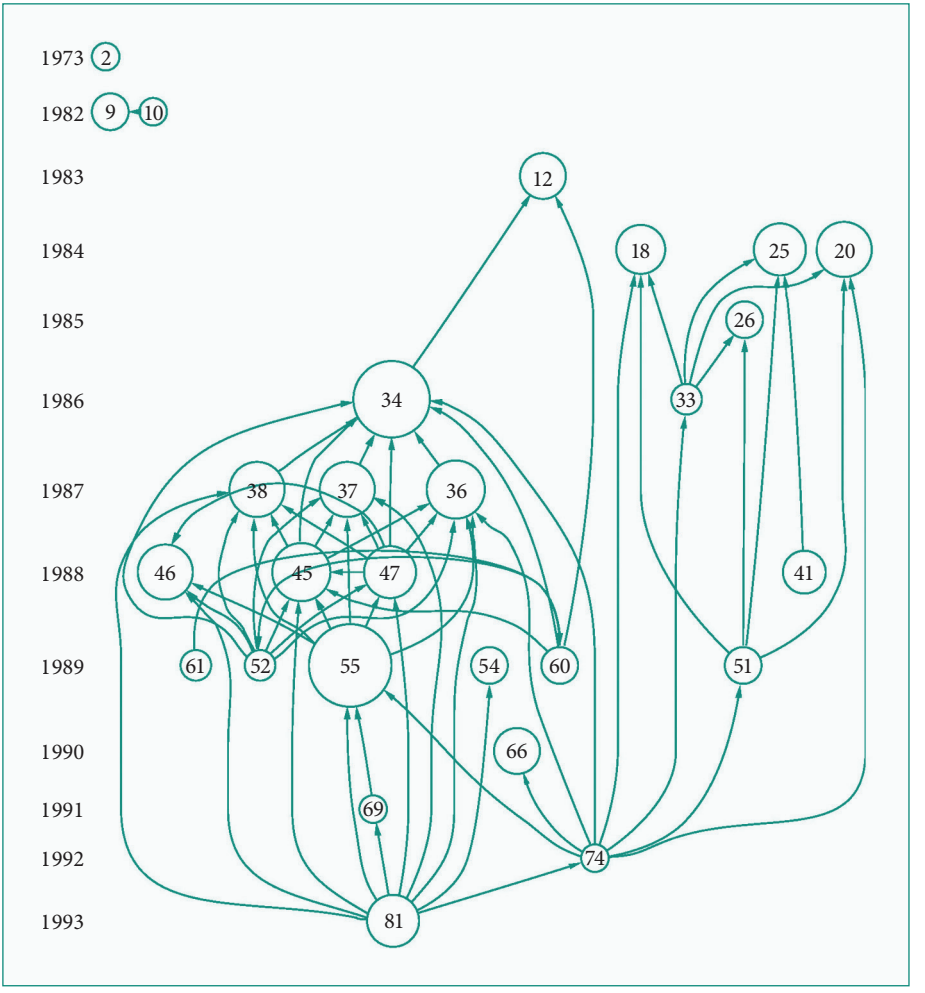


Figure 3: Default output of HistCite on the basis of 172 documents authored by András Schubert. The figure shows the top layer (n = 30) in the internal ("local") citation structure of his œuvre.

HistCite provides a legend in a separate (split) screen (Table 1). Node 34, for example, is self-cited eight times; this paper, coauthored with Tibor Braun (Schubert & Braun, 1986), seems to have been constitutive for the research program thereafter.

HistCite can also be used to generate a complete citation network by setting the limit above the size of the set under study (instead of the 30 which are the default for the graph in HistCite, for example, 172 in our case). This network is exported in the Pajek (.net) format that can be used in many network analysis and visualization programs such as UCInet, Gephi, and VOSviewer. Pajek furthermore offers the option to study the main path in the network.

b. Analysis and visualization of the citation network

The network file exported from HistCite contains the 172 documents as nodes and the citation relations among them as links; 95 nodes are linked into a largest component. This largest component can be visualized as a citation network (Figure 4). By choosing the layout of Fruchterman & Reingold (1992), we can observe the two constitutive clusters of the ISSRU research program to the left in the bottom half. One cluster is dominated by papers with Tibor Braun as first author and the other by papers with András Schubert as first author. Wolfgang Glänzel joined the Budapest group first as a PhD student and then became a third (co-)author in the second half of the 1980s. Most of the papers are coauthored by at least two of these three authors.

At the top right of Figure 4, one can see that the recent work of Schubert (since 2005) is only weakly related to earlier work in terms of citation relations; references to papers coauthored with Glänzel as lead author provide the relationship with evaluation studies. In 2005, Jorge Hirsch published his study of the *h*-index which opened a whole new set of questions for bibliometric investigation. Thirteen of the 44 papers in the period 2005-2015 (that is, 30%) contain the words "h-index" or "Hirsch" in the *title*. Within this cluster of most recent papers, Tibor Braun is the lead author in two cases.

One advantage of network analysis and visualization programs is the availability of algorithms for the decomposition and further statistics, whereas HistCite™ has remained descriptive. In Figure 4, for example, seven clusters were distinguished by using the decomposition algorithm of Blondel *et al.* (2008). The modularity *Q*—a measure for the dividedness between 0 and 1—of the network is 0.578. Thus, the clusters are weakly distinct. Similarly, one can feed the Pajek file into VOSviewer and obtain a comparable network. The algorithm then reveals a finer distinction of 11 clusters in

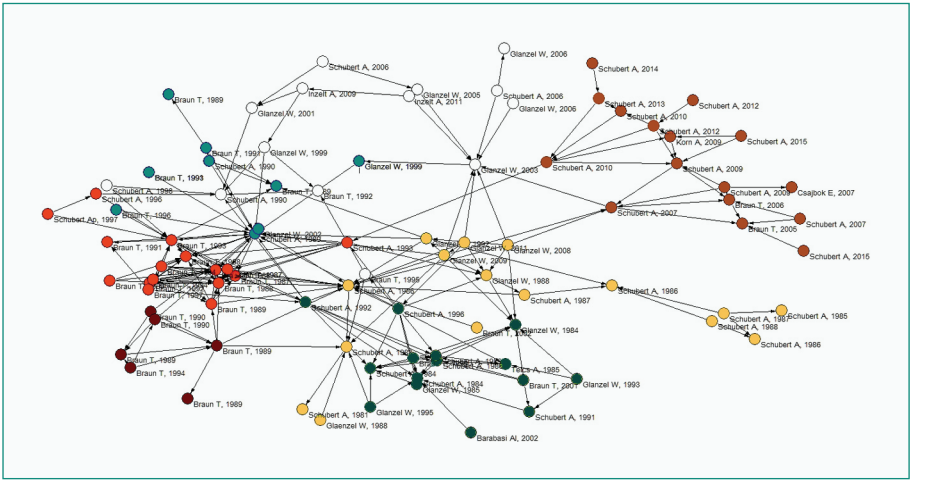


Figure 4: Seven clusters in the main component of the citation matrix (n = 95), distinguished using Blondel et al. (2008) in Pajek; Fruchterman & Rheingold (1992) was used for the layout.

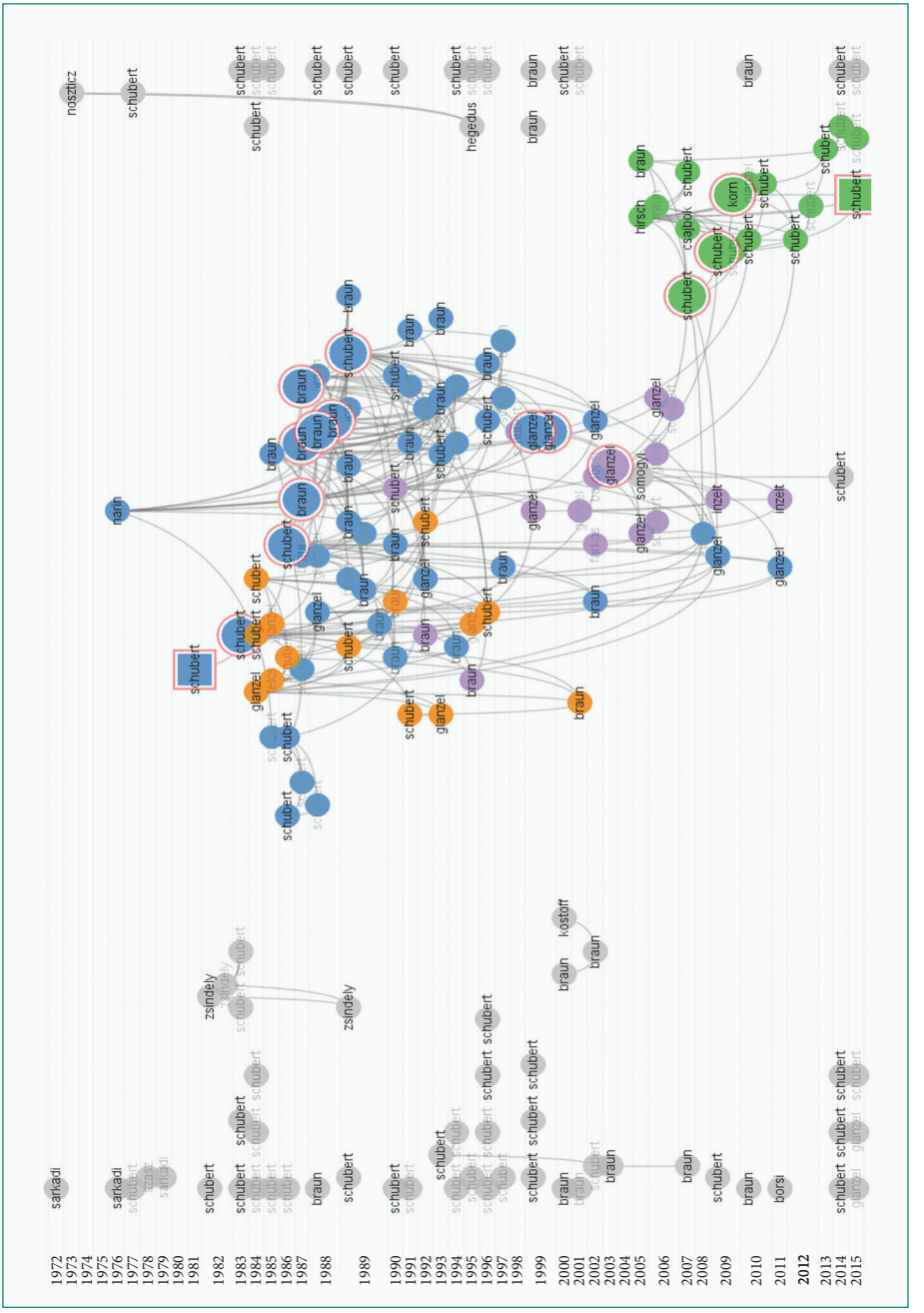


Figure 5: Clustering and indication of shortest path applying CitNetExplorer to the citation network of Schubert's œuvre (at http://www.citnetexplorer.nl/).

the large component, and one obtains other options for the visualization (not shown here). More specifically developed for citation network analysis is the program CitNetExplorer of the same group at the Center for Science and Technology Studies CWTS (van Eck & Waltman, 2014).

Figure 5 shows the results of feeding the original WoS download (n = 172) into CitNetExplorer. By default, the program analyzes only the articles with a times-cited score equal to or larger than ten. As against the default of making only 40 nodes (papers) visible, we chose to make all the remaining papers visible. This includes a number of papers which are not connected and therefore colored grey in Figure 5.

The clustering algorithm of CitNetExplorer distinguishes four main groups with a minimum size of ten. One of them is the recent group of papers (colored green) and is discrete from the other three which are more mixed. Although one can distinguish the Braun-dominated cluster from the Schubert-dominated one during the late 1980s and 1990s, the division is fuzzy. The third group in the first decade of the 2000s is dominated by Glänzel's papers (lilac). The visualization of CitNetExplorer not only labels with the citing papers, but includes the cited first authors; for example, Hirsch (2005).[8]

Within CitNetExplorer, the analyst can mark two nodes and ask for the shortest path between them. In Figure 5, we marked Schubert & Braun (1981) as the first paper in the common cluster, and Schubert (2015) at the bottom as the last paper. These two nodes are marked on the map with (orange) squares. More than a single shortest path (in six steps) was reported in this data. The papers on a shortest path are indicated with orange circles around the nodes.

c. Main path analysis in Pajek

Unlike the shortest path between two nodes selected by the analyst, the main path is defined as a systemic property. Citing previous literature and being cited by subsequent literature position a paper in relation to other papers in the set (Hummon & Doreian, 1989). When a set of documents represents a self-contained field—not significantly building on knowledge from other fields—the citation network among the key documents (the most highly cited ones) can be expected to contain at least one main path (Carley *et al*., 1993). Main-path algorithms enable us to make the structural backbone of a literature visible (Lucio-Arias & Leydesdorff, 2008).

The main path is reconstructed by calculating the connectivity of the links in terms of their degree centrality and outlining the path formed by the nodes with the highest degree. In terms of a citation network, this degree measure considers the number of citations a document receives (indegree) as well as the number of cited references in the documents (outdegree). The main path is constructed by selecting those connected documents with the highest scores until an end document is reached (Batagelj, 2003). This can be either a document that is no longer cited or one that contains no further references within the set.

---

[8] Insofar as the cited references are to citing papers in the set, the title-field is imported into the documentation of the visualization.

Figure 6: Main path in the citation network of András Schubert's œuvre. Blondel et al. (1998) was used for the decomposition and Kamada & Kawai (1989) for the layout.

The main path shown in Figure 6 can be extracted in Pajek as a partition from the citation network. Although we did not add the years as a temporal dimension to the documents (as in the shortest path analyzed above), the algorithm itself sorts the references along a time line. Using Blondel's *et al.* (1998) algorithm for the decomposition, four clusters are robustly indicated ($Q = 0.588$).[9] The first cluster (yellow) shows the initial period of institutionalization of the ISSRU unit and the journal *Scientometrics* during the 1980s. The second period represents the 1990s; the third (red) period begins after Glänzel left the unit for Louvain in 2002. Schubert himself, however, begins new research lines since 2010. These latter papers are all first-authored by him, whereas in the previous periods coauthorship with Glänzel was also common on the main path.

Note that these are distinctions within the construct of the main path. They inform us about the network structure of citation relations, potentially including relations among different research lines. We refrain from rationalizing the transitions indicated in Figure 6 in terms of intellectual changes, but return to this issue more extensively in the discussion section.

d. RPYS

RPYS plots the cumulative distribution of cited references in terms of the referenced publication years. The peaks in the graph are often discrete and thus indicate specific publications which were highly cited within the domain of the sample. But this is not the case at the research front—that is, the most recent years—because the citation classics are not yet sorted out in that part of the domain (Price, 1970). Baumgartner & Leydesdorff (2014) distinguish between transient knowledge claims at the research front and sticky ones which

[9] We formulate "robustly" because this analysis can be repeated.

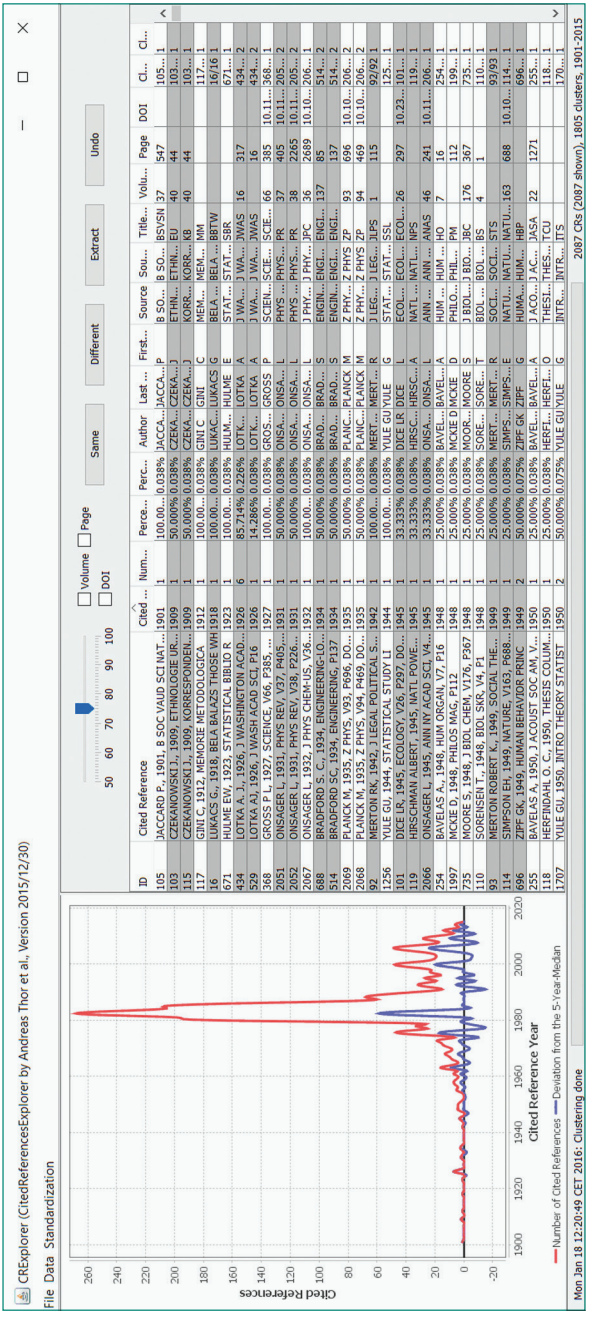


Figure 7: User interface of CRExplorer after importing the œuvre of Schubert.

remain highly cited after more than ten years. One can also consider the latter citations as concept-symbols (Small, 1978) and the former as citation currency.

Figure 8 shows the results of using CRExplorer for RPYS. The red line visualizes the number of cited references per referenced publication year during the period 1900-2015. In order to identify those publication years with significantly more cited references than other years, the deviation of the number of cited references in each year from the median of the number of cited references in the two previous, the current, and the two following years (t—2; t—1; t ; t + 1; t + 2) is visualized as a blue line. This deviation from the five-year median provides a smoother curve than one in terms of absolute numbers.

The disadvantage of the figure in the left pane is the possibility that several papers may be adding up to a peak in a specific year. Inspection of the listing in the right pane teaches us that the first peak in the figure to the publication year 1926 points to Lotka (1926), which is a citation classic in this field; but the 1963 peak, for example, is composed of several classics: Price

(1963), Irwin (1963), and Kessler (1963), cited four, three, and three times, respectively. Furthermore, Lotka (1926) is cited six times as "LOTKA A. J., 1926, J WASHINGTON ACAD SC, V16, P317", but also once as "LOTKA AJ, 1926, J WASH ACAD SCI, P16".

Although Thomson Reuters standardizes the cited references of papers included in WoS, the problem of variants of the same cited references remains, potentially disturbing the results of RPYS and citation analysis more generally. If cited references are available with several variants, it is no longer possible to produce a reliable list or ranking of the most frequently cited publications. Evaluation studies are very susceptible to this type of error. The problem of variants is especially urgent for document types other than journal papers (such as books and book chapters). Can the cited references be disambiguated?

CRExplorer offers the possibility to cluster the variants of cited references. A detailed description of the clustering and merging methods used in the program can be found in Thor *et al*. (2016, in press). After a first round of automatic cleaning, one can proceed with manual cleaning. Since the automatic clustering of variants can also be a source of error, one is advised to check and possibly correct the results of the automatic clustering manually. Note that not all errors can be corrected because references may be incomplete (Leydesdorff, 2008: 285, Table 4).

Table 2: After disambiguation (CRExplorer), Glänzel (1988) is added to the publications referenced more than five times in the set; Narin (1976) and Braun (1987) are ranked at a higher position.

CR	LCS
BRAUN T, 1985, SCIENTOMETRIC INDICA	24
SCHUBERT A, 1989, SCIENTOMETRICS, V16, P3	19
SCHUBERT A, 1986, SCIENTOMETRICS, V9, P281	16
HIRSCH JE, 2005, P NATL ACAD SCI USA, V102, P16569	13
BRAUN T, 1987, SCIENTOMETRICS, V11, P9	10
BRAUN T, 1988, SCIENTOMETRICS, V13, P181	10
NARIN F., 1976, EVALUATIVE BIBLIOMET	**10**
BRAUN T, 1987, LIT ANAL CHEM SCIENT	**9**
BRAUN T, 1987, SCIENTOMETRICS, V11, P127	9
BRAUN T, 1987, SCIENTOMETRICS, V12, P3	9
BRAUN T, 1988, SCIENTOMETRICS, V14, P3	9
SCHUBERT A, 1984, SCIENTOMETRICS, V6, P149	9
BRAUN T, 1988, SCIENTOMETRICS, V14, P365	8
BRAUN T, 1993, SCIENTOMETRICS, V28, P137	8
GLANZEL W, 1984, Z WAHRSCHEINLICHKEIT, V66, P173	8
GLANZEL W, 2003, SCIENTOMETRICS, V56, P357	8
GARFIELD E, 1972, SCIENCE, V178, P471	7
PRICE DJD, 1976, J AM SOC INFORM SCI, V27, P292	7

SCHUBERT A, 1984, J RADIOANAL NUCL CH, V82, P215	7
GLANZEL W, 1988, J INFORM SCI, V14, P123	**6**
HAITUN SD, 1982, SCIENTOMETRICS, V4	6
HAITUN SD, 1982, SCIENTOMETRICS, V4, P5	6
HAITUN SD, 1982, SCIENTOMETRICS, V4, P89	6
IRWIN JO, 1975, J ROY STAT SOC A STA, V138, P18	6
IRWIN JO, 1975, J ROY STAT SOC A STA, V138, P204	6
IRWIN JO, 1975, J ROY STAT SOC A STA, V138, P374	6
LOTKA A. J., 1926, J WASHINGTON ACAD SC, V16, P317	6
SCHUBERT A, 1983, SCIENTOMETRICS, V5, P59	6
SCHUBERT A, 1990, SCIENTOMETRICS, V19, P3	6
TAGUE J, 1981, J AM SOC INFORM SCI, V32, P280	6

Table 2 lists the publications referenced more than five times by András Schubert's publication set after careful (automatic and manual) clustering of the cited references using CRExplorer. Two publications change positions in the hierarchy, and one (Glänzel & Schubert, 1988) was added to the set of 29 publications referenced more than five times. Francis Narin's (1976) book on the use of bibliometrics in evaluation, for example, is referenced with four variants. It is cited ten instead of seven times in the publications of András Schubert after the disambiguation process.

e. Multi-RPYS

Multi-RPYS provides an extension of standard RPYS methodology and was developed to make possible comparative analysis among different years and/or different sets. This objective is accomplished by applying a rank-transformation to the standard RPYS outputs and by visualizing the results as a heat map. Multi-RPYS has hitherto been used to investigate (1) communal intellectual histories across different journals, and (2) the temporal dynamics of historical influences (Comins & Hussey, 2015a; Comins & Hussey, 2015b; Comins & Leydesdorff, in press). Specifically, this latter technique segments the set of citing articles by publication year and generates a Multi-RPYS heat map across these segments to track *when* and *how consistently* references were cited by researchers. Below we use this approach to consider shifts in the intellectual influences driving András Schubert's œuvre.

The largest peak in the RPYS plot of Schubert's works occurs in 1982 (see Figure 7), and is driven by Haitun's (1982) three papers about "Stationary Scientometric Distributions" published as different parts in *Scientometrics*. The band (B) in Figure 8 corresponds to 1982 as the referenced publication year. It shows that most citations to this year occurred from citing documents—chronologically sorted along the y-axis—published in the first half of Schubert's career. By splitting (in the lower part of the figure)

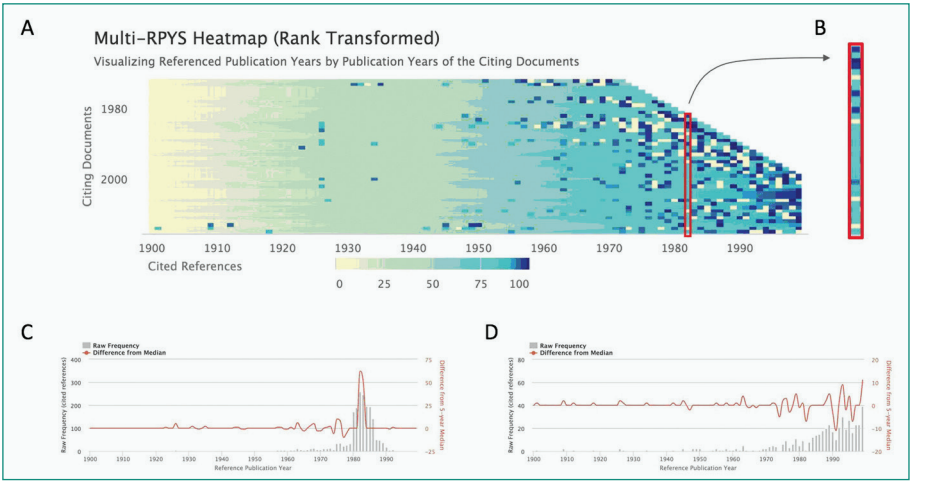


Figure 8. Multi-RPYS heatmap computing RPYS results for Schubert's œuvre segmented by publications year of the citing documents (along the y-axis).

the works of Schubert into those published from 1972-1993 (C) and 1994-2015 (D), the absence of 1982 as a peak reference year in the latter set becomes visible. In other words, Haitun's work was cited by Schubert only during the first part of his career.

RPYS and bibliographic coupling

The data used for RPYS and citation network studies (CR) can also be used for bibliographic coupling (Kessler, 1963). What is the difference? In citation network studies and RPYS, cited references across the sets under study are binned in years; in studies of bibliographic coupling one uses the citing documents as units of analysis. Using years, heterogeneous sets in terms of cognitive contents and social relations are potentially lumped together. Figure 9, for example, shows the clear structure that can instead be found in Schubert's œuvre when these same cited references are used for a map of the bibliographic coupling among the co-authors of Schubert.

We shall not discuss Figure 9 here; but show it in order to make the point that diachronic analysis and static analysis can lead to very different results. One cannot easily map the relations among 44+1 (co-)authors diachronically. Using a dynamic optimization among multi-dimensional scaling outputs for subsequent years, however, Leydesdorff & Schank (2008) have developed a version of visone[10] (visone v2.3.X at http://www.leydesdorff.net/visone) that allows for combining social and cognitive attributes of documents in animations (e.g., Leydesdorff, 2010a and b).

[10] Visone is a network analysis and visualization program, freely available at http://visone.info .

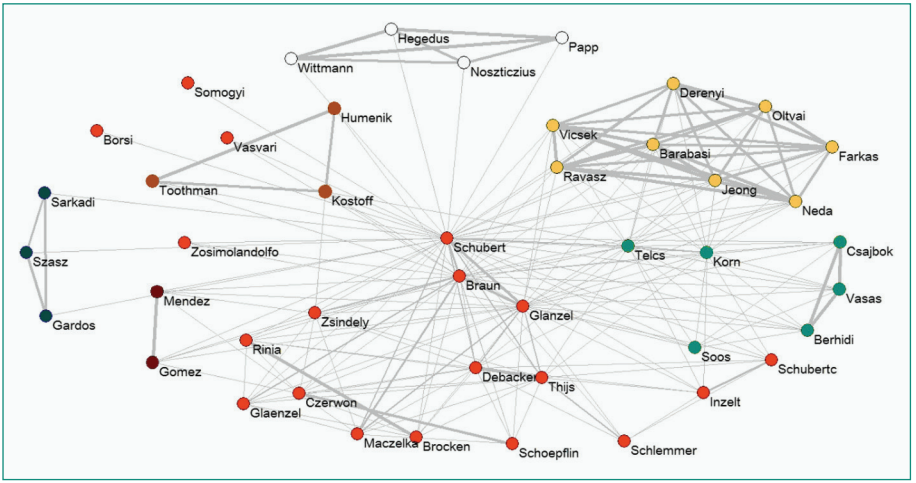


Figure 9. Bibliographic coupling of 44 co-authors of Schubert's 172 publications; seven clusters were distinguished by the algorithm of Blondel et al. (2008); Q = 0.639; Kamada & Kawai (1989) was used for the visualization; the output is cosine-normalized.

One disadvantage of focusing on cited references in terms of referenced publication years is the neglect of the knowledge content which structures citation networks in the development of the sciences. One risks studying the dynamics of citations instead of the dynamics of science. Combining the referenced publication years with the cited journals may provide a perspective for the further development of Multi-RPYS in a direction that will show the development of socio-cognitive structures in the data over time (cf. Leydesdorff & Goldstone, 2014).

Conclusions

RPYS is a recently introduced method for the study of the historical roots of research fields or researchers. It is based on the analysis of cited references and especially cited reference years. The occasion of a *Festschrift* for András Schubert's 70th birthday provides us with the opportunity to discuss the different options for RPYS in relation to the longer-term research program of algorithmic historiography using Schubert's publications and the references cited therein as a bibliographic domain. The results show that RPYS allows for the reconstruction of the shoulders on which a researcher stands. Without disambiguation, however, the CR field remains an unreliable source. Using it for evaluation purposes requires disambiguation. CRExplorer offers a partial solution to this problem.

The largest peak in the RPYS plot of Schubert's publications (which indicates the works with the largest influence on Schubert's research) occurs in 1982, and is driven by Haitun's (1982) three papers about "Stationary Scientometric Distributions". The results of Multi-RPYS revealed, however, that Haitun's papers were primarily refer-

enced during the first half of Schubert's career. These and further results in this study based on András Schubert's publications demonstrate that RPYS is a useful addition to the already available bibliometric techniques for algorithmic historiography (such as *HistCite*™, CitNetExplorer, visone, etc.).

Acknowledgement

We are grateful to Wolfgang Glänzel for his comments on an earlier version of this paper.

