NBER WORKING PAPER SERIES

THE NBER-RENSSELAER SCIENTIFIC PAPERS DATABASE: FORM, NATURE, AND FUNCTION

James D. Adams
J. Roger Clemmons

Working Paper 14575
http://www.nber.org/papers/w14575

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2008

The Andrew W. Mellon Foundation has generously supported this research. We thank Nancy Bayers and Henry Small of Thomson-Reuters for assistance with the data that we describe in this article. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

The NBER-Rensselaer Scientific Papers Database: Form, Nature, and Function
James D. Adams and J. Roger Clemmons
NBER Working Paper No. 14575
December 2008, Revised December 17th
JEL No. D2,O3

ABSTRACT

This article is a guide to the NBER-Rensselaer Scientific Papers Database, which includes more than 2.5 million scientific publications and over 21 million citations to those papers. The data cover an important sample of 110 top U.S. universities and 200 top U.S.-based R&D-performing firms during the period 1981-1999. This article describes the file system which comprises the database, explains the variables included in the files, and discusses the functions of the various files. It includes numerous descriptive tables, as well as graphs of the data in the time series dimension. In addition, it discusses limitations and strengths of the data as well as some questions that the data might be used to address.

The data discussed in this paper can be found at http://www.nber.org/RPI-sci-pap

James D. Adams
Department of Economics
Rensselaer Polytechnic Institute
3406 Russell Sage Laboratory
Troy, NY 12180-3590
and NBER
adamsj@rpi.edu

J. Roger Clemmons
Institute for Child Health Policy, College of Medicine
The University of Florida
PO Box 100147
Gainesville, FL 32610-0147
jrc@ichp.ufl.edu

I. Introduction

The following article discusses a large database of scientific papers, assembled and processed over a period of time, which might prove useful in understanding the role of science in the economy. The study of science is interwoven with the study of industrial innovation (Jaffe, 1989; Adams, 1990; Mansfield, 1991). Furthermore, it is clearly in the air that the economic role of science, measured by its role in commercialization, has increased markedly during the 20th and 21st centuries (Audretsch and Stephan, 1996; Narin, Hamilton, and Olivastro, 1997; Zucker, Darby, and Brewer, 1998; Adams, Chiang, and Starkey, 2001), suggesting that a country's scientific research is increasingly a key to growth. At the same time, the precise role of science in industrial invention and new product development remains largely unknown. Given the differences between scientific discovery and industrial innovation, controversy continues to surround the question as to whether the knowledge contained in papers matters to the same degree as the scientific training and human capital of industrial researchers. We hope that the database described in these pages will contribute to an eventual resolution of this and other puzzles concerning the economic role of science.

We refer to the collection of files and data as the NBER-Rensselaer Scientific Papers Database. Included are more than two million papers, written to varying degrees in the United States during 1981-1999, as well as citations made to and received by the papers. The institutions whose scientists and engineers author the papers consist of top U.S. universities and R&D-performing firms. The data also include collaborations between scientific institutions. This is significant because co-authorships typically indicate large investments of time and resources in the production of scientific research. Finally, the

data incorporate our efforts to fractionalize papers and citations, thereby reflecting collaborative research and avoiding multiple counting of scientific output in the economy as a whole.

Questions that the data could be used to address include the following: Who thinks about whose research, and in what fields? Who works with whom, and Why does this happen? What are the effects of thinking about and working with other scientists and engineers on invention and real output? How do these behaviors and their outcomes change over time?

This project originated in conversations between one of us (Adams) and the late Zvi Griliches as to how one might undertake database construction on the economics of science that would promote research on the topic at the NBER and elsewhere. It was an honor to be recruited in this way by so peerless an economist, as those who know best must realize. Later, Adams was fortunate in bringing Clemmons on board. Clemmons' skill in handling data, his insistence on asking the right questions, and his work ethic have kept the project going when it would otherwise have ended badly.

In many ways the project was intended to parallel and complement the research on patents that had been ongoing at NBER over a period of time, as this is reflected first in Griliches (1986) and later in Jaffe and Trajtenberg (2002). It was immediately understood that the project would be risky and that it would not be fashionable. The idea of working with citations is of course, not new. To our knowledge, maps that link industries together go back to at least Terleckyj (1974) and Scherer (1982a, b). However, the development of citations in economics to measure the importance of ideas and their

flow is contained in the work of Trajtenberg (1990) on CT scanner patents and the compendium of work in Jaffe and Trajtenberg (2002).

Papers that have used the data include Adams, Clemmons and Stephan (2004 and forthcoming 2006; and 2006), Adams, Black, Clemmons, and Stephan (2005); and Adams and Clemmons (2008a, 2008b; and forthcoming 2009). The reader may find additional information concerning the data in these papers.

The rest of the paper consists of three sections. Section II describes the data from several perspectives. In part A. of the section we discuss the form and function of the files comprising the database. Part B. discusses the distribution of papers by fields of science, and it does so separately for the top 110 U.S. universities and the top 200 U.S. R&D-performing firms. The papers data are graphed in Part C., the citations data are graphed in Part D., and finally, Part E. presents graphs of the collaborations data. Section III is a discussion and assessment of limitations and strengths of the data. Section IV concludes.

II. Description of the Database [1]

A. The File System

The database consists of eight files. Table 1 provides an overview, listing file names, numbers of observations, and file functions. We shall keep referring to this table as a means of organizing our tour of the data. The eight files are a careful reworking of archival data from Thomson-Reuters on scientific papers and citations. Arranged by publication date their time period is January 1, 1981 to December 31, 1999. The data begin in 1981, because in this year the company implemented a new and more comprehensive data processing system that is not entirely incompatible with earlier data.

[1] For access to the data, go to http://www.nber.org/RPI-sci-pap

The data end in 1999, because during the subsequent extraction process for the project, 1999 was the latest full year available.

The information specifically derive from Thomson-Reuters' Current Contents database, which at the time covered an expanding set of 5,507 journals across the sciences[2]. In addition, we use a set of 1,630 discontinued or renamed journals that were cited by Current Contents journals. Originally included are 2,836,700 scientific papers written in one or more of the Top 110 U.S. Universities, as well as 238,277 papers written in one or more of the Top 200 U.S. R&D Firms. Some overlap takes place between the universities and firms, because of scientific collaboration between institutions. A total of 21,386,007 citation pairs occur between *groups* of papers defined by citing and cited institutions, fields (defined below), and years[3]. Likewise 797,348 collaboration pairs occur between *groups* of papers, defined by collaborating and collaborated pairs of institutions, in turn arranged by field and year[4].

The designation of Top 110 University is Thomson-Reuters'. It is based on publication volume, and indeed these universities account for the majority of academic papers written in the U.S. The list of Top 110 universities appears as Table A.1 of the Appendix.

The designation of Top 200 Firm was developed for this study. With one exception it refers to the 200 publicly traded corporations who performed the most R&D in 1998, that were based in the United States, for which histories of reasonable length

[2] At the start of the project, the consulting group in Evaluative Bibliometrics at Thomson-Reuters was known as the Institute for Scientific Information (ISI). We shall use this original name for the company from time to time.

[3] Put differently, the 21,386,007 citation pairs represent counts of citations at the *six*-dimensional level of citing and cited institutions, fields, and years. We discuss this in more detail below.

[4] Since papers are assigned to fields and years according to the journal where they appear, collaborations occur in the same field and year. This leads to a *four*-dimensional file classified by field, year, and collaborating and collaborated institutions.

could be constructed. The one exception is Bell Communications Research (Bell CORE; ticker BELLC), which spins off from AT&T in 1984. The list of Top 200 R&D firms appears as Table A.2 of the Appendix.

Divestiture (and acquisition) can change the nature and amount of scientific research in a firm. We note two major examples of this in the data. Table A.3 discusses the treatment of the AT&T and General Motors families of firms, for which the divestiture problem is paramount. This is because, for General Motors and AT&T, divestiture significantly alters the practice of science and R&D in each firm. Note that we treat papers and citations of Lucent (Bell Laboratories) and Bell Communications Research as separate from AT&T in all years and that we treat papers of Delphi Automotive Systems as separate from General Motors in all years. This strategy allows the user to construct the definition of the firm that is most suitable for them. One such definition, which we have used in our papers, retains spinoffs prior to divestiture as divisions of the main firm, and afterwards treats them as separate firms.

As we have said, Table 1 is an overview of the database. We discuss the eight files in order of their appearance. The first file, UNIVERSITIES, is described in Table 2. It includes three variables: STANDALONE, an indicator variable equal to one if a university is a standalone campus or zero if it is a multi-campus system; UNIVID, or the ID of the university; and UNIVNAME, the name of the university in NSF's CASPAR database. UNIVID is the modified Federal FICE Code for a given university. In its original sense, the FICE code is an identifier assigned by the Federal Interagency Committee on Education (FICE). As with any identification scheme, though, the coding system adjusts to suit the users. Our version of the FICE code harmonizes with the

CASPAR database of universities, a collection of university data assembled by the National Science Foundation. Thomson-Reuters also uses the CASPAR definitions of standalone campuses and university systems in building its institutional dictionaries for the Top 110 universities. The CASPAR FICE codes and university definitions are the ones used in the database.

The top 110 universities include 26 university systems. For these 26 universities the file UNIVERSITY_SYSTEMS includes three variables: BRANCH, the name of the branch campus included in each system; UNIVID, the university ID; and UNIVNAME, the university name. Since there are 142 branch campuses in the 26 systems, the average number of branches is 5.5. Table 3 describes the file UNIVERSITY_SYSTEMS.

The third file, UNIVERSITY_DESCRIPTION, summarizes the paper and citation statistics for the 110 universities. Table 4 shows that these statistics are arranged by university (UNIVID), Thomson-Reuters ISI 88 science field (ISI88), and publication year (YEAR). The number of observations is 143,119. Besides UNIVID, ISI88, and YEAR, the file includes ten variables. CITSFIRM is the total number of forward (future) citations received from firms for a given university, field, and year. The citations occur from the publication year through 1999. CITSFIRM is a citation window of variable length. The window shortens as publication year approaches 1999, so that forward citations, because they are right-truncated, must eventually decline. CITSFIRM5 is the number of forward citations received from firms starting with the year of publication and including the next four years. Since CITSFIRM5 is a fixed five-year window of citations and the data end in 1995, it is not defined after 1995. CITSFIRM5 has the advantage of

being a fixed window, but it has the disadvantage of truncating citations received, especially for longer-lived papers.

CITSUNIV is the total number of forward citations received from other universities arranged by cited university, field, and year. The citations occur from the publication year through 1999. Since the citation window contracts as the publication year approaches 1999, CITSUNIV eventually declines. CITSUNIV5 is the number of forward citations received from other universities, starting with year of publication and including the next four. Since CITSUNIV5 is a fixed five-year window, it is not defined after 1995. CITSUNIV and CITSUNIV5 exclude institutional self-citations from a university to itself.

The variable PAPERS is the total number of papers written in a university, field and year. Remaining variables are fractional counterparts to the previous variables. FRPAPERS is the "fractional" version of PAPERS. The relationship between the two is this: FRPAPERS is the sum over all papers of the institutional fraction for each paper that is accounted for by the university in question, for a given field and year. To understand this, consider some examples. If Harvard writes a paper by itself, it is assigned a fraction of 1.0. If it writes a paper with Yale and IBM, it is assigned a fraction of 1/3. And if it writes a paper with MIT, Princeton, Biogen, and Merck, then it receives a fraction of 1/5. Summing over paper fractions yields FRPAPERS. By definition this is less than or equal to PAPERS.

The same idea applies to citations. This yields FRCITSFIRM, the fractional version of CITSFIRM: fractional citations received from firms on each paper, summed over papers. Clearly FRCITSFIRM is less than or equal to CITSFIRM.

The same is true of FRCITSFIRM5 and CITSFIRM5, of FRCITSUNIV and CITSUNIV, and of FRCITSUNIV5 and CITUNIV5. Our reason for offering fractional counterparts to "whole" citations is simple: fractional citations preserve totals over the entire system of universities and firms, whereas "whole" papers and citations count papers and citations multiple times and overstate totals in the system as a whole.

The fourth file in Table 1 marks a transition from universities to firms. FIRMS lists the top 200 companies and Table 5 describes its contents. The three variables are FIRMID, the ticker symbol of the firm in 1998; FIRMNAME, the name of the firm in Compustat; and SIC4, the largest four digit industry of the firm in Compustat, based on the 1987 Standard Industrial Classification (SIC) classification system.

FIRM_DESCRIPTION appears fifth. It is the analogue of file UNIVERSITY_DESCRIPTION. As Table 6 explains, the file includes three classifying variables: FIRMID, ISI88, and YEAR, as well as descriptive variables. The latter include: PAPERS, the number of papers; CITSFIRM, total forward citations received from other firms; CITSFIRM5, forward citations received from other firms in the first five years; CITSUNIV, total forward citations received from universities; and CITSUNIV5, forward citations received from universities in the first five years. Included besides are five "fractional" descriptive variables: FRPAPERS, FRCITSFIRM, FRCITSFIRM5, FRCITSUNIV, and FRCITUNIV5. Since these variables are the same as those in UNIVERSITY_DESCRIPTION, we refer the reader to the discussion of Table 4 for further details.

The file FIELDS appears sixth in Table 1. FIELDS describes the 88 Thomson-Reuters (ISI) field codes that we use throughout the database. These are known

collectively as ISI88. The table maps the detailed 88 field codes into more aggregative CASPAR NSF12 and NSF20 field codes, and describes the codes[5]. Table 7 lists the variables in FIELDS. Appendix Table A.4 records field codes and descriptive labels.

In Table 7 ISI88_DESCRIPTION labels the variable ISI88. For the 12 main NSF fields in NSF12 we provide NSF12_DESCRIPTION. And for the details of engineering and earth science fields in NSF20, we provide NSF20_DESCRIPTION. It is inevitable that an element of judgment should enter the mapping between ISI88, NSF12, and NSF20. This is due to field overlap. The point is especially pertinent for the life sciences. The major fields of biology and medicine, for example, clearly share similar scientific research. The mapping that we offer is a compromise: it assigns agricultural ISI88 fields to agriculture, basic biomedicine in ISI88 to biology, and clinical biomedicine in ISI88 to medicine. Thus for example, we interpret the ISI88 field, CGX as the fundamental biology of cancer and ONC as clinical intervention and cancer treatment. Because of this overlap, it is important to take note of the following design feature. Since the data are classified throughout in term of the detailed ISI88 fields, the user is free to pursue an alternative mapping of detailed fields into aggregates than the ones we have chosen in Table 7 and Table A.4.

The final two files in Table 1 link citing-cited and collaborating-collaborated observations. CITATION_PAIRS does this for citing-cited observations. It consists of 21,386,007 observations in six dimensions consisting of citing and cited institutions, fields, and years. The file includes ten variables. These are: citations made, or backward citations (CITATIONS), and potentially citing and cited scientific papers (PAPERSCTG, PAPERSCTD) all by citing and cited university or firm (INSTCTG, INSTCTD), citing

[5] We thank Paula Stephan for discussions concerning the mapping procedure.

and cited ISI 88 field (ISI88CTG, ISI88CTD), and citing and cited year (YEARCTG, YEARCTD). In addition, it includes a character variable, CTG_CTD, which identifies the type of citing and cited institution, university (UNV) or firm (FRM). It follows, for instance, that to select a sample of firms citing universities, one would apply the substring operator to CTG_CTD to choose CTG='FRM' and CTD='UNV'.

