DISCUSSION PAPER SERIES

No. 3094

THE NBER PATENT CITATIONS DATA FILE: LESSONS, INSIGHTS AND METHODOLOGICAL TOOLS

Bronwyn H Hall, Adam B Jaffe
and Manuel Trajtenberg

INDUSTRIAL ORGANIZATION

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP3094.asp**

ISSN 0265-8003

THE NBER PATENT CITATIONS DATA FILE: LESSONS, INSIGHTS AND METHODOLOGICAL TOOLS

Bronwyn H Hall, University of California, Berkeley
Adam B Jaffe, Brandeis University
Manuel Trajtenberg, Tel Aviv University and CEPR

Discussion Paper No. 3094
December 2001

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **Industrial Organization**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as a private educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions. Institutional (core) finance for the Centre has been provided through major grants from the Economic and Social Research Council, under which an ESRC Resource Centre operates within CEPR; the Esmée Fairbairn Charitable Trust; and the Bank of England. These organizations do not give prior review to the Centre's publications, nor do they necessarily endorse the views expressed therein.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

CEPR Discussion Paper No. 3094

December 2001

ABSTRACT

The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools*

This Paper describes the database on US patents that we have developed over the past decade, with the goal of making it widely accessible for research. We present main trends in US patenting over the last 30 years, including a variety of original measures constructed with citation data, such as backward and forward citation lags, indices of 'originality' and 'generality', self-citations, etc. Many of these measures exhibit interesting differences across the six main technological categories that we have developed (comprising Computers and Communications, Drugs and Medical, Electrical and Electronics, Chemical, Mechanical and Others), differences that call for further research. To stimulate such research, the entire database — about 3 million patents and 16 million citations — is now available on the NBER website. We discuss key issues that arise in the use of patent citations data, and suggest ways of addressing them. In particular, significant changes over time in the rate of patenting and in the number of citations made, as well as the inevitable truncation of the data, make it very hard to use the raw number of citations received by different patents directly in a meaningful way. To remedy this problem we suggest two alternative approaches: the fixed-effects approach involves scaling citations by the average citation count for a group of patents to which the patent of interest belongs; the quasi-structural approach attempts to distinguish the multiple effects on citation rates via econometric estimation.

JEL Classification: O30
Keywords: citations, data construction, patents and truncation

Bronwyn H Hall
Department of Economics
549 Evans Hall
UC Berkeley
Berkeley CA 94720-3880
USA
Tel: (1 510) 642 3878
Fax: (1 510) 548 556
Email: bhhall@econ.berkeley.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new~dps/dplist.asp?authorid=102247

Adam B Jaffe
Department of Economics
Brandeis University
Waltham
MA 02454-9110
USA
Tel: (1 781) 736 2251
Fax: (1 781) 736 2269
Email: ajaffe@brandeis.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new~dps/dplist.asp?authorid=121850

Manuel Trajtenberg
Eitan Berglas School of Economics
Tel-Aviv University
Tel-Aviv 69978
ISRAEL
Tel: (972 3) 640-9911
Fax: (972 3) 6409908
Email: manuel@post.tau.ac.il

For further Discussion Papers by this author see:
www.cepr.org/pubs/new~dps/dplist.asp?authorid=118771

* This Paper pulls together material from multiple research projects spanning approximately a decade, and hence it would be impossible to thank everyone who contributed to it. Still, our co-authors Rebecca Henderson and Michael Fogarty, and research assistants Meg Fernando, Abi Rubin, Guy Michaels and Michael Katz deserve special thanks. Above all, this Paper – indeed much of the patent-related research of the past two decades – could never have come to pass without the inspiration, support and criticism of our mentor Zvi Griliches. Financial support was provided by the National Science Foundation via Grants SBR-9320973 and SBR-9413099. Additional support came from the Alfred P Sloan Foundation, via a grant to the NBER Industrial Technology and Productivity Program.

Submitted 13 November 2001

I. Introduction

The goal of this paper is to describe the data base on U. S. patents that we have developed over the past decade, so as to make it widely accessible for research. In so doing we discuss key issues that arise in the use of patent citations data, and suggest ways of addressing them. We also present some of the main trends in patenting over the last 30 years, including a variety of original measures constructed with citation data, such as indices of "originality" and "generality", self-citations, backward and forward citation lags, etc. Many of these measures exhibit interesting differences across the six main technological categories that we have developed (comprising Computers and Communications, Drugs and Medical, Electrical and Electronics, Chemical, Mechanical and Others).

Broadly speaking, the data comprise detailed information on almost 3 million U. S. patents granted between January 1963 and December 1999, all citations made to these patents between 1975 and 1999 (over 16 million), and a reasonably broad match of patents to Compustat (the data set of all firms traded in the U. S. stock market). As it stands now, the data file is fully functional, and can be used with relative ease with standard software such as SAS or Access. We hope that the availability of patent data in this format will encourage researchers to use these data extensively, thus making patent data a staple of research in economics.

This represents the culmination of a long-term research and data-creation effort that involved a wide range of researchers (primarily the present authors, Rebecca Henderson, and Michael Fogarty), institutions (the NBER, REI at Case-Western, Tel-Aviv University), programmers (Meg Fernando, Abi Rubin, and Adi Raz), research assistants (notably Guy Michaels and Michael Katz), and financial resources (primarily from various NSF grants). Hopefully, the contribution of these data to present and future research in economics will justify the magnitude of the investment made.

Patents have long been recognized as a very rich and potentially fruitful source of data for the study of innovation and technical change. Indeed, there are numerous advantages to the use of patent data:

- Each patent contains highly detailed information on the innovation itself, the technological area to which it belongs, the inventors (e.g. their geographical location), the assignee, etc. Moreover, patents have very wide coverage (in terms of fields, types of inventors, etc.), and in the course of the last three decades U. S. patents increasingly reflect not only inventive activity in the U. S. itself, but also around the world.[1]
- There are a very large number of patents, each of which constitutes a highly detailed observation: the "stock" of patents is currently in excess of 6 million, and the flow is of over 150,000 patents per year (as of 1999-2000). Thus the wealth of data potentially available for research is huge.
- Patents have been granted in the U. S. continuously since the late 18th century. The current numbering and reporting system dates to the 1870s, meaning that there are (in principle) over 100 years of consistently reported data.
- In contrast to other types of economic information, the data contained in patents are supplied entirely on a voluntary basis, and the incentives to do so are plain and clear. After all, the whole idea of patents is that they constitute a "package deal," namely, the grant of temporary monopoly rights in exchange for ***disclosure***.
- Patent data include citations to previous patents and to the scientific literature. These citations open up the possibility of tracing multiple linkages between inventions, inventors, scientists, firms, locations, etc. In particular, patent citations allow one to study spillovers, and to create indicators of the "importance" of individual patents, thus introducing a way of capturing the enormous heterogeneity in the "value" of patents.

There are also serious limitations to the use of patent data, the most glaring being the fact that not all inventions are patented. First, not all inventions meet the patentability

[1] The percentage of U. S. patents awarded to foreign inventors has risen from about 20% in the early sixties, to about 45% in the late1990s.

criteria set by the USPTO (the invention has to be novel, non-trivial, and has to have commercial application). Second, the inventor has to make a strategic decision to patent, as opposed to rely on secrecy or other means of appropriability. Unfortunately, we have very little idea of the extent to which patents are representative of the wider universe of inventions, since there is no systematic data about inventions that are not patented. This is an important, wide-open area for future research. Another problem that used to be a serious hindrance stemmed from the fact that the patent file was not entirely computerized. Furthermore, until not long ago it was extremely difficult to handle those "chunks" that were computerized, because of the very large size of the data. In fact, the whole feasibility of this data construction project was called into question (certainly at the beginning of this endeavor, in the early 1990s), in view of these problems. However, rapid progress in computer technology has virtually eliminated these difficulties, so much so that at present the whole data reside in personal computers, and can be analyzed with the aid of standard PC software.

The idea of using patent data in a large scale for economic research goes back at least to Schmookler (1966), followed by Scherer (1982), and Griliches (1984).[2] The work of Schmookler involved assigning patent counts to industries (by creating a concordance between patent subclasses and SICs), whereas Griliches' research program at the NBER entailed matching patents to Compustat firms. In both cases the only data item used, aside from the match itself, was the timing of the patent (i.e. the grant or application year), such that in the end the patent data available for research consisted of patent counts by industries or firms, by year. Of course, it is the ***linking out*** of such data that made it valuable, since it could then be related to the wealth of information available on the industries/firms themselves. The project that Scherer undertook involved classifying a sample of 15,000 patents into industry of origin and industries of use, by the textual examination of each patent. The result was a detailed technology flow matrix, that again could be linked to other, external data, such as R&D expenditures on the one hand, and productivity growth on the other hand.

One of the major drawbacks of these and related research programs, extremely valuable as they had been, was that they relied exclusively on simple patent counts as indicators of some sort of innovative output. However, it has long been known that innovations vary enormously in their technological and economic "importance", "significance" or "value", and moreover, that the distribution of such "values" is extremely skewed. The line of research initiated by Schankerman and Pakes (1986) using patent renewal data clearly revealed these features of the patent data (see also Pakes and Simpson, 1991). Thus, simple patent counts were seriously and inherently limited in the extent to which they could faithfully capture and summarize the underlying heterogeneity (see Griliches, Hall and Pakes, 1987). A further (related) drawback was of course that these projects did not make use of any of the other data items contained in the patents themselves, and could not do so, given the stringent limitations on data availability at the time.

Keenly aware of the need to overcome these limitations on the one hand, and of the intriguing possibilities held by patent citations on the other hand, we realized that a major data construction effort was called for. Encouraged by the novel finding that citations appear to be correlated with the value of innovations (Trajtenberg, 1990), we undertook work aimed primarily at demonstrating the potential usefulness of citations for a variety of purposes: as indicators of spillovers (Jaffe, Trajtenberg and Henderson, 1993, Caballero and Jaffe, 1993), and as ingredients in the construction of measures for other features of innovations, such as "originality" and "generality" (Trajtenberg, Jaffe and Henderson, 1997). We used for each of these projects samples of patent data that were acquired and constructed with a single, specific purpose in mind. As the data requirements grew, however, we came to the conclusion that it was extremely inefficient if not impossible to carry out a serious research agenda on such a piece-wise basis.

[2] This is by no means a survey of patent-related work, rather we just note the key data-focused research projects that put forward distinctive methodologies, and had a significant impact on further research. For a survey of research using patent data, see Griliches (1990).

In particular, the “inversion” problem that arises when using citations received called for an all-out solution. The inversion problem refers to the fact that the original data on citations come in the form of citations ***made*** (i.e. each patent lists references to previous patents), whereas for many of the uses (certainly for assessing the importance of patents) one needs data on citations ***received***. The trouble is that in order to obtain the citations received by any ***one*** patent granted in year *t*, one needs to search the references made by ***all*** patents granted after year *t*. Thus, any study using citations received, however small the sample of patents is, requires in fact access to the whole citations data, in a way that permits efficient search and extraction of citations. The latter means in fact being able to “invert” the citations data, sorting it not by the patent number of the ***citing*** patent, but by the patent number of the ***cited*** patent. This inherent indivisibility led us to aim for a comprehensive data construction effort.[3]

The paper is organized as follows: Section II describes the data in detail, and presents summary statistics (primarily via charts) for each of the main variables. Since these statistics are computed on the basis of the ***whole*** data, the intention is both to provide benchmark figures that may be referred to in future research, as well as to highlight trends and stylized facts that call for further study. Section III discusses the problems that arise with the use of citation data, because of truncation and other changes over time in the citation process. We outline two ways of dealing with these issues, a “fixed-effects” approach, and a structural-econometric one.

II. Description of the Data

II.1 Scope, Contents and Sources of the Data

The main data set extends from January 1, 1963 through December 30, 1999 (37 years), and includes all the ***utility*** patents granted during that period, totaling 2,923,922

[3] It is interesting to note that in the early 1990s this enterprise seemed rather far-fetched, given the state (and costs) of computer technology at the time: the patent data as provided then by the Patent Office occupied about 60 magnetic tapes, and the inversion procedure (of millions of citations) would have necessitated computer resources beyond our reach. However, both computers and data availability improved along the way fast enough to make this project feasible.

patents;[4] we shall refer to this data set as **PAT63_99**. This file includes two main sets of variables, those that came from the Patent Office ("original" variables), and those that we created from them ("constructed" variables). The citations file, **CITE75_99**, includes all citations made by patents granted in 1975-1999, totaling 16,522,438 citations. In addition, we have detailed data on inventors, assignees, etc. The patent data themselves were procured from the Patent Office, except for the citations from patents granted in 1999, which come from MicroPatent. The **PAT63_99** file occupies less than 500 MB (in Access or in SAS), the **CITE75_99** about 260 MB. The contents of these files are as follows:

1. *PAT63_99*

(i) Original Variables:[5]

1. Patent number
2. Grant year
3. Grant date[6]
4. Application year (starting in 1967)
5. Country of first inventor
6. State of first inventor (if U. S.)
7. Assignee identifier, if the patent was assigned (starting in 1969)
8. Assignee type (i.e., individual, corporate, or government; foreign or domestic)
9. Main U.S. patent class
10. Number of claims (starting in 1975)

(ii) Constructed variables:

1. Technological category
2. Technological sub-category
3. Number of citations made
4. Number of citations received
5. Percent of citations made by this patent to patents granted since 1963[7]
6. Measure of "generality"

[4] In addition to utility patents, there are three other minor patent categories: Design, Reissue, and Plant patents. The overwhelming majority are utility patents: in 1999 the number of utility patents granted reached 153,493, versus just 14,732 for Design patents, 448 Reissue, and 421 Plant. Our data do not include these other categories.

[5] We also have the patent subclass, and the SICs that the Patent Office matched to each patent. However, we have not used these data so far, and they are not included in the PAT63_99 file.

[6] Number of weeks elapsed since January 1, 1960.

[7] That is, for each patent we compute the following ratio: number of citations made to patents granted since 1963 divided by the total number of citations made. The point is that older citations are not in our data, and hence for purposes such as computing the measure of originality, the actual computation is done only on the basis of the post-63 citations. However, one needs to know to what extent such calculations are partial.

7. Measure of “originality”
8. Mean forward citation lag
9. Mean backwards citations lag
10. Percentage of self-citations made –upper and lower bounds

2. *CITE75_99*

1. Citing patent number
2. Cited patent number

3. *The “Inventors” file*

This file contains the full names and addresses of each of the multiple inventors listed in each patent (most patents have indeed multiple inventors, the average being over 2 per patent). Both the names of the inventors and their geographical locations offer a very rich resource for research that has yet to be fully exploited.

4. *The “Coname” file*

1. Assignee identifier (numerical code, as it appears in PAT63_99)
2. Full assignee name

5. *The Compustat match file* *(see II.11 below)*

II.2 Dating of Patents, and the Application – Grant Lag

Each patent document includes the date when the inventor filed for the patent (the ***application*** date), and the date when the patent was granted. Our data contains the grant ***date*** and the grant ***year*** of all patents in the file (i.e., of all utility patents granted since 1963) and the application ***year*** for patents granted since 1967.[8] Clearly, the actual timing of the patented inventions is closer to the application date than to the (subsequent) grant date. This is so because inventors have a strong incentive to apply for a patent as soon as possible following the completion of the innovation, whereas the grant date depends upon the review process at the Patent Office, which takes on average about 2 years, with a

[8] Actually the grant year can be retrieved from the patent numbers, since these are given sequentially along time. Moreover, the Patent Office publishes a table indicating the first and last patent number of each grant year.

significant variance (see Table 1). Indeed, the mode of operation of the Patent Office underwent significant changes in the past decades, thereby introducing a great deal of randomness (that have nothing to do with the actual timing of the inventions) into any patent time series dated by grant year.

Thus, and whenever possible, the application date should be used as the relevant time placer for patents.[9] On the other hand one has to be mindful in that case of the ***truncation*** problem: as the time series move closer to the last date in the data set,[10] patent data timed according to the application date will increasingly suffer from missing observations consisting of patents filed in recent years that have not yet been granted. Table 1 shows the distribution of application-grant lags for selected sub-periods, as well as the mean lag and its standard deviation.[11] Overall the lags have shortened significantly, from an average of 2.4 years in the late 1960s to 1.8 years in the early 1990s, at the same time as the number of patents examined (and granted) more than doubled. Notice however that the trend was not monotonic: during the early 1980s the lags in fact lengthened, but shortened again in the second half of the 1980s and early 1990s. Notice also that the percentage granted 2 years after filing is about 85% (for recent cohorts), and after 3 years about 95%. Thus, it is advisable to take at least a 3-year "safety lag" when dating patents according to application year, and/or to control for truncation, for example by including dummies for years.

II.3 Number of Patents

Figure 1 shows the annual number of granted patents by application year, and Figure 2 the number of patents by grant year. The extent of the truncation problem can be clearly seen in Figure 1, for the years 1996-99: the sharp drop in the series is just an artifact reflecting the fact that the data include patents granted up to the end of 1999, and hence for the years just before that we only observe those patent applications that were

[9] The series for the patent variables that we present below are indeed mostly by application year, and include data up to 1997: given that we have patents granted only up to December 1999, there are too few applications for 1998 and 1999.

[10] For our data this date is December 1999.

[11] The figures presented there may still suffer slightly from truncation: there probably are patents applied for in 1990-92 that still were not granted by 12/1999.

granted relatively fast, but not all those other patents that will be granted afterwards. The series in Figure 1 are smoother than those in Figure 2, reflecting the changing length of the examination process at the Patent Office, which causes the series dated by granting date to vary from year to year in a rather haphazard way.

Figure 1 shows that the total number of successful patent applications remained roughly constant up to 1983, oscillating around 65,000 annually, and then took off dramatically, reaching almost 140,000 in the mid 1990's. In terms of patents granted, the single most pronounced changed occurred between 1997 and 1998, when the number of patents granted increased by almost 1/3 (from 112K to 148K). In terms of composition, the number of patents granted to U. S. inventors actually declined up to 1983, but such decline was almost exactly compensated by the increase in the number of patents granted to foreigners. Despite these differences for the pre-1983 period, the acceleration that started in 1983 applies both to U. S. and to foreign inventors (see Kortum and Lerner, 1998). Note in Figure 2 that the turning point there (i.e. according to grant year) would appear to have occurred in 1979, but that just reflects the application-grant lag (and changes in that respect) and not a "real" phenomenon.

II.4 Types of Assignees

The USPTO classifies patents according to the type of assignees, into the following seven categories (the figures are the percentages of each of these categories in our data):

1 – Unassigned	18.4%
Assigned to:	
2 – U. S. non-government organizations (mostly corporations)	47.2%
3 – Non-U. S., non-government organizations (mostly corporations)	31.2%
4 – U. S. individuals	0.8%
5 – Non-U. S. individuals	0.3%
6 – The U. S. Federal Government	1.7%
7 – Non-U. S. Governments	0.4%

"Unassigned" patents are those for which the inventors have not yet granted the rights to the invention to a legal entity such as a corporation, university or government agency, or to other individuals. These patents were thus still owned by the original inventors at the time of patenting, and they may or may have not transferred their patent rights at a later time (we do not have data on transfers done after the grant date). By far the vast majority of patents (78.4%) are assigned to corporations,[12] and another 18.4% are unassigned. Of the remaining ones, 2.1% are assigned to government agencies, and 1.1% to individuals. This later category is thus unimportant, and for practical purposes can be regarded as part of the "unassigned" category. As Figure 3 shows, the percentage of corporate patents for U. S. inventions increased slightly over the period from 72% to 77%, whereas for foreign patents the increase was much steeper, from 78% in 1965 to 90% in 1997. The increase in the share of corporate inventions reflects the long-term raising dominance of corporations as the locus of innovation, and the concomitant relative decline of individual inventors.

II.5 Technological Fields

The USPTO has developed over the years a highly elaborate classification system for the technologies to which the patented inventions belong, consisting of about 400 main (3-digit) patent classes,[13] and over 120,000 patent subclasses. This system is being updated continuously, reflecting the rapid changes in the technologies themselves, with new patent classes being added and others being reclassified and discarded.[14] Each patent is assigned to an "original" classification (class and subclass), and to any number of subsidiary classes and subclasses. For the vast majority of uses one is likely to resort only to the original, 3-digit patent class, and hence we include only it in the **PAT63_99** file.

Furthermore, even 400 classes are far too many for most applications (such as serving as controls in regressions), and hence we have developed a higher-level

[12] The category refers as said to "non-government organizations", which consists overwhelmingly of business entities (i.e. corporations), but includes also universities.

[13] There were 417 classes in the 1999 classification, which is the one we use.

[14] From time to time the Patent Office reassigns patents retroactively to patent classes according to the most recent patent classification system. Therefore, one has to be careful when using jointly data files created at different times, or when adding recent patents to older sets.

classification, by which the 400 classes are aggregated into 36 two-digit technological sub-categories, and these in turn are further aggregated into 6 main categories: Chemical (excluding Drugs); Computers and Communications (C&C); Drugs and Medical (D&M); Electrical and Electronics (E&E); Mechanical; and Others (see Appendix 1). Of course, there is always an element of arbitrariness in devising an aggregation system and in assigning the patent classes into the various technological categories, and there is no guarantee that the resulting classification is "right", or adequate for most uses. For example, we found that within the category Drugs and Medical there is a high degree of heterogeneity between sub-categories in some of the dimensions explored: the sub-category Drugs (no. 31) exhibits a much higher percentage of self-citations than the others, and Biotechnology (no. 32) scores significantly higher in terms of generality and originality. Thus, we suggest that while convenient, the present classification should be used with great care, and reexamined critically for specific applications.