References

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications, 311*(3), 590-614.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 8*(10), 10008.

Carley, K. M., Hummon, N. P., & Harty, M. (1993). Scientific Influence An Analysis of the Main Path Structure in the Journal of Conflict Resolution. *Science Communication, 14*(4), 417-447.

Comins, J. A., & Hussey, T. W. (2015a). Compressing multiple scales of impact detection by Reference Publication Year Spectroscopy. *Journal of Informetrics, 9*(3), 449-454.

Comins, J. A., & Hussey, T. W. (2015b). Detecting seminal research contributions to the development and use of the global positioning system by reference publication year spectroscopy. *Scientometrics*, 1-6.

Comins, J. A., & Leydesdorff, L. (in preparation). Identification of long-term concept-symbols among citations: Can documents be clustered in terms of common intellectual histories? , http://arxiv.org/abs/1601.00288.

Comins, J.A. & Leydesdorff, L. (in press), RPYS i/o: A web-based tool for the historiography and visualization of citation classics, sleeping beauties, and research fronts, *Scientometrics* (in press).

Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-directed replacement. *Software—Practice and Experience, 21*, 1129-1166.

Garfield, E., Pudovkin, A. I., & Istomin, V. I. (2003). Mapping the Output of Topical Searches in the Web of Knowledge and the case of Watson-Crick. *Information Technology and Libraries, 22*(4), 183-187.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). The use of citation data in writing the history of science. Philadelphia, PA: Institute for Scientific Information.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science, 14*(2), 123-127.
Haitun, S.D. (1982a). Stationary scientometric distributions: Part I. Different Approximations, *Scientometrics*, *4*(1), 5-25.
Haitun, S.D. (1982b). Stationary scientometric distributions: Part II. Non-Gaussian nature of scientific activities. *Scientometrics, 4*(2), 89-104.
Haitun, S.D. (1982c). Stationary Scientometric Distributions : Part III. The Role of the Zipf Distribution. *Scientometrics, 4*(3), 181-194.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA, 102*(46), 16569-16572.
Irwin, J. O. (1963). The place of mathematics in medical and biological statistics. *Journal of the Royal Statistical Society. Series A (General), 126*, 1-45.
Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11*(1), 39-63.
Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31*(1), 7-15.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*, 10-25.
Leydesdorff, L. (2010a). Eugene Garfield and Algorithmic Historiography: Co-Words, Co-Authors, and Journal Names. *Annals of Library and Information Studies, 57*(3), 248-260.
Leydesdorff, L. (2010b). What Can Heterogeneity Add to the Scientometric Map? Steps towards algorithmic historiography. In M. Akrich, Y. Barthe, F. Muniesa & P. Mustar (Eds.), *Débordements: Mélanges offerts à Michel Callon* (pp. 283-289). Paris: École Nationale Supérieure des Mines, Presses des Mines.
Leydesdorff, L., Bornmann, L., Marx, W., & Milojević, S. (2014). Referenced Publication Years Spectroscopy applied to iMetrics: Scientometrics, Journal of Informetrics, and a relevant subset of JASIST. *Journal of Informetrics, 8*(1), 162-174.
Leydesdorff, L., & Goldstone, R. L. (2014). Interdisciplinarity at the Journal and Specialty Level: The changing knowledge bases of the journal *Cognitive Science*. *Journal of the Association for Information Science and Technology 65*(1), 164-177.
Leydesdorff, L., & Schank, T. (2008). Dynamic Animations of Journal Maps: Indicators of Structural Change and Interdisciplinary Developments. *Journal of the American Society for Information Science and Technology, 59*(11), 1810-1818.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy of Sciences, 16*(12), 317-323.
Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite™-based historiograms. *Journal of the American Society for Information Science and Technology, 59*(12), 1948-1962.
Marx, W. (2011). Special features of historical papers from the viewpoint of bibliometrics. *Journal of the American Society for Information Science and Technology, 62*(3), 433-439.
Marx, W., & Bornmann, L. (2014). Tracing the origin of a scientific legend by reference publication year spectroscopy (RPYS): the legend of the Darwin finches. *Scientometrics, 99*(3), 839-844.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology, 65*(4), 751-764.

Narin, F. (1976). *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Washington, DC: National Science Foundation.

Price, D.J. de Solla (1963). *Little Science, Big Science*. New York: Columbia University Press.

Schubert, A. (2015). Rescaling the h-index. *Scientometrics, 102*(2), 1647-1653.

Schubert, A. (2013). Measuring the similarity between the reference and citation distributions of journals. *Scientometrics, 96*(1), 305-313.

Schubert, A., & Braun, T. (1981). Some scientometric measures of publishing performance for 85 Hungarian research institutes. *Scientometrics, 3*(5), 379-388.

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics, 9*(5), 281-291.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science 8*(3), 113-122.

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (in press). Introducing CitedReferencesExplorer: A program for Reference Publication Year Spectroscopy with Cited References Disambiguation. *Journal of Informetrics* (in press). Preprint available at http://arxiv.org/abs/1601.01199 .

van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics, 8*(4), 802-823.

András' Contribution to Scientometrics

PAUL WOUTERS
Centre for Science and Technology Studies, Leiden University

I met András Schubert in the course of my study of the history of the Science Citation Index and of scientometrics as a field. Apart from the occasional meeting at conferences and workshops that I started to attend, what made the strongest impression on me was his accessiblity and energy during the interview in his office in Budapest. He seemed quite convinced of the possibility of the development of the field as a hard science about science. This he clearly shared with the founding editor of the journal *Scientometrics* and with Derke de Solla Price. András used the Price Index in his analysis (together with Hajnalka Maczelka) of the first ten years of the journal to verify empirically that the field was indeed moving towards this ideal state of knowledge which was published in the mostly qualitatively oriented journal *Social Studies of Science* (Schubert & Maczelka, 1993). Their conclusion was that the field had moved a bit into the harder direction, although it was still "pre-paradigmatic" and somewhere in between the hard and soft sciences. During the interview, he was clearly one of the pioneers who had the vision that a thorough and empirical analysis of scientific research by computerised means and large databases should not only be possible but should also be taken as the basis for a scientifically based science policy and research evaluation. With hindsight, it is clear that the interview inspired me to write one of the chapters in my historical thesis (for which the research was performed together with Loet Leydesdorff) (Wouters & Leydesdorff, 1994) about the development of the field of scientometrics, in which, inter alia, we concluded that the field displayed most characteristics of a social science, rather than an exact science.

So, at the occasion of the 70th birthday of András, I wondered what the overall contribution of this wonderful scientometrician has been to the journal to which he has devoted so much energy. It seemed fitting to do this in a scientometric way. So, instead of

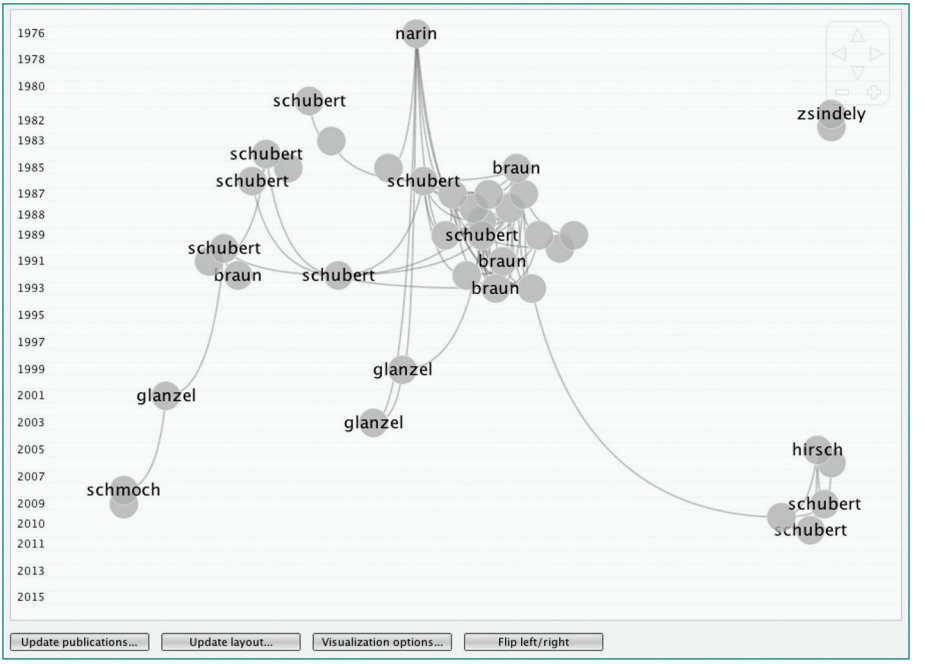


Figure 1: Overall historical development of Schubert in Scientometrics, produced with CitNetExplorer (http://www.citnetexplorer.nl) [threshold of 5 citations] (Eck & Waltman, 2014).

reading the 120 articles that András published in *Scientometrics*, according to the Web of Science (the archive of the journal itself is ironically deficient, it only shows 40 articles written by András and seems to have deleted its history from before it became a Springer title), I am offering a bibliometric portrait of András' contributions to the core journal of the field.

Let us start with the historical development.

Several strands can be distinguished. Starting with the oldest common publication (Evaluative Bibliometrics by Francis Narin from 1976), the bottom right strand focuses on evaluation indicators and journals, including the Hirsch Index. The bottom left strand (ending in Schmoch 2008) is about co-authorship and international collaboration. The publication connecting this strand with the central cluster is a publication together with Tibor Braun on developing countries. The fine structure of the central cluster is shown in Figure 2. The "hanging" two publications with Glänzel are about classification schemes in science.

These publications are all focusing on data files and facts and figures of research performance. They deal both with national research systems and with the technical requirements of the data and indicators. One could call this the cluster of enabling scientometric technologies. It is telling that this takes such a central place in András oevre in *Scientometrics*.

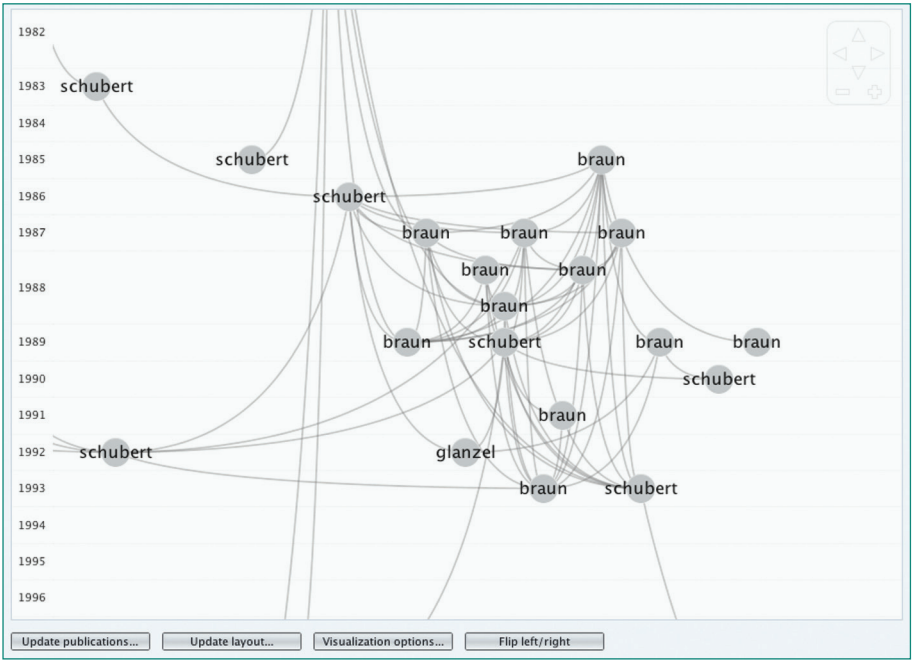


Figure 2: The central cluster in the publication set in Scientometrics, produced with CitNetExplorer (http://www.citnetexplorer.nl).

This view is carried by the citation relations between the documents. But what about their content? To analyze this without reading, I loaded the titles and abstracts of all articles into the VOSViewer (http://www.vosviewer.com). The resulting map is shown in Figure 3.

We see eight clusters of topics that represent the scientific content of András' publications in *Scientometrics*. Interestingly, publications belonging to the different clusters intermingle quite freely, there is no clear spatial separation between the clusters with the exception of the eight cluster. The latter cluster does indeed differ in characteristics since represents bibliographic work. Apparently, our author has done a fair amount of community service work. The other clusters are all related to research. Six of them are fairly comparable in size (between 20 and 30 terms), two are smaller. Figure 4 gives an indication of the density of the map which indicates the extent to which terms are close to each other.

Looking at the items that constitute the different clusters, we see clusters focused respectively on: science policy and national research output; Hirsch-type of indexes; classification issues; international collaboration; research evaluation; national science policy and research output; authorship issues; and bibliographies. However, what is most perhaps striking is the dispersion of clusters over the map. There is no strict spatial separation. I think this can be interpreted as indicator of an underlying dimension that binds the clusters relatively strong together.

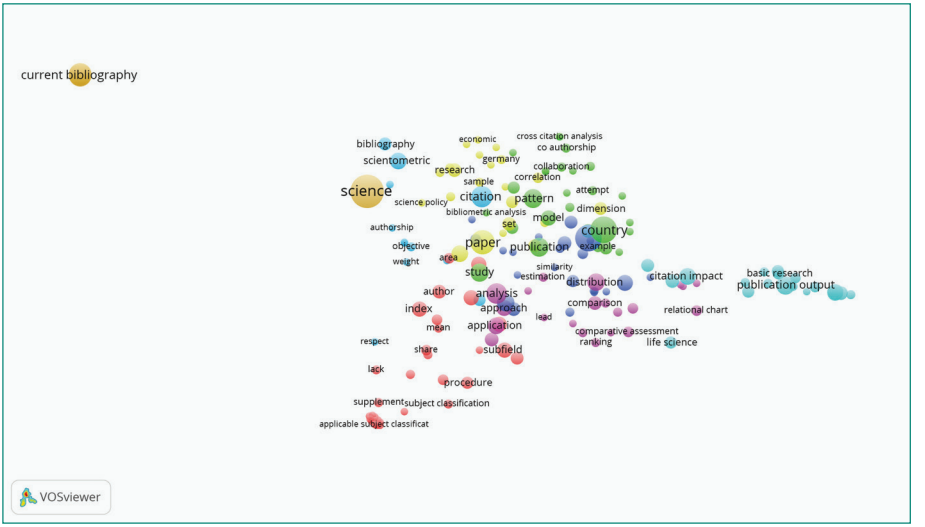


Figure 3: Text mining Schubert in Scientometrics, produced with VOSViewer (http://www.vosviewer.com) (Eck & Waltman, 2014).

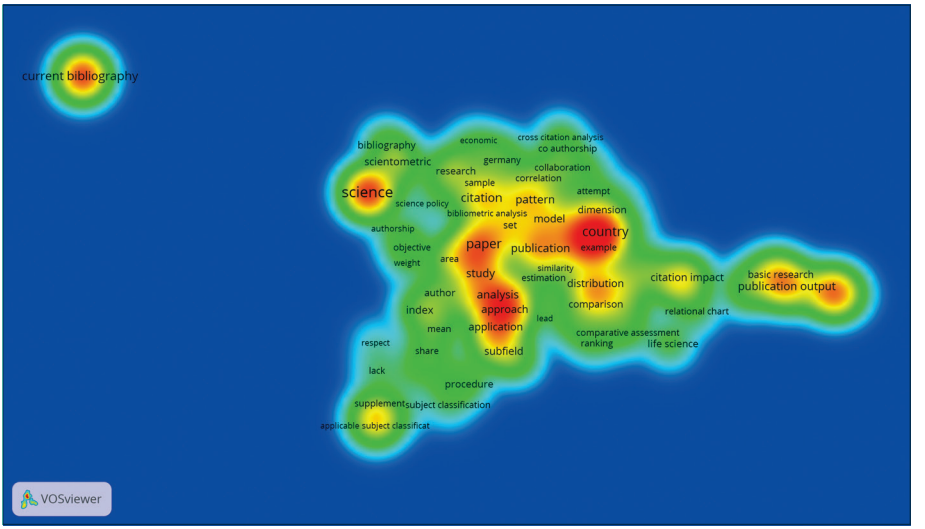


Figure 4: Density map of Schubert in Scientometrics, produced with VOSViewer (http://www.vosviewer.com).

I may be wrong, but my guess would be that this integration is the result of the technical curiosity of András. In the social science field that scientometrics is and will remain, technically oriented explorers like András Schubert have played, and will continue to play, an important engineering and enabling role. Something to cherish and celebrate.

References

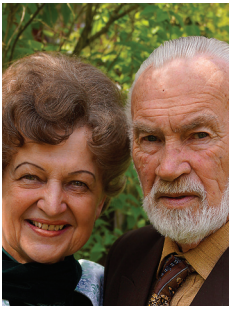
Eck, N. J. Van, & Waltman, L. (2014). *Visualizing bibliometric networks. Measuring scholarly impact: Methods and practice*. doi:10.1007/978-3-319-10377-8_13

Schubert, A., & Maczelka, H. (1993). Cognitive Changes in Scientometrics during the 1980s, as Reflected by the Reference Patterns of its Core Journal. *Social Studies of Science*, *23*, 571–581.

Wouters, P., & Leydesdorff, L. (1994). Has Price's Dream Come True: Is Scientometrics a Hard Science? *Scientometrics*, *31*(2), 193–222.

Emergence of 3-D Order in Regular Shapes of Co-Author Patterns Mirrored in "András Schubert–Google Scholar Citations"

HILDRUN KRETSCHMER & THEO KRETSCHMER
COLLNET Center, Borgsdorfer Str. 5, Hohen Neuendorf, Germany
kretschmer.h@onlinehome.de



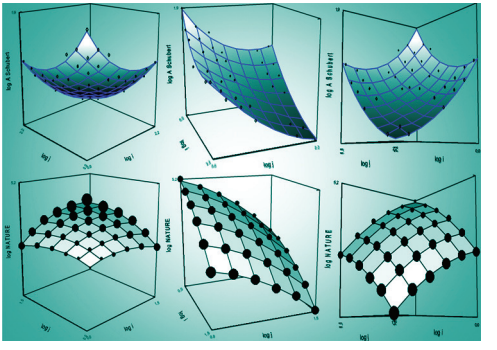
Abstract: Three-dimensional visualization and animation of emerging patterns ("Social Gestalts") by the process of self-oganization in collaboration networks are already presented and described in several papers published by the two authors H. and T. Kretschmer, sometimes in combination with co-authors. In contrast to a single power function distribution (2-D graphs) the new mathematical model of "Social Gestalts" visualizes 3-D graphs, using animation in form of rotation of these graphs. In the former time these 3-D graphs are visualized on the level of large social networks only (journals or large institutions). The number of authors per large social network was ranging between 91 and 111,447. The median was equal to 13,609. In average the degree-centrality of a whole network was low, indicating that many authors are not connected.

On the other hand, the structure of the new here presented study shows strong differences in relation to the former studies:

- The degree-centrality of András Schubert is equal to the number of all of his co-authors presented in "András Schubert—Google Scholar Citations", i.e. all of these authors are connected with him.
- The new studied social network is rather small (60 authors)

Are there important differences visible between the emerging 3-D graphs of the former studied large collaboration networks and the new small network centered by one person?

Keywords: Social network analysis, Self-organization, Complementarities, Co-authorship, Mathematical model, 3-D computer graphs, Animation, Visualization



Graphical Abstract
Co-authorship networks: Rotated 3-D computer graphs. The "A Schubert Shapes" are on the first row and the shapes of the journal NATURE on the second. The black dots are co-author pair's frequencies in logarithmic version.

1. Introduction

Collaboration is increasing in science and in technology. The study of the frequency of pairs or triples of co-authors is highly relevant as well as other kind of studies (de B Beaver 2001; de Solla Price 1963; Glänzel, 2002; Glänzel & de Lange 1997; Luukkonen, Persson & Silvertse 1992; Miquel, Okubo 994; Okubo, Miquel, Frigoletto, Doré 1992; Tijssen & Moed 1989; Zitt, Bassecoulard & Okubo 2000; Newman 2001).

Since more than two decades social network analysis (SNA) can be used successfully in the information sciences, as well as in the studies of collaboration in science. A variety of application possibilities of SNA is available (Wassermann & Faust 1994, Otte & Rousseau 2002) both for studies in large and in small networks. Using SNA is very common both on the micro (actor-centered) level and on the macro (whole-network) level (Borgatti et al 2009, Rousseau and Zhao 2015).

We could find many interesting publications on the topic of collaboration, co-authorship and network analysis in "András Schubert—Google Scholar Citations". Some of them are mentioned here. The names of the authors/co-authors and the titles are given:

1. A Inzelt, A Schubert: Collaboration between researchers from academic and non-academic organisations. A case study of co-authorship in 12 Hungarian universities
2. A Inzelt, A Schubert, M Schubert: Incremental citation impact due to international co-authorship in Hungarian higher education institutions
3. W Glänzel, A Schubert: Domesticity and internationality in co-authorship, references and citations
4. W Glänzel, A Schubert: Analyzing scientific networks through co-authorship
5. I Farkas, I Derényi, H Jeong, Z Neda, ZN Oltvai, E Ravasz, A Schubert, AL Barabási, T Vicsek: Networks in life: Scaling properties and eigenvalue spectra
6. AL Barabási, H Jeong, Z Neda, E Ravasz, A Schubert, T Vicsek: Evolution of the social network of scientific collaborations
7. AL Barabási, H Jeong, R Ravasz, Z Neda, T Vicsek, A Schubert: On the topology of the scientific collaboration networks
8. AL Barabási, H Jeong, Z Neda, E Ravasz, A Schubert, T Vicsek: Scale free topology of e-mail networks
9. W Glänzel, A Schubert: Double effort = double impact? A critical view at international co-authorship in chemistry
10. T Braun, W Glänzel, A Schubert: Publication and cooperation patterns of the authors of neuroscience journals
11. A Schubert, T Braun: International collaboration in the sciences 1981–1985
12. T Braun, I Gómez, A Méndez, A Schubert: International co-authorship patterns in physics and its subfields, 1981–1985
13. T Braun, A Schubert: Analytical viewpoint. International collaboration in analytical chemistry

Conclusion for future studies:

- First, the selection of this "Small co-authorship network (SCN)" above on the topic of collaboration, co-authorship and network analysis is useful for the first step of explanation the basic methods for the description of "Social Gestalts" in this paper.
- Second, the results of a special study of all of the papers found in "András Schubert—Google Scholar Citations" will be presented. We have called all of these papers together: "Large co-authorship network (LCN)".
- Third, we compare the "Social Gestalt" structure of the "Large co-authorship network (LCN)" with the emerging "Social Gestalts" found in 52 international journals published in the OA paper Kretschmer, H & T. Kretschmer (2013, Invited Paper, open access): Who Is Collaborating with Whom in Science? Explanation of a Fundamental Principle. *Social Networking*. 2, 99-137, http://dx.doi.org/10.4236/sn.2013.23011. Published online July 2013. DOI: 10.4236/sn.2013.23011

In Section 2 we explain some details and hypotheses based on the results of the OA paper.

2. The Social Gestalt Model in Brief and Hypothesis

The Social Gestalt model adds a new dimension to studies on *interactions* in social networks. It allows researchers to identify and to examine special regularities of network structures based on *interpersonal attraction* and *characteristic features of the people*.

Since 2009 the mathematical function of the Social Gestalt model is called "*Intensity Function of Interpersonal Attraction*" (Kretschmer & Kretschmer 2009, 2012).

This function for describing Social Gestalts of distributions of co-author pairs' frequencies (N_{ij}) results in the logarithmic version ($\log N_{ij}$) in:

$$\log N_{ij} = c + \alpha \cdot \log(|X-Y|+1) + \beta \cdot \log(4-|X-Y|) + \gamma \cdot \log(X+Y+1) + \delta \cdot \log(7-X-Y) \qquad (1)$$

with X = log i and Y = log j and with c = constant
log i: logarithm of the number of publications i
log j: logarithm of the number of publications j
$\log N_{ij}$: logarithm of co-author pairs' frequencies

The corresponding details and explanation for the derivation of the mathematical function for describing "Social Gestalts" of distributions of co-author pairs' frequencies ($\log N_{ij}$) can be found in the Appendix. The corresponding basic methods for visualizing theoretical and empirical 3-D patterns are explained in Section 4. The data for the studies are presented in Section 3.

Compared with the Appendix, the extended theoretical approach of the mathematical model of Social Gestalts is presented in section 2 of the open access (OA) paper by Kretschmer et al. 2015: http://dx.doi.org/10.1016/j.joi.2015.01.004 and the application in: http://dx.doi.org/10.1016/j.joi.2015.01.009.

This theoretical approach is an essentially improved description and interpretation of the original model and analysis published in the open access (OA) paper by Kretschmer & Kretschmer 2013.

The study of the co-authorship networks presented in this paper is a part of all of our studies on these networks by the new mathematical model of Social Gestalts (3-D graphs). This model has been applied to 52 large co-authorship networks (Kretschmer & Kretschmer 2013, open access (OA), http://dx.doi.org/10.4236/sn.2013.23011). The visualized Social Gestalts in the form of 3-D computer graphs are almost identical with the corresponding empirical distributions. After regression analysis, for 96% of them the squared multiple R is larger than 0.98% and for 77% of the 52 networks even equal or larger than 0.99 (cf. Fig. 1, upper part). The corresponding 40 Social Gestalts (with $R^2 \geq 0.99$) in combination with empirical data are presented in the Appendix of the open access paper mentioned above, cf. pages 117-137.

We have continued these studies (Kretschmer et al 2012 and Ozel et al 2014) resulting up today in 62 Social Gestalts in total. For 80% of these networks the squared multiple R is larger than 0.99 and for 97% larger than 0.98. The median is equal to 0.994.

In continuing these studies we expect a general validity of this mathematical model for research on co-authorship networks.

Additionally to the above mentioned squared multiple R, the F-ratios of the regression analyses are increasing with the total number of co-author pairs (cf. Fig. 2).

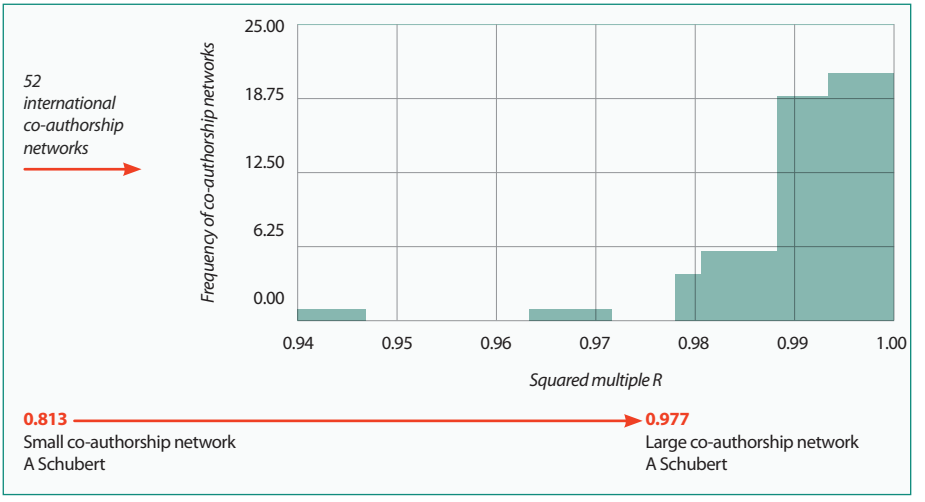


Fig. 1. Upper part: Copy of Fig. 13, Page 110, (Kretschmer & Kretschmer 2013). Frequency of co-authorship networks (ordinate) depend on the squared multiple R (abscissa) after regression analysis (empirical distribution of co-author pairs' frequencies (logNij) and social Gestalt). Note: The squared multiple R ranges between 0.944 and 1.000 and the median is equal to 0.993. For 96% of the co-authorship networks the squared multiple R is larger than 0.98. For comparison, the part on the bottom is showing the results of the "Small co-authorship network (SCN)": Squared multiple R=0.813 (in red color) and the "Large co-authorship network (LCN)": Squared multiple R=0.977 (in red color).

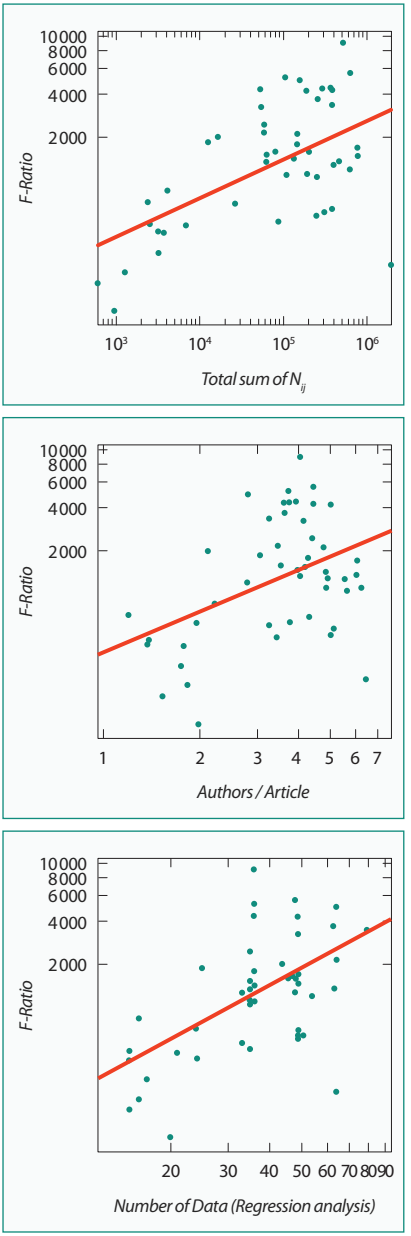


Figure 2. Continuation of Fig. 1: The F-ratios of the regression analyses are increasing with the total number of Nij (first pattern), with the relative number of authors per article (middle) and with the number of data (right pattern)

Therefore we expect both, increasing squared multiple R from SCN to LCN and increasing F-ratios from SCN to LCN.