COLLABORATION_PAIRS performs a similar linking function for collaborating-collaborated observations. It consists of 797,348 observations in four dimensions consisted of collaborating and collaborated institutions, field and year, since collaboration occurs in the same field and year. The eight variables consist of "collaborating" and "collaborated" institutions (INSTCLBG, INSTCBD), ISI 88 field (ISI88), and year (YEAR). Included are collaborations (COLLABORATIONS), numbers of potentially collaborating and collaborated papers (PAPERSCLBG, PAPERSCLBD), and a character variable, CLBG_CLBD, which identifies type of collaborating and collaborated institution, university (UNV) or firm (FRM). So to select a sample of universities collaborating with firms, apply the substring operator to CLBG_CLBD to select CLBG='UNV' and CLBD='FRM'.

B. Distribution of Scientific Papers

Table 10 displays distributions of fractional and whole scientific papers in universities and firms[6]. This is done by the 12 main fields included in NSF12 in the National Science Foundation classification scheme. Field-specific totals and percentages are shown above; grand totals are shown in the bottom row.

[6] Recall that fractional papers are the sum of institutional fractions on all papers to which a university or firm contributes. Thus, for each paper, the fraction is 1.0 if the paper is written entirely within an institution, ½ if it is coauthored with another institution, and so on. Again, fractional papers are the sum of such fractions for a given "cell".

Notice that whole scientific papers overstate "true" scientific papers because of institutional collaborations. When we compare whole with fractional papers we see that 3,074,977 whole papers are written across all institutions and fields, but that the actual total, measured by fractional papers is 2,604,324. Thus whole papers overstate scientific "output" in the system as a whole by 18 percent (3,074,977/2,604,324≈1.18).

When we examine the university data by field, it is clear that most papers are written in the life sciences (agriculture, biology, and medicine). Together these fields account for 61.5 percent of all (fractional) papers. Second largest are papers in the natural sciences (chemistry and physics) and "technology" (computer science and engineering), which together account for another 24.8 percent. The remaining 13.7 percent of fractional university papers consists of astronomy, earth sciences, economics and business, mathematics and statistics, and psychology, which are trace elements in the universe of scientific papers. Note that percentages contributed by the fields of agriculture, chemistry, and engineering are higher among fractional than whole papers. Conversely, percentages of medicine and physics are lower among fractional papers than among whole. This is because institutional collaboration occurs less frequently than average in agriculture, chemistry, and engineering, while it occurs more frequently in medicine and physics.

Turning to the firm data, we observe a quite different distribution by field. Life sciences (agriculture, biology, and medicine) account for 31.4 percent of (fractional) industrial papers, compared with 61.5 percent in universities. Of course, the share of the life sciences in firm papers has increased in recent years, but overall the share is smaller in industry. Conversely, the share of natural sciences (chemistry and physics) and

technology (computer science and engineering) is 65.2 percent compared with 24.8 percent in universities. Observe that chemistry and engineering take up a larger share of fractional papers than whole papers, whereas the reverse is true of medicine. As before, this is because institutional collaboration is less common in industrial chemistry and engineering and more common in industrial medicine.

C. Graphical Depiction of the Papers Data

The following sections present time series graphs consisting of Figures 1-9. Time is represented in calendar years or as a lag between calendar years. In several cases, when we wish to present the data by sector or science field, graphs appear as multiples. Primarily for this reason, the following sections contain 27 separate graphs of the data. Figures 1-3 pertain to papers and are discussed in this section.

Figure 1 presents time series of university and total (university plus firm) fractional papers on the left scale, and time series of firm papers on the right scale. Scientific publishing slows down in universities and as whole in the United States after 1992, and it falls in absolute terms in industry (Adams, 2007), reflecting the downsizing and disappearance of some large industrial laboratories performing basic science research during this period, notably AT&T Bell Laboratories.

Figures 2 presents shares in all papers of science fields for universities while Figure 3 does the same for firms. Fields appear as "strata" covering seven major areas (agriculture, biology, chemistry, computer science, engineering, medicine, and physics), plus a residual "other" field category. The flatness of the strata for universities suggests little change in relative shares during 1981-1999. And yet a slight increase in the share of

biology and medicine, a decline in agriculture's share, and an increase in the engineering and physics share can be seen in the university data.

Figure 3, for industry, stands in sharp contrast to Figure 2. The share of biology and medicine doubles from less than 20 percent of industrial papers in 1981 to almost 40 percent in 1999. Computer science also gains share though from a small base. And while the share of chemistry is stable, the share of engineering and most notably physics decline strongly in this picture. The topsy-turvy nature of the shares in Figure 3 is due to the rise of industrial scientific research in pharmaceuticals and biotechnology, and the decline in large industrial laboratories that specialize in natural science and technology.

D. Graphical Depiction of the Citations Data

Figures 4-7 illustrate the citations data. Figure 4 depicts counts of scientific citations made (backward citations) and citations received (forward citations) over time. The figure covers all fields. It combines citations involving university papers on the left scale, with those involving firm papers on the right scale. Backward citations increase as time passes, because later papers have more generations of earlier papers to cite. This is partly an artifact of the left truncation of the data in 1981, since citations to papers before 1981 are eliminated. But it is partly real, reflecting the growth of fields and the growing ease of generating citations.

Citations received, or forward citations, at first grow and then decline. This is due to the combined action of three effects. As time passes the results contained in papers diffuse to readers, and this causes citations to go up. Second, the relevance of scientific research often decays, and this causes citations to decline. And third, right truncation of citations after 1999 eventually cuts citations received to zero in this window of data.

Figures 5A through 5G report the four curves in Figure 4 (citations made and received by universities and firms) separately for seven major fields. As before, the left scale refers to university citations made and received, while the right refers to firm citations made and received.

The general shape of the curves is similar across fields, but with some notable differences. The date at which citations received peak varies by sector and field. Peaking occurs sooner in sectors and fields where growth is less and where diffusion and decay are greater, and later when the reverse is true.

Figure 6 presents lagged or backward citation rate curves by the lag between citing and cited years. The figure covers all fields. To reduce complexity, the figure depicts citation curves within sectors (universities citing other universities, firms citing other firms), but the between sector curves appear very similar to those shown. The curves depict citation rates: these are citations made divided by papers that could be cited. It is thus a weighted citation rate, where weights are shares of potentially cited papers[7]. The university data are referred to the left axis, while the firm data are referred to the right. Note that both curves peak at a lag of two years. The university curve is higher because the aggregate citing population of university papers is larger relative to the number of papers that could be cited than is true of firms.

Figures 7A through 7G are graphs of the citation rates for seven major fields ranging from agriculture to physics. Consider Figure 7A. The figure shows that university citations to agriculture peaks in the third year after publication—slower than

[7] The reader needs to be aware that entire families of citation curves exist. The particular family depends on the level of aggregation, the sectors involved, and whether the data are weighted or un-weighted, so that the appearance of the curves can vary markedly. All the curves shown in this article are weighted curves that pertain to all fields within a sector or to individual fields within a sector.

average. The firm citation curve choppy and irregular and exhibits multiple peaks. This occurs because citing observations are few, especially at long lags. Other features of the field-specific diagrams are that citation rates in physics, biology, and chemistry peak more rapidly, that citation rates in technology, defined as computer science and engineering, peak more slowly, and that citation rates in rapidly peaking fields decay at a higher rate and conversely, at a lower rate for slowly peaking fields.

For computer science in universities the citation curve reaches a plateau that is almost unchanged between lags of four and ten years. At long lags the firm-firm computer science curve exhibits the same choppiness as agriculture. This occurs for the same reason, that there are relatively few industrial papers at long lags.

E. Graphical Depiction of the Collaborations Data

Figure 8 presents line graphs of institutional collaborations between universities, between universities and firms, and between firms over time. The figure covers all fields of science. The university-university curve exceeds the university-firm curve by tenfold. The height of the university-firm curve is again ten times that of the firm-firm curve. These size differences dictate the logarithmic scale on the vertical axis of the figure. Underlying these differences is the fact that academic papers are ten times as many as industrial papers. Thus, a roughly similar propensity to collaborate results in a collaboration count among university papers that is ten times the count among firm-university papers. For the same reason, university-firm collaborations are ten times the firm-firm count, where one tenth as many papers are potentially collaborating with one tenth as many papers. Figure 8 is steeply trended, with trends about the same in the

different curves, but the small number of firm-firm collaborations, some of them initiated in graduate school, suggests that such collaborations are a second order phenomenon.

Figures 9A to 9G describe similar graphs for the seven major fields of science. The field-specific curves are in the same order as for Figure 8, with relative differences depending on comparative frequency of collaboration across sectors. Figure 9D illustrates for computer science. Firm collaboration counts are more than a tenth of university counts, and thus the university-university and university-firm curves are closer than average. Also, university-firm collaborations grow more rapidly, so the curves converge over time. In Figure 9F, for medicine, the two curves are further apart than average, but converge over time. These patterns and those in other figures largely reflect field-specific publication frequency in firms and universities as well as changes in these frequencies.

III. Discussion, Comparison, and Assessment

Having described the data, we would like to assess their limitations and strengths. We can think of two important limitations. First, we would like to point out that each of the roughly 7,000 journals is assigned to a single science field. This assignment is accurate for the vast majority of specialized journals. But the method does produce serious errors for up to one percent of journals (approximately 70) that fall into Thomson-Reuters' Multidisciplinary category, some of which are highly influential. The category is treated as part of biology, because biology accounts for the largest fraction of papers. To see why the problem matters, note that Multidisciplinary journals include **Nature**, **Science**, **Proceedings of the National Academy of Sciences USA**, and **Philosophical Transactions of the Royal Society**. Clearly wholesale assignment of articles here to

biology is wrong. But to correct the problem would require article (not journal) assignment to fields. Moreover, some Multidisciplinary journals really are linked to biology. Therefore, the problem applies to less than one percent of ISI journals[8].

The main alternative to journal-field assignment is to assign papers according to perhaps multiple fields of the authors. But current practice effectively rules this out[9]. This is unlike patents, where multiple class assignments are common. To carry out such assignments would require clear criteria that would have to be acted on by a single Scientific Papers Office, much like the Patent Office does today. In the near future neither condition will be met.

A second limitation is that science citation data include publication date but not the date of first submission, or even better, the date of completion of the research. Use of publication date produces an upward bias in the observed lag between citing and cited papers. The true lag is the gap between cited publication date and citing first submission date. The extra "frictional" lag, between first submission and publication date of the citing paper, necessarily overstates the lag in scientific influence. Moreover, the problem produces greater upward biases in fields with greater frictional lags.

Science citations refer to prior literature and yet their motivations for doing so are not always clear. For example, they could measure influence of earlier ideas or seek to place limits on the problem being addressed. They could seek to refute earlier findings or constitute a strategy to raise the odds of acceptance. Of these motives, the first two seem

[8] Examples include **Bioinformatics**, **Biomaterials**, **Biometrics**, **Biometrika**, **Journal of Mathematical Biology**, **Journal of Theoretical Biology** and many others.

[9] We tested an alternative method of assignment using roughly 100,000 Harvard papers. We tried to assign each paper to one of the NSF 12 main science fields using information on authors' departmental addresses. A third of the papers could not be assigned to a field using this information, leading us to abandon the effort. More could be done on this problem, provided that across science, journals were to adopt a uniform approach to encoding fields of authors.

most likely to truly represent scientific influence. Given negative or strategic citations of the last two types, however, we must regard science citations as measuring prior influence with error[10].

One strong point of the database is that science citations are controlled by authors. While referees and editors can suggest references, including them requires authors' assent, suggesting that observed references are known to authors. In contrast, patent citations are often suggested by examiners and attorneys and are unknown to inventors.

Suppose that science citations reflect credible investments of time in searching the literature for useful knowledge. What would the earmarks of such investments be? For starters, the number of citations would set the marginal benefit of another citation equal to its marginal cost. This suggests that citations would span larger fractions of smaller disciplines, since similar marginal benefit and cost relationships across disciplines would lower the proportion of large literatures that is cited. Furthermore, literatures that require larger investments of time per cited paper would yield a lower citation rate holding size of the literature constant. Adams, Clemmons, and Stephan (2004, forthcoming 2006) find patterns very similar to those suggested above.

While voluminous, the papers and citations data are only a window on scientific research. Since they are truncated on the left and right in time, we lack most citations to papers from the late 1990s, which are not yet observed. And we know little about papers that influence research in the early 1980s since citations to these papers are left truncated and missing. The data are limited besides by sector and country, since they must have at least one author from a top 110 U.S. university or a top 200 firm. Citations made and

[10] See Jaffe, Fogarty, and Banks (1998) for an analysis that uses a set of NASA patents, as well as expert opinion on the patents, to test the validity of patent citations, answered in the affirmative, as an indicator of the importance of patents.

received by papers wholly authored in non-U.S. institutions are excluded. And so many international interactions are left out of the analysis. But the science citations and papers data described here are still a substantial improvement over much of the evidence that we have had.

IV. Conclusion

This paper has introduced a new database on academic and industrial science that covers the last two decades of the 20th century as well as a sample of the largest universities and R&D performing firms in the United States. Basic science and applied industrial research can and do overlap. It is for this very reason that the data described in this paper could make a difference to economic research devoted to the study of growth and technological progress. We sincerely hope that this is the case, and that these data will foster theoretical and empirical research into economic aspects of scientific research, both at the NBER and beyond, both now and in the future.