Figure 4 shows the number of patents in each of the six technological categories over time by application year, Figure 5 expresses these numbers as shares of total patents. The changes are quite dramatic: the three traditional fields (Chemical, Mechanical and Others) have experienced a steady decline over the past 3 decades, from about 25% to less than 20% each. The big winner has been Computers and Communications, which rose steeply from 5% in the 1960s to 20% in the late 1990s, and also Drugs and Medical, which went from 2% to over 10%. The only stable field is Electrical and Electronics, holding steady at 16-18%. All told the 3 traditional fields dropped from 76% of the total in 1965 to 54% in 1997 by application year. (Their share of 1999 grants was just 51%.) This clearly reflects the much-heralded "technological revolution" of our times, associated with the rise of Information Technologies on the one hand, and the growing importance of Health Care Technologies on the other hand.

Figure 4 reveals yet another aspect of these changes: The ***absolute*** number of patents in the traditional fields (Chemical, Mechanical and Others) declined slightly up to 1983 (certainly during the late seventies), and then increased by 20-30%. By contrast, the emerging fields of Computers and Communications and Drugs and Medical increased throughout the whole period, with a marked acceleration after 1983. All told, the absolute number of patents in C&C experienced a ***5-fold*** increase since 1983, and similarly for those in D&M. This makes clear both the extent to which there was a turning point in the early 1980s (across the board), and the dramatic changes in the rates of growth of innovations in emerging versus traditional technologies. Comparing patents of U. S. versus non-U. S. inventors, the only significant difference is that the field of D&M grew significantly faster in the U. S.: by the mid 1990s the share of D&M for U. S. inventors was 12%, versus 8% for non-U. S..

II.6 Citations Made and Received

A key data item in the patent document is “References Cited – U. S. Patent Documents” (hereafter we refer to these just as “citations”). Patent citations serve an important legal function, since they delimit the scope of the property rights awarded by the patent. Thus, if patent B cites patent A, it implies that patent A represents a piece of previously existing knowledge upon which patent B builds, and over which B cannot have a claim. The applicant has a legal duty to disclose any knowledge of the “prior art,” but the decision regarding which patents to cite ultimately rests with the patent examiner, who is supposed to be an expert in the area and hence to be able to identify relevant prior art that the applicant misses or conceals. The presumption is thus that citations are informative of links between patented innovations. First, citations made may constitute a “paper trail” for spillovers, i.e. the fact that patent B cites patent A may be indicative of knowledge flowing from A to B; second, citations ***received*** may be telling of the “importance” of the cited patent.[15] The following quote provides support for the latter presumption:

> “..the examiner searches the…patent file. His purpose is to identify any prior disclosures of technology… which might be similar to the claimed invention and limit the scope of patent protection...or which, generally, reveal the state of the technology to which the invention is directed. If such documents are found...they are “cited”... if a single document is cited in numerous patents, the technology revealed in that document is apparently involved in many developmental efforts.

[15] See Jaffe, Trajtenberg and Fogarty (2000) for evidence from a survey of inventors on the role of citations in both senses.

> Thus, the number of times a patent document is cited may be a measure of its technological significance." (OTAF, 1976, p. 167)

Beyond that, one can construct citations-based measures that may capture other aspects of the patented innovations, such as "originality", "generality", "science-based", etc. (see Trajtenberg, Jaffe and Henderson, 1997). We discuss below some of these measures.

Our data include citations made starting with grant year 1975, and to the best of our knowledge there are no computerized citations data prior to that.[16] Figure 6 shows the mean number of citations made and received over time. Notice the steep rise in the number of citations made: from an average of about 5 citations per patent in 1975, to over 10 by the late 1990s.[17] This increase is partly due to the fact that the patent file at the USPTO was computerized during the 1980s, and hence patent examiners were able to find potential references much more easily.[18] Beyond that, we cannot tell the extent to which some of the rise may be "real" as opposed to being purely an artifact that just reflects changing practices at the USPTO. Thus, one has to be very careful with the time dimension of citations, and use appropriate controls for citing years.

The decline in the number of citations received in recent years as shown in Figure 6 is a result of truncation: patents applied for in say 1993 can receive citations in our data just from patents granted up to 1999, but in fact they will be cited by patents in subsequent years as well, only that we do not yet observe them. Obviously, for older patents truncation is less of an issue; in general, the extent to which truncation is a problem depends on the distribution of citation lags, which we examine below. Notice

[16] Citations were made before 1975, and may have resided within the PTO in some computerized form. However, we have not been able to establish when precisely the current citation practices started at the USPTO, and moreover, no publicly available electronic data of which we are aware contains pre-1975 (grant year) citations.

[17] The decrease in the mean number of citations made after 1995 in the series plotted by *application* year is somewhat puzzling, in view of the fact that the series keeps rising when *plotted* by grant year. The divergence may be due to the fact that patent applications that make fewer citations are less "complex" and hence are granted relatively quickly.

[18] Another reason may be the steep rise in the number of patents granted since 1983, which means that there are many more patents to cite.

that patents applied for prior to 1975 also suffer from truncation, but in a different way: a 1970 patent will have all the citations received from patents granted since 1975, but none of the citations from patents granted in 1970-74. Truncation thus reinforces the need to use appropriate controls for the timing of citations, beyond the aforementioned problem of the rising number of citations made.

Figure 7 shows the number of citations made by technological categories, and Figure 8 does the same for citations received. Clearly, patents belonging to different technological categories diverge far more in terms of citations ***received*** than in terms of citations ***made***. In general, the traditional technological fields cite more and are cited less, whereas the emerging fields of C&C and D&M are cited much more but are in between in terms of citations made. Thus, the category Others displays the highest number of citations made, Electrical and Electronics the lowest, Computers and Communications makes as many citations as Chemicals, whereas Drugs and Medical went from making the lowest number of citations to making the second highest.

On the receiving side, the distinction between traditional and advanced fields is clear-cut, and the differences are very large. Thus, C&C received up to 12 citations per patent (twice as many as Mechanical), D&M about 10, E&E over 7, whereas the traditional fields received just about 6. Once again, we do not know whether the differences in citations ***made*** reflect a “real” phenomenon (e.g. fields citing less are truly more self-reliant, and perhaps more “original”), or rather different citation practices that are somehow artifactual. On the other hand the differences in citations received are more likely to be “real”, since it is hard to believe that there are widespread practices that systematically discriminate between patents by technological fields when making citations.

II.7 Citation Lags

There are two ways to look at citation lags, backwards and forward. The backward lags focus on the time difference between the application or grant year of the ***citing*** patent, and that of the ***cited*** patents. For patents granted since 1975 we have the complete list of citations ***made***, we know their timing, and therefore we can compute for them the entire distribution of backward citation lags. When we look at citations received and hence at forward lags the situation is very different, because of truncation: for patents granted in 1975 the citations lags may be at most of 24 years, and for more recent patents the distribution of lags is obviously truncated even earlier.

Figure 9 shows the frequencies of backward citation lags up to 50 years back, and separately the remaining tail for lags higher than 50; Figure 10 shows the cumulative distribution up to 50 years back.[19] The striking fact that emerges is that citations go back very far into the past (some even over a hundred years!), and that to a significant extent patents seem to draw from old technological predecessors. Thus, 50% of citations are made to patents at least 10 years older than the citing patent, 25% to patents 20 years older or more, and 5% of citations refer to patents that are at least 50 years older than the citing one! Reversing the perspective, if this distribution and the number of patents granted were to remain stable over the long haul, patents granted in year 2,000 will receive just half of their citations by 2,010, 75% by 2,020, and even by 2,050 they will still be receiving some. Of course, we know very little about the stability of the lag distribution (strictly speaking it is impossible to ascertain it), but there is some indication that the lags have been shortening lately, as evidenced by the following figures for various cohorts of citing patents:

[19] These distributions are computed by taking each citation to be an observation, rather than by taking the average lag for each patent. The backward lags are computed from the grant year of the citing patent to the grant year of the cited patent: we do not have the application year for patents granted prior to 1967, and hence could not compute the lags from application to application years. For the forward lags we do have the application year for both citing and cited patents (starting with the 1975 cited patents), and hence they are computed from application year to application year.

	Mean Backward Lag (in years)[20]	
Cohort	**by citations**	**by citing patents**
1975-77	15.22	14.30
1983-85	16.44	15.22
1989-91	15.96	14.52
1997-99	14.08	12.66

Thus, starting in the early 1980s the backward citation lag has shortened significantly (by over 2 years). As discussed further below, however, this trend could simply be due to the fact that the rate of patenting has accelerated since then, meaning that the "target" population to cite is, on average, younger than it used to be.

Turning now to forward citation lags, Figure 11 shows the frequency distribution of lags for patents from selected application years. An interesting feature of these distributions is that they are quite flat, particularly those for the earlier years. This is simply the result of the steep rise both in the number of citations ***made*** per patent and in the number of patents granted (and hence citing). Take the distribution for 1975 patents: after the first 3 – 4 years, and as time advances, these patents should have been getting fewer citations. In fact though, the number of citations that the 1975 patents received did not fall, because the number of citations made by later patents kept rising (and among others they were citing the 1975 patents), and the number of citing patents kept growing. These trends compensated for the fact that the 1975 were getting older and hence becoming less likely to be cited. Of course, as the distribution approaches the maximum lag possible (of 24 years for the 1975 patents), the number of citations has to fall because of truncation.

Another feature of interest is that it took over 10 years for the 1975 patents to receive 50% of their (forward) citations. Thus, even with truncation it is clear that the citation process is indeed a lengthy one, however one looks at it. It is therefore imperative

[20] The mean lag "by citation" is computed by taking the lag of each citation to be an observation and computing the mean for all of the citations; the mean lag "by citing patent" means that we first compute the

to take quite a wide time window in order to get significant coverage of forward citations. This does not imply that citation analysis has to be confined to old patents, but that one needs to carefully control for timing in using citations.

II.8 "Self" Citations

One of the interesting issues in this context is whose patents are cited, and in particular, to what extent they cite previous inventions patented by the same assignee (we refer to these as "self citations"), rather than patents of other, unrelated assignees. This has important implications, *inter alia*, for the study of spillovers: presumably citations to patents that belong to the same assignee represent transfers of knowledge that are mostly internalized, whereas citations to patents of "others" are closer to the pure notion of (diffused) spillovers.

We compute the percentage of self-citations made as follows: for each patent that has an assignee code we count the number of citations that it made to (previous) patents that have the same assignee code, and we divide the count by the total number of citations that it made.[21] This is in fact a lower bound, because the assignee code variable starts only in 1969, and hence for citations to patents granted earlier we cannot establish whether they are self-citations or not.[22] We also compute an upper bound, dividing the count of self-citations by the number of citations that have an assignee code, rather than by the total number of citations.[23]

The mean percentage of self-citations made is 11% for the lower bound, and 13.6% for the upper bound. However, there are wide differences across technological

mean lag for each citing patent, and then take the mean for all citing patents.

[21] We exclude from the computation citing patents that are unassigned (about 25% of patents), since by definition there is no "match" possible to any other assignee of the cited patents.

[22] There is a further reason for this to be a lower bound: the assignee code is not "consolidated", that is, the same firm may appear in different patent documents under various, slightly different names, one assignee may be a subsidiary of the other, etc. Thus, if for example we were to compute the percentage of self-citations using the Compustat CUSIPs (after the match) rather than the assignee codes, we would surely find higher figures.

[23] This is presumably an upper bound because we know that self-citations occur earlier on average than citations to unrelated assignees; given that patents with missing assignee codes are relatively old (i.e.

fields, as shown in Figure 12 (computed for the lower bound). The fact that the percentages are much higher in Chemical and in Drugs and Medical corresponds well with what we know about these fields: innovation is concentrated there in very large firms, and hence the likelihood that they will cite internally is higher.[24] Others and Mechanical are at the other extreme: in those fields innovation is much more widely spread among highly heterogeneous assignees (in terms of size, types of products, etc.), and hence self-citations are on average less likely.

Self-citations occur much more rapidly than citations to other patents: for the cohort of patents granted in 1997-99, the overall mean backward citation lag was of 14.1 years, and the median of 9 years. For self-citations the mean was of just 6.5 years, and the median 5 years. These differences are part of a more general phenomenon: citations to and from patents that are "closer" in terms of geography, technology, or institutional belonging occur earlier than citations to and from patents that are further removed along those dimensions (see Jaffe, Trajtenberg, and Henderson, 1993).

Figure 13 examines how the fraction of self-citations made has varied over time. There was a gradual increase over the decade of the 1970s. After 1980 there are some movements up and down but no clear trend. This may reflect some kind of increase in competition in invention in the last two decades, but that is pure conjecture at this point. More detailed examination of these variations in self-citation rates might provide valuable insights into the cumulative and competitive aspects of dynamic innovation.

Just as we have looked at the fraction of self-citations *made*; we can also examine the fraction of the citations *received* by a given patent that come from the same assignee. Self-citations received are, however, potentially distorted by the truncation of our data series, interacting with the phenomena that self-citations come sooner. That is, because they come sooner, self-citations are less affected by truncation than non-self-citations,

granted prior to 1969), citations to them would be less likely to be self-citations. However, the issue raised in the previous footnote still remains open, and hence this is not an upper bound in that sense.

[24] There is a huge difference between "Drugs" and "Medical" in this respect: the percentage of self-citations in Drugs is about 20%, that in the remaining D&M sub-categories just 8%.

causing the calculated percentage of self-citations received for recent cohorts to be biased upward. This is seen clearly in Figure 14, which is analogous to Figure 13 but calculated on the basis of percent of self-citations received. It shows the same slight upward trend in the 1970s, followed by a leveling off, and then a rapidly rising rate as we approach the truncation of the data in the 1990s.

II.9 Measures of "Generality" and "Originality"

A wide variety of citations-based measures can be defined and computed in order to examine different aspects of the patented innovations and their links to other innovations. We have computed and integrated into the data two such measures, "Generality" and "Originality," as suggested in Trajtenberg, Jaffe and Henderson, 1997:[25]

$$Generality_i = 1 - \sum_j^{n_i} s_{ij}^2 \ ,$$

where s_{ij} denotes the percentage of citations received by patent i that belong to patent class j, out of n_i patent classes (note that the sum is the Herfindahl concentration index). Thus, if a patent is cited by subsequent patents that belong to a wide range of fields the measure will be high, whereas if most citations are concentrated in a few fields it will be low (close to zero). Thinking of forward citations as indicative of the impact of a patent, a high generality score suggests that the patent presumably had a widespread impact, in that it influenced subsequent innovations in a variety of fields (hence the "generality" label). "Originality" is defined the same way, except that it refers to citations made. Thus, if a patent cites previous patents that belong to a narrow set of technologies the originality score will be low, whereas citing patents in a wide range of fields would render a high score.[26]

[25] Note that these measures depend of course upon the patent classification system: a finer classification would render higher measures, and conversely for a coarser system.

[26] As indicated earlier, we included in the data a variable indicating the % of citations made by each patent to patents granted since 1963, which in the present context means the percentage of cited patents that have a patent class. Since "originality" was computed on the basis of these patents only (rather than on the total number of citations made), this is an indicator of the extent to which the computation is accurate.

These measures tend to be positively correlated with the number of citations made (for originality) or received (for generality): highly cited patents will tend to have higher generality scores, and likewise patents that make lots of citations would display on average higher originality. In effect, where there are more citations, there is a built-in tendency to cover more patent classes. How one thinks about this tendency is to some extent a matter of interpretation. To some degree, the tendency of highly cited patents to also have a more general impact is presumably real. It can, however, lead to potentially misleading inferences, particularly when comparing patents or groups of patents that have different numbers of citations because they come from different cohorts and are therefore subject to differing degrees of truncation. If one views the observed distribution of citations across patent classes as a draw from an underlying multinomial distribution, then it can be shown that the observed concentration is biased upward (and hence the generality and originality measures are biased downward), due to the integer nature of the observed data. In effect, it is likely that many of the classes in which we observe zero citations do have some non-zero expected rate of citation. The resulting bias will be particularly large when the total number of citations is small. Appendix 2 (due to Bronwyn Hall) shows how to calculate the magnitude of the bias, and hence bias-adjusted measures, under fairly simple assumptions about the structure of the process.

Figure 15 shows the averages over time for both generality and originality. The steep decline in generality at the end of the period is almost surely due to truncation, which reduces the number of observed citations; the adjustment described in the previous paragraph mitigates but does not eliminate this decline. The decline remaining *after* adjustment may be due to the tendency of citations that are “nearer” in technology space to come sooner, so that even after adjusting for the number of citations, generality is biased downward when based only on “fast” citations.[27] Figures 16 and 17 present these measures over time by technological fields. The traditional fields Mechanical and Others are at the bottom in terms of generality, whereas Computers and Communications is at

[27] The slight decline in the mean originality during 1996-97 may also be due to truncation, in the sense that the number of citations made may be indicative of the “complexity” of the patent, and hence patents that are granted relatively fast probably make fewer citations; since originality is correlated with number of

the top, with Chemical and Electrical and Electronics in between. Surprisingly perhaps, Drugs and Medical is also at the bottom, both in terms of generality and of originality. However, a closer look reveals that the sub-category of Biotechnology stands much higher than the rest of D&M both in generality and originality, and hence that the aggregation in this case may be misleading in terms of these measures. Also somewhat surprisingly, Chemical (that we regard as a traditional field) stands high in both measures, being second to C&C in generality, and even higher than C&C in terms of originality.

The fact that Computers and Communications scores highest in terms of generality fits well the notion that this field may be playing the role of a "General Purpose Technology" (see Bresnahan and Trajtenberg, 1995), and its high originality score reinforces the view that it is breaking traditional molds even within the realm of innovation. Likewise, the low scores of Mechanical and Others correspond to expectations, in terms of the low innovativeness and restricted impact of those fields. In that sense, this constitutes a sort of "validation" of the measures themselves. At the same time, we should be aware of the fact that both originality and generality depend to a large extent upon the patent classification system, and hence there is an inherent element of arbitrariness in them. Thus, a "finer" classification within a field, in terms of number of 3-digit patent classes available, will likely result ceteris paribus in higher originality and generality measures, and one may justly regard that just as an artifact of the classification system (that may be the case for example with Chemicals). In terms of field averages, there is the further issue of degree of heterogeneity within fields, as for example with Drugs and Medical. Further exploration of these issues, and the possible role played by the calculation bias in them, is a fruitful area for future research.

II.10 Number of Claims

A further item in our data is "number of claims", as it appears in the front page of each patent. The claims specify in detail the "components", or building blocks of the patented invention, and hence their number may be indicative of the "scope" or "width"

citations made, and for those years we have only those patents that were granted relatively quickly (by application year), we would observe indeed a decline in originality for recent years.

of the invention (see for example Lanjouw and Schankerman, 1999). The average number of claims made has risen substantially over time, from 9.3 in 1974 to 14.7 in 1996. Figure 18 shows the averages by technological fields over time: traditional fields make fewer claims than advanced fields, with Chemical crossing from high to low in the 1990s. The differences are very substantial: the average for Computer and Communications (the top field) in 1995 was 16.8, the average for Others (the lowest) just 13.7.

II.11 Match to Compustat

In order to take full advantage of the wealth of information contained in patent data, one needs to be able to link patents to outside data of various sorts – otherwise the analysis would be self-contained, with all the limitations that implies. Thus for example the information on the location of inventors (state/city/counties for U. S. inventors, country/city for foreign ones) allows one to place each patent in geography space, and hence link out with location-specific data. Similarly for data items such as technological field, time, and institutional belonging.

One of the potentially most fruitful linkages is through the identity of the assignee: if one could relate each patent to the corporation that owns it, and bring together data about the corporations and about the patents, the scope of analysis would be greatly expanded. This is indeed what Zvi Griliches envisioned in setting up the NBER R&D, patents, and productivity project in the early 1980s (see Griliches, 1984). At that time though the only data item available about patents was patent counts by assignees, which were then attached to Compustat. Linking out our data allows one to use ***all*** the patent data fields, not just their count.

As already mentioned, about 80% of patents are assigned to non-government organizations, which are in fact mostly corporations, and our data contains both the name of the assignee, and an assignee code. The trouble is that there are about 150,000 such names, and their corresponding code is internal to the patent system, with no outside linkages. We undertook to match these assignee names to the names of corporations as they appear in Compustat, which comprises all firms traded in the U. S. stock market (about 36,000 of them). This was one of the most difficult and time-consuming tasks of the entire data construction project.

Figure 19 shows the percentage of patents matched, out of the total number of ***assigned*** patents. Not surprisingly, the percentage of foreign patents matched is very small, given that the overwhelming majority of foreign assignees are not traded in the U. S. stock market and hence do not appear in Compustat. For U. S. patents, though, the percentage is quite high up to the early 1980s, hovering around 70%. The steep decline from then on probably reflects both the fact that the match was done for the 1989 Compustat file, and the rapidly changing composition of patents. Indeed, and as mentioned above, the technological composition of patents has changed quite drastically since the mid-1980s, with traditional fields declining to less than 50% of all patents. It is quite likely that these changes have been accompanied by a large turnover in the composition of assignees, with many of the new entrants not yet traded by 1989, the year of the match.

III. "Benchmarking" of Citation Data

III.1 Overview

Although the previous section have demonstrated intriguing trends and contrasts visible in data on patent citations, it must be acknowledged that there is no natural scale or value measurement associated with citations data. Standing by itself, the fact that a given patent has received 10 or 100 citations does not tell you whether that patent is "highly" cited. Intrinsically, information on patent citations is meaningful ***only*** when used comparatively. That is, the evaluation of the patent intensity of an invention, an inventor, an institution, or any other group of patents, can only be made with reference to some "benchmark" citation intensity.

The determination of the appropriate benchmark is complicated by several phenomena that are inherent to the patent citations data. First, as already mentioned, the

[28] Bronwyn Hall was the main driving force behind the matching process, and it is only thanks to her monumental efforts that this task was accomplished.

number of citations received by any given patent is truncated in time because we only know about the citations received so far. More importantly, patents of different ages are subject to differing degrees of truncation. For example, it is not obvious whether a 1990 patent that received 5 citations by 1999 should be thought of as more or less highly cited than a 1985 patent that received 10 citations by 1999. Second, differences in Patent Office practices across time or across technological areas may produce differences in citation intensities that are unrelated to the true impact for which we use citations as a proxy. As shown above, the average patent issued in 1999 made over twice as many citations as the average patent issued in 1975 (10.7 versus 4.7 citations). At first blush, this would seem to imply something about the meaning or value of a given number of citations.

