Which Scientific Elite?
On the Concentration of Funding, Productivity and Citations

Vincent Larivière[1], Benoît Macaluso[1], Éric Archambault[2] and Yves Gingras[1]

[1] lariviere.vincent@uqam.ca; macaluso.benoit@uqam.ca; gingras.yves@uqam.ca
Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, Case Postale 8888, Succursale Centre-Ville, H3C 3P8 Montréal, Québec (Canada)

[2] eric.archambault@science-metrix.com
Science-Metrix, 1335A avenue du Mont-Royal Est, H2J 1Y6 Montréal, Québec (Canada)

Abstract
Using a very large population of university professors and researchers (N=13,479), this paper analyses the concentration of funding, papers and citations at the level of individual researchers. It shows that each of these distributions is different: citations being the most concentrated, followed by funding, papers published and finally number of funded projects. Concentration measures also vary between fields, social sciences and humanities being generally the most concentrated. The paper also shows that the correspondence between the elites defined by each of these measures is limited. In fact, only 3.2% of the researchers are in the top 10% on all indicators, while about 20% are in the top 10% in at least one of the indicators. The paper concludes with a discussion on the causes of these observed differences and formulates a few hypotheses.

Introduction

Distributions of researchers' productivity and citations have been studied fairly extensively by early information scientists. For instance, as early as 1917, Cole and Eales analysed the distribution of productivity of researchers in the field of comparative anatomy. A decade later, Lotka (1926) found that a minority of scientists were responsible for the majority of the scientific papers published. Similar patterns of concentration were also found for journals in which researchers published (Bradford, 1934) as well as for words used in scientific texts (Zipf, 1949). With the development of the sociology of science and the advent of the SCI, these distributions were analyzed in light of the inequalities of the reward system of science (Merton, 1973), the scientific elite (Zuckerman, 1977) as well as in the debate surrounding the "Ortega hypothesis" (see, among others, Cole & Cole, 1972; 1973; MacRoberts & MacRoberts, 1987). More recently, following Pratt (1977), bibliometricians such as Egghe (1988), Rousseau (1992) and Burrell (1991) worked on the mathematical aspects of concentration measures. More specifically, Rousseau (1992) showed, using Allison's (1980), data on U.S. scientists, that citations were more concentrated than publications. Along the same line, Ioannidis (2006), Evans (2008) and Larivière, Gingras & Archambault (2009) provided empirical measures of concentration at the paper and journal levels. However, no study has yet combined, for a large population of researchers, measures of concentration of publications, citations and research funding.

Using a very large dataset of publications and citations for the full population of professors and university-based researchers in Quebec (N=13,479), this paper presents concentration measures of research funding, publications and citations at the level of individual researchers. More specifically, this paper aims at answering the following research questions: 1) what are the fields in which research funding productivity and impact are the most concentrated 2) which distribution (funding, publications or citations) is the most concentrated and 3) do we find the same group of scientists in the top ranking for these measures of funding, and scientific output and impact.

Two indicators are used to measure the concentration of research funding, productivity, and citations. The first is the percentage of researchers who have published at least a paper or have received at least a citation over the 2000-2007 period. The second indicator is given by the cumulative Pareto distributions (Lorenz curve) of funding, publications and citations. It shows the percentage of researchers needed to account for any given percentage (20%, 50%, 80%, etc.) of the variable under study (funding, publications and citations). The lower the percentage of researchers needed to account for any percentage of a variable the more concentrated that variable is.