References

Adams, James D., "Fundamental Stocks of Knowledge and Productivity Growth," **Journal of Political Economy** 98 (August 1990): 673-702

_____, "Comparative Localization of Academic and Industrial Spillovers," **Journal of Economic Geography** 2 (July 2002): 253-278, and reprinted in Stefano Breschi and Franco Malerba, editors, **Clusters, Networks, and Innovation**, Oxford, UK: Oxford University Press, 2005

_____, "Learning, Internal Research, and Spillovers: Evidence from a Sample Of R&D Laboratories," **Economics of Innovation and New Technology** 15 (January 2006): 5-36

_____, "Recent Trends in U.S. Science and Engineering: Challenges, Implications, and Opportunities," in **Perspectives on U.S. Competitiveness in Science and Technology**, RAND Conference Proceedings, 2007

_____, Eric P. Chiang, and Katara Starkey, "Industry-University Cooperative Research Centers," **Journal of Technology Transfer** (January 2001): 73-86

Adams, James D., Grant C. Black, J. Roger Clemmons, and Paula E. Stephan, "Scientific Teams and Institutional Collaborations: Evidence from U.S. Universities, 1981-1999," **Research Policy** 34 (April 2005): 259-285

_____, J. Roger Clemmons, and Paula E. Stephan, "Standing on Academic Shoulders: Measuring Scientific Influence in Universities," Cambridge, Mass.: NBER Working Paper No. 10875, October 2004; and forthcoming in **Les Annales D'Economie et de Statistique** 79/80 (2006)

_____, "How Rapidly Does Science

Leak Out?" Cambridge, Mass.: NBER Working Paper No. 11997, January 2006

Adams, James D., and J. Roger Clemmons, "The Origins of Industrial Scientific Discoveries," NBER Working Paper No. 13823, February 2008a

_____, "Science and Industry: Tracing Basic Research through Manufacturing and Trade," **Economics of Innovation and New Technology** 17 (July 2008b): 473-495

_____, "The Growing Allocative Inefficiency of the U.S. Higher Education Sector," NBER Working Paper No. 12683, November 2006; and forthcoming in Richard B. Freeman and Daniel Goroff, editors, **Science and Engineering Careers in the U.S.**, Chicago, Illinois: University of Chicago Press for NBER, forthcoming 2009

Audretsch, David B., and Paula E. Stephan, "Company-Scientist Locational Links: The Case of Biotechnology," **American Economic Review** 86 (June 1996): 641-652

Griliches, Zvi, editor, **R&D, Patents, and Productivity**, Chicago, Illinois: University of Chicago Press for NBER, 1986

Jaffe, Adam B., "Real Effects of Academic Research," **American Economic Review**79 (December 1989): 957-970

_____, Michael S. Fogarty, and Bruce A. Banks, "Evidence from Patents and Patent Citations on the Impact of NASA and Other Federal Labs on Commercial Innovation," **Journal of Industrial Economics** XLVI (June 1998): 183-205

_____, and Manuel Trajtenberg, **Patents, Citations and Innovations: A Window on the Knowledge Economy**: Cambridge, Mass.: MIT Press, 2002

Mansfield, Edwin, "Academic Research and Industrial Innovation," **Research Policy** 20 (February 1991): 1-12

Narin, Francis, Kimberly S. Hamilton, and Dominic Olivastro, "The Increasing Linkage Between U.S. Technology and Public Science," **Research Policy** 26 (October 1997): 317-330

Scherer, F. Michael, "Inter-industry Technology Flows in the United States, **Research Policy** 11 (August 1982a): 227-245

_____, "Inter-industry Technology Flows and Productivity Growth," **Review Of Economics and Statistics** 64 (November 1982b): 627-634

Terleckyj, Nestor E., "Effects of R&D on the Productivity of Industries," (1974) Washington, DC: National Planning Association

Trajtenberg, Manuel, "A Penny for Your Quotes: Patent Citations and the Value of Innovations," **RAND Journal of Economics** 21 (1990): 172-187

Zucker, Lynn G., Michael R. Darby, and Marilynn B. Brewer, "Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises," **American Economic Review**88 (March 1998): 290-306

Figure 1--Papers of the Top 110 U.S. Universities And the Top 200 U.S. R&D Firms

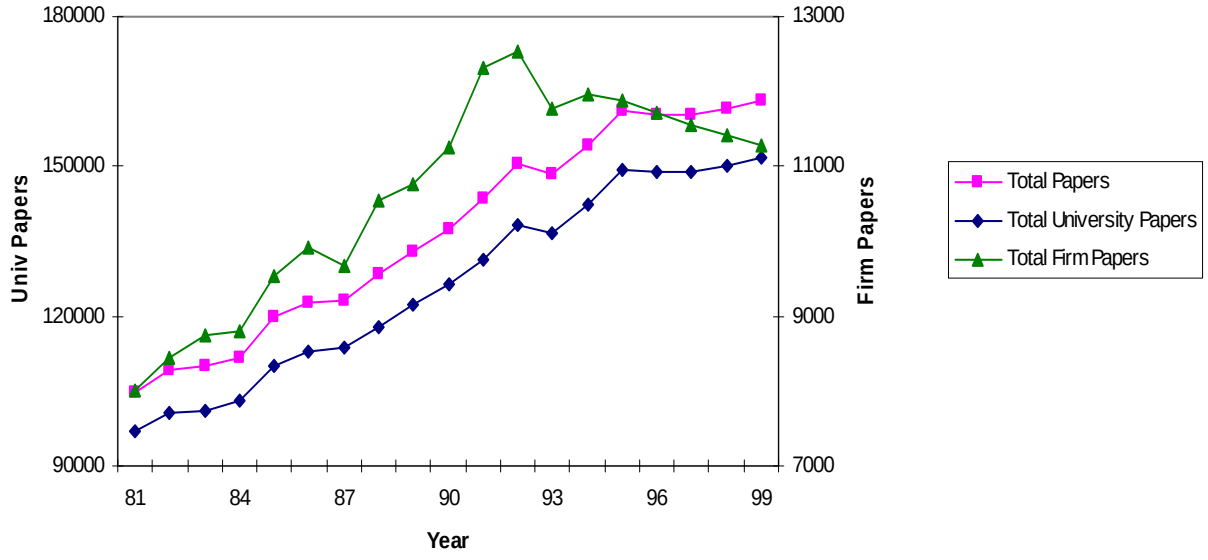
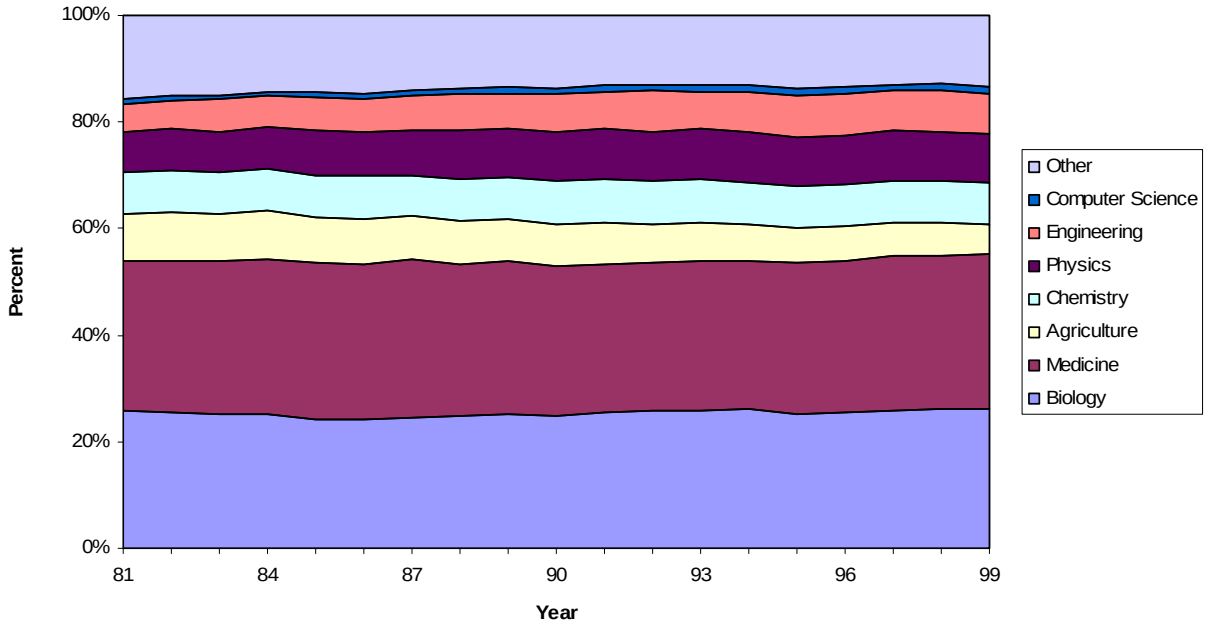


Figure 2--Shares of Fields in U.S. University Papers



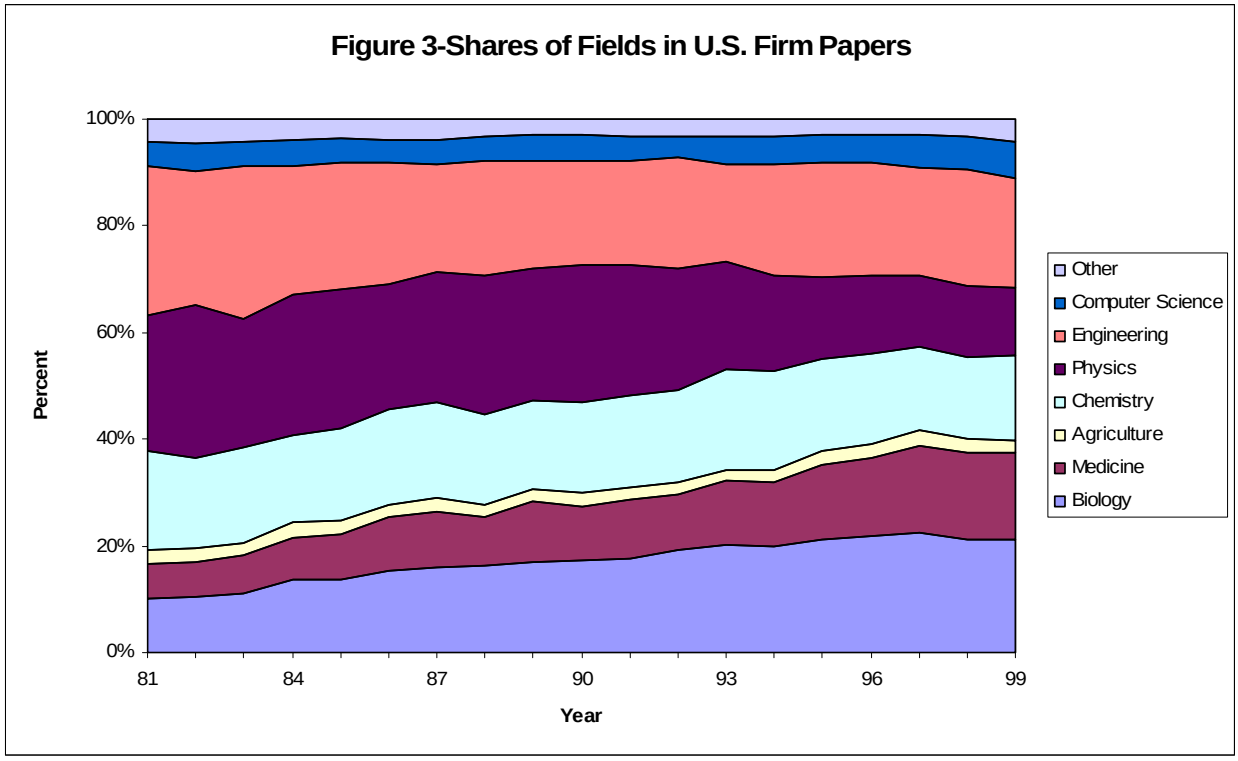


Figure 3-Shares of Fields in U.S. Firm Papers
100%
80%
60%
40%
20%
0%
Percent
81
84
87
90
93
96
99
Year
Other
Computer Science
Engineering
Physics
Chemistry
Agriculture
Medicine
Biology

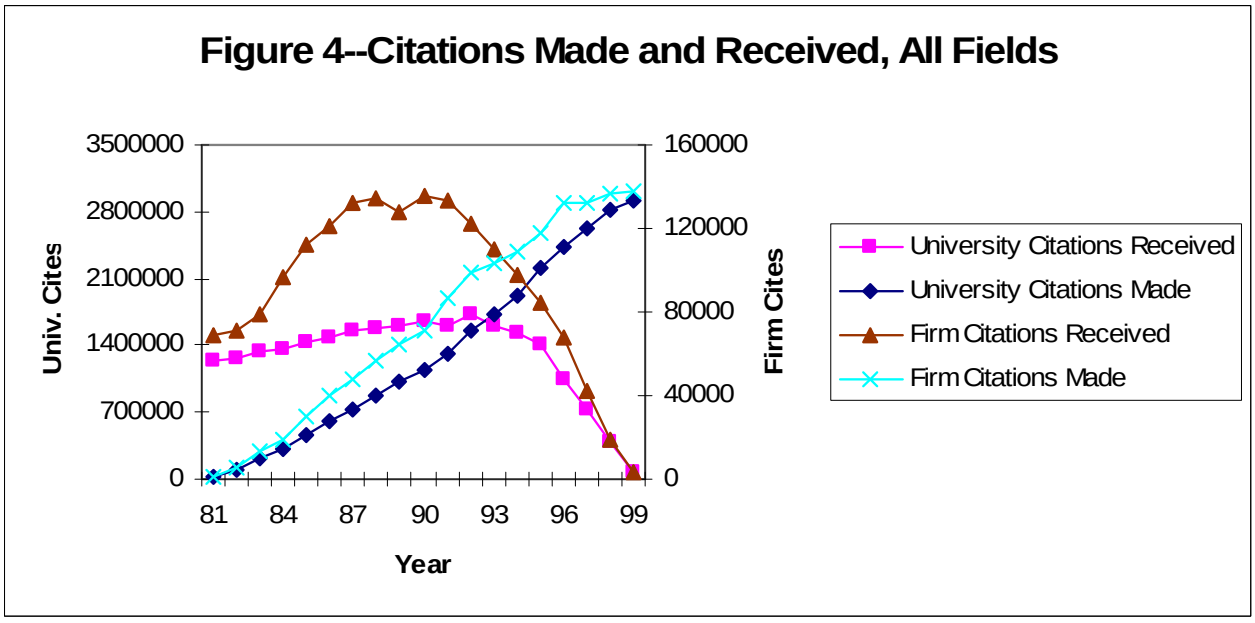


Figure 4--Citations Made and Received, All Fields
3500000
2800000
2100000
1400000
700000
0
Univ. Cites
160000
120000
80000
40000
0
Firm Cites
81
84
87
90
93
96
99
Year
University Citations Received
University Citations Made
Firm Citations Received
Firm Citations Made

