Science as a Punctuated Equilibrium

Haiko Lietz | netsci2012@haikolietz.de | Institute for Research Information and Quality Assurance, Competence Centre for Bibliometrics, Berlin, Germany | University Duisburg-Essen, Institute of Sociology, Duisburg, Germany | Mittweida University, Faculty of Mathematics, Physics, and Computer Science, Mittweida, Germany

Introduction

Self-similarity signals the **emergence of identity** which re-enacts itself on different length-scales [1-2]. There is tremendous evidence for the self-similarity of the science system [3-6], both in terms of its social (collaboration) [7-8] and cultural (citation) [9-10] structure. Classical models [11-12], recently re-formulated and proven in network language [13-14], explain the emergence of these scale-invariant formations through Preferential Attachment, the **Matthew Effect**, *i.e.*, scientists or ideas gain in prestige or centrality the more prestigious or central they already are.

Other evidence for heavy tails in science has been discussed in the framework of **Self-Organized Criticality** [15-17]. It has been proposed [18] that **Kuhn's phase model of scientific advance** [19] resembles a **Punctuated Equilibrium** [18], a self-organized critical process used to explain biological macroevolution [20]. According to Kuhn's model, periods of normal science, during which paradigms ensure a continuous and cumulative production of knowledge, are interrupted by scientific revolutions, during which an old canon of knowledge is replaced by a new one.

Natural punctualism presents itself, among other ways, in the scale-invariance of genera lifetimes (Fig. 4) and bursts of extinction events (Fig. 5) [21]. We model scientific advance using citations as concept symbols [22] and, inspired by natural punctualism, test **whether knowledge macroevolution is bursty**, using three hypotheses.

H1: A few publications are highly cited for a long time (Power-law distribution of citation lifetimes).

H2: A few extinct publications are highly cited (Power-law distribution of extinct publication citations).

H3: A highly cited publication is more likely followed by a highly cited publication than by a low cited publication and *vice versa* (Power-law correlation of maximum citation variations).

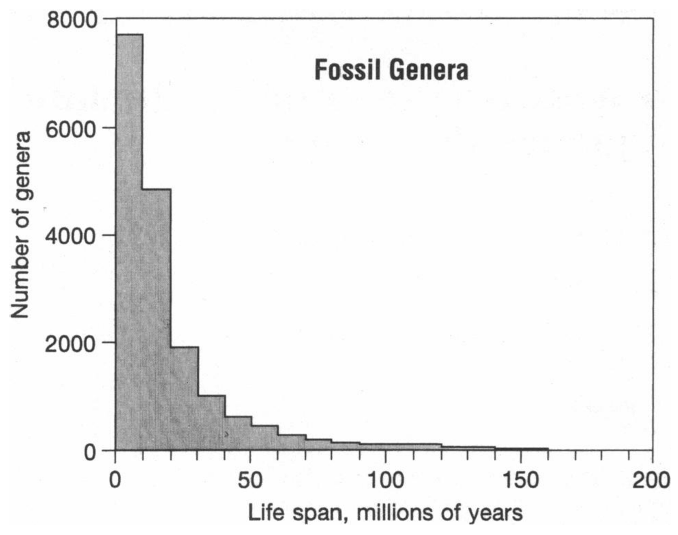

Fig. 4: Power-law distribution ($\gamma \approx 2$) of genera lifetimes. The Image is reproduced from [21] in fair use.

