Distribution of Knowledge Production in the Chemical Sciences in the US

by

Peter A. Lohse

Ph.D. Organic Chemistry
Federal Institute of Technology, Zürich, Switzerland, 1992

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FUFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF BUSINESS ADMINISTRATION

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2011

Signature of Author: __

Sloan School of Management
May 6, 2011

Certified by:

Fiona Murray
Associate Professor of Management
Thesis Supervisor

Accepted by: ____________

Stephen Sacca
Director, Sloan Fellows Program
MIT Sloan School of Management

DISTRIBUTION OF KNOWLEDGE PRODUCTION IN THE CHEMICAL SCIENCES IN THE US

by

Peter A. Lohse

Submitted to the MIT Sloan School of Management
on May 6, 2011 in partial fulfillment of the requirements for the Degree of
Master in Business Administration

ABSTRACT

A citation analysis was carried out to gain an understanding of the geographical and institutional distribution of highly cited articles in the chemical sciences in the US over the last thirty years. The contribution of US chemistry departments was determined by quantifying the number of highly cited articles published by individual authors or groups of authors from the same department. Articles stemming from collaborative research across schools were not considered.

The results show that a dilution in intradepartmental knowledge production has occurred both on a geographical and institutional level. Three chemistry departments have emerged as strong producers of high impact articles over the last thirty years: the University of North Carolina, Texas A&M University and the University of Utah. In terms of aggregate numbers of highly cited articles these three schools are in the top ten of over seventy schools which were evaluated; their chemistry departments are *en par* in terms of scientific impact with those from Ivy League schools like Stanford University, Harvard University and the California Institute of Technology.

While the literature reports increasing concentration for the US research base, the present analysis shows a dilution in chemical knowledge production when collaborative efforts across departments and schools are excluded. This finding suggests that the increase in concentration in the US science base is not a uniform trend when studied on a more granular level.

Thesis Supervisor: Fiona Murray

Title: Associate Professor of Management

Acknowledgements

I would like to thank Professor Fiona Murray for introducing me to the academic field of Technological Innovation & Entrepreneurship, for her friendship and advice, and for working with me during a demanding period of study.

My academic year at Sloan and this thesis would not have been possible without the support and love of my wife, Laura. She absorbed many of my family duties and created the space and time for me to attend the Fellows Program. I could not have done it without her!

Contents

Introduction

This year (2011) nations are celebrating the international year of chemistry – an unlikely event considering chemistry's absence from the headlines these days. The times when chemistry was celebrated as a central discipline for industry and business go back several decades. Today the scientific discipline "Chemistry" is taking a backset when compared to physics and the life sciences which are perceived to play a more prominent role in modern life. Still, when taking a more analytical perspective it's hard to ignore that chemistry has remained the central science which connects the fundamental theoretical concepts of physics with the more empirical problems of biology. Chemistry has been and will continue to be key to solving the most pressing global issues we are facing today: energy shortage, finding cures for unmet medical needs and devising new materials which achieve function without burdening the environment when products are made or expire.

Considering chemistry's role in tackling fundamental problems it is appropriate to celebrate but, perhaps more importantly, society needs to think about how the nation can support those who are advancing this key discipline through scientific research. Understanding where chemical knowledge is created will help to make the appropriate funding decisions on a national, state and institutional level.

The present thesis contributes to the understanding of knowledge creation in chemistry through analysis of highly cited articles in the US from 1980-2005. In contrast to the concentration which is seen for the US science base in general the results of the present analysis show a dilution in the production of intra-departmental chemical research on both a geographical as well as an institutional level. The observed dilution is mostly a result of a few chemistry departments which emerged as new and strong contributors of cutting edge chemistry research over the time period studied. The reasons for the observed diversification in high impact chemistry have not been investigated and are likely multifold.

The observed dilution observed for high-quality science from intra-departmental research suggests that the increase in concentration in the US science base is not a uniform trend when studied on a more granular level.

A brief history of chemistry

Chemistry is an old science going back to Robert Boyle (1627–1691) who refined the modern scientific method for alchemy and separated chemistry further from alchemy. Modern chemistry started flourishing shortly before the French revolution, when Antoine Lavoisier discovered the law of conversation of mass, and his refutation of the phlogiston theory of combustion in 1783.

After the nature of combustion was settled, Friedrich Wöhler's accidental synthesis of urea from inorganic substances in 1828 opened organic chemistry as a new field of research in chemistry which focused on carbon-based compounds and their derivatives. Organic compounds are widely found in nature and are highly diverse and as such constituents of many products including plastics, drugs, food and industrial materials like paints.

Systematic studies of organic compounds in the latter half of the 19^{th} century and the early twentieth century paved the way for synthesis of highly complex molecules. The art of organic synthesis peaked with the total synthesis of the complex natural product vitamin B_{12} by research groups around Albert Eschenmoser and Robert Burns Woodward in 1972.

With a detailed understanding of structure and reactivity scientists started to apply their discipline increasingly to interdisciplinary research mostly in the life sciences (e.g. biology) and physical sciences (e.g. material science) and many would argue that the most interesting chemistry is happing these days when applied interdisciplinary to further the molecular understanding in other fields.

Exponential growth of chemical knowledge

Chemistry is by far the most productive science when considering the total number of publications (1), or the number of papers abstracted by *Chemical Abstracts*. The annual Nobel awards emphasize the most important discoveries and may give the impression that the growth of knowledge in chemistry is growing linearly. This however is far from reality: Price determined 1961 that scientific output is doubling every 12-16 years (2). Chemists produce more papers than all other natural and social scientists together (3). Today over 400 times as many chemistry papers are published as in 1901, when van't Hoff received the first Nobel Prize in Chemistry for his work with solutions (Table 1).

Accessing chemistry knowledge

Today's numbers of publications in chemistry are such that no one can claim to be able to keep track of the all the knowledge produced, not even of all the knowledge produced in a sub discipline. Gone are the times when single scientists were able to understand and contribute broadly to the field of chemistry in general (like a Linus Pauling , 1901-1994). Keeping track of a comprehensive collection of scientific information today requires access to large databases. Powerful algorithms allowing the search of either text or structure (like SciFinder which allows to mine the vast Chemical Abstract Database (4)) combined with fast computer supported processing and retrieving of information remotely through the internet are the modern way for chemists to find up-to date chemical information in a comprehensive way.

Hypothesis and Research Question

Easy access to scientific information is known to promote scientific progress. For example in the late 1990s when Celera Genomics and the public sponsored Human Genome Project competed for completing the first full sequence of the human genome, Celera fed its sequences into its proprietary database with limited access, while all sequences created from the publicly funded Genome Project were made immediately available to the public with minimal restrictions. A comparison of the rate of research measured in outcomes including patents, papers published, and commercially developed diagnostic tests showed that Celera's IP led to reductions in subsequent scientific research and product development outcomes on the order of 30 percent. The reduced access to Celera's sequences thus appeared to have had a negative effect on subsequent innovation relative to a counterfactual of the Human Genome Project's sequences being available in the public domain (5)

Another example which demonstrates how openness and free information sharing can accelerate scientific problem solving and promote the generation of new knowledge involves InnoCentive Inc, an organization which facilitates problem solving by providing the solution seeker access to large community of scientists. A study which involved broadcasting problem information to a large group of outside solvers showed that providing open access to a broad range of scientific knowledge (in this case scientists) is an effective means of solving scientific problems (6).

These examples demonstrate that open access to data, information and knowledge has a stimulating effect on research in the field where new knowledge is openly shared. Considering that electronic communication media, in particular the internet, facilitate access to knowledge it can be expected that knowledge production becomes more distributed when compared to thirty years ago a time when scientific knowledge was mostly stored in large libraries of wealthy research institutions in the academic, corporate and government sector.

To the extent that the ability to produce new knowledge is grounded in access to prior scientific knowledge, materials and expertise, then improved or more open access to scientific knowledge can be expected to have a stimulating effect on R&D activities away from the traditionally known ivory towers of knowledge (7) (8). In other words, reducing or eliminating barriers to data, information and knowledge expands the population that can use and create new knowledge. For example, a lower cost of participation can be expected to induce the emergence of scientific research in places which were barred from getting involved in the past. This might have two distinctive effects – one is an increase in the level of knowledge being produced outside the elite universities. A second is an increase in the number of high quality ideas being produced outside the elite i.e. the mapping between ideas quality and institution quality may be distinctive. If scientific knowledge today is indeed being created in a more distributed fashion, and more specifically, if a larger diversity of research institutions contribute to high quality cutting-edge know-how creation today than in the past, then it should be possible to test this empirically by comparing the distribution of high quality research outcomes today with e.g. the distribution from 30 years ago.

In this thesis, the results of such an analysis are being presented for the chemical sciences in the US. A shift in output of highly-cited articles from the traditional centers of excellence to lesser know places was identified. A dilution effect was determined on both the geographical and institutional level. Geographic leveling was most apparent on a regional level and could be ascribed to strong output from schools which have been less prominent contributors to high impact chemical knowledge in the past.

Setting for studying the distribution of chemical knowledge

The present analysis considers only the top 10% of the most cited articles so that any changes are indicative of true shifts in knowledge production while ignoring the effects created by information with little impact for advancing the scientific frontier in chemistry.

Analysis of highly-cited articles in multidisciplinary chemistry was conducted for three five year time periods between 1980-2005 for articles exclusively published in the US. Bibliometric data of high impact articles was obtained by querying the Science Citation Index Expanded (SCI-EXPANDED) through the "Web of Science" database interface provided by Thomson Reuters (9).