For comparison, the part on the bottom of Figure 1 is showing the results of SCN and LCN. SCN: Squared multiple R = 0.813 (in red color) and LCN: Squared multiple R = 0.977 (in red color). The corresponding differences between the small and the large co-authorship networks are presented and discussed in the following sections.

But based on the large co-authorship networks, in average the squared multiple R of the former studied 62 Social Gestalts is higher than the squared multiple R of SCN and LCN (cf. Fig. 1).

3. Data

The data are obtained from "András Schubert—Google Scholar Citations".
Papers: 221 are studied. All of these papers are connected with A Schubert as author or co-author. Number of authors in these 221 papers: 60

4. Methods (Partly Presented in Previous Studies of Social Gestalts)

Some of the methods are partly presented in previous studies: Kretschmer, Hildrun & Theo Kretschmer (2009), Kretschmer, H, Kundra, R., deB. Beaver, D.& Kretschmer, T. (2012), Bülent Özel, Hildrun Kretschmer & Theo Kretschmer (2014), Hildrun Kretschmer, Donald deB Beaver & Theo Kretschmer (2015),

The method of counting co-author pairs based on social network analysis (SNA), the logarithmic binning procedure and the method of visualizing the 3-D collaboration patterns are presented in this paper.

Method of Counting Co-author Pairs, Based on Social Network Analysis (SNA):

For the purposes of analysis, a social network can be considered as consisting of two sets, a set of n nodes (individuals) and a set of m edges (undirected relations) between pairs of the nodes.

The degree of a node F_x with x (x = 1, 2....n) is equal to the number of nodes (or edges) that are attached to the node F_x. In co-authorship networks between two authors (nodes) F_x and F_y, there exists an edge if both were acting as co-authors one time at least. In other words the degree centrality of a node F_x is equal to the number of his/her co-authors.

The meaning of "undirected" relations between pairs of nodes is as follows: Under the condition the node F_a is attached to the node F_b, the node F_b is also attached to the node F_a. Symmetrical patterns are emerging.

The "Small co-authorship network (SCN)", based on the 13 articles presented in the Introduction, is selected for explanation.

In general, an author's productivity is measured by his number of publications (cf. Table 1, left side). The number of publications i per author F_x or j per possible co-author F_y respectively are determined by using the ´normal count procedure´. Each time the name of an author appears, it is counted. The n authors F_x are grouped according to their productivities i (cf. Table 1).

Table 1: Authors of the "Small co-authorship network (A Schubert)", ordered according to the number of their publications (i). (Instead of the name of an author the attached letter (cf. below) is used from now on.)

Name of Author	Number of Publications (i)	Letter of Author with # of Publications(i)
M Schubert	1	A(1)
I Farkas	1	B(1)
I Derényi	1	C(1)
ZN Oltvai	1	D(1)
I Gómez	1	E(1)
A Méndez	1	F(1)
R Ravasz	1	G(1)
A Inzelt	2	H(2)
E Ravasz	3	I(3)
W Glänzel	4	J(4)
T Braun	4	K(4)
H Jeong	4	L(4)
AL Barabási	4	M(4)
Z Neda	4	N(4)
T Vicsek	4	O(4)
A Schubert	13	P(13)

The co-author pairs of authors F_{xi}, who have the number of publications i in co-authorship with authors F_{yj} who have the number of publications j (cf. Table 2) are counted.

The resulting sum of co-author pairs N_{ij} is equal to the sum of degrees of the authors F_{xi} to the co-authors F_{yj}. Therefore, the matrix of N_{ij} is symmetrical (cf. the Tables 4 and 5). The data in Table 4 are the results of the data in Table 2.

First example (Table 2): The authors with number of publications i = 1 are attached by co-authors with the number of publication j = 1. The resulting sum of co-author pairs is equal to $N_{11} = 8$. (cf. Table 4)

Second example (Table 2): The authors with number of publications i = 4 are attached by co-authors with the number of publication j = 1. The resulting sum of co-author pairs is equal to $N_{41} = 18$ and because of symmetry: $N_{14} = 18$. (cf. Table 4).

In other words: N_{ij} is equal to the sum of co-author pairs of authors who have the number of publications i in co-authorship with authors who have the number of publications j.

N is equal to the total sum of degrees of all n nodes (all authors F_x) in a network, equal to the total sum of pairs. The Table 4 shows the co-author pairs N_{ij}, selected from the "Small co-authorship network SCN" and Table 5 the co-author pairs N_{ij}, selected from the "Large co-authorship network LCN".

An artificial full table of c-author pairs N_{ij} can be found in Table 3.

Table 2: Co-authors with number of publications j attached to authors with number of publications i. Example: The author A(1) with i = 1 is attached by the co-author H(2) with j = 2. Vice versa: The author H(2) with i = 2 is attached by the co-author A(1) with j = 1. Conclusion: The matrix of co-author pairs N_{ij} is symmetrical (cf. the Tables 3, 4 and 5).

Authors with # of publications (i)	Attached co-authors with # of publications (j)
A (1)	H(2), P(13)
B (1)	C(1), D(1), I(3), L(4), M(4), N(4), O(4), P(13)
C (1)	B(1), D(1), I(3), L(4), M(4), N(4), O(4), P(13)
D (1)	B(1), C(1), I(3), L(4), M(4), N(4), O(4), P(13)
E (1)	F(1), K(4), P(13)
F (1)	E(1), K(4), P(13)
G (1)	L(4), M(4), N(4), O(4), P(13)
H (2)	A(1), P(13)
I (3)	B(1), C(1), D(1), L(4), M(4), N(4), O(4), P(13)
J (4)	K(4), P(13)
K (4)	E(1), F(1), J(4), P(13)
L (4)	B(1), C(1), D(1), G(1), I(3), M(4), N(4), O(4), P(13)
M (4)	B(1), C(1), D(1), G(1), I(3), L(4), N(4), O(4), P(13)
N (4)	B(1), C(1), D(1), G(1), I(3), L(4), M(4), O(4), P(13)
O (4)	B(1), C(1), D(1), G(1), I(3), L(4), M(4), N(4), P(13)
P (13)	A(1), B(1), C(1), D(1), E(1), F(1), G(1), H(2), I(3), J(4), K(4), L(4), M(4), N(4), O(4)

Table 3: Artificial table of co-author pairs N_{ij}. Note: $N_i = \Sigma_j N_{ij}$ is the sum of co-authors of all authors with i publications per author. $N_j = \Sigma_i N_{ij}$ is the sum of co-authors of all authors with j publications per author. N = Total sum of degrees of all nodes in a network, equal to the total sum of pairs including F_x each, with x (x = 1, 2 ... n).

i/j	1	2	3	N_i
1	30	20	10	60
2	20	25	5	50
3	10	5	2	17
N_j	60	50	17	N = 127

Table 4: Table of the co-author pairs N_{ij}, (Sum = 104) selected from the "Small co-authorship network" Note: $N_i = \Sigma_j N_{ij}$ is the sum of co-authors of all authors with i publications per author. $N_j = \Sigma_i N_{ij}$ is the sum of co-authors of all authors with j publications per author. N = Total sum of degrees of all nodes in a network, equal to the total sum of pairs including F_x each, with x (x = 1, 2 ... n).

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	N_j
1	8	1	3	18									7	37
2	1												1	2
3	3			4									1	8
4	18		4	14									6	42
5														
6														
7														
8														
9														
10														
11														
12														
13	7	1	1	6										15
N_i	37	2	8	42									15	**N = 104**

Logarithmic Binning Procedure:

Distributions of this kind of co-author pairs' frequencies (N_{ij}) have already been published (Kretschmer & Kretschmer 2007; Kundra, deB. Beaver., Kretschmer & Kretschmer 2008, Guo, Kretschmer, Liu 2008). However, these former distributions were restricted to $i_{max} = 31$.

Usually the stochastic noise increases with higher productivity because of the decreasing number of authors. We intend to overcome this problem in this paper with help of the *logarithmic binning procedure*. Newman has already proposed in 2005 using the logarithmic binning procedure for the log-log scale plot of power functions. To get a good fit of a straight line (log-log scale plot of power functions, for example Lotka's

Table 5: Table of the co-author pairs N_{ij}, (Sum = 360) selected from the "Large co-authorship network" Note: $N_i = \Sigma_j N_{ij}$ is the sum of co-authors of all authors with i publications per author. $N_j = \Sigma_i N_{ij}$ is the sum of co-authors of all authors with j publications per author. N = Total sum of degrees of all nodes in a network, equal to the total sum of pairs including F_x each, with x (x = 1, 2 ... n).

i/j	1	2	3	4	5	6	7	8	9	10	11	12-91	92	93	94	95	96-220	221	N_j
1	76	2	14	13			1	2					9			4		41	158
2	2	4						1		1			3			4		7	20
3	14		6	6									1					4	31
4	13		6	2												1		3	25
5							1	1					2			2		4	10
6																			
7	1				1								1					1	4
8	2	1			1											1		1	6
9																			
10		1									1		1			1		1	5
11										1			1					1	3
12-91																			
92	9	3	1		2		1			1	1			1				1	19
93													1					1	2
94																			
95	4	4		1	2			1		1									13
96-220																			
221	41	7	4	3	4		1	1		1	1		1	1					64
N_i	158	20	31	25	10		4	6		5	3		19	2		13		64	**N = 360**

distribution), we need to bin the data i into exponentially wider bins. Each bin is a fixed multiple wider than the one before it. For example, choosing the multiplier of 2 we receive the intervals 1 to 2, 2 to 4, 4 to 8, 8 to 16, etc.... For each bin we have ordered the corresponding first value of i (or j) to this bin. Thus, the sequence of bins i' or j' is: i' (i' = 1,2,4,8,16,32, 64, 128, 256...). The same holds for the bins j'. The sizes or widths of the bins (Δi') are: 1, 2, 4, 8, 16 etc.... The same holds for (Δj').

However, because of the bivariate presentation the width of a bin ($cell_{i'j'}$) in the matrix is the product of Δi' and Δj' = (Δi'·Δj'). The sum of co-author pairs in a bin ($cell_{i'j'}$) is called N^S_{ij}, cf. the Tables 6 and 7. The total sum of $N^S_{ij} = \Sigma_{ij} N^S_{ij}$ is equal to the total number of co-author pairs N of a co-authorship network: $N = \Sigma_{ij} N^S_{ij}$

Method of Visualizing the 3-D Collaboration Patterns

The "Small Co-authorship Network SCN" and the "Large Co-authorship Network LCN" are used as examples for comparison.

Remarks:

The following methods will be presented:
- Visualizing empirical patterns,
- Visualizing theoretical patterns and overlay of empirical and theoretical patterns into a single frame

Visualizing Empirical Patterns:

For visualizing the original data we use the sum of co-author pairs in a bin ($cell_{i'j'}$), i.e. N^S_{ij} directly in dependence on i'(bin) and j'(bin), (cf. Tables 6 and 7). Because log 0 is not given, we are using the value "0" for presentation of N^S_{ij} in the tables but not for regression analysis.

The data in Table 4 are used for creating the data in Table 6. The data in Table 4 are combined according to the i'(bin) and j'(bin) of Table 6.

Example: The co-author pairs $N_{21} = 1$ and $N_{31} = 3$ in Table 4 are combined to $N^S_{21} = N_{21} + N_{31} = 4$ in Table 6.

Table 6: Matrix of N^S_{ij} in dependence on i'(bin) and j'(bin), (SCN) with N = 104.

i'(bin)/j'(bin)	1	2-3	4-8	8-15	Sum
1	8	4	18	7	37
2-3	4	0	4	2	10
4-8	18	4	14	6	42
8-15	7	2	6	0	15
Sum	37	10	42	15	N = 104

The matrices of N^S_{ij} (SCN and LCN), Tables 6 and 7, are used as examples for explanation the following steps of the methods.

As the next step in the logarithmic binning procedure: N^S_{ij} of a cell ($cell_{i'j'}$) has to be divided by the width of the bin: ($\Delta i' \cdot \Delta j'$), matrix of the width, cf. Table 8. In other words, the new value in a bin (Example, cf. the Table 9 for SCN and Table 10 for LCN) is simply the arithmetic average of all the points in the bin. This new value, i.e. the ratio, is called the average co-author pairs' frequency $N^{'}_{ij}$.

Using the log-log-log presentation after the logarithmic binning procedure, the sequence of log i' (rows) is as follows: log i'(log i' = 0, 0.301, 0.602, 0.903,, ...); the same

Table 7: Matrix of N^{S}_{ij} in dependence on i'(bin) and j'(bin), (LCN) with N = 360.

i'(bin)/ j'(bin)	1	2-3	4-7	8-15	16-31	32-63	64-127	128-255	256-511	
1	76	14	14	2	0	0	12	40	0	
2-3	14	10	6	2	0	0	8	11	0	
4-7	14	6	4	1	0	0	6	8	0	
8-15	2	2	1	2	0	0	4	3	0	
16-31	0	0	0	0	0	0	0	0	0	
32-63	0	0	0	0	0	0	0	0	0	
64-127	12	8	6	4	0	0	2	2	0	
128-255	40	11	8	3	0	0	2	0	0	
256-511	0	0	0	0	0	0	0	0	0	
SUM	158	51	39	14	0	0	34	64	0	**N = 360**

holds for log j' (columns) resulting in a square matrix. An example of the matrix of the logarithm of the average co-author pairs' frequencies log $N^{'}_{ij}$ is shown in the Tables 11 and 12. The values are obtained from SCN and LCN.

Table 8: Matrix of the Width of the Bin: Δi'·Δj' (As example from 1- 8 only)

Δi'/Δj'	1	2	4	8
1	1	2	4	8
2	2	4	8	16
4	4	8	16	32
8	8	16	32	64

Additionally to the Tables 11 (SCN) and 12 (LCN) two other matrices of the logarithm of the average co-author pairs' frequencies log $N^{'}_{ij}$ are shown in the Tables 13 and 14. The values are obtained from PWQ (Table 13) and from NATURE (Table 14).

"PWQ" is the network of the journal "Psychology of Women Quarterly", 1976-2011. Papers: 1146; authors: 2569.

"NATURE" shows a large network, 1980-98. Papers: 20,673; authors: 52,937

These additional Tables are selected:

- for comparison of smaller networks as SCN and LCN with larger networks (PWQ and NATURE) and
- PWQ is used as an example for explanation basic methods for visualization theoretical and empirical patterns.

Table 9: Matrix of the Average Co-author Pairs' Frequencies N'_{ij} in dependence on i'(bin) and j'(bin) (SCN)

i'(bin)/j'(bin)	1	2	4	8
1	8	2	4.5	0.875
2	2	0	0.5	0.125
4	4.5	0.5	0.875	0.188
8	0.875	0.125	0.188	0

Table 10: Matrix of the Average Co-author Pairs' Frequencies N'_{ij} in dependence on i'(bin) and j'(bin) (LCN)

i / j	1	2	4	8	16	32	64	128	256
1	76	7	3.5	0.25	0	0	0.1875	0.3125	0
2	7	2.5	0.75	0.125	0	0	0.0625	0.04297	0
4	3.5	0.75	0.25	0.03125	0	0	0.02344	0.01563	0
8	0.25	0.125	0.03125	0.03125	0	0	0.00781	0.00293	0
16	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0
64	0.1875	0.0625	0.02344	0.00781	0	0	0.00049	0.00024	0
128	0.3125	0.04297	0.01563	0.00293	0	0	0.00024	0	0
256	0	0	0	0	0	0	0	0	0

Table 11: Matrix of log N'_{ij} in dependence on log i' and log j' (SCN)

log i'/log j'	0	0.30103	0.60205999	0.90308999
0	0.90309	0.30103	0.65321	-.058
0.30103	0.30103		-0.301	-0.9031
0.60205999	0.65321	-0.301	-0.058	-0.727
0.90308999	-0.058	-0.9031	-0.727	

Table 12: Matrix of log N'_{ij} in dependence on log i' and log j' (LCN)

log i/ log j	0	0.301	0.602	0.903	1.204	1.505	1.806	2.107
0	1. 881	0.845	0.544	-0.602			-0.727	-0.505
0.301	0.845	0.398	-0.125	-0.903			-1.204	-1.37
0.602	0.544	-0.125	-0.602	-1.505			-1.630	-1.806
0.903	-0.602	-0.903	-1.505	-1.505			-2.107	-2.533
1.204								
1.505								
1.806	-0.727	-1.204	-1.630	-2.107			-3.311	-3.612
2.107	-0.505	-1.367	-1.806	-2.533			-3.612	

Table 13: Matrix of log N'_{ij} in dependence on log i' and log j' (PWQ)

log i'/log j'	0	0.30103	0.60205999	0.90308999
0	3.34791519	2.39707055	1.66745295	1.25224605
0.30103	2.39707055	1.58546073	0.82118588	0.51188336
0.60205999	1.66745295	0.82118588	0.32735893	-0.12493874
0.90308999	1.25224605	0.51188336	-0.12493874	-0.72699873

In 3-D presentations log i' is placed on the X-axis, log j' on the Y-axis and log N'_{ij} on the Z-axis, cf. Figure 3 as example.

The view at the three patterns on the left column of Fig. 3 and at the bottom pattern on the right column is given from the bottom right corner of the matrix, Table 13, to the top left corner (i.e. along the main diagonal).

One can follow the process of making these patterns visible starting with the upmost pattern at the left column of Figure 3 followed by the other two patterns below. The empirical values (log N'_{ij}) are presented as dots on the top of the corresponding vertical spikes (But empirical values (dots) can also be used separately without any vertical spikes).

On the upmost pattern (left column) one can see the dots on the main diagonal for

- log N'_{ij} = -0.72699873 with log i' = log j' = 0.90308999 in front,
- in the middle: log N'_{ij} = 0.32735893 with log i' = log j' = 0.60205999 and
- log N'_{ij} = 1.58546073 with log i' = log j' = 0.30103 and
- log N'_{ij} = 3.34791519 with log i' = log j' = 0 in the background.

On the second pattern (left column) all of the 16 empirical values (dots on the top of the corresponding vertical spikes) are plotted (But empirical values (dots) can also be separately used without any vertical spikes).

Visualizing the Theoretical Pattern and Overlay of Empirical and Theoretical Patterns into a Single Frame:

For better understanding; as the first step we show *examples* after overlay of empirical and theoretical patterns into a single frame presented in Fig. 3. Explanation about visualizing the theoretical pattern and the method of overlay of empirical and theoretical patterns into a single frame are following afterwards.

Examples:

The bottom pattern at the left column of Fig. 3 is presenting the overlay of the empirical data (dots) taken from the middle pattern of the left column and the corresponding theoretical pattern (lines). But the overlay of empirical dots and theoretical lines in combination with the appearance of the corresponding white colored 3-D surface can be found on the three patterns, at the right column.

As mentioned above, the Social Gestalt shows well-ordered three-dimensional bodies, totally rotatable around and their manifold shapes are visible in the space from all possible points of view. Thus two examples, i.e. the rotation twice in succession of the bottom pattern, right column, are selected. The view at the pattern in the middle is given from the lower left entry of the matrix (Table 13) to the upper right entry (i.e. along the secondary diagonal). The view at the upmost pattern (right column) is given from the top left corner of the matrix, Table 13, to the bottom right corner (i.e. along the main diagonal).

Method of Visualizing Theoretical Patterns and Overlay:

Theoretical patterns are obtained by regression analysis based on the mathematical model for the intensity function of interpersonal attraction (Equation 1 in Appendix). For visualizing the theoretical patterns (lines and/or the 3-D surfaces as in the Figure 3) in combination with the empirical values (dots) we use the Function Plot of SYSTAT for the theoretical and the Scatterplot for the empirical patterns.

After regression analysis using the Equation 1, cf. Appendix or Section 2, after logarithmic binning, we obtain 4 parameters α, β, γ, and δ plus a constant c which are entered into the Function Plot (Z is the dependent variable and X and Y are the independent variables.

The parameter values and the constants for PWQ and for NATURE can be found in Table 15.

Table 14 shows the matrix of log N'_{ij} in dependence on log i' and log j' (NATURE).

Scale Range: The maximum and minimum values to appear on the axis are specified, i.e. both all of the empirical and corresponding theoretical data have to be presented. Any data values outside these limits will not appear on the display. The minimum for the X-axis is in principle specified as 0 ($(\log i')_{min} = 0$) and the maximum is equal to $(\log i')_{max}$ of the empirical data. For example, in Table 13: $(\log i')_{max} = \log 8$. The same holds for the Y-axis $(\log j')_{max} = \log 8$ in Table 13.

Table 14: Matrix of log N'_{ij} in dependence on log i' and log j' (NATURE)

log i'/log j'	0	0.301	0.602	0.903	1.204	1.505
0	5.188	4.552	3.939	3.188	2.122	0.833
0.301	4.552	4.049	3.455	2.693	1.677	0.530
0.602	3.939	3.455	2.881	2.127	1.181	- 0.014
0.903	3.188	2.693	2.127	1.382	0.420	- 0.785
1.204	2.122	1.677	1.181	0.420	- 0.408	- 1.505
1.505	0.833	0.530	- 0.014	- 0.785	- 1.505	- 2.709

The minimum and maximum values for the Z-axis are selected according to the minimum and maximum values of the whole Gestalt produced by the function. In case there are empirical values greater or less than these two theoretical values, the minimum or maximum of the Z-axis has to be extended accordingly so that all of the empirical values become visible.

The Surface and Line Style dialog box is used to customize the appearance of lines or surfaces. The used XY Cut Lines are in two directions. The number of cuts in the grid has to be specified by the number of bins i' (or j' respectively) minus 1 in the data set. For example, a special data set has 4 bins i' as in Table 13 (PWQ); the number of cuts in the grid is specified by 4-1 = 3. The resulting number of lines of the theoretical pattern (Gestalt) is equal to the double of the number of bins i' (2•4 = 8, cf. Fig. 3). The number of points where two of the lines intersect, is equal to the square of the number of bins i'($4^2 = 16$). The Scale Range of the empirical pattern has to be equal (or slightly less) to the theoretical Gestalt (cf. Figure 3).

Most important Remarks:
After the overlay of the empirical distribution and the theoretical pattern into a single frame as in the Figures 3 the goodness-of-fit is highest in the case where the empirical values (dots) are directly placed on the points where two of the theoretical lines intersect. In the case the distance between the intersection points and the dots increases, the goodness-of-fit decreases.

For simplification we use dots in future for presentation of the empirical values but not the vertical spikes.

5. Results

The statistical results of the Social Gestalts from PWQ and NATURE can be found in Table 15 and the statistical results from SCN and LCN in Table 16. These results are based on the mathematical function of the Social Gestalt model called "*Intensity Function of Interpersonal Attraction*", cf. Section 2 and the Appendix (Theory and Mathematical Model for the Intensity Function of Interpersonal Attraction").

*Table 15: Statistical Results of Social Gestalts from PWQ and NATURE (***p < 0.001; **p < 0.01; *p < 0.05)*

	PWQ	NATURE
c	9.616**	-35.571***
α	-2.914*	3.547***
β	-10.853**	13.065***
γ	-8.836***	2.891***
δ	0.321	38.785***
R^2	0.998	0.999
Adj.R^2	0.998	0.999
n df (Regression) df (Residual)	16 4 11	36 4 31
F – Stat	1,354.5	9,024.2
p – value <	9.99E-15	9.99E-16

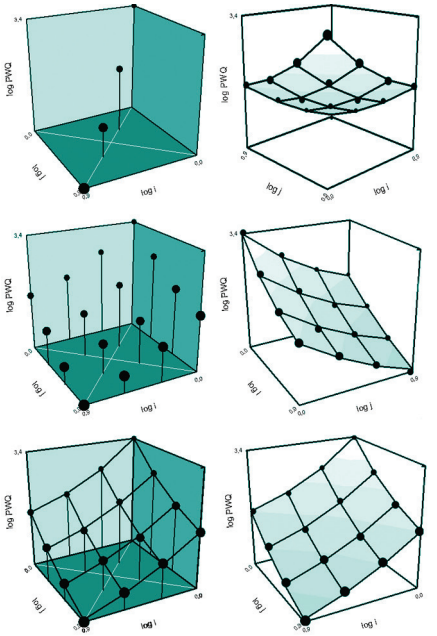


Fig. 3: Visualizing 3-D Collaboration Patterns (PWQ) on the basis of Table 13. Patterns on the left column: Empirical patterns on the upmost and middle patterns. Bottom pattern on the left column: Overlay of empirical and theoretical patterns into a single frame with theoretical pattern in form of lines. Patterns on the right column: Overlay of empirical and theoretical patterns into single frames with theoretical patterns in form of coloured 3-D surfaces.

With permission of the copyright owner the patterns on the right side are partially reproduced from Fig 2 in Kretschmer et al. (2012). With Permission of the copyright owner the patterns are partially reproduced from Fig 2 and 3 in Kretschmer et al. (2015).

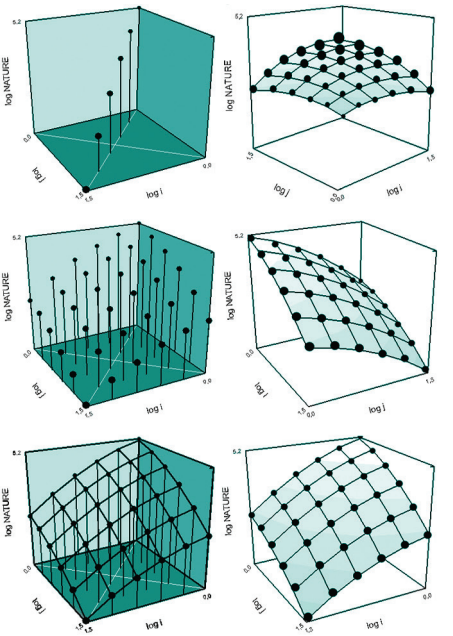


Fig. 4: Visualizing 3-D Collaboration Patterns (NATURE) on the basis of Table 14: Matrix of log N'ij in dependence on log i' and log j (NATURE). Patterns on the left column: Empirical patterns on the upmost and middle patterns. Bottom pattern on the left column: Overlay of empirical and theoretical patterns into a single frame with theoretical pattern in form of lines. Patterns on the right column: Overlay of empirical and theoretical patterns into single frames with theoretical patterns in form of coloured 3-D surfaces. With permission of the copyright owner the patterns on the right side are partially reproduced from Fig 2 in Kretschmer et al. (2012). With Permission of the copyright owner the patterns are partially reproduced from Fig 2 and 3 in Kretschmer et al. (2015).

Whereas the 3-D collaboration patterns from PWQ and NATURE show a similar quality in the relation to the similarity between theoretical patterns and the distribution of the corresponding empirical black dots, the collaboration patterns from SCN and LCN are more or less different, especially SCN. The quality of LCN is higher than SCN (cf. Figures 5 and 6.)

Comparing the "Social Gestalt" structures of the "Small co-authorship network (SCN)" and the "Large co-authorship network (LCN) with the emerging "Social Gestalts", found in 52 international journals published in the OA (open access) paper, we

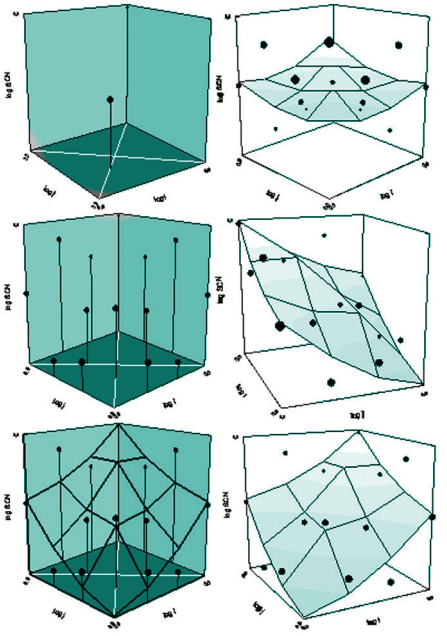


Fig. 5: Visualizing 3-D Collaboration Patterns (SCN): Patterns on the left column: Empirical patterns on the upmost and middle patterns. Bottom pattern on the left column: Overlay of empirical and theoretical patterns into a single frame with theoretical pattern in form of lines. Patterns on the right column: Overlay of empirical and theoretical patterns into single frames with theoretical patterns in form of colored 3-D surfaces.

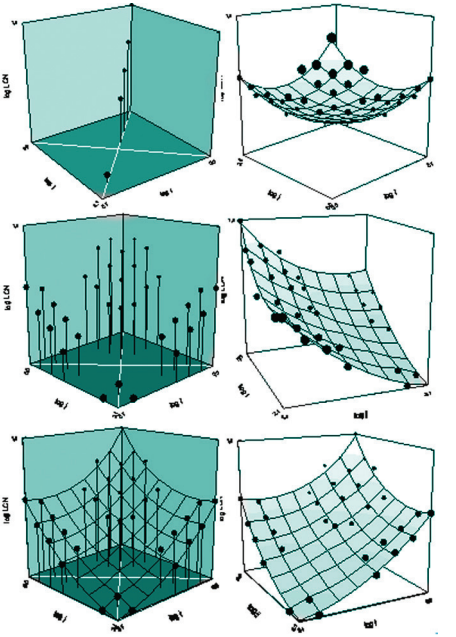
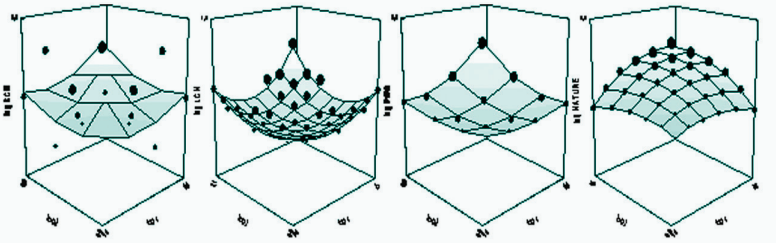


Fig. 6: Visualizing 3-D Collaboration Patterns (LCN): Patterns on the left column: Empirical patterns on the upmost and middle patterns. Bottom pattern on the left column: Overlay of empirical and theoretical patterns into a single frame with theoretical pattern in form of lines. Patterns on the right column: Overlay of empirical and theoretical patterns into single frames with theoretical patterns in form of colored 3-D surfaces.