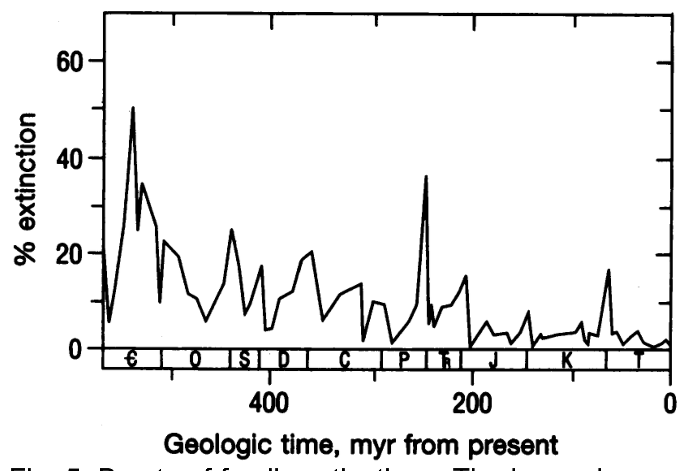

Fig. 5: Bursts of family extinctions. The image is reproduced from [21] in fair use.

Data & Methods

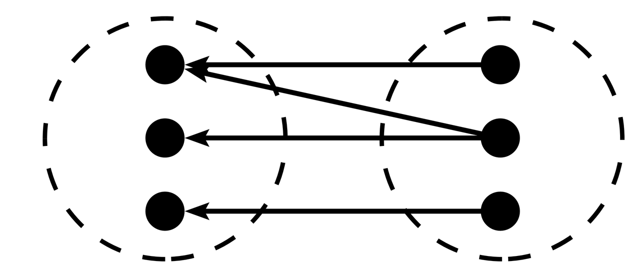

Fig. 6: Cited publications are the units of analysis.

Data is from the *Web of Science*® (WoS). The all science system, its 6 sciences and 32 disciplines, created by aggregating 260 available subjects, are the **objects of study**. Publications from <1980, cited by publications from 1980-2007 and indexed by the WoS, are the **units of analysis** (Fig. 6). Unique cited publications are identified by transforming each reference in a citing publication into a string consisting of the first 4 initials of the first author's name, the publication year, the first initial of the source title, and the first page [23].

Citation and citation lifetime distributions are based on absolute citations. The **citation lifetime** is defined as the number of years that a publication belonged to the top 100%, 10%, 1%, 0.1%, or 0.01% of the most highly cited publications in the respective citing year. For disciplines only the lifetimes of highly cited publications (belonging to the Top 1%) are computed. **Extinct publications** have accumulated at least 90% of their absolute citations in 1980-1993. **Scaling analysis** is the interpretation of exponents γ in the fit function $P(x) \sim x^{-\gamma}$. Exponents are fitted using logarithmic binning and least squares.

Long-term correlation analyses are based on **mean normalized citation rates**, *i.e.*, for each publication and citing year 1980-2007 the absolute citation rate of a publication is divided by its corresponding expected citation rate and finally averaged over all 28 citing years. To account for the growth of science, expected citation rates are computed for exponentially growing normalization windows. Citation variations are studied using **peak-over-threshold analysis** [24]. Maximum normalized citations (Fig. 7) are first detrended by taking the common logarithm and subsequently subtracting a moving, 13 year average (Fig. 8). Values ≤ 0 are removed and the remaining values normalized and discretized (Fig. 9). For each threshold q = 0.2, 0.4, 0.6, 0.8, 1 return intervals r and the average return intervals R_q are computed. If variations are correlated, the distributions $P_q(r)$ are congruent stretched exponentials where the stretching exponent γ is the correlation exponent between 0 and 1 [24].

