

Measuring science-technology interaction using rare inventor-author names

Kevin W. Boyack^a and Richard Klavans^b

^a SciTech Strategies, Inc., Albuquerque, NM 87122 USA (kboyack@mapofscience.com)

^b SciTech Strategies, Inc., Berwyn, PA 19312 USA (rklavans@mapofscience.com)

Abstract

The relationship between science and technology has been extensively studied from both theoretical and quantitative perspectives. Quantitative studies typically use patents as proxy for technology and scientific papers as proxy for science, and investigate the relationship between the two. Most such studies have been limited to a single discipline or country. In this paper, we investigate science-technology interaction over a broad range of science and technology by identifying and validating a set of 13,440 inventor-authors through matching of rare names obtained from paper and patent data. These 13,440 inventor-authors are listed as inventors on over 42,000 US patents between 2002 and 2006. Analysis of the distribution of these patents over classes shows that this 5% sample is a suitable sample for further analysis. In addition, a map of 290 IPC patent subclasses was created, showing the relationship between patent classes and industries as well as the distribution of patent classes with high science orientation and low science orientation.

Keywords: science-technology interaction; inventor-author matching; patent mapping; patent distributions

1. Introduction

The relationship between science and technology has been extensively studied from multiple perspectives. On one hand a variety of models or descriptive frameworks (e.g. Triple Helix of university-industry-government relations) have been proposed to characterize this relationship. On another hand, data have been used to develop metrics for the purpose of quantifying the relationship. Most of these studies, whether conceptual or quantitative, are policy oriented. Many have been concerned with knowledge production and the nature, mechanism, directionality, and/or magnitude of the transfer of that knowledge between science and technology.

Quantitative studies have largely focused on the non-patent references (NPR) on the front pages of US patents, the majority of which refer to scientific or technical documents, such as journal articles or conference papers. Early work in this area, such as that performed by Narin and colleagues (Carpenter & Narin, 1983; Narin & Olivastro, 1992, 1998), was based on the assumption that a patent-to-paper reference was an indicator that a technology (in the form of a patent) directly descended from science (in the form of a paper). More recent work, particularly that of Meyer, has shown that the inference of a direct linkage from science to a resulting technology is unfounded in a large fraction of cases. Rather, the majority of patent-to-paper references are there for other reasons. This does not discount the fact that the patent and paper are related, but merely calls into question the directionality and mechanism of that relationship. While many publications refer to “science-technology linkage”, Meyer has eschewed that terminology in favor of “science-technology interaction” (M. Meyer, 2000), and we will follow that convention.

While many studies have used NPR data with the specific intent of linking economic benefit to public science through technology development or patents, we suggest there is great value in simply understanding the overlap between science and technology, regardless of the direction of or mechanism for the interactions. We believe

there is a much less messy¹ and more comprehensive way to show the interaction between science and technology – through inventor-authors. This is not a new idea. Inventor-authors have been studied for the better part of two decades. However, we now have the means to scale such a study to cover “all of science” and “all of technology”, and to quantify the overlaps or interactions between the two. In this paper, we identify a large number of inventor-authors using rare names. Rare names that occur in both a literature and a patent database are far more likely to identify a single person, an inventor-author, than are more common names.

The balance of the paper will proceed as follows. First, we give a brief background on relevant studies of science-technology interaction, and more particularly on inventor-author studies. We then describe the data and methods used to identify and verify inventor-authors. This is followed by a discussion of results and implications, introduction of a patent map showing the science orientation of different technologies and industries, and finally by a short summary and suggestions for future work.

2. Background

Several approaches to quantifying science-technology interaction have been used over the years. Primary among these has been citation analysis using NPR's. In most of these studies, patent references are linked to specific papers in SCI (Science Citation Index) journals. Francis Narin and his company, CHI Research², generated significant business out of this type of study (Hicks, Tomizawa, Saitoh, & Kobayashi, 2004), and eventually linked all science-based NPR's to SCI papers starting with the 1983 patent year (Narin & Olivastro, 1998). We presume this took a great deal of manual work. Meyer also has a substantial body of work based on NPR's (M. Meyer, 2000, 2002), much of which focuses on the area of nanotechnology (M. Meyer, 2001).