*Table 16: Statistical Results of Social Gestalts from SCN and LCN (***p < 0.001; **p < 0.01; *p < 0.05)*

	SCN	LCN
c	-8.213	6.568***
α	-8.9374	-3.22*
β	-20.588	-9.002***
γ	2.825	-7.203***
δ	25.502	0.771
R^2	0.7901	0.9767
Adj.R^2	0.6969	0.9736
n df (Regression) df (Residual)	16 4 9	64 4 30
F – Stat	8.47	314.38
p – value <	0.0038	0.0000

FROM LEFT TO RIGHT:
INCREASING POPULATION, INCREASING SQUARED MULTIPLE R AND INCREASING F-STAT



	SCN	LCN	PWQ	NATURE
R^2	0.7931	0.9767	0.998	0.999
F – Stat	8.624	314.38	1354.5	9024.2

From left to right: Increasing F-Stat and increasing Squared Multiple R

Fig. 7: Increasing population, increasing Squared Multiple R and increasing F-Stat in connection with a special increasing order of the networks from left to right (SCN, LCN, PWQ, NATURE).

From left to right the distance between the intersection points and the dots decreases and following the goodness-of-fit is increasing. Vice versa, the distance between the intersection points and the empirical dots is largest at the "Small co-authorship network SCN".

have seen that the squared multiple R = 0.813 of SCN is very strongly smaller than the smallest value from the OA paper. But the squared multiple R = 0.977 of LCN is already belonging to the "Social Gestalts".

The OA paper has also shown, the F-ratios of the regression analysis are increasing with the total number of N_{ij} and with the increasing relative number of authors per article (Fig. 2). We could find a similar result in Fig. 7 of the present paper. Independently of the former combinations of different 3-D collaboration patters a new version is presented.

6. Order and Symmetry in the Regular Shapes of 3-D Graph Images of "Social Gestalts" and Conclusion

According to the remarks by Barabási (2000), most complex systems do not offer a high degree of order; many complex systems are often random and unpredictable. But the discovery of the (2-D) power- law degree distribution by Barabási and collaborators offered the first evidence that large networks self-organize into a scale-free state. 2-D power law distributions of co-author pairs' frequencies are first time shown in 2007 by Morris and Goldstein (cf. Fig. 10).

Compared with the straight lines of the 2-D power law distributions—as an extension—the emerging "Social Gestalts" offer the evidence that large collaboration networks are self-organizing given the high degree of spatial order and special symmetry

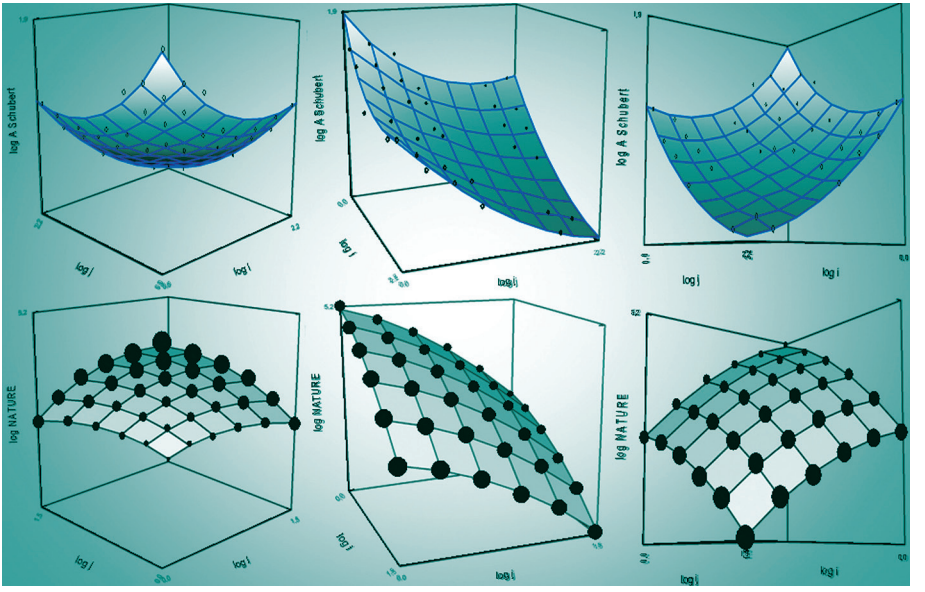


Fig.8: Rotated "Social Gestalt" obtained by LCN (Squared multiple R=0.977) shown at the first row, in comparison with the rotated "Social Gestalt" obtained by NATURE (Square multiple R=0.999)

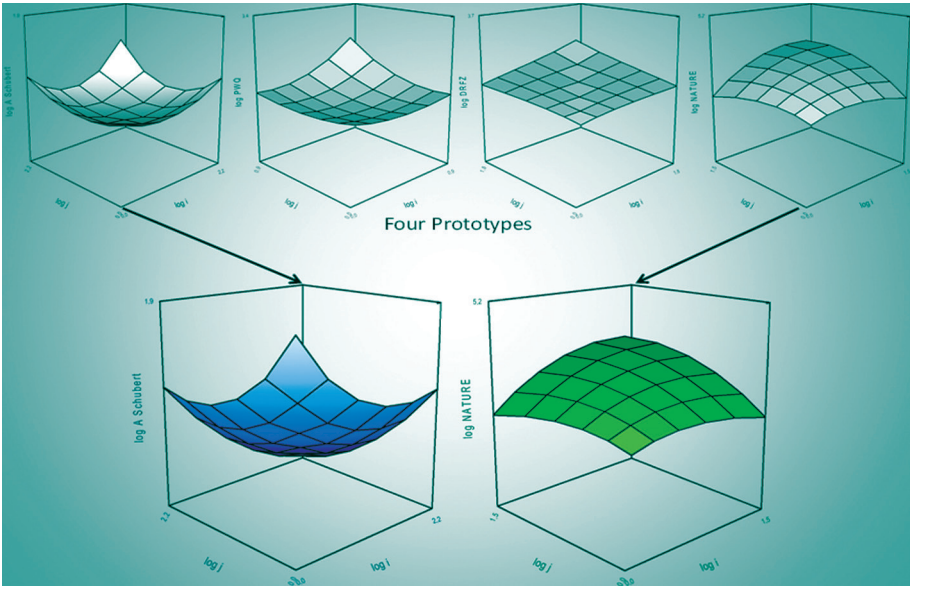


Fig.9: First row: Four varying prototypes of Social Gestalts. Left side to the right: LCN (R2=0.977), PWQ (R2=.998), DRFZ (R2=0.996), NATURE (R2=0.999)
Second row: Comparison of the Gestalts by LCN (left) and NATURE (right)
(DRFZ: Deutsches Rheuma-Forschungs-Institut)

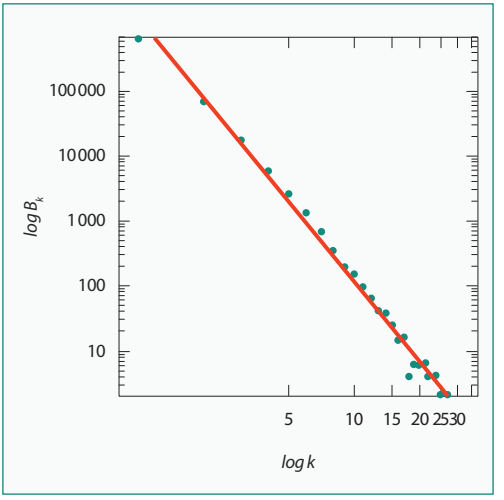


Fig.10: Journal of Biochemistry: Power Law Distributions of Co-author Pairs' Frequencies B_k with k publications per co-author pair

characterizing the manifold 3-D graph images of "Social Gestalts". Each Gestalt of special size (large networks) can be graphed as a 3-dimensional array of co-authorships (cf. Figures 3, 4, 6, 7, 8, 9, except Fig. 5). SCN (Fig. 5) is a "Small co-authorship network".

The "Social Gestalt" model is a new parametric model visualizing 3-D graphs on the level of large networks, using animation to show these graphs from different points of view, cf. Figures (Kretschmer 1999, 2002, 2015) and the Figures 3, 4, 6, 8, 9. As already shown in the present paper, this new model leads to well-ordered rotatable 3-D graphs of co-author pairs' frequencies explaining "Who is collaborating with whom", cf. the two open access (OA) papers: Kretschmer et al. 2015, http://dx.doi.org/10.1016/j.joi.2015.01.004 and http://dx.doi.org/10.1016/j.joi.2015.01.009

According to our question at the beginning of our study, whether there are differences visible between the emerging 3-D graphs of our former studied large collaboration networks and the small collaboration network centered by one person, we can assume probably very small networks are different from larger networks. But the research in future could be of interest to find out from which size on collaboration networks are self-organizing given the high degree of spacial order and special symmetry characterizing the manifold 3-D graph images of "Social Gestalts".

Appendix. Theory and Mathematical Model for the Intensity Function of Interpersonal Attraction (Reproduced from Kretschmer et al. (2012) with permission of the publisher.)

The mathematical function for describing the three-dimensional distribution of co-author pairs' frequencies (N_{ij}) is a special case derived from Kretschmer's *mathematical model for the intensity function of interpersonal attraction (Who is attracting whom? "Intensity" means the extent of this attraction).* This function is already presented in another version in Kretschmer & Kretschmer 2007.

Interpersonal attraction is a major area of study in social psychology.

Whereas in physics, attraction may refer to gravity or to the electromagnetic force, *interpersonal attraction can be thought of force acting between two people tending to draw them together.*

When measuring interpersonal attraction, one must refer to the *qualities of the attracted* as well as *the qualities of the attractor. That means one must refer to their per-*

sonal characteristics. For example, in terms of the degree of the node F_x and the degree of the node F_y (Newman 2002) or in terms of productivity: X = log i of co-author F_x and Y = log j of co-author F_y (Kretschmer & Kretschmer 2007, 2009).

> The notion of "birds of a feather flock together" points out that *similarity is a crucial determinant of interpersonal attraction.*
>
> But: Do birds of a feather flock together or do opposites attract?
>
> This leads to a *model of complementarities: Complementarities are a crucial determinant* of the Intensity Function of Interpersonal Attraction.

Derivation of the Intensity Function of Interpersonal Attraction:

We assume the *intensity structure of mutual* attraction Z_{XY} can be described by a function of a special power functions' combination (X is the value of a special personality characteristic (quality) of an attracted and Y is the value of the same personality characteristic (quality) of the attractor and in case of mutual attraction also vice versa).

The *crucial determinant of interpersonal attraction (similarity or dissimilarity)* suggests considering the *distance A* between the qualities of persons (A = |X-Y|) as the independent variable of a power function:

$$Z^* = c_1 \cdot (A+1)^{\alpha} \qquad (2)$$

with c_1 = constant; the 1 is added because log A is not possible in case A = 0. We see that *as A increases, dissimilarity increases.*

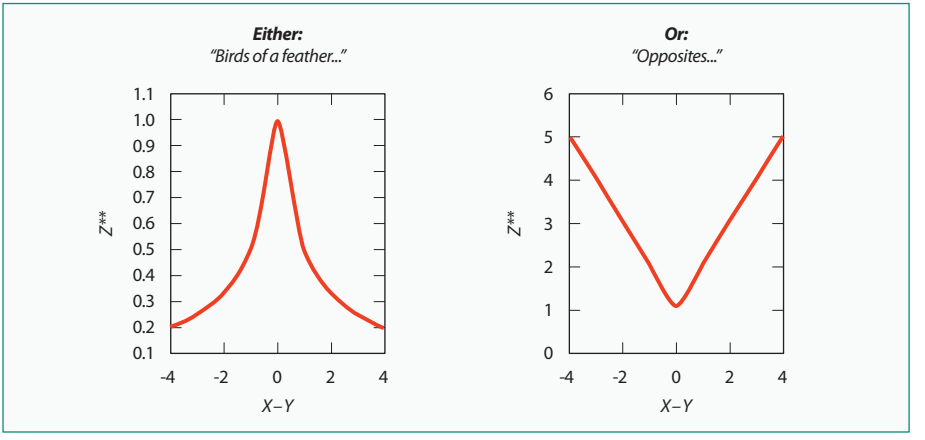


*Fig. 11: Power functions with different values of parameter α (Non-log presentation). In both patterns X-Y is the abscissa with X-Y=0 (Similarity is highest) in the middle and Z*is the ordinate. On the left pattern, the parameter α is negative: "Birds of a feather flock together", i.e. decrease of interpersonal relations with increasing dissimilarity. On the right pattern, the parameter α is positive: "Opposites attract", i.e. increase of interpersonal relations with increasing dissimilarity (This Figure is a copy of Figure 12 in Kretschmer & Kretschmer 2007)*

A power function with only one parameter (unequal to zero) is either only monotonically decreasing or only monotonically increasing; when referred to both proverbs we obtain: *either "birds of a feather flock together" or" the opposites attract"*, cf. Fig 11.

In order to fulfil the inherent requirement that both proverbs with their extensions can be included in the representation, the second step of approximation follows.

Information in brief: There is a *complementary variation of similarity and dissimilarity.* As dissimilarity increases between persons, similarity decreases, and vice versa. Similarity is greatest at the minimum of A and least at the maximum and vice versa, dissimilarity is greatest at the maximum and least at the minimum.

A is a variable with the two opposite poles A_{min} and A_{max}. The sum of A_{min} and A_{max} is a constant. Thus,

$$A_{complement} = A_{min} + A_{max} - A \tag{3}$$

That means, the variable $A_{complement}$ increases by the same amount as the variable A decreases and vice versa, cf. Table 17.

Table 17: Example: Amin = 0, Amax = 3

A	$A_{complement}$
0	3
1	2
2	1
3	0

The model of *complementarities* leads to the conclusion to use additionally the"*complement of the distance A*" = $A_{complement}$ as the independent variable of a secondpower function:

$$Z^{**} = c_2 \cdot (A_{complement} + 1)^{\beta} \tag{4}$$

$$Z_A = \text{constant} \cdot (A + 1)^{\alpha} \cdot (A_{complement} + 1)^{\beta} \tag{5}$$

The relationships of the two parameters α and β to each other determine the expressions of the complementarities (similarities, dissimilarities) in each of the 8 shapes, cf. Fig. 12. In correspondence with changing relationships of the two parameters α and β to each other a systematic variation is possible from "Birds of a feather flock together" to "Opposites attract" and vice versa.

While in the upmost pattern "Birds of the feather flock together" is more likely to be in the foreground, the bottom pattern reveals that "Opposites attract" is more likely to be salient.

Starting pattern by pattern counter clockwise from the upmost pattern towards the bottom pattern, "Birds of the feather flock together" diminishes as "Opposites attract" emerges. Vice versa, starting pattern by pattern counter clockwise from the bottom pattern towards the upper pattern, "Opposites attract" diminishes as "Birds of the feather flock together" emerges.

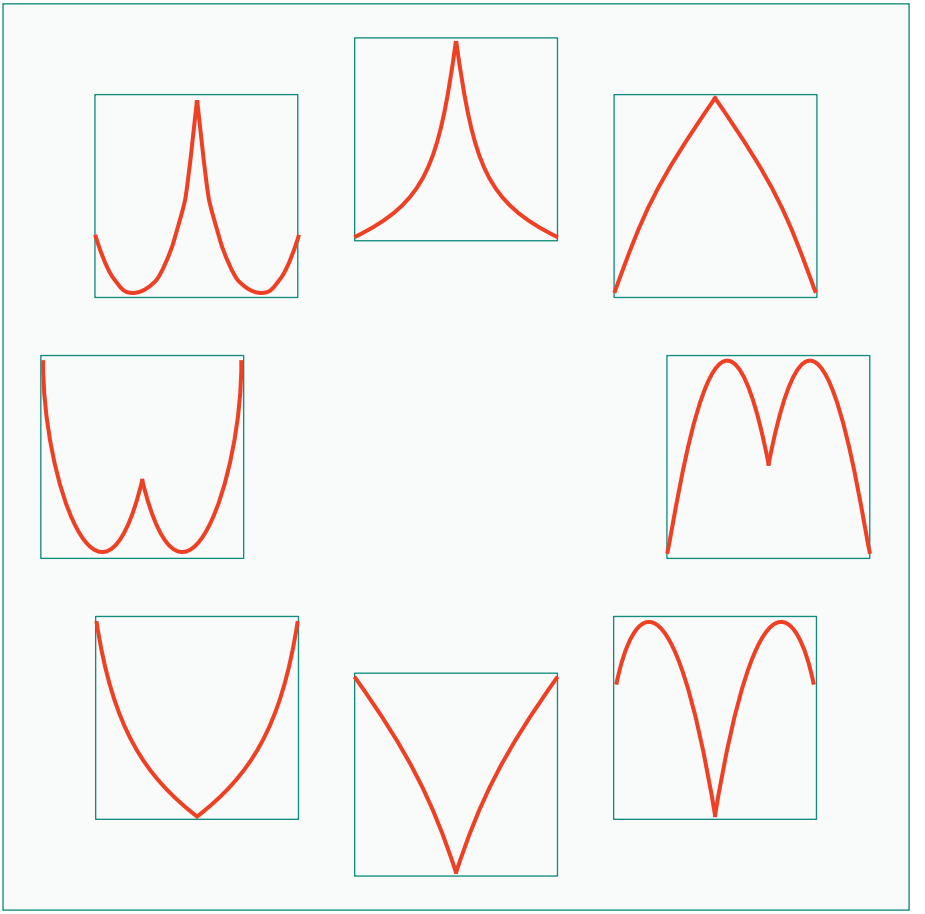


Fig.12: Patterns with varying combinations of the two parameters α and β (Non-log presentation). In all of the 8 patterns X-Y is the abscissa with X-Y=0 in the middle and Z_A is the ordinate. (This Figure is a copy of Figure 7 in Kretschmer & Kretschmer 2007)

For the *purpose of completion,*

- Let the addition ($B = X + Y$) as the *opposite* of subtraction ($A = |X - Y|$), be the independent variable of the third power function

  $$Z^{***} = c_3 \cdot (B+1)^{\gamma} \quad (6)$$

- and the complement ($B_{complement}$) be the independent variable of the fourth power function

  $$Z^{****} = c_4 \cdot (B_{complement}+1)^{\delta} \quad (7)$$

In analogy to A and $A_{complement}$:

$$B_{complement} = B_{min} + B_{max} - B \quad (8)$$

$$Z_B = (B+1)^{\gamma} \cdot (B_{complement}+1)^{\delta} \quad (9)$$

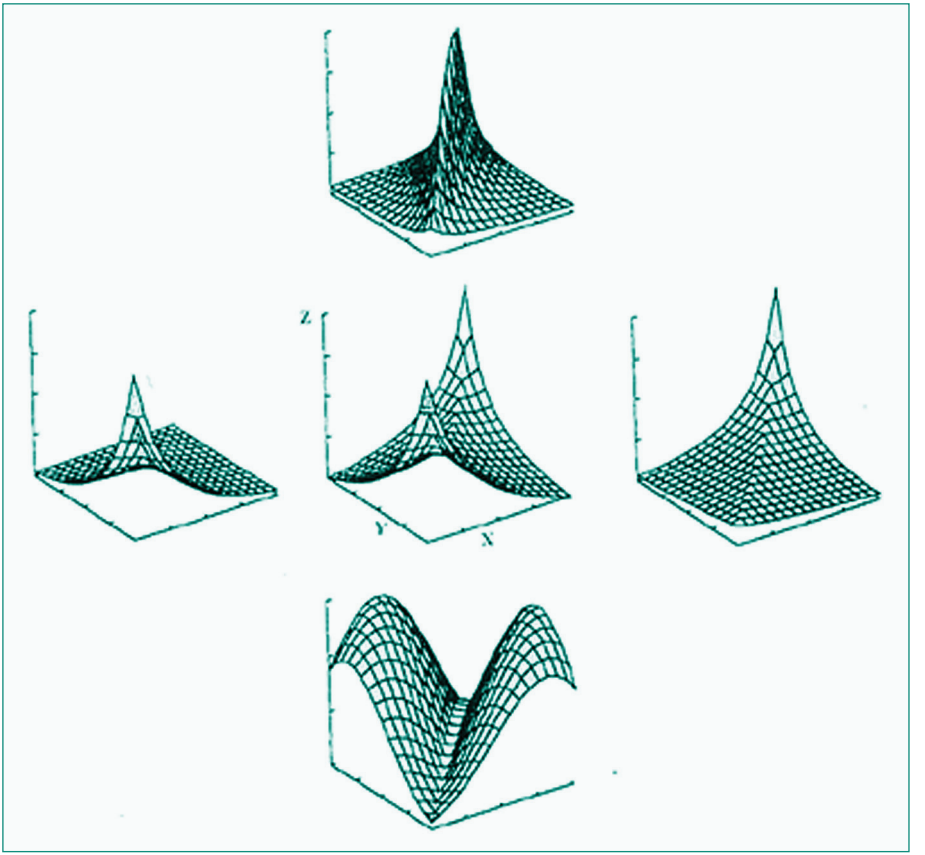


Fig. 13: Prototypes of social Gestalts (non-logarithmic presentation). Several empirical patterns matching the 5 Prototypes were already taken out and presented in Fig. 3, Kretschmer 2002 and in Fig. 5, Kretschmer et al. 2007. The distribution of co-author pairs' frequencies N_{ij} is one of the empirical patterns. The non-logarithmic presentation is similar to the left prototype. However, in this paper we are showing the corresponding log-log-log presentation only (log N_{ij} with log i and log j).

Because the function Z_A can vary independently from the function Z_B we assume the intensity of mutual attraction Z_{XY} is proportional to the product of the two functions Z_A and Z_B:

$$Z_{XY} \sim Z_A \cdot Z_B \tag{10}$$

Therefore, the Intensity Function of Interpersonal Attraction (Social Gestalt) can be formalized as follows (Prototoypes of Social Gestalts, cf. Fig. 13):

$$Z_{XY} = \text{constant} \cdot (A+1)^{\alpha} \cdot (A_{\text{complement}}+1)^{\beta} \cdot (B+1)^{\gamma} \cdot (B_{\text{complement}}+1)^{\delta} \tag{11}$$

with $A = |X - Y|$) and $B = X + Y$

$$A_{complement} = A_{min} + A_{max} - A \quad (12)$$

$$B_{complement} = B_{min} + B_{max} - B \quad (13)$$

$$A_{min} = (|X - Y|)_{min} \quad (14)$$

$$A_{max} = (|X - Y|)_{max} \quad (15)$$

$$B_{min} = (|X + Y|)_{min} \quad (16)$$

$$B_{max} = (|X + Y|)_{max} \quad (17)$$

Measurement of the variables X, Y and Z_{XY} including $X_{min} = Y_{min}$ and $X_{max} = Y_{max}$ depends on the subject being studied.

Examples (types) of social interactions (Z_{XY}) are collaboration, friendships, marriages, etc., while examples (types) of characteristics or of qualities of these individual persons (X or Y) are age, labor productivity, education, professional status, degree of a node in a network, etc.

Whereas Z_A and Z_B each alone produce two-dimensional patterns, the bivariate function Z_{XY} shows three-dimensional patterns (Non-logarithm presentation).

We show one example of how to measure the variables X and Y in relation to the function of the distribution of co-author pairs' frequencies $Z_{XY} = N_{ij}$. The physicist and historian of science de Solla Price (1963) conjectured that the logarithm of the number of publications has greater importance than the number of publications per se.

Thus, using the logarithm of the number of publications (log i or log j respectively) as an indicator of the personal characteristic 'productivity', we define:

$$X = \log i \quad (18)$$

$$Y = \log j \quad (19)$$

$$A = |\log i - \log j| \quad (20)$$

$$B = \log i + \log j \quad (21)$$

Consequently:

$$A_{min} = |X\text{-}Y|_{min} = 0 \text{ with } \log i = \log j \quad (22)$$

$$A_{max} = |X\text{-}Y|_{max} = |(\log i)_{max} — \log 1| = |\log 1 — (\log j)_{max}| = (\log i)_{max} = (\log j)_{max} \quad (23)$$

$$B_{min} = (X+Y)_{min} = \log 1 + \log 1 = 0 \quad (24)$$

$$B_{max} = (X+Y)_{max} = (\log i)_{max} + (\log j)_{max} = 2(\log i)_{max} = 2(\log j)_{max} \quad (25)$$

Let us assume a specific value for the maximum possible number of publications i (or j respectively) of an author as a standard for such studies, which does not vary depending upon the given sample. We assume that the maximum possible number of publications of an author is equal to 1000, i.e.

$$A_{max} = \log 1000 = 3 \quad (26)$$

$$B_{max} = 2\ A_{max} = 6 \quad (27)$$

Thus it follows that:

$$A_{COMPLEMENT} = 3 - |\log i — \log j|, \text{ with } A_{COMPLEMENT} + 1 = 4 - |\log i — \log j| \quad (28)$$

$$B_{COMPLEMENT} = 6 - (\log i + \log j), \text{ with } B_{COMPLEMENT} + 1 = 7 - (\log i + \log j) = 7 - \log i — \log j \quad (29)$$

Thus, the theoretical mathematical function for describing the social Gestalts of the distribution of co-author pairs' frequencies results in the previously mentioned logarithmic version ($\log N_{ij}$):

$$\log Nij = c + \alpha \cdot \log(|X-Y|+1) + \beta \cdot \log(4-|X-Y|) + \gamma \cdot \log(X+Y+1) + \delta \cdot \log(7-X-Y) \quad (1)$$

with X = log i and Y = log j and with c = constant.

References

"András Schubert—Google Scholar Citations"

Barabási, A.-L., R. Albert, H. Jeong (2000): Scale-free characteristics of random networks: The topology of the world-wide web. Physica A, 281, 69-77.

Borgatti, S. P., A. Mehra, D. J. Brass, G. Labianca (2009): Network Analysis in the Social Sciences. Science, 323, 892; DOI: 10.1126/science.1165821

DeB Beaver, D. (2001). Reflections on scientific collaborations (and its study): Past, present and prospective. Scientometrics. 52, 365-377

Glänzel, W. (2002). Co-authorship patterns and trends in the sciences (1980-1998): Abibliometric study with implications for database indexing and search strategies. Library Trends, 50, 461-473

Glänzel, W., de Lange, C., (1997): Modelling and measuring multilateral co-authorship in international scientific collaboration. Part II. A Comparative study on the extent and change of international scientific collaboration links. Scientometrics, 40, 605-626.

Guo Hanning, Hildrun Kretschmer and Zeyuan Liu (2008): Distribution of co-author pairs'frequencies of the Journal of Information Technology. COLLNET Journal of-Scientometrics and Information Management. Vol. 2, No.1,73-81

Kretschmer, H. (1999). A New Model of Scientific Collaboration. Part I: Types of Two-Dimensional and Three-Dimensional Collaboration Patterns. Scientometrics. Vol.46.No.3, 501-518

Kretschmer, H. (2002, Invited Paper): Similarities and Dissimilarities in Co-authorship-Networks; Gestalt Theory as Explanation for Well-ordered Collaboration Structures and Production of Scientific Literature. Library Trends. Vol. 50 No.3, 474-497

Kretschmer, H. & T. Kretschmer (2007): Lotka's Distribution and Distribution of Co-Author Pairs' Frequencies. Journal of Informetrics. 1, 308-337

Kretschmer, H. & T. Kretschmer (2009, Invited Keynote Speech): Who is collaborating with whom? Explanation of a fundamental principle: In: Haiyan Hou, Bo Wang, Shengbo Liu, Zhigang Hu, Xi Zhang, Mingzi Li (Eds.): Proceedings of the 5th International Conference on Webometrics, Informetrics and Scientometrics and 10th COLLNET Meeting, 13-16 September 2009, Dalian, China (CD-ROM for allparticipants and for libraries)

Kretschmer, H, Kundra, R., deB. Beaver, D.& Kretschmer, T. (2012): Gender Bias in Journalsof Gender Studies. Scientometrics 93, 135–150; DOI 10.1007/s11192-012-0661-5

Kretschmer, H & T. Kretschmer (2013, Invited Paper, open access): Who Is Collaboratingwith Whom in Science? Explanation of a Fundamental Principle.Social Networking. 2,99-137, http://dx.doi.org/10.4236/sn.2013.23011. Published online July 2013. DOI: 10.4236/sn.2013.23011

Kretschmer, H; D. deB. Beaver; B. Ozelc; T. Kretschmer (2015): Who is collaborating with whom? Part I. Mathematical model and methods forempirical testing. Journal of Informetrics. http://dx.doi.org/10.1016/j.joi.2015.01.004

Kretschmer, H; D. deB. Beaver; B. Ozelc; T. Kretschmer (2015): Who is collaborating with whom? Part II. Application of the Methods to Male and toFemale Networks. Journal of Informetrics http://dx.doi.org/10.1016/j.joi.2015.01.009

Kundra, Ramesh; Donald deB. Beaver Hildrun Kretschmer and Theo Kretschmer (2008): Co-author pairs' frequencies distribution in journals of gender studies. COLLNET Journalof Scientometrics and Information Management. Vol. 2, No. 1, 63-71

Luukkonen, T.Persson O., & Silvertes, G. (1992). Understanding patterns of internationalscientific collaboration. Science, Technology. Q Human Values, 17, 101-126

Miquel J.F., Okubo, Y. (1994): Structure of international collaboration in science: Comparisons of profiles in countries using a link indicator. Scientometrics. 29(2), 271-294

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

Newman, M. E. J. (2001), The structure of scientific collaboration networks. Proceedings ofthe National Academy of Sciences of the USA, 98 : 404—409

Newman, M.E.J. (2002): Assortative mixing in networks. Physical review letters, 89, 208701

Newman, M. E. J. (2005): Power Laws, Pareto Distributions and Zipf's Law. ContemporaryPhysics, Vol. 46, No. 5, 2005, pp. 323-351. doi:10.1080/00107510500052444

Okubo, Y, Miquel, J.F., Frigoletto, L. & Doré, J.C. (1992). Structure of internationalcollaboration in science; typology of countries through multivariate techniques using alink indicator. Scientometrics. 25, 3121-351

Otte, E., R. Rousseau (2002), Social network analysis: a powerful strategy, also for theinformation sciences. Journal of Information Science, 28: 443-455.
Ozel, B., Kretschmer, H. and Kretschmer, Th. (2014): Co-authorship Pair Distribution-Patterns by Gender. Scientometrics. Volume 98, Number 1, 703-723; DOI 10.1007/s11192-013-1145-y
Price, D.J., De Solla: (1963): Little Science, Big Science. New York: Columbia UniversityPress, 1963
Rousseau, R. & S. X. Zhao (2015): A general conceptual framework for characterizing the egoin a network. Journal of Informetrics, Volume 9, Issue 1, pages 145–149
Tijssen, R.J.W.,& Moed, H.F. (1989) Science and technology indicators. Leiden: DSWO Press
Wasserman, S. & K. Faust (1994): Social network analysis. Methods and applications. Cambridge: Cambridge University Press, 1994
Zitt, M., Bassecoulard, E., Okubo, Y., (2000): Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science, Scientometrics, 47, 627-657.

articles III: metrics

András Schubert's Altmetric Footprint –December 2015

JUDIT BAR-ILAN
Bar-Ilan University, Department of Information Science, Ramat-Gan, Israel

Introduction

András is going to celebrate his birthday in a few months, and as a birthday present I decided to look at his altmetric footprint. Of course this cannot be done without comparing altmetric indicators to traditional bibliometric measurements. I carried out an extensive search in both bibliometric and altmetric data sources.