The problem created by the increase in the number of citations made *per patent* is exacerbated by the fact that the number of *patents* issued has also been rising steeply since 1983. Even if each patent issued made the same number of citations as before, the increase in the universe of citing patents would increase the total number of citations made. The combination of more patents making more citations suggests a kind of citation "inflation" that may mean that later citations are less significant than earlier ones. As a result, if we compare the citations received by a 1994 patent 5 years forward (i.e. up to 1999) with those received by a 1975 patent up to 1980, we cannot be sure that these totals are comparable. Thus even such "fixed-window" comparisons—which do not suffer from truncation bias—may be hard to make.

In addition to varying over time, the number of citations made per patent varies by technological field (See Figure 7). Thus, one might suspect that a given number of citations received from patents in Computers and Communications (which typically *make* fewer citations than those in other fields) is indicative of a larger impact than the same number of citations received from other fields. On the other hand, differences in citations ***received*** per patent (across time, fields, etc.) could be indicative of "real" differences in technological impact (see Figure 8).

The way in which we treat any of these systematic differences in citation intensities when developing appropriate benchmarks for analyzing citation data will depend on maintained hypotheses as to which of them are to be regarded as “real” and which as “artifacts.” For example, we might believe that the increase in the rate of patenting represents a real increase in the rate of invention, so that *its* contribution to changes in the number of citations received by patents is part of the real technological impact of the cited patents. At the same time, we might believe that the increase in the number of citations made *per patent* is a pure artifact of changes in patent examination practices, so that the best measure of “real” technological impact would be citation intensity “purged” of any differences due to the changing citation propensity. If so, we would want to control for the latter, but not make any adjustment for changes in the rate of patenting. Or we may be agnostic, and try to infer the nature of these effects by constructing citation-based impact measures with and without first purging the citations data of these effects, and then examining which measures are more highly correlated with non-patent indicators of technological or economic impact.

This discussion assumes implicitly that it is possible to identify and quantify the changes in citation intensity that are associated with the different effects. But this is actually harder than it may seem. Consider, for example, the increase in the average number of citations made per patent. It might seem that if each patent is making twice as many citations, that means each citation is “worth” half as much. But since the stock of patents available to be cited has been growing at a rapid (and accelerating) rate, this is not clear. Since there are so many potentially cited patents “competing” for the citations, you might think that getting one means as much as it did before, not withstanding the increase in the flow of citations.

To begin to think systematically about this set of issues, consider the following stylized facts that hold in our data: (*i*) the average number of citations received by patents in their first 5 years has been rising over time; (*ii*) the average number of citations *made* per patent has been rising over time (see Figure 6); and (*iii*) the observed citation-lag distributions for older cohorts have fatter “tails” than those of more recent cohorts (see

Figure 11). Considering the first fact in isolation, one might conclude ***either*** that more recent cohorts are more "fertile," ***or*** that the citation-lag distribution has shifted to the left (citations are coming sooner than they used to.) Considering the second fact in isolation, one might conclude that there has been an artifactual change in the propensity to make citations.[29] The last fact, taken by itself, seems to suggest that the citation-lag distribution has shifted to the right. Without further assumptions one cannot tell apart which of these competing scenarios is "correct", and hence one cannot make any statistical adjustments to the citations data, including adjustments for truncation of lifetime citations.

In this section we discuss two generic approaches to these problems. The first, which we call the *fixed-effects* approach, involves scaling citation counts by dividing them by the average citation count for a group of patents to which the patent of interest belongs.[30] This approach treats a patent that received say 11 citations and belongs to a group in which the average patent received 10 citations, as equivalent to a patent that received 22 citations, but happens to belong to a group in which the average was 20. Likewise, such a patent would be regarded as inferior to a patent receiving just 3 citations but for which the group average was only 1. The advantage of this approach is that it does not require one to make any assumptions about the underlying processes that may be driving differences in citation intensities across groups. The disadvantage is that, precisely because no structure is assumed, it does not distinguish between differences that are "real" and those that are likely to be artifactual.

The second or "quasi-structural" approach attempts to distinguish the multiple effects on citation rates via econometric estimation.[31] Once the different effects have thereby been quantified, the researcher has the option to adjust the raw citation counts to remove one or more of the estimated effects. If the assumptions inherent in the econometric estimation are correct, this approach permits the extraction of a stronger signal from the noisy citation data than the non-structural, fixed-effects approach.

[29] Another, more subtle interpretation could be that the rising propensity to cite is itself merely a reflection that more recent cohorts have been more fertile.

[30] Empirical analyses based on this approach include Henderson, Jaffe and Trajtenberg, 1998, and Jaffe and Lerner, 2001.

III.2 The "fixed-effects" approach

The fixed-effects approach assumes that *all* sources of systematic variation over time in citation intensities are artifacts that should be removed before comparing the citation intensity of patents from different cohorts. That is, we "re-scale" all citation intensities, and express them as ratios to the mean citation intensity for patents in the same cohort.[32] If we want to compare a 1990 patent with 2 citations to a 1985 patent with 4 citations, we divide each by the average number of citations received by other patents in the cohort. This rescaling purges the data of effects due to truncation, effects due to any systematic changes over time in the propensity to cite, and effects due to changes in the number of patents making citations. Unfortunately, it also purges the data of any systematic movements over time in the importance or impact of patent cohorts. It is possible that the *typical* 1985 patent has more citations than the *typical* 1990 patent (partly) because it is indeed more "fertile". Conversely, it could be that the 1990 patent is in fact "better" than the 1985 patent, once the effects of truncation are removed. Under the fixed-effects approach we do not attempt to separate "real" differences among cohorts from those due to truncation and propensity to cite effects, so any "real" effects that may be there are lost.

An issue arises as to how to treat technological fields in applying the fixed-effects correction. As with any fixed-effect approach, one can "take out" year effects, field effects, and/or year-field interaction effects.[33] As discussed above, there are systematic differences across fields in the frequency of citations made and received. If one believes that such effects are "real," then it is not appropriate to remove them when rescaling. To the extent that they are artifacts of, for example, the disciplinary training of patent

[31] An example of analysis based on this approach is Hall, Jaffe and Trajtenberg, 2001.

[32] Henderson, Jaffe and Trajtenberg examined the citation intensity of university patents by comparing it to the citation intensity of corporate patents from the same year. Since most patents are corporate patents, this is similar in effect to comparing the university patents to the overall mean.

[33] An obvious question to ask is why we propose to rescale the citations data rather than simply including the corresponding fixed-effect in whatever regression or other statistical analysis we are going to use the citations in. The reason is that such analyses typically have as a unit of observation entities that in any given year hold patents from many different cohorts. Hence the rescaling described here does not correspond to a simple fixed effects regression.

examiners in different fields, one may want to remove them. Further, the empirical lag distribution of citations vary by technological field, which means that the extent of truncation of a patent of given vintage depends on its technological field.

Tables 2 shows the average number of citations received by patents of each cohort (Table 2a according to application year and 2b by grant year) in each technological field, and the overall means. In order to remove all year, field and year-field effects, one can take the number of citations received by a given patent and divide by the corresponding year-field mean. Alternatively, to remove only pure year effects, one can divide by the yearly means (calculated without regard to field). Finally, one can envision the removal of year effects and year-field interaction effects but *not* the main field effect. This can be accomplished by dividing the entries in Table 2 by the overall mean for each technological category (bottom row). Each cell in the resulting matrix is then the year-field mean *relative to* the overall mean for the field. If actual citation counts are then divided by the appropriate entry from this adjusted matrix, overall differences in mean intensities across fields are *not* removed. This permits the correction for truncation to vary by field, while allowing the overall average differences in citation intensity by fields to remain in the rescaled data.[34]

To summarize, the fixed-effects rescaling aims to increase the signal-to-noise ratio in the data and allow comparability of citation counts over time by removing from the data variance components that are associated with truncation and also with possibly artifactual aspects of the citations generation process. Unfortunately, there is no way to do so without also removing variance components that might be real. The only way to tune this more finely is to put more structure on the problem, with a model that, under additional assumptions, allows separate identification of different sources of variation.

[34] Note that we have calculated these rescaling factors by the technological field of the cited patent. One could imagine constructing similar factors by technological field of the citing patent. Indeed, one might believe that variations by field of the citing patent are more likely to be pure artifacts than variations by field of the cited patent. As a practical matter, rescaling by the field of the citing patent is computationally much more difficult. The rescaling factors that we propose can be applied directly to the total citations

III.3 The “quasi-structural” approach

If the citation-lag distribution, the fertility of different patent cohorts, and the propensity to cite have all been varying over time, there is no general way to identify separately the contribution of each of these to variations in observed citation rates. The fixed-effects approach accepts this reality and simply removes variance components that are likely to be contaminated to some degree. To go any further one must impose additional structure, and in particular one must commit to some identifying assumptions. The assumptions that we make here are as follows:

- *Proportionality:* the shape of the lag distribution over time is independent of the total number of citations received, and hence more highly cited patents are more highly cited at all lags.

- *Stationarity:* the lag distribution does not change over time, i.e., does not depend on the cohort (application or grant year) of the cited patent.

These assumptions accomplish two objectives. First, stationarity means that we can estimate a time-invariant citation-lag distribution, which tells us the fraction of lifetime citations that are received during any specified time interval in the life of the patent. With proportionality, the observed citation total at a point in time for any patent can then be corrected for truncation, simply by “scaling up” the observed citation total by dividing it by the fraction of the lifetime citations that are predicted to occur during the lag interval that was actually observed. Second, these assumptions allow us to estimate changes in the propensity to cite over time in a way that controls for the citation lag distribution, as well as for changes in the “fertility” of the cited cohorts (at least to some extent). In principle, this allows a researcher who believes that the “citing year” effects are artifactual but “cited” year effects are real to remove the former but not the latter. In contrast, the fixed-effects approach implicitly takes out both.

Of course, we cannot know whether these identifying assumptions are really valid. As to proportionality, we found some (still weak) supporting evidence in the fact

received by a given patent. Rescaling factors tied to the field of the citing patent would have to be applied individually to each *citation* rather than simply to each cited patent.

that there is virtually zero correlation between the average forward citation lag per patent, and the number of citations received.[35] That is, the average citation lag for patents with few citations is virtually identically to the mean lag for patents that receive lots of citations. Stationarity is a more complex issue, since the observed citation-lag distribution could shift over time for different reasons, and without making *other* identifying assumptions it is difficult to test this in data while other things are also changing over time.

To implement this approach, let P_{ks} be total patents observed in technological field *k* in year *s*. Let C_{kst} be the total number of citations *to* patents in year *s* and technology field *k*, coming *from* patents in year *t*. The ratio C_{kst}/P_{ks} is then the average number of citations received by each *s-k* patent from the aggregate of all patents in year *t*. Consistent with our proportionality assumption, we model this citation frequency as a multiplicative function of cited year (*s*) effects, citing-year (*t*) effects, field (*k*) effects, and citation lag effects. Denoting the citation lag (*t-s*) as *L*, we can write this as:

$$C_{kst}/P_{ks}=\alpha_0{}'\ \alpha_s{}'\ \alpha_t{}'\ \alpha_k{}'\ exp[f_k(L)]$$

or, equivalently,

$$log[C_{kst}/P_{ks}]=\alpha_0+\alpha_t+\alpha_k+f_k(L)$$

where $\alpha_j=log(\alpha_j{}'),$ and $f_k(L)$ indicates some function, perhaps varying by technological field, that describes the shape of the citation-lag distribution. It could be a parametric function such as the double exponential used by Caballero and Jaffe (1993) and Jaffe and Trajtenberg (1999), or it could be different proportions estimated for each lag. We impose the constraint that the summation of $exp[f_k(L)]$ over *L* (*L*=1...35) is unity. We also normalize $\alpha_{t=1}=\alpha_{k=1}=0.$[36]

[35] The correlation coefficient is of 0.03 for all patents, and of 0.015 for patents with 5 citations or more.

[36] Note that since *L=t-s*, all of the α_s and α_t may not be identified, depending on the functional form of $f_k(L)$. We discuss this further below.

This equation can be estimated by OLS, at least for some forms of $f_k(L)$, or by non-linear methods, as in Jaffe and Trajtenberg (1999). The α parameters can be interpreted as the proportional difference in citation intensity for a given year or field relative to the base group. These parameters can therefore be used directly to adjust or normalize observed citations for these effects, if desired. The estimated $f_k(L)$ can be used to adjust patent totals for differential truncation across cohorts.

Implementation of this approach is illustrated in Table 3, which updates through 1999 estimates originally presented in Hall, Jaffe and Trajtenberg (2001). In this model, $f_k(L)$ is given by:

$$f_k(L)=exp(-\beta_{1k}L)(1-exp(-\beta_{2k}))$$

where the parameter β_{1k} captures the depreciation or obsolescence of knowledge and β_{2k} captures its diffusion.[37] Because this function is non-linear, it is possible to identify distinct α_s and α_t effects, at least in principle. In practice, we found that estimation was difficult with a full set of unconstrained cited year and citing year effects. Because we believe that the true "fertility" of invention changes only slowly, we grouped the cited years and estimated separate α_s coefficients for five-year intervals. The α_t effects are allowed to vary every year.

The estimates in the first column constrain the diffusion parameter β_2 to be the same across different fields k, while allowing the obsolescence parameter β_1 to vary. The second column reverses this, holding obsolescence constant but allowing diffusion to vary. The column labeled "Full Model" allows both of these parameters to vary across fields. Although allowing the βcoefficients to vary does not have a large effect on the overall fit, it does affect somewhat the estimated shape of the lag distributions; this can be seen in summary form in the variations in the simulated modal citation lags shown in the bottom of the table. We will focus herein on the Full Model results in the last column

[37] For a motivation of this parameterization, see Caballero and Jaffe, 1993.

The results show that the citing year effects are indeed significant.[38] After controlling for the effects of the lag distribution, the number of patents available to be cited, and cited year fertility, the number of citations made roughly *tripled* between 1975 and 1995. Note that this is the combination of effects due to the increased number of citing patents and the increased rate of citations made per patent. The part that is due to the increased rate of citations made per patent, because it has been purged of other effects, can be thought of as a measure of changes in the "pure" propensity to make citations. To focus on this, Table 4 takes the series of increasing citing year effects and decomposes it between the rise in the number of citing patents, and the pure propensity to cite effect. We see in Column 2 that the number of citing patents *by application year* peaks in 1995 in our data at about twice the number in 1975.[39] Column 3 is just the α_t coefficients from Table 3. Column 4 divides this series by the index of the number of potentially citing patents by application year (Column 2), thus removing the effect due to the rising number of citing patents. We find that the pure propensity to cite was also rising until 1995, accounting for about a 50% increase in citations made.

In looking at totals of citations *made*, one could similarly divide the number made by the entry in the table that corresponds to the application year of the patent(s) of interest.

It is interesting to compare this estimated "pure" propensity to cite effect with the "raw" change in the average number of citations made by each patent. The latter increased by about 100% between 1975 and 1995 (from about 5 to about 10). Our estimates say that roughly half of this increase was due to rising "pure" propensity to cite, and the other half was due to the fact that there were many more patents out there available to be cited.

[38] There is less variation in the cited year effects, and no clear pattern over time.

[39] This was already seen in Figure 1. To emphasize again, the decline in the application year numbers in the late 1990s is due to the truncation in the application-year series based on patents granted by the end of

After 1995, both the number of patents and the estimated overall citing year effect decline. Indeed, the citing year coefficients from the regression decline *faster* than the number of patents, causing the rise in the pure propensity to cite to reverse itself. Now, this latter effect *is not* due to truncation. It says that the patents issued in the late 1990s made fewer citations, after controlling for the size and fertility of the stock of patents available to be cited, than those before.[40] This finding is very consistent with the general notion that the patent office has been overwhelmed by the dramatic upsurge in patent applications in the last few years, with patent examiners having less time to review each application, and, therefore, being less thorough in finding prior art that should be cited.

The series presented in Columns 3 or 4 of Table 4 can be interpreted as "deflators" that can be used to purge citation totals of effects due to the rising tide of citations made. For a given patent or set of patents, one can divide the number of citations received from each application year by the appropriate entry in the table. Dividing counts of citations by the deflator in Column 3 removes ***all*** citing year effects. Dividing by Column 4 removes only the effect due to the changing propensity to cite, thus implicitly treating the effect due to the rising patenting rate as real. Either way, the resulting deflated totals of citations received can be interpreted as "real 1975 citations," in the same way that nominal dollar amounts divided by a base year 1975 price index are interpreted as real 1975 dollars.[41]

If one were interested in "deflating" the number of citations *made* by a given patent or set of patents, one does not need to worry about effects due to the rising number of patents. But one might be interested in removing the pure propensity to cite effect. This could be accomplished by dividing the number of observed citations made by the entry in Column 4 corresponding to the application year of the patent(s) of interest.

1999. Once the rest of the applications from the late 1990s are processed, we will no doubt see a continued increase in successful applications per year.

[40] This effect is even visible in the raw averages of citations made per patent, which also turn downward in the late 1990s after the earlier increases already noted.

Analogous "deflators" derived from Table 3 can be used across technology fields, if one believes that the average difference in citation rates across technology fields is an artifact of field practices rather than a real difference across fields in knowledge flows. One can simply "deflate" citation totals by dividing by the α_k coefficients for the different fields. This would have the interpretation of converting citation totals into equivalent numbers of citations for the "Other" technology field (the base group, whose α-coefficient is normalized to unity). We have not employed such adjustments in our work, because we believe that field effects are likely to contain a significant real component. But this is a topic for further research. If field effects are real, then deflating citation totals by field effects ought to *decrease* the signal to noise ratio in citations data, implying that the correlation of citations with other indicators of technology impact (e.g. market value) ought to be reduced by deflation. If the opposite is true, it would suggest that much of the variance in the citation intensity across fields is artifactual.

The estimates in Table 3 can also be used to correct the citation totals of any given patent for truncation. As shown in Figure 20, the estimates for β_1 and β_2 can be used to construct the citation-lag distribution by field (normalized to unity over 35 years), after removing cited and citing year effects. The contrast between Figure 20 and Figure 11 illustrates the dramatic impact of the citing and cited year effects on the shape of the citation-lag distributions. The variations across field are also quite apparent. Citations in Computers and Communications come the fastest, followed by Electronics. Drugs and Medical and "Other" are the slowest, with Chemicals and Mechanical falling in the middle. This has some effect on corrections for truncation. The estimates imply, for example, that if we have citation data truncated at 5 years after the initial application, we are seeing about 33% of the "lifetime" (actually, of the first 35 years) citation total for an average C&C patent, but only 22% of the "lifetime" citations for a Drug and Medical patent.

[41] Because is "purged" of truncation effects, the deflator in Column 4 applies (in principle) to citation totals no matter how derived. Column 3, however, reflects the truncation by application year in our data, and so is appropriate only for citation totals derived from within this dataset.

The yearly fractions underlying Figure 20 are presented in cumulative form in Table 5. These can be used directly to adjust citation totals, based on the observed interval, whether the truncated or unobserved portion is at the end, at the beginning (cited patents applied for before 1975), or both. For example, for a patent applied for in 1973, we observe only years 2 through 25 of the citation lag distribution (1975-1999). If this were a Chemical patent, we see from Table 5 that for the typical Chemical patent, 87% of the estimated or predicted "lifetime" citations occur in this interval (.906 - .037), so we would divide the observed total by 0.87 to yield the truncation-adjusted total.

Finally, under the proportionality assumption that we have made, all corrections or adjustments are purely multiplicative. This makes it possible, in principle, to correct or adjust for any combination of effects. If, for example, one wants totals corrected for pure propensity-to-cite effects and for truncation, one would divide the number of citations received from each year by column 4 of Table 4, and then take the resulting total for each patent and normalize using Table 5. If one also wanted to remove technology field effects, one could then divide by the estimated α_k reported in the last column of Table 3. Of course, none of these adjustments should be taken as gospel or applied mechanically; we present them to illustrate the approach and encourage further research on the best ways to maximize the signal-to-noise ratio in these data.

IV. Conclusion

It has been a major theme of the NBER since its inception that good economic research depends on the generation of appropriate and reliable economic data. It is generally agreed that the 21st century economy is one in which knowledge—particularly the technological knowledge that forms the foundation for industrial innovation—is an extremely important economic commodity. The inherently abstract nature of knowledge makes this a significant measurement challenge. We believe that patents and patent citation data offer tremendous potential for giving empirical content to theorizing about the role of knowledge in the modern economy. We hope that by constructing the NBER

Patent Citations Data File, demonstrating some of the uses to which it can be put, and making it available to other researchers, we can provide a broader and deeper measurement base on which to build the economics of technological change.

References

Bresnahan, T. and M. Trajtenberg, "General Purpose Technologies - Engines of Growth?," Journal of Econometrics, January 1995, 65(1), pp. 83-108.

Caballero, R. and A. Jaffe, "How High are the Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth," in O. Blanchard and S. Fischer, eds., National Bureau of Economic Research Macroeconomics Annual, Vol. 8, MIT Press, 1993

Griliches, Zvi, "Patent Statistics as Economic Indicators," Journal of Economic Literature 92: 630-653, 1990.

Griliches, Zvi, (ed.) R&D, Patents, and Productivity, NBER Conference Proceedings. University of Chicago Press, 1984

Griliches, Z., Hall, B.H. and A. Pakes, "The Value of Patents as Indicators of Inventive Activity," in P. Dasgupta and P. Stoneman, eds., Economic Policy and Technological Performance, Cambridge, England: Cambridge University Press, 1987, 97-124.

Hall, B.H., Jaffe, A. and M. Trajtenberg, "Market Value and Patent Citations: A First Look," University of California, Berkeley, Dept. of Economics Working Paper No. XXXX, August 2001.

Henderson, R., A. Jaffe and M. Trajtenberg, "Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965-1988," Review of Economics and Statistics, 80: 119-127, 1998.