The analysis was focused on articles in multidisciplinary chemistry. According to Thomson Reuters this includes articles which report mostly a general or interdisciplinary approach in the chemical sciences. Articles are derived from a total collection of about 150 scientific journals including well known publications like the ANGEWANDTE (10), HELVETICA CHIMICA ACTA (11), JOURNAL OF THE AMERICAN CHEMICAL SOCIETY (12), ACCOUNTS OF CHEMCIAL RESEARCH (13) and CHEMICAL COMMUNICATIONS (14). Publications having a primary focus on e.g. analytical, inorganic, organic or physical chemistry are placed in their own categories and are therefore not included in the analysis.

The "Web of Science" search capability was used to mine the expanded Science Citation Index. Searches were conducted for topic: chemistry and time period. Resulting reference sets were refined for "articles" only (while ignoring other type of publications like e.g. proceedings), "chemistry, multidisciplinary" and "US author" addresses. The resulting focused dataset typically including a hundred to a few thousand references was then exported for further analysis into a Microsoft Excel document.

Two large datasets were generated. The initial dataset was used to establish an understanding of the distribution of articles published between 1983 - 2005 comparing annual article outputs.

The second dataset combines article references and citation data from three five year time periods. This second dataset is larger and allowed to compare publication and citation data for identical periods of time. Only articles from single authors, single research groups or from within-department collaborations with an address in the US were used for analysis to ensure that geographical and institutional rankings are independent from the effects of collaborative research betweens schools.

References identified in the search were ranked based on total citations (TC). In the first dataset the top quartile of articles was used in the analysis. For the second dataset the top decile (10%) of references indentified for each of the five year periods was used for further analysis. Hence both analyses were using only top cited articles while ignoring the less cited articles or articles which had no citations.

Using the total citation count for measuring impact of knowledge

Recently new methods employing surrogate measures of scientific significance were developed for qualifying scientific output. The discipline of "Bibliometrics" uses for example citation analysis to determine the popularity and impact of specific articles and provides a measure to gauge the importance of a scientific contribution. Although citation analysis has been around for decades (the Science Citation Index began publication in 1961) it's only since 1997 that , when the *Institute for Scientific Information* (ISI) (15), now part of Thomson Reuters, made it possible to look up the complete citation record for any publication in a matter of seconds.

Using bibliometric data for judging the quality of science and scientists as well as their institutions has become increasingly popular by granting agencies whose task is obviously simplified by extensive recourse to bibliometrics. This trend has met some criticism though, as publication behavior is different from scientist to scientists: While some focus on publishing many small papers others wait for decades before they publish one large and likely more meaningful contribution. It has been argued that ranking science based on the citation indices is too simplistic and does not distinguish between hype and real scientific value which is better appraised on the responsible practice of peer review (16).

Results of the empirical analysis

The total citation count as a valid measure of article impact

The recent literature concerned with the microeconomics of knowledge generation often employs citation to academic papers to estimate the impact or quality of prior knowledge on current advances (8). While citations are certainly not the only means to measure the cumulative impact of scientific publication or article, citations are a useful though noisy indicator of the extent to which knowledge is being used by subsequent researchers. Fuhrman & Stern argue that citations are likely more informative in the life sciences where research papers typically are short, focused and with few extraneous references to the literature beyond those directly impacting the specific results described (7). The same can be said for the physical and chemical sciences and therefore, in the present study, the citation count was used as a measure for determining the scientific impact or quality of an article. Since the study involved comparison of citations of articles published in the 1980s with those published after year 2000, only the citations accumulated over a constant time frame (e.g. five years) after the publication of an article were considered in the citation count initially. However, initial calculations showed that determining the aggregate number of citations for five years following publication is cumbersome and time consuming if the datasets are large (the five year citation count cannot be obtained from the Web of Science as the necessary algorithm is not available). Hence, it was investigated if the total citation count (TC) can be used as a proxy for the number of citations accumulated over a constant time period after publication? The total citation count (TC= aggregate citations for year of publication until present) is a number available from the Science Citation Index (SCI) for every referenced article.

First, the distribution of citations for articles published within a selected year was investigated. A first set of references was generated including articles published in 1985, 1995 and 2005 (Table 2). This first data set provided a number of insights: First, the number of citations typically peaks after 3-4 years of publication and then declines steadily (Table 3). Secondly, a ranking of articles based on aggregate citations for a five year time period following publication is more or less identical with a ranking of the same articles based on total citation count (TC) (Table 4): The graphs in the second column show that the normalized counts of total citations (y-axis) closely follows normalized citations aggregated over five years for each of the publication years investigated. The third column shows aggregate citations when articles are binned based on their quartile distribution. The charts indicate that the normalized aggregate citation counts for the top quartile articles for a given publication year are more or less independent from the time period of aggregation: aggregate citation counts for the top quartile are within 6.6% for articles published in 1985, within 4.4% for articles published in 1995 and identical for articles published in 2005 (see row "Q1" in Table 2).

This finding suggests that high quality articles are typically recognized within the first five years of publication (no hidden gems). It also suggests that the order of articles following this five year period does not change much over the next 20 years (for articles published in the eighties), meaning that article ranking based on citation count does not change over time. Considering that the normalized aggregated citation count for the top cited articles for a given publication year is more or less independent from the time period of aggregation (difference clearly < 10% for shown time periods, see Table 2) the total citation count (TC – aggregate citations from publication until present) was used in all subsequent analyses as a measure of article impact/quality as this number was easily obtained from the Science Citation Index (SCI).

Growth in chemical knowledge production 1981-2005

Informed by the results from the analysis of the fist data set a new set of references (second data set) was generated for further analysis (Table 5). This second set of references considered articles published in either of three distinct five year time periods in the eighties (1981-1985), nineties (1991-1995) and in the first decade of this century (2001-2005). Compared with the first data set larger time periods (five years each) were chosen so that articles from organizations which publish less frequently are given a chance to qualify. The analysis considers only the top 10% of the most cited articles (based on total citations = TC) so that any changes are indicative of true shifts in knowledge production while ignoring the effects created by generation of information with little impact for advancing the scientific frontier in chemistry. Articles originating from industrial laboratories or independent research institutes were disaggregated. Only articles from single authors, single research groups or from within-department collaborations with an address in the US were considered in the subsequent analysis. This was done to ensure that rankings of the department or school are independent from the performance of between-school collaborations.

Assembling this second dataset provided a number of notable insights. First, the total number of articles published during the three five year periods increased continuously and was more than six-fold larger for the most recent time period (2001-2005) compared with the number of articles published during the first time period (1981-1985). This increase in articles output is slightly lower when compared with the increase of article output determined for the first data set over the same period of time (ca. eight-fold for 1985-2005, see Table 2). However, the six-fold increase in output determined for the second data is likely more reliable as articles are aggregated over longer time periods (five years for second dataset vs. one year for first data set). A six-fold increase in article output over twenty years is still high and about twice the growth rate determined by Price (2) for science in general, suggesting a faster growth in publication volume in chemistry over the last thirty years compared to other scientific disciplines.

Further analysis of the top 10% top cited articles showed that the percentage of inter-departmental works is considerable and growing faster than articles published by a single author or stemming from intra-departmental collaborations (Figure 1). The percentage of articles published by a single department or research group declined significantly from 58.3% for 1981-1985 to 43.0% for 2001-2005 (Table 5). For the most recent time period (2001-2005) articles from inter-departmental collaborations exceeded those form intra-departmental collaborations or articles from single authors (Figure 2). This finding compares well with the results by Jones et al. which showed that team work in sciences increasingly spans university boundaries and that multi-university collaborations are the fastest and the only steadily growing type of authorship structure (17).

Geographical distribution of highly cited articles across US states

In order to obtain a picture of the geographical distribution of top cited articles over time, citation counts were plotted against US state location of the author addresses (Table 6). Only highly cited articles from intra-departmental university research was considered. The departmental address was used to determine the US state.

Even though top cited articles are originating from an increasing number of states the charts do not show an obvious dilution of knowledge production. The increase in number of originating states from 14 states (for 1981-1985), to 25 states (for 1991-1995) to 29 states (for 2001-2005) can be attributed to the increasing number of articles per time period (see Table 5).

The charts in Table 6 clearly illustrate the dominant aggregate contribution of chemistry departments from universities in California (CA). CA institutions contribute ~ 20-30% of all papers. The productivity of this Western state is matched by no other and is largely independent from the point in time, meaning that CA was dominating in the nineteen eighties as it is today.

An interesting data point which is worth mentioning and which sticks out is the contribution of Colorado (CO) in the time period 1981-1985 where the total citation count clearly exceeds the number of publication by an unprecedented ca. six-fold (Table 6). This difference is due to a single publication from 1981 which has been very highly cited since (TC = 829, average TC for other articles published in 1981 is TC = 160). The highly cited article is a contribution by MATTEUCCI & CARUTHERS in the JOURNAL OF THE AMERICAN CHEMICAL SOCIETY and a seminal contribution in the field of nucleotide chemistry (18). This paper outlines the chemistry for automated nucleotide synthesis which is used today in nucleic acid synthesis (DNA & RNA) worldwide.

Besides CA, other US states hosting large numbers of Ivy League universities including Massachusetts (MA), Illinois (IL) and Ney York (NY) are well represented in the charts (Table 6). Interestingly a number of states which are lesser known for their academic institutions are emerging over the 20 year time period; Texas (TX), North Carolina (NC) and Utah (UT) show a steadily increasing contribution to top-cited articles and are prominently represented during the last time period (2001-2005). Staggering the plots for the three different time periods produces a three dimensional chart which shows the publication trends over time (Figure 3). A more quantitative analysis of top article production per state confirms the qualitative trend (Table 7).

Not surprisingly the states hosting a larger number of Ivy League universities lead the ranking including CA, MA, IL and NY. Interestingly states which traditionally have been less known for their productivity in the chemical sciences have entered these ranks. TX, NC and UT are states which show a steady increase in their contribution of top cited papers over the three time periods evaluated.