Figure 5A--Citations Made and Received in Agriculture

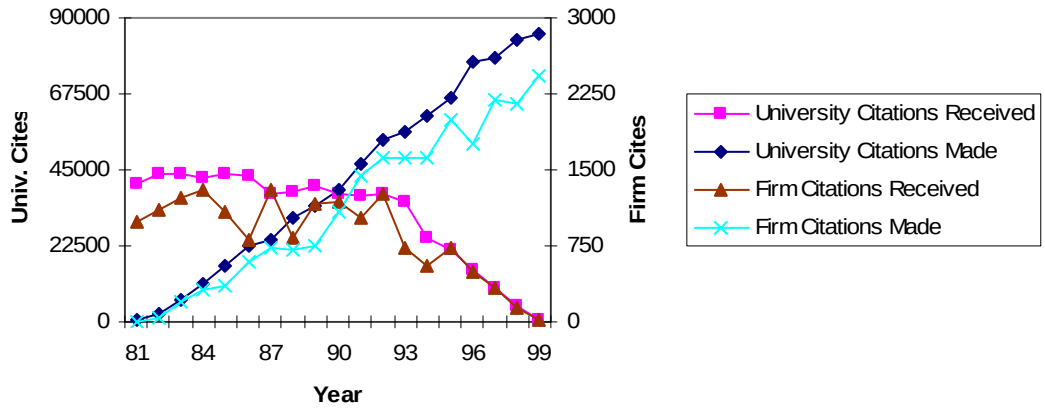


Figure 5B--Citations Made and Received in Biology

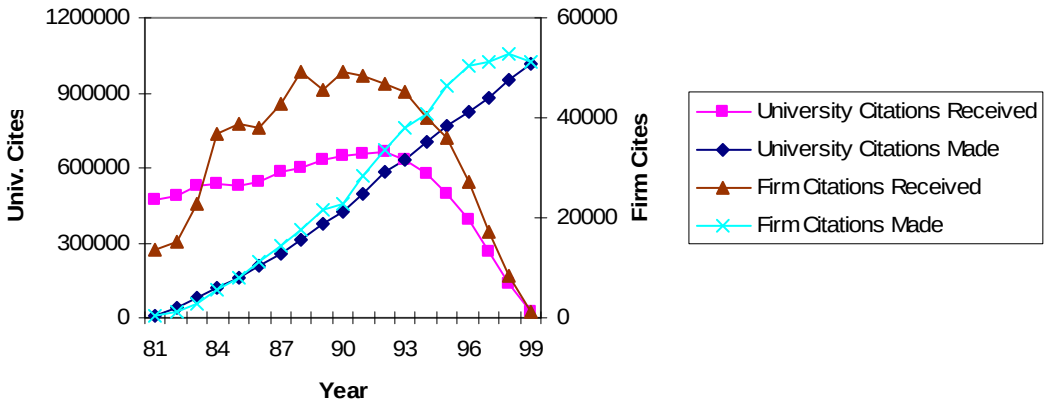


Figure 5C--Citations Made and Received in Chemistry

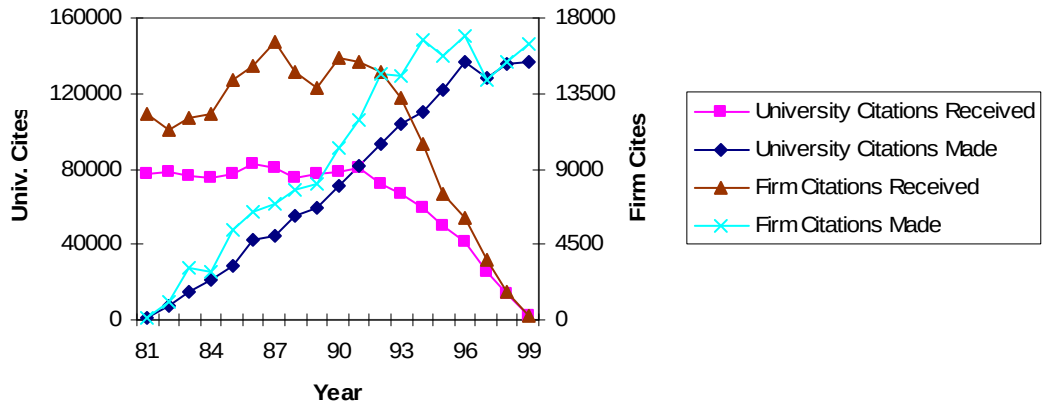


Figure 5D--Citations Made and Received in Computer Science

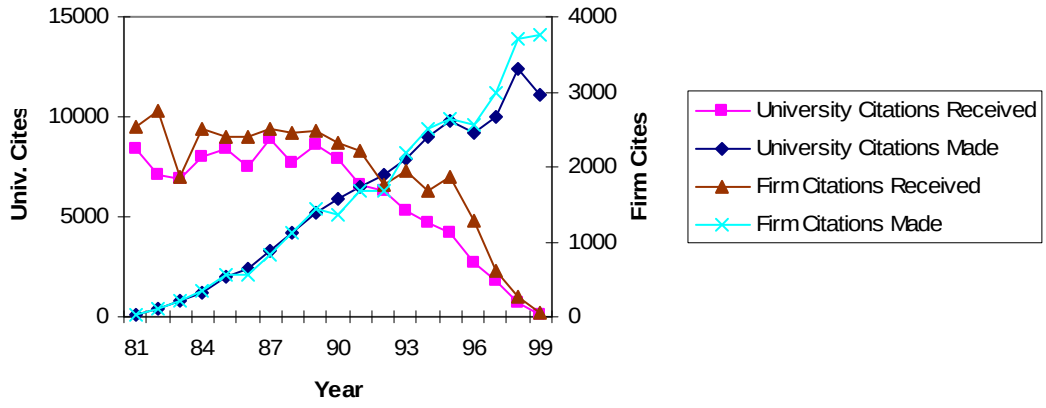


Figure 5E--Citations Made and Received in Engineering

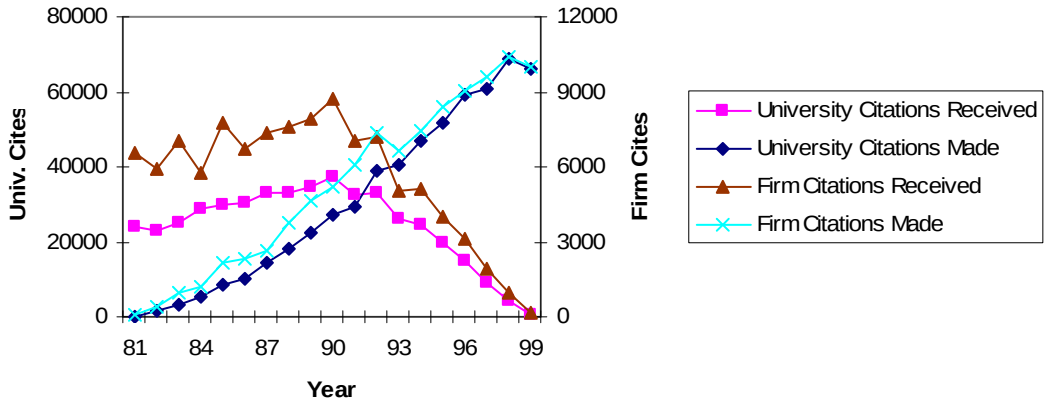
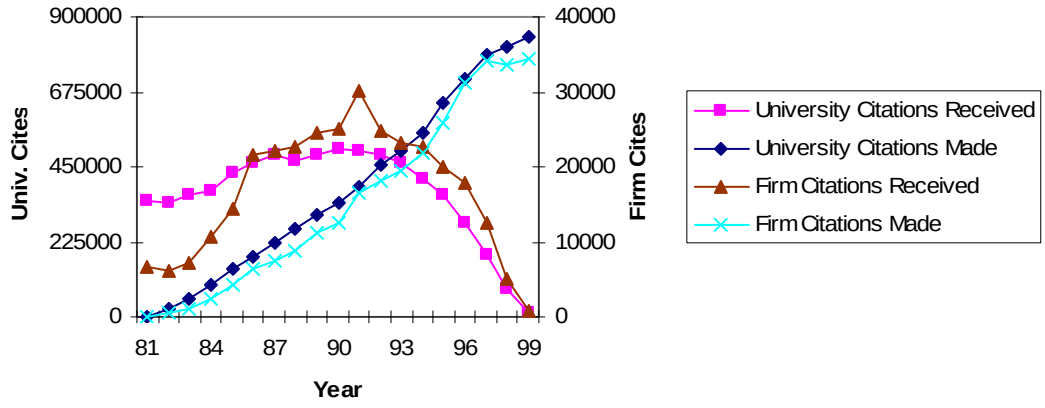


Figure 5F--Citations Made and Received in Medicine



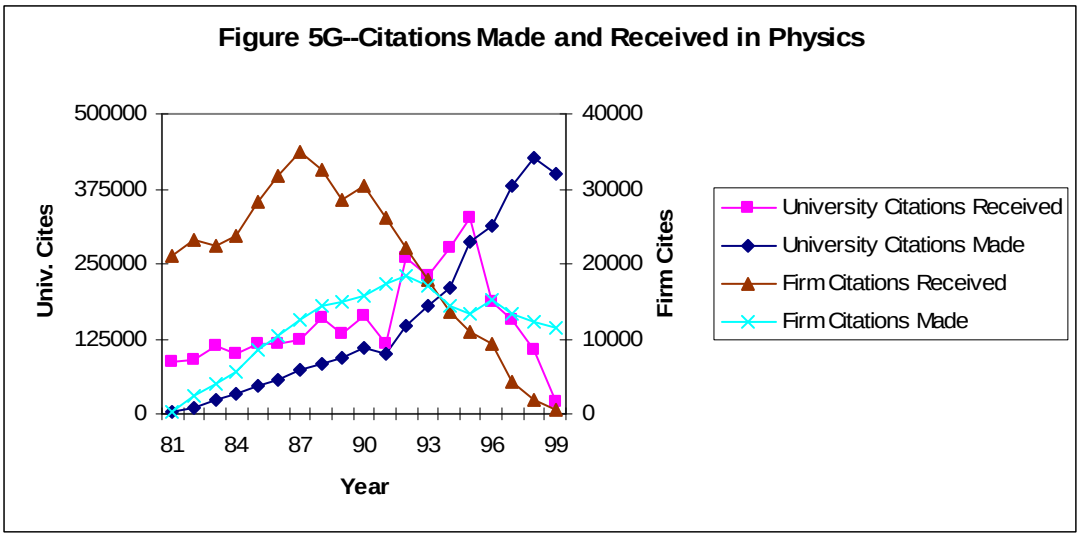


Figure 5G--Citations Made and Received in Physics
500000
375000
250000
125000
0
40000
30000
20000
10000
0
Univ. Cites
Firm Cites
81
84
87
90
93
96
99
Year
University Citations Received
University Citations Made
Firm Citations Received
Firm Citations Made

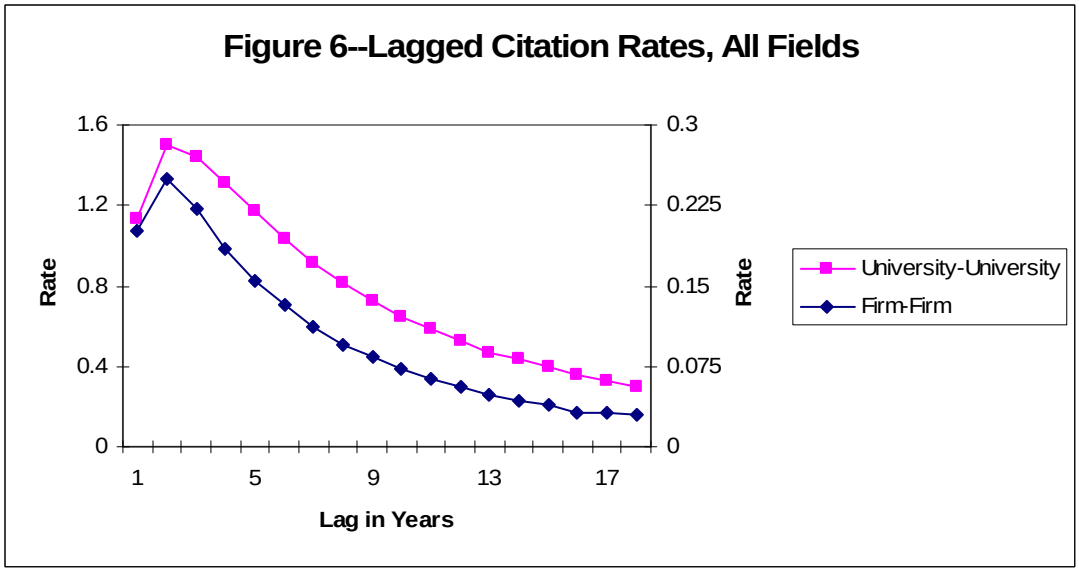


Figure 6--Lagged Citation Rates, All Fields
1.6
1.2
0.8
0.4
0
0.3
0.225
0.15
0.075
0
Rate
Rate
1
5
9
13
17
Lag in Years
University-University
Firm-Firm

Figure 7A--Lagged Citation Rates in Agriculture

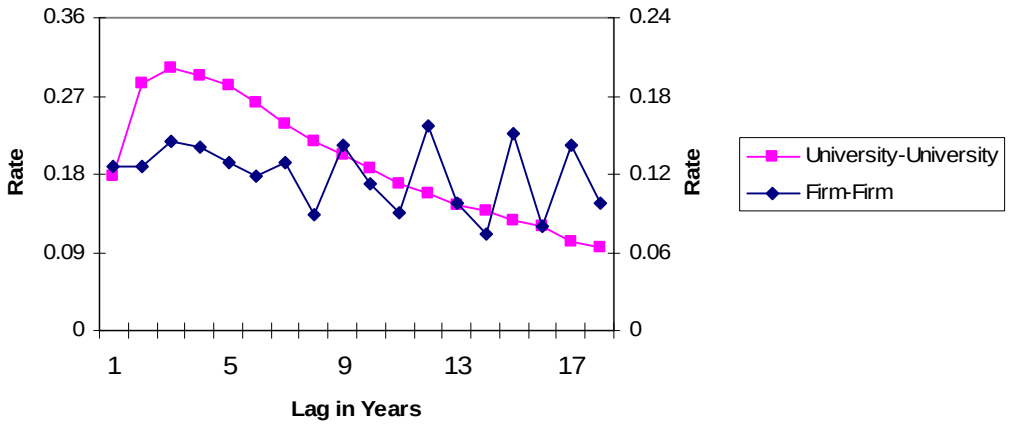


Figure 7B--Lagged Citation Rates in Biology

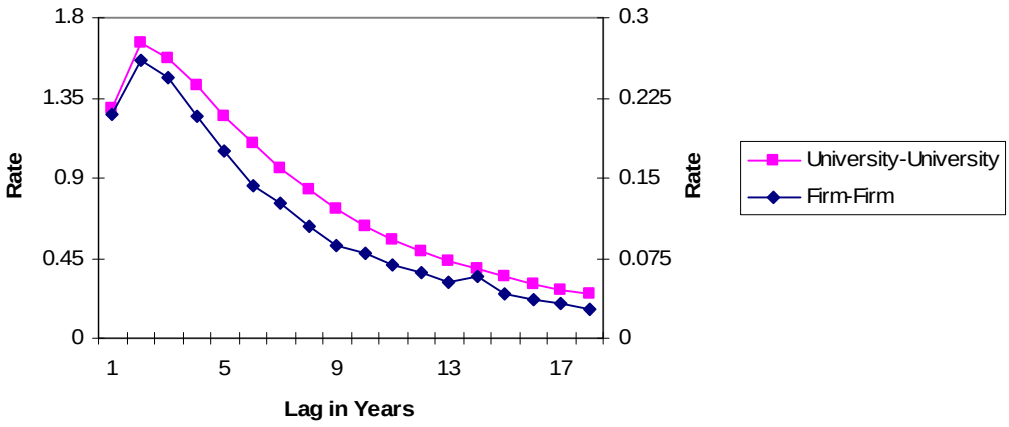
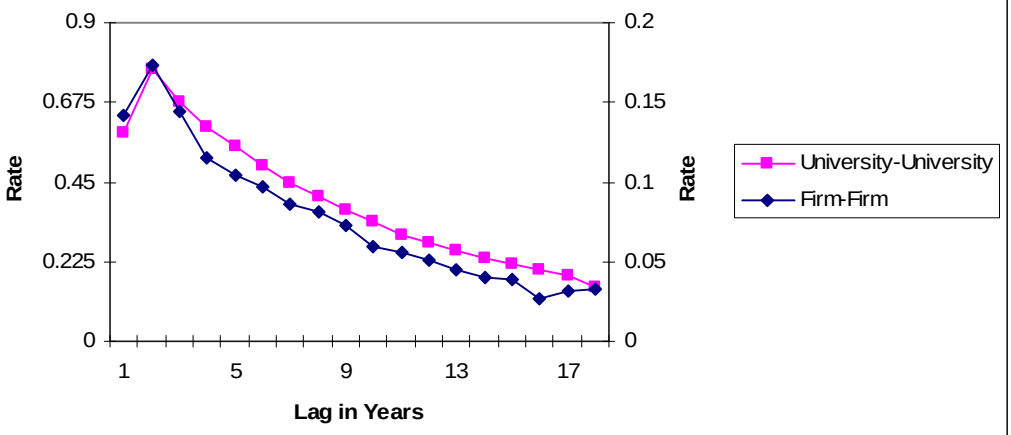