Jaffe, A., and J. Lerner, "Reinventing Public R&D: Patent Policy and the Commercialization of National Laboratory Technologies," Rand Journal of Economics 32, No. 1, pp 167-198 (Spring 2001)

Jaffe, A., Trajtenberg, M. and M. Fogarty, "Knowledge Spillovers and Patent Citations: Evidence from A Survey of Inventors". American Economic Review, Papers and Proceedings, May 2000, pp. 215-218.

Jaffe, A., Trajtenberg, M. and R. Henderson, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," Quarterly Journal of Economics, August 1993, pp. 577-598.

Jaffe, A., and M. Trajtenberg, "International Knowledge Flows: Evidence from Patent Citations," Economics of Innovation and New Technology, 1999, 8, pp. 105-136.

Kortum, S. and J. Lerner, "Stronger Protection or Technological Revolution: What is Behind the Recent Surge in Patenting?" Carnegie-Rochester Conference Series on Public Policy, 48 (June 1998) 247-304. Abridged version reprinted as “What is Behind the Recent Surge in Patenting?” in Research Policy 28 (January 1999) 1-22.

Lanjouw, Jean O. and Mark Schankerman, “The Quality of Ideas: Measuring Innovation with Multiple Indicators,” National Bureau of Economic Research Working Paper No. 7345, 1999.

Pakes, Ariel and Margaret Simpson, “The Analysis of Patent Renewal Data,” Brookings Papers on Economic Activity, Microeconomic Annual, 1991, pp. 331- 401.

Schankerman, M. and A. Pakes, “Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period,” Economic Journal, Vol. 96, No. 384, December 1986, pp. 1052-1077.

Scherer, F.M. “Inter-Industry Technology Flows and Productivity Growth,” Review of Economics and Statistics, 64, November 1982.

Schmookler, J. Invention and Economic Growth. Cambridge: Harvard University Press, 1966.

Trajtenberg, M. "A Penny for Your Quotes: Patent Citations and the Value of Innovations," The Rand Journal of Economics, Spring 1990, 21(1), 172-187.

Trajtenberg, M., Jaffe, A. and R. Henderson, "University versus Corporate Patents: A Window on the Basicness of Invention," Economics of Innovation and New Technology, 1997, 5 (1), pp. 19-50.

Appendix 1

Classification of Patent Classes into Technological Categories and Sub-Categories[42]

Cat. Code	Category Name	Sub-Cat. Code	Sub-Category Name	Patent Classes
1	Chemical	11	Agriculture, Food, Textiles	8, 19, 71, 127, 442, 504
		12	Coating	106,118, 401, 427
		13	Gas	48, 55, 95, 96
		14	Organic Compounds	534, 536, 540, 544, 546, 548, 549, 552, 554, 556, 558, 560, 562, 564, 568, 570
		15	Resins	520, 521, 522, 523, 524, 525, 526, 527, 528, 530
		19	Miscellaneous-chemical	23, 34, 44, 102, 117, 149, 156, 159, 162, 196, 201, 202, 203, 204, 205, 208, 210, 216, 222, 252, 260, 261, 349, 366, 416, 422, 423, 430, 436, 494, 501, 502, 510, 512, 516, 518, 585, 588
2	Computers & Communications	21	Communications	178, 333, 340, 342, 343, 358, 367, 370, 375, 379, 385, 455
		22	Computer Hardware & Software	341, 380, 382, 395, 700, 701, 702, 704, 705, 706, 707, 708, 709, 710, 712, 713, 714
		23	Computer Peripherals	345, 347
		24	Information Storage	360, 365, 369, 711
3	Drugs & Medical	31	Drugs	424, 514
		32	Surgery & Medical Instruments	128, 600, 601, 602, 604, 606, 607
		33	Biotechnology	435, 800
		39	Miscellaneous-Drug&Med	351, 433, 623
4	Electrical & Electronic	41	Electrical Devices	174, 200, 327, 329, 330, 331, 332, 334, 335, 336, 337, 338, 392, 439
		42	Electrical Lighting	313, 314, 315, 362, 372, 445
		43	Measuring & Testing	73, 324, 356, 374
		44	Nuclear & X-rays	250, 376, 378
		45	Power Systems	60, 136, 290, 310, 318, 320, 322, 323, 361, 363, 388, 429
		46	Semiconductor Devices	257, 326, 438, 505
		49	Miscellaneous-Elec.	191, 218, 219, 307, 346, 348, 377, 381, 386

[42] Based on the Patent Classification System as of 12/31/1999. The list of patent classes as of that date includes 8 additional new classes that are not to be found in the data: 532, 901, 902, 930, 968, 976, 984, and 987.

Appendix 1 - Continued

Cat. Code	Category Name	Sub-Cat. Code	Sub-Category Name	Patent Classes
5	Mechanical	51	Materials Processing. & Handling	65, 82, 83, 125, 141, 142, 144, 173, 209, 221, 225, 226, 234, 241, 242, 264, 271, 407, 408, 409, 414, 425, 451, 493
		52	Metal Working	29, 72, 75, 76, 140, 147, 148, 163, 164, 228, 266, 270, 413, 419, 420
		53	Motors, Engines & Parts	91, 92, 123, 185, 188, 192, 251, 303, 415, 417, 418, 464, 474, 475, 476, 477
		54	Optics	352, 353, 355, 359, 396, 399
		55	Transportation	104, 105, 114, 152, 180, 187, 213, 238, 244, 246, 258, 280, 293, 295, 296, 298, 301, 305, 410, 440
		59	Miscellaneous-Mechanical	7, 16, 42, 49, 51, 74, 81, 86, 89, 100, 124, 157, 184, 193, 194, 198, 212, 227, 235, 239, 254, 267, 291, 294, 384, 400, 402, 406, 411, 453, 454, 470, 482, 483, 492, 508
6	Others	61	Agriculture, Husbandry, Food	43, 47, 56, 99, 111, 119, 131, 426, 449, 452, 460
		62	Amusement Devices	273, 446, 463, 472, 473
		63	Apparel & Textile	2, 12, 24, 26, 28, 36, 38, 57, 66, 68, 69, 79, 87, 112, 139, 223, 450
		64	Earth Working & Wells	37, 166, 171, 172, 175, 299, 405, 507
		65	Furniture, House Fixtures	4, 5, 30, 70, 132, 182, 211, 256, 297, 312
		66	Heating	110, 122, 126, 165, 237, 373, 431, 432
		67	Pipes & Joints	138, 277, 285, 403
		68	Receptacles	53, 206, 215, 217, 220, 224, 229, 232, 383
		69	Miscellaneous-Others	1, 14, 15, 27, 33, 40, 52, 54, 59, 62, 63, 84, 101, 108, 109, 116, 134, 135, 137, 150, 160, 168, 169, 177, 181, 186, 190, 199, 231, 236, 245, 248, 249, 269, 276, 278, 279, 281, 283, 289, 292, 300, 368, 404, 412, 428, 434, 441, 462, 503

Appendix 2
A Note on the Bias in Herfindahl-type Measures Based on Count Data
by Bronwyn H. Hall

1. Introduction

Measures based on citations obtained by patents in individual patent classes or held by individual firms often suffer from bias due to the count nature of the underlying data. The source of the bias is the fact that cells with small numbers of expected citations have a non-zero probability that no citations will actually be observed. When this happens, the cell is removed from the analysis, implying that measures of diversification will be biased downward and measures of concentration will be biased upwards. In the cases considered in the text, patent generality or originality measures take the form of diversification measures and will therefore be biased downward when the total number of citations to or from the patent are small. If the bias is not corrected for, patents with few forward or backward citations will be more likely to be considered less "general" or "original" than those with many.

This appendix suggests a method for correcting the bias that is valid under a set of simple but fairly general assumptions. The two key assumptions are the following:

1. Either we treat the total number of citations (or patents) on which the measure is based as given (that is, we condition on them) or the number is large enough relative to the individual cell counts so that it can be treated as non-random.
2. The probability that a given citation or patent falls in a cell is independent of the probability that it falls in another cell. That is, there is no causal connection between the deviation of the observed outcome from the expected outcome in a particular cell and what happens in another cell (other than the adding up constraint). We can therefore describe the probability distribution over a set of cells of multinomial probabilities.

Given these assumptions, we are able to compute a simple correction for the bias that depends only on the total number of counts in the measure. This correction is large when the number of counts is small and quickly converges to zero as the number of counts increases.

Mathematically, the statement of the problem is the following: Suppose a researcher uses a Herfindahl-type measure to describe the concentration of patents or cites across patent classes, patent holders, or some other set. Here we use patents as an example, but all the same arguments apply to citation counts. For a set of N patents falling into J classes, with N_j patents in each class *($N_j \geq 0, j=1,\ldots,J$)*, the sample Herfindahl index *(HHI)* is defined by the following expression:

$$HHI = \sum_{j=1}^{J} (\frac{N_j}{N})^2$$

However, the population Herfindahl is given by

$$\eta = \sum_{j=1}^{J} \lambda_j^2$$

where the λ_js are the multinomial probabilities that the N patents will be classified in each of the J classes. Under reasonable assumptions,

$$E\left[\frac{N_j}{N}\right] = \lambda_j$$

Unfortunately, this does NOT imply that

$$E[HHI] = \eta$$

because of nonlinearity. In fact, in general the measured *HHI* will be biased upward when *N* is small, due to Jensen's inequality and the properties of the count distribution.

2. Computing and adjusting for the bias

Assume a multinomial distribution with parameters (λ_j, $j=1,\ldots,J$) for the $\{N_j\}$; then the expectation for each N_j^2 is the following (Johnson and Kotz, Discrete Distributions):

$$E\left[N_j^2\right] = N\lambda_j + N(N-1)\lambda_j^2$$

Conditional on the total number of patents, *N*, this implies the following relation between the estimated and true Herfindahl measure:[43]

$$E[HHI \mid N] = E\left[\sum_{j=1}^{J}\left(\frac{N_j}{N}\right)^2\right] = \sum_{j=1}^{J}\frac{E(N_j^2)}{N^2} = \sum_{j=1}^{J}\frac{N\lambda_j + N(N-1)\lambda_j^2}{N^2} = \frac{1}{N} + \frac{N-1}{N}\sum_{j=1}^{J}\lambda_j^2 =$$
$$= \frac{1}{N} + \frac{N-1}{N}\eta$$

Note that as $N \uparrow \infty, E[HHI \mid N] \rightarrow \eta$, as we would expect. The bias in this estimator is

[43] Conditioning on *N* is innocuous unless the process that generates the total number of draws (patents or citations) is related to the particular set of multinomial parameters with which we are working. For example, the procedure outlined here may not be precisely valid if "general" patents (patents whose cites are widely distributed across patent classes) are also highly cited patents. I am grateful to Tom Rothenberg for a discussion of this point.

$$E[HHI \mid N] - \eta = \frac{1-\eta}{N}$$

The bias declines at a rate N as the number of counts grows and as concentration increases. Both results are intuitive.[44]

Under the assumptions given in the introduction, it is straightforward to correct for this bias. Consider the following estimator for the Herfindahl:

$$\hat{\eta} = \frac{N \cdot HHI - 1}{N - 1}$$

For a given *N*, and under the assumption that the underlying process is multinomial with parameters λ_j, $j=1,\ldots,J$, this estimator is an unbiased estimator of η:

$$E[\hat{\eta} \mid N] = \frac{N \cdot E[HHI \mid N] - 1}{N-1} = \frac{1 + (N-1)\eta - 1}{N-1} = \eta$$

3. The generality index

For many problems, the measure used is one minus the Herfindahl rather than the Herfindahl itself. In particular, we define “generality” as

$$G_i = 1 - \sum_{j=1}^{J} \left(\frac{N_{ij}}{N_i} \right)^2$$

where N_i denotes the number of forward citations to a patent, and N_{ij} is the number received from patents in class *j,* and use a similar formula to measure “originality” based on the distributions of citations made. Patents with a high value of G_i are cited across a broad range of patent classes.

This measure is also a biased estimate of the true measure γ_i=1- η_i:

[44] It is also true that standard error estimates obtained in the conventional way will be biased, but it is also possible to compute the exact relationship between the standard error estimated from biased measures and that estimated for the unbiased measures. The standard error of the estimated mean of the Herfindahl will be biased downward by *(N-1)/N.* This is large if *N* is small and does not depend on the estimated Herfindahl. An unbiased estimator for the variance of the mean Herfindahl over a set of *M* observations is the following:

$$Var(\bar{\hat{\eta}}) = \frac{1}{M} \sum_{k=1}^{M} \frac{N_k^2 \cdot Var(HHI_k)}{(N_k - 1)^2}$$

where HHI_k is the *k*th biased estimate of the Herfindahl. Of course, if one uses the unbiased estimator to form the mean, one does not need to perform this correction in addition.

$$E[G_i \mid N_i] = 1 - E\left[\sum_{j=1}^{J} \left(\frac{N_{ij}}{N_i} \right)^2 N_i \right] = 1 - \frac{1+(N_i-1)\eta_i}{N_i} = \frac{N_i - 1}{N_i}\gamma_i$$

The bias is the following:

$$E[G_i \mid N_i] - \gamma_i = -\frac{\gamma_i}{N_i}$$

Again, the absolute size of the bias declines as the sample size increases and as generality decreases. The generality index will be biased downward in general and this effect is larger for small *N*. Appendix Figure 1 plots the bias versus the index for three values of *N* (3, 10, and 100). Clearly the magnitude is largest when *N* is small or generality is high.

Once again, one can form an unbiased estimator of γ_i:

$$\hat{\gamma}_i = \frac{N_i}{N_i - 1} G_i$$

The same arguments as the previous apply to standard error estimates of the generality index. The true standard errors will be *N/(N-1)* larger than the estimated standard errors. When the number of cites to a patent is small, generality will be underestimated and it will be more likely that significant differences among generalities of different patents will be found. But as we have indicated, correcting for the bias is straightforward.

4. Additional Reference

Johnson, Norman L., and Samuel Kotz. 1969. *Discrete Distributions.* New York: John Wiley and Sons.

Table 1

Application-Grant Lag Distribution by 3-Year Sub-periods

	Application Years							
	1967-69	**1970-72**	**1973-75**	**1976-79**	**1980-82**	**1983-85**	**1986-89**	**1990-92**
Lag (years)	**(*i*) Distribution of Lags (in %)**							
0	0.4	0.1	1.0	1.0	0.2	1.0	1.8	2.6
1	11.3	19.9	40.1	32.5	18.0	26.6	40.4	40.4
2	48.7	59.8	48.2	51.0	51.1	49.4	43.6	42.0
3	32.0	16.2	8.0	11.9	24.1	16.7	10.6	11.1
4	5.6	2.4	1.5	2.0	4.0	3.7	2.5	2.3
5	1.0	0.9	0.6	0.8	1.2	1.5	0.7	0.7
6	0.4	0.3	0.2	0.4	0.7	0.7	0.2	0.4
7+	0.5	0.3	0.3	0.3	0.7	0.4	0.2	0.4
Total	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*
(*ii*) Mean and Standard Deviation of the Lag, in Years								
Mean	2.39	2.08	1.74	1.88	2.25	2.05	1.76	1.76
s.d.	1.02	0.93	0.91	0.93	1.02	1.02	0.90	0.95

Table 2a
Citations Received by *Application* Year and Technological Category

App. Year	Chemical	Computers & Comm.	Drugs & Medical	Elect. & Electronics	Mechanical	Others	All
1967	4.35	4.30	7.27	3.80	3.87	4.52	4.24
1968	4.72	4.68	7.70	4.20	4.17	4.76	4.57
1969	5.00	5.80	7.84	4.66	4.50	5.03	4.94
1970	5.55	6.56	8.37	5.40	4.85	5.44	5.46
1971	5.90	7.55	8.82	5.82	5.22	5.78	5.87
1972	6.10	8.02	9.77	6.29	5.58	6.02	6.22
1973	6.39	8.70	9.64	6.73	5.70	6.26	6.50
1974	6.42	9.34	10.58	6.80	5.91	6.54	6.74
1975	6.58	10.06	8.90	7.12	5.90	6.61	6.85
1976	6.71	10.40	9.32	7.20	5.94	6.52	6.93
1977	6.69	10.63	9.26	7.30	5.79	6.39	6.89
1978	6.57	10.62	9.20	7.11	5.80	6.28	6.82
1979	6.59	10.96	9.63	7.32	5.79	6.23	6.92
1980	6.70	11.55	9.75	7.31	5.84	6.10	7.04
1981	6.62	12.06	9.99	7.15	5.80	6.10	7.10
1982	6.49	11.77	10.04	7.22	5.82	6.18	7.11
1983	6.77	11.96	10.30	7.40	5.70	6.17	7.24
1984	6.66	12.21	10.13	7.40	5.80	6.21	7.25
1985	6.56	11.82	10.64	7.15	5.74	6.12	7.19
1986	6.32	12.01	10.44	7.23	5.66	5.99	7.14
1987	6.05	11.42	9.95	6.94	5.34	5.65	6.84
1988	5.44	11.06	9.10	6.69	5.09	5.21	6.45
1989	4.93	10.63	8.26	6.24	4.78	4.80	6.03
1990	4.39	9.75	7.59	5.73	4.38	4.35	5.53
1991	3.82	8.29	6.79	5.40	3.89	3.91	4.98
1992	3.34	7.04	5.45	4.55	3.38	3.31	4.26
1993	2.59	5.62	3.77	3.66	2.67	2.62	3.35
1994	1.71	3.96	2.18	2.60	1.87	1.81	2.34
1995	0.93	2.08	0.95	1.49	1.10	1.03	1.28
1996	0.40	0.75	0.41	0.60	0.49	0.44	0.53
1997	0.11	0.18	0.10	0.16	0.14	0.13	0.14
All	4.62	6.44	5.99	4.75	4.17	4.46	

Table 2b
Citations Received by *Grant* Year and Technological Category

Grant Year	Chemical	Computers & Comm.	Drugs & Medical	Elec. & Electronics	Mechanical	Others	All
1963	2.86	1.98	4.89	2.21	2.77	3.36	2.90
1964	3.08	1.99	5.35	2.30	2.93	3.43	3.01
1965	3.47	2.20	5.75	2.44	3.08	3.67	3.20
1966	3.63	2.47	5.21	2.72	3.24	3.90	3.40
1967	3.71	2.92	6.40	2.89	3.39	4.07	3.61
1968	3.85	3.25	6.57	3.24	3.62	4.23	3.82
1969	4.11	3.19	6.95	3.51	3.78	4.42	4.02
1970	4.41	3.93	7.72	3.82	4.07	4.73	4.35
1971	4.85	5.20	8.71	4.59	4.41	4.93	4.83
1972	5.41	6.74	8.03	5.42	4.85	5.45	5.45
1973	5.81	7.27	8.56	5.89	5.20	5.73	5.82
1974	5.92	8.03	9.27	6.40	5.51	6.06	6.16
1975	6.17	8.65	10.20	6.78	5.80	6.40	6.54
1976	6.44	9.25	9.59	6.82	5.97	6.58	6.73
1977	6.57	10.10	9.10	7.23	5.95	6.73	6.92
1978	6.75	10.64	8.56	7.27	5.87	6.57	6.91
1979	6.76	10.11	9.27	7.32	5.90	6.42	6.92
1980	6.46	10.62	9.30	7.17	5.75	6.24	6.81
1981	6.77	10.86	9.15	7.28	5.85	6.22	6.90
1982	6.63	11.28	10.02	7.21	5.91	6.26	7.05
1983	6.72	11.56	10.14	7.26	5.96	6.24	7.10
1984	6.72	12.66	10.14	7.24	5.70	6.13	7.08
1985	6.72	11.91	10.09	7.40	5.71	6.18	7.11
1986	6.67	11.75	10.91	7.27	5.80	6.07	7.17
1987	6.59	12.07	11.46	7.38	5.80	6.08	7.33
1988	6.27	11.81	10.40	7.12	5.63	6.00	7.09
1989	5.82	11.18	9.69	6.79	5.20	5.37	6.67
1990	5.33	11.18	9.20	6.63	4.97	4.97	6.34
1991	4.84	10.26	8.64	6.14	4.58	4.66	5.87
1992	4.43	10.06	7.83	5.69	4.24	4.23	5.48
1993	3.73	9.17	6.52	5.23	3.72	3.69	4.90
1994	3.17	7.92	5.47	4.37	3.13	3.08	4.22
1995	2.37	6.05	3.85	3.50	2.50	2.40	3.30
1996	1.61	4.43	2.40	2.47	1.74	1.63	2.34
1997	0.85	2.45	1.09	1.40	0.99	0.90	1.28
1998	0.32	0.87	0.33	0.51	0.39	0.34	0.48
1999	0.03	0.06	0.02	0.05	0.03	0.03	0.04
All	4.62	6.44	5.99	4.75	4.17	4.46	

Table 3: Estimation of Citation Probabilities

	Constant Diffusion		Constant Obsolescence		Full Model	
	Coef.	**S.E.**	**Coef.**	**S.E.**	**Coef.**	**S.E.**
Tech Field Effects (base=other)						
Chemicals exc. Drugs	1.004	0.026	0.867	0.020	0.526	0.030
Computers & Comm.	2.281	0.058	1.451	0.033	1.495	0.094
Drugs & Medical	1.295	0.035	1.818	0.051	0.724	0.042
Electrical & Electronics	1.374	0.035	0.896	0.021	0.678	0.038
Mechanical	0.937	0.026	0.742	0.019	0.444	0.025
Citing Year Effects (base=1975)						
1976	0.742	0.036	0.812	0.038	0.871	0.040
1977	0.764	0.037	0.828	0.038	0.878	0.039
1978	0.839	0.041	0.900	0.041	0.943	0.041
1979	0.905	0.044	0.962	0.043	0.997	0.042
1980	0.956	0.045	1.008	0.044	1.034	0.041
1981	0.967	0.048	1.010	0.047	1.026	0.043
1982	1.022	0.052	1.059	0.050	1.064	0.045
1983	1.010	0.055	1.037	0.051	1.030	0.045
1984	1.110	0.061	1.130	0.056	1.111	0.048
1985	1.230	0.070	1.243	0.063	1.209	0.053
1986	1.360	0.080	1.362	0.071	1.312	0.059
1987	1.545	0.094	1.530	0.083	1.459	0.069
1988	1.728	0.111	1.692	0.097	1.600	0.079
1989	1.855	0.123	1.800	0.106	1.684	0.085
1990	1.931	0.132	1.856	0.112	1.724	0.090
1991	2.018	0.143	1.919	0.120	1.769	0.096
1992	2.256	0.165	2.119	0.137	1.940	0.109
1993	2.551	0.195	2.365	0.159	2.151	0.127
1994	3.053	0.241	2.799	0.197	2.529	0.155
1995	3.947	0.321	3.583	0.261	3.218	0.205
1996	3.382	0.284	3.033	0.227	2.709	0.180
1997	2.816	0.246	2.495	0.193	2.217	0.152
1998	0.701	0.069	0.612	0.054	0.542	0.044
1999	0.030	0.003	0.026	0.002	0.023	0.002
Cited Year Effects (base=1963-64)						
1965-69	0.635	0.018	0.710	0.022	0.814	0.021
1970-74	0.637	0.018	0.741	0.022	0.886	0.029
1975-79	0.602	0.022	0.724	0.027	0.911	0.038

1980-84	0.555	0.027	0.700	0.033	0.926	0.049
1985-89	0.511	0.032	0.686	0.040	0.937	0.062
1990-94	0.433	0.033	0.624	0.046	0.866	0.068
1995-99	0.287	0.029	0.434	0.041	0.604	0.063
Beta1: Obsolescence by Technology Field						
Chemicals exc. Drugs	1.007	0.020			0.689	0.025
Computers & Comm.	1.297	0.026			1.099	0.034
Drugs & Medical	0.760	0.018			0.503	0.024
Electrical & Electronics	1.235	0.025			0.850	0.027
Mechanical	1.040	0.022			0.653	0.025
Beta1 (Base=Other)	0.102	0.003	0.104	0.004	0.111	0.003
Beta2: Diffusion by Technology Field						
Chemicals exc. Drugs			1.639	0.105	3.404	0.362
Computers & Comm.			2.358	0.156	2.200	0.203
Drugs & Medical			0.783	0.048	2.919	0.287
Electrical & Electronics			2.615	0.188	3.815	0.390
Mechanical			2.091	0.144	4.572	0.527
Beta2 (Base=Other)	0.436	0.016	0.225	0.012	0.162	0.011
R-squared	0.950		0.941		0.956	
Standard error of regression	0.0595		0.0653		0.0561	
Simulated Modal Lag						
Chemicals exc. Drugs	3.81		4.10		3.82	
Computers & Comm.	3.35		3.41		3.83	
Drugs & Medical	4.34		5.62		4.75	
Electrical & Electronics	3.43		3.22		3.27	
Mechanical	3.74		3.63		3.26	
Other	3.82		5.11		5.55	

Note: The dependent variable is citations (by citing year, cited year, cited field) divided by potentially citable patents (by cited year and cited field). Cited years run from 1963-99 and citing years from 1975-99, for a total of 3,600 observations [6*(12*25+(24*25)/2)].