A closer analysis of the underlying references showed that a short list of universities were responsible for the increasing article output observed in these states. The University of North Carolina, Texas A&M University and the University of Utah all published seven or more top-cited articles in chemistry during the time periods investigated (Table 8).

These numbers are particularly impressive when compared with the average number of top-cited articles published per university in states well known for hosting Ivy League schools. For example in Massachusetts, three universities (Harvard, MIT and to a lesser extent University of Massachusetts, Amherst) published about 8 articles in average (Table 9).

Clearly the reason for CA's impressive productivity are the many universities this large state harbors. However, when ranked by the numbers of articles published per university in each state, CA's ranking slips somewhat to rank 4, after MA, UT and IL but before NC. TX with 6 contributing universities is on rank 13 while NY with 19 universities and less than 2 papers per school is on rank 20. Clearly there is some asymmetry in the sense that one or a few top universities per state contribute the majority of highly cited articles.

The large contribution in top cited articles of these lesser known states during the most recent time period contributes to a more balanced distribution of articles across regions when compared to the earlier time periods in the eighties and nineties (Figure 4). The imbalance in terms of regional distribution in the eighties is likely the result of the comparatively small numbers of articles during that time period which may not be representative. For example the productivity of one laboratory (Prof. T.J. Marks, NORTHWESTERN UNIV, DEPT CHEM,EVANSTON,IL 60201) results in the dominance of the Midwestern region for the time period 1983-1985.

The Western Region is emerging as the most productive region for highly cited articles in multidisciplinary chemistry. While an increase in number of highly cited articles is also found for the Midwestern and southern regions, the Northwestern region lost its dominant position in total article output to the Western region during the most recent time period (2001-2005).

The US-wide dilution in article output is supported by the decreasing Herfindahl indexes for each time period suggesting a dilution of chemical knowledge production across the 50 US states (Table 10). The "moderately concentrated" index for 1981-85 maybe a result of the limited dataset which was used for this early period. The larger number of articles published in the two more recent time periods show a significant contribution by states which were absent (e.g. TX) or barely producing top cited articles (e.g. NC) in 1981-85 suggesting that some dilution of chemical knowledge production has taken place over the last twenty years.

Ranking of schools based on number of highly cited articles published

Address analysis of the combined 237 references of highly cited articles in multidisciplinary chemistry (see last row in Table 5) showed that the articles had originated from 78 different chemistry departments. The schools were ranked by the number of articles they contributed to the top 10% most cited articles. Of a total of 78 schools only those which contributed at least two articles were considered in the final ranking of 41 schools. Schools which contributed the same number of articles were ranked based on the number of aggregate citations of those articles (Table 11).

The school ranking in Table 11 is based on number of highly cited publications each school published. The ranking order was expected to change when based on number of total citations. However, the change in order was expected to be small because only highly cited articles were considered.

Comparison of the top ten schools based on number of publications versus total number of citations (TC) showed minimal differences indeed. The only significant change in rank order was seen for the California Institute of Technology which moved from place 9 to 1 when ranked based on citations (Table 12). Overall however the comparison shows that even though the rank order is somewhat different, all schools which rank among the top ten based on number of publications also rank in the top ten when ranking is based on aggregate citation counts demonstrating that the ranking is robust as it is mostly consistent for two different measures.

Plotting the numbers of highly cited articles against schools ranked as shown above provides a graphical illustration of the distribution of article production (Figure 5). While the majority of highly-cited articles is published by the renowned universities of the Ivy League, a dilution in output becomes apparent for the most recent time period 2001-2005. The dilution in chemical knowledge production can be captured numerically by calculating the Herfindahl index for each time period (Table 13). Even though knowledge production across 41 schools is "unconcentrated", the numerical values of the Herfindahl indexes for each time period decrease significantly (2.3-fold) from 1981-85 to 2001-05.

The decrease in articles published by the top ten schools underlines the shift of chemical knowledge production away from the prestigious schools to universities which are less known for their contributions to the chemical sciences. The top ten schools published ca. 20% fewer highly cited articles in 2001-2005 compared to 1991-1995 (Table 14). The numbers for the early time period (1981-1985) are likely less meaningful as the dataset is limited (a total of only 22 top cited publications across 42 schools).

Use of funds by emerging chemistry departments

Departmental output of highly cited articles can be expected to show a positive correlation with departmental funding as larger research funds can be expected to translate into more impactful research outcomes for the following reasons: Better funding allows involving a larger number of graduate students thereby increasing the chance of new findings. Larger funds also allow for better equipment and resources which can be expected to speed up research. Finally, well-funded research departments are likely to take on more risk which eventually will lead to more impactful research outcomes and more highly cited papers.

Based on the assumptions above, and considering that more highly cited papers lead to more funding, chemistry departments which want to enter the ranks of Ivy League schools must be more efficient with using their funds compared to the already established schools.

This hypothesis was tested by plotting aggregate R&D funding for each of the 41 chemistry departments indentified above against output of highly cited papers (Figure 6). As expected a positive correlation was found between funding and output of highly cited papers.

As expected the three chemistry departments which were identified as emerging centers of high impact chemistry, including the University of North Carolina, Texas A&M University and the University of Utah, were clearly above the trend line indicating that these three departments used funds very efficiently. Calculation of USD spent per highly cited publication ranked the chemistry departments of these three schools within the top ten most cost effective research fund users (Table 15).

Discussion & Conclusions

The hypothesis underlying the present analysis was that the recent emergence of electronic communication media, in particular the internet, has had a stimulating effect on research activities at universities which traditionally did not have the infrastructure or financial means to access scientific knowledge easily. This hypothesis was tested in the current thesis by analyzing the distribution of highly cited articles for three five year time periods between 1981 and 2005 in the United States.

The results suggest that a dilution in knowledge production has occurred in the chemical sciences over the last thirty year in the US. Even though the expected trend could be confirmed empirically the relationship between knowledge creation and technology facilitating access to recorded knowledge is complex and a link between cause and effect could not established in the context of this thesis. In fact there is evidence which suggests that despite the greater geographic interconnectedness of universities the production of outstanding scientific knowledge today is taking place in fewer rather than more centers of scientific excellence (17). In the light of these findings, the increasing number of centers of high impacts chemistry identified for the most recent time period (2001-2005) is unexpected and likely the result of exogenous factors related to funding and/or policies.

The chemistry departments which have emerged as strong producers of highly cited articles over the last thirty years are from the three following schools: the University of North Carolina, Texas A&M University and the University of Utah. As efficient users of research funds these three departments had a major impact on the distribution of chemical knowledge production in the US both on the geographical and institutional level. Each of these chemistry departments published as many highly-cited papers in the most recent time period (2001-2005) as the Ivy League schools Stanford University, Harvard University and the California Institute of Technology. In terms of aggregate publication numbers the

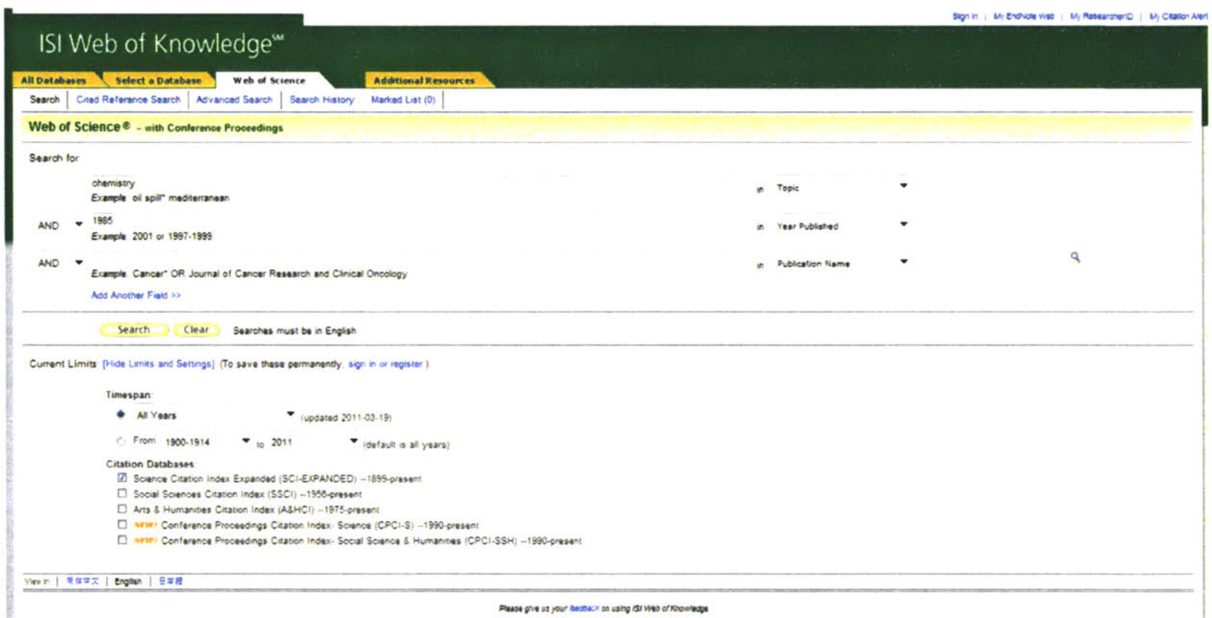
University of North Carolina, Texas A&M University and the University of Utah were in the top ten of over seventy schools which had contributed. The high impact science produced at the University of North Carolina and at Texas A&M University propelled the states NC and TX into the top ten US states with the highest impact articles in chemistry.