Results

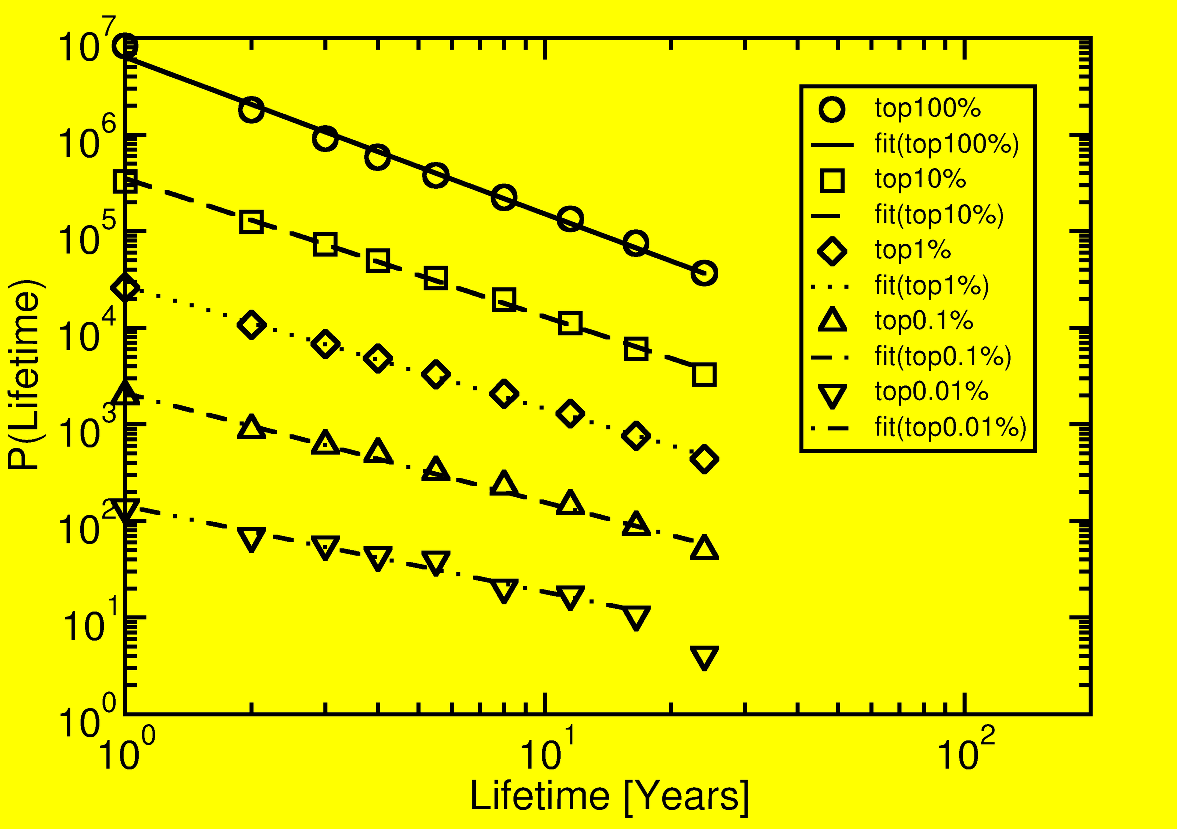

Fig. 1: Citation lifetime distribution of the system of all sciences. Many publications are only cited for a few years while few publications are cited for many years. Scaling exponents range from 1.62 for all publications (Top 100%) over 1.27 for highly cited publications (Top 1%) to 1.01 for extremely highly cited publications (Top 0.01%). Distributions for disciplines are scale-free but not universal (Tab. 1). H1 doesn't need to be rejected.