Methods

This paper draws on a very large dataset comprising funding, publication and citation data of each professor or university-based researcher in the Canadian province of Quebec over the 2000-2007 period (1999-2006 for funding). In order to compile such data, the list of all of Quebec university researchers and professors (N=13,479) has been provided by the *Ministère du développement économique, de l'innovation et de l'exportation* (MDEIE), as well as from Quebec's three research councils[1]. Coming from four different sources, this list included several double counts which were carefully eliminated. In addition to including the names (family name and given name) of the researchers, this list also included their university and department, which proved very helpful for the reconstitution of researchers' publication files. Each professor and university-affiliated researcher in Quebec was categorized into one of 9 disciplinary category that cover all fields of university research, which were adapted from the Classification of instructional programs (CIP – see Figure 1)[2]. Some professors affiliated with more than one department were categorized into more than one disciplinary category (N= 215).
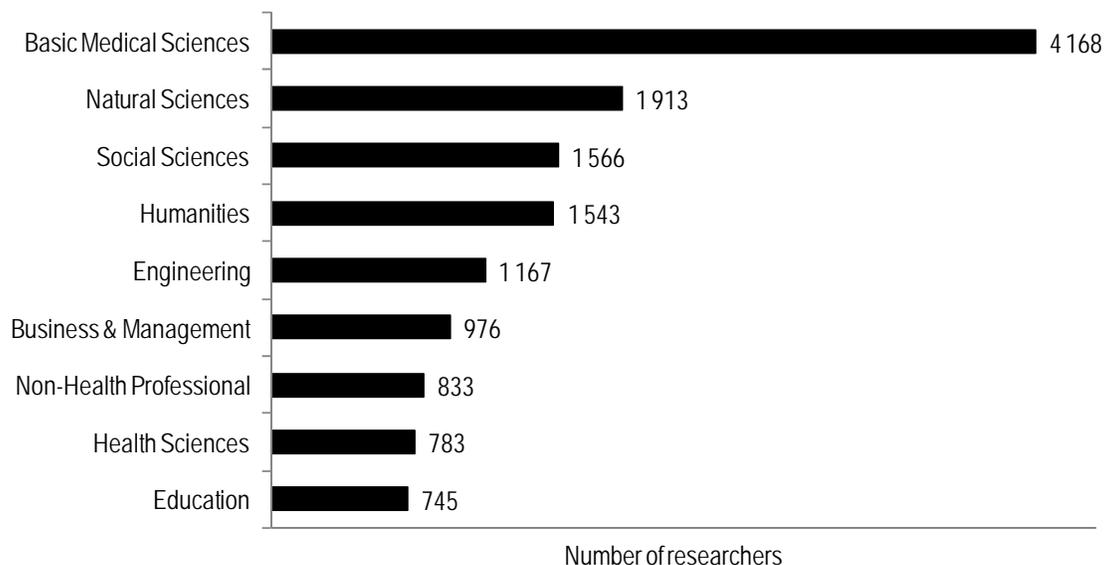
Figure 1. Number of researchers, by field

About one third of these are active in basic medical sciences. The other important groups are researchers in the natural sciences, social sciences and those in the humanities, each of these

three groups being of comparable size. Non-health professionals comprise researchers in planning and architecture, media and communications, social work, library and information science and law. Health science researchers comprise, for example, those active in public health and health administration, kinesiology and so forth (see Appendix 1 for the list of subfields included in the disciplinary categories).

All bibliometric indicators in this paper are constructed using bibliometric data from Thomson Reuters' Web of Science, (WoS) which cover about 9,000 journals annually in all fields of the natural sciences, medicine, social sciences and humanities. These databases list several types of scientific documents but, as usual in bibliometric studies, the statistics presented here are limited to articles, research notes, and review articles, which are the main forms of original publication (Moed, 1996). Thomson Reuters' databases don't cover all published research. Some scientific discoveries can be diffused in non-indexed journals, but also in other types of documents such as conference proceedings, grey literature and books. That being said, this database contains research that is published in core international journals, that is, journals that are most visible to both national and international scientific communities and that have the highest citations rates in their respective subfields (Garfield, 1990).

In addition to the obvious fact that two or more researchers can have the same name, Thomson's databases have two shortcomings when it comes to compiling statistics on individual authors. First, until very recently, bibliometric databases did not include the first names of authors of papers, but only their initials. In other words, in the bibliographic record of his papers, John William Dawson would appear as Dawson-JW or Dawson-J, which creates many namesakes, especially for very common surnames. The same applies for the citations John William Dawson might receive. The other limitation is caused by the fact that the WoS did not contain, at the time this research was carried, any information on the relationship between the authors' names (Allen-DS, Smith,-J, etc.) and their institutional addresses (Laval University, University of Montreal University, etc)[3]. Thus, for a given paper signed by John, Jack, Jane and Jacky and on which Laval and Montreal universities appear, it is impossible to know which researcher belongs to which research organization: indeed, several different combinations are possible.

Using, on the one hand, the surname and initials of professors and, on the other hand, the surname and initials of authors of Canadian scientific articles indexed by Thomson Reuters, a database of 125,656 distinct articles (and 347,421 author-article combinations) authored by these researchers and their namesakes was created. When papers were written in collaboration, one paper was attributed to each of the co-authors. In order to remove the papers authored by namesakes, each article was manually validated. This time-consuming but essential step reduced the number of distinct papers by 51% to 62,026 distinct articles and by 70% to 103,376 author-article combinations. In order to have data on « inactive » researchers, non-publishing university researchers were kept in the dataset. This constitutes a significant improvement over previous studies which generally start with a database of papers published and therefore do not have any data on non-publishing researchers. On the whole, these 62,026 distinct papers received 1,189,423 citations over the 2000-2007 period (including self-citations). Hence, citations are only counted for the set of WoS indexed papers published between 2000 and 2007. A more complete approach would be to compile all the citations received by these authors including non-source material, such as books, as well citations to source material published before 2000. Compiling citation to non-source material at the

---

[3] This has been implemented in the 2008 WoS data.

author level for such a large dataset would be prohibitively time consuming and expensive and is therefore not practical (Butler & Visser, 2006). Counting citations to source material published before 2000 would have been possible but would have meant enlarging substantially the work to be performed, which was already a time-consuming undertaking, given the available resources.