Table 4
Potential "Deflators" for Citing Patent Totals

Application Year	(1) Total Patents	(2) Index of Patent Total (1975=1)	(3) Citing Year Coefficient (from Table 3)	(4) Pure Propensity to Cite Effect [(3)/(2)]
1975	65888	1.000	1.000	1.000
1976	65804	0.999	0.871	0.872
1977	65978	1.001	0.878	0.877
1978	65601	0.996	0.943	0.947
1979	65726	0.998	0.997	0.999
1980	66491	1.009	1.034	1.025
1981	63910	0.970	1.026	1.058
1982	65009	0.987	1.064	1.078
1983	61563	0.934	1.030	1.103
1984	67071	1.018	1.111	1.091
1985	71442	1.084	1.209	1.115
1986	75088	1.140	1.312	1.151
1987	81458	1.236	1.459	1.180
1988	90134	1.368	1.600	1.170
1989	96077	1.458	1.684	1.155
1990	99254	1.506	1.724	1.145
1991	100016	1.518	1.769	1.165
1992	103307	1.568	1.940	1.237
1993	106848	1.622	2.151	1.326
1994	120380	1.827	2.529	1.384
1995	137661	2.089	3.218	1.540
1996	131450	1.995	2.709	1.358
1997	114881	1.744	2.217	1.271
1998	33780	0.513	0.542	1.057

Table 5
Simulated Cumulative Lag Distributions by Technology Field

Lag	Chem. exc. Drugs	Comp & Comm	Drugs & Medical	Electrical & Electronic	Mechanical	Other
1	0.037	0.045	0.026	0.048	0.043	0.026
2	0.091	0.112	0.067	0.115	0.101	0.069
3	0.152	0.188	0.114	0.187	0.164	0.123
4	0.214	0.266	0.165	0.259	0.226	0.182
5	0.275	0.342	0.216	0.327	0.285	0.244
6	0.333	0.413	0.265	0.390	0.341	0.306
7	0.387	0.479	0.314	0.448	0.393	0.366
8	0.438	0.538	0.360	0.502	0.442	0.424
9	0.485	0.592	0.404	0.550	0.487	0.479
10	0.529	0.640	0.446	0.594	0.530	0.530
11	0.569	0.683	0.486	0.635	0.569	0.578
12	0.607	0.721	0.524	0.671	0.606	0.622
13	0.642	0.755	0.560	0.705	0.640	0.662
14	0.674	0.785	0.593	0.735	0.671	0.699
15	0.704	0.812	0.625	0.763	0.701	0.732
16	0.732	0.835	0.656	0.788	0.728	0.763
17	0.758	0.856	0.684	0.811	0.753	0.790
18	0.782	0.875	0.711	0.832	0.777	0.815
19	0.804	0.891	0.737	0.851	0.799	0.837
20	0.824	0.906	0.761	0.868	0.820	0.858
21	0.843	0.919	0.784	0.884	0.839	0.876
22	0.861	0.930	0.806	0.898	0.856	0.892
23	0.877	0.940	0.826	0.911	0.873	0.907
24	0.892	0.949	0.845	0.923	0.888	0.920
25	0.906	0.957	0.864	0.934	0.902	0.932
26	0.919	0.964	0.881	0.943	0.916	0.942
27	0.931	0.970	0.897	0.952	0.928	0.952
28	0.942	0.976	0.913	0.960	0.939	0.961
29	0.952	0.981	0.928	0.968	0.950	0.968
30	0.962	0.985	0.941	0.975	0.960	0.975
31	0.971	0.989	0.954	0.981	0.969	0.981
32	0.979	0.992	0.967	0.986	0.978	0.987
33	0.987	0.995	0.978	0.991	0.986	0.992
34	0.994	0.998	0.990	0.996	0.993	0.996
35	1.000	1.000	1.000	1.000	1.000	1.000

Figure 1
Number of Patents by Application Year

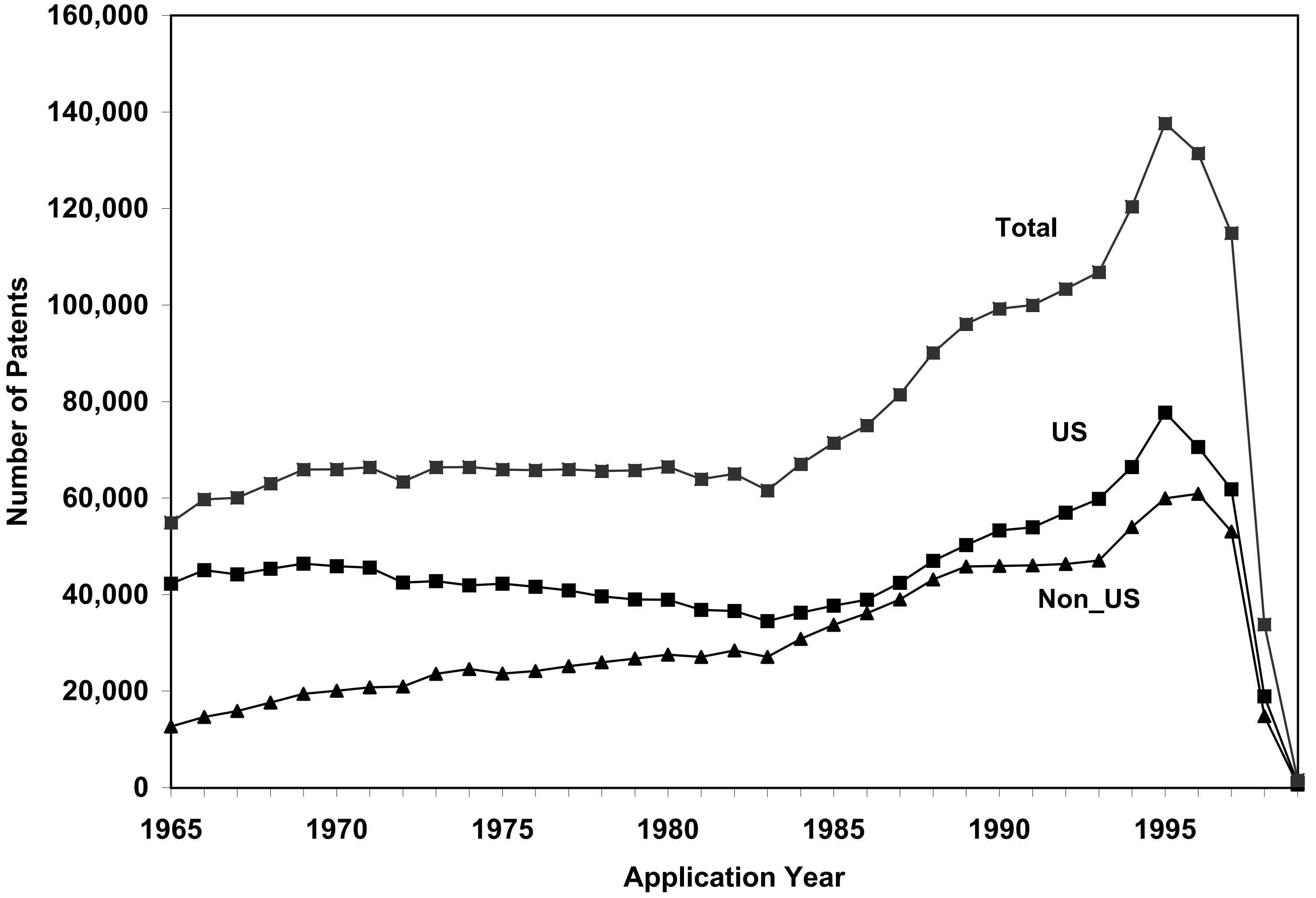


Figure 2
Number of Patents by Grant Year

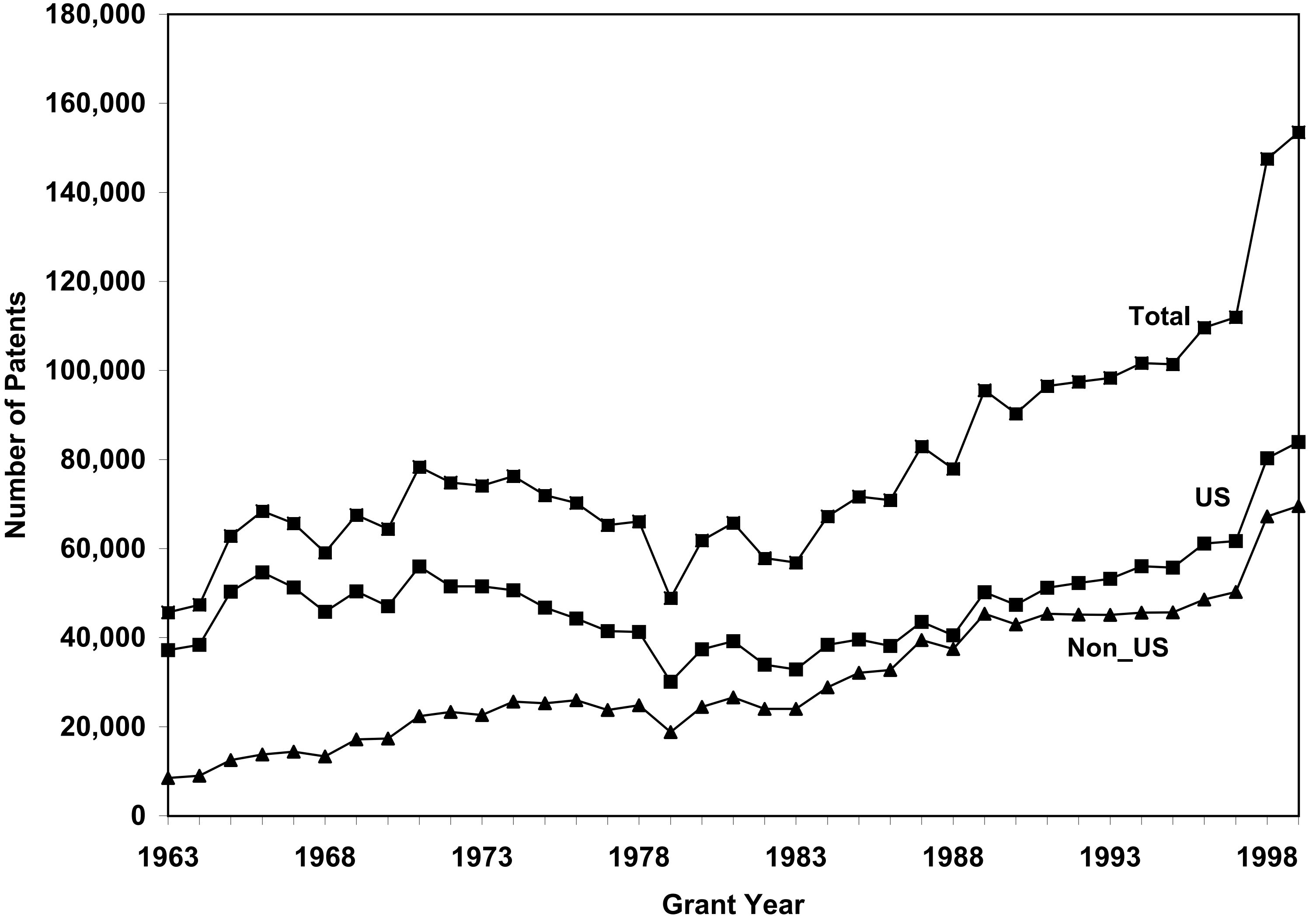


Figure 3
Share of Patents Assigned to Corporations

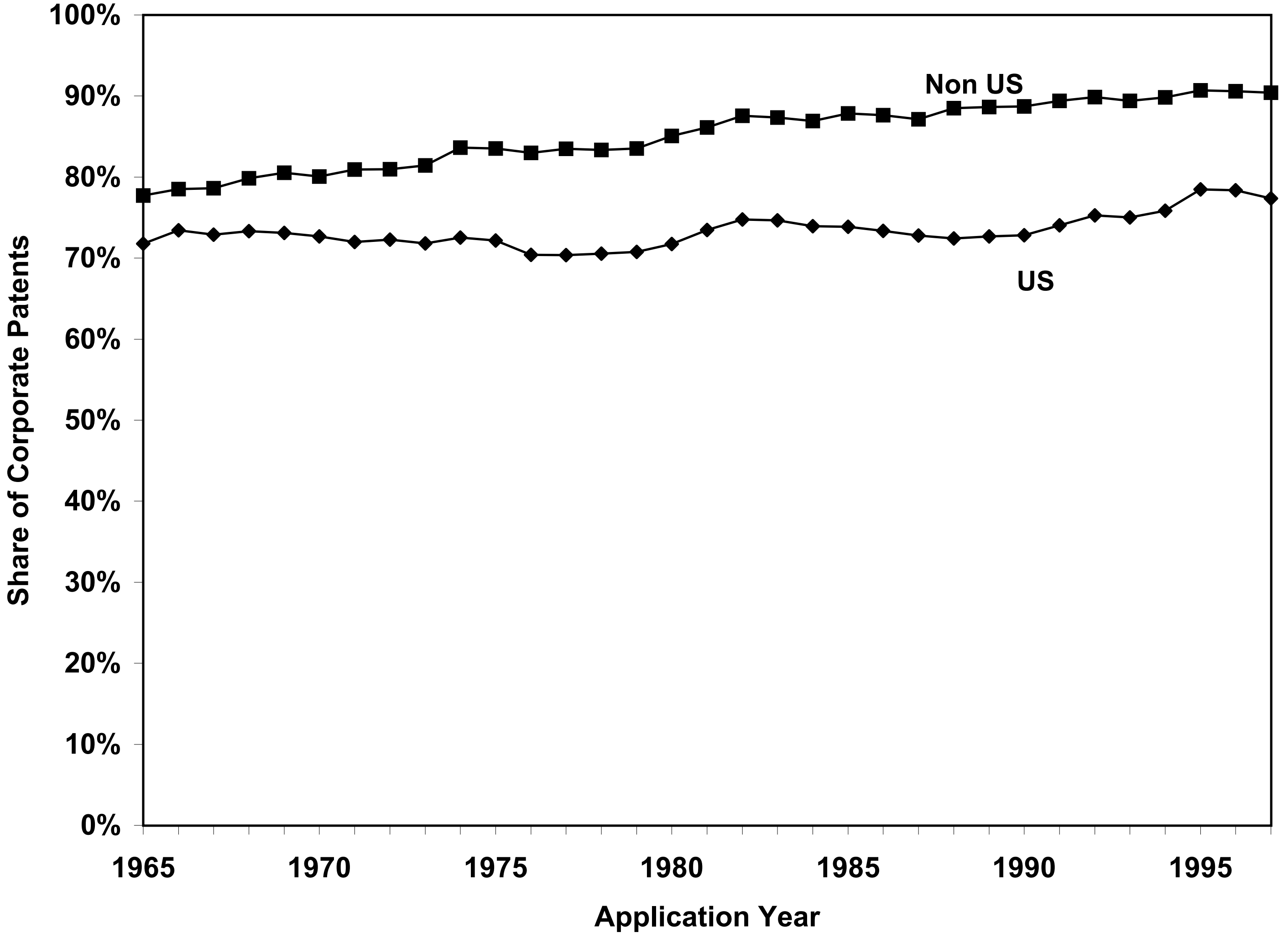


Figure 4
Distribution of Patents by Technological Categories
(absolute numbers)