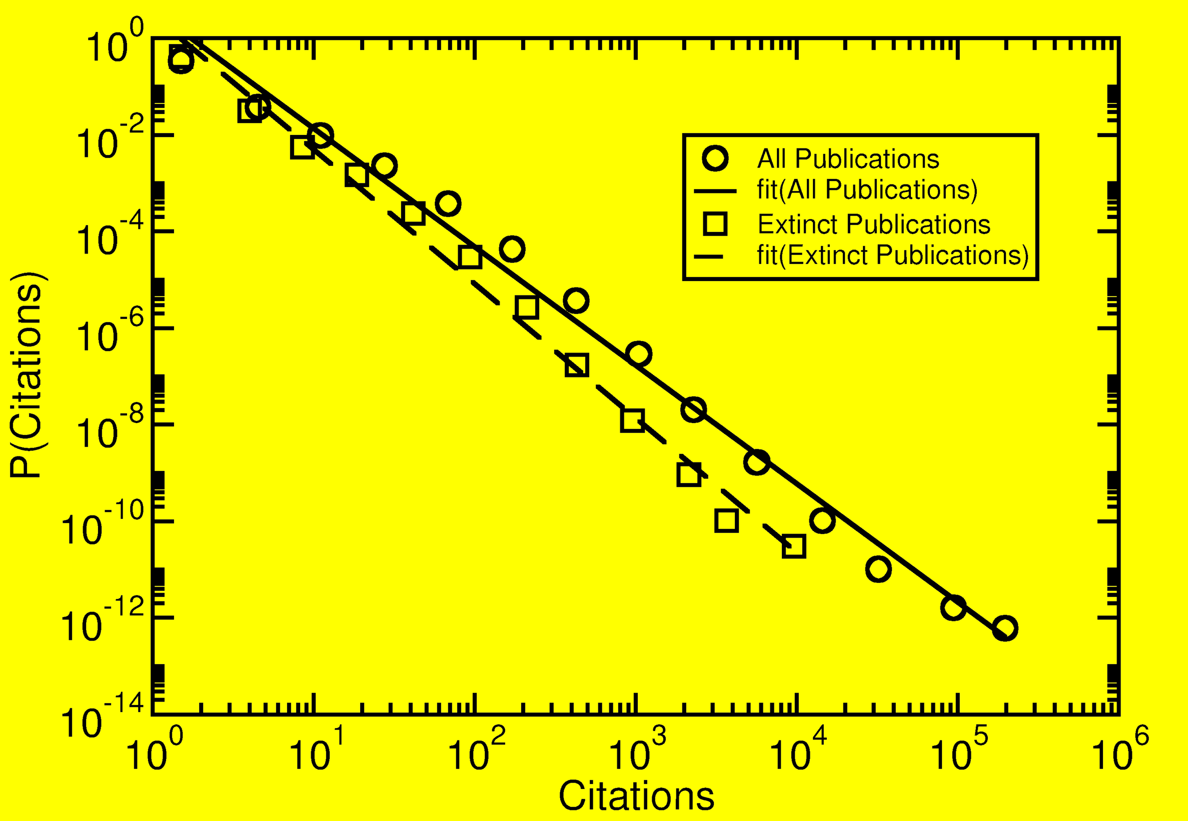

Fig. 2: Citation distributions of the system of all sciences. The distribution for all publications is scale-free, as expected [10, 14]. But also extinct publications (hardly getting cited since 1994) are scale-free. There are large extinction events. Regarding disciplines, exponents only exceed 4 in the humanities (Tab. 1). H2 doesn't need to be rejected.