The data on research funding comes from the Information System on University Research[4], which contains all funded research in Quebec's universities and comprises data such as the projects' title, the full name of the investigators, and the sources and amounts of funding. The matching of this funding database with the list of university researchers and professors proved to be easier than the matching of bibliometric data, as both lists included a unique researcher ID to match each researcher to its funded projects. On the whole 8,787 researchers had at least one project funded, for at total of 133,273 distinct projects and a funding of Can$ 6,760,445,931 over the 1999-2006 period.

Results

Figure 2 presents the proportion of researchers who have received funding for at least one project during the last 8 years. It is interesting to note that not all researchers receive funding from external sources (that is, over their salaries). The field with the lowest proportion of researchers funded (55%) is basic medical sciences, with health sciences not far above and on a par with education at 60%. At the top of the scale, one can find engineering where more than 80% of the researchers receive funding for at least one project. These differences between fields can probably be explained by the presence in medical sciences of large research teams with a few leaders securing the lion share of the funding as principal investigators and a greater proportion of researchers who are hierarchically dependent on them and, hence, have no funding of their own.
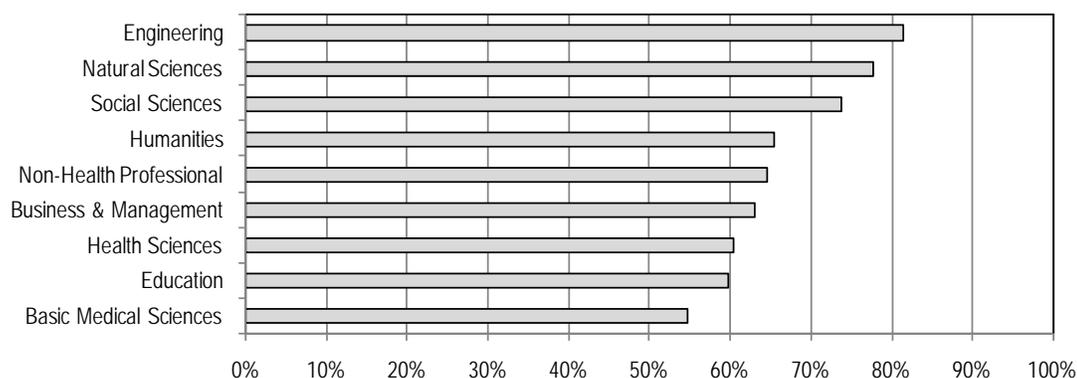
Figure 2. Percentage of researchers with at least one funded project, by field, 2000-2007

Figure 3 presents the Pareto distributions (Lorenz curves) of the number of funded research projects as well as the total amount of funding (Can$) by researcher. It is immediately obvious that, in every field, the dollar amount of research funding is considerably more concentrated than the number of project funded, which is expected given the huge variability in the size of research grants. These data also show that funding received is highly concentrated. For instance, 80% of research funds are concentrated between 11% and 18% of

---

[4] Système d'information sur la recherche universitaire (SIRU). More details on the SIRU database can be found at http://www.meq.gouv.qc.ca/stat/siru/accueil.htm.