Figure 7C--Lagged Citation Rates in Chemistry



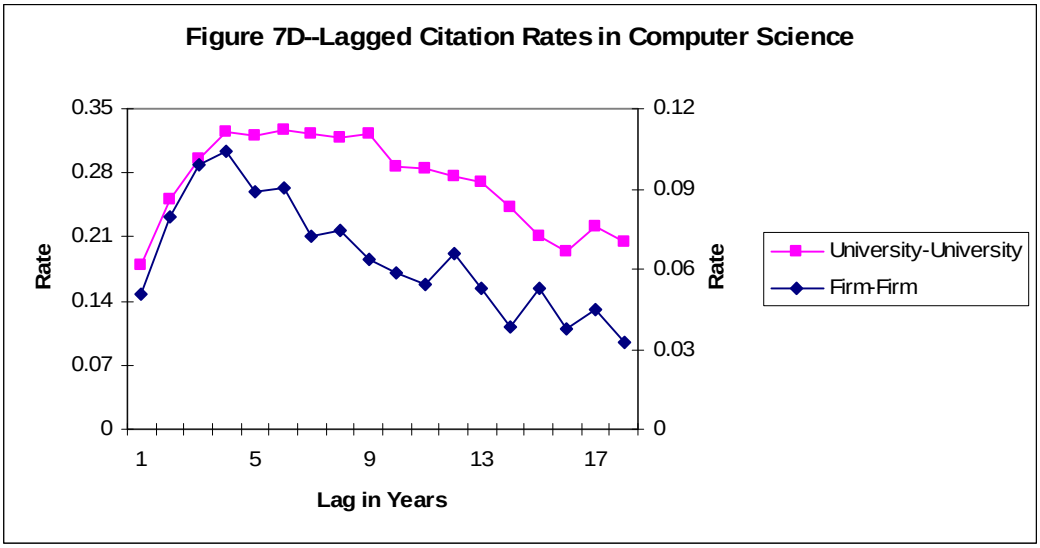


Figure 7D--Lagged Citation Rates in Computer Science
0.35
0.28
0.21
0.14
0.07
0
0.12
0.09
0.06
0.03
0
Rate
Rate
1
5
9
13
17
Lag in Years
University-University
Firm-Firm

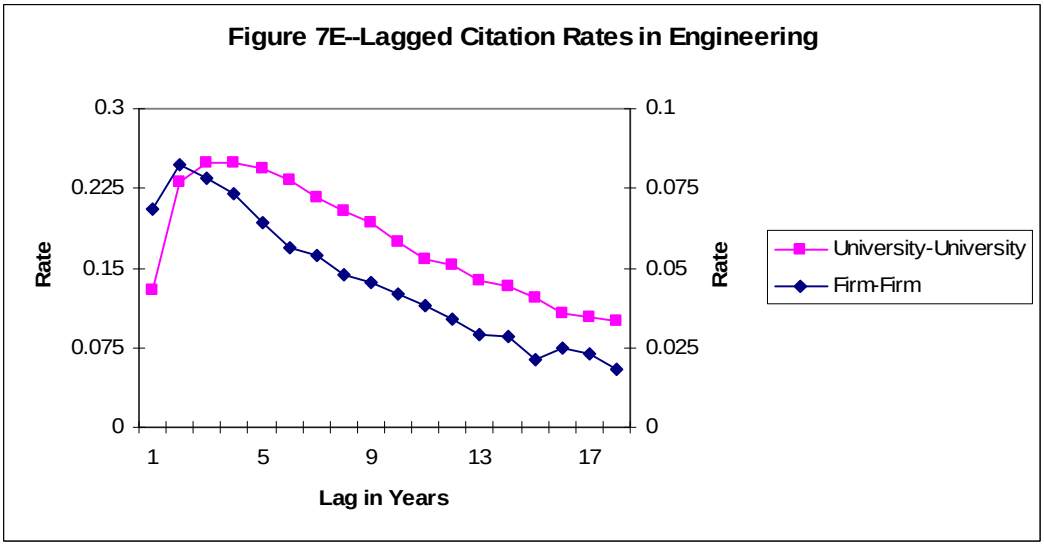


Figure 7E--Lagged Citation Rates in Engineering
0.3
0.225
0.15
0.075
0
0.1
0.075
0.05
0.025
0
Rate
Rate
1
5
9
13
17
Lag in Years
University-University
Firm-Firm

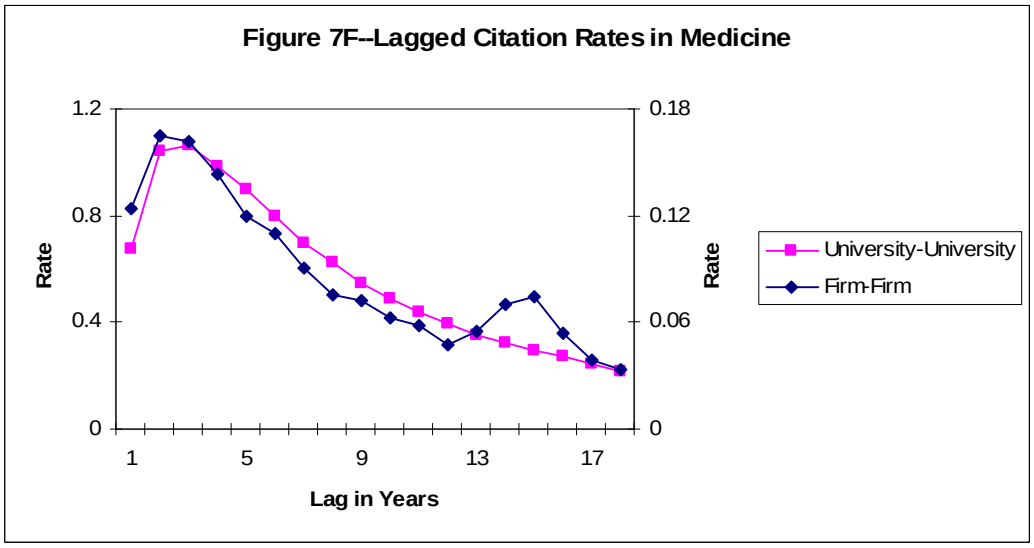


Figure 7F--Lagged Citation Rates in Medicine
1.2
0.8
0.4
0
0.18
0.12
0.06
0
Rate
Rate
1
5
9
13
17
Lag in Years
University-University
Firm-Firm

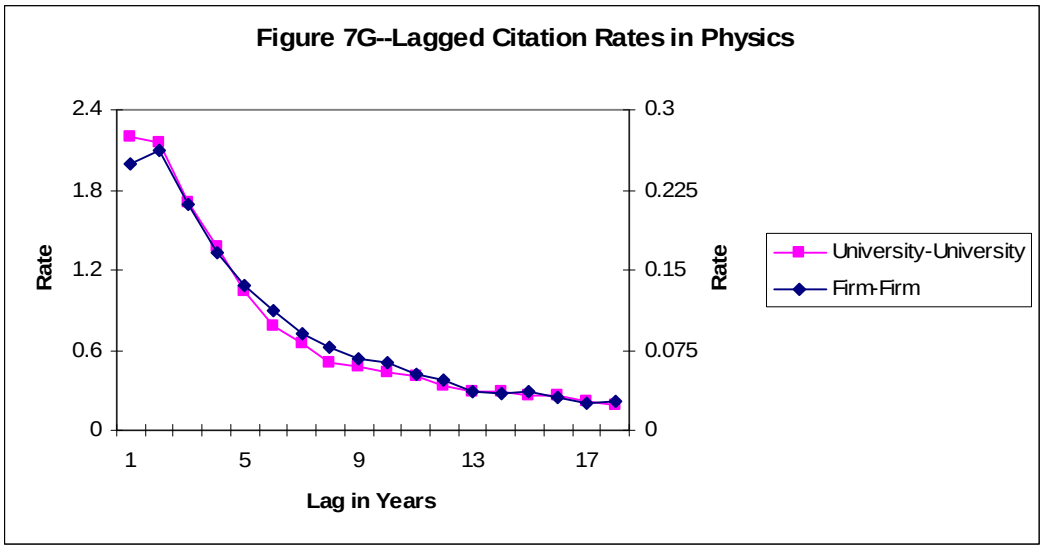


Figure 7G--Lagged Citation Rates in Physics
2.4
1.8
1.2
0.6
0
Rate
0.3
0.225
0.15
0.075
0
Rate
1
5
9
13
17
Lag in Years
University-University
Firm-Firm

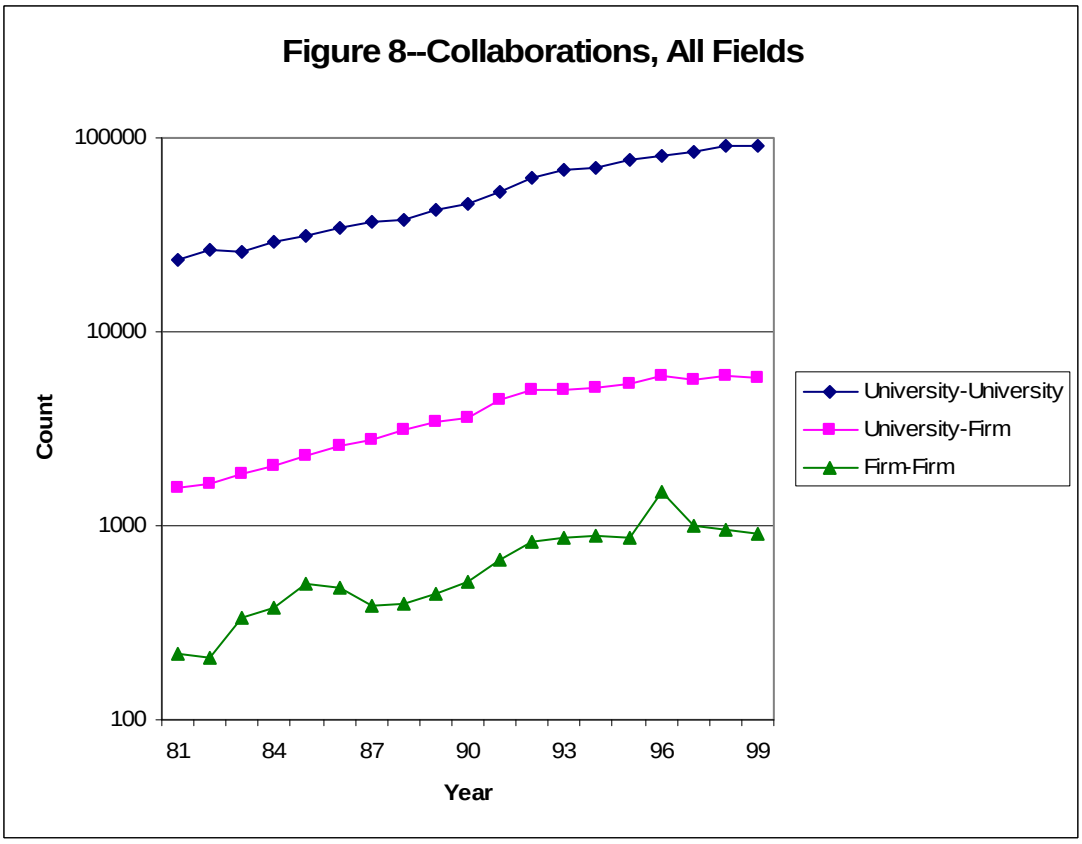


Figure 8--Collaborations, All Fields
100000
10000
1000
100
Count
81
84
87
90
93
96
99
Year
University-University
University-Firm
Firm-Firm

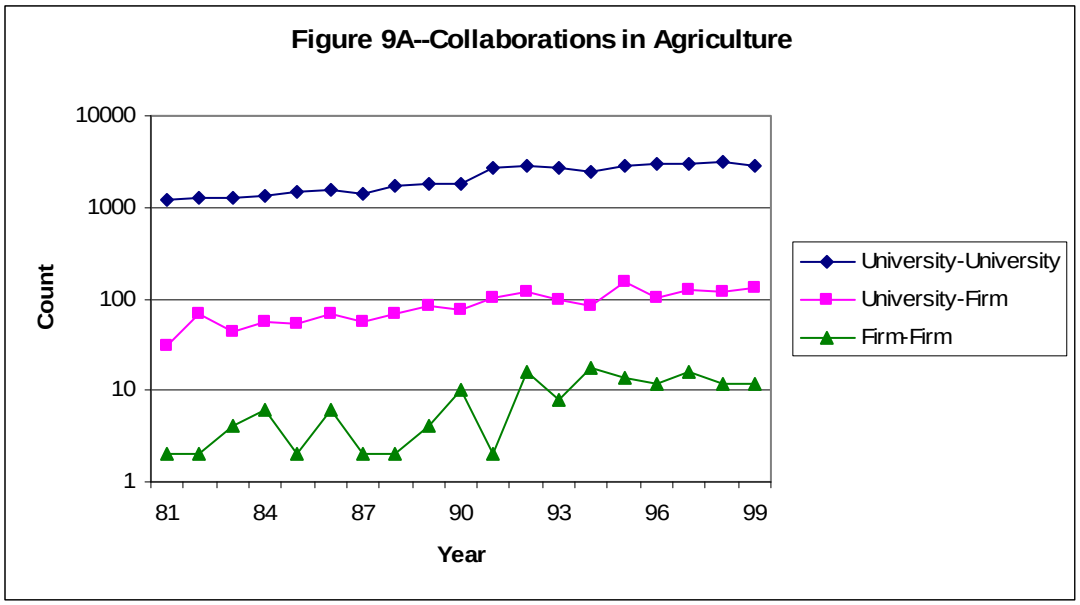


Figure 9A--Collaborations in Agriculture
10000
1000
100
10
1
Count
81
84
87
90
93
96
99
Year
University-University
University-Firm
Firm-Firm

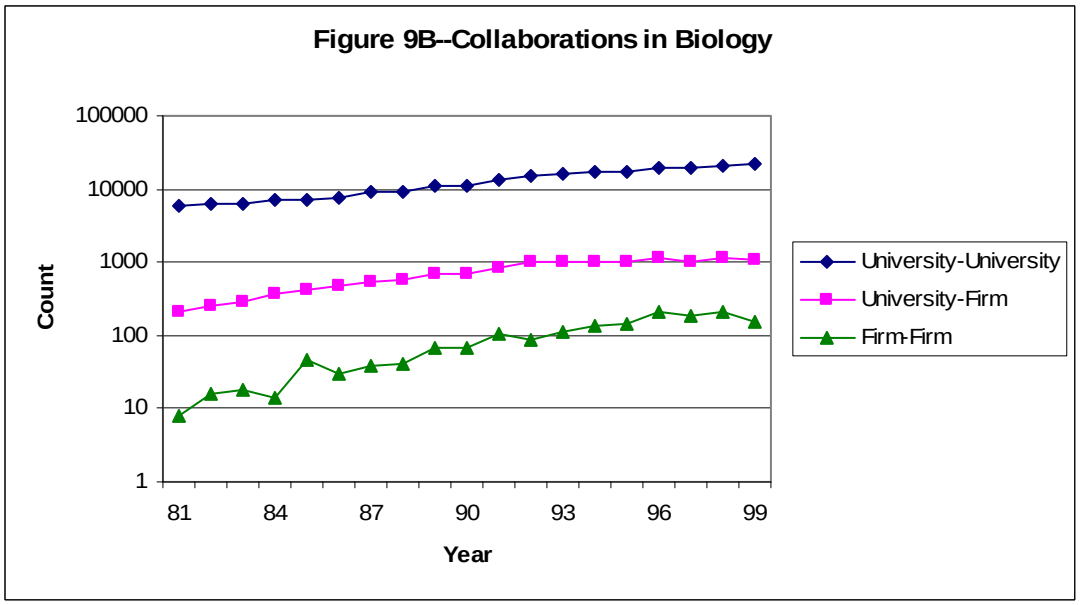


Figure 9B--Collaborations in Biology
100000
10000
1000
100
10
1
Count
81
84
87
90
93
96
99
Year
University-University
University-Firm
Firm-Firm