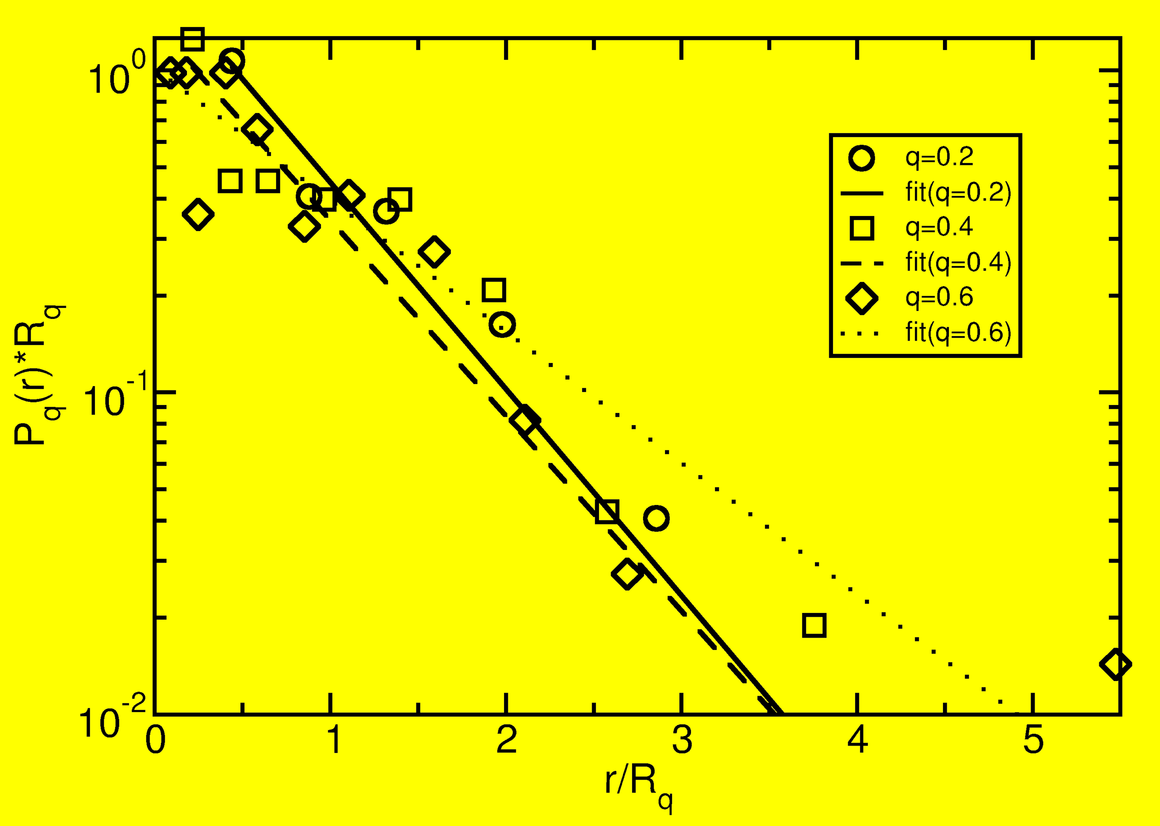

Fig. 3: Normalized distributions of return intervals of maximum citations for different thresholds for the system of all sciences. The distributions, also for disciplines, decay exponentially. Therefore, there are no long-term correlations of maximum citation variations. H3 must be rejected.

Discussion

It is well known that distributions of absolute citations that publications have accumulated in a certain period of time are scale-free [10, 14]. This was confirmed and also shown to be true for extinct publications which have accumulated 90% of all their citations in the first half of the citing window 1980-2007. In other words, there are large extinction events (Fig. 1). There are differences between the sciences. Exponents in the humanities are much bigger than in human science (Tab. 1). Hence, change is not as pronounced in the humanities. In the following, **change** is defined as the inverse of the extinction exponent.

Studying lifetimes it was shown that a few publications are highly cited for a long time, in the all science system as well as in scientific disciplines (Fig. 2). Exponents (Tab. 1) can directly be interpreted as canonic **plurality**. Plotting plurality against change (Fig. 10) proofs a central message explicit in Kuhn's theory: Change in the soft sciences can never be big because their disciplines have not organized around canonical knowledge. **Only in the hard monolithic sciences change can amount to scientific revolutions.**

Highly cited publications are not long-term correlated (Fig. 3). The most likely reason is that the data is inadequate. After all, it's still a look back from 28 recent years. Historical publications following a breakthrough publication will likely remain in the shadow of the latter. Correlations may also be hidden by noise due to high levels of aggregation. But studying subjects instead of disciplines is not possible for reasons of statistical validity. The string used to identify unique cited publications, which requires first pages to match, is also expected to have an influence on the results. In general, the analyses are hampered by what Merton called **Obliteration by Incorporation** [25]: Many important concepts are „undercited“ because they have become too institutionalized to still be cited [26].

Summing up, even though long-term correlations were not found, evidence for the burstiness of knowledge macroevolution was found in power-law distributions of citation lifetimes and citations to extinct publication. The **ubiquity of self-similarity** indicates that science is an ever-dynamic system with births and deaths occurring on all spacial and temporal lenght scales. There are differences between different sciences and disciplines. Exponents smaller than those generated by pure Preferential Attachment processes call for models combining Preferential Attachment and Self-Organized Criticality.