Paper-to-patent citation analysis has also been investigated (Glänzel & Meyer, 2003), although the numbers of patents cited in papers is quite low relative to other citation types. Paper-to-patent citation volume is dominated by chemistry references, thus this technique may be less applicable in other disciplines. Lexical approaches have also been used to establish semantic linkages between patents and scientific articles. Outputs from this approach include correspondence tables between patent classes and scientific disciplines (Bassecoulard & Zitt, 2004).

The idea of a common knowledge base giving rise to joint or concurrent development of science and technology through inventor-authors has also been studied. Bonaccorsi and Thoma (2007) explore some of the reasons why individuals are useful as a unit of study as opposed to using citations. Coward and Franklin (1989) were among the first to match inventors with authors using literature and patent data. They explored the field of semiconductor-related science using a database of over 100k papers and 2,452 patents, with the objective of matching patent data to their bibliometric model of the field. They identified 247 inventor-authors, and found that this method gave better results (more matches) than either institutional matching or linking through NPRs.

Most studies of inventor-authors since that time have been small scale, dealing with specific fields of science such as laser medicine research (Noyons, van Raan, Grupp, & Schmoch, 1994) or nanotechnology (M. Meyer, 2006),

¹ Non-patent references are notoriously messy, difficult to clean and parse using automated methods, and difficult to match to journal articles. For example, Verbeek et al. (2002) were able to link only 9.3% of 1.15 million NPR to specific articles in the Science Citation Index. They had better luck (26%) identifying journal titles. This is still well short of the 50% or more of NPRs that are commonly assumed to be science-related.

² CHI Research was sold and divided several years ago. The patent research from CHI is now located at Ipiq.

or with the output of a country (M. Meyer, 2003; Tijssen & Korevaar, 1997). Murray (2002) took one seminal and highly cited paper-patent pair in tissue engineering, and investigated the network of researchers and inventors who subsequently cited those works.

Although most work in matching inventors with authors has also required an institutional match to validate the inventor-author, there are cases in which academic authors have inventions that are assigned to an organization other than the academic institution. Several studies have attempted to match these academic inventors with their patents that are assigned to other institutions (Balconi, Breschi, & Lissoni, 2004; M. Meyer, 2003; Noyons et al., 2003). The work by Noyons et al. (2003) is the largest scale and most comprehensive inventor-author study to date. Although their work was limited to EU countries, they were able to link over 15,000 inventor-authors (combination of full and partial matches), roughly 60% of which also showed institutional matches, and the other 40% of which linked the academic inventor with patents assigned to other institutions.

As with all types of studies involving names, synonymy (multiple name variations for one person) and homonymy (multiple persons with the same name) are issues in inventor-author studies as well. Different approaches to reducing the effects of these issues include restricting matches to same-country matches (M. Meyer, 2006; Noyons et al., 2003), limiting the domain of study (M. Meyer, 2006), logic involving variations on name parts (Noyons et al., 2003), and textual analysis using vector space matching on extracted terms (Cassiman, Glennisson, Verbeek, & van Looy, 2007). In this study, we approach the synonymy and homonymy issues by restricting matches to rare names occurring in literature and patent data.

3. Data and Methods

3.1 Data

The purpose of this study is to quantify the interaction between science and technology through inventor-author linkages in such a way that those interactions can ultimately be compared by scientific discipline for policy purposes.³ To do this, we need to cover “all of science” and “all of technology”, or at least as large a segment of each as is possible, rather than to focus at the level of a single discipline. Large databases were thus needed for this study.