researchers except in engineering, natural sciences and social sciences where the proverbial 80:20 rule can be observed. Business & management is the field where the funding is the most concentrated. As mentioned previously, the *number* of research project funded is distributed much more evenly—especially in the fields of natural science and in engineering, where 33-34% of researchers are on the receiving end of 80% of the projects funded. These fields are followed, in increasing order of concentration, by fields of the social sciences and humanities (SSH), and then by health sciences and basic medical sciences.
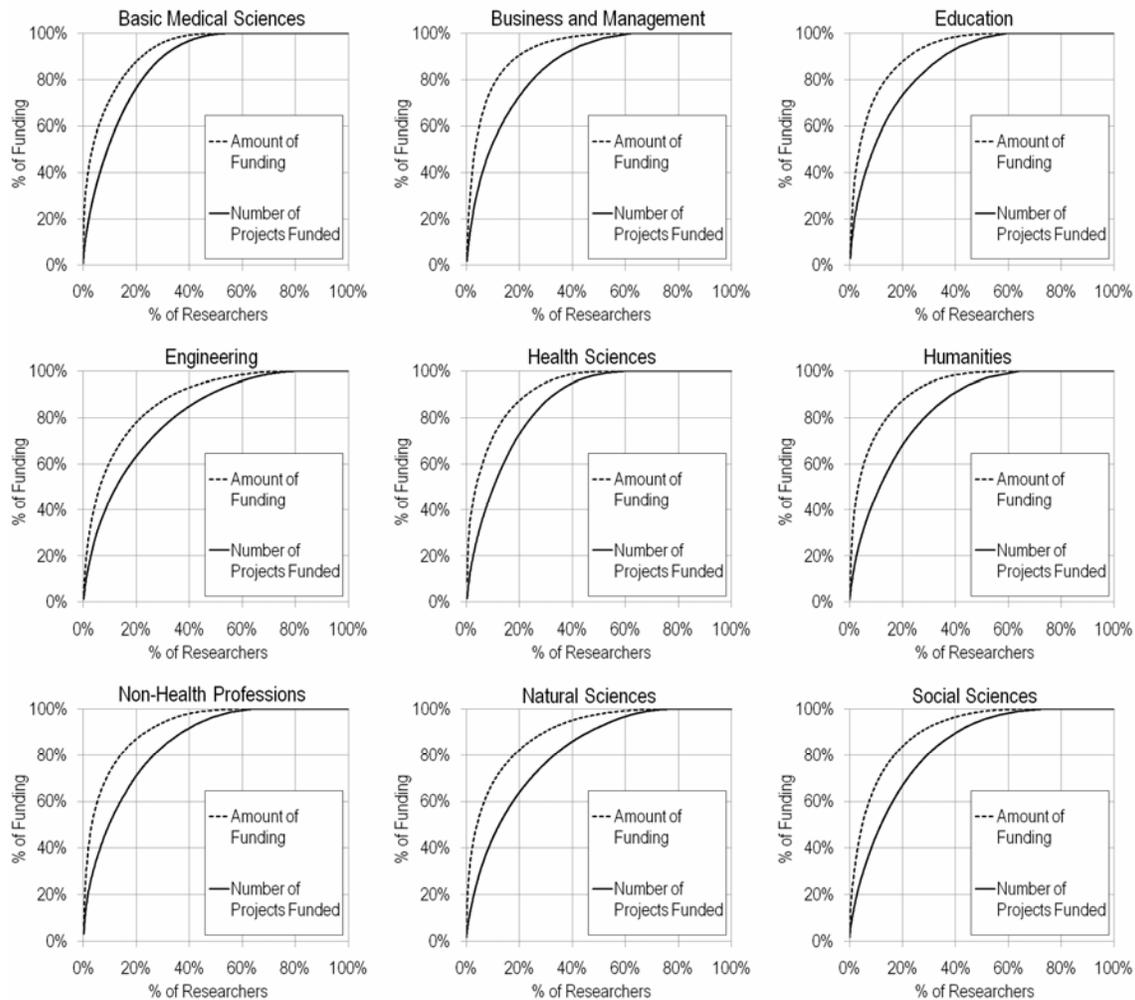
Figure 3. Pareto cumulative distributions of research funding (amount and number of projects funded per researcher) by field, 2000-2007

Figure 4 presents data on the proportion of researchers who have published at least one *paper* indexed in the WoS between 2000 and 2007 and the proportion of researchers who have received at least one citation to any one of these papers. As could be expected, the results vary considerably between fields—much more considerably than the proportion of researchers who have received research funding. There is a group of academic fields comprising the natural sciences, engineering, basic medical sciences, health sciences, and the social sciences where more than 50% of researchers have published at least one paper during the last eight years. However, fields like business and management, non-health professionals, humanities, and education, present a substantially different portrait as fewer than 50% of the researchers published at least one paper that was indexed in the WoS in the last eight years. These results

are consistent with previous research (Nederhof, Zwaan, Debruin, & Dekker, 1989, Hicks, 1999, Larivière et al., 2006) which showed that in fields such as the humanities, journal articles are not the mainstream outlet for knowledge diffusion—books being still much more cited than papers. Also, there is a linguistic factor that must be considered, as many researchers in the social sciences and humanities in Quebec will prefer to publish in French-language journals, and these are known to be under-represented in Thomson's databases (Archambault et al., 2006). Despite the existence of mitigating factors, it is interesting to note that so many researchers—close to one third of the population—have not published a single paper in eight years in the 9,000 mainstream journals indexed by Thomson.