Thomson Reuter's Global Research Report from 2010 suggests that the US research base has been concentrated 30 years ago and is even more concentrated today (19). Jones et al. had concluded similarly that an increase in multi-university collaborations tend to embed the production of outstanding scientific knowledge in fewer rather than more centers of high impact science (17). The present analysis indentified the same trend in chemistry; while the majority of highly cited articles were mostly the product of intra-departmental collaborations or single authors during the eighties and nineties inter-departmental collaborations were the major source of high impact papers 2001-2005. The distribution of articles from inter-departmental collaboration was not investigated.

While the 2010 report by Thomson Reuter (19) finds a concentration in US science in general, similar to the trend Jones et al. (17) determined for multi-university collaborations in particular, the present analysis which excludes collaborative efforts across departments points in the other direction and suggests a dilution in chemical knowledge production in the US. This finding is not unintuitive when considering the possibility that increasing collaboration between schools unavoidably will lead to some harmonization in thinking for those involved. Scientists not involved in collaborations will be less exposed to group think and more likely to forge new insights based on their own experience and creativity. As Root-Bernstein's has observed before: new science often originates at the geographical periphery where new institutions and new programs are sometimes more easily established than at existing ones (20). The present analysis shows that scientific talent combined with funding can lead to high impact science outside the Ivy League schools and result in a more diverse research base on both a geographical and institutional level.

Methods

Paper references and citation data was downloaded from ISI Web of Knowledge data base provided by Thompson Reuters. Specifically the search window for Web of Science was used to identify and download data in the field chemistry for different time periods.

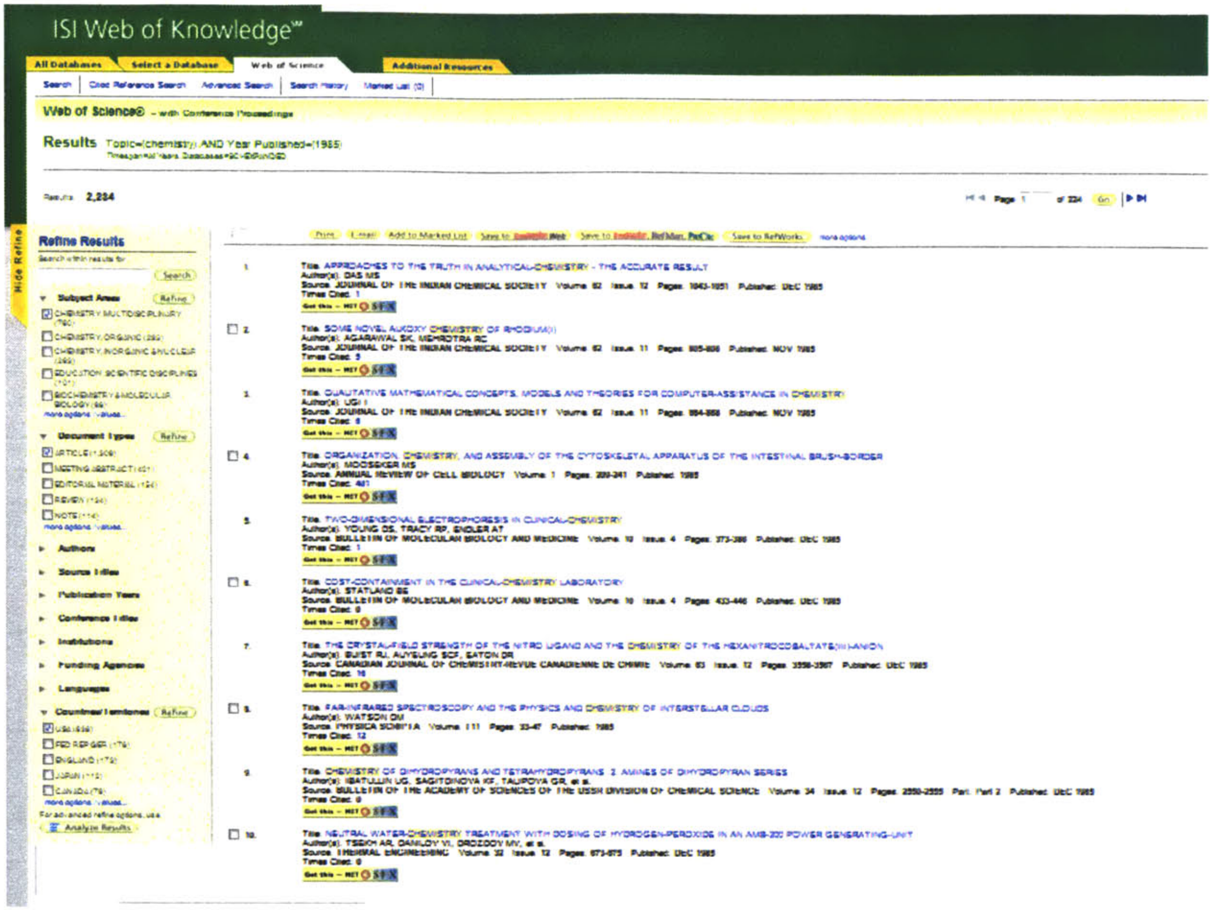


Example for a search for chemistry publications in 2008:

Search for topic [chemistry] & publication year [1985], timespan: all years, database=SCI-EXPANDED.

The search yielded 2234 hits. Results were further refined by selecting:

1. Subject areas: chemistry, multidisciplinary includes resources having a general or interdisciplinary approach to the chemical sciences.
2. Document types: articles – This means that only articles are being considered. Proceedings, review, book chapters are excluded document types.
3. Countries/Territories: USA – This means that at least one author must have an US address.

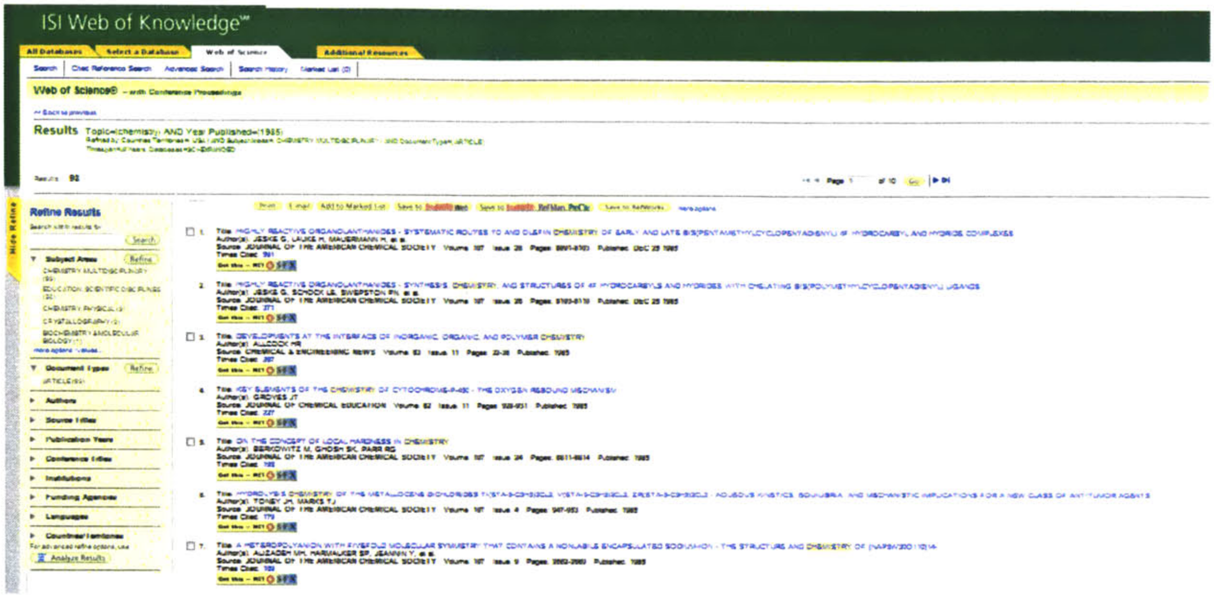


According to Thomson Reuters “chemistry, multidisciplinary” is an Institute for Scientific Information (ISI) topic which includes resources having a general or interdisciplinary approach to the chemical sciences. Special topic chemistry resources that have relevance to many areas of chemistry are also included in this category. Resources having a primary focus on analytical, inorganic and nuclear, organic, physical, or polymer chemistry are placed in their own categories. Journals categorized under "Chemistry, Multidisciplinary" are rarely co-categorized in a Chemistry sub-category; that is why there is generally no overlap between, say, "Chemistry, Multidisciplinary" and "Chemistry, Organic." Subject categories are assigned at the journal level, not the individual article level. Therefore, an article on organic chemistry that appears in a journal that has been categorized exclusively under "Chemistry, Multidisciplinary" will not be coded for "Chemistry, Organic" (21).