For the literature part of the study, we use Scopus data from a five-year period, 2002-2006, comprising 5.96 million papers and 23.6 million authors with institutions. The advantage of using Scopus data over Thomson Scientific (or Web of Science) data for author studies is that Scopus links individual authors and institutions at the paper level. By contrast, with Thomson data, although all authors and all institutions are listed, institutional affiliations of each author must be inferred except for first author/first institution combinations and papers with only one authoring institution.

For the patent part of the study, we use data from the US Patent and Trademark Office (USPTO) for the same five-year period, 2002-2006, comprising over 907,500 patents and 2.15 million inventors. Weekly front-page files are available on-line from USPTO⁴ that contain inventor, assignee, and reference information, among other fields. Although the format of these files has changed over the years (from tagged record to XML), full inventor and

³ A detailed disciplinary comparison is planned, but is beyond the scope of this paper.

⁴ <ftp://ftp.uspto.gov/pub/patdata/>

assignee data are available and can be easily parsed from the source data. Inventors are not specifically linked to assignees in these data. Rather, relationships between inventors and assignees must be inferred. Over the five-year time period, 11.8% of US patents have no assignee, 85.8% have one assignee, and only 2.4% have multiple assignees.

3.2 Methodology

Our matching of inventors with authors is based on the simple assumption that if a name is rare and occurs in the inventor data, and in the author data, it is very possible that both instances are referring to the same person. Conversely, if a name is common, and could represent tens, or even hundreds, of different authors or inventors, it will take much more effort to find an accurate match. We wish to avoid this effort. Further, we do not believe finding all inventor-authors is necessary. For policy purposes, and to show the overlap between science and technology on a disciplinary basis, a representative sample is all that is required.

Let us now define what we mean by a rare name. Certainly, the most extreme case of rare would be where there is only one institutional affiliation for a given name. However, if we were to limit the analysis to this definition of rare, it might be difficult to obtain a sufficiently large and representative sample of inventor-authors. We have chosen to have a variable definition of rare, one that will allow us to not only gather a larger sample, but to compare thresholds of rareness. Consider the case in Table 1 for one author string.

Table 1. Institutional distribution of papers for a unique author string.

Author string	Institution	# papers	fraction
ABARBANEL_H	University of California, San Diego	39	0.907
ABARBANEL_H	Scripps Institution of Oceanography	3	0.070
ABARBANEL_H	Universidad Autonoma de Madrid	1	0.023

Author ABARBANEL_H is associated with 3 different institutions in the Scopus data. However, nearly 91% of the publications assigned to this author name are at one institution. We thus make the assumption that this name is rare at a threshold of 0.90. Note that we have truncated the author name to include only the first initial because neither authors, inventors, publishers, nor database vendors are consistent in their use of multiple initials. Although this choice increases the effect of homonymy in our analysis, it reduces the synonymy dramatically. We rely on the rareness fraction to reduce the effect of homonymy.

The method used to process and match the author and inventor data is as follows:

- Using all author/institution pairs from the Scopus data, all author names were converted to authfi strings (last name with first initial), and the Scopus orgid was used for the institution affiliation.
- For each unique authfi, the fraction of papers by institution was calculated, following the example given in Table 1. The sum of the fractions for each unique author string should be equal to 1.0.
- Using all inventor/assignee pairs from the patent data, all inventor names were converted to strings in last name with first initial format, to match the author format. Assignee data were used for institutional affiliation. In the case of one assignee, the inventor was given that institutional affiliation. In the case of no assignee, the affiliation was left null. In the case of multiple assignees, additional logic was used to link inventors with assignees – first using common city, then common state, and finally common country.

- For each unique inventor string, the fraction of patents by assignee was calculated (NULL was considered an assignee), following the example given in Table 1. The sum of the fractions for each unique inventor string should be equal to 1.0.
- Author strings and inventor strings were then matched at different rareness fraction thresholds. For example, using a threshold of 1.0, the author and inventor strings were required to match exactly, and both fractions were required to be 1.0. For a threshold of 0.7, the author and inventor strings were required to match exactly, and both fractions were required to be 0.7 or higher.