Because cited researchers are a subset of publishing researchers, the percentage of researchers with at least one citation is lower than the percentage of researchers with at least one paper. Unsurprisingly, the fields in which there is a high percentage of researchers without any citation are also the ones in which the proportion of uncited papers is the highest (Larivière, Gingras & Archambault, 2009). For instance, uncited researchers over the 8-year period covered, represent more than half of the population in SSH fields, while they represent less than half of the population in the fields of natural sciences, engineering and health (NSE). Obviously, these two indicators (publications and citations) are highly concentrated, especially in some fields—such as the humanities, non-health professions and education—where less than 40% of researchers published a paper and 20% or less obtained at least one citation over the period.
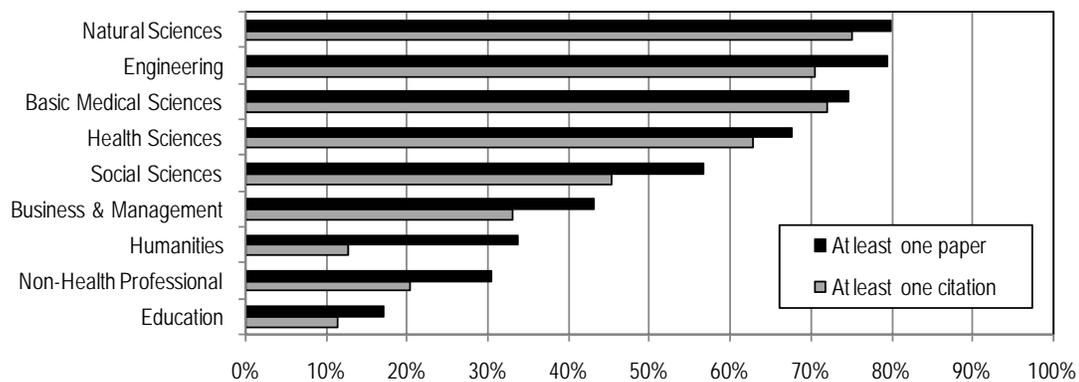
Figure 4. Percentage of researchers with at least one paper and at least one citation, by field, 2000-2007

In light of the fact there is such a large difference between fields in the proportion of researchers who do not publish, it is relevant to present Lorenz curves for the whole population of researchers as well as for the subset comprising researchers with at least one paper. Figure 5 shows that while publications are significantly more concentrated in SSH than in NSE when non-publishing researchers are included, the opposite is true when they are excluded. For example, in education, the percentage of researchers needed to account for 80% of the publications goes from 9.2% to 54% when non-publishing researchers are excluded, and goes from the most concentrated field to the most evenly distributed. Because the majority of researchers in NSE have published at least one paper over the period, the difference between the two curves is smaller. It is important to bear in mind that many researchers in SSH could in fact be publishing books or papers in journals not covered by Thomson and that the results for these fields are certainly less complete than the ones obtained for natural and medical sciences. From this point of view, the curve obtained by

eliminating the "non publishing" researchers may provide a more valuable representation of these distributions.