Although it is going to be András' 16th birthday, his first publications in scientometrics date back to the beginning of the 80's. There was no Twitter, Facebook or Wikipedia back then. There was telnet, ftp and email, but papers were submitted to journals by "regular" (snail) mail and written using a typewriter. There was only a single citation database produced by the ISI (Institute of Scientific Information) in three parts: the Science Citation Index, the Social Science Citation Index and the Arts & Humanities Index. These appeared in print for the general public, and only became available on CDROM as of 1988 (Thomson Reuters, 1988); although some institutions, including the Hungarian Academy of Sciences received the data on computer tapes.

There were no altmetrics or social media back then. The term "altmetrics" was only coined a little more than five years ago (Priem, Taraborelli, Groth & Neylon, 2010). András is well represented on social media, he has a blog (http://schubaa.weebly.com/english.html), where he publishes his literary texts (in Hungarian) and music, but also links to ResearchGate for fulltexts and links to his Researcher ID. However being visible on the Web is only one side of altmetrics. What is usually measured is the attention articles get from others. Since Twitter, Mendeley and Facebook did not exist in the 80's and users of these platforms are usually more interested in recent publications, András Schubert is not an altmetric

star—we only found 10 tweets and 1 facebook post as reported by altmetric.com on the Scopus website. However he is doing quite well on Mendeley and on ResearchGate. Thus we will concentrate on the publications and the citations indexed by WOS, Scopus and Google Scholar vs the readers and reads on Mendeley and ResearchGate. Data were collected at the end of December 2015.

Results and Discussion

I was able to identify 227 publications, not including book reviews, combining several sources: WOS, Scopus, Google Scholar Citation profile, ResearchGate profile, RESEARCHER ID and András Schubert's publication list from the Hungarian Repository of Research Publications (https://vm.mtmt.hu//www/index.php?AuthorID=10049931). None of the sources were completely comprehensive. Table 1 displays the summary data for citations and/or the number of readers in the different sources together with the number of items indexed by each source. The "read-index" is the same as the h-index, but instead of citations the number of readers are counted. Not surprisingly, we see considerable differences between the sources (Bar-Ilan, 2008). Figure 1 displays the yearly number of publications, while Figure 2 shows the number of citations/reads the items published in each year received as of the end of December 2015.

Table 1: Summary data

	# publications indexed	Coverage (out of 227)	Total # of citations/reads	Avg. nr. of cit./ reads per source	h-index/ "read index"
WOS	165	73%	4,225	25.61	32
Scopus	145	64%	4,443	30.64	31
Google Scholar	206	91%	9,131	44.33	48
Mendeley	113	50%	2,236	19.79	23
ResearchGate	148	65%	2,532	17.11	23

We see that 1989 is the peak year in the number of publications, while 2002 is the peak year in the number of citations. This is caused by the top cited/read item in all the sources, except ReseachGate: "Evolution of the social network of scientific collaborations" by Barabási, Jeong, Néda, Ravasz, Schubert and Vicsek, published in Physica A. On ResearchGate, "Analysing scientific networks through co-authorship" by Glänzel and Schubert published in 2005 in the Handbook of Quantitative Science and Technology was read most.

As can be seen in Figure 2, publications from the last century are not very highly read by users of Mendeley and ResearchGate. For items published in 2000 or after, the general trend of all the sources is more or less the same.

During data collection we noticed that the number of reads on ResearchGate is totally unreliable. When a signed-in user clicks several times on a publication within a

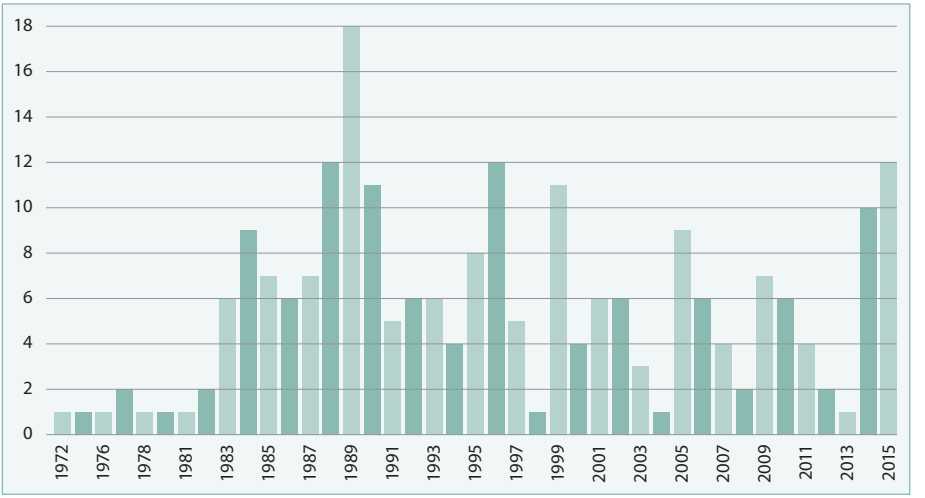


Figure 1: Yearly number of publications

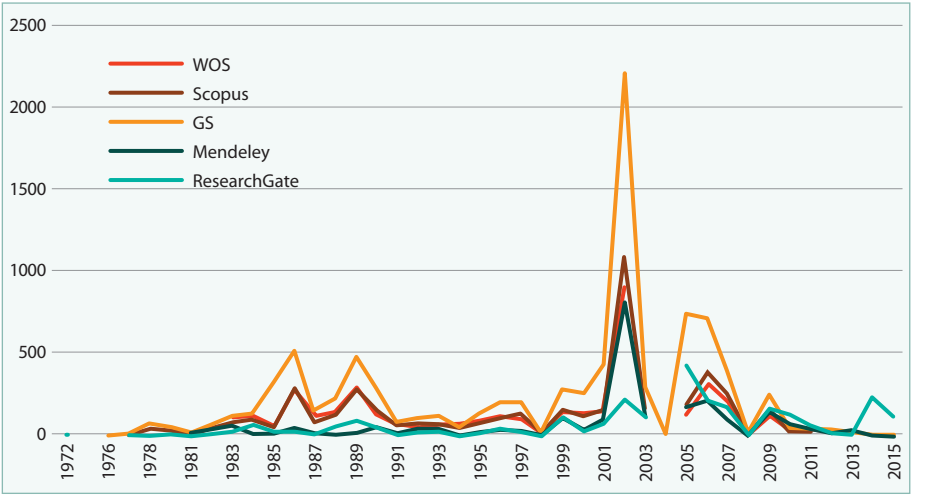


Figure 2: Number of citations/reads per publication year

short period of time, the number of reads increases accordingly (doesn't work for own publications). First of all it means that it does not count readers (persons or groups) as Mendeley does, but it counts act of viewing the metadata record, and this number can be very easily manipulated, perhaps not by the authors themselves, but by friends as can be seen from Figures 3-6

Some articles are cited much more of Google Scholar than read on Mendeley. The ones with the highest differences are displayed in Table 2 and those that were read more than cited on Google Scholar in Table 3. We see that

Table 2: Cited more on Google Scholar than read on Mendeley

Authors	Title	Year	Source title	GS	M.
Braun T., Glänzel W., Schubert A.	A Hirsch-type index for journals	2006	Scientometrics	475	65
W Glänzel, A Schubert	Analysing scientific networks through co-authorship	2005	Handbook of quantitative science and technology	387	105
Braun T., Glänzel W., Schubert A.	A Hirsch-type index for journals [1]	2005	Scientist	244	11
Barabási A.L et al.	Evolution of the social network of scientific collaborations	2002	Physica A	1998	605
Schubert A., Braun T.	International collaboration in the sciences 1981-1985	1990	Scientometrics	217	16
Schubert A., Glänzel W., Braun T.	Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981-1985	1989	Scientometrics	323	23
SCHUBERT, A; BRAUN, T	Relative indicators and relational chart for comparative assessment of publication output and citation impact	1986	Scientometrics	356	25

Table 3: Read more on Mendeley than cited on Google Scholar

Authors	Title	Year	Source title	GS	M.
Schubert A.	Measuring the similarity between the reference and citation distributions of journals	2013	Scientometrics	7	19
Braun T., Schubert A.	Journal of radioanalytical and nuclear chemistry, 2005-2009: A citation-based bibliography and impact analysis using Hirsch-type statistics	2010	Journal of Radioanalytical and Nuclear Chemistry	2	14
Schubert A., Schubert M.	Outperform your neighbors	2009	Scientometrics	1	12
Farkas I., et al.	Networks in life: Scaling properties and eigenvalue spectra	2002	Physica A	89	136
Schubert, A	Scientometrics: A citation based bibliography 1994-1996	1999	Scientometrics	7	17

Finally the Spearman correlations. The Spearman correlation between Google Scholar citations and Mendeley readership counts for the 112 publications indexed by both of them is 0.764 ($p<0.01$); between Scopus and Mendeley (for 106 publications) is 0.743 ($p<0.01$) and between WOS and Mendeley (for 103 publications) is 0.692 ($p<0.01$). The correlations are higher than most of the cases reported in the literature, where the correlation is around 0.5 (e.g. Zahedi, Costas & Wouters, 2014).

Conclusion

The conclusion here is straightforward: Happy birthday, many happy returns and lots of more publications, citations and readers!! And even more important are the other things, including music, literary writing and family. Wish you health and happiness.

Figure 3: Captured from ResearchGate at 20:27 on 24/01/2016

Figure 4: Captured from ResearchGate at 20:28 on 24/01/2016

Figure 5: Captured from ResearchGate at 20:29 on 24/01/2016

References

Bar-Ilan, J. (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. Scientometrics, 74(2), 257-271.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics manifesto. Retrieved from http://altmetrics.org/manifesto/

Thomson Reuters. (2015). 50th Anniversary Science Citation Index. Retrieved from http://wokinfo.com/sci-anniversary.html

Zahedi, Z., Costas, R. & Wouters, P. (2014) How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. Scientometrics, 101(2), 1491-1513.

András Schubert at a Glance: A Portrait Drawn from his Most-Frequently Cited Publications

MARÍA BORDONS & ISABEL GÓMEZ
CSIC, Spain

András Schubert is unquestionably a distinguished member of the scientometric research community and a remarkable contributor to the advancement of this field. Initially geared with an academic training in Chemistry, he soon switched over and has devoted his professional career to information science and scientometric research at the Hungarian Academy of Sciences (Budapest, Hungary) for more than three decades.

András Schubert is the author of more than 160 scientific publications covered by the Web of Science database from the late 70s to the present day. He was particularly prolific in the 80s and the early 90s, but has since been regularly publishing his works attaining high impact scores as measured by the number of citations received with an upward-sloping trend up to recent years. He is credited with a high h-index value (32, in WoS core collection; 47, in Google Scholar) bearing witness to his outstanding role in the field.

A glance at András's most-cited publications enables us to glean interesting information about his research lines and most relevant scientific contributions to the field. In particular, he is the author of seven papers which have received 100 or more citations (WoS core collection), attaining the most cited one almost 800 citations in the Web of Science, bringing it close to 2000 citations in Google Scholar. Most of these papers were published in the journal *Scientometrics* in collaboration with his colleagues Tibor Braun and Wolfgang Glänzel.

Amongst these most-cited papers, two, dating from his 'early period', were published in the late 80s and were considered a breakthrough in the development of publication and citation indicators at the macro level to assess performance by country. In the first one, the need to develop relative indicators of activity and impact (activity and attractivity indexes) was advocated and empirically demonstrated in the field of Chemistry for a sample of 25 countries (Schubert & Braun, 1986). The second one was published a few years later (Schubert et

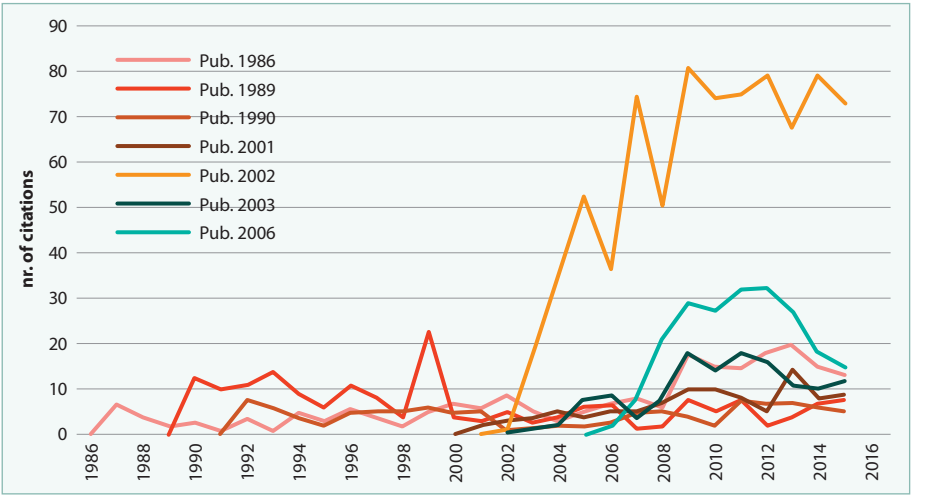


Figure 1: Yearly evolution of citations received by Schubert's most-cited publications (WoS)

al., 1989) providing an extensive set of indicators for 96 countries across all fields where publication and citation profiles by country were described with an unprecedented level of detail and comprehensiveness. For many years, it has been an essential tool for scientometricians, stimulating both debate and further research.

International collaboration has also been a major interest in the research agenda of András Schubert and two of his most-cited papers were devoted to collaborative issues. Besides providing empirical data which made apparent interfield differences in international collaboration, he has made significant contributions to different topics such as the assessment of the strength of co-authorship links between countries (Schubert & Braun, 1990), the asymmetry of collaboration links, and the relationship between international collaboration and citation impact (Glänzel & Schubert, 2001).

Three of András Schubert's most-cited papers were published in his, so to speak 'modern period' spanning from 2002 to the present day. They have in common a certain methodological nature which has made them useful to many researchers in the field. That is the case, for instance, of his proposal of a classification scheme of science fields, which undertakes the challenge of classifying multidisciplinary papers (Glänzel & Schubert, 2003); or of the paper putting forward the suggestion of applying the h-index at journal level (Braun et al., 2006). Finally, an incursion of his in the field of Physics resulted in an interesting and highly-cited paper published in *Physica A* where social network analysis is used to characterize co-authorship links between authors determining network topology and changes over time (Barabási et al., 2002).

András Schubert´s research impact bears heavily on the Information Science and Library Science field as attested by the distribution of the citations received by his most-cited papers (more than 70% of such citations are sourced from IS&LS journals), and *Scientometrics* is the most frequent citation channel for his works, in particular, for the older papers.

However, a completely different situation emerges from the research published in *Physica A*, which was a collaborative paper with Physics researchers. This paper pulls a huge number of citations from the Computer Science and Physics fields, and shows a high dispersion of citations considering both journals and institutions. It is a clearly interdisciplinary paper which crosses disciplinary borders and attains recognition both from researchers in the scientometric community and from researchers well beyond its boundaries.

Concerning the age distribution of citations, it is interesting to remark that his most-cited papers remain influential today. Moreover, the highest absolute number of citations per year is obtained by his most recent papers, thus providing further evidence of the successful evolution of András Schubert's research track. On the other hand, the case of the 1986 publication is particularly worth mentioning, since after a long and quite stable citation performance, a sudden swell in citations took place 27 years after its publication. This event may be accounted for in the context of an upsurge of interest and debate on relative indicators and normalization procedures in the scientometric community.

The citations received by András Schubert's top-ranking papers are spread over a wide variety of countries and institutions. A higher concentration of citations coming from a reduced number of European institutions such as CWTS, Leuven University and the Hungarian Academy of Sciences is observed for his most specialized or methodological papers. On the other hand, the distribution of citations across institutions and countries shows higher dispersion rates in the case of topics, such as the h-index or international collaboration issues, which have gathered interest from a wider range of different readers.

In summary, back in 1993, András was deservedly distinguished with the Derek John de Solla Price Award for being a pioneer in the field of scientometric studies and he has successfully managed to retain the brilliance of his works throughout his professional career.

References

Schubert, A.; Braun, T. (1986). Relative indicators and relational charts for comparative-assessment of publication output and citation impact. *Scientometrics* 9(5-6): 281-291.

Schubert, A. ; Glänzel, 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.

Schubert, A.; Braun, T. (1990). International collaboration in the sciences, 1981-1985. *Scientometrics* 19(1-2): 3-10.

Glänzel, W.; Schubert, A. (2001). Double effort = Double impact? A critical view at international co-authorship in chemistry. *Scientometrics* 50(2): 199-214.

Barabasi, A.L.; Jeong, H.; Neda, Z; Ravasz, E.; Schubert, A.; Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A-Statistical mechanics and its applications* 311 (3-4): 590-614.

Glänzel, W.; Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics* 56 (3): 357-367, 2003.

Braun, T.; Glänzel, W.; Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics* 69(1): 169-173.

Scientometrics and Musicometrics

HENK F. MOED
Independent researcher and senior scientific advisor,
Amsterdam, The Netherlands. Email: hf.moed@gmail.com

András Schubert has made impressive contributions to the field of quantitative science studies He played a key role in the foundation of the journal Scientometrics, and was associate editor of this journal for many years. How does one recognize a good scholar? From his or her publications. This is perhaps one of the base assumptions in the field of scientometrics. And indeed, András' list of publications is impressive.

Scientometricians have developed methodologies to visualise and study the impact that research publications in a field make to surrounding research activities, or to scientific progress in general. Developing such methods is a genuine endeavour. Although none of the methodologies fully captures a publication's value, and practitioners in the field agree that impact and quality are by no means identical concepts, many of us believe that citation counts, when properly used, provide a useful and valid tool in the assessment of a publication's impact.

Of course, the theoretical foundation of citation analysis is still heavily debated. It is essential that methodologies and indicators applied in policy studies of scholarly activity and performance are properly tested and theoretically founded. Quantitative science and technology studies is a multi-disciplinary field, and even within a discipline fundamentally distinct paradigms were developed. The existence of distinct, to some extent competing theoretical positions is not uncommon in the social sciences. It is therefore invalid to assume that a theoretical foundation is sound only when there is a strict consensus among practitioners involved, and that, whenever various, competing theoretical positions exist, it follows that there is no theoretical foundation at all.

Let us look at the citation rates of András' publications. Figure 1 displays the citation rate of his 20 most frequently cited publications. The list of titles—in some cases abbreviated –at the left had side of the chart is informative. Publi-

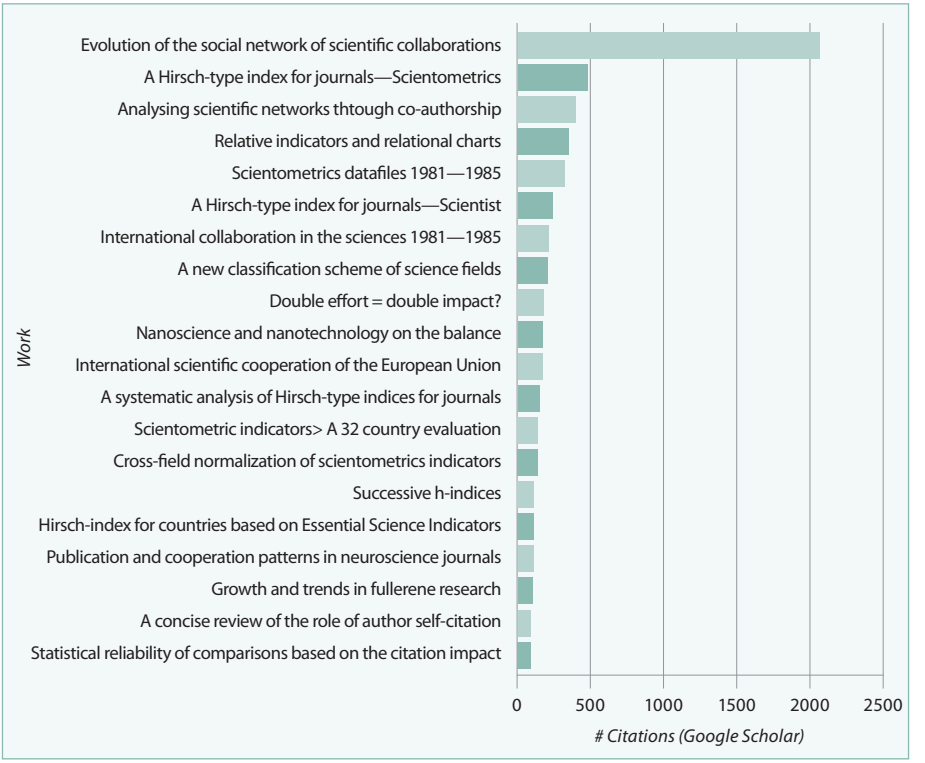


Figure 1: Number of citations to the 20 most frequently cited publications authored by András Schubert. Data from Google Scholar, collected on 16 February 2016

cations in at least the following four main themes seem to have attracted the largest attention in the scientific-scholarly literature indexed in Google Scholar: the study of scientific collaboration as a social network; statistical properties of bibliometrc indicators, especially the Hirsch Index and field-normalized measures; macro-views of global scientific activity; and the emergence of new research fields.

But there is more than scientometrics alone. I remember many social events during international scientometric conferences in which András practiced his great passion: making music. In fact, playing the clarinet is for him more than a hobby: it is his second life. And he is playing the instrument very well, at the level of a full professional. In this way his performance at conferences established a direct link between the domain of science and that of the arts, particularly music.

There are other ways to link science and music. Scientometricians could learn from the study of music. In the quest for citation theories, it could be fruitful to further enlarge the horizon, and analyze for instance differences and similarities between scientific and musical performance, and between the ways in which performance is being assessed. In both domains, the notion of quality plays a key role. How can we obtain a better

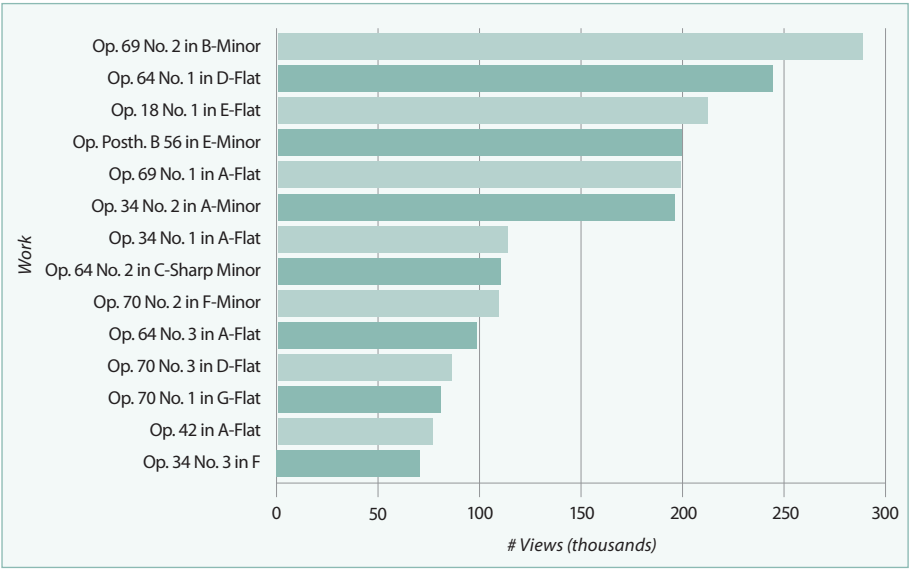


Figure 2. Number of views in Youtube of 17 waltzer by Chopin, performed by Arthur Rubinstein. Data collected from Youtube on 19 December 2015 from https://www.youtube.com/watch?v=laSh3D_77ZM&list=PLD3C0C5CF92D4C7B3&index=1. All pieces were posted in July or August 2009 by the same user.

understanding of what bibliometricians measure, if we confront ourselves with the assessment processes, funding policies and management tools in the domain of the arts?

Conversely, the study of the arts could learn from scientometricans, or, more generally from informetricians as well. This is not a new viewpoint. Focusing on music, there are already quite a few organizations, both in the academic and the private sector, dealing with the development and application of business analytics for the music industry. This industry is becoming more and more data-driven in its decision making. Coining a term is hard to do. The term 'musicmetric' is already in use—for instance, there is a music analytics company with this name, based in the UK and bought in 2015 by Apple. If one aims to construct a term analogously to 'scientometrics', 'musicometrics' is perhaps the best candidate. As an internet domain name it is apparently still available. But what is in a name?

One approach could be the systematic exploration of the application of informetric methods developed within the context of quantitative science studies in the domain of music. A challenging object of research would be YouTube. Figure 2 presents a typical example of an informetric result obtained from YouTube. It gives the number of views of 14 waltzes composed by Frederic Chopin and performed by Arthur Rubinstein. They were all posted on YouTube approximately at the same date and by the same user, and grouped into one single playlist. One could hypothesize that each work in this list has the same probability of being viewed.

Figure 2 shows large differences in the number of views among the various pieces. Six of them have approximately 200,000 views of more, and eight of about 100,000 or

less. Many questions could be raised. A first relates to stability. Do other performances of these 15 waltzes, by other piano performers, show the same pattern? And is this grouping of pieces into two classes stable over time? Other questions relate to validity and interpretation. Are there any musicological explanations for the observed pattern? Is there information available on the profiles of the viewers? Are there differences in the way the various types of viewers behave? Are there external factors at stake that have little to do with the intrinsic properties—quality, if one wishes—of the pieces, but, for instance, relate to the visibility of particular piece on the web or elsewhere? Is there a tendency that the first piece in a playlist is more often viewed than the other contributions?

A comparison of the results for the 14 waltzes in Figure 2 with those related to 19 of Chopin's Nocturnes, performed by the same pianist, and posted in YouTube by the same user at approximately the same time, reveals that the distribution of views among pieces is more skewed for Nocturnes than it is for Waltzes. It also shows that Nocturnes are viewed on average about 30 per cent more often than Waltzes. What does this outcome tell us about the differences in reception of these two series of pieces?

The quantitative study of art performance, and especially music performance, is an interesting subject for scientometricians and informetricians. Actually, it constitutes a part of the field of informetrics. Scientometrics and musicometrics have a lot to offer to one another. More explicitly linking these two would be beneficial for both. In this way the link between science and music established by András at our international conferences further evolves. The master showed us the way.

Scientometrics and its Institutionalization: The Role of András Schubert

SUJIT BHATTACHARYA[1]
CSIR-National Institute of Science, Technology and Development Studies and Academy of Scientific and Innovative Research; Pusa Campus, Dr. K.S. Krishnan Marg, New Delhi-110012, India, sujit@nistads.res.in

I feel proud and privileged to be invited by Prof Wolfgang Glanzel to contribute to the Festschrift volume of Dr. András Schubert to be published on his 70th Birthday. My interactions with András were brief, in the big ISSI conferences. However, even in the brief interactions I could sense his warmth and felt that he was glad of my presence. I feel that scholars like András know the special efforts needed for researchers from developing countries to be part of international research community. However, better funding and acceptance of diversity of views provides more possibilities for researchers from developing countries to participate in global forums nowadays.

Quantitative studies of Science primarily emerged from intellectual curiosity of Derek Price[2], and also due to demand for more objectivity in science funding (see for example Price, 1963[3]). Eugene Garfield's creation of citation index provided the needed tool for demonstrating the application of this new field of inquiry. However, I argue that institutionalization of the field happened in September 1978 with the first issue of the journal Scientometrics. The footmarks of András is visible from the first few issues onwards of this journal and has continued over the years covering various contemporary top-

---

[1] Also Editor-in-Chief of Journal of Scientometric Research (www.jscires.org). This journal published by Wolter-Kluwer Health in association with SciBiol-Med. For correspondence editor.jscires@gmail.com

[2] Price work made seminal contribution to establish the proposition 'Science as a social institution", a proposition which was primarily articulated by J.D. Bernal, see for example J.D. Bernal "The Social Function of Science" published in 1939. Price developed statistical model of science through the new approach. Among others, Vasilii Nalimov,Moscow University and Gennady Debrov, Ukraine Academy of Science, contemporary of Derek strengthened this new methodological approach.