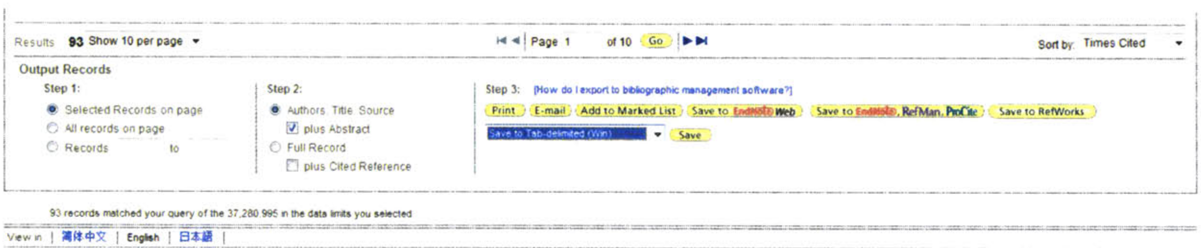
Below is a table listing all journals categorized under "Chemistry, Multidisciplinary":

ADV FUNCT MATER	ADV MATER	ABSTR PAP AM CHEM S	ACCOUNTS CHEM RES
ACS NANO	ACTA CHIM SINICA	ACTA CHIM SLOV	ACTA PHARMACOL SIN
ACTUAL CHIMIQUE	AFINIDAD	ANGEW CHEM-GER EDIT	ANGEW CHEM INT EDIT
ANN CHIM-SCI MAT	ARCH PHARM	ARZNEIMITTEL-FORSCH	ASIAN J CHEM
AUST J CHEM	BIOCONJUGATE CHEM	BULG CHEM COMMUN	B CHEM SOC ETHIOPIA
B CHEM SOC JPN	B KOREAN CHEM SOC	CAN J CHEM	CENT EUR J CHEM
CHEM SCI	CHEMIJA	CHEM CENT J	CHEM-ASIAN J
CHEMSUSCHEM	CHEM ENG NEWS	CHEM PHARM BULL	CHEM COMMUN
CHEM J CHINESE U	CHEM PAP	CHEM REC	CHEM RES CHINESE U
CHEM RES TOXICOL	CHEM REV	CHEM SOC REV	CHEM LISTY
CHEM UNSERER ZEIT	CHEM BIODIVERS	CHEM LETT	CHEM WORLD-UK
CHEM-EUR J	CHIMIA	CHIM OGGI	CHINESE CHEM LETT
CHINESE J CHEM	COLLECT CZECH CHEM C		CR CHIM
CROAT CHEM ACTA	CRYST GROWTH DES	CRYSTENGCOMM	DOKL CHEM
DRUG CHEM TOXICOL	E-J CHEM	ENERG ENVIRON SCI	ENVIRON CHEM LETT
FIBRE CHEM+	GREEN CHEM	HELV CHIM ACTA	HETEROATOM CHEM
HYLE	INDIAN J CHEM A	INT J MOL SCI	IRAN J CHEM CHEM ENG
ISR J CHEM	J CHEM INF MODEL	J COMPUT THEOR NANOS	J EXP NANOSCI
J SAUDI CHEM SOC	J SULFUR CHEM	J IRAN CHEM SOC	J MEX CHEM SOC
J CHEM ENG DATA	J CHEM EDUC	J CHEM RES	J CHEM SCI
J CHEM TECHNOL BIOT	J COMB CHEM	J COMPUT CHEM	J CONTROL RELEASE
J INCL PHENOM MACRO	J IND ENG CHEM	J MATH CHEM	J MOL MODEL
J NANOPART RES	J NANOSCI NANOTECHNO	J PHARM SCI-US	J PHYS CHEM REF DATA
J PHYS CHEM SOLIDS	J PORPHYR PHTHALOCYA	J AM CHEM SOC	J BRAZIL CHEM SOC
J CHEM SOC PAKISTAN	J CHIL CHEM SOC	J CHIN CHEM SOC-TAIP	J INDIAN CHEM SOC
J SERB CHEM SOC	J THEOR COMPUT CHEM	KOREAN J CHEM ENG	LAB CHIP
LANGMUIR	MACED J CHEM CHEM EN	MAGN RESON CHEM	MAIN GROUP CHEM
MAR CHEM	MATCH-COMMUN MATH CO	MENDELEEV COMMUN	MOL DIVERS
MONATSH CHEM	NANO TODAY	NANOSCALE	NAT CHEM
NACHR CHEM	NANO LETT	NEW J CHEM	OXID COMMUN
PHARM RES-DORDR	PHARMAZIE	PROG CHEM	PRZEM CHEM
PURE APPL CHEM	QUIM NOVA	RES J CHEM ENVIRON	RES CHEM INTERMEDIAT
REV COMP CH	REV CHIM-BUCHAREST	REV ROUM CHIM	RUSS CHEM B+
RUSS CHEM REV+	RUSS J GEN CHEM+	SAR QSAR ENVIRON RES	SCI CHINA CHEM
SEP SCI TECHNOL	SMALL	SOLID FUEL CHEM+	SOLVENT EXTR ION EXC
SOLVENT EXTR RES DEV	S AFR J CHEM-S-AFR T	STUD U BABES-BOL CHE	STRUCT CHEM
SUPRAMOL CHEM	THEOR EXP CHEM+	TOP CURR CHEM	TURK J CHEM
ULTRASON SONOCHEM			

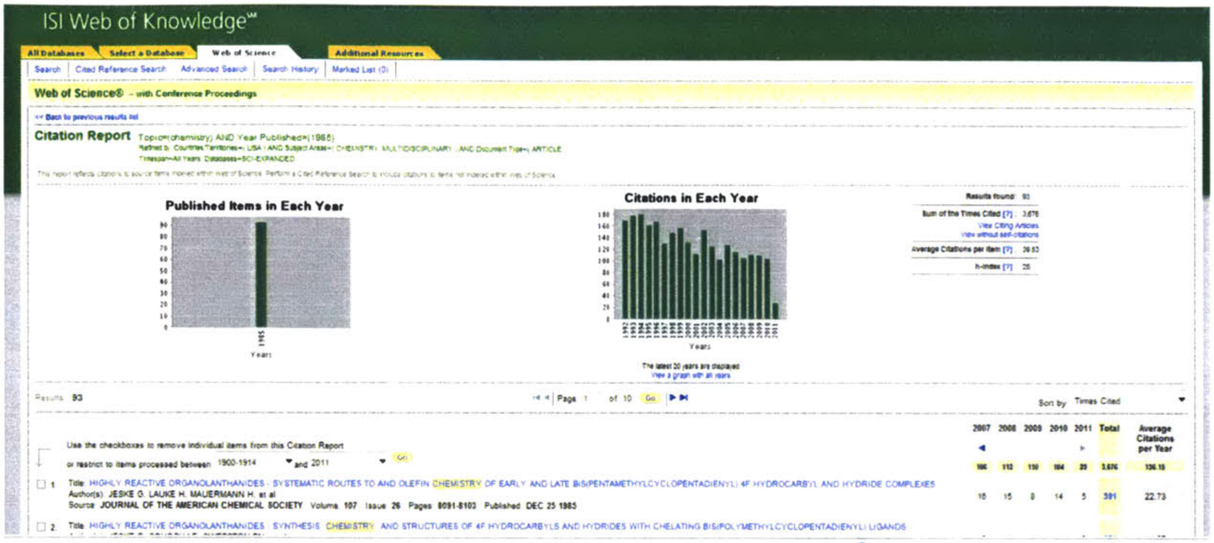
Refinement yielded 93 results which were sorted by "times cited":



All results (93) were then exported and saved to a tab-delimited Windows file:



A citation report was created clicking in the corresponding link "Create Citation Report":



Any further analysis was conducted in Microsoft Excel using the exported file.

Data, Figures and Tables

Year	# Publications & Proceedings
1901	51
2010	22,166

Table 1 Knowledge production in Chemistry today and at the beginning of the last century. Search in ISI Web of Science, Topic=(chemistry) AND Year Published=(2011), Timespan=All Years. Databases=SCI EXPANDED.

Publication year	1985		1995		2005	
Articles published	93		360		743	
Aggregation period for citations	1985-1990	1985-2011	1995-2000	1995-2011	2005-2010	2005-2011
Q1	0.815	0.869	0.661	0.690	0.677	0.677
Q2	0.156	0.104	0.222	0.209	0.209	0.209
Q3	0.029	0.024	0.102	0.086	0.093	0.093
Q4	0	0.003	0.015	0.015	0.021	0.021

Table 2 First data set: Articles published in the subject area "chemistry, multidisciplinary" published with at least one US author address. Articles were ranked based on their citation count for given time period (third row). Normalized aggregate citation counts were calculated for each quartile (Q1 - Q4).

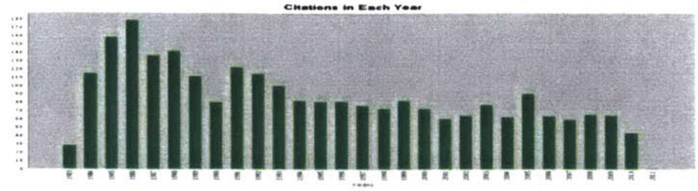
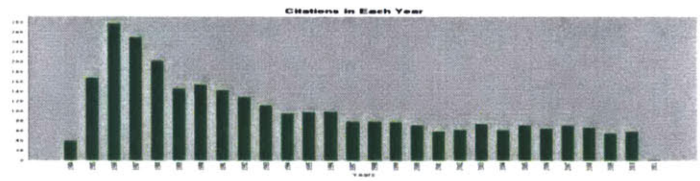
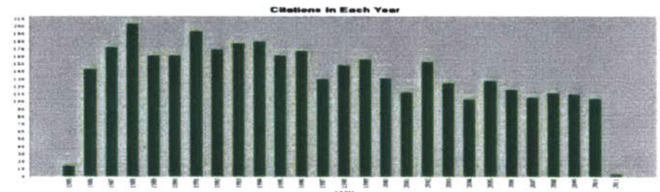
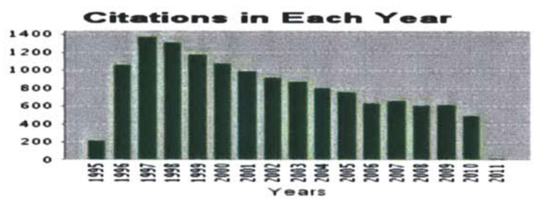
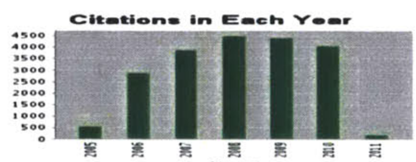
Year of publication	Distribution of annual aggregate citation counts - year of publication until 2010
1983	
1984	
1985	
1995	
2005	

Table 3 Distribution of annual aggregate citation counts over time for selected publication years.