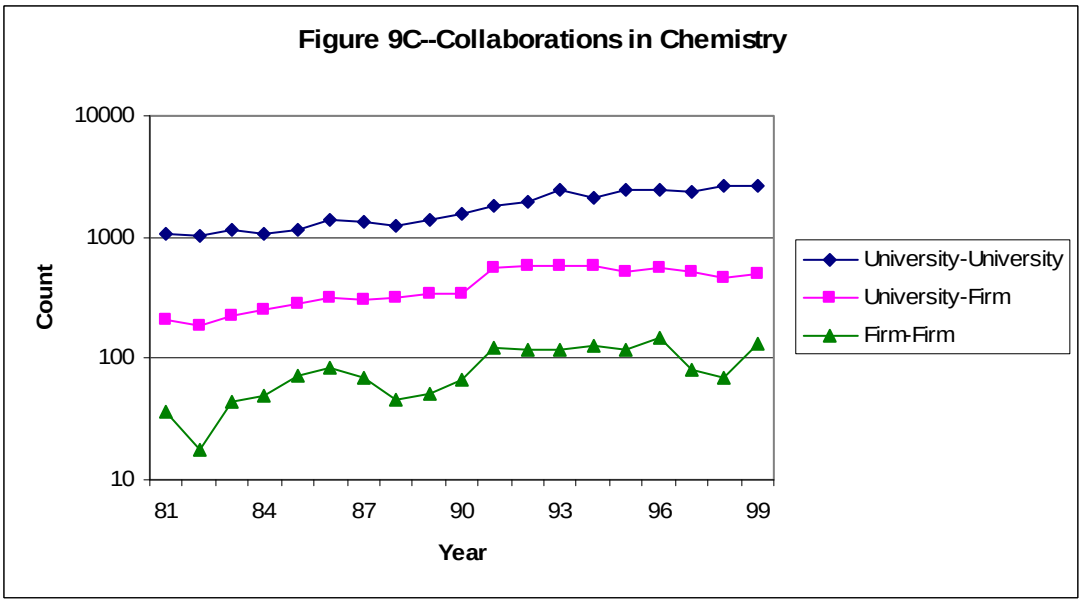


Figure 9C--Collaborations in Chemistry
10000
1000
100
10
Count
81
84
87
90
93
96
99
Year
University-University
University-Firm
Firm-Firm

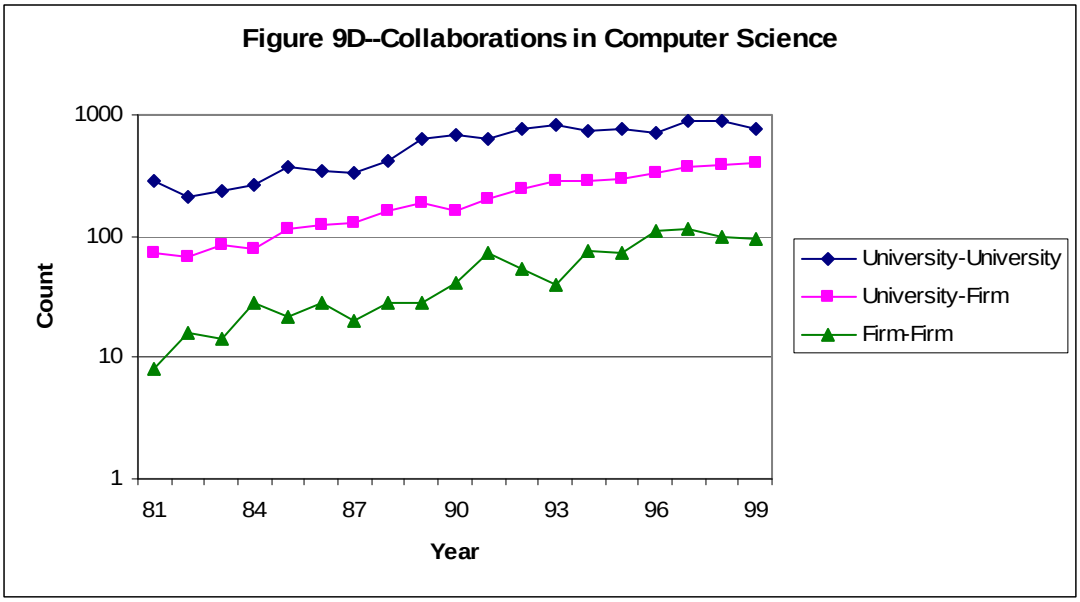


Figure 9D--Collaborations in Computer Science
1000
100
10
1
Count
81
84
87
90
93
96
99
Year
University-University
University-Firm
Firm-Firm

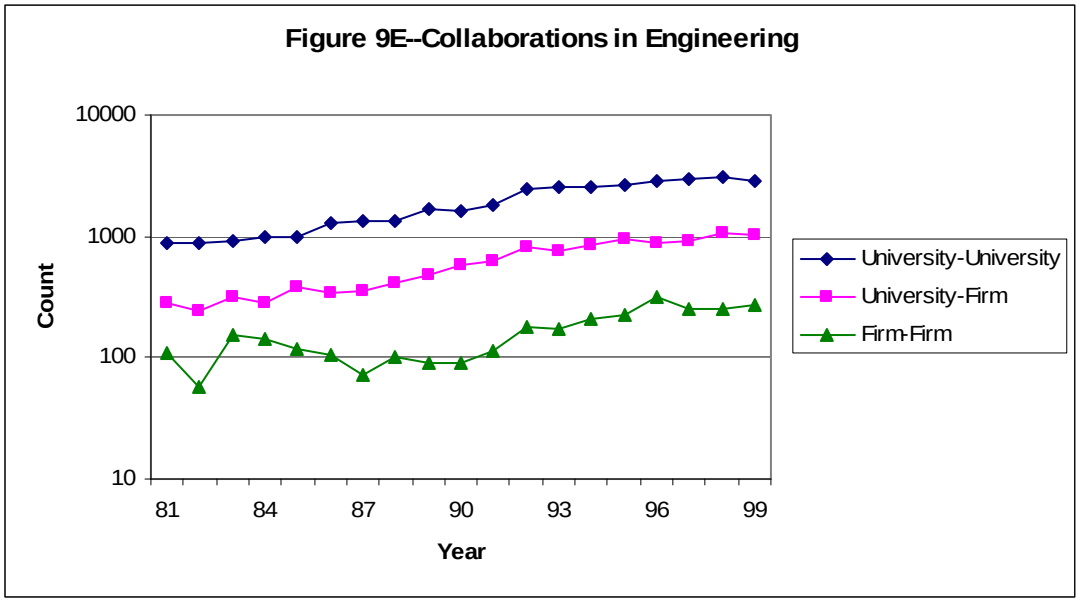


Figure 9E--Collaborations in Engineering
10000
1000
100
10
Count
81
84
87
90
93
96
99
Year
University-University
University-Firm
Firm-Firm

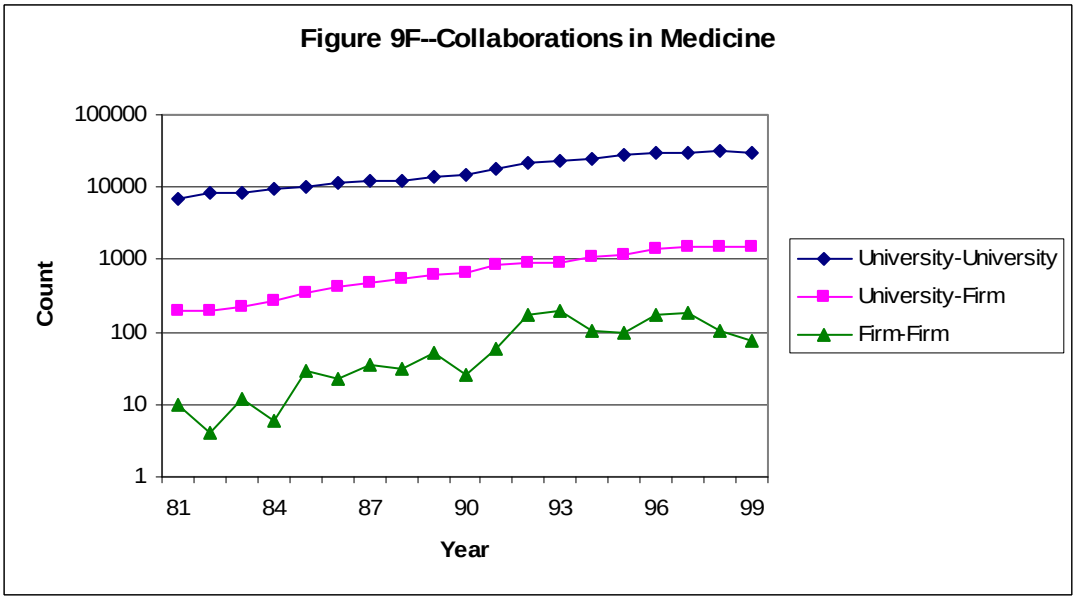


Figure 9F--Collaborations in Medicine
100000
10000
1000
100
10
1
Count
81
84
87
90
93
96
99
Year
University-University
University-Firm
Firm-Firm

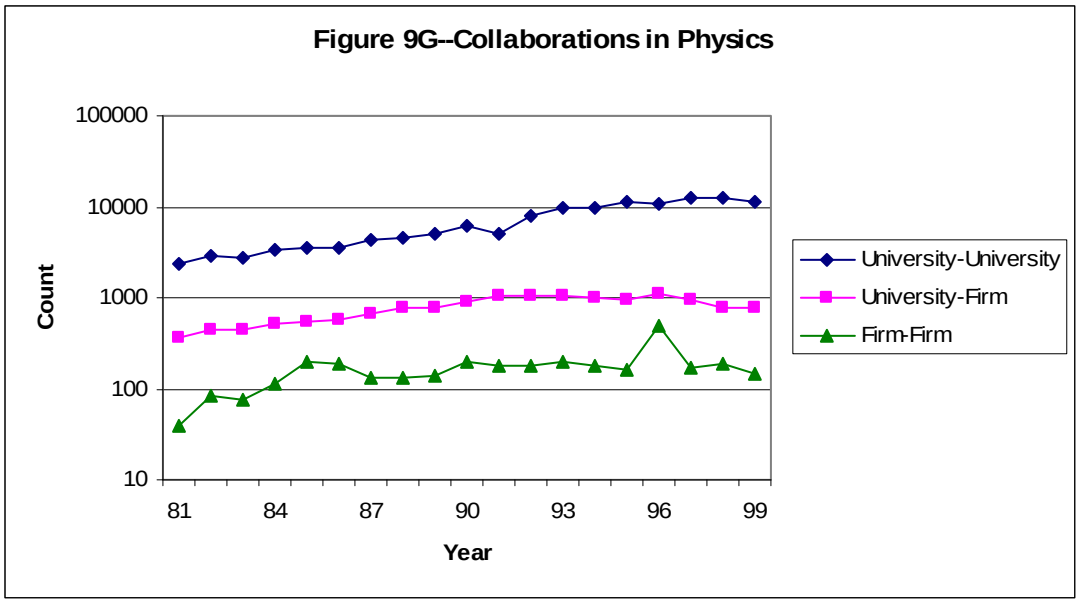


Figure 9G--Collaborations in Physics
100000
10000
1000
100
10
Count
81
84
87
90
93
96
99
Year
University-University
University-Firm
Firm-Firm

Table 1
Files of the NBER-Rensselaer Scientific Papers Database

File Name	Number of Observations	Function
UNIVERSITIES	110	Defines the set of Top 110 U.S. Universities and whether or not they belong to a multi-campus system
UNIVERSITY_SYSTEMS	142	Defines the set of branch campuses for universities that are multi-campus systems
UNIVERSITY_DESCRIPTION	143,119	Describes papers and citations received by university, Thomson-Reuters ISI 88 science field, and year [a]
FIRMS	200	Defines the set of Top 200 U.S. R&D-performing firms
FIRM_DESCRIPTION	36,689	Describes papers and citations received by firm, ISI 88 science field, and year
FIELDS	88	Defines Thomson-Reuters ISI 88 science fields and provides a cross-walk to NSF 12 main fields and NSF 20 fields [a]
CITATION_PAIRS	21,386,007	Describes science citations and numbers of citing and cited papers, by institution, Thomson-Reuters 88 science field, and year [a]
COLLABORATION_PAIRS	797,348	Describes science collaborations and numbers of collaborating and collaborated papers, by institution, Thomson-Reuters 88 science field, and year [a]

Notes: Thomson-Reuters was formerly known as the Institute for Scientific Information (ISI), hence the use of the acronym ISI to refer to the 88 relatively detailed fields used in the data.

Table 2—Contents of File UNIVERSITIES

Variable Name	Label	Format
STANDALONE	1 if stand alone campus, 0 if part of a multi-campus system	Numeric
UNIVID	Modified Federal FICE Code [a]	Character
UNIVNAME	University Name	Character

Notes: Here and in the following tables the FICE Code refers to the Federal Interagency Committee on Education Code that identifies a university.