[3] Price, D.J.D.S., *Little Science Big Science*, Columbia University Press, USA, 1963.

ics of the time, providing new pathways to understand science dynamics and its intellectual contents. He is among a few whose work spans a rich repository covering vast terrains over a long period of time in Scientometrics.

András played a major role in developing research community in this field through this journal. The Scientometric research community we have today spans scholars and young researchers from across the globe i.e. a wide representation covering North as well as South countries. This would not have been possible if persons like András were not there from the beginning. I provide some evidences in support of this claim.

Research Question and the Scope of the Study

The study argues that establishment of a journal in a new field of inquiry is one of the key process of institutionalization of a field. Keeping this argument, the study posits that Scientometrics journal played a major role in institutionalizing this field. The role of András Schubert is examined in this context. Keeping this research context, the study examines the influence of András Schubert in the journal for the period 1978 to 1992. This covers 15-year period from the start of the Scientometrics journal in September 1978.

The Initial Years of Scientometrics

It is interesting to read what David Edge[4] had to say about his experience on the first day of his assignment to start the science studies unit at Edinburg University:

> "In the early morning of March 1, 1966, I arrived at Waverly station in Edinburgh, on the night train from Landon, to start the Science Studies Unit at Edinburgh University. Later that day, I was shown my bare office: no phone, no books, no bibliographical resources, no files, no staff—indeed, it was temping to think, *no subject*!"

I imagine that this was similar picture at many centers that decided to start scientometrics unit. It is important to trace the initial years to have a proper perspective of the field as the voices from the past have a very contemporary ring. While doing so the role of key actors who helped shape the intellectual and institutional domain of this field need to be properly acknowledged. Institute for Research Policy Studies (ISSRU) established at the Library of the Hungarian Academy of Science in 1978 was possibly the first or among the first organized unit for scientometric research globally. A major outcome of this research centre was the establishment of Scientometrics journal in September 1978. *Establishment of a journal in a new field is a very important part of*

---

[4] Edge, D., *Reinventing the wheel.* In Handbook of Science and Technology Studies, revised edition. (ed. Jasanoff, S., Markle, G., Petersen, J., Pinch, T.), SAGE Publication, Thousand Oaks, CA Inc., 2006, pp. 3-25

institutionalization of a field[5]. It helps develop the research community. Bringing a journal in this field was shared vision of pioneers at that time[6] but from all available evidences I claim that the credit for translation into a product i.e. a journal in this field goes to Hungarian Academics of Science and particularly ISSRU[7].

Hungarian Academy of Science had all ingredients to take this field forward. An established scholar Tibor Braun was given headship of this new centre i.e. ISSRU. Andrus Schubert, PhD in chemistry similar to Tibor Braun joined ISSRU in 1979. Later another key scholar Wolfang Glanzel joined ISSRU. Tibor Braun was Managing Editor from the inaugural issue of the journal and later became Editor-in-Chief. András Schubert in the early days of the journal was Editor of the Bibiliography section and later became the Associate Editor and presently he is Editor of the journal. Wolfgang Glanzel was Editorial Advisor and Editor of Book review section in the early days. Later he became Editor and presently he is Editor-in-Chief of the journal. Intellectual partnership of Tibor, András and Wolfgang and journal editorial responsibility they have shared has been a striking feature of the journal.

The influence of Hungarian Academy of Science can also be observed from the editorial responsibility of the journal. Mikolas Orban, Technical Editor of the journal at that time was also member of the Hungarian Academy of Science. Interestingly he was also from Chemistry, showing the strong intellectual connectivity of this field in Hungary in exploring this new area. J. Farkas, editor of News section at that time was also from Hungarian Academy of Science where he was senior research associate. Another pioneer Peter Vinkler was the Director of scientific publication Data Bank of the Hungarian Academy of Science. One of the Editors-in-chief of the initial period from the inaugural issue onwards was M.T. Beck from Hungary who was also closely associated with Hungarian Academy of Science. He was from Department of Physical Chemistry, Kossuth University, Hungary and surprisingly another scholar from Chemistry!

András Schubert and the Institutionalization of Scientometrics

Unlike strong orientation of many international journals towards research emerging from North countries, from the beginning Scientometrics journal adopted an inclusive approach by encouraging studies from developing countries. In my view, the origin of the journal from Hungary itself a developing country was an influential factor in this regard. One can see the involvement of many scholars from developing econo-

---

[5] Writing on the Editorial Statements in the first issue of *Scientometrics* 1, 1978 Price makes this point "I hope this new stage in the institutionalization of a scientific subfield will produce a positive cybernetic feedback and help us all to be aware of each other's work."

[6] Editorial Statements, *Scientometrics* 1, 1978, pp. 3-8.

[7] See for example the Notes to Contributors in the earlier issues that explicitly informs, "Scientometrics is edited within the Department of Informetrics and Science Analysis of the Library of the Hungarian Academy of Science".

mies and third world countries in this journal from the beginning (from the editorial board composition to articles published from third world countries). Articles were not rejected because English was not up to international standard[8]. Reflections from articles of that period indicate interest of the journal to encourage research that provided research insight of developing countries[9].

The most important phase of any journal is in the beginning years. The journey becomes more difficult when the field is not yet established. There would be a few researchers scattered across some institutes globally driven by their personal research interest in a new area. Only a few lucky ones would get some funding support. All these were typical to Scientometrics and the new journal.

András' key contribution in institutionalizing this field can be seen when one examines his contributions in the initial years. In Volume 16, Issue 1-6, 1981, I see his first published work in this journal along with Tibor Braun on 'Some scientometric measures of publishing performance for 85 Hungarian research institutes'. Then in Volume 4, Issue 2, March 1982 he published a Book Review with Inhalver. *But what I see as his valuable contribution is his influential role in developing the research community in scientometrics. Over a period of time he had individually published bibliographies and with his two intellectual partners Tibor and Wolfgang, a very important series 'World flash on basic research'*. I posit that these two types of contributions were very important in providing the wherewithal for researchers who were working in this field.

Bibliography is a very important document particularly when the field is emerging as literature is scattered across journals and provides the first entry point for any researcher who intends to work in the field. The Scientometrics journal also wanted bibliography to be an important part of the journal. This can be seen from 'Call for bibliographies' in the journal; one announcement was made in Vol 3, Issue 5, September 1981 and subsequently in volume 3, Issue 6, November 1981. This announcement was required as there were only three bibliographies till that date, all by J. Valachy on 'Lotka's law and related phenomenon' in the first issue itself and on 'Mobility in science' in Vol 1, issue 2 January 1979 and 'Nobel prizes' in volume 1, issue 3, March 1979.

After this 'Call for bibliographies', we find one bibliography by Hjeppe 'Supplement to a bibliography of bibliometrics and citation indexing & analysis' in Volume 4, Issue 3, May 1982. Subsequently from Volume 5, Issue 2, March 1983 onward we find András Schubert dedicated effort in bringing out series of bibliographies on 'Quantitative studies of science a current bibliography'. Over the fifteen years period of this study, 17 bibliographies have been published by dedicated individual efforts of András. The bibliographies are an exhaust compilation and provide the needful reference point for scholars. For a new field, one can imagine the 'value' these bibliographies would

---

[8] I am sad to make this point but in many instances I have seen that nowadays articles in journals and conferences inspite of excellent content are rejected because reviewers do not find English as per standard.

[9] It was not only researchers from developing countries who were publishing on scientific trends in developing countries. Many contributions can be seen from researchers from developed economies, see for example J.D. Frame article on "Measuring scientific activity in lesser developed countries" in Volume 2(2), 1980.

have provided and they are still so highly relevant. András was the Editor of the bibliography section from 1984 onwards. One can see András influence in this section as only one bibliography is seen from 1983 onwards by Dabrov and Haitun in 1989.

During 1978-1992 i.e. this study period, 19 contributions have been published under the series 'World flash of basic research'. Two periods have been covered for this study, 1978-1980 and 1981-1985. The first contribution was in Volume 11 (1), 1987 on "Facts and figures on publication output and relative citation of countries of 107 countries 1978-1980". Along with covering the above title in the two time periods for different group of countries, the series also covered data on journal distribution in SCI, international collaboration in science, landscape of national performance in science, indicator datafiles, etc. Along with providing rich data and analysis of key domains of investigation within the field, it also addressed methodological aspects. Under this series many of us particularly recall the key role played by 'Scientometric datafiles" that were published in the journal in 1989 by Schubert, Glanzel and Braun. The datafiles provided a comprehensive set of indicators on 2649 journals covered in all the five years of the 1981-85 period by the SCI. Furthermore, the journals were clustered into subfields, subfields into fields, and each paper was classified into the field/subfield of the journal. This resulted in data files on science fields and subfields and data files on country. In 1990, Volume 16(3) "Scientometric datafiles supplementary indicators on 96 countries for the period 1981-85" was published. I am sure these valuable data files influenced every research scholar who was involved in scientometric research during those days.

In India during the early 1990s, there was a big debate on Science in India going down. A few of us came together to give a more informed view by applying scientometric approach. *The data files were the only source available to us to construct a proper contribution of India in the global research landscape*. Scientometric studies were undertaken around the theme of Indian scientific activity. The journal Scientometrics was the outlet for publication of these research studies. The scientific community in India noticed the importance of these types of studies and in the process Scientometrics also established in the research and policy community in India. NISTADS in particular emerged as key loci of scientometric research in India and received liberal funding support, which helped to develop competency in this field and establish this area of research in policy studies. I imagine similar influence in many developing countries at that time.

Discussion and Conclusion

I always had high regard for András Schubert. But after analysis of first fifteen years of journal Scientometrics my intellectual gratitude to him has increased multi-fold. I realized that how much effort a few persons like András made to institutionalize our field. I also saw broadly his work spanning latter years as well as his research reflected in other journals. He has published over 200 articles and provided new insights to contemporary research topic of a period. His work has been highly cited and he is the only person from our field to be listed in the ISI highly cited researcher database.

András influential contribution is observed in key research debates of the field. In *h-index*, for example, the scientometric community found a new indicator, which could capture the dimension of quantity and quality together. Inspite of elegance of this new indicator, scientometricians have been critically examining its implications and scope and what further refinement can make this indicator more useful. András has also been very active in this and so far I have observed 14 contributions of his under this topic. In other prominent topics also, one finds András extensive contributions, for example, in international collaboration, co-authorship studies, social network analysis, performance analysis, etc. On the other hand he is continuously involved in developing bibliography of this field The title of the bibliographies are now called "Scientometrics: A citation based bibliography". Scientometrics journal is periodically publishing 'World flash on basic research'. The team has expanded supporting the three core partners of the series Tibor, Wolfgang, and András. These two themes I have argued are seminal in building the scientometric research community.

I conclude by admitting that I could scratch only the surface of the intellectual repository of András. A further detailed study is called for to draw more informed insight of András contribution. However, even this limited study reveals the gratitude scientometric community owes to him. In this happy occasion of his 70th Birthday, I am glad we are bringing out this Festschrift in his honour. I am sure he will continue to guide us and make our research relevant to the global science community.

A Brief Reception History on András Schubert's Contribution to the Study of Nanotechnology

MARTIN MEYER
Kent Business School, University of Kent,
Canterbury CT2 7PE, United Kingdom, m.s.meyer@kent.ac.uk

Introduction

This note offers a brief appreciation of the contribution András Schubert has made to the social science study of nanotechnology. Arguably, this will be one of Schubert's less well-known contributions to our field. As it is the contribution that ultimately got the author of this note involved in bibliometrics, it does hold some personal importance.

'Nanoscience and nanotechnology on the balance'

In 1997, Schubert, together with Braun and Zsindely, published a brief paper in Scientometrics that put 'Nanoscience and nanotechnology on the balance' (Braun et al., 1997). This is most likely the first bibliometric study of this emergent area. We explore the reception history of this paper drawing on Thomson-Reuters' Web of Science database, having identified 95 papers citing Braun, Schubert and Zsindely's work which in turn contain 4057 references.

Table 1. Papers cited at least 15 times.

Work	Citations
Braun T, 1997, V38, P321	94
Meyer M, 1998, V42, P195	44
Schummer J, 2004, V59, P425	40
Hullmann A, 2003, V58, P507	34
Meyer M, 2001, V51, P163	27
Porter A, 2008, V10, P715	21
Zitt M, 2006, V42, P1513	18
Zhou P, 2006, V35, P83	18
Meyer M, 2000, V48, P151	16
Huang Z, 2004, V6, P325	15
Leydesdorff L, 2007, V70, P693	15

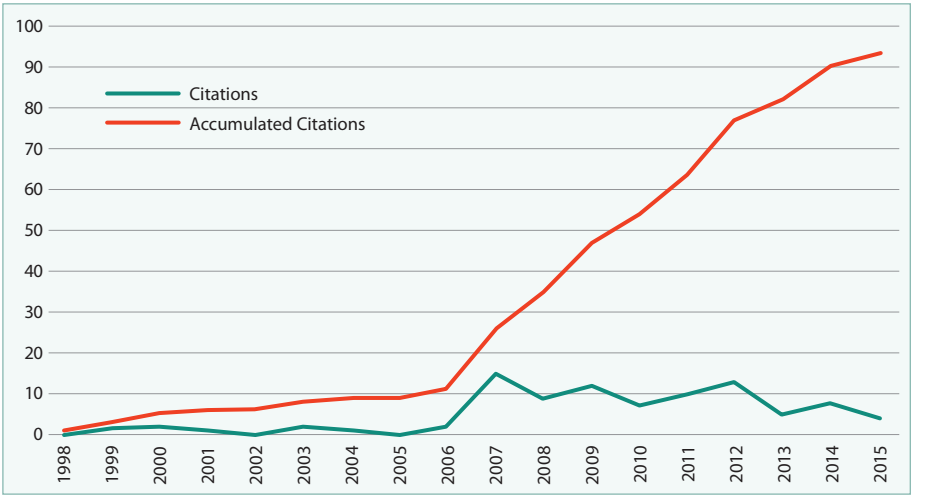


Figure 1. Citation counts, accumulated and by year.

Citations

The paper has an interesting reception history. A look at citation counts (Figure 1) illustrates this. While in the 8 years after its publication there has been a steady stream of individual citations, there has been a pronounced shift in the level of citation from 2007 onwards. As one would expect, Scientometrics as a journal is the most prominent source of citations to the work by Schubert and his colleagues accounting for a total of 41 and nearly half of the 95 citations. While this remains the case before and after 2007, it is noteworthy to see that citation after 2007 were also received from papers in journals outside the library and information science field, such as the *Journal of Nanoparticle Research* as well as *Technological Forecasting & Social Change* (see Appendix 1 for details).

Overlay

An overlay analysis following Leydesdorff and Rafols (2009) reinforces the point of the paper's impact beyond the discipline and specialty (see Figure 2). We can recognise the main area of contribution clearly in the Information and Library Science area as well as in Interdisciplinary Computer Science but there is also considerable impact in the sciences—chemistry, applied physics, medicine—that are likely areas of application in nanotechnology. And it is these areas in which Schubert's work enjoyed increased citation from 2007.

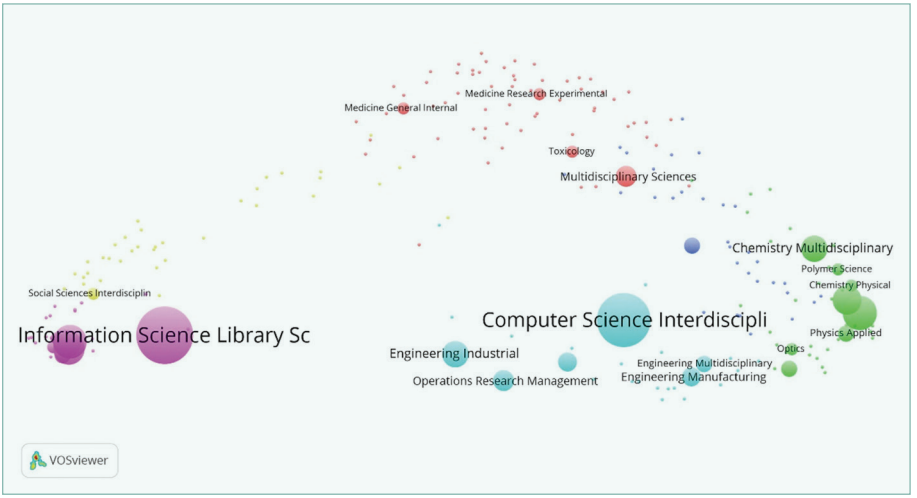


Figure 2. Overlay Map of citing papers.

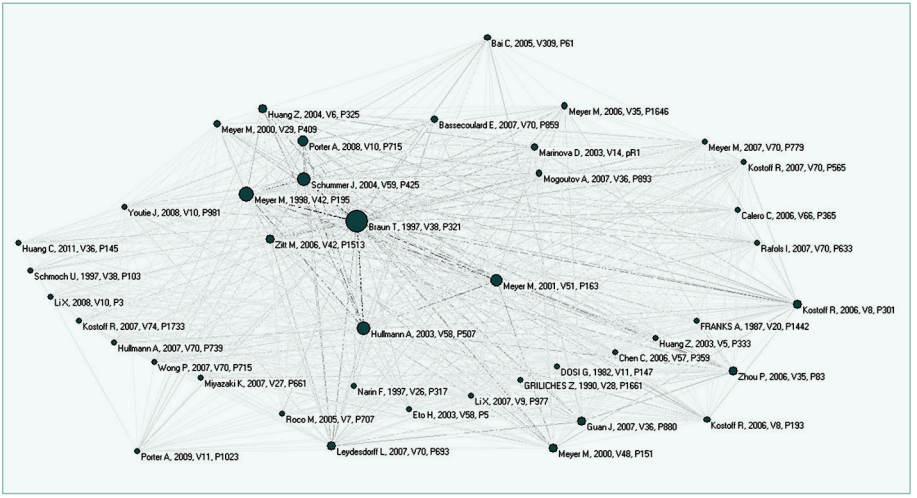


Figure 3. Co-citation map.

Co-citation analysis

A co-citation analysis indicates locates the contribution András and colleagues have made in the context of other related work. The map displayed in Figure 3 is based on papers co-cited more than 5 times. Table 1 lists the most highly cited papers. The map and table show the prominent and influential role András and colleagues' contribution has played.

Conclusions

This brief note aimed to shed some light on one of András' many contributions, arguably one less well known but still quite impactful. It explores a paper that in many ways is typical of András as it is very much the result of a collaborative effort and tracks how it has been received over a period of nearly 20 years. The brief analysis shows that the impact of András' does not stop at the boundaries of our specialty but can be found in other disciplines as well.

References

Braun, T; Schubert, A; Zsindely, S (1997). Nanoscience and nanotechnology on the balance. Scientometrics, 38 (2), 321-325; DOI 10.1007/BF02457417

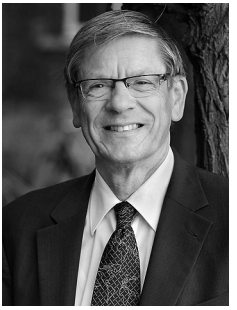
Leydesdorff, L.; Rafols, I. (2009). A Global Map of Science Based on the ISI Subject Categories, Journal of the American Society for Information Science and Technology 60 (2), 348-362.

Appendix 1: Citing Journals & Proceedings

Source Title	1998-2006	Since 2007	Total
SCIENTOMETRICS	7	34	41
JOURNAL OF NANOPARTICLE RESEARCH	0	7	7
TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	0	6	6
JOURNAL OF INFORMETRICS	0	2	2
JOURNAL OF TECHNOLOGY TRANSFER	0	2	2
MALAYSIAN JOURNAL OF LIBRARY & INFORMATION SCIENCE	0	2	2
PICMET '12: PROCEEDINGS—TECHNOLOGY MANAGEMENT FOR EMERGING TECHNOLOGIES	0	2	2
Proceedings of ISSI 2007	0	2	2
RESEARCH POLICY	1	1	2
TECHNOVATION	0	2	2
ADVANCED ENGINEERING MATERIALS II, PTS 1-3	0	1	1
CURRENT SCIENCE	0	1	1
ENGINEERING SOLUTIONS FOR MANUFACTURING PROCESSES IV, PTS 1 AND 2	0	1	1
ENVIRONMENT INTERNATIONAL	0	1	1
ENVIRONMENTAL TOXICOLOGY AND CHEMISTRY	0	1	1
FUTURES	0	1	1
ICIM2014: PROCEEDINGS OF THE TWELFTH INTERNATIONAL CONFERENCE ON INDUSTRIAL MANAGEMENT	0	1	1
INFORMATION PROCESSING & MANAGEMENT	1	0	1
JOURNAL OF THE PAKISTAN MEDICAL ASSOCIATION	0	1	1
KINETICS AND CATALYSIS	0	1	1
KNOWLEDGE ORGANIZATION	0	1	1
MATERIALS RESEARCH-IBERO-AMERICAN JOURNAL OF MATERIALS	0	1	1
NANO	1	0	1
NANOTECHNOLOGY	1	0	1
NEW DIRECTIONS IN REGIONAL ECONOMIC DEVELOPMENT	0	1	1
OPTICAL MATERIALS	0	1	1
PICMET '07: PORTLAND INTERNATIONAL CENTER FOR MANAGEMENT OF ENGINEERING AND TECHNOLOGY, VOLS 1-6, PROCEEDINGS: MANAGEMENT OF CONVERGING TECHNOLOGIES	0	1	1
PROCEEDINGS OF ISSI 2009—12TH INTERNATIONAL CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS, VOL 1	0	1	1
PROCEEDINGS OF ISSI 2011: THE 13TH CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS, VOLS 1 AND 2	0	1	1
PROCEEDINGS OF THE 3RD INTERNATIONAL CONFERENCE ON MATERIAL, MECHANICAL AND MANUFACTURING ENGINEERING	0	1	1
PROCEEDINGS OF THE 5TH INTERNATIONAL CONFERENCE ON INFORMATION ENGINEERING FOR MECHANICS AND MATERIALS	0	1	1
PROGRESS IN POLYMER SCIENCE	0	1	1
R & D MANAGEMENT	0	1	1
SCIENCE AND PUBLIC POLICY	0	1	1
SOUTH AFRICAN JOURNAL OF SCIENCE	0	1	1
SPRINGERPLUS	0	1	1
TECHNOLOGY ANALYSIS & STRATEGIC MANAGEMENT	0	1	1
Grand Total	***11***	***84***	***95***

Schubert A. versus Schubert A.

RONALD ROUSSEAU



Who?

On the one hand we have Schubert A.: a scientometrician and a chemical scientist, aka a chemist, naka a druggist, pharmacist or apothecary; on the other hand we have Schubert A.: not a scientometrician, but maybe he or she (who knows?) is a chemist. Actually he/she has the potential to publish in any field. Does he/she? That said I wonder who is the better? Schubert A. or Schubert A.?

A search in the Web of Science (WoS) will solve everything! Let us refer to the first Schubert A. as András, even if this sounds somewhat familiar. The second one is then just Schubert A. Scientometricians are used to representations, so in this contribution Schubert A. (András) and Schubert A. will be represented by their respective, eh … representations.

András is represented by the set of publications resulting from the following WoS query:

> (AU=Schubert A* AND SO=Scientometrics) OR (AU=Schubert A* AND AU=Glanzel W*) OR (AU=Schubert A* AND AU=Braun T*).

To this set we added another one (31 publications), the result of a secret query which revealed András as a chemist (alchemist?), a lobbyist, a godollist, a collaborator of Albert-László Barabási, a discometrician, and performer of other lesser known activities.

Total catch in the WoS: 182 publications, to which we will refer as the András set.

The other set, the "Schubert A."–set is the result of the query:

> AU=Schubert A* AND PY= (1972-2016), where the restriction on the publications years is chosen in such a way that it matches András'. From this set we removed the András set, leading to the final "Schubert A."- set. Total catch: 738 publications.

Has András any chance to win this battle against an adversary who has an army which is 4 times bigger?

Composition of the two sets

Before starting any form of comparison we look at the composition of the two sets, see Table 1. Some publications are assigned to more than one type so that totals do not match. Percentages are calculated with respect to the total number of different publications. András is more a bibliographer, a reviewer and a book reviewer than Schubert A., who writes more proceedings papers and meeting abstracts.

Table 1. Types of publications

Type	András		Schubert A.	
	numbers	%	numbers	%
Article	119	65,4	431	58,4
Bibliography	19	10,4	0	0
Proceedings paper	18	9,9	173	23,4
Review	11	6,1	13	1,8
Book review	6	3,3	11	1,5
Note	5	2,8	9	1,2
Letter	4	2,2	16	2,2
Editorial material	4	2,2	14	1,9
Meeting abstract	4	2,2	108	14,6
Item about an individual	2	1,1	0	0
Correction	1	0,6	2	0,3
Biographical item	1	0,6	0	0
Correction addition	0	0	4	0,5

Another interesting point to look into is the fields (WoS Subject categories) where these items are published. Table 2 shows for each of the sets the five most-used subject categories. This table shows that András works much more focused in terms of subjects than Schubert A., whose interest is quite dispersed.

And what about nationality? While the András set is clearly Hungarian (96 %), followed by far by Belgium (17 %), the "Schubert A."-set is largely German (52,8%), followed by the USA (26,1%) and Italy (12,5%). Hungary does not occur in the top-10 of countries participating in Schubert A.'s research.

Table 2. Subject categories

Subject category	Numbers	Percentages
	András	
Information science library science	129	70,9
Computer science interdisciplinary applications	116	63,7
Computer science information systems	11	6,0
Chemistry analytical	10	5,5
Chemistry multidisciplinary	9	4,9
	Schubert A.	
Anesthesiology	89	12,1
Engineering electrical electronic	59	8,0
Materials Science multidisciplinary	56	7,6
Plant Sciences	54	7,3
Physics particles fields	34	4,6

A citation study

First, we collected the total number of received citations (as on February 1, 2016), the h-index and the average number of citations per item. Next we restricted the item set to those of the following types: article, review, (contributions to) conference proceedings and note. It is no surprise that for the absolute indicators Schubert A. has higher values than András, and similarly for the h-index. Yet, when it comes to the number of citations per item, András is the better of the two. Moreover, when comparing ratios we observe a clear decreasing trend, consistent with better results for András.

Table 4. Scientometric indicators

	András	Schubert A.	ratio
All			
# items	182	738	4,05
Received citations	4.249	12.676	2,98
h-index	32	53	1,66
Citations/item	23,35	17,18	0,74
Restricted set			
# items	141	583	4,13
Received citations	4.114	12.555	3,05
h-index	32	53	1,66
Citations/item	29,18	21,54	0,74

Scientists are often associated with their most cited or best known publications. Consequently we collected these for the two antagonists. Table 5 shows the top three articles, in terms of received citations, of András and Schubert A.

Table 5. Most-cited articles

	Bibliographic record	Times cited
	András	
1	Barabási, Jeong, Neda, Ravasz, Schubert & Vicsek (2002). Evolution of the social network of scientific collaborations. *Physica A – Statistical Mechanics and its Applications*, 311(3-4), 590-614.	798
2	Schubert & Braun (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, 9(5-6), 281-291.	219
3	Braun, Glänzel & Schubert (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1), 169-173.	213
	Schubert A.	
1	Human Microbiome Project Consortium [consisting of 248 co-authors] (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402), 207-214.	1200
2	Casey, Trainor, Orendi, Schubert, Nystrom, Giedd, Castellanos, Haxby, Noll, Cohen, Forman, Dahl & Rapoport (1997). A developmental functional MRI study of prefrontal activation during performance of a Go-No-Go task. *Journal of Cognitive Neuroscience*, 9(6), 835-847.	569
3	Casey, Castellanos, Giedd, Marsh, Hamburger, Schubert, Vauss, Vaituzis, Dickstein, Sarfatti & Rapoport (1997). Implication of right frontostriatal circuitry in response inhibition and attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36(3), 374-383.	503

This leads us to a refinement of the scientometric indicators shown in Table 4. Indeed, Table 4 is calculated using so-called inflated counts, as each co-author of each article received a full score. Professional scientometricians know that one should better use a form of fractional counting. If the exact contribution of each co-author is not known this is preferably complete-normalized counting (equal credit to each co-author). Consequently we calculated fractional scores for András and Schubert A., based on their top three publications.

For András the score is: 798/6 + 219/2 + 213/3 = 313.5

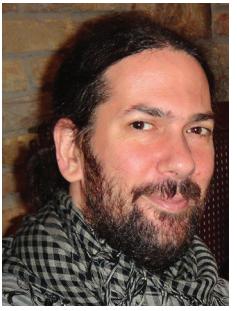
For Schubert A. the corresponding score is: 1200/248 + 569/13 + 503/11 = 94.34

Conclusion and discussion

It is clear that, when using the proper methodology, András is the better scientist. Surely András and Schubert A. have also publications outside the ISI-Thomson Reuters empire. A study of these contributions is kept for a following publication, noting that celebrating an octogenarian might be a good opportunity for such an endeavour.

Performer Name Length in Music as a Factor of Success

GÁBOR SCHUBERT[1] & MIHÁLY SCHUBERT[2]
[1] *Stockholm University Library, Stockholm, Sweden*
[2] *ELTE Radnóti Miklós School, Budapest, Hungary*



Abstract: Billboard Hot 100 year-end charts were analyzed in order to find a possible relationship between the length of performer names and success of musical works.