The results of this process are shown in Table 2. We matched name strings down to a fractional threshold of 0.7. At this level, 56,774 unique author strings matched inventor strings. Additional matching could be done down to fraction > 0.5 , and still preserve the feature that each unique name string will occur only once. If matching were done at fraction = 0.5, it would introduce the case where many author strings will be given twice, each with a fraction of 0.5. We would expect the precision of matching to drop drastically at this point.

Table 2. Results of the inventor-author name matching process at different fractional thresholds.

Rareness fraction range	# auth names	# inv names	Inv-auth Matches	Inv-auth + institution matches	Fraction valid matches	# patents	NULL assignees
$f = 1.00$	1,106,404	278,146	35,360	7,843	0.222	18,816	3,708
$1.0 \geq f \geq 0.9$	1,138,340	281,214	38,842	9,068	0.233	25,370	3,973
$1.0 \geq f \geq 0.8$	1,222,530	292,594	47,454	11,362	0.239	34,653	4,653
$1.0 \geq f \geq 0.7$	1,305,848	304,527	56,774	13,440	0.237	42,129	5,391
$1.0 \geq f \geq 0.6$	1,462,269	330,886	76,027				
$1.0 \geq f > 0.5$	1,512,207	335,987	84,402				
$1.0 \geq f \geq 0.5$	1,971,180	425,546	148,532				
UNIQUE (authfi)	2,182,303	436,521					
UNIQUE (authfi+inst)	8,712,536	1,049,650					

Note that we did not require a country or institutional match to this point in the process. This was done for two reasons; first, we wanted to measure the precision of matching based on rare names only, without other filtering; second, the institutional names used in the Scopus and USPTO databases are often quite different for the same institution (e.g. University of California, San Diego vs. Regents of the University of California), and we were keen to avoid the cleaning steps.

To validate inventor-author pairs, each of the 56,774 inventor-author matches was visually inspected for an institutional match to determine the actual matches; the resulting numbers of validated pairs (those with institutional matches) is given in Table 2. Interestingly, at a rareness fraction of 1.0, the valid matching rate is only 22%. There were another 10.5% of the inventor-author pairs where the patent assignee was NULL. (This is only slightly lower than the overall NULL assignee rate of 11.8%.) Some of these would undoubtedly be matches if the patent owner were named. However, if the rate at matching null assignees is similar to that of matching named assignees, the resulting matching rate would only be $.222/.895 = 25\%$. This suggests that even among the

rarest of names, those where there is only one affiliated authoring institution and only one patenting institution, roughly $\frac{3}{4}$ of the presumed inventor-author pairs are actually two different people, one who writes papers at one institution, and another who patents at a separate institution. When the rareness threshold is lowered to 0.8, the valid matching rate actually increases from 22% to nearly 24%. For the next rareness increment, for thresholds down to 0.7, the matching rate decreases slightly.

It is important to mention that our method does not match inventor-authors in cases where the author is an academic researcher whose patents are assigned to companies rather than to the academic institution. Research on this effect was reviewed in section 2. However, our results, showing that even among the rarest of names, $\frac{3}{4}$ of inventor-author name matches are false, suggest that the academic inventor matching techniques used by Noyons et al. (2003), Meyer (2006), and others, may actually overstate academic involvement to some degree. Those methods assume that a name match between an academic author and an inventor of a non-academic patent in the same country are the same person, and do not account for the fact that there may actually be an inventor at the assignee institution with the same name. To be fair, those methods do match on multiple initials, and are thus more precise in that regard, undoubtedly giving rise to fewer false hits than in our method. Yet, rare names are often rare because they are indigenous to specific countries or regions. Thus, limiting matches to within-country matches does not assure accurate matching.