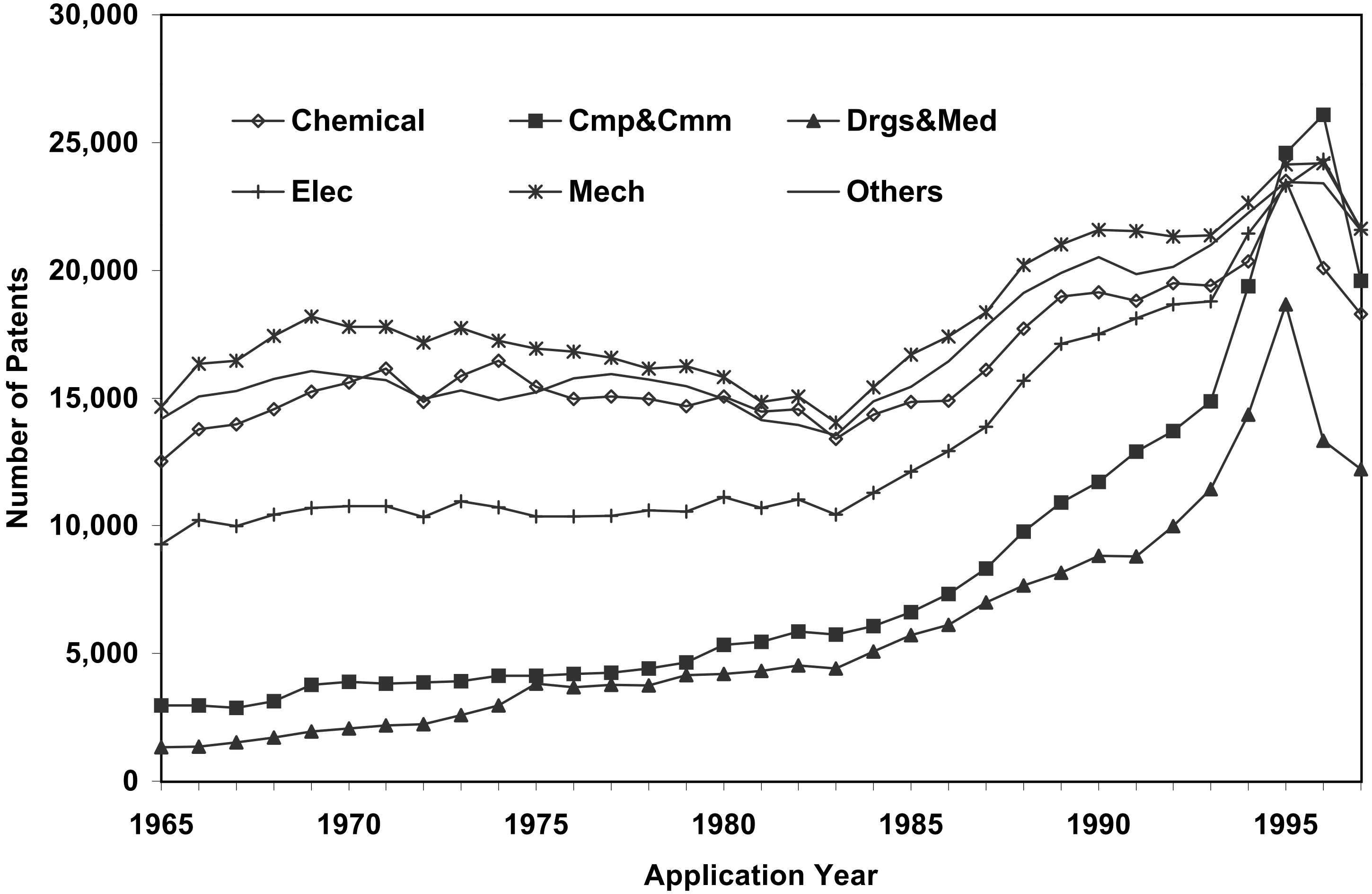


Figure 5
Distribution of Patents by Technological Categories - Shares

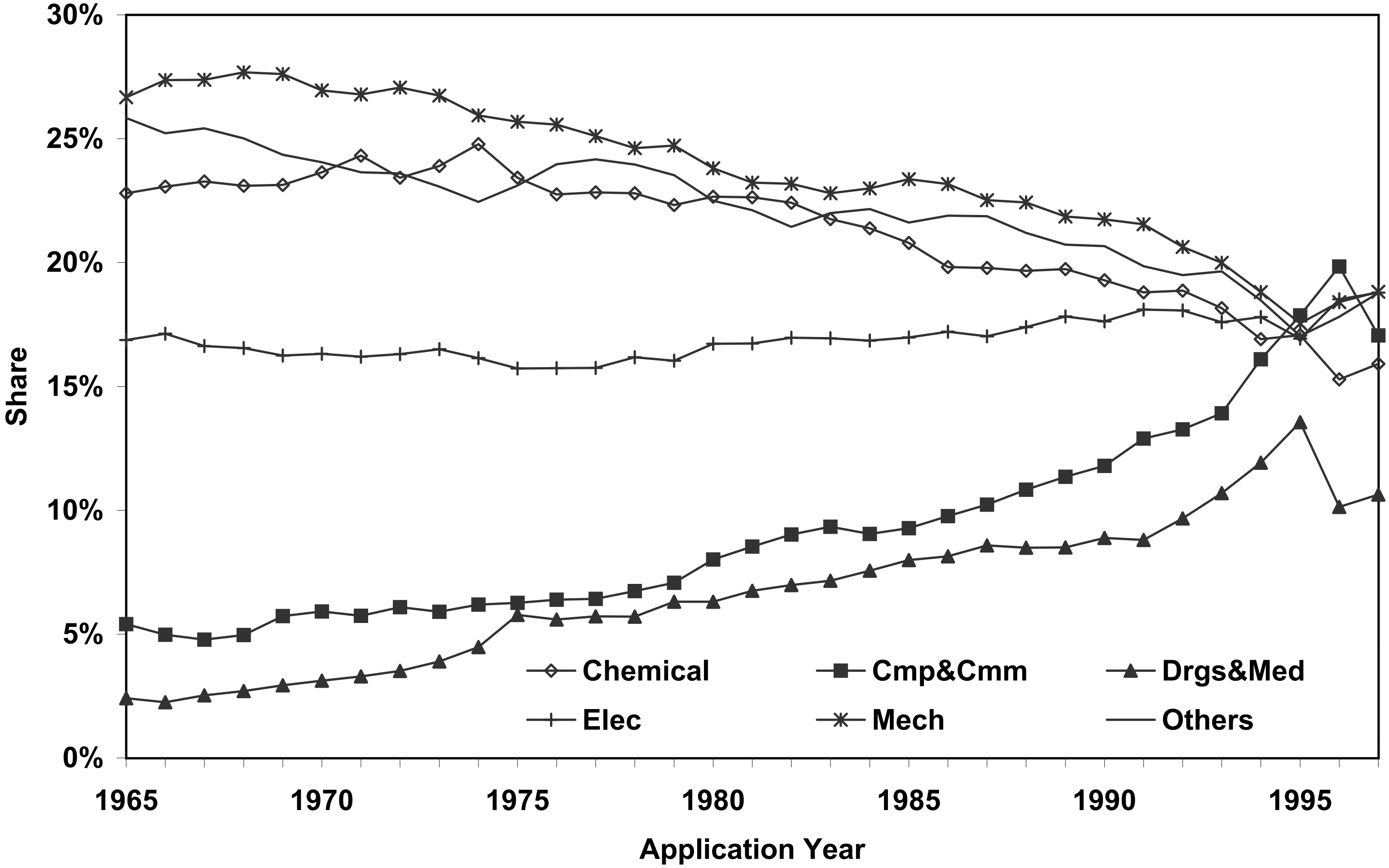


Figure 6
Mean Citations Made and Received by Application Year and Grant Year

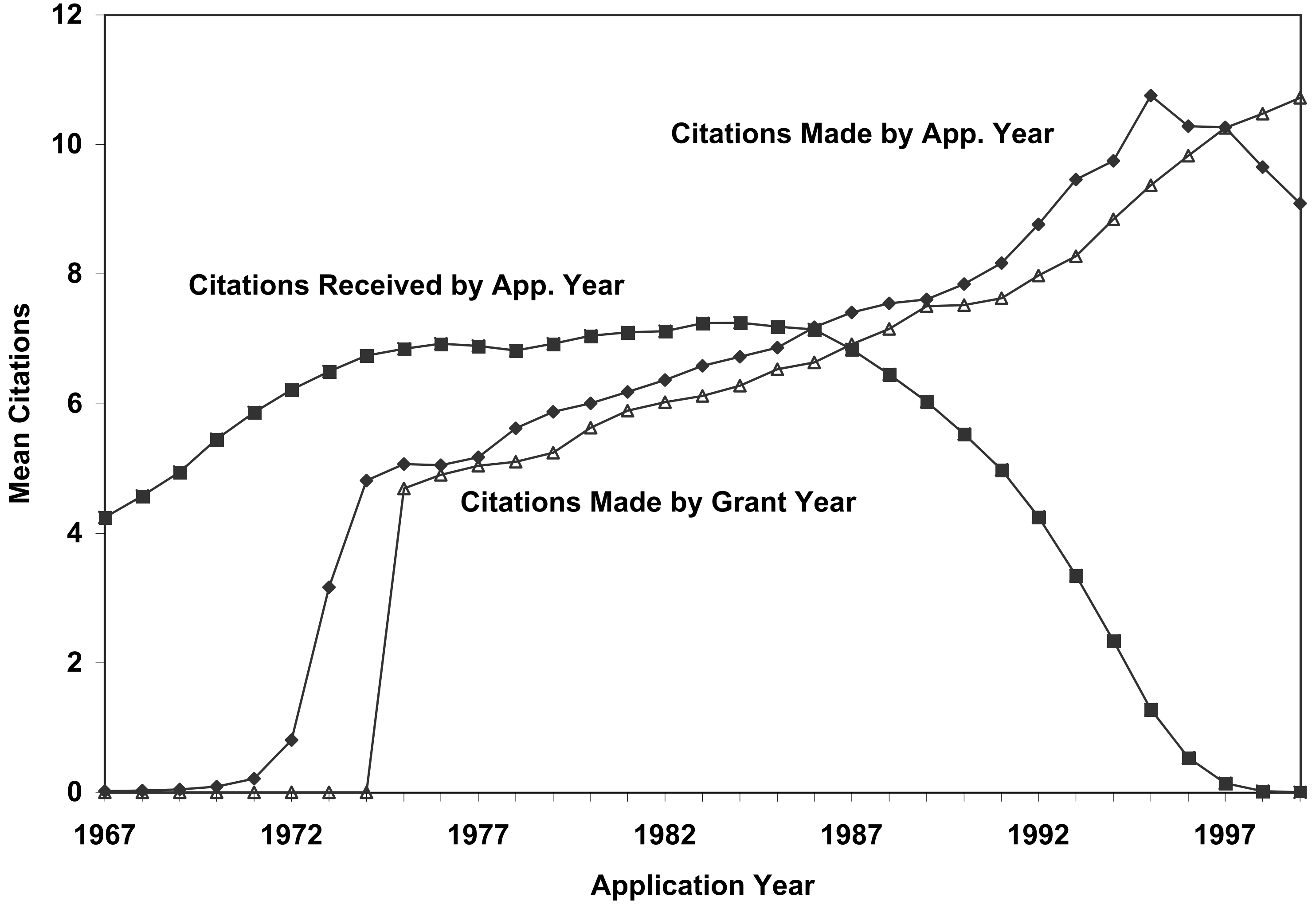


Figure 7
Mean Citations Made by Technological Categories

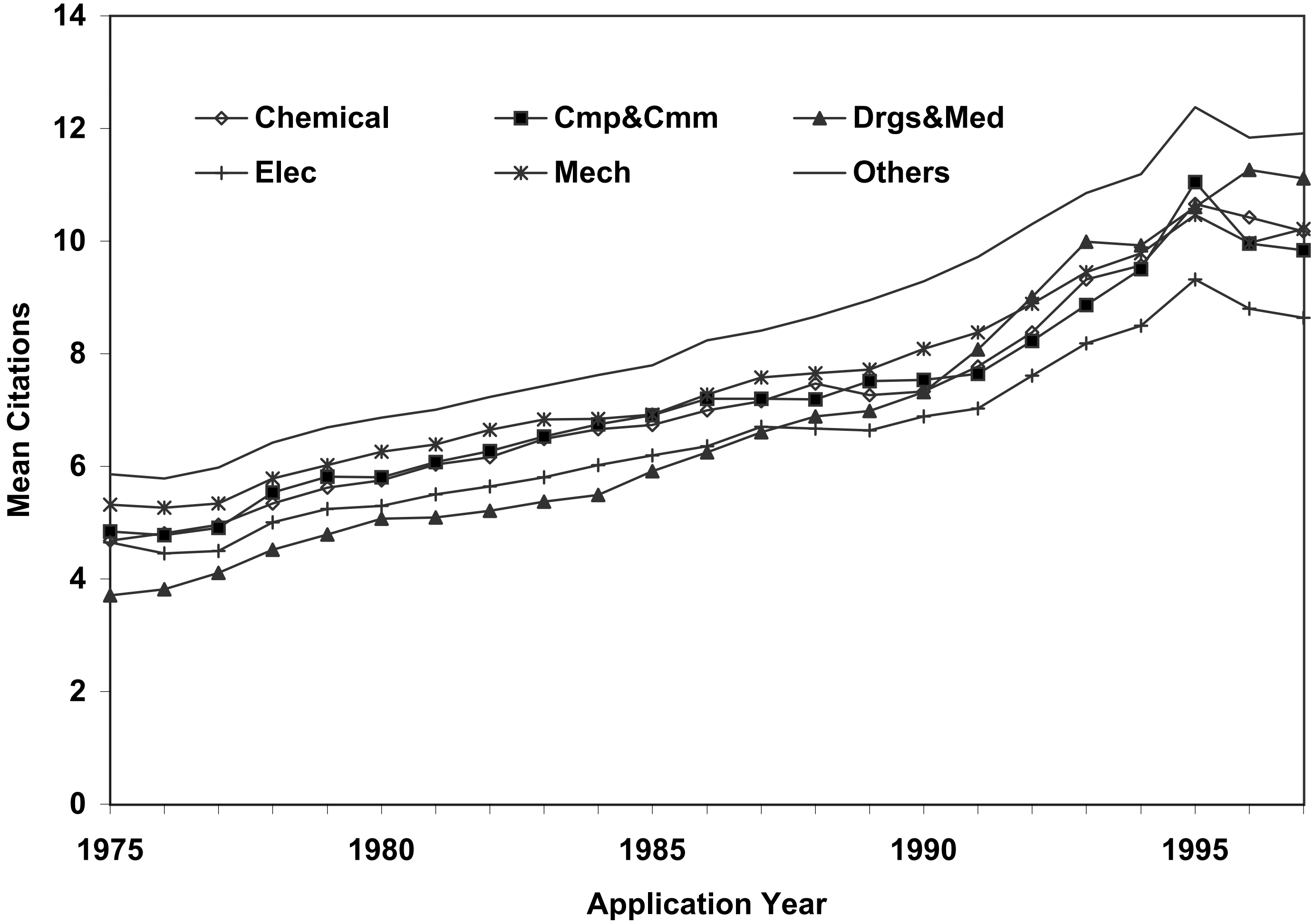


Figure 8
Mean Citations Received by Technological Categories

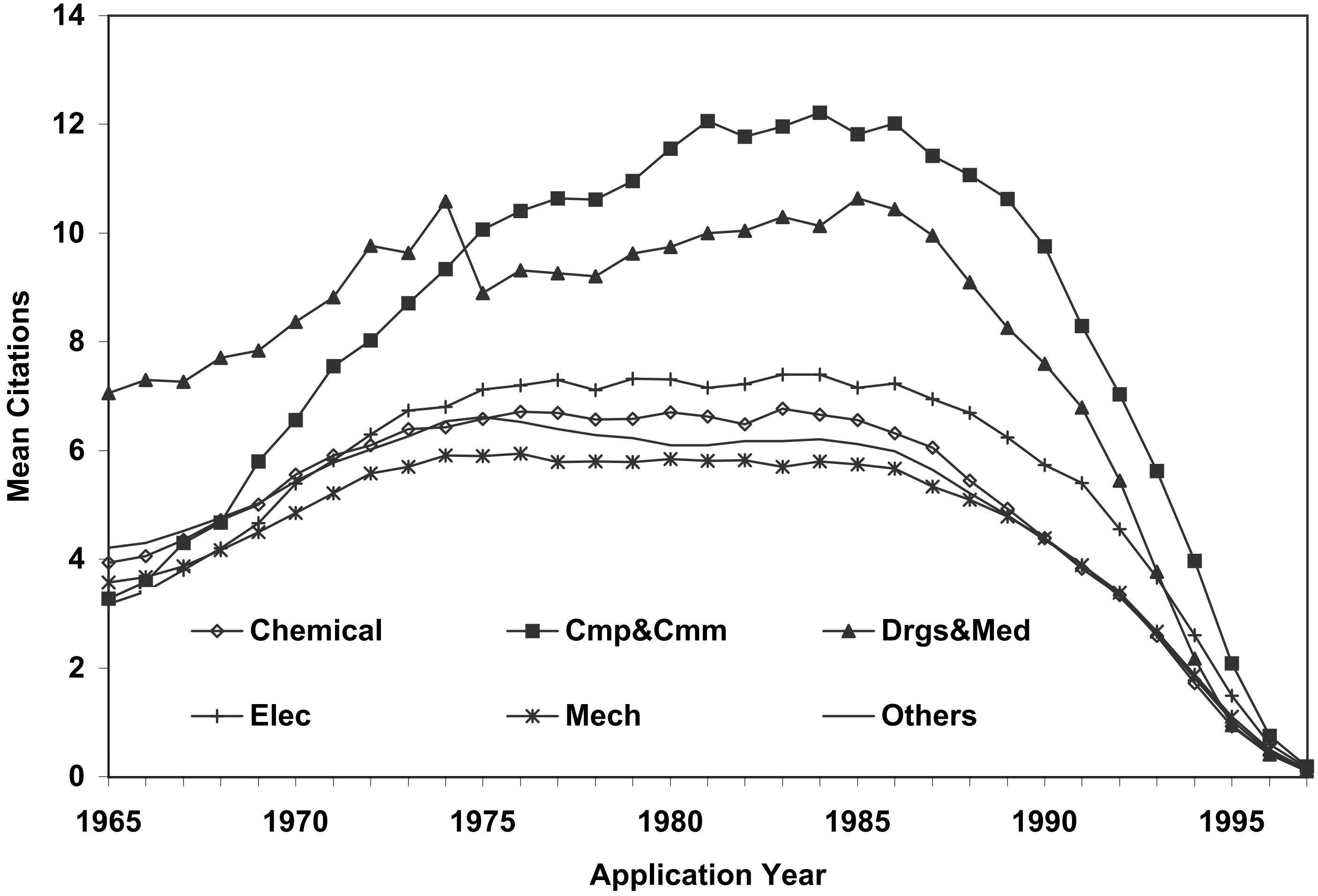


Figure 9
Distribution of Backward Citation Lags

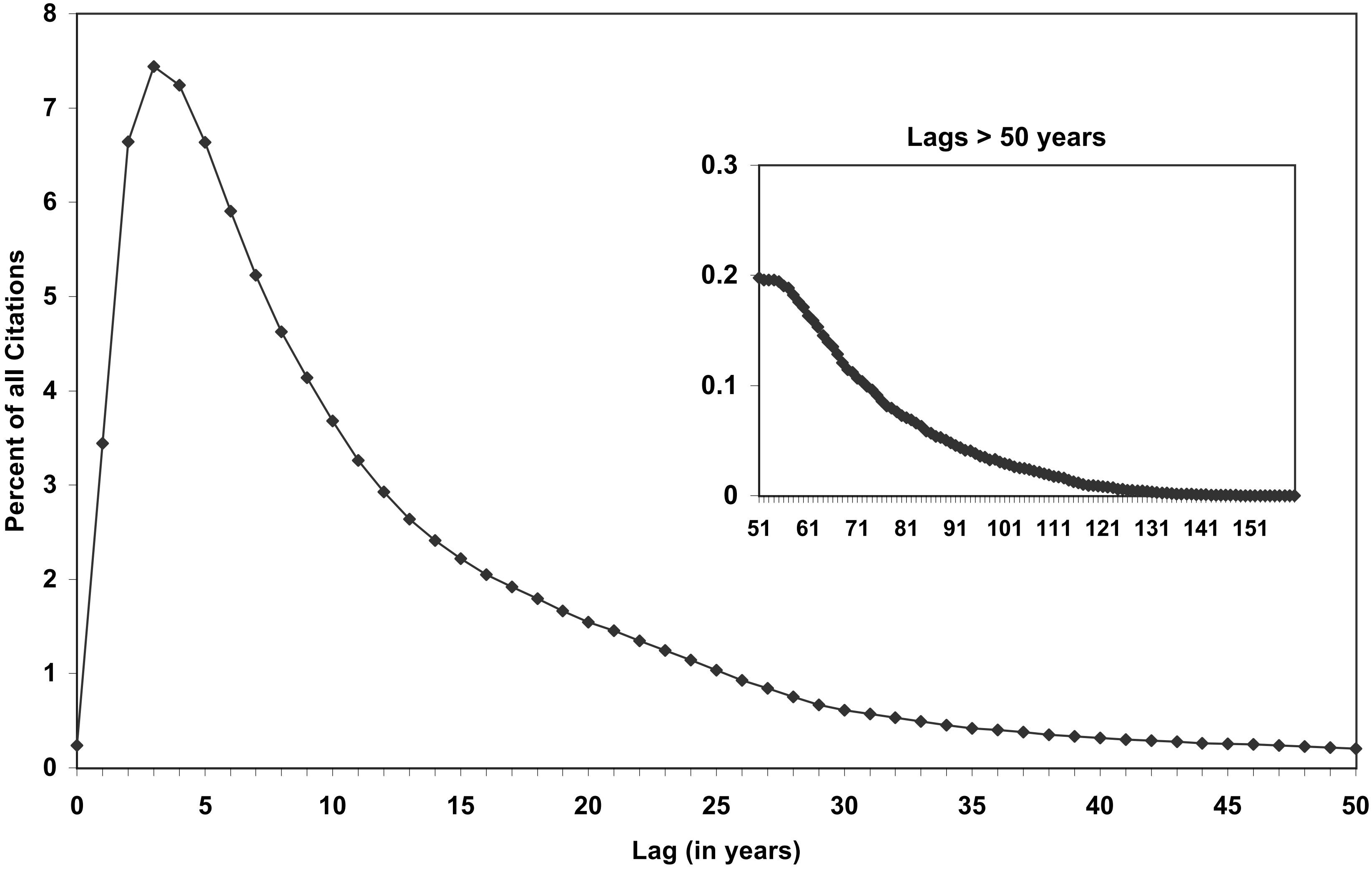


Figure 10
Cumulative Distribution of Backward Citation Lags (up to 50 years back)

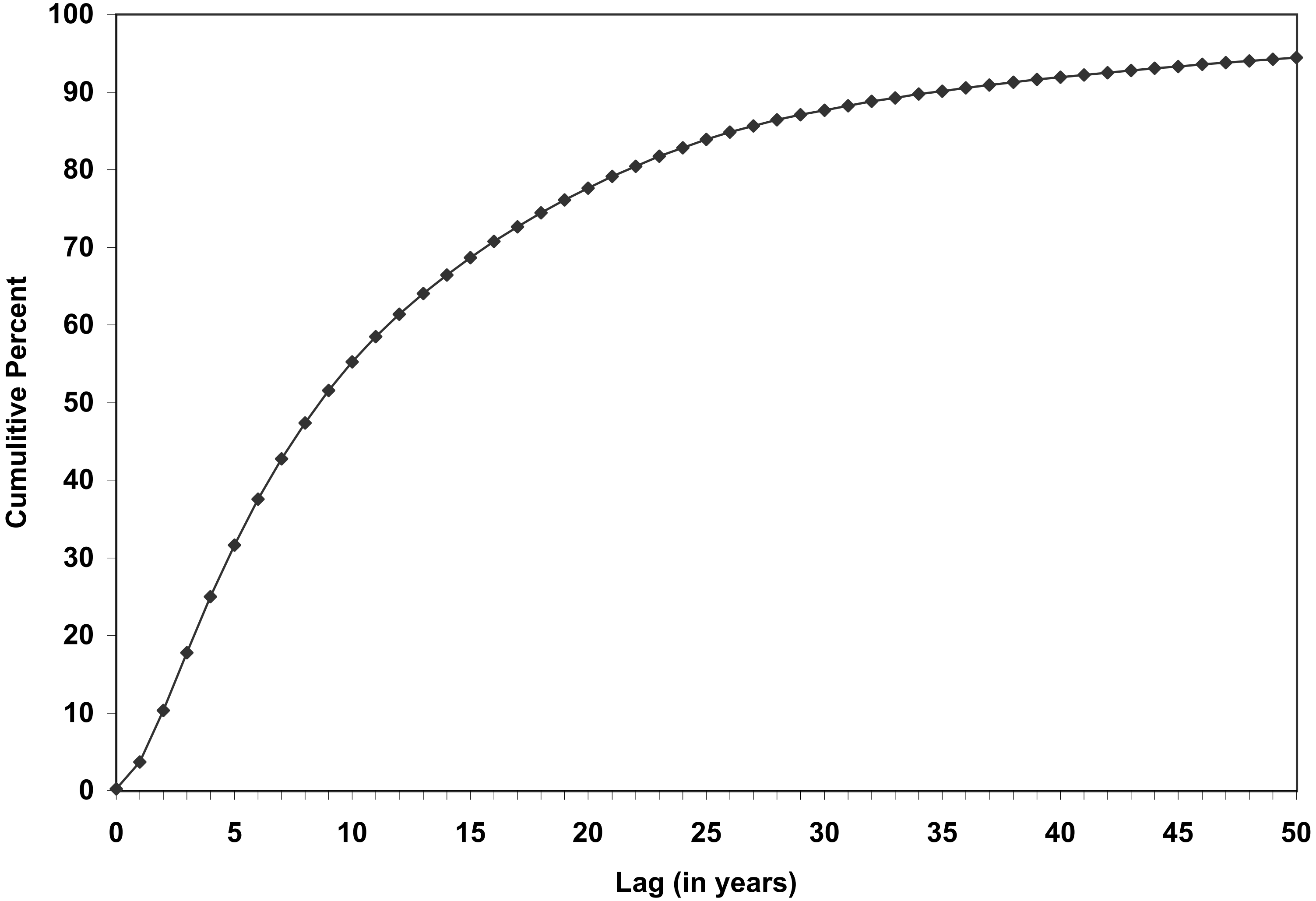


Figure 11
Distribution of Forward Citation Lags
Selected Cohorts: 1975, 1980, 1985 and 1990