Introduction

In this study we are trying to find a relationship between the length of performer names and the success of musical works. We are also attempting to assess the musical success of András Schubert (and his band): "MedveCukor Jazz (Band)".

There are not many studies available about the relevance of the length of different names in general, or specifically in the context of music.

Among the studies of person names we should mention the seminal works of Cabe from the 1960s [Cabe 1967, Cabe 1968], who found no evidence that the length of family names operate as a factor in mate selection.

Interesting correlations were found by several authors between the length and popularity of given names. One study suggests that longer given names are associated by success [Mehrabian 1993].

Another researcher could found a relationship between given name length and popularity, although he could not show a correlation between popularity and the frequency of letters in a name from the right side of the QWERTY keyboard. [Thogmartin 2013].

There has been attempts to use the name length of presidential candidates of the Unites States in order to foresee the winner of the presidential elections, but the use "the longest last name wins"-method is not suggested in the election forecasts [Lewis-Beck 1985]

In another study no lengthening of performer names were observed during the last decades [Lamere 2012]. The longest

performer name ever according to this study is "*Tim and Sam's Tim and the Sam Band with Tim and Sam*" with 51 characters (spaces included). It should be noted that the original name of András Schubert´s band "*Medvecukor Jazz Band*" with 20 characters would qualify into the top 20 longest artist names between 2000 and 2004 according to the previously mentioned study.

A recent founding from the analysis of the artist names in the Spotify streaming music service suggest that shorter artist names are more often used ambiguously by several different artists. [Machado Gonzalez 2012]

Data collection

We chose the year-end top 100 singles charts published by the Billboard magazine. This chart considered to be the industry standard record chart in the United States and has probably the largest influence in the world of popular music. The year-end chart contains the 100 most popular songs from a given year.

The time frame was chosen as 1946-2016 because of two reasons: 1) this time interval overlaps with the first 70 active years of András Schubert, and 2) the Billboard singles charts were first published in the 1940s.

Data was downloaded from the website "billboardtop100of.com/" [Billboard 2016]. This data source was chosen instead of the official Billboard website because of the simple table presentation of the charts. The downloaded data was used as is, no further text manipulation or correction was applied.

In this first study we chose to sample one year per decade to reduce the amount of data. The following charts were used: 1946, 1956, 1966, 1976, 1986, 1996, 2006, and 2015. It should be noted that the chart from 1946 contained only 48 songs, and 2015 was chosen instead of 2016, because that is the last available chart. All data is available in the supplementary information.

The collected data set contains 748 songs and 585 unique performer names. The names were treated separately even if the same person was involved in different constellations, for example "*Nicki Minaj feat. Drake and Lil Wayne*" and "*Nicki Minaj feat. Drake, Lil Wayne and Chris Brown*" were counted as distinct performers. Some performers appear multiple times in the same year-end chart, and some appear even in several different decades.

Performer name lengths were calculated by counting the number of letters in the performer name string, including spaces and any special characters.

Results

Overall statistics

The average name length of the performers of the 748 songs in the data set were 15. A histogram of all performer name length is shown in Figure 1.

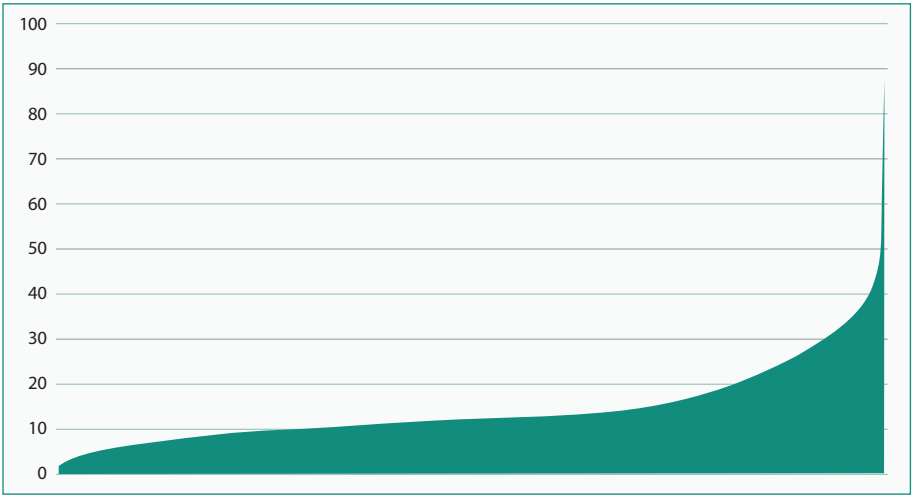


Figure 1. Histogram of performer name lengths for all the 748 song in the observed data set

The shortest performer name was 2 characters: "3T", this performer reached position 43 with a song called "Anything" in the 1996 top chart.

The longest performer name was "*Macklemore and Ryan Lewis feat. Eric Nally, Melle Mel, Kool Moe Dee and Grandmaster Caz*" with 81 characters, reaching position 84 in 2015 with a song called "Downtown". The longest performer name, which was not in the format of "*XXX feat. YYY*" or "*XXX and the YYY*" was 37 characters: "*Wing and a Prayer Fife and Drum Corps*" a band which reached position 68 with a song called "Baby face" in the 1976 chart.

There were 20 performers who appeared on charts from different decades, using exactly the same performer name, and 2 of them managed to appear on charts from three consecutive decades: *Madonna* (1986, 1996, 2006), and *Frank Sinatra* (1946, 1956, 1966).

Name length and success

There are no well-defined quantitative measures for success. Although the order of songs in the chart gives a natural ranking: number 1 is probably more successful than number 100, that is also true that all the performers who appear on these year-end charts should be considered successful. Therefore we investigated different types of success.

Performers with the highest number of songs during the entire investigated period

110 of the 585 unique performers succeeded to have at least two songs on any of the charts. 10 of the performers had at least 4 songs on the chart, the name of these performers can be seen in Table 1.

Table 1 Performers with the most songs on any of the charts

Name	Number of songs
Perry Como	8
Elvis Presley	5
Madonna	5
The Beatles	5
Fall Out Boy	4
Frank Sinatra	4
Janet Jackson	4
Pat Boone	4
Taylor Swift	4
The Beach Boys	4

According to this measure it is possible to define three distinct levels of success:
- Top level success: Performers with at least 4 songs on any of the charts.
- Medium level success: Performers with 2 or 3 songs on any of the charts.
- Low level success: Performers with just one song on a chart.

Table 2 shows some statistics about the length of performer names in these three distinct groups:

Table 2

Success level	Number of performers	Name length			
		Average	Median	Min	Max
Top	10	11	12	7	14
Medium	100	13	12	3	41
Low	475	16	13	2	87

The results presented in the previous table suggest that there is a weak reciprocal relationship between the success of a performer and the length of its name according to this measure.

Performers with highest rankings within a single year-end chart

In this approach we defined two distinct sets of songs in each year-end chart: one for the top 10 ranked songs and one for the rest of the songs between rank 11 and 100 (11 and 48 for the 1946 chart).

The average performer name lengths for each investigated years are shown in Table 3.

Table 3 Average performer name length for top 10 and the rest of the charts

Year	Top10	11-100
1946	15	15
1956	11	14
1966	16	14
1976	15	13
1986	20	12
1996	15	14
2006	23	18
2015	13	18

According to this measure it is hard to find a significant relationship between success and the length of performer name. Although there were more years when the top10 performers had longer average name, than the rest of the chart, and only on three charts had the top 10 performers longer average names than the overall average.

Average name length for different chart positions

Another approach could be to calculate the average name length of the performers for all the 100 distinct positions of the Top 100 charts. The results are shown on Figure 2.

These results show that the performers for number one hits in the observed data set has the second longest average name (22) compared to the average performer name

Figure 2. Average performer name length at different chart positions.

length for the other position of the chart. This is partially due to the fact that the 1986 number one song was called "That´s what friends are for" performed by "*Dionne and Friends (Dionne Warwick, Gladys Knight, Elton John and Stevie Wonder)*"

Predictions for András Schubert´s performance on the Billboard charts

The current performer name used by András is "MedveCukor Jazz" according to his personal homepage [Medvecukor, 2016]. The length of his performer name is 15 characters, which is actually the same as the average performer name length of the 748 songs in the observed data set. This shows that he carefully chose his performer name in order to maximize the probability to obtain a chart position on the Billboard Hot 100 year-end chart.

Concluding remarks

In this short study we tried to elucidate the complex nature of success in musical performance via a simple analysis of Billboard Top 100 charts. We used a basic approach and analyzed only a part of the available data. The results are promising, but further research is needed to obtain a practical guide to find the optimal performer name length.

References

Billboard 2016. http://billboardtop100of.com. Accessed on 2016-01-19

Cabe 1967: Cabe PA, "Name length as a factor in mate selection", Psychological Reports, 21, 678 (1967)

Cabe 1968: Cabe PA, "Name length as a factor in mate selection: age controlled", Psychological Reports, 22, 794 (1968)

Lamere 2012. Lamere P, "Have artist names been getting longer?", Music Machinery blog, (2012). Accessed on 2016-01-19. http://musicmachinery.com/2012/01/07/have-artist-names-been-getting-longer/

Lewis-Beck 1985. Lewis-Beck MS, "Election Forecasts in 1984: How Accurate Were They?", Political Science & Politics, 18, 53-62 (1985)

Machado González A, "Recognizing artist ambiguity with machine learning techniques", Master´s Thesis, Lund University (2012)

Medvecukor 2016. http://schubaa-mus.weebly.com/medvecukor-toumlrteacutenelem.html. Accessed on 2016-01-19

Mehrabian 1993. Mehrabian A, "Affective and Personality Characteristics Inferred from Length of First Names", Personality and Social Psychology Bulletin, 19, 755-758 (1993)

Thogmartin 2013. Thogmartin WE, "The QWERTY Effect Does Not Extend to Birth Names", Names, 61, 47-52 (2013)

The h-index of German Nobel Laureates in Physics: Historical Contexts*

HANS-JÜRGEN CZERWON
Neumühler Str. 23, D-16348 Wandlitz, Germany
hans.czerwon@gmx.de

Abstract: In the last ten years, the h-index proposed by Jorge E. Hirsch has become a powerful indicator for evaluating the publication performance of individual scientists quantitatively. In the present paper, investigations are made on the publication activity of German Nobel laureates in physics so far as it is visible in the Web of Science. The author suggests that in addition to the introduction of reference standards for different time periods, greater consideration should be given to details of a researcher's life in order to evaluate his or her performance objectively.

Introduction

Bibliometric indicators, i.e. publication and citation indicators, have been proven to be an essential tool to assess the performance of researches in natural and life sciences. This is especially true for the past two to three decades. The standard set of indicators reflecting the impact of scientific publications (total citation count, mean number of citations per paper, number of highly cited papers etc.) was supplemented in 2006 with the h-index introduced by Hirsch: "*A scientist has index h if h of his or her N_p papers have at least h citations each and the other (Np—h) papers have ≤ h citations each*" (Hirsch, J.E., 2006). Therefore, this indicator describes the most productive core of the publication output of a scientist and informs about the number of papers in the core. Papers outside of this core are not considered. From the definition follows immediately that an overestimation of single or few highly cited paper is avoided. In particular, the index favors researchers who continuously publish influential articles. It is obvious that the h-index is time-dependent and can only increase with the years elapsed since the first published paper of a scientist till the present. Hirsch concluded from empirical data

* Dedicated to András Schubert on the occasion of his 70th birthday

that the index follows approximately a linear behavior with time (especially during the active period of a researcher) and depends on the specific research field (Hirsch, J.E., 2007).

Already in his seminal paper Hirsch computed h-indices for individual physicists. He classified physicists with an h-index of 20 after 20 years of scientific activity as successful researchers. According to Hirsch, an h-index of 40 after 20 years of scientific activity characterizes outstanding and a value of 60 after 20 years truly unique scientists. For physicists who received the Nobel Prize between 1985 and 2005, Hirsch found h-indices in the range from 22 to 79 with an average value of 41 and a median of 35.

A consequence of the h-index defined above is that it underestimates the importance of authors with a small number of papers of which many received a high citation rate.

An extreme case of underrated scientists (by bibliometric indicators) is the British biochemist Frederick Sanger (1918-2013) who won the Nobel Prize for chemistry twice. On the one hand, "only 75" of his papers are covered in *Thomson Reuters/ISI Web of Science* database in the timespan 1942-2004 (h-index=43), and therefore he was not included in the h-index ranking of living chemists including those with a score of 55 and higher (Schaefer, H., Peterson, A., 2007). On the other hand, Sanger published in 1977 one of the most cited papers in the history of science—on DNA sequencing with more than 66 000 citations (Sanger, F. et al., 1977).

Many variants of Hirsch's h-index have been proposed since 2006 in order to extend and overcome the shortcomings and limitations of the original index (Alonso, S. et al., 2009). In this context, especially Egghe's g-index is to be mentioned (Egghe, L., 2006). Actually, there are hundreds of articles on this subject, theoretical as well as applications to a variety of problems of practical interest, for example, to identify the most influential scientific journals (Braun, T. et al., 2006).

In the following we investigate in a brief report the publication activity of German Nobel Prize winners in physics. The laureates are in accordance with Harriet Zuckerman members of the "ultra-elite" of science, which is characterized by specific patterns of productivity, publication practices and coauthorships (e.g. Zuckerman, H., 1967, 1977). Eugene Garfield concluded from many of his studies that author citation rankings are an effective method for identifying both past and present Nobelists as well as laureates-to-be (e.g. Garfield, E., Welljams-Dorof, A., 1992).

Due to the fact that the German Nobel Prize winners lived in different eras of science history, the question arises, whether and to what extent data which we obtained are comparable. Furthermore, the problem should be discussed in what way specific circumstances have influenced the scientific productivity of Nobel laureates.

Data collection

The bibliographic data presented in this study are based on the *Thomson Reuters/ISI Web of Science* (WoS) which covers a carefully selected set of important scientific journals dating back to 1864. Note that the set of source items has been extended continuously by *Thomson Reuters* in recent years. The WoS database in the *General Search mode* was used. In some

cases, it was a difficult task to identify the correct set of publications, especially among different authors with the same surname and the same first initial (e.g. W. Paul). Once the publications of a scientist are identified, WoS provides the h-index and other indicators.

In all cases, the period has been determined in which the scientist was "visible" by WoS publications. In general, this period can be regarded as the most productive life stage of a scientist.

Biographical and other information about the Nobel laureates who we have investigated in more detail were taken from different sources.

German Nobel laureates in physics

In our study, all German-born Nobel Prize winners in physics were taken into account, especially also those who did not have the German citizenship in the year in which they were honored. Throughout history 29 German physicists have been awarded the Nobel Prize. The first one was Conrad Wilhelm Röntgen in 1901. The last prize winner was Peter Grünberg in 2007 "for the discovery of giant magnetoresistance".

Among the German laureates are four emigrants to the Unites States after the Nazis' Machtergreifung (seizure of power) on 30 January 1933: Hans Bethe after a stay in England in 1935, Albert Einstein, James Franck, Otto Stern. Maria Goeppert-Mayer (née Maria Göppert) married in 1930 the U.S. chemist Joseph E. Mayer and left with him Germany. Max Born emigrated in 1933 to the United Kingdom. After his retirement, he returned to Germany in 1954.

Hans Georg Dehmelt and Herbert Kroemer (born as Herbert Krömer) left Germany during the 1950s and worked since then in different positions in the U.S. After receiving his PhD, Horst L. Störmer moved to the U.S. in 1977. In 1990, Wolfgang Ketterle came to MIT, Cambridge, Mass., as a postdoc.The German laureates reached the zenith of their scientific productivity in different periods of the 20th century, which considerably differ with respect to the patterns of scientific communication. It is evident that publication and citation indicators for physicists of the early 20th century are not simply comparable to those obtained for scientists from the second half of the century. Therefore, for certain periods of time typical publication and citation cultures should be taken into account in an appropriate way to evaluate scientists. In a study by W. Marx et al. reference standards and reference multipliers have been proposed, which enable the comparison of the citation impact of physics papers from the period at the beginning of the 20th century and the impact of contemporary papers (Marx, W. et al., 2010). With reference to Derek de Solla Price, the authors divide the history of physics into two eras: the era of "Little Science" (beginning about 1900, mainly individual science) and the era of "Big Science" (beginning in the mid-1950s, mainly team science). An example for the latter are the large multinational teams of physicists in high-energy physics at CERN and other research centers. This phenomenon emerged in the late 1950s.

It should be noted that Nobel Prizes are often awarded long after the maximum productivity and the essential discoveries of scientists. For example, Ernst Ruska received the prize about half a century after "his fundamental work in electron optics, and for the design of the first electron microscope". Other physicists received the Nobel Prize relatively

shortly after their discoveries, e.g. Georg Bednorz in 1987 already one year after the "important break-through in the discovery of superconductivity in ceramic materials" in 1986.

Table: German recipients of the Nobel Prize in physics: year of award, number of WoS publications, h-index (Laureates are arranged in reverse chronological order of the year of Nobel Prize awarding.)

Nobel laureate	Nobel Prize (year)	WoS publications	h-index
Peter Grünberg (* 1939)	2007	194	38
Theodor W. Hänsch (* 1941)	2005	652	87
Wolfgang Ketterle (* 1957)	2001	217	78
Herbert Kroemer (* 1928)	2000	269	54
Horst L. Störmer (* 1949)	1998	253	73
Hans Georg Dehmelt (* 1922)	1989	155	43
Wolfgang Paul (1913-1993)	1989	50	19
Georg Bednorz (* 1950)	1987	134	43
Gerd Binnig (* 1947)	1986	197	53
Ernst Ruska (1906-1988)	1986	53	16
Klaus von Klitzing (* 1943)	1985	365	56
Hans A. Bethe (1906-2005)	1967	314	63
J. Hans D. Jensen (1907-1973)	1963	47	18
Maria Goeppert-Mayer (1906-1972)	1963	30	19
Rudolf L. Mößbauer (1929-2011)	1961	147	34
Walther Bothe (1891-1957)	1954	118	22
Max Born (1882-1970)	1954	202	41
Otto Stern (1888-1969)	1943	53	22
Werner Heisenberg (1901-1976)	1932	135	41
Gustav Hertz (1887-1975)	1925	38	12
James Franck (1882-1964)	1925	95	30
Albert Einstein (1879-1955)	1921	172	58
Johannes Stark (1874-1957)	1919	254	16
Max Planck (1858-1947)	1918	90	14
Max von Laue (1879-1960)	1914	74	12
Wilhelm Wien (1864-1928)	1911	56	11
Ferdinand Braun (1850-1918)	1909	18	5
Philipp Lenard (1862-1947)	1905	36	17
Wilhelm Conrad Röntgen (1845-1923)	1901	6	5

All German Nobel Prize winners are listed in the Table. As can be seen from the Table, there are large differences in the publication productivity and Hirsch index. Even within an age cohort of physicists, large differences occur. In particular, this is valid for

the German laureates who were born in the 19th century or even for prize winners who were honored jointly for a discovery (James Franck and Gustav Hertz).

Similar observations were made by Cardona[1] and Marx for Russian Nobel laureates (Cardona, M., Marx, W., 2006). It should also be mentioned in this context that our results differ slightly from the results of Cardona and Marx for some German Nobelists in physics (Cardona, M., Marx, W., 2008). This is partly due to the fact that scientific papers of a few of the laureates from the first half of the 20th century are still frequently cited. The Hirsch index of Albert Einstein has grown over the last years from 50 to 58 (cf. Hirsch, J.E., 2011)[2].

To make clear the intention of the present paper, we compare in the following the publication activity of two Nobelists.

A comparison: Hans A. Bethe and Maria Goeppert-Mayer

Both Nobel laureates in physics, the theoreticians Maria Goeppert-Mayer and Hans Albrecht Bethe were born in the same year 1906 and received the prize in the 1960s. However, the careers of the laureates were very different.

At Munich in the winter of 1929 the young Bethe wrote what he considered to be his best paper: "On the theory of the passage of fast corpuscular rays through matter" (Brown, G.E., Lee, S., 2009). It was Bethe's habilitation thesis published in *Annalen der Physik* (Bethe, H., 1930). Up to and including 2015, this publication has been cited more than 3100 times. It became Bethe's most cited paper followed by a publication in *Zeitschrift für Physik* "Zur Theorie der Metalle" with about 2350 citations (Bethe, H., 1931).

Hans Bethe published about 314 articles covered by WoS in the time period from 1927 till 2007, i.e. his scientific publication activity spanned eight decades. This is an exceptionally long period. His academic career at Cornell University (Ithaca, N.Y.) and his publication activity were interrupted only 1941-1945 during the Second World War and a stay in Los Alamos, New Mexico, because of his involvement in the Manhattan Project and the necessary confidentiality requirements. His last paper was published by colleagues posthumously in 2007. Bethe's Hirsch index of 63 is extremely high for physicists of his generation, and he has written numerous highly cited articles—many of them as a single author.

The Nobel Prize in physics 1967 was awarded to Hans Bethe "for his contributions to the theory of nuclear reactions, especially his discoveries concerning the energy production in stars". Bethe said in his speech at the Nobel banquet: "You have given me the Prize I believe for a lifetime of quiet work in physics rather than for any spectacular single contribution." This self-assessment of Bethe is probably correct, because among his many pioneering contributions to physics the paper from 1939 on the energy production in stars is only one among numerous seminal publications.

---

[1] According to the ISI Citations Web Database, Philadelphia, Penn., Manuel Cardona was one of the eight most-cited physicists constantly since 1970; he died in 2014.

[2] By the way, Einstein's paper about the so-called EPR paradox is his most cited one (currently 6750 citations).

One half of the Nobel Prize 1963 was awarded jointly to Maria Goeppert-Mayer and J. Hans D. Jensen „for their discoveries concerning nuclear shell structure". Goeppert-Mayer was the second female Nobel Prize winner in physics, after Marie Curie, who had received this prize sixty years earlier. Goeppert-Mayer is also the last woman, who won the Nobel Prize in physics.

She was a student of Max Born in Göttingen and was early well trained in the mathematical concepts required to understand quantum mechanics. Together with Max Born, she published in 1931 during a summer stay in Göttingen (she had already moved to the U.S.) an article in the *Handbuch der Physik*: "Dynamische Gittertheorie der Kristalle" (Sachs, R.G., 1979).

After her final move to the U.S., she did not receive a regular appointment to a staff position in a university, primarily due to the economic situation during the Great Depression. She followed her husband to the Johns Hopkins University (Baltimore, Maryland), where she had a very modest assistantship, which gave her access to the University facilities. She participated in the scientific activities of the University and had also the opportunity to present some lecture courses for graduate students. During this time, she specialized in the field of theoretical chemical physics and published a few, but very important papers (e.g. on double beta decay).

In 1939 Goeppert-Mayer and her husband both received appointments in chemistry at Columbia University, New York. At first her position was even more tenuous than at the Johns Hopkins University. In 1940, she published with J.E. Mayer their well-known monograph *Statistical Mechanics*. In December 1941, she was offered a position as a lecturer at a college, and since spring 1942 she worked on the separation of uranium isotopes for the atomic bomb project in a research group led by Harold Urey. After the war Goeppert-Mayer's interests centred increasingly on nuclear physics, and in 1946 she became a voluntary Associate Professor of Physics in the Institute for Nuclear Studies (later Enrico Fermi Institute) at the University of Chicago. At the same time, she joined the newly established Argonne National Laboratory as Senior Physicist. During this very productive period of her life she made her major contribution to the field of nuclear physics, the nuclear shell model and the importance of spin-orbit coupling for explaining this model. A similar theory was developed almost simultaneously and independently by J.H.D. Jensen and coworkers in Heidelberg. In 1951, during a visit to Germany, she and Jensen had the opportunity to start a collaboration, which culminated in the publication of their book *Elementary Theory of Nuclear Shell Structure* (Goeppert-Mayer, M. et al., 1955). In 1960 she accepted a regular appointment as a full Professor of Physics at the University of California (San Diego).

Unlike Hans Bethe, she published in the period from 1929 (her first publication) to 1965 (her last publication) only a few papers. Her bibliography compiled by Robert G. Sachs includes 41 items (Sachs, R.G., 1979). 30 items published 1929-1964 are covered by the WoS database. These publications are distributed discontinuously during her academic career. Thus, between 1941 and 1946, for example, she published only two papers. Because of her low publication productivity, one could not expect a very high Hirsch index. On the other hand, her h-index has a quite high value of 19, also in comparison with other German laureates.

It is a difficult task to assess in detail the achievments of scientists of the Nobel elite. Our comments are intended to show that also the members of this elite should not be lumped all together.

References

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F. (2009). H-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4), 273-289.

Bethe, H. (1930). Zur Theorie des Durchgangs schneller Korpuskularstrahlen durch Materie. *Annalen der Physik*, 397(3), 325-400.

Bethe, H. (1931). Zur Theorie der Metalle. I. Eigenwerte und Eigenfunktionen der linearen Atomkette. *Zeitschrift für Physik*, 71(3-4), 205-226.

Braun, T., Glänzel, W., Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1), 169-173.

Brown, G.E., Lee, S. (2009). Hans Albrecht Bethe, 1906-2005—A Biographical Memoir. *National Academy of Sciences*, Washington D.C.

Cardona, M., Marx, W. (2006). Vitaly L. Ginzburg—a bibliometric study. *Journal of Superconductivity and Novel Magnetism*, 19(3-5), 459-466.

Cardona, M., Marx, W. (2008). Max Born and his legacy to condensed matter physics. *Annalen der Physik*, 17(7), 497-518.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131-152.

Garfield, E., Welljams-Dorof, A. (1992). Of Nobel class: a citation perspective on high impact research authors. *Theoretical Medicine*, 13(2), 117-135.

Goeppert Mayer, M., Jensen, J.H.D. (1955). *Elementary Theory of Nuclear Shell Structure*. New York, J. Wiley.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, 102(46), 16569-16572.

Hirsch, J.E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences of the USA*, 104(49), 19193-19198.

Hirsch, J.E. (2011). On the value of author indices. *Physics Today*, 64(3), 9.

Marx, W., Bornmann, L., Cardona, M. (2010). Reference standards and reference multipiers for the comparison of the citation impact of papers published in different time periods. *Journal of the American Society for Information Science and Technology*, 61(10), 2061-2069.

Sachs, R.G. (1979). Maria Goeppert Mayer, 1906-1972—A Biographical Memoir. *National Academy of Sciences*. Washington D.C.

Sanger, F.. Nicklen, S., Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA*, 74(12), 5463-5467.

Schaefer, H. (2007). H-index ranking of living chemists. *Chemistry World*, 23 April 2007 (online, updated 12 December 2011).

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

Zuckerman, H. (1977). Scientific elite: Nobel laureates in the United States. New York: Free Press.

The Eponym's Curse

VIRGINIA TRIMBLE
Dept. of Physics & Astronomy, Univ. of California, Irvine CA 92697-4575 USA
and Queen Jadwiga Observatory, Rzepinnik Biskupi, Poland

Abstract: Your dictionary's eponym is a real or legendary personage from whom a nation, city, epoch, theory, etc. is reputed to derive its name, like Pelops for Peloponesus.

The scientific cases include effects, laws, equations, rules, theories, processes, units, chemical elements, and devices, and the people are all, we think, real, though by no means all rightly credited. The sections that follow look briefly at many, carefully at a few, examples. An effort was made to decide whether things with both every-day, descriptive names and eponymous ones are better known by the eponyms by asking Google (c) how many listings it found for each. Those numbers follow some of the names and equivalent descriptions below. For most of them, at least the first page picks up the right idea, but in no case have I checked all the entries, which range from hundreds to millions, and I do not encourage you to try to find out about gravitational radiation detectors called Weber bars or about the physicist A. Dose this way. Late in the preparation of this paper (and a closely related talk for the American Chemical Society), I encountered "Eponym as Placebo" by psychologist Edwin G. Boring (1964), who had beaten me to many of the ideas, and whom I have, I hope, properly credited below. Most of the examples following come from chemistry (because it is the honoree's original field) and astronomy (because it is mine).
Keywords: eponym, effect, process (etc.), Stigler's law, kindle cole principle

1. Introduction and landscape of examples

Yes, that's an Erlenmeyer flask (787,000), but which is the Plucker tube (330,000), the Pitot tube (516,000) and the Buchner funnel (345,000)? Giambattista Riccioli, an Italian Jesuit, probably started it all by blanketing his 1651 lunar map with the names of dozens of Greek, Roman, early Christian, medieval, and Renaissance scholars, deftly assigning the largest, best illuminated craters to Ptolemy and other geocentrists, while Galileo and Hypatia are lucky to get even tiny ones. He gave himself a "medium".

The 21st century version has been the naming of mountains, craters, plains, and everything else on planets and other moons according to complex rules devised by the International Astronomical Union (which a1so names comets and asteroids under different rules, and exop1anets coming next). But it was the International Union of Pure and Applied Chemistry that gave 13 of the most recent 19 elements to people rather than places, properties, or characteristic reactions (well, how do you decide if one atom is a noble gas?). They have four more (113, 115,117, 118) to play with soon.

The Heck process (59,900,000 but only 33,500 for Mizoroki-Heck) you probably remember, because he just died in October, 2015, and perhaps Ostwald (345,000), Pasteur (11,600,000), and Haber (1,140,000). But Birkeland-Hyde and Birkeland-Eyde (30,700 together and confused), Bucher (869,000), Caro & Franke (82,800), Coslett (419,000), Fischer-Tropsch (411,000), Mercerized (497,000), Parker (157,000,000 so clearly confused with Parker House rolls or something), Schoop (73,600), and Serpek (4270)? All of these are things that were either developed or expanded to industrial scale during Wor1d War I and are, variously, for nitrogen fixation, rust proofing, producing hydrogen gas, or making thread smoother.

Hall has both a process (555,000,000) and a current (890,000,000), though they are not the same Hall. Degrees centigrade have become Celcius and the atomic mass unit (amu) is fading into the Dalton. Nearly all the units of electromagnetism, from Ampere to Weber, are eponyms, with very little correlation between what folks did and what is named for them.