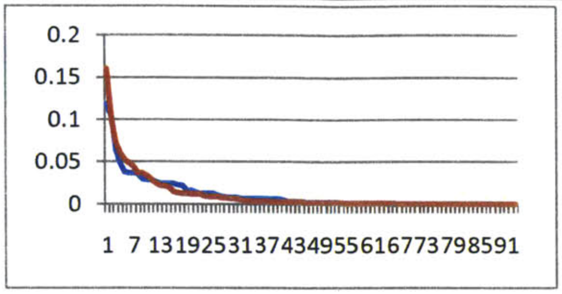
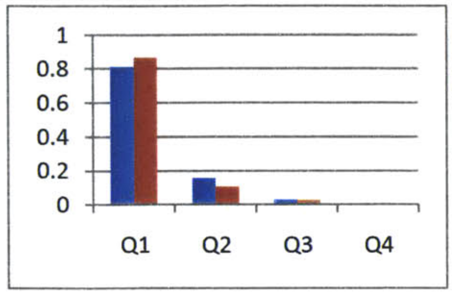
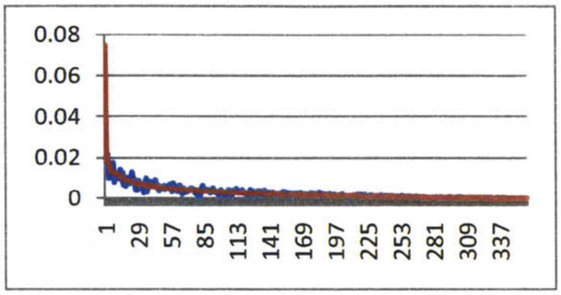
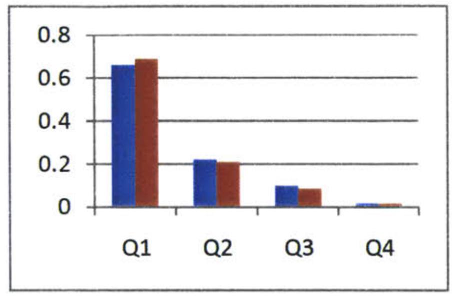
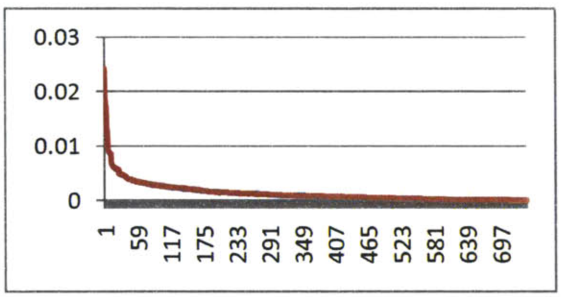
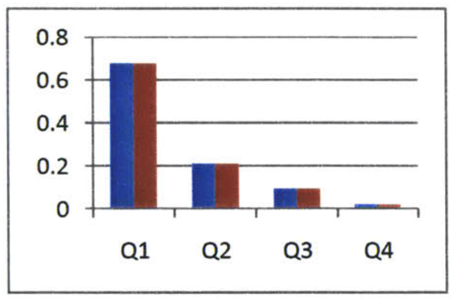
Publication Years	Ranking of articles based on citation counts	Quartile distribution of aggregate citations
1983-85		
1995		
2005		

Table 4 Ranking of articles based on citations aggregated over five years following publications (blue) or based on total number of citations (TC – red). The graphs in the second column show the normalized aggregate citation counts for articles (x-axis: articles ordered based on number of citations. The article with the highest number of citation is furthest to the right. y-axis: normalized citation count). The graphs in the third column show the normalized aggregate citation counts for each quartile.

Time period	1981-1985	1991-1995	2001-2005
total number of articles published	482	1677	3017
top 10% of highly cited articles	**48**	**168**	**302**
non-collaborative articles	28 (58.3%)	96 (57.1%)	130 (43.0%)
non-collaborative articles, universities only	27 (56.3%)	85 (50.6%)	125 (41.4%)

Table 5 Second data set: Articles published in the subject area "chemistry, multidisciplinary" in the US. Article references were obtained from SCI according to the method outlined above. The last row shows the number of articles originating from non-collaborative research at academic institutions for each five year time period.

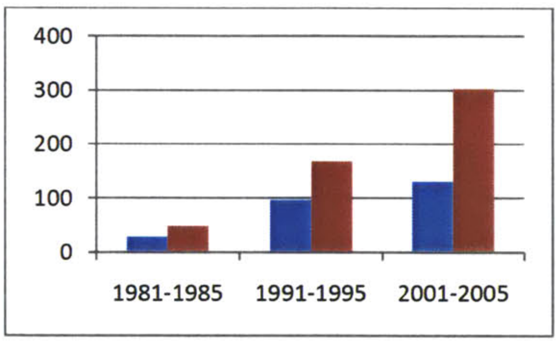


Figure 1 Articles published per time period (second data set). Total number of articles per time period in red, articles from non-collaborative research in blue.

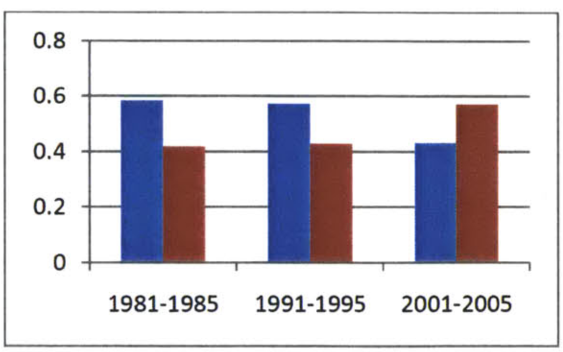
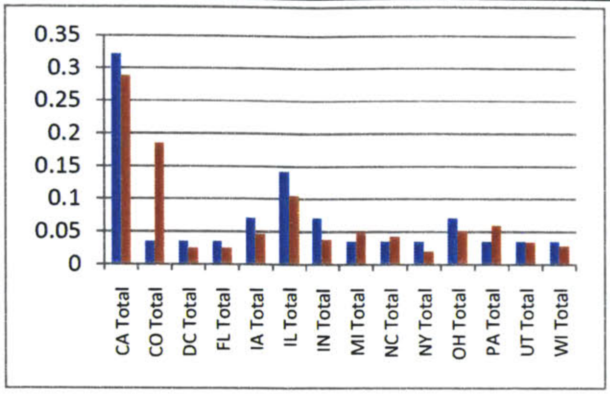


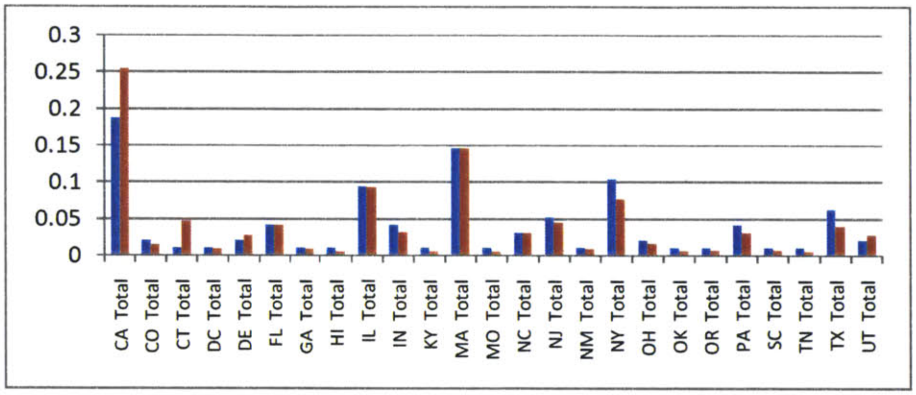
Figure 2 Fraction of articles published per time period (second data set): Total number of articles per time period in red, articles from non-collaborative research in blue.

Time period	Distribution of highly cited articles per US state

1981-1985



1991-1995



2001-2005

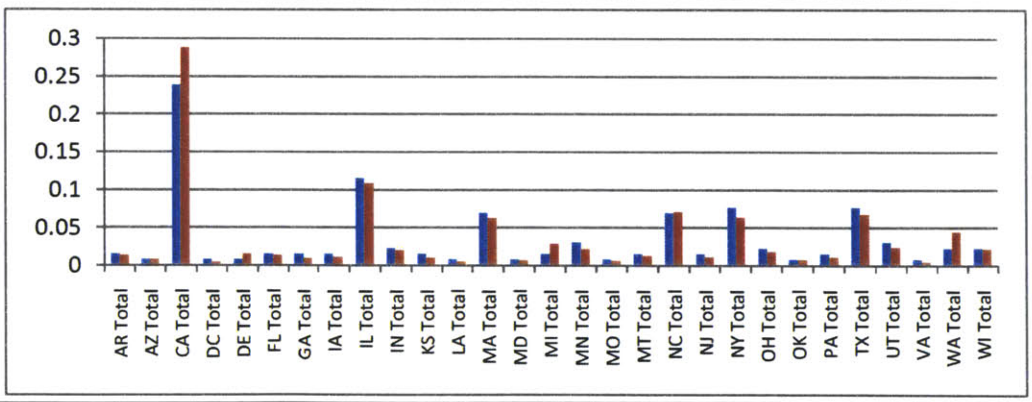


Table 6 Geographical distribution of top cited articles across US states for the three different time periods. The blue bars show the normalized number of aggregate publications while the red bar shows the normalized number of aggregate citations for a given state.