4. Discussion

4.1 Distribution by IPC sections

The numbers of unique patents for which the validated inventor-authors were listed as inventors is given in Table 2. At the fraction ≥ 0.7 level, 42,129 patents were identified. Of these, 41,803 patents were assigned to IPC subclasses (4-character) shown in the IPC8 classification guide (WIPO, 2006). This is 5.09% of all U.S. patents issued (in those same IPC subclasses) during the years 2002-2006. The remaining few hundred patents are design patents and other patents not classified in one of the standard IPC subclasses.

The IPC8 guide (WIPO, 2006) lists 8 sections (A-H), 129 classes (3 character codes such as A01), and 639 subclasses (4 character codes such as A01B). In addition, the 8 sections are subdivided into named groupings that loosely correspond to 2 character codes, such as A0. In order to quantitatively link patents to scientific disciplines, we need to know if our 5% sample of the patents is representative of the actual distribution of science-oriented patents.

Figure 1 shows the distribution of our sample of 41,803 patents with respect to the distribution of all patents, and to the distribution of patents containing non-patent references. Results are shown by IPC subsection (two-character codes) using only the primary IPC code for each patent. Secondary classifications are not considered here. Category B8, related to nanotechnology, is a very new category, with few patents, and has been omitted from the chart.

Perusal of the Figure 1 suggests that our sample of patents is not representative of the distribution of all patents. The differences become even greater if viewed at the level of the 129 IPC3 codes. There may be many reasons why our 5% sample is not representative. Here we explore two of those reasons.