One can also note that citation distributions are much more concentrated than publication distributions, and that the difference between the two curves (*all* and *active*) is less pronounced. In both cases, citations are much more concentrated for fields of the SSH than for the NSE. In all fields of the SSH, except business and management, 80% of the citations are obtained by less than 20% of the *active* researchers. If *inactive* researchers are included, this figure drops to less than 10%, and as low as 3-4% for education, non-health professions and humanities. In NSE, 80% of the citations are obtained by 22-25% of *active* researchers, and by 15-20% of *all* researchers. Citations are thus much more dispersed across researchers in NSE than in SSH, whether or one includes or not the non-publishing researchers.
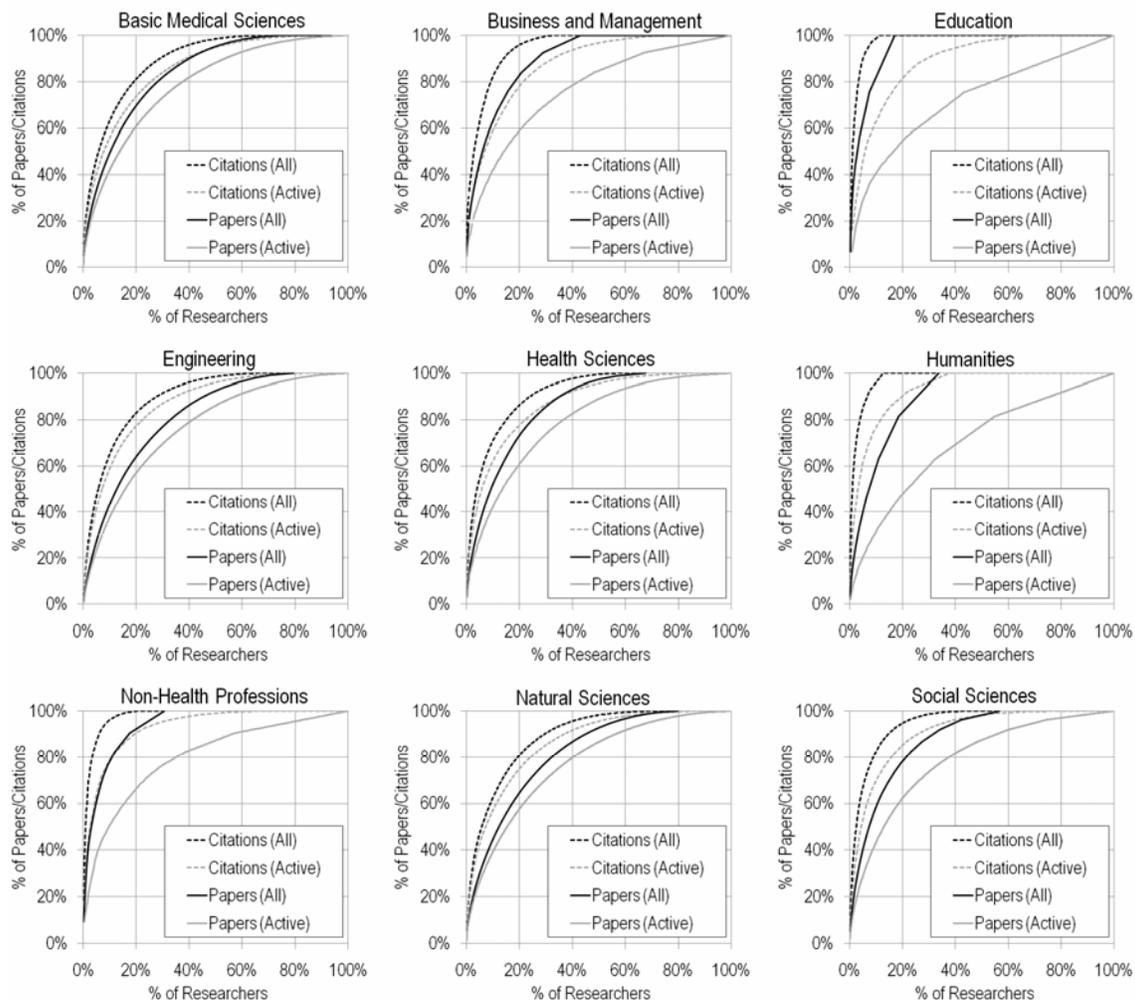
Figure 5. Pareto cumulative distributions of papers and citations, using all professors and professors who published at least a paper (active) as denominators, by field, 2000-2007

Taken globally, these figures all show that the majority of scientific resources, output and impact are obtained by a small minority of researchers. One might ask, then, if this scientific elite is composed of the same group of researchers for all three measures of activity. In other words, is the group of researchers receiving the majority of research funds also the group publishing the majority of the papers and receiving the majority of citations? Table 1 presents,

for each field as well as for all fields combined (field-normalized), the percentage of researchers who are in the top 10% in 1) at least one of the four indicators, 2) two indicators, 3) three indicators and 4) all four indicators. As one might expect, the percentage of researchers that are in the top 10% at least one of the four categories is higher than the percentage of those who are in this top ranking in all four indicators. Although slightly more than one fifth of all researchers are in the top 10% in at least one of the indicators, only 3.2% are in the top 10% in all four indicators. As one might expect, the percentage of researchers who are in the top 10% in both citations and impact (6.9%) is higher than that of researchers in the top 10% in either publications and funding or citations and funding (~4%). Though not shown, data for the top 20% showed similar patterns: about one third of all researchers are in the top 20% in at least one of the indicators, 15% are in the top in publications and citations, 10% are in the top for three indicators and about 8% are in the top in all four indicators. The data also show that the "elite" thus defined (top in all these four performance indicators) is generally more concentrated into a smaller proportion of the researchers in the SSH than in NSE, which is, again, likely to be at least in part a reflection of mismeasurement of research outputs in SSH.