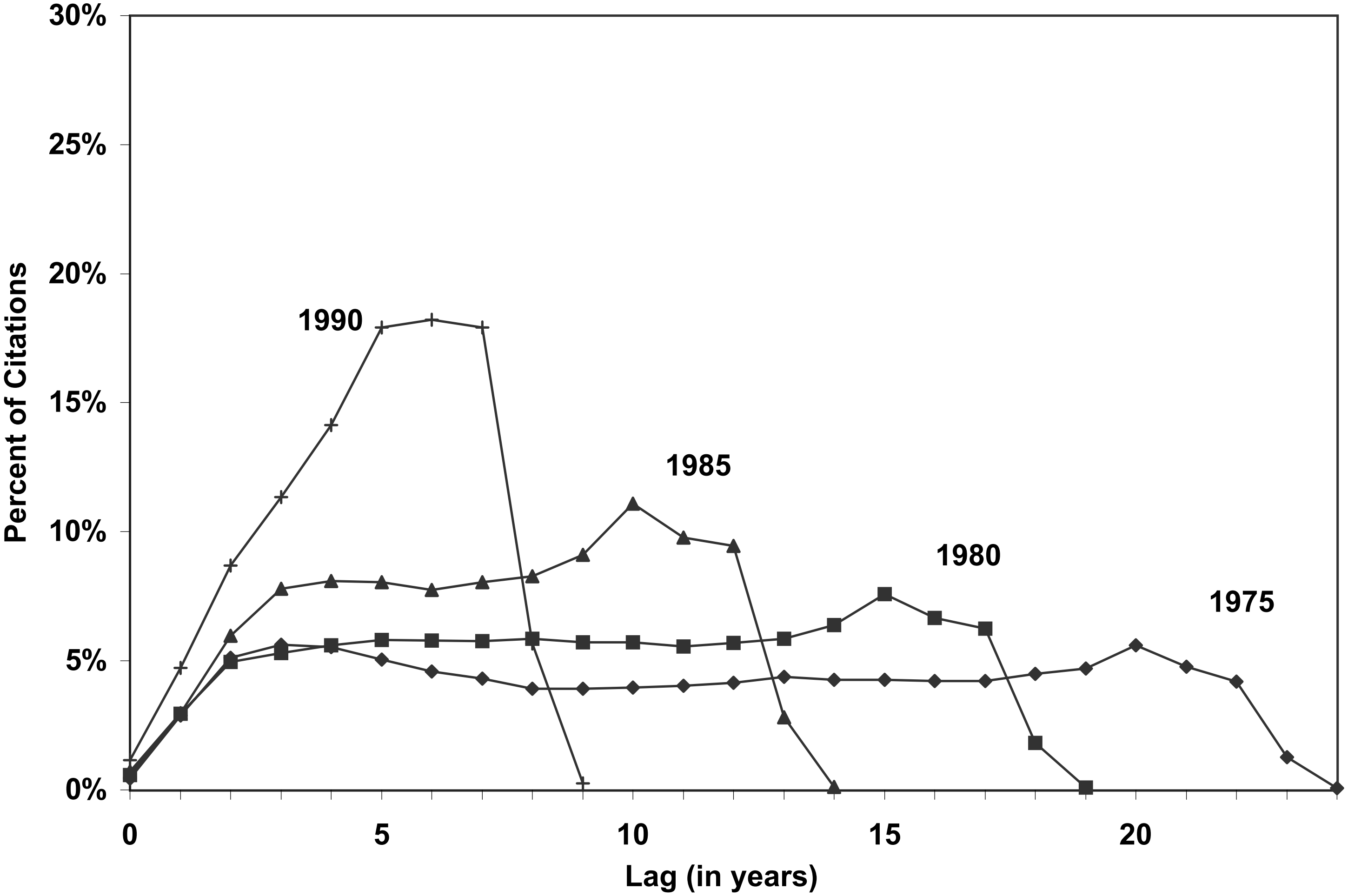


Figure 12
Self-Citations Made by Technological Category

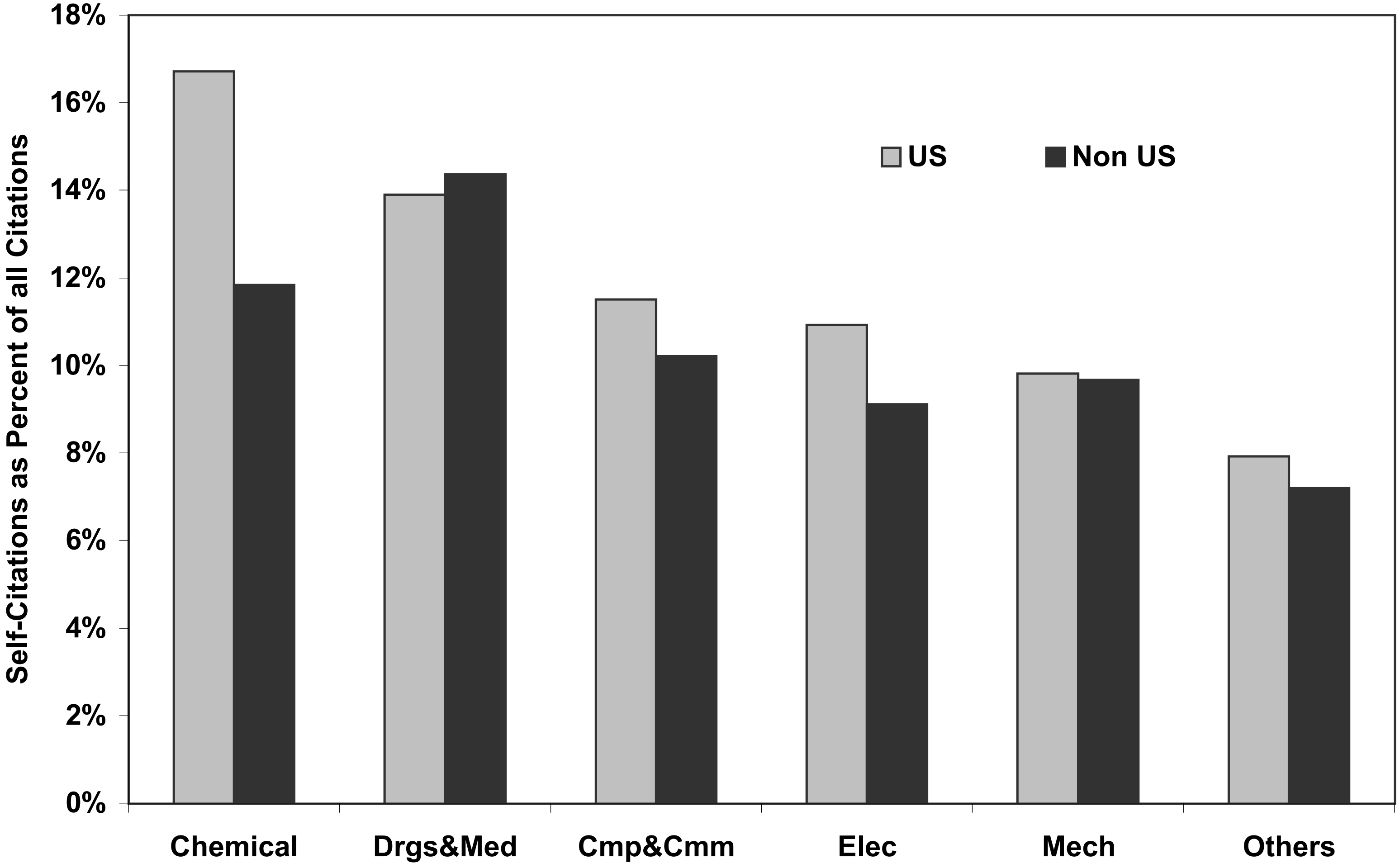


Figure 13
Self Citation as a Percentage of Total Citations Made by Application Year

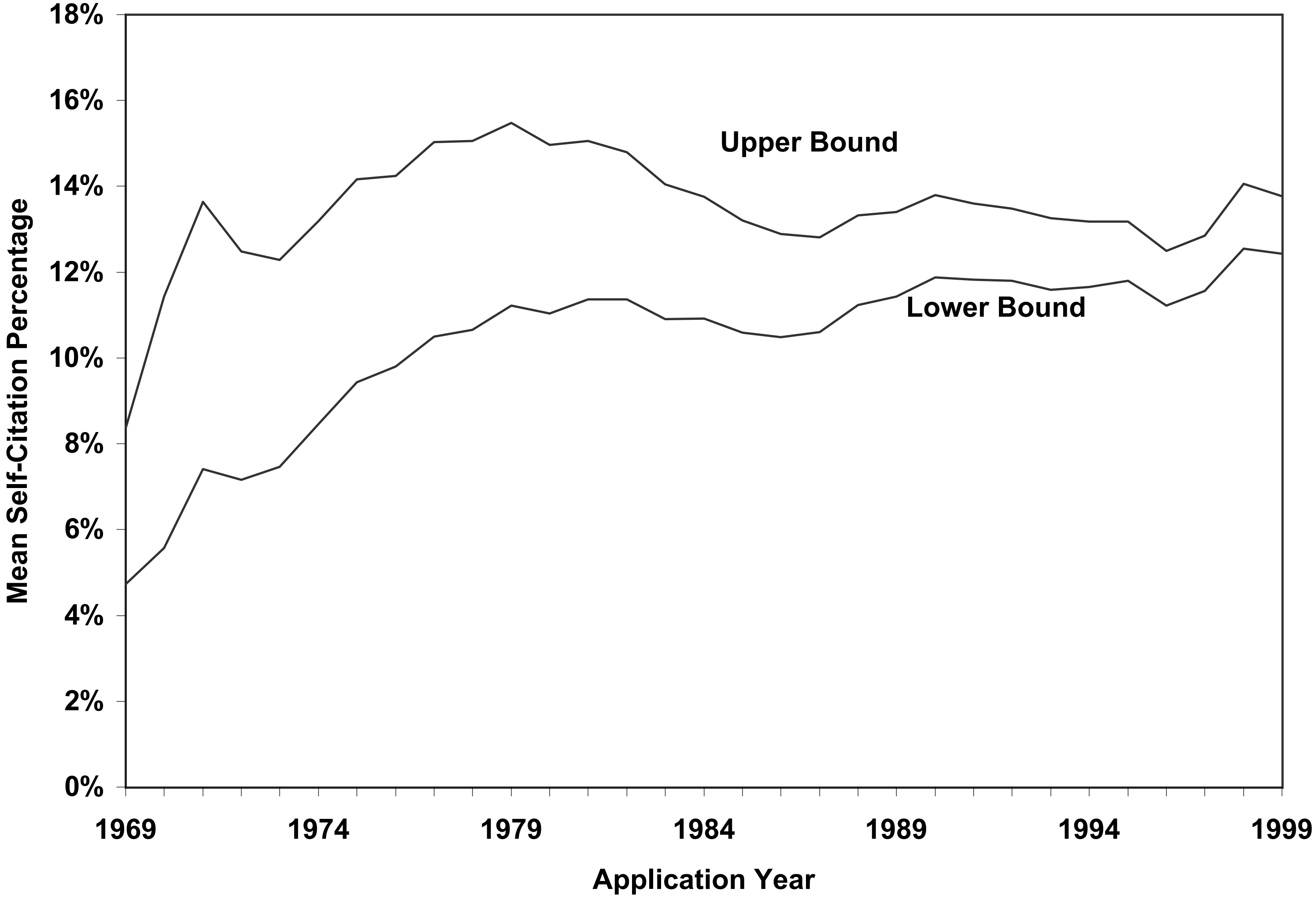


Figure 14
Self Citation as a Percentage of Total Citations Received by Application Year