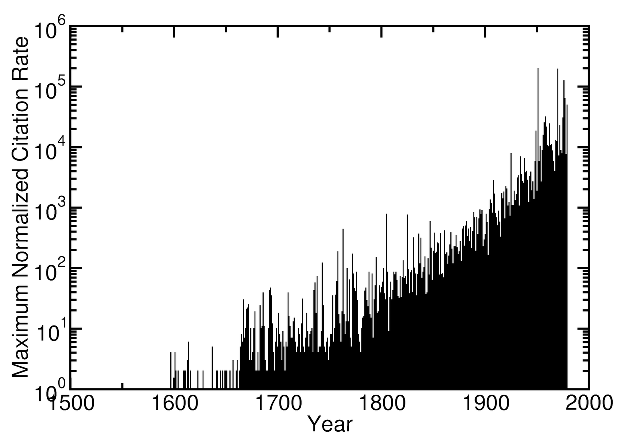

Fig. 7: Super-exponential growth of maximum normalized citation rates in the system of all sciences.

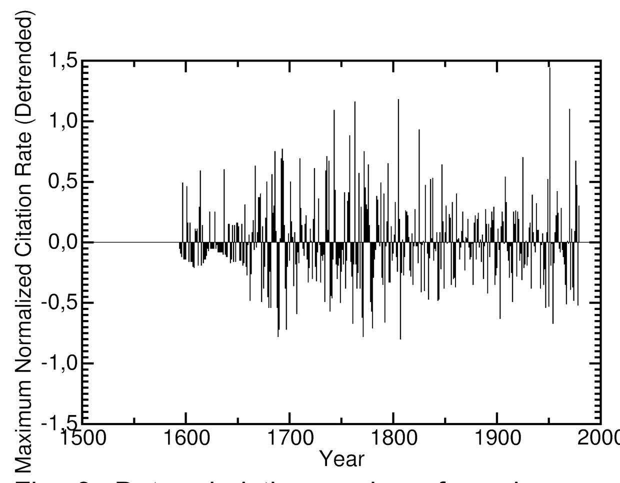

Fig. 8: Detrended time series of maximum normalized citation rates for the system of all sciences.

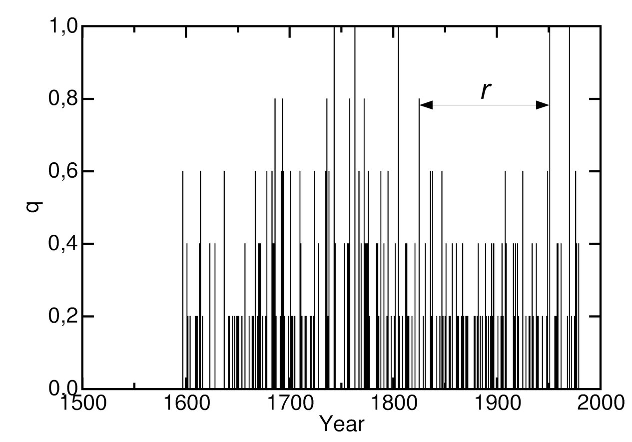

Fig. 9: Time series of maximum normalized citation rates after detrending, normalization, and discretization for the system of all sciences. A return interval r for the threshold q = 0.8 is shown.