Table 1. Percentage of researchers in the top 10% number of papers, citations, projects funded and total funding received, by field, 2000-2007

Indicator / Field	Basic Medical Sciences	Business & Management	Education	Engineering	Health Sciences	Humanities	Non-Health Professional	Natural Sciences	Social Sciences	All Fields
Either one of the indicators	19,1%	21,3%	21,5%	21,3%	19,7%	17,9%	21,2%	20,5%	20,0%	20,4%
Papers & Citations	6,9%	6,4%	6,8%	6,8%	7,0%	4,7%	7,6%	6,8%	7,7%	6,9%
Papers & Fundings ($)	5,0%	3,4%	3,1%	4,5%	5,0%	2,6%	3,6%	4,1%	4,2%	4,2%
Papers & Funding (N)	4,7%	4,4%	3,2%	3,9%	5,0%	2,7%	3,4%	4,0%	4,0%	4,1%
Citations & Funding ($)	4,9%	2,8%	3,0%	3,6%	4,6%	2,7%	3,0%	3,5%	3,8%	3,9%
Citations & Funding (N)	4,4%	3,2%	3,1%	3,0%	4,2%	2,4%	2,4%	3,3%	3,4%	3,6%
Citations & Papers & Funding ($)	5,0%	2,3%	2,3%	4,5%	4,1%	1,7%	2,8%	3,8%	3,6%	3,8%
Citations & Papers & Funding (N)	4,4%	2,8%	2,6%	4,0%	4,2%	1,7%	2,3%	3,5%	3,2%	3,5%
Citations & Papers & Funding ($) & Funding (N)	4,2%	2,2%	2,0%	3,9%	3,7%	1,4%	1,9%	3,2%	2,7%	3,2%

Discussion and Conclusion

This paper provides concentration measures of research funding, production and citations at the researchers' level. It reveals important differences between fields in the percentage of researchers who have published papers and received citations and funding, as well as clear differences between fields in the percentage of researchers needed to account for the majority of papers, citations and funding. Although concentration is greatest in the SSH when all researchers are included, the opposite is true when non-publishing faculty are excluded, which is likely a reflection of the fact that in those fields, a smaller share of researchers are publishing papers, books still being the privileged knowledge diffusion medium in the humanities for example. The pattern is different for funding, where the majority of researchers have had at least a grant over the period under study and where SSH researchers are, globally, on a par with their colleagues in the NSE. This is at least in part a reflection of the limitations of our indicators: whereas our funding database is fairly exhaustive and should not

discriminate in favour of one group over the other[5], our output indicator measures only one part of the scientific output, especially in the fields of SSH.

In opposition to the tendency to see "universal" distributions everywhere, these data highlight the peculiar nature of these distributions, which are both field- and indicator-dependant. Indeed, while for all fields combined, 27% of the all researchers account for 80% of the projects funded, 20%, 14% and 10% of researchers respectively account for 80% of the papers published, total funding and citations received. This percentage increases to 42% and 19% for publications and citations, respectively, when one excludes researchers that have not published at least a paper, which clearly shows the effect of keeping (or not) such "zero" cases in the distributions. The observed difference between publications and citations is also consistent with Rousseau's (1992) observations.

Although we do not have a complete explanation for these differences, some hypotheses can be suggested. As mentioned previously, the fact that researchers in SSH diffuse their research in media other than WoS indexed journals clearly increases the concentration of publications in these fields. The low citation rates and long half-lives of citations to papers published in those fields also affect the concentration of citations, as only a minority of papers are cited in the immediate years following their publication (Larivière, Gingras & Archambault, 2009). Hence, we can infer from this relationship that high productivity, high citation rates and low uncitedness seem to decrease concentration. Along the same lines, the size of the field in terms of both papers published and citations received seems to be negatively related with concentration. On the other hand, the size of research teams might affect the concentration of funding—especially in basic medical sciences—where principal investigators might have the largest grants registered under their name, but then redistribute the funding to their collaborators. On the opposite end of the spectrum, on can find the SSH where researchers work less in large teams and thus where researchers are more likely to have their own grants.

Finally, this paper also suggests that one should take care in defining the "elite" of a scientific field as it can differ appreciably depending on the indicator used. While about 20% of all researchers are in the top 10% in at least one of the chosen indicators, only about 3% are in the top 10% on all indicators. Although there are some correlations between these variables, the variance is large enough so that the individuals identified by each variable are often different, and 50% of the top funded researchers are neither in the most productive groups, nor in the most cited ones. One can also observe a greater stability of the elites identified with the different indicators in the NSE than in the SSH.

Acknowledgments

References

Allison, P.D. (1980). Inequality and scientific productivity. *Social Studies of Science*, 10, 163-179.

---

[5] Except possibly for those researchers who would receive contract money as consultants thus using their own private company instead of their employer university.