Table 3—Contents of File UNIVERSITY_SYSTEMS

Variable Name	Label	Format
BRANCH	Name of Branch Campus	Character
UNIVID	Modified Federal FICE Code	Character
UNIVNAME	University Name	Character

Table 4—Contents of File UNIVERSITY_DESCRIPTION

Variable Name	Label	Format
CITSFIRM	Total Forward Citations from Firms	Numeric
CITSFIRM5	Forward Citations from Firms, Five Years	Numeric
CITSUNIV	Total Forward Citations from Other Universities	Numeric
CITSUNIV5	Forward Citations from Other Universities, Five Years	Numeric
FRCITSFIRM	Fractional CITSFIRM	Numeric
FRCITSFIRM5	Fractional CITSFIRM5	Numeric
FRCITSUNIV	Fractional CITSUNIV	Numeric
FRCITSUNIV5	Fractional CITSUNIV5	Numeric
FRPAPERS	Fractional PAPERS of a Firm	Numeric
ISI88	Thomson-Reuters ISI 88 Field Code	Character
PAPERS	Number of Papers	Numeric
UNIVID	Modified Federal FICE Code	Character
YEAR	Year	Character

Table 5—Contents of File FIRMS

Variable Name	Label	Format
FIRMID	1998 Ticker Symbol of a Firm	Character
FIRMNAME	Firm Name	Character
SIC4	4 Digit 1987 SIC Code	Numeric

Table 6—Contents of File FIRM_DESCRIPTION

Variable Name	Label	Format
CITSFIRM	Total Forward Citations from Other Firms	Numeric
CITSFIRM5	Forward Citations from Other Firms, Five Years	Numeric
CITSUNIV	Total Forward Citations from Universities	Numeric
CITSUNIV5	Forward Citations from Universities, Five Years	Numeric
FIRMID	1998 Ticker Symbol of Firm	Character
FRCITSFIRM	Fractional CITSFIRM	Numeric
FRCITSFIRM5	Fractional CITSFIRM5	Numeric
FRCITSUNIV	Fractional CITSUNIV	Numeric
FRCITSUNIV5	Fractional CITSUNIV5	Numeric
FRPAPERS	Fractional PAPERS of a Firm	Numeric
ISI88	Thomson-Reuters ISI 88 Field Code	Character
PAPERS	Number of Papers	Numeric
YEAR	Year	Character

Table 7—Contents of File FIELDS

Variable Name	Label	Format
ISI88_DESCRIPTION	Thomson-Reuters ISI 88 Field Description	Character
NSF12	NSF-CASPAR 12 Field Code	Character
NSF20	NSF-CASPAR 20 Field Code	Character
NSF12_DESCRIPTION	NSF-CASPAR 12 Field Description	Character
NSF20_DESCRIPTION	NSF-CASPAR 20 Field Description	Character
ISI88	Thomson-Reuters ISI 88 Field Code	Character

Table 8—Contents of File CITATION_PAIRS

Variable Name	Label	Format
CITATIONS	Number of Citations from Citing to Cited	Numeric
CTG_CTD	Citing and Cited Type, UNV or FRM	Character
INSTCTD	Cited UNIVID or FIRMID	Character
INSTCTG	Citing UNIVID or FIRMID	Character
ISI88CTD	Cited Thomson-Reuters ISI 88 Field Code	Character
ISI88CTG	Citing Thomson-Reuters ISI 88 Field Code	Character
PAPERSCTD	Potentially Cited Number of Papers	Numeric
PAPERSCTG	Potentially Citing Number of Papers	Numeric
YEARCTD	Cited Year of Publication	Character
YEARCTG	Citing Year of Publication	Character

Table 9—Contents of File COLLABORATION_PAIRS

Variable Name	Label	Format
CLBG_CLBD	Collaborating and Collaborated Type, UNV or FRM	Character
COLLABORATIONS	Number of Collaborations	Numeric
INSTCLBD	Collaborated UNIVID or FIRMID	Character
INSTCLBG	Collaborating UNIVID or FIRMID	Character
ISI88	Thomson-Reuters ISI 88 Field Code	Character
PAPERSCLBD	Potentially Collaborated Number of Papers	Numeric
PAPERSCLBG	Potentially Collaborating Number of Papers	Numeric
YEAR	Year of Publication	Character

Table 10
Distribution of Fractional and Whole Papers, 1981-1999
Universities and Firms
(Column Percentages in Parentheses)

NSF 12 Main Science Field	Top 110 Universities		Top 200 R&D Firms	
	Fractional Papers	Whole Papers	Fractional Papers	Whole Papers
Agriculture	180,427 (7.5%)	199,045 (7.0%)	4,927 (2.4%)	5,720 (2.4%)
Astronomy	35,534 (1.5%)	47,593 (1.7%)	777 (0.4%)	1,128 (0.5%)
Biology	609,732 (25.4%)	717,213 (25.3%)	35,506 (17.6%)	42,053 (17.7%)
Chemistry	190,108 (7.9%)	208,604 (7.4%)	34,682 (17.2%)	38,820 (16.3%)
Computer Science	26,647 (1.1%)	32,821 (1.2%)	10,258 (5.1%)	12,367 (5.2%)
Earth Sciences	72,541 (3.0%)	88,306 (3.1%)	2,799 (1.4%)	3,616 (1.5%)
Economics and Business	43,767 (1.8%)	53,292 (1.9%)	485 (0.2%)	642 (0.3%)
Engineering	167,191 (7.0%)	188,100 (6.6%)	43,883 (21.7%)	50,356 (21.1%)
Mathematics and Statistics	59,739 (2.5%)	70,730 (2.5%)	1,994 (1.0%)	2,574 (1.1%)
Medicine	686,459 (28.6%)	821,534 (29.0%)	23,039 (11.4%)	29,200 (12.3%)
Physics	212,414 (8.8%)	267,697 (9.4%)	42,791 (21.2%)	50,588 (21.2%)
Psychology	117,695 (4.9%)	141,765 (5.0%)	932 (0.5%)	1,213 (0.5%)
All Fields	2,402,255 (100%)	2,836,700 (100%)	202,069 (100%)	238,277 (100%)

Appendix:
Universities, Firms, and Fields

Table A1
The Top 110 U.S. Universities

University Name	Modified Federal FICE Code	Observation
ARIZONA STATE UNIV	1081	1
UNIV ARIZONA	1083	2
CALTECH	1131	3
UNIV TEXAS HOUSTON HLTH SCI CTR	11618	4
STANFORD UNIV	1305	5
UNIV CALIF BERKELEY	1312	6
UNIV CALIF DAVIS	1313	7
UNIV CALIF IRVINE	1314	8
UNIV CALIF LOS ANGELES	1315	9
UNIV CALIF RIVERSIDE	1316	10
UNIV CALIF SAN DIEGO	1317	11
UNIV CALIF SAN FRANCISCO	1319	12
UNIV CALIF SANTA BARBARA	1320	13
UNIV CALIF SANTA CRUZ	1321	14
UNIV SO CALIF	1328	15
COLORADO STATE UNIV	1350	16
YALE UNIV	1426	17
UNIV DELAWARE	1431	18
GEORGETOWN UNIV	1445	19
FLORIDA STATE UNIV	1489	20
UNIV FLORIDA	1535	21
UNIV MIAMI	1536	22
EMORY UNIV	1564	23
UNIV GEORGIA	1598	24
UNIV HAWAII	1610	25
LOYOLA UNIV	1710	26
NORTHWESTERN UNIV	1739	27
UNIV CHICAGO	1774	28
UNIV ILLINOIS URBANA	1775	29
UNIV ILLINOIS CHICAGO	1776	30
IOWA STATE UNIV	1869	31
UNIV IOWA	1892	32
TULANE UNIV	2029	33
JOHNS HOPKINS UNIV	2077	34
UNIV MARYLAND COLLEGE PARK	2103	35

Table A1
The Top 110 U.S. Universities

University Name	Modified Federal FICE Code	Observation
UNIV MARYLAND BALTIMORE	2104	36
BOSTON UNIV	2130	37
BRANDEIS UNIV	2133	38
HARVARD UNIV	2155	39
MIT	2178	40
TUFTS UNIV	2219	41
WOODS HOLE OCEANOG INST	2230	42
MICHIGAN STATE UNIV	2290	43
WAYNE STATE UNIV	2329	44
WASHINGTON UNIV	2520	45
DARTMOUTH COLL	2573	46
UNIV NEW HAMPSHIRE	2589	47
PRINCETON UNIV	2627	48
UNIV NEW MEXICO	2663	49
NEW YORK UNIV	2785	50
ROCKEFELLER UNIV	2807	51
UNIV ROCHESTER	2894	52
YESHIVA UNIV	2903	53
UNIV ALASKA	29094	54
DUKE UNIV	2920	55
N CAROLINA STATE UNIV	2972	56
UNIV N CAROLINA CHAPEL HILL	2974	57
WAKE FOREST UNIV	2978	58
CASE WESTERN RESERVE UNIV	3024	59
OREGON STATE UNIV	3210	60
UNIV OREGON	3223	61
CARNEGIE MELLON UNIV	3242	62
LEHIGH UNIV	3289	63
UNIV PENN	3378	64
BROWN UNIV	3401	65
VANDERBILT UNIV	3535	66
RICE UNIV	3604	67
UNIV TEXAS AUSTIN	3658	68
UNIV TEXAS SAN ANTONIO HLTH SCI CTR	3659	69
UNIV TEXAS SOUTHWESTERN MED CTR DALLAS	3660	70
UNIV UTAH	3675	71
UTAH STATE UNIV	3677	72
UNIV VERMONT	3696	73
VIRGINIA COMMONWEALTH UNIV	3735	74
VIRGINIA POLYTECH INST	3754	75
UNIV WASHINGTON	3798	76

Table A1
The Top 110 U.S. Universities

University Name	Modified Federal FICE Code	Observation
WASHINGTON STATE UNIV	3800	77
W VIRGINIA UNIV	3827	78
UNIV WISCONSIN MADISON	3895	79
OREGON HLTH SCI UNIV	4882	80
BAYLOR COLL MED	4949	81
NEW MEXICO STATE UNIV	8773	82
UNIV CINCINNATI	8805	83
SUNY STONY BROOK	9555	84
UNIV ALABAMA	X1051	85
UNIV MISSOURI	X2515	86
CUNY	X2686	87
TEXAS A&M UNIV	X3632	88
UNIV VIRGINIA	X3745	89
COLUMBIA UNIV	X7963	90
UNIV NEBRASKA	X8025	91
UNIV TENNESSEE	X8051	92
UNIV COLORADO	X8717	93
UNIV CONNECTICUT	X8718	94
GEORGIA INST TECHNOL	X8723	95
INDIANA UNIV	X8731	96
PURDUE UNIV	X8732	97
UNIV KENTUCKY	X8744	98
LOUISIANA STATE UNIV	X8745	99
UNIV MASSACHUSETTS	X8755	100
UNIV MINNESOTA	X8761	101
RUTGERS STATE UNIV	X8771	102
CORNELL UNIV	X8779	103
SYRACUSE UNIV	X8789	104
OHIO STATE UNIV	X8802	105
PENN STATE UNIV	X8813	106
UNIV PITTSBURGH	X8815	107
UNIV KANSAS	X9001	108
UNIV MICHIGAN	X9091	109
SUNY BUFFALO	X9554	110

Table A2
The Top 200 U.S. R&D Firms in 1998

Firm Name	Firmid [a]	Observation
ALCOA INC	AA	1
APPLE COMPUTER INC	AAPL	2
ABBOTT LABORATORIES	ABT	3
ADOBE SYSTEMS INC	ADBE	4
ADC TELECOMMUNICATIONS INC	ADCT	5
ANALOG DEVICES	ADI	6
ADAPTEC INC	ADPT	7
AUTODESK INC	ADSK	8
ALLERGAN INC	AGN	9
AMERICAN HOME PRODUCTS CORP	AHP	10
ALLIEDSIGNAL INC	ALD	11
APPLIED MATERIALS INC	AMAT	12
ADVANCED MICRO DEVICES	AMD	13
AMGEN INC	AMGN	14
AMP INC	AMP	15
AMERICA ONLINE INC	AOL	16
AIR PRODUCTS & CHEMICA LS INC	APD	17
APPLIED MAGNETICS CORP	APM	18
AMERN STANDARD CO INC	ASD	19
ASCEND COMMUNICATIONS INC	ASND	20
ATMEL CORP	ATML	21
AUTOMATIC DATA PROCESSING	AUD	22
AVID TECHNOLOGY INC	AVID	23
BOEING CO	BA	24
BAXTER INTERNATIONAL INC	BAX	25
BRUNSWICK CORP	BC	26
BARD (C.R.) INC	BCR	27
BLACK & DECKER CORP	BDK	28
BECTON DICKINSON & CO	BDX	29
BEA SYSTEMS INC	BEAS	30
BECKMAN COULTER INC	BEC	31
BELL COMMUNICATIONS RESEARCH INC [b]	BELLC	32
BIOGEN INC	BGEN	33
BAKER-HUGHES INC	BHI	34
BMC SOFTWARE INC	BMCS	35
BRISTOL MYERS SQUIBB	BMY	36
BAUSCH & LOMB INC	BOL	37
BOSTON SCIENTIFIC CORP	BSX	38
COMPUTER ASSOCIATES INTL INC	CA	39
CATERPILLAR INC	CAT	40

Table A2
The Top 200 U.S. R&D Firms in 1998

Firm Name	Firmid [a]	Observation
CADENCE DESIGN SYS INC	CDN	41
CERIDIAN CORP	CEN	42
CHIRON CORP	CHIR	43
CHEVRON CORP	CHV	44
COLGATE-PALMOLIVE CO	CL	45
COMVERSE TECHNOLOGY INC	CMVT	46
CENTOCOR INC	CNTO	47
COMPAQ COMPUTER CORP	CPQ	48
CIRRUS LOGIC INC	CRUS	49
CABLETRON SYSTEMS	CS	50
CISCO SYSTEMS INC	CSCO	51
CUMMINS ENGINE	CUM	52
CONVERGYS CORP	CVG	53
CYPRESS SEMICONDUCTOR CORP	CY	54
DANA CORP	DCN	55
DU PONT (E I) DE NEMOU RS	DD	56
DETROIT DIESEL CORP	DDC	57
DEERE & CO	DE	58
DELL COMPUTER CORP	DELL	59
DEXTER CORP	DEX	60
DATA GENERAL CORP	DGN	61
DANAHER CORP	DHR	62
GENENTECH INC	DNA	63
DOVER CORP	DOV	64
DOW CHEMICAL	DOW	65
DELPHI AUTOMOTIVE SYS CORP	DPH	66
EASTMAN KODAK CO	EK	67
EMC CORP/MA	EMC	68
EASTMAN CHEMICAL CO	EMN	69
EMERSON ELECTRIC CO	EMR	70
ELECTRONIC ARTS INC	ERTS	71
EATON CORP	ETN	72
FORD MOTOR CO	F	73
FMC CORP	FMC	74
FEDERAL-MOGUL CORP	FMO	75
GILLETTE CO	G	76
GENERAL DYNAMICS CORP	GD	77
GENERAL ELECTRIC CO	GE	78
GENZYME GENERAL	GENZ	79
CORNING INC	GLW	80
GENERAL MOTORS CORP	GM	81

Table A2
The Top 200 U.S. R&D Firms in 1998

Firm Name	Firmid [a]	Observation
GOODRICH (B F) CO	GR	82
GOODYEAR TIRE & RUBBER CO	GT	83
GTE CORP	GTE	84
HALLIBURTON CO	HAL	85
HASBRO INC	HAS	86
HONEYWELL INC	HON	87
HERCULES INC	HPC	88
HARRIS CORP	HRS	89
HEWLETT-PACKARD CO	HWP	90
INTL BUSINESS MACHINES CORP	IBM	91
ICOS CORPORATION	ICOS	92
INTEGRATED DEVICE TECH INC	IDTI	93
INTL FLAVORS & FRAGRANCES	IFF	94
INFORMIX CORP	IFMX	95
ITT INDUSTRIES INC	IIN	96
IMATION CORP	IMN	97
IMMUNEX CORP	IMNX	98
INTERGRAPH CORP	INGR	99
INTEL CORP	INTC	100
INTUIT INC	INTU	101
IOMEGA CORP	IOM	102
INTL PAPER CO	IP	103
INGERSOLL-RAND CO	IR	104
I2 TECHNOLOGIES INC	ITWO	105
JOHNSON CONTROLS INC	JCI	106
EDWARDS J D & CO	JDEC	107
JDS UNIPHASE CORP	JDSU	108
JOHNSON & JOHNSON	JNJ	109
KELLOGG CO	K	110
KLA-TENCOR CORP	KLAC	111
KIMBERLY-CLARK CORP	KMB	112
LYCOS INC	LCOS	113
LEAR CORP	LEA	114
LITTON INDUSTRIES INC	LIT	115
LILLY (ELI) & CO	LLY	116
LOCKHEED MARTIN CORP	LMT	117
LAM RESEARCH CORP	LRCX	118
LSI LOGIC CORP	LSI	119
LUCENT TECHNOLOGIES INC	LU	120
LEXMARK INTL GRP INC	LXK	121
LUBRIZOL CORP	LZ	122