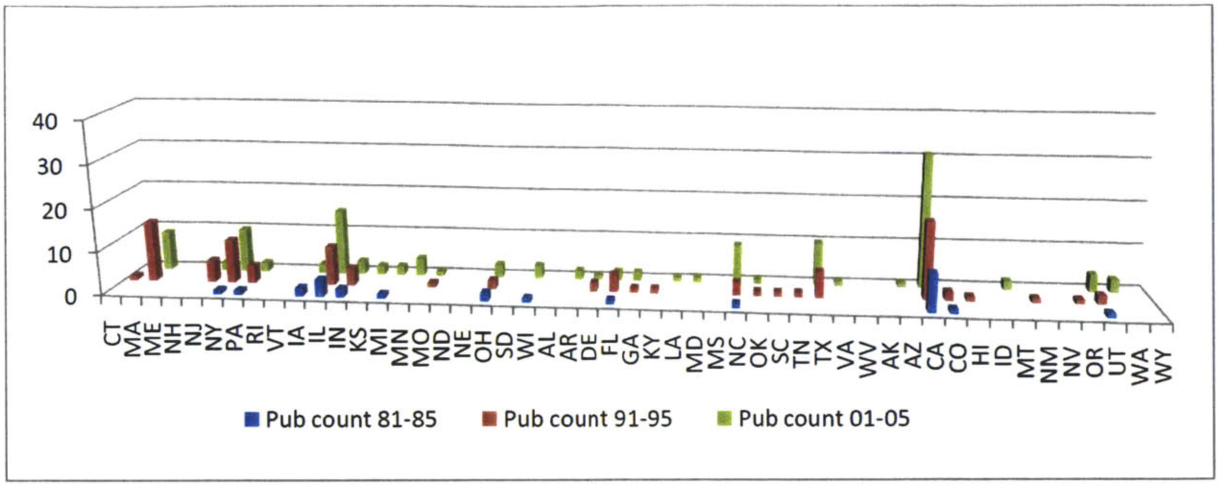


Figure 3 Distribution of top-cited article production across US states. The states are lined up alphabetically within the region affiliation.

Rank	State	# of publications per time period			Total # of articles
		1981-85	1991-95	2001-05	
1	**CA**	**9**	**18**	**31**	58
2	**IL**	**4**	**9**	**15**	28
3	MA		14	9	23
4	NY	1	10	10	21
5	**TX**		**6**	**10**	16
6	**NC**	**1**	**3**	**9**	13
7	IN	2	4	3	9
8	NJ		5	2	7
9	PA	1	4	2	7
10	OH	2	2	3	7
11	FL	1	4	2	7
12	**UT**	**1**	**2**	**4**	7

Table 7 Ranking of US states based on top-cited article output aggregated over the three five year periods. States which show a steady increase in article production are in bold.

State	University	# of articles
North Carolina	UNIV N CAROLINA	9
	DUKE UNIV	3
Texas	TEXASA&MUNIV	8
	UNIVTEXAS	3
Utah	UNIVUTAH	7

Table 8 Key contributing universities in US states which have emerged as top knowledge producers in chemistry

Rank	State	Total # of articles	Universities contributing	Articles per University
1	MA	23	3	7.67
2	UT	7	1	7.00
3	IL	27	5	5.40
4	CA	55	11	5.00
5	NC	13	3	4.33
6	MN	4	1	4.00
7	WI	4	1	4.00
8	IN	7	2	3.50
9	PA	7	2	3.50
10	DC	3	1	3.00
11	MI	3	1	3.00
12	WA	3	1	3.00
13	TX	15	6	2.50
14	FL	7	3	2.33
15	OH	7	3	2.33
16	AR	2	1	2.00
17	IA	4	2	2.00
18	MT	2	1	2.00
19	OK	2	1	2.00
20	NY	19	10	1.90

Table 9 Ranking of US states based on average number of top-cited articles contributed per school

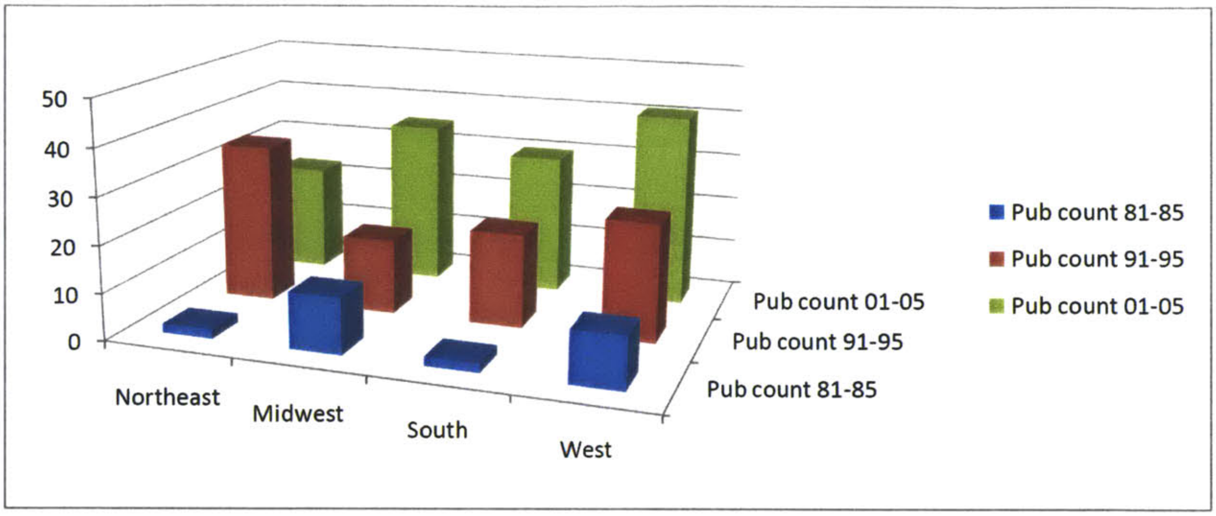


Figure 4 Regional distribution of top cited articles in the US for the three timer periods indicated. States were assigned to regions according to the US Census Bureau (22):

Time period	$H = \sum_{i=1}^{N} s_i^2$	Type of concentration
1981-1985	0.160494	moderate concentration
1991-1995	0.093629	unconcentrated
2001-2005	0.099694	unconcentrated

Table 10 Herfindahl index [H] (23) determined for the 50 US states in each of the three time periods. N = 50, s_i = normalized publication count for state i

Rank	University	# Pub 81-85	# Pub 91-95	# Pub 01-05	# Pub tot
1	MIT		9	2	11
2	UNIVCALIFBERKELEY	2	3	5	10
3	UNIVILLINOIS	1	4	5	10
4	NORTHWESTERNUNIV	3	4	2	9
5	UNIVNCAROLINA	1	1	7	9
6	STANFORDUNIV		4	5	9
7	HARVARDUNIV		5	3	8
8	TEXASA&MUNIV		4	4	8
9	CALTECH	1	4	2	7
10	UNIVUTAH	1	2	4	7
11	UNIVCALIFLOSANGELES	4	1	2	7
12	UNIVCALIFDAVIS		1	6	7
13	CORNELLUNIV		4	3	7
14	SCRIPPSRESINST		1	5	6
15	PURDUEUNIV		3	3	6
16	UNIVCHICAGO			6	6
17	UNIVSFLORIDA		3	2	5
18	UNIVPITTSBURGH		3	2	5
19	UNIVMASSACHUSETTS			4	4
20	UNIVWISCONSIN	1		3	4
21	UNIVMINNESOTA			4	4
22	UNIVWASHINGTON			3	3
23	UNIVMICHIGAN	1		2	3
24	DUKEUNIV		1	2	3
25	OHIOSTATEUNIV	1	1	1	3
26	UNIVCALIFSANDIEGO	1	1	1	3
27	CASEWESTERNRESERVEUNIV	1	1	1	3
28	GEORGETOWNUNIV	1	1	1	3
29	UNIVIOWA	1		2	3
30	UNIVTEXAS		1	2	3
31	UNIVCALIFRIVERSIDE		1	1	2
32	UNIVSOCALIF		1	1	2
33	PENNSTATEUNIV	1	1		2
34	UNIVOKLAHOMA		1	1	2
35	COLORADOSTATEUNIV		2		2
36	EMORYUNIV		1	1	2
37	UNIVARKANSAS			2	2
38	MONTANASTATEUNIV			2	2
39	NYU		1	1	2
40	SUNYBUFFALO	1		1	2
41	UNIVROCHESTER			2	2

Table 11 The 41 top schools with two or more highly-cited articles in multidisciplinary chemistry. Articles from Schools which contributed the same number of articles were ranked based on the number of aggregate citations of those articles.

Rank	University ranking based on # of publication	University ranking based on aggregate citations
1	MIT	CALTECH
2	UNIVCALIFBERKELEY	MIT
3	UNIVILLINOIS	UNIVCALIFBERKELEY
4	NORTHWESTERNUNIV	UNIVILLINOIS
5	UNIVNCAROLINA	NORTHWESTERNUNIV
6	STANFORDUNIV	HARVARDUNIV
7	HARVARDUNIV	UNIVNCAROLINA
8	TEXASA&MUNIV	STANFORDUNIV
9	CALTECH	UNIVUTAH
10	UNIVUTAH	TEXASA&MUNIV

Table 12 Top ten schools based on number of highly-cited publications and number of total citations (TC).