Archambault, É,. Vignola-Gagné, É., Côté, G., Larivière, V. & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, 68, 329-342.

Bradford, S.C. (1934). Sources of information on scientific subjects. *Engineering*, 137, 85-86.

Burrell, Q.L. (1991). The Bradford distribution and the Gini index. *Scientometrics*, 21, 181-194.

Butler, L. & Visser, M.S. (2006). .Extending citation analysis to non-source items. *Scientometrics*, 66, 327-343.

Cole, J.R. & Cole, S. (1972). The Ortega hypothesis. *Science*, 178, 368-375.

Cole, S. & Cole, J.R. (1973). *Social stratification in science*. Chicago: University of Chicago Press.

Cole, F.J. & Eales, N.B. (1917). The history of comparative anatomy, Part I: A statistical analysis of the literature. *Science Progress in the Twentieth Century*, 11, 578-596.

Egghe, L. (1988). The relative concentration of a journal with respect to a subject and the use of online services in calculating it. *Journal of the American Society for Information Science*, 39, 281-284.

Evans, J.A. (2008). Electronic publication and the narrowing of science and scholarship. *Science*, 321, 395-399.

Garfield, E. (1990). How ISI selects journals for coverage: Quantitative and qualitative consideration. *Essays of an Information Scientist*, 13, 185-193.

Gingras, Y., Larivière, V., Macaluso, B. & Robitaille, J.P. (2008). The effects of aging on researchers' publication and citation patterns. *PLoS ONE*, 3, e4048.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44, 193-215.

Ioannidis J.P.A. (2006). Concentration of the most-cited papers in the scientific literature: Analysis of journal ecosystems. *PLoS ONE*, 1, e5.

Larivière, V., Archambault, É., Gingras, Y. & Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57, 997-1004.

Larivière, V., Gingras, Y. & Archambault, É. (2009). The decline in the concentration of citations. *Journal of the American Society for Information Science and Technology*, 60, 858-862.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, 16, 317-323.

Macroberts, M.H. & Macroberts, B.R. (1987). Testing the Ortega hypothesis: Facts and artifacts. *Scientometrics*, 12 , 293-295.

Merton, R.K.(1968). The Matthew effect in science. *Science*, 159, 56–63.

Merton, R.K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: Chicago University Press.

Moed, H.F. (1996). Differences in the construction of SCI based bibliometric indicators among various producers: a first overview. *Scientometrics*, 35, 177-191.

Morgan, K.J. (2004) The research assessment exercise in English universities, 2001, *Higher Education*, 48, 461-482.

Nederhof, A.J., Zwaan, R.A., Debruin, R.E., & Dekker, P.J. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioral sciences: A comparative study. *Scientometrics*, 15,423-435.

Pratt, A.D. (1977). A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, 28, 285-292.

Price, D.J.D. (1963). Little science, big science. New York: Columbia University Press.

Rousseau, R. (1992). Concentration and diversity of availability and use in information systems: A positive reinforcement model. *Journal of the American Society for Information Science*, 43, 391-395.

Williams, B. (1993). Research management for selectivity and concentration -With what effect? *Higher Education Quarlerly*, 47, 4-16.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*, Cambridge: Addison-Wesley Press.

Zuckerman, H.A. (1977). *Scientific Elite: Nobel Laureates in the United States*. New York: Free Press.

Appendix 1. Composition of fields and subfields (based on CIP)

Basic Medical Sciences
  Surgical Specialties
  Medical Specialties
  Laboratory Medicine
  General Medicine
Business & Management
Education
Engineering
  Mechanical & Industrial Engineering
  Other Engineering
  Electrical & Computer Engineering
  Civil Engineering
  Chemical Engineering
Health Sciences
  Public Health & Health Administration
  Kinesiology / Physical Education
  Rehabilitation Therapy
  Nursing
  Dentistry
  Other Health Sciences
Humanities
  History
  Fine & Performing Arts
  Philosophy
  French/English
  Religious Studies & Vocations
  Foreign Languages Literature, Linguistics & Area Studies

Non-Health Professional
  Planning & Architecture
  Media & Communication Studies
  Social Work
  Library & Information Sciences
  Law & Legal Studies
Natural Sciences
  Resource Management & Forestry
  Agricultural & Food Sciences
  Earth & Ocean Sciences
  Computer & Information Science
  Biology & Botany
  Mathematics
  Physics & Astronomy
  Chemistry
Social Sciences
  Other Social Sciences & Humanities
  Political Science
  Economics
  Psychology
  Geography
  Anthropology, Archaeology & Sociology