Table A2
The Top 200 U.S. R&D Firms in 1998

Firm Name	Firmid [a]	Observation
MATTEL INC	MAT	123
MCKESSON HBOC INC	MCK	124
MEDTRONIC INC	MDT	125
MENTOR GRAPHICS CORP	MENT	126
MALLINCKRODT INC	MKG	127
MINNESOTA MINING & MFG CO	MMM	128
PHILIP MORRIS COS INC	MO	129
MOBIL CORP	MOB	130
MOLEX INC	MOLX	131
MOTOROLA INC	MOT	132
MERCK & CO	MRK	133
MICROSOFT CORP	MSFT	134
MONSANTO CO	MTC	135
MICRON TECHNOLOGY INC	MU	136
MYLAN LABORATORIES	MYL	137
NAVISTAR INTERNATIONL	NAV	138
NCR CORP	NCR	139
NABISCO GROUP HLDGS CORP	NGH	140
NORTHROP GRUMMAN CORP	NOC	141
NOVELL INC	NOVL	142
NETSCAPE COMMUNICATIONS CORP	NSCP	143
NATIONAL SEMICONDUCTOR CORP	NSM	144
NOVELLUS SYSTEMS INC	NVLS	145
PITNEY BOWES INC	PBI	146
PACCAR INC	PCAR	147
PFIZER INC	PFE	148
PROCTER & GAMBLE CO	PG	149
PIONEER HI-BRED INTERNATIONL	PHB	150
PARAMETRIC TECHNOLOGY CORP	PMTC	151
PHARMACIA & UPJOHN INC	PNU	152
PPG INDUSTRIES INC	PPG	153
POLAROID CORP	PRD	154
PEOPLESOFT INC	PSFT	155
QUALCOMM INC	QCOM	156
QWEST COMMUNICATION IN TL INC	QWST	157
RALSTON PURINA CO	RAL	158
READ-RITE CORP	RDRT	159
ROHM & HAAS CO	ROH	160
ROCKWELL INTL CORP	ROK	161
RAYTHEON CO -CL B	RTN.B	162
SUNGARD DATA SYSTEMS I NC	SDS	163

Table A2
The Top 200 U.S. R&D Firms in 1998

Firm Name	Firmid [a]	Observation
SEAGATE TECHNOLOGY	SEG	164
SCIENTIFIC-ATLANTA INC	SFA	165
SILICON GRAPHICS INC	SGI	166
SCHERING-PLOUGH	SGP	167
SHELL OIL CO	SHELL	168
S3 INCORPORATED	SIII	169
SHARED MEDICAL SYSTEMS CORP	SMS	170
SYNOPSYS INC	SNPS	171
SUNDSTRAND CORP	SNS	172
SEQUENT COMPUTER SYSTEMS INC	SQNT	173
ST JUDE MEDICAL INC	STJ	174
STORAGE TECHNOLOGY CP	STK	175
SUN MICROSYSTEMS INC	SUNW	176
SYBASE INC	SYBS	177
STRYKER CORP	SYK	178
SYMANTEC CORP	SYMC	179
AT&T CORP	T	180
TERADYNE INC	TER	181
TELLABS INC	TLAB	182
THERMO ELECTRON CORP	TMO	183
TRW INC	TRW	184
TEXACO INC	TX	185
TEXAS INSTRUMENTS INC	TXN	186
TEXTRON INC	TXT	187
UNIGRAPHICS SOLUTIONS INC	UGS	188
UNISYS CORP	UIS	189
UNION CARBIDE CORP	UK	190
UNITED TECHNOLOGIES CO RP	UTX	191
VARIAN MEDICAL SYTEMS INC	VAR	192
VLSI TECHNOLOGY INC	VLSI	193
WORLD ACCESS INC	WAXS	194
WESTERN DIGITAL CORP	WDC	195
WHIRLPOOL CORP	WHR	196
WARNER-LAMBERT CO	WLA	197
XILINX INC	XLNX	198
EXXON CORP	XON	199
XEROX CORP	XRX	200

Notes: [a,b] Firmid is the 1998 ticker symbol of the firm, except for BELL COMMUNICATIONS RESEARCH INC, which is assigned the artificial ticker BELLC.

Table A.3
Treatment of AT&T and General Motors
Families of Companies

Name of Family or Firm	FIRMID	Spinoff Date	Treatment of Papers and Citations
AT&T Family			
AT&T CORP	T	N.A.	Separate in all years
BELL COMMUNICATIONS RESEARCH INC	BELLC	1984	"
LUCENT TECHNOLOGIES INC	LU	1996	"
General Motors Family			
GENERAL MOTORS CORP	GM	N.A.	Separate in all years
DELPHI AUTOMOTIVE SYS CORP	DPH	1998	"

Notes: [a,b] Firmid is the 1998 ticker symbol of the firm, except for BELL COMMUNICATIONS RESEARCH INC, which is assigned the artificial ticker BELLC.

Table A4
Mapping Between ISI88 Fields and NSF12 and NSF20 Fields

ISI88	ISI88 Field Description	NSF12	NSF12 Field Description	NSF20	NSF20 Field Description
A_A	AGRICULTURE/AGRONOMY	AGRI	Agriculture	AGRI	Agriculture
AN	ANIMAL & PLANT SCIENCES	AGRI	Agriculture	AGRI	Agriculture
AQU	AQUATIC SCIENCES	AGRI	Agriculture	AGRI	Agriculture
AS	ANIMAL SCIENCES	AGRI	Agriculture	AGRI	Agriculture
CMA	AGRICULTURAL CHEMISTRY	AGRI	Agriculture	AGRI	Agriculture
ENT	ENTOMOLOGY/PEST CONTROL	AGRI	Agriculture	AGRI	Agriculture
F	FOOD SCIENCE/NUTRITION	AGRI	Agriculture	AGRI	Agriculture
PL	PLANT SCIENCES	AGRI	Agriculture	AGRI	Agriculture
VET	VETERINARY MEDICINE/ANIMAL HEALTH	AGRI	Agriculture	AGRI	Agriculture
SP	SPACE SCIENCE	ASTR	Astronomy	ASTR	Astronomy
BEH	NEUROSCIENCES & BEHAVIOR	BIOL	Biology	BIOL	Biology
BIL	BIOCHEMISTRY & BIOPHYSICS	BIOL	Biology	BIOL	Biology
BIO	BIOLOGY	BIOL	Biology	BIOL	Biology
BTC	BIOTECHNOLOGY & APPLIED MICROBIOLOGY	BIOL	Biology	BIOL	Biology
CEL	CELL & DEVELOPMENTAL BIOLOGY	BIOL	Biology	BIOL	Biology
CGX	ONCOGENESIS & CANCER RESEARCH	BIOL	Biology	BIOL	Biology
ENV	ENVIRONMENT/ECOLOGY	BIOL	Biology	BIOL	Biology
EXP	EXPERIMENTAL BIOLOGY	BIOL	Biology	BIOL	Biology
IMM	IMMUNOLOGY	BIOL	Biology	BIOL	Biology
MBG	MOLECULAR BIOLOGY & GENETICS	BIOL	Biology	BIOL	Biology
MCB	MICROBIOLOGY	BIOL	Biology	BIOL	Biology
PHM	PHARMACOLOGY & TOXICOLOGY	BIOL	Biology	BIOL	Biology
PSL	PHYSIOLOGY	BIOL	Biology	BIOL	Biology

Table A4
Mapping Between ISI88 Fields and NSF12 and NSF20 Fields

ISI88	ISI88 Field Description	NSF12	NSF12 Field Description	NSF20	NSF20 Field Description
CML	CHEMISTRY & ANALYSIS	CHEM	Chemistry	CHEM	Chemistry
CMP	CHEMISTRY	CHEM	Chemistry	CHEM	Chemistry
INC	INORGANIC & NUCLEAR CHEMISTRY	CHEM	Chemistry	CHEM	Chemistry
ORG	ORGANIC CHEMISTRY/POLYMER SCIENCE	CHEM	Chemistry	CHEM	Chemistry
PHC	PHYSICAL CHEMISTRY/CHEMICAL PHYSICS	CHEM	Chemistry	CHEM	Chemistry
SIA	SPECTROSCOPY/INSTRUMENTATION/ANALYTICAL SCIENCES	CHEM	Chemistry	CHEM	Chemistry
CSE	COMPUTER SCIENCE & ENGINEERING	COMP	Computer Science	COMP	Computer Science
IST	INFORMATION TECHNOLOGY & COMMUNICATIONS SYSTEMS	COMP	Computer Science	COMP	Computer Science
ECO	ECONOMICS	ECON	Economics	ECON	Economics
MTH	MATHEMATICS	MATH	Mathematics and Statistics	MATH	Mathematics and Statistics
XY	STATISTICS	MATH	Mathematics and Statistics	MATH	Mathematics and Statistics
AIC	ANESTHESIA & INTENSIVE CARE	MEDI	Medicine	MEDI	Medicine
CAR	CARDIOVASCULAR & RESPIRATORY SYSTEMS	MEDI	Medicine	MEDI	Medicine
CVS	CARDIOVASCULAR & HEMATOLOGY RESEARCH	MEDI	Medicine	MEDI	Medicine
DEN	DENTISTRY/ORAL SURGERY & MEDICINE	MEDI	Medicine	MEDI	Medicine
DER	DERMATOLOGY	MEDI	Medicine	MEDI	Medicine
DGX	"MEDICAL RESEARCH, DIAGNOSIS & TREATMENT"	MEDI	Medicine	MEDI	Medicine
END	"ENDOCRINOLOGY, NUTRITION & METABOLISM"	MEDI	Medicine	MEDI	Medicine
GAS	GASTROENTEROLOGY AND HEPATOLOGY	MEDI	Medicine	MEDI	Medicine

Table A4
Mapping Between ISI88 Fields and NSF12 and NSF20 Fields

ISI88	ISI88 Field Description	NSF12	NSF12 Field Description	NSF20	NSF20 Field Description
GNC	GENERAL & INTERNAL MEDICINE	MEDI	Medicine	MEDI	Medicine
HEM	HEMATOLOGY	MEDI	Medicine	MEDI	Medicine
HLT	HEALTH CARE SCIENCES & SERVICES	MEDI	Medicine	MEDI	Medicine
INF	CLINICAL IMMUNOLOGY & INFECTIOUS DISEASE	MEDI	Medicine	MEDI	Medicine
MED	RESEARCH/LABORATORY MEDICINE & MEDICAL TECHNOLOGY	MEDI	Medicine	MEDI	Medicine
MGN	MEDICAL RESEARCH, GENERAL TOPICS	MEDI	Medicine	MEDI	Medicine
MUL	MULTIDISCIPLINARY	MEDI	Medicine	MEDI	Medicine
NEU	NEUROLOGY	MEDI	Medicine	MEDI	Medicine
NUT	ENDOCRINOLOGY, METABOLISM & NUTRITION	MEDI	Medicine	MEDI	Medicine
OGS	MEDICAL RESEARCH, ORGANS & SYSTEMS	MEDI	Medicine	MEDI	Medicine
ONC	ONCOLOGY	MEDI	Medicine	MEDI	Medicine
OPH	OPHTHALMOLOGY	MEDI	Medicine	MEDI	Medicine
ORT	ORTHOPEDICS, REHABILITATION & SPORTS MEDICINE	MEDI	Medicine	MEDI	Medicine
OTO	OTOLARYNGOLOGY	MEDI	Medicine	MEDI	Medicine
PED	PEDIATRICS	MEDI	Medicine	MEDI	Medicine
PMC	PHARMACOLOGY/TOXICOLOGY	MEDI	Medicine	MEDI	Medicine
PSY	CLINICAL PSYCHOLOGY & PSYCHIATRY	MEDI	Medicine	MEDI	Medicine
RAD	RADIOLOGY, NUCLEAR MEDICINE & IMAGING	MEDI	Medicine	MEDI	Medicine
REP	REPRODUCTIVE MEDICINE	MEDI	Medicine	MEDI	Medicine
RHU	RHEUMATOLOGY	MEDI	Medicine	MEDI	Medicine
SOC	ENVIRONMENTAL MEDICINE & PUBLIC HEALTH	MEDI	Medicine	MEDI	Medicine

Table A4
Mapping Between ISI88 Fields and NSF12 and NSF20 Fields

ISI88	ISI88 Field Description	NSF12	NSF12 Field Description	NSF20	NSF20 Field Description
SUR	SURGERY	MEDI	Medicine	MEDI	Medicine
URO	UROLOGY & NEPHROLOGY	MEDI	Medicine	MEDI	Medicine
APP	APPLIED PHYSICS/CONDENSED MATTER/MATERIALS SCIENCE	PHYS	Physics	PHYS	Physics
O_A	OPTICS & ACOUSTICS	PHYS	Physics	PHYS	Physics
PHS	PHYSICS	PHYS	Physics	PHYS	Physics
PSI	PSYCHIATRY	PSYC	Psychology	PSYC	Psychology
PSO	PSYCHOLOGY	PSYC	Psychology	PSYC	Psychology
AER	AEROSPACE ENGINEERING	TENG	Total Engineering	AERE	Aerospace Engineering
ARA	AI, ROBOTICS & AUTOMATIC CONTROL	TENG	Total Engineering	INDE	Industrial Engineering
CIV	CIVIL ENGINEERING	TENG	Total Engineering	CIVE	Civil Engineering
CME	CHEMICAL ENGINEERING	TENG	Total Engineering	CHEE	Chemical Engineering
EEE	ENVIRONMENTAL ENGINEERING & ENERGY	TENG	Total Engineering	OENG	Other Engineering
EL	ELECTRICAL & ELECTRONICS ENGINEERING	TENG	Total Engineering	ELEE	Electrical Engineering
EMA	ENGINEERING MATHEMATICS	TENG	Total Engineering	OENG	Other Engineering
GNE	ENGINEERING MANAGEMENT/GENERAL	TENG	Total Engineering	INDE	Industrial Engineering
GPM	GEOLOGICAL, PETROLEUM & MINING ENGINEERING	TENG	Total Engineering	OENG	Other Engineering
I_M	INSTRUMENTATION & MEASUREMENT	TENG	Total Engineering	INDE	Industrial Engineering
IG	BIOMEDICAL ENGINEERING	TENG	Total Engineering	OENG	Other Engineering
IJ	INDUSTRIAL ENGINEERING	TENG	Total Engineering	INDE	Industrial Engineering
MEC	MECHANICAL ENGINEERING	TENG	Total Engineering	MECE	Mechanical Engineering
MET	METALLURGY	TENG	Total Engineering	MATE	Materials Science

Table A4
Mapping Between ISI88 Fields and NSF12 and NSF20 Fields

ISI88	ISI88 Field Description	NSF12	NSF12 Field Description	NSF20	NSF20 Field Description
MTR	MATERIALS SCIENCE & ENGINEERING	TENG	Total Engineering	MATE	Materials Science
NCL	NUCLEAR ENGINEERING	TENG	Total Engineering	OENG	Other Engineering
EAR	EARTH SCIENCES	TGEO	Total Earth Sciences	EART	Earth Science
SI	OCEANOGRAPHY	TGEO	Total Earth Sciences	OCEA	Oceanography

Notes: ISI88 is a set of 88 detailed fields developed by Thomson-Reuters (formerly, the Institute for Scientific Information) for the assignment of scientific journals to disciplines. NSF12 is a set of 12 main fields used by the National Science Foundation to assign research expenditures, graduate students, and other survey data to universities and sciences. NSF20 is a slightly more detailed breakdown of fields that considers engineering and earth science sub-fields separately.