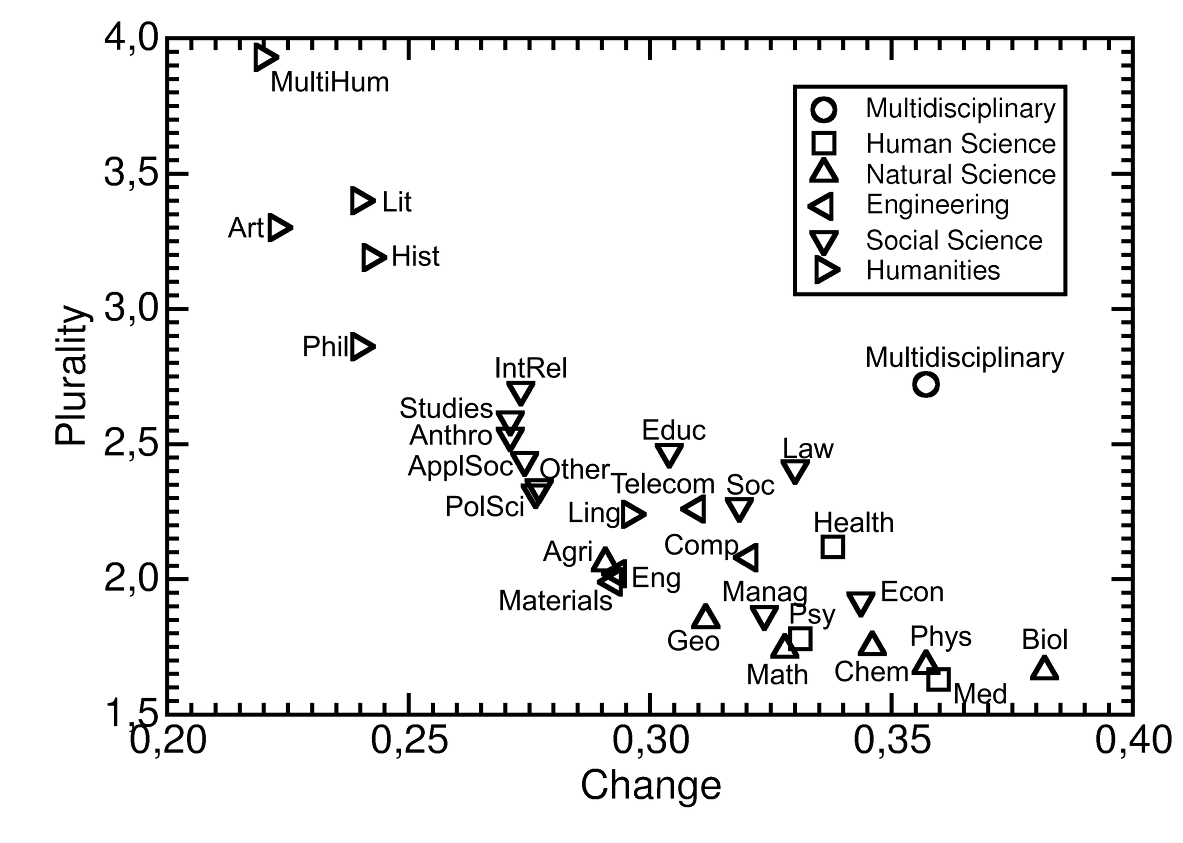

Fig. 10: Scientific disciplines along the dimensions change (inverse exponent of citations to extinct publications) and plurality (citation lifetime exponent). Only for disciplines in the lower right quadrant change can amount to scientific revolutions due to these disciplines' organization around a canon of knowledge.