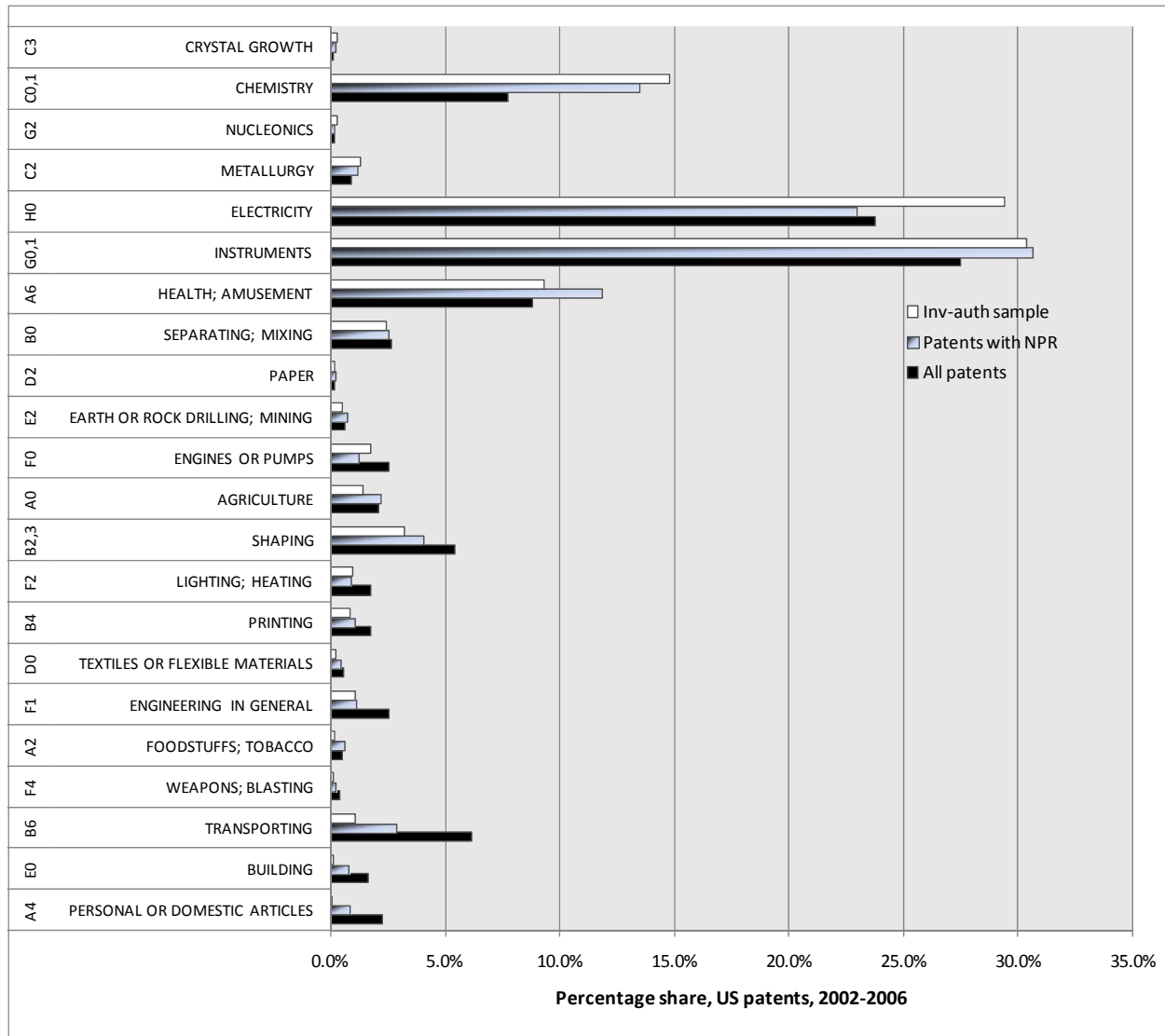


Figure 1. Distribution of patents assigned to validated inventor-authors by IPC subsection. Subsections are ordered (top to bottom) by the ratio of patents by our inventor-author sample to all patents to show an ordering of relative science orientation.

First, there may not be equal likelihoods for different patents categories to link to science. If one assumes that the sample distribution should be the same as the actual patent distribution, this contains the implicit assumption that each patent (and thus each patent class) is equally likely to have an inherent linkage to the science that is being published today. We believe that this is a poor assumption for intuitive reasons. As an example, we would expect the computer industry to be highly linked to science because increases in computing are tied to semiconductor processes whose advances are taking place at the micro-scales that are the subject of current scientific endeavors worldwide. By contrast, new methods of assembling clothing or furniture, or a new tool used in the construction industry, are most likely due to advances in engineering, rather than to advances in the science that is published in peer reviewed journals. Verbeek et al. (2002) also found that some technology fields are highly science-dependent while others are not.

Our intuition about the likelihood for different patent classes to be inherently linked to science is reflected in the sample distribution, and also in the NPR distributions shown in Figure 1. For example, categories A4 (Personal and domestic articles), B6 (Transporting), and E0 (Building) are highly under-represented by our sample. The categories in section F (Mechanical engineering) are also under-represented, but to a lesser extent. These are examples of categories that are engineering based rather than science based. By contrast, categories C0,1 (Chemistry), G0,1 (Instruments) and H0 (Electricity) are over-represented. These categories are all tied to basic sciences on an intuitive level. Although at first glance one might think that G0,1 (Instruments) is engineering based, it is the major category within section G (Physics), and thus does tie back to basic science.

The second reason we wish to mention is that these differences may represent the different propensities for inventors in different industries to publish. Some companies and industries have an unwritten policy not to publish (e.g. the American automobile industry). This is likely a confounding factor, but not the primary one, given the arguments listed for the first reason above. Taken by itself, this explanation would require the assumption that all patents are equally likely to link to science, which we have already shown is not likely. These arguments suggest that our 5% sample of patents, while not representative of the actual patent distribution, is reasonably representative of the distribution of patents that actually link to science from a patent classification standpoint.

4.2 Distribution by country

It is also informative to look at the country distribution of the sample using inventor addresses. Country shares and rank are shown in Table 3 for all patents, patents with NPR, and our inventor-author based sample. Fractional counting at the patent level was used in all three cases. Comparison of the shares and rankings for the three different cases gives rise to some interesting observations.