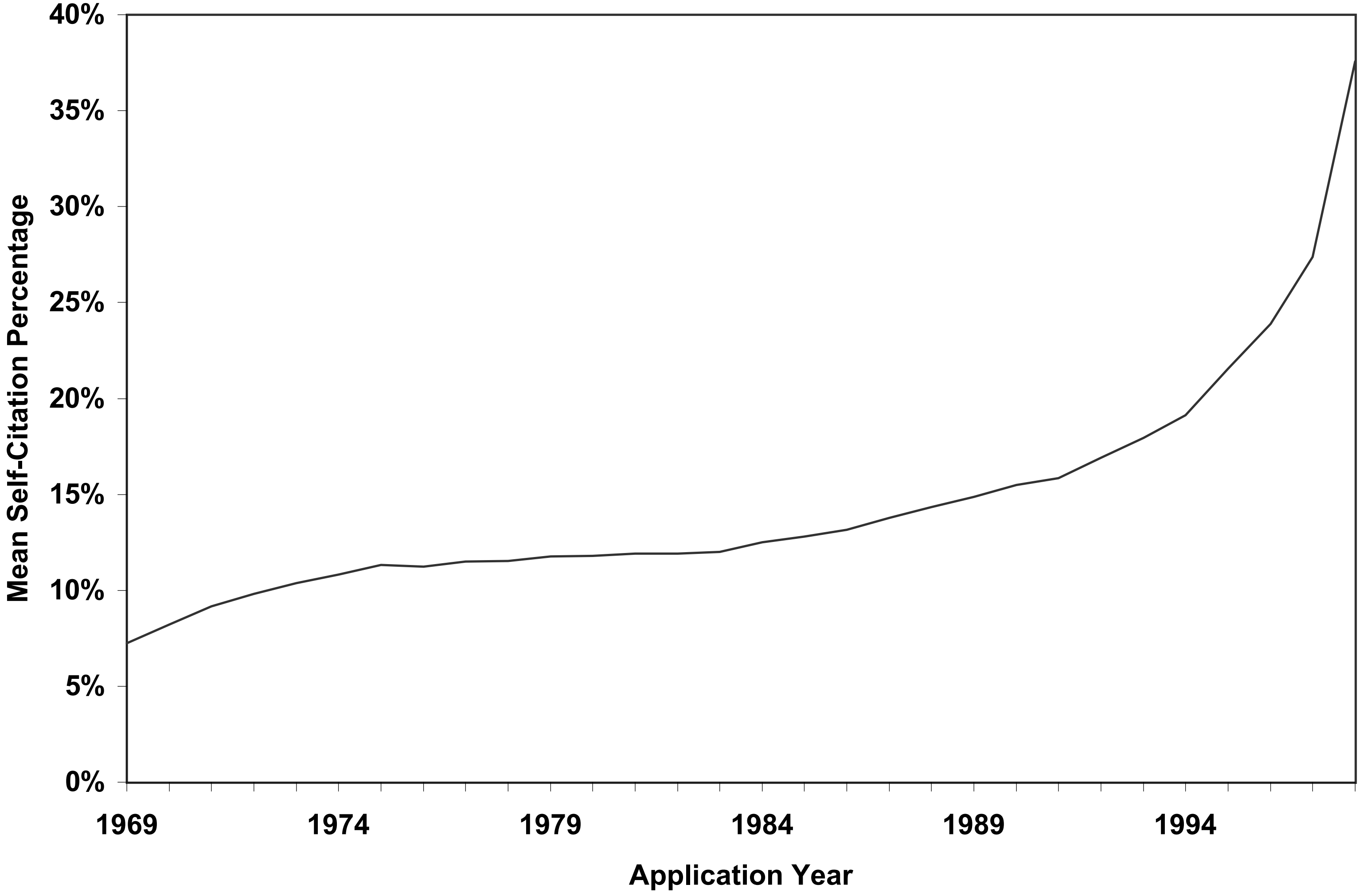


Figure 15
Measures of Generality and Originality
Yearly Averages

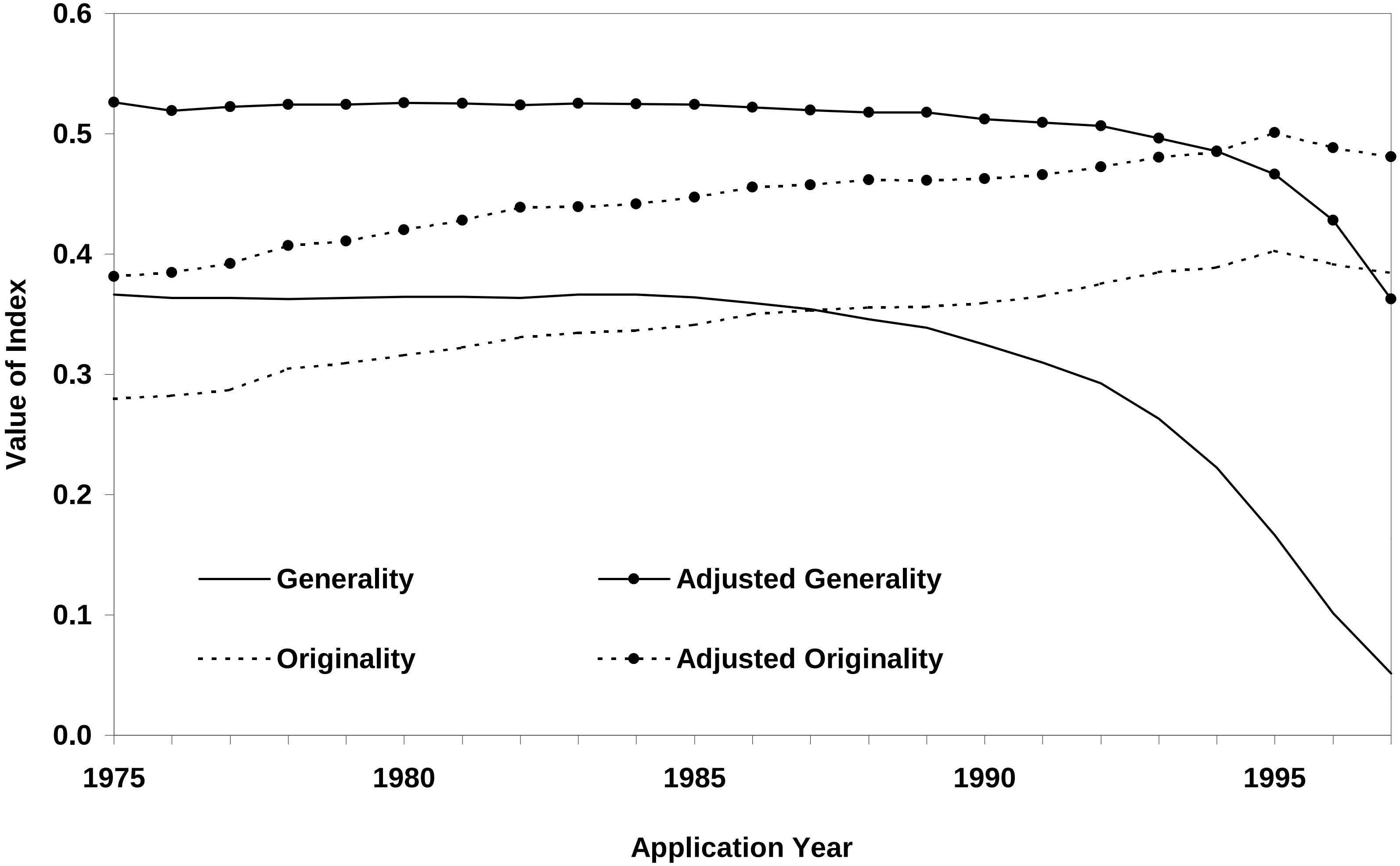


Figure 16
Generality by Technological Category
Yearly Averages

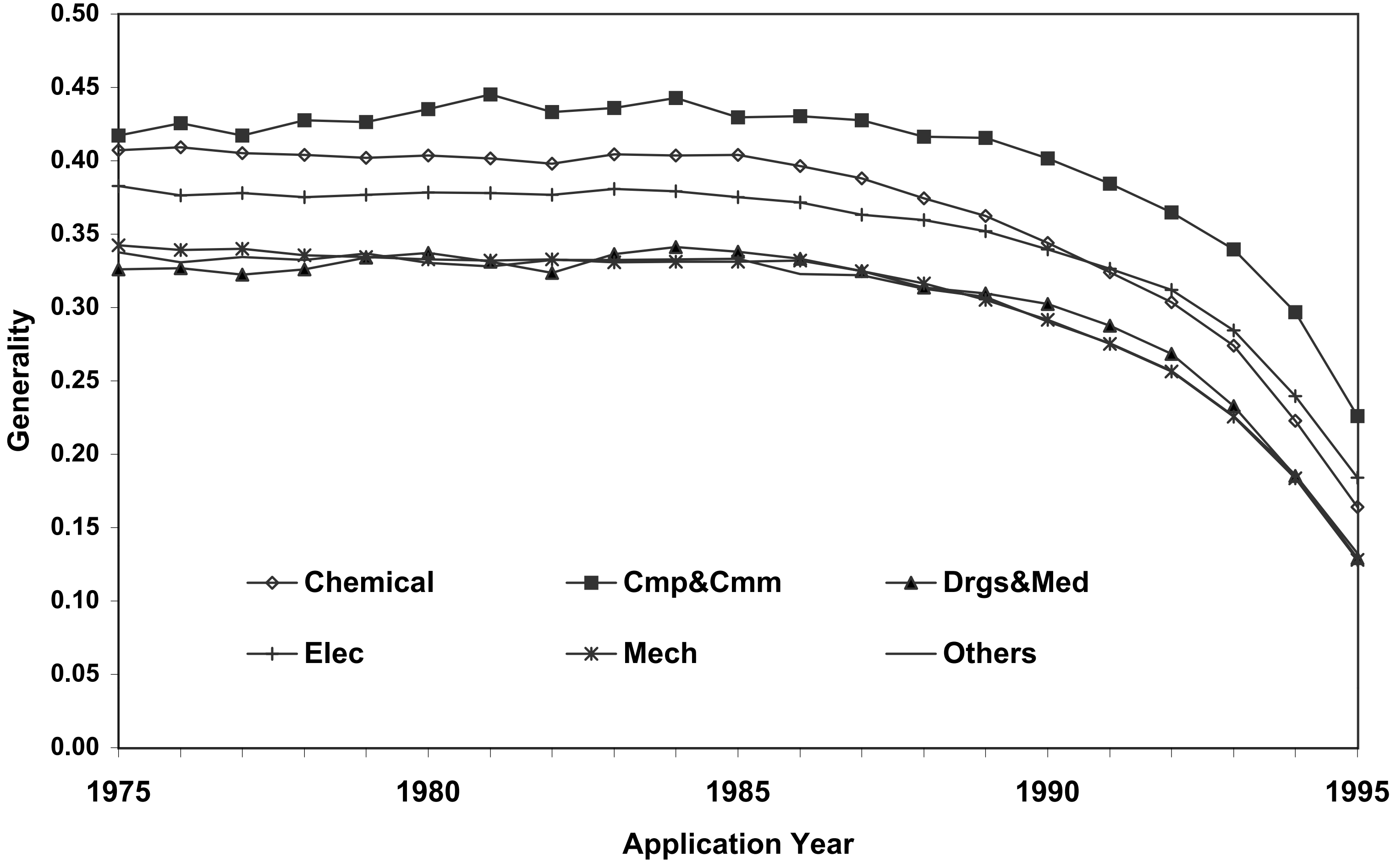


Figure 17
Originality by Technological Category
Yearly Averages

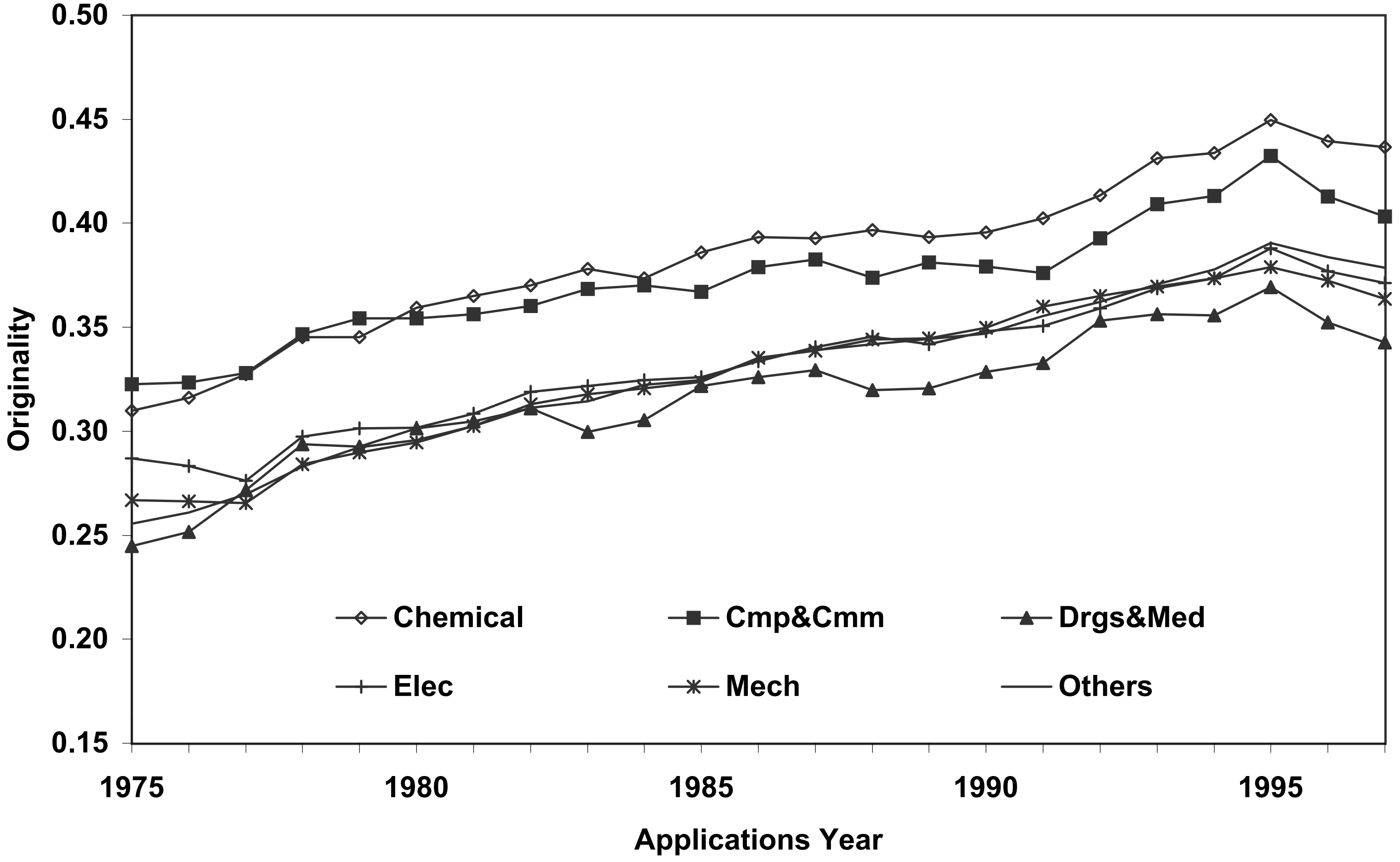


Figure 18
Claims Made by Technological Categories
Yearly Averages

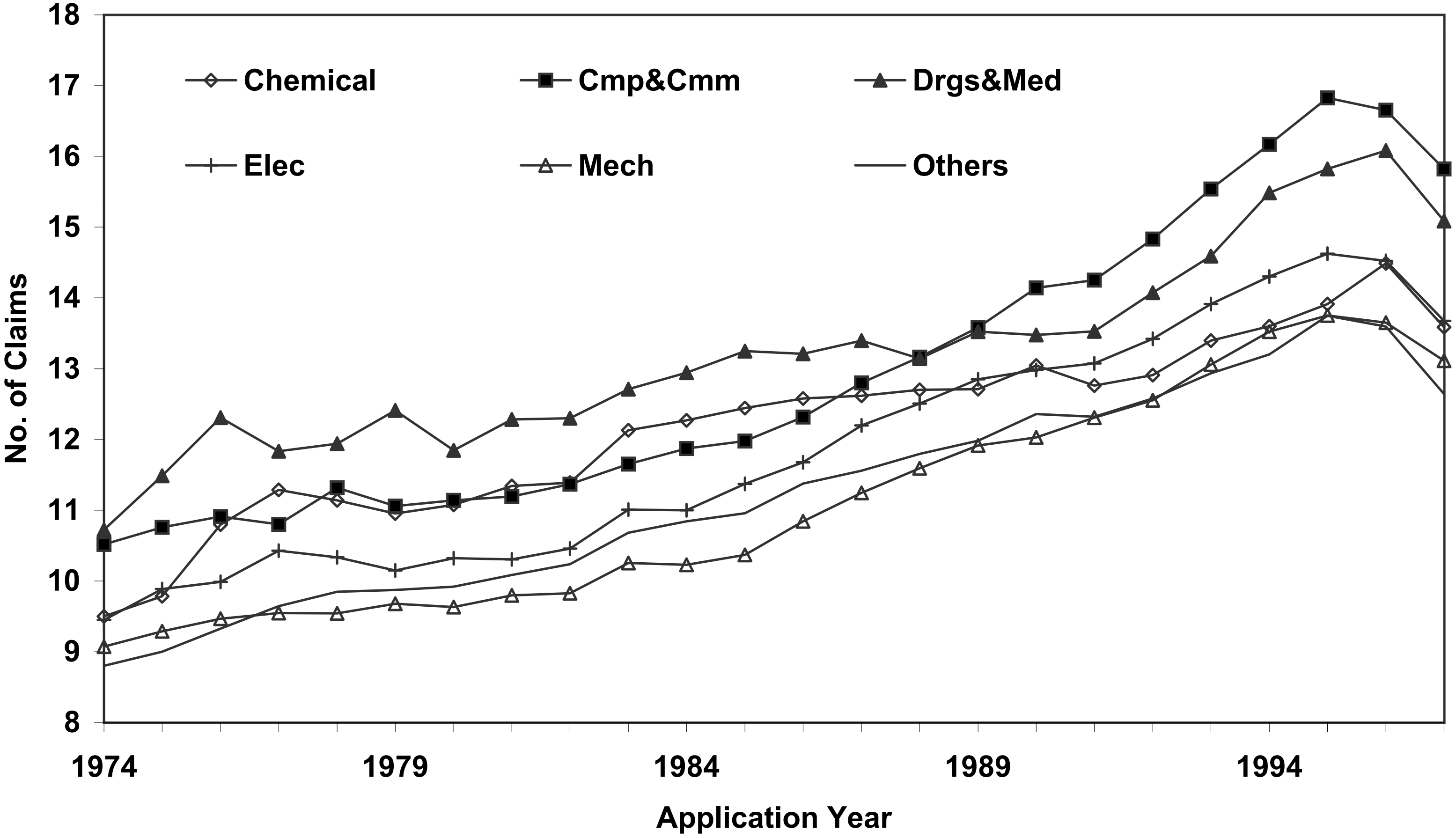


Figure 19
Percentage of Patents Matched to Compustat
(out of assigned patents)

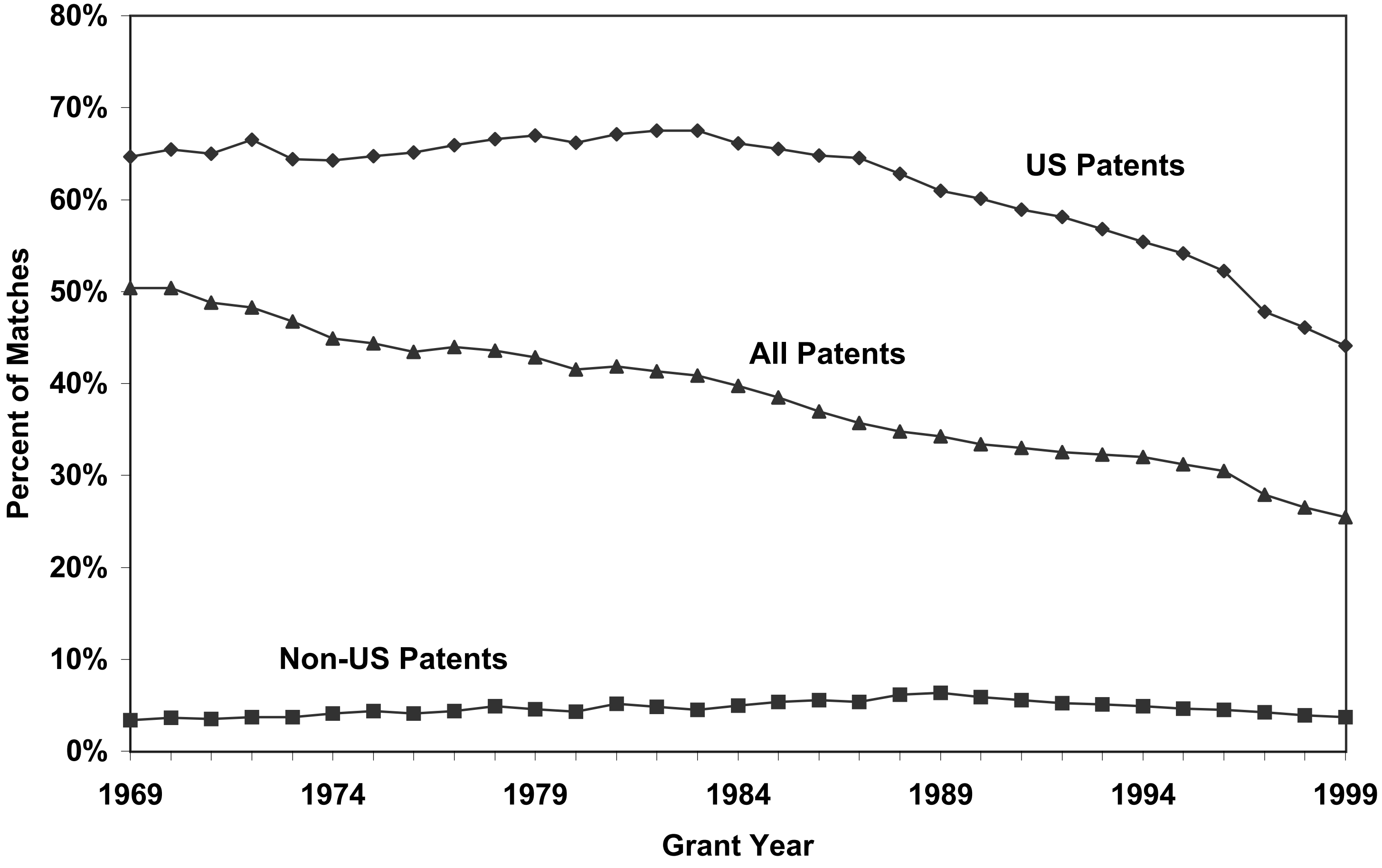
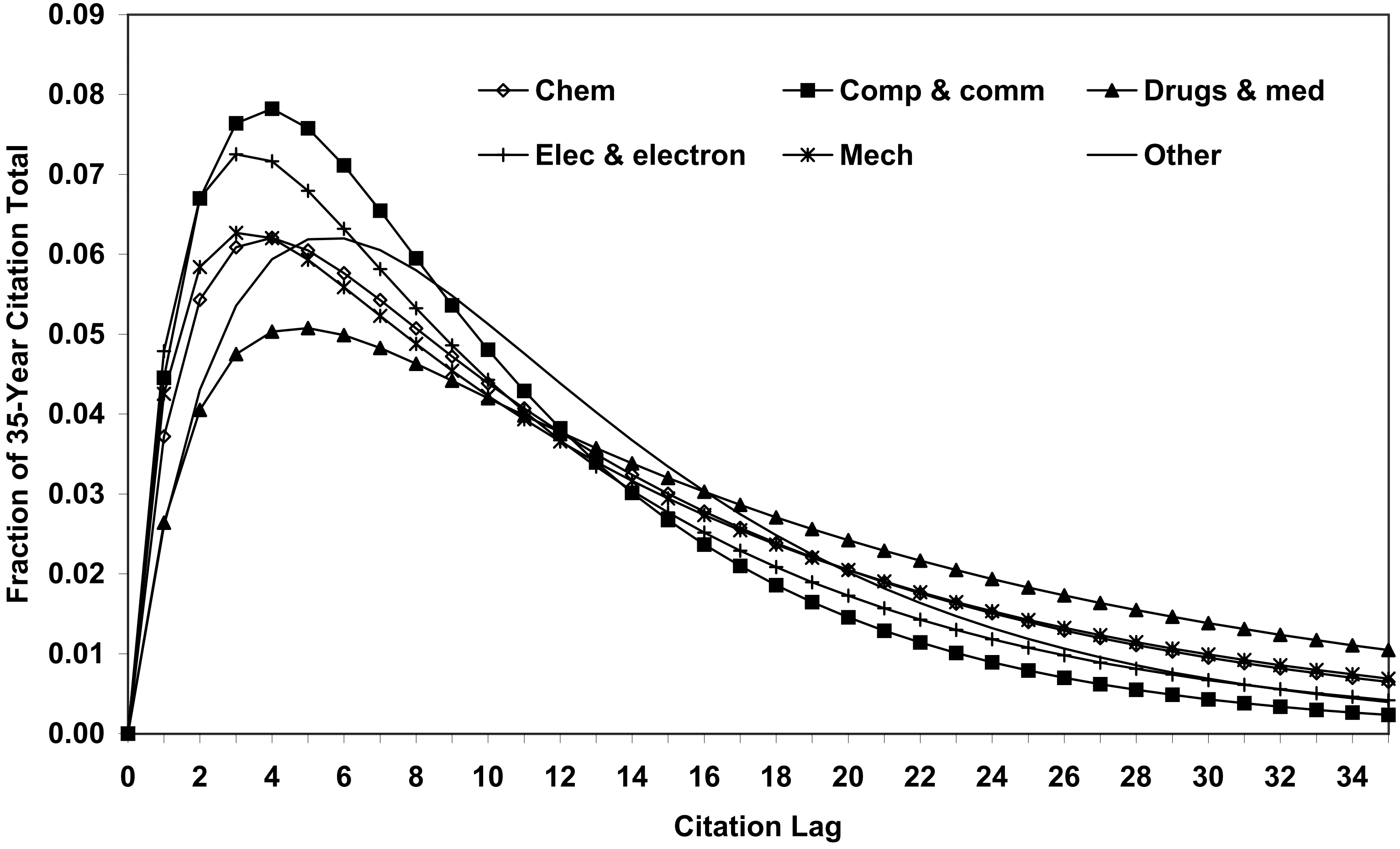
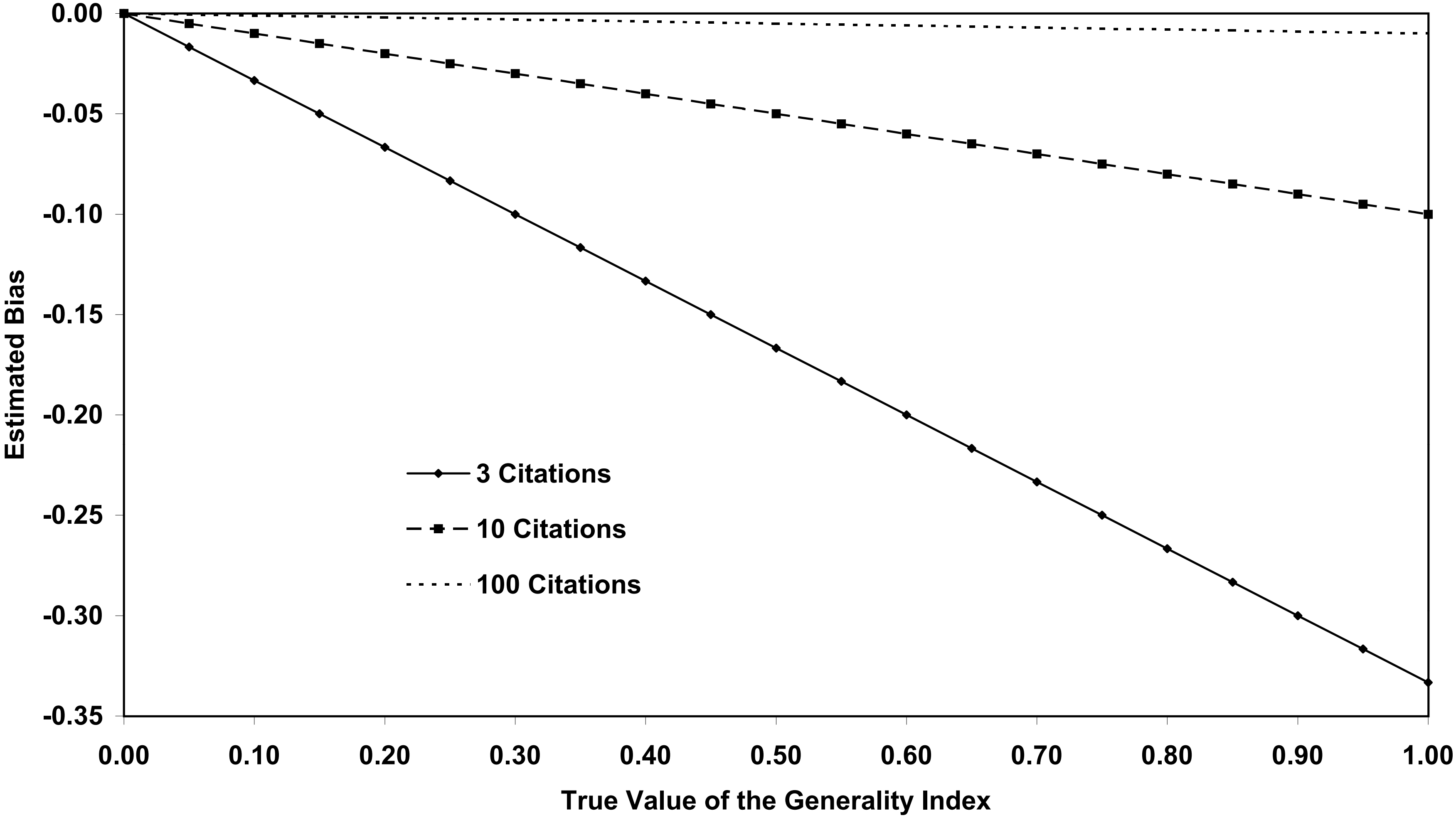


Figure 20
Simulated Citation Lag Distributions by Field



Appendix Figure 1
Bias of the Generality Index Based on Patent Citations


Estimated Bias
True Value of the Generality Index
3 Citations
10 Citations
100 Citations