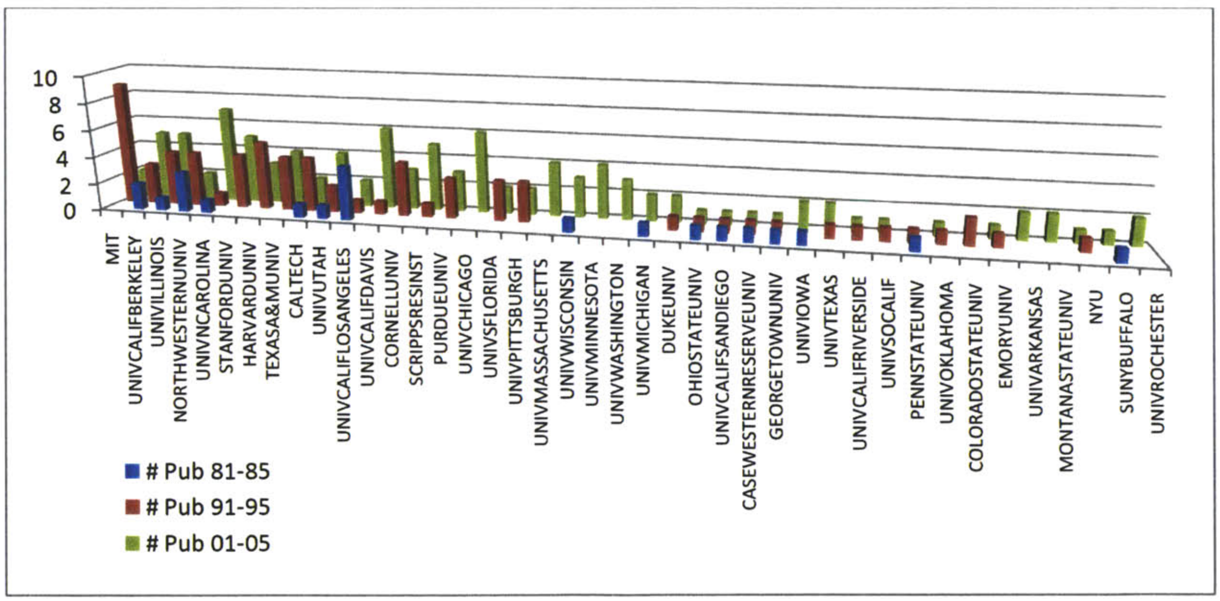


Figure 5 Distribution of highly-cited articles across schools ranked based on chemical knowledge productivity

Time period	$H = \sum_{i=1}^{N} s_i^2$	Type of concentration
1981-1985	0.081285	unconcentrated
1991-1995	0.052172	unconcentrated
2001-2005	0.034888	unconcentrated

Table 13 Herfindahl index [H] determined for the 41 schools/chemistry departments in each of the three time periods. N = 41, s_i = normalized publication count for school i.

Time period	Articles published by schools ranked 1-10	Articles published by schools ranked 11-41	% of articles published by top ten schools
1981-1985	9	13	41%
1991-1995	40	30	57%
2001-2005	39	67	37%

Table 14 Fraction of highly-cited articles published by top ten schools.

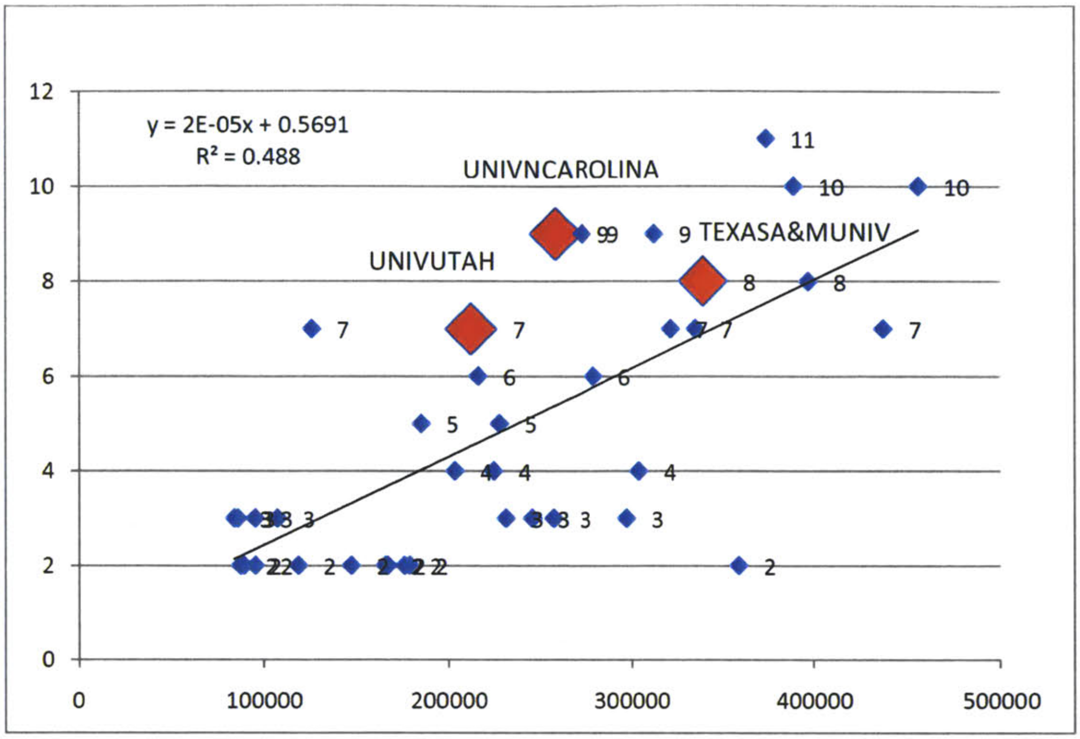


Figure 6 Output of highly cited articles per chemistry department in dependence of departmental R&D funding, in thousand [USD]. For R&D funding data see (24).

Rank	School	# of highly cited papers aggregated over 1991-95 + 2001-05	Relative cost of funding per highly cited article
1	UNIVCALIFDAVIS	7	1.00
2	UNIVNCAROLINA	8	1.79
3	MIT	11	1.89
4	STANFORDUNIV	9	1.93
5	UNIVUTAH	6	1.96
6	DUKEUNIV	3	1.99
7	UnivChicago	6	2.00
8	UNIVPITTSBURGH	5	2.06
9	GEORGETOWNUNIV	2	2.34
10	TEXASA&MUNIV	8	2.36

Table 15 Ranking of chemistry departments based on USD spent per highly cited article published.

Bibliography

1. *Coping with the Growth of Chemical Knowledge: Challenges for Chemistry Documentation, Education, and Working Chemists.* **Schummer, J.** 2, 1999, Educación Química, Vol. 10, pp. 92-101.

2. **Price, D.J. de Solla.** *Science Since Babylon.* New Haven and London : Yale Univ. Press, 1961.

3. *The law of exponential growth: Evidence, implications, and forecasts.* **Tague, J., Beheshti, J. and Rees-Potter, L.** 1981, Library Trends, Vol. 30, pp. 125-150.

4. [Online] http://www.cas.org/index.html.

5. *Intellectual Property Rights and Innovation: Evidence from the Human Genome.* **Williams, H.L.** s.l. : National Bureau of Economic Research, 2010, NBER Working Paper Series. http://www.nber.org/papers/w16213.pdf.

6. *The Value of Openness in Scientific Problem Solving.* **Lakhani, K.R., et al.** 2007, Harvard Business School Working Paper. http://www.hbs.edu/research/pdf/07-050.pdf.

7. *Climbing atop the shoulders of giants: The impact of institutions on cumulative research.* **Furman, J.L. and Stern, S.** s.l. : US National Bureau of Economic Research, 2006, NBER Working Paper Series. http://www.nber.org/papers/w12523.pdf.

8. *Of Mice and Academics: Examining the Effect of Openness on Innovation.* **Murray, F., et al.** s.l. : US National Bureau of Economic Research, 2009, NBER Working Paper Series. http://www.nber.org/papers/w14819.pdf.

9. [Online] http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/.

10. [Online] http://www.wiley-vch.de/publish/en/journals/alphabeticIndex/2002/.

11. [Online] http://www.helvchimacta.ch/.

12. [Online] http://pubs.acs.org/journal/jacsat.

13. [Online] http://pubs.acs.org/journal/achre4.

14. [Online] http://www.rsc.org/publishing/journals/cc/article.asp.

15. [Online] http://en.wikipedia.org/wiki/Institute_for_Scientific_Information.

16. *Bibliometrics as Weapons of Mass Citation.* **Molinie, A. and Bodenhausen, G.** 2010, Chimia, Vol. 64, pp. 78-89.

17. *Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science.* **Jones, B. F., Wuchty, S. and Uzzi, B.** 2008, Science, Vol. 322, pp. 1259-1262.

18. *NUCLEOTIDE CHEMISTRY .4. SYNTHESIS OF DEOXYOLIGONUCLEOTIDES ON A POLYMER SUPPORT.* **MATTEUCCI, M.D. and CARUTHERS, M.H.** 1981, J. Am. Chem. Soc., Vol. 103, pp. 3185-3191.

19. **Adams, J. and Pendlebury, D.** *Global Research Report United States.* s.l. : Thomson Reuters, 2010. http://researchanalytics.thomsonreuters.com/m/pdfs/globalresearchreport-usa.pdf.

20. **Root-Bernstein, R.S.** *Discovering.* Cambridge MA : Harvard University Press, 1989.

21. Personal communication, Rodney Chonka, Editor, Physical, Chemical & Earth Sciences, Healthcare & Science, Thomson Reuters.

22. [Online] http://en.wikipedia.org/wiki/List_of_regions_of_the_United_States.

23. [Online] http://en.wikipedia.org/wiki/Herfindahl.

24. *Chinese graduate students and US scientifci productivity.* **Gaule, P. and Piacentini, M.** 2010. NBER Conference Department - Summer Institute. http://www.nber.org/confer/2010/SI2010/PRIPE/Gaule_Piacentini.pdf.