First, a comparison of patents with NPR to all patents can show the relative propensity of a nation to patent in fields that have an interaction with science. For example, the United States, although it has the largest share of US patents by any measure, has a larger share of patents with NPR than would be expected from the shares of all patents. This would suggest that the United States has a higher than average propensity to patent in areas related to science. The same thing can be seen, although to a lesser degree, for the United Kingdom, Canada, Israel, and several other countries further down the list.

By contrast, Japan, South Korea, and Taiwan have shares of patents with NPR that are far lower than would be expected from their shares of all patents. This is most pronounced for Taiwan, whose share of patents with NPR is only about 20% of the expected value. It would be easy to assume that these countries have a propensity to patent in fields that have lower interaction with science. However, other factors suggest a closer look. It is well known that the science base in these countries is heavily weighted toward engineering and the physical sciences (see, for example, Figure 3 in King (2004) for a citation share distribution for Japan)⁵. Since these are precisely the fields with the highest science-technology interaction, corresponding to IPC sections C, G, and H (see Figure 1), it is difficult to understand why these countries would not have larger fractions of patents with NPR. We note that these are all Pacific Rim countries with very different patenting systems and cultures than are found in the United States and Europe. There could be some systematic reason related to culture that results in lower than expected NPR rates, even when these patents are examined and granted within the US system.

⁵ Unpublished work by the authors shows very similar science profiles for Japan, China, Korea, and Taiwan.

Table 3. Distribution of patents by country. The top 25 countries are ordered by share of patents with NPR (non-patent references).

Country	All patents		Patents with NPR		Sample		Shares /
	Rank	Share	Rank	Share	Rank	Share	Share _{NPR}
United States	1	51.5%	1	59.0%	1	64.5%	1.09
Japan	2	21.1%	2	16.7%	2	13.3%	0.79
Germany	3	6.5%	3	6.1%	3	8.7%	1.42
United Kingdom	6	2.2%	4	2.6%	6	1.4%	0.54
France	7	2.2%	5	2.3%	4	2.6%	1.13
Canada	8	2.0%	6	2.2%	5	1.6%	0.75
South Korea	5	2.7%	7	1.7%	19	0.1%	0.07
Netherlands	10	0.9%	8	0.9%	8	1.1%	1.14
Israel	13	0.7%	9	0.9%	10	0.9%	0.99
Sweden	11	0.8%	10	0.8%	12	0.4%	0.50
Switzerland	12	0.8%	11	0.8%	9	1.1%	1.36
Italy	9	1.0%	12	0.8%	7	1.2%	1.57
Taiwan	4	3.4%	13	0.7%	24	0.1%	0.09
Australia	14	0.6%	14	0.7%	17	0.2%	0.28
Finland	15	0.5%	15	0.6%	11	0.6%	0.91
Belgium	16	0.4%	16	0.5%	13	0.4%	0.86
Denmark	18	0.3%	17	0.4%	14	0.4%	1.03
India	21	0.2%	18	0.3%	16	0.3%	0.91
Austria	17	0.3%	19	0.3%	15	0.3%	1.23
Singapore	20	0.3%	20	0.2%	22	0.1%	0.37
China	19	0.3%	21	0.2%	29	0.0%	0.18
Russian Federation	25	0.1%	22	0.2%	23	0.1%	0.53
Spain	22	0.2%	23	0.2%	20	0.1%	0.60
Norway	23	0.1%	24	0.1%	18	0.1%	0.89
Ireland	26	0.1%	25	0.1%	28	0.0%	0.40