Science	Discipline	Publications All	Publications Extinct	Lifetimes
Multidisciplinary		2.61	2.80	2.72
Human Science	Applied Health	2.52	2.96*	2.12
	Medicine	2.46	2.78*	1.63
	Psychology	2.30*	3.02*	1.78
Natural Science	Agriculture	2.50	3.44*	2.06
	Biology	2.35	2.62	1.66
	Chemistry	2.46	2.89	1.75
	Geoscience	2.45*	3.21*	1.85
	Mathematics	2.49*	3.05*	1.74
	Physics	2.26*	2.80*	1.68
Engineering	Computer Science	2.32	3.12	2.08
	Engineering	2.59*	3.41*	2.02
	Materials Science	2.47	3.42	1.99
	Telecommunications	2.48	3.23	2.26
Social Science	Anthropology/Archaeo.	2.62	3.69	2.53
	Applied Social Science	2.53	3.65*	2.44
	Economics	2.17	2.91	1.92
	Education	2.73	3.29	2.47
	International Relations	2.76	3.66	2.70
	Law	2.55	3.03	2.41
	Management	2.31	3.09*	1.87
	Other Social Sciences	2.61	3.61	2.34
	Political Science	2.49*	3.62	2.32
	Sociology	2.55	3.14*	2.27
	Studies	2.78	3.69	2.59
Humanities	Art	3.22	4.49	3.30
	Ethics	2.68		2.93
	History	3.34	4.13	3.19
	Humanities, Multidisc.	3.53	4.55	3.93
	Linguistics	2.63	3.38	2.24
	Literature	3.37	4.17	3.40
	Philosophy/Religion	2.80	4.17	2.86

Tab. 1: Exponents for the citation distribution of all publications, of extinct publications, and for the citation lifetime distribution. A * marks fits to logarithmically binned data where the coefficient of determination is between 0.96 and 0.98. Otherwise it is better.

Literature

[1] HC White, *Identity and Control* (2008) | [2] S Fuchs, *Against Essentialism* (2001) | [3] DdS Price, *Little Science, Big Science* (1963) | [4] JS Katz, *Research Policy* **28** (1999) 501-517 | [5] J Leskovec *et al.*, *Proc ACM SIGKDD* (2005) 177-187 | [6] LMA Bettencourt *et al.*, *Scientometrics* **75** (2008) 495-518 | [7] AJ Lotka, *J Wash Acad Sci* **16** (1926) 317-323 | [8] MEJ Newman, *Lect Notes Phys* **650** (2004) 337-370 | [9] SC Bradford, *Engineering* **137** (1934) 85-86 | [10] PO Seglen, *J Am Soc Inf Sci* **43** (1992) 628-638 | [11] RK Merton, *Science* **159** (1968) 56-63 | [12] DdS Price, *J Am Soc Inf Sci* **27** (1976) 292-306 | [13] A-L Barabási & R Albert, *Science* **286** (1999) 509-512 | [14] H Yeong *et al.*, *Europhys Lett* **61** (2003) 567-572 | [15] AFJv Raan, *Scientometrics* **47** (2000) 347-362 | [16] GG Brunk, *Scientometrics* **56** (2003) 61-80 | [17] A Mazloumian *et al.*, *PLoS ONE* **6** (2011) e18975 | [18] SJ Gould & N Eldridge, *Nature* **366** (1996) 223-227 | [19] TS Kuhn, *The Structure of Scientific Revolutions* (1996) | [20] P Bak, *How Nature Works* (1996) | [21] K Sneppen *et al.*, *Proc Natl Acad Sci* **92** (1995) 5209-5213 | [22] H Small, *Soc Stud Sci* **8** (1978) 327-340 | [23] T Braun *et al.*, *Scientometric Indicators* (1985) | [24] A Bunde *et al.*, *Phys Rev Lett* **94** (2005) 048701 | [25] RK Merton, *ISIS* **79** (1988) 606-623 | [26] KW McCain, *J Am Soc Inf Sci Tech* **62** (2011) 1412-1424

Acknowledgements

Thanks to Diego Rybski and Daniel Sirtes for technical help and fruitful discussions. Data is from the bibliometric database of the German Competence Centre for Bibliometrics and derived from the 1980 to 2007 *Science Citation Index Expanded* (SCIE), *Social Sciences Citation Index* (SSCI), *Arts & Humanities Citation Index* (AHCI), *ISI Proceedings – Science and Technology* (ISTP), and *ISI Proceedings – Social Sciences and Humanities* (ISSHP) prepared by Thomson Reuters (Scientific) Inc. (TR®), Philadelphia, Pennsylvania, USA: ©Copyright Thomson Reuters (Scientific) 2012.