Where's the harm if the scientific community wishes to honor its own in this fashion, or even, as less generous commentators have said, things get named for the first person who fails to credit his predecessors? This last is one version of Stigler's law of eponyms (Stigler 1980) which he credits to the late Robert K. Merton. First is the loss of information, as per Celcius, Dalton, and the threatened renaming of the Cepheid period-luminosity relation (57,900) as Leavitt's law (47,800). An ideal gas law (3,930,000) tells you what it is good for far better than Boyle's (464,000) and Charles' (41,300,000) law. Second is the simplified, even erroneous, history encapsulated, a point made firmly by Boring (1964) and explained by him as a failure of human memory.

A classic astronomical example is Hubble's law (122,000) for the velocity—distance relation (2,750,000) in cosmology, so called because it was discovered by Knut Lundmark. This one has a considerable literature, including Seitter & Duerbeck (1990) who support Wirtz; Nussbaumer & Bieri (2009, 2011) making the case for Le Maître; and a summary of earlier pleadings by Trimble (2014), who thinks Lundmark was the discoverer, but favors keeping Hubble's name. Cross-checking at this point revealed an Inconvenient Truth. I had realized that one had better check numbers of Google entries for eponym and straight versions of laws etc. at the same moment, because the Google readers prowl unceasingly and are bound to find additional entries. But in fact the numbers are not stable: search on two different days for Hubble's law and velocity-distance relation came up with the pairs (113,000; 2,830,000), (122,000; 2,750,000). That is, one shrunk and the other grew. But I would also not want to have to defend the laws of Biot-Savarin (66,400, having some ambiguation with a chef,

and 166,000 for Biot-Savart), Clausius-Clapeyron (140,000), Dulong-Petit (59,000) or Guy-Lussac (226,000, or maybe Gay-Lussac).

And the third problem, what happens when either the persona or the concept is declared non-grata? A few years ago, Debye had a posthumous narrow escape from losing his institute, his Prize, and perhaps even his degree. Luckily the fuss had died down before his length (679,000 for Debye length) was endangered. The excessive coziness of which he was accused was with Nazis not female graduate students. Meanwhile, anyone for Blanc's rule (40,700) or Mattauch's law (257,000), both now known to be false.

2. What if you are the person eponymized?

Here are two cases from astronomy and one from physics where I know the answer, and one from chemistry, where I do not. The situation seems to have changed a bit with time. The late Martin Schwarzschild was the son of Karl Schwarzschild, best known for the Schwarzschild solution (154,000 vs. 5,980,000 for black hole equation) to Einstein's equations (19,400,000), oh, all right, the Einstein-Hilbert equations (125,000). But this story belongs to the Schwarzschild criterion (90,700) for convective instability, which Martin always called simply the criterion for convective instability (412,000). A student forced him into a nomenclaturial corner in about 1965, hoping to hear "Schwarzschild criterion," but got instead "my father's criterion!".

A few years later, I asked Richard Feynman "what do you call the diagrams?" Everybody else calls them Feynman diagrams (375,000), and they are an important way of describing reactions in particle physics. "The diagrams," he said.

Moving on to the present, the Sunyaev-Zeldovich effect (89,200) is a slight distortion of the spectrum of the cosmic microwave background radiation (CMB) left over from the hot Big Bang. Yakov Borisovich Zel'dovich is sadly no longer with us. He would be 102 this year, but died at 73. Rashid Sunyaev, however, is still a very active member of the cosmology community, often asked to talk about the CMB and its meanings. He says Sunyaev-Zeldovich effect. This is not a problem I am ever likely to have, but I note that, in the second edition of the Biographical Encyclopedia of Astronomers (Hockey et al 2014), Zel'dovich is preceded immediately by Edvard Hugo von Zeipel (who had a theorem, 4920), which was also derived by Edward Milne of the Milne-Eddington approximation (14,500), and Zeeman's effect (409,000) was explained by Konrad Lorentz of the Lorentz Fitzgerald contraction (36,200).

Tiresomely, Jeffrey I. Seeman (2016), though he quotes from his interviews with Roald Hoffmann, does not say what the latter called the Woodward-Hoffman rules (124,000) which predict the outcome of certain reactions based on molecular orbital symmetry (conservation of orbital symmetry, 83,200).

Whimsey is possible. Jan Oort (1900-1992) has associated with his name a limit to the local mass density, rotation constants for the Milky Way, and a cloud of potential comets in the outer reaches of the galaxy. Contemporary Erik Holmberg (1908-2000) was perhaps asked whether he was sorry that there was only the Holmberg radius. "No, no!" he supposedly replied, "there is also the Holmberg diameter!" In fact there is also a Holmberg effect,

pertaining to the location of small satellite galaxies around big ones like our Milky Way; later data have shown an opposite correlation, not generally called the anti-Holmberg effect.

3. What if your name is taken?

Long ago, when we were graduate students, Jim (James Edward) Gunn worried that there was no use his discovering anything important, because there was already a Gunn effect (24,500,000) in solid state (now condensed matter) physics. A year or two later, he did, but luckily fellow grad student Bruce Peterson was a willing collaborator on the paper (Gunn & Peterson 1965) on what is now called the Gunn-Peterson effect (339,000). It is a limit to the density of hydrogen in intergalactic space, deduced from the lack of absorption of light from distant quasars by intervening hydrogen gas. This limit made it very difficult to close the universe with ordinary matter (and indeed it is not so closed).

That the same effect was published somewhat earlier by George Field and Peter A. G. Scheuer (separately) just goes to show. There are Field effect transistors (not an eponym), and what might have been named Scheuer's method is called P(D) and has had to be rediscovered and misunderstood many times in different branches of astronomy. The method uses Poisson statistics (39,500,000).

Our honoree is not to be confused with Gerald Schubert of the Go1dreich-Schubert instability (4490) or Go1dreich-Schubert-Fricke instability (3850). In fact he is surely one of the least unstable editors around!

My late husband worked under the burden of a unit used in magnetic field measurements being called the Weber (though he sometimes noted, correctly, that a Weber was worth 10,000 Gauss, another such unit). But his work on detectors for gravitational radiation was sufficiently high-profile that Weber bars are generally recognized in the field. Google brings you mostly other sorts of Weber bars, but gravity wave detectors score 341,000 and gravity wave detectors Weber bars 49,000.

As for Trimble entities, a science fiction novel by Robert L. Forward (who was Weber's advisee) makes use of Trimble temblors, which no one can pronounce five times quickly thereby limiting popularity. That they were supposed to happen on neutron stars probably didn't help, and Google finds zero, not even retrieving the original book.

I can't help but feel that this is somehow a higher distinction than the "Google-whackbit", where there is exactly one reference found. Aldebarium (a non-existence element named for the star Aldebaran) fell briefly in this category. Aldebaran means "the follower," and so is non-eponymic, but the Pleiades collectively and their seven individual names were daughters of Aeolus, the mythical god of the winds.

4. The three witches' rule

Human memory obviously has limits. I can remember two things (for instance to pick up at the grocery store on the way home); for three I have to make a list. This is not

unique. Certain primitive cultures were reputed to count "one, two, many". And it really is true that astronomers describe the chemical compositions of stars and galaxies as percentages of hydrogen, helium, and metals ([1]). A hypothetically-similar chemist would, of course, speak of the series "methane, ethane, paraffin."

But three is commoner. Many journals include in their format instructions that papers with many authors shall be cited as Szczepanowska, Martinez-Garcia, Sackmann -Christy et al. The Nebuchadnezzarian victims were three by name, Shadrach, Meshach, and Abednigo (Daniel, Chapter 3), while the uncounted wise men from the east (Matthew 2, 1-12) with their gold, frankincense, and myrrh early transmogrified into Three Kings named Balthazar, Melchior, and Caspar. And I would be remiss not to point out that the Greek names for the gifts are chryson, libanon, smyrnan (suggesting connections with other people and places).

Three also seems to be easily enough remembered that most physicists (well, anyhow most physicists interested in general relativity) can tell you the difference between Einstein-Infeld-Hoffmann (dealing with the geometry and evolution of, appropriately, many-body systems) and Einstein-Podolsky-Rosen (related to what Einstein called "spooky action at a distance"). Recent published examples includes Child-Pugh-Turcotte class B (NEJM 373, 216, a class of decompensated cirrhosis) and Hong-Ou-Mandel interference (Nature 527, 74), both in contexts without further explanation, indicating that the relevant readers are supposed to know what is meant.

Most practitioners of mathematical sciences will recognize the WKB (Wentzel-Kramers-Brillouin) approximation, arising from a quantum mechanical context. Only careful Brits will go for WKB-J, meaning that Sir Harold Jeffreys did it first, in a classical context. Curiously, the Google-count is 132,000 for WKB-J approximation and only 81,500 for WKB approximation. Kramers has also an opacity and Brillouin had zones.

Comets can carry the names of at most three discoverers. Merton (1993) dropped back to the kindle cole principle when the eponym grew to Vives-Hooke-Newton-Merton. The underlying idea is that attaching people's names (at least living people's names) to thing merely adds fuel to the fire.

Most bitterly this has been true for Nobel Prizes (unless you are a committee, like the Red Cross, which won Peace in both 1917 and 1944, well, the wars did end the next year). Now, pull out an almanac that lists Nobels with what they were given for and pick out your favorite trios that should have been quartets. A common choice is 1962 medicine or physiology (Crick, Watson, and Wilkins, but Rosalind Franklin had died some years before). Another is physics 1965 (Tomonaga, Schwinger, and Feynman, where Freeman Dyson perhaps also belonged). While the almanac is open to Nobels, take a look at economics. George Stigler, 1982, is the father of the "law of eponyms" Stigler (1980) and Robert Merton, 1997, is the son of OTSOG Merton (1993). Oh, and in the references to this paper, Trimble (2014) is, of course, the daughter of Trimble (1945).

---

[1] This is not quite as odd as it sounds. Typical numbers are (by weight) 74% hydrogen, 24% helium, and 2% everything else. Admittedly more than half of that "everything else" is CNO, but early stellar spectroscopy, with blue-sensitive plates, readily showed lines of Fe, Ca, Na, Ti, etc. which really are metals.

The section title looks back (or anyhow sideways) at Macbeth, who does not attempt to remember the names of the "dark and midnight hags" who placed one-two-three in the Edinburgh Ugly Contest. But how many of the names of the attendant fairies in Midsummer Night's Dream can you remember? Um. Peaseblossom… ? As for the witches' shopping, I'm sure they made a list. Eye of newt …

5. The boring landscape

The Boring figure (71,200,000) was called "my wife and my mother in law" by the cartoonist who drew it and as "young girl old woman drawing" scores 8,680,000 Google mentions. It is an extreme example of a two-dimensional figure that can be seen as two different images, either switching spontaneously from one to the other as you stare at it or reversible by focussing, for instance, on the girl's eye or the old woman's chin. Senior colleagues who took introductory psychology courses in the 1950s and 60s still often remember being asked to "take out your Boring textbook."

Edwin Garrigues Boring (1886-1968) was a largely-Harvard-based psychologist and historian of psychology. I met him as "footnote 1" in Rutherford (2015), which cites his 1963 talk (Boring 1964 as published) on "Eponym as Placebo," just as I was finishing the collection of data for this commentary. She is writing on "Maintaining masculinity in mid-twentieth-century American psychology: Edwin Boring, scientific eminence, and the "woman problem", and quotes a passage from the "eponym paper" for its description of an ideal scientist, making him (definitely) sounds like an obsessive-compulsive Asperger's case. Boring also wrote on the moon illusion, perception of color and sound, women in psychology, and delays in publication. His views on women were clearly conflicted; he married a student and had at least one paper with her, but felt that women were too interested in the particular, the suffering, applications, the young, and treatment to be suitable for theoretical psychology. His statements about Jews in psychology were very similar. None of this is atypical for his time, background, and host institutions.

Now about the eponyms. He made three points (1) memory is short and so indeed "every great event in science (must) have an owner" and "every fundamental paradigm in science (will) create an eponym" but that "the history of science ought not to be hung on enormous eponymous pegs," (2) everybody wants to belong to a winning team and so we are prone to celebrate our captains, (3) scientists all desire to assume leadership, and so encourage attitude (2). In favor of having scientific heroes he claims they will inspire and stimulate neophytes, which is good. Conversely unique crediting will add fuel to fires, distort history, and keep us from recognizing how science is really done. He credits similar views to Derek de Solla Price (whom all readers of Scientometrics must know) and Gerald Holton, and mentions Newton vs. Leibnitz and Adams vs. Le Verrier on choices of credit-receivers.

A couple of other Boring thoughts with which I agree: (1) authors should have to pay page charges to keep them from rambling on too long (he suggests $8/page, but this was a long time ago and might well translate up to the current Astrophysical Journal

$120/page or so) and (2) "intelligence is what the tests test", coming from his experience with military intelligence tests during his work in World War I. And one view with which I disagree (even though he credits it to R.K. Merton and Gerald Holton), that science is becoming less competitive, so that we might look forward to a future Utopia, in which histories of science discuss nameless advances and no more eponyms are coined.

The now-constant quests for funding, jobs, promotions, recognition, prizes, and all clearly says the opposite. It is, however, perhaps worth noting that, while Nobel sticks at three, some other fairly prestigious prizes, like the Gruber Cosmology Prize, now sometimes go to leaders plus their teams. Of course all the really good prizes—Nobel, Gruber, Shaw, Kavli, Ambartsumian, Crafoord, and even the lesser ones (like the Weber Award in astronomical instrumentation of the American Astronomical Society and about half of the awards for 2016 from the American Chemical Society listed in C&EN, Chemical and Engineering News for 2016 January 4)—carry someone's name, which you may or may not recognize as a donor or honoree (the latter a sort of eponym).

6. Untied threads

Sometimes the person being eponymized has no choice. A Vandyke was an intermediate stage between a draftsman's original drawing and a blue print in the aircraft industry roughly four centuries after the painter (1499–1541) died (Trimble, 1945). The connecting concept is dark brown pigment, still sometimes known as Vandyke brown. It must have been important at the time because casual civilian travel was not encouraged, but father presented his work at a technical conference in New York on 17 October 1944. He was working for Lockheed Aircraft Corporation in Burbank California at the time.

Clark refractors are not eponymous. They were made by Alvan Clark and his sons, Alvan G, and George B. Clark in 19th century Massachusets. Compare, for instance, Strauss waltzes and Sousa marches. Galileian refractors, Newtonian reflectors, and Maksutov telescopes carry names of original designers, and half-breeds like Mak-Cass and Mak-Newt are frequently made by Meade. These are clearly eponymous.

Is the plague increasing? Probably. In addition to the cases of the chemical elements, Daltons, degrees Celcius, and Leavitt's law noted above, something has happened to our space missions. The first 1500 or more rose (or sometimes failed to) carrying names like Sputnik, Vanguard, Explorer, Luna, Apollo, Ariel, and Cosmos (Seaborn 1968). Then one fine day in 1972, the US and UK launched an ultraviolet and X-ray observing satellite and named it Copernicus (he was not at the launch). Pretty soon we had Einstein (X-ray), ROSAT (for Roentgen), BeppoSAX, HIPPARCOS (Hipparchus), Compton GRO, Spitzer and Herschel Space Telescopes (infrared), Chandra and XMM-Newton (both X-ray observatories and neither honoree very closely connected to X-rays), RXTE (Rossi X-ray Timing Explorer), the very well-known Hubble Space Telescope ("Hubble finally got the telescope he deserved" said a senior colleague when the mirror flaw was discovered), the James Webb Space Telescope (finally on track for 2018 October?), and the progressing of Cosmic Background Explorers from COBE to Wilkinson Microwave

Anisotropy Project, to Planck. These are all some combination of US and European missions. A comparable Russian one was MIR/KVANT. And Japanese satellites all still carry names that mean "flapping bird" and similar designation. In addition, they are named only after successful launches, so that Japan has never fully lost a named mission, just as the University California Irvine football team has never lost a game since 1965.

A colleague just drifted by and said that at least Eponyms were clear and/or less ambiguous than other descriptions, for instance, Newton's laws, rather than laws of motion. This is not always the case. Penning traps actually pen particles in magnetic fields, but if you ask a speaker who mentions free energy whether he means the Gibbs free energy or the Helmholtz free energy, he generally has to contemplate for a while. A short answer is that the Gibbs free energy or potential is the one that is useful for chemical reactions.

A few more favorite items include the Kaliapparat (15,500) vs Liebig tube (298,000) Maillard reactions (455,000) vs non-enzymatic browning (149,000), Bessemer process (343,000) vs pneumatic conversion process (957,000).

And finally one more case where "Garry" Boring got there first. That some people have more prizes, honorary degrees, and things named for them than others undoubtedly reflects relative merits, but it is not the whole story. I thought I had invented the concept that "honors are bosons" (meaning particles that like to huddle together, like photons, and opposed to Fermions, which like to space themselves out). But, said, Boring, "professional prestige is autocatalytic," which is, as nearly as possible, the same idea from the point of view of a chemist rather than an astrophysicist. It is probably also significant that he picked "prestige," which is a collective noun, while I picked elementary particles, which can be counted.

Acknowledgements

I am grateful to Wolfgang Glänzel for the invitation to write these comments, to András Schubert for providing the occasion, and to Gerald Holton for offering to consider reading these pages. He makes a cameo appearance in Sect. 5.

References

Boring, E.G. (1964). Eponym as placebo. *Acta Psychologica,* 23, 9-23.

Gunn, J.E., Peterson, B.A. (1965). On the density of neutral hydrogen in intergalactic space. *Astrophysical Journal,* 142, 1633-1641.

Livingston, A. (2015). The man who put the names on the moon. *Sky and Telescope*, May, p. 27.

Lundmark, K. (1924). The determination of the curvature of space-time in de Sitter's world. *Monthly Notices Royal Astronomical Society,* 84, 747-770.

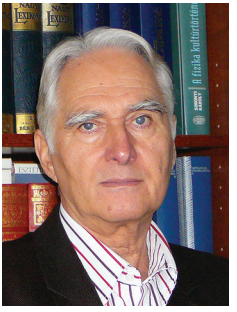
Lundmark, K. (1925). The motions and the distances of spiral Nebulae. *Monthly Notices Royal Astronomical Society*, 85, 865.

Merton, R.K. (1993). *On the Shoulders of Giants*, University Chicago Press.

Nussbaumer, H., Bieri, L. (2009). *Discovering the Expanding Universe.* Cambridge University Press.
Nussbaumer, H., Bieri, L. (2011). Who discovered the expansion of the Universe? *Observatory Magazine,* 131, 394-398.
Rutherford, A. (2015). Maintaining masculinity in mid-twentieth-century American psychology: Edwin Boring, scientific eminence, and the "woman problem". *Osiris,* 30, 250-271.
Seaborn, H.T. Ed. (1968). TRW Space Log 7, No.4
Seitter, W.C., Duerbeck, H.W. (1990). *Carl Wilhelm Wirtz—an early observational cosmologist*. In: Cosmology in Retrospect. Eds. B. Bertotti et al., Cambridge University Press. p. 365-399
Stevens, S.S. (1973). *Edwin Garrigues Boring, 1886-1968.* US National Academy of Sciences Biographical Memoir, vol. 40
Stigler, S. (1980). *Stigler's law of eponymy.* Transactions of the New York Academy of Sciences, 39, 147-158.
Trimble, L.S. (1945), A New Medium for the Production of Vandykes. *Journal of the Society of Motion Picture Engineers*, 45(1), 54-64.
Trimble, V. (2014). Anybody but Hubble! *Asian Journal of Physics,* 23 (1-2), 91-100.

Error Calculations, András Schubert, and the Wheat Beer

PETER VINKLER
Research Centre for Natural Sciences,
Hungarian Academy of Sciences, Budapest, Hungary



There are many aspects of both the professional and private life of any scientist. Man would like to be happy in life but creative people also want to exert influence on the world. After a long career people use to draw up a balance, whether their impact would be large enough. To do that for a scientometrician, it seems to be very simple. We have (too) many indicators for characterizing scientific impact both quantitatively and qualitatively. The poor outsiders could have only faint ideas where, what and how to apply them. I try to present some indices here which may represent the high standard of the scientific results of András Schubert, properly.

Selected scientometric indicators were calculated from the data of publications of András (Table 1). He published 129 articles on scientometrics referenced in WoS in 1975-2015. His production seems to be outstanding not only quantitatively but also as far as the scientific impact of his publications is concerned. This may be concluded from the high value of both the *h*, *g*, and π-index (Table 1).

The publication performance of András was related to a standard. The standard was calculated as the mean of 10 Price medalists (P.M.) who were member of the editorial board of both Scientometrics and Journal of Informetrics in 2014. (For details, see P. Vinkler: Core indicators and professional recognition of scientometricians, Journal of the Association for Information Science and Technology, DOI: 10.1002/asi.23589).

In my view, the mean indices of the total impact derived from the total set of publications of individuals, teams or countries do not characterize the real impact of published information properly. The real impact of the results on the scientific community and science may be characterized by the *most influential part* of the publications. Accordingly, the task of scientometricians is to select the *core* (or *elite*) *set* publications within a total set. This may be done by selecting *h*,

g, or π-core of publications. The two former mentioned indices are well known. The number of core publications is definied as the measure of the *h*-index or *g*-index. The π-core = $\sqrt{P}$, i.e. the square root of total papers. The π-index can be obtained as one hundredth of the number of citations received by π-core papers. The papers should be ranked in the decreasing citation frequency.

The ratio of the indices in Table 1 reveals that the publication performance of András is significantly higher than the selected standard.

Table 1. Some scientometric indicators of András Schubert (A. S.) in 1975-2015 (I) and in 1975-2014 (II) compared to the mean of selected 10 Price medallists (P.M.).

	A. S./I	A. S./II	P. M./II mean	Ratio
P	129	125	87.00	1.44
C	3030	2856	1722.30	1.66
h	30	27	21.30	1.27
g	51	50	39.20	1.28
π	14.49	13.58	8.59	1.58
πr	131.73	123.45	95.44	1.29

P/I: Total number of publications (P) in 1975-2015 in WoS filtered as Information and Library Science. P/II: P in 1975-2014. C: Total number of citations to P publications. h: h-index. g: g-index. π: π-index. πr: π-rate, i.e. mean citation frequency of publications in the π-core.

The most cited paper of András (together with T. Braun and W. Glanzel) is: "Hirsch-type index for journals" (Scientometrics, 69, 169-173, 2006) with 225 citations. The second (with T. Braun) is: "Relative indicators and relational charts for comparative assessment of publication output and citation impact" (Scientometrics, 9, 281-291, 1986). It has been cited 224 times (!) until now. I appreciate highly this paper as one of the first papers on relative scientometric indices. In my view, the introduction of relative indices would be at least as, or even more significant in scientometrics than the introduction of the *h*-index. But, many of András' papers have made also significant impact. Some topics of his articles: statistical reliability and normalization of scientometric indices, scientometric distributions, collaboration network of individuals and countries, *h*-index studies, etc.

András graduated as a chemical engineer, and started his scientific career at the Gödöllő University for Agriculture. He was interested in theoretical chemistry, and in 1976 published a book entitled: Kinetics of Homogeneous Reactions (Műszaki Könyvkiadó). But later his interest turned towards scientific information and scientometrics. He joined the Library of the Hungarian Academy of Sciences, and became a member of the Information Science and Scientometric Research Unit, ISSR established by Tibor Braun.

András published his first scientometric paper referenced by WoS in 1981 together with T. Braun (Some scientometric measures publishing performance for 85 Hungarian research institutes, Scientometrics, 3, 379-388). Consequently, his Publication Life Time: PLT = 2016–1981 = 35 year.

I published my first scientometric paper in English only in 1986 entitled "Evaluation of some methods for the relative assessment of scientific publications" (Scientometrics, 10, 157-177). Accordingly, my PLT = 2016–1986 = 30 year. (Although I am just 5 years older than András.)

In 1991 (my PLT was that time: 1991–1986 = 5 year) took place the European Workshop on Scientometric Methods of Research Evaluation in the Sciences, Social Sciences and Technology in Potsdam, Germany. I delivered a lecture on research contribution, authorship and team cooperativeness. I presented the results of a questionnaire study concerning the contribution share of co-authors of publications with 2, 3, 4, 5 and 6 co-authors. After the lecture a scientist from the audience put a question concerning the *reliability* of the data. I answered that the data would be reliable, because I made the interviews with the authors personally.

'I understand and respect that, but have you made any error calculations?'—I was asked immediately.

First I did not understand the question. Why should I do "error calculations"? I answered that I made no error calculations, because I thought the data were reliable.

After the session András (his PLT was that time: 1991–1981 = 10 year) invited me to drink a Berlin white (Berlin Weisse, wheat beer). He realized that I was desperate. I was ashamed. András consoled me:

'Well, the set you analyzed seemed to be rather homogeneous. I think, your conclusions would be reliable. Nevertheless, you need to prove that the mean of the contribution of first authors really differs from that of the second authors. And the share you found for second authors differs from that of the other authors etc. Well, I can recommend you some books dealing with statistics and error calculations.'

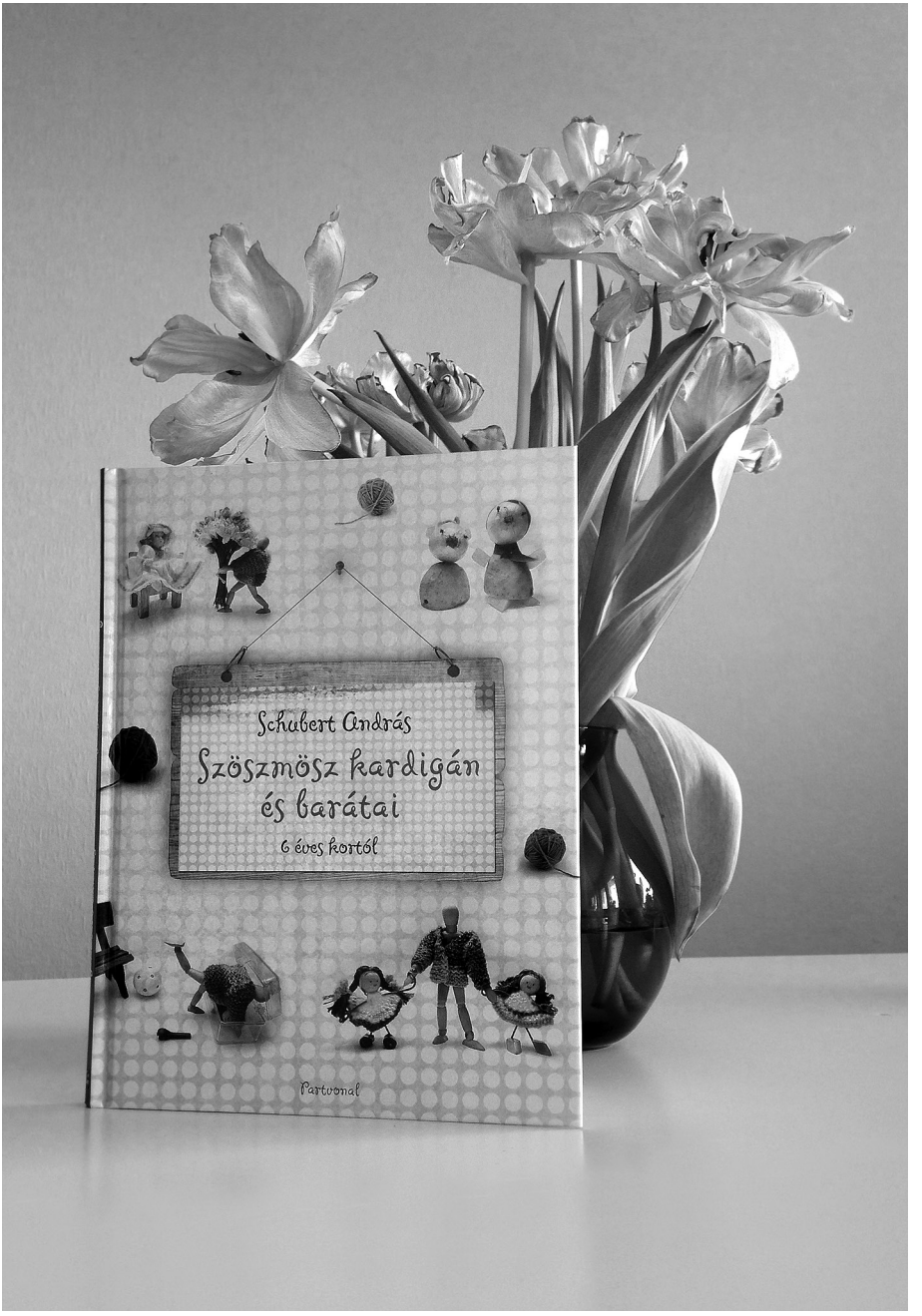
I was very grateful to András for his friendly suggestions—and making me acquainted with the excellent wheat beer.

András is really a renaissance man. In 2006 the whole scientometric community could enjoy his clarinet play at the evening party in Leuven. Together with Balázs Schlemmer (piano) they played excellent jazz and pop-music. Moreover, András plays clarinet in a band in a pub regularly.

I am always highly impressed realising that somebody who is excellent in his profession can offer also other activities at high standard. In 2010 András published a book entitled "Fluffy cardigan and his friends", which is a fairy tale for children and adults with child-like soul.

I think, the children are sincere and they have fantasy. As time goes, we lose both, except for András.

I wish him good luck, health, further successes in scientometrics, and enjoyment in music and book writing.



András Schubert's first child book "Fluffy cardigan and his friends" (cf. contributions of Wolfgang Glänzel et al., Guillaume Cabanac, Gábor Schubert & Mihály Schubert, Péter Vinkler). Photo © Balázs Schlemmer