Second, comparison of the sample share with the share of patents with NPR, the ratio of which is shown in the last column of Table 3, leads to observations about the usefulness of using rare names to identify inventor-authors by nation. Countries with a share ratio greater than 1.0 are over-represented using our rare name method of identifying inventor-authors, while those with a share ratio less than 1.0 are under-represented. The most over-represented countries percentage-wise are Italy, Germany, Switzerland, and Austria. Three of these four are dominated by Germanic names; thus a case can be made that Germanic names have a higher level of rareness than names from other languages. On the other end of the scale are the Asian names. South Korea, Taiwan, and China each have share ratios of less than 0.2. Japan has far more rare names than its Pacific Rim counterparts, but still less than many other countries, with a share ratio of 0.79. English names show diverse behavior; the United States is unexpectedly over-represented with a ratio of 1.09, while the United Kingdom (0.54), Canada (0.75), Ireland (0.40), and Australia (0.28) have far fewer rare names among inventor-authors. We have no means of explaining this difference among English-speaking countries. In short, our method of identifying inventor-authors using rare names does not provide a representative sample when considered by nation or language. However, this does not

seem to be a problem when considering the purpose of the study – to quantify science-technology overlap on a broad basis. The results of section 4.1 and Figure 1 suggest that our sample is sufficiently representative for this purpose.

4.3 Map of IPC classes

Figure 1 showed science-technology interaction at the level of 22 IPC subheadings. A much more detailed view of the interaction is possible if one maps IPC subclasses. Figure 2 shows a visual map of 290 IPC subclasses (4-character codes), generated from co-classification of IPC subclasses. The patent map was generated using the same process and algorithms we have used previously to generate maps of scientific journals (Boyack, Klavans, & Börner, 2005) and papers (Klavans & Boyack, 2006b). Co-classification counts were assigned for a pair of IPC subclasses if code₁ was primary and code₂ was secondary for a single patent. Counts were then summed by subclass pairs over all patents, the matrix was made symmetric by adding the upper and lower halves, and a K50 similarity measure (Klavans & Boyack, 2006a) was calculated. Given the number of very small subclasses, we kept only those subclasses with a total of 150 or more co-classification counts. Thus, only 290 of the original 639 IPC subclasses were represented in the final co-classification matrix. Of the 9,966 pair-wise similarities between classes (23.8% of the matrix elements were non-zero), only 940 edges remained after the edge pruning that is part of our layout process (Klavans & Boyack, 2006b). The resulting layout, visualized in Pajek (Batagelj & Mrvar, 1998), is shown in Figure 2. Node sizes reflect the relative numbers of patents by class.

Labeling of the map was done by hand at two different levels. First, most visually identifiable clusters in the map were labeled using examination of both the IPC subclass names and the dominant institutional assignees in the clusters. Second, the map was examined to see if it could be labeled by industry. Using industry names from Hoover's taxonomy of 37 industries,⁶ we found that the map could indeed be segmented by industry into contiguous segments. We do not claim that the industry boundaries on this patent map are absolutely precise, but rather show fuzzy boundaries between industries. IPC subclasses are relatively coarse distinctions, and in many cases a single subclass may belong to multiple industries, much as many large companies serve multiple industries. Seventeen of Hoover's 37 industries are represented on the patent map. In two cases industries have been combined: *Automotive* and *Aerospace & Defense* in one case, and *Computer Hardware + Software* in the other. In these two cases, any division between the industries based on locations of IPC subclasses on the map would have been artificial. Some industries are completely surrounded by others in patent space. For example, the *Telecommunications* and *Computer* industries are bounded by *Electronics* on both sides. All but one of the industries have drawn boundaries on the map. The one exception is the *Consumer Products* industry, which comprises all of the unbounded space.

The patent map shown in Figure 2 could be the subject of an entire paper. Many stories about technology and industry linkages could be told. Although such an analysis would undoubtedly be interesting, we choose to maintain our focus here on science-technology interaction, and limit our interpretation of the map to that end. In Figure 2, node color is related to science-technology interaction. Here, relative science-technology interaction values were calculated as the ratio of the fraction of patents in the sample to the fraction of all patents by class. Darker nodes, those with ratios over 1.15, show areas where science-technology interaction is high, while lighter nodes (white or very light) with ratios lower than 0.87 show areas where the interaction is low.

⁶ Hoover's is a Dun & Bradstreet company that collects and sells detailed business information about companies and industries. Their free industry taxonomy is available at <http://www.hoovers.com/free/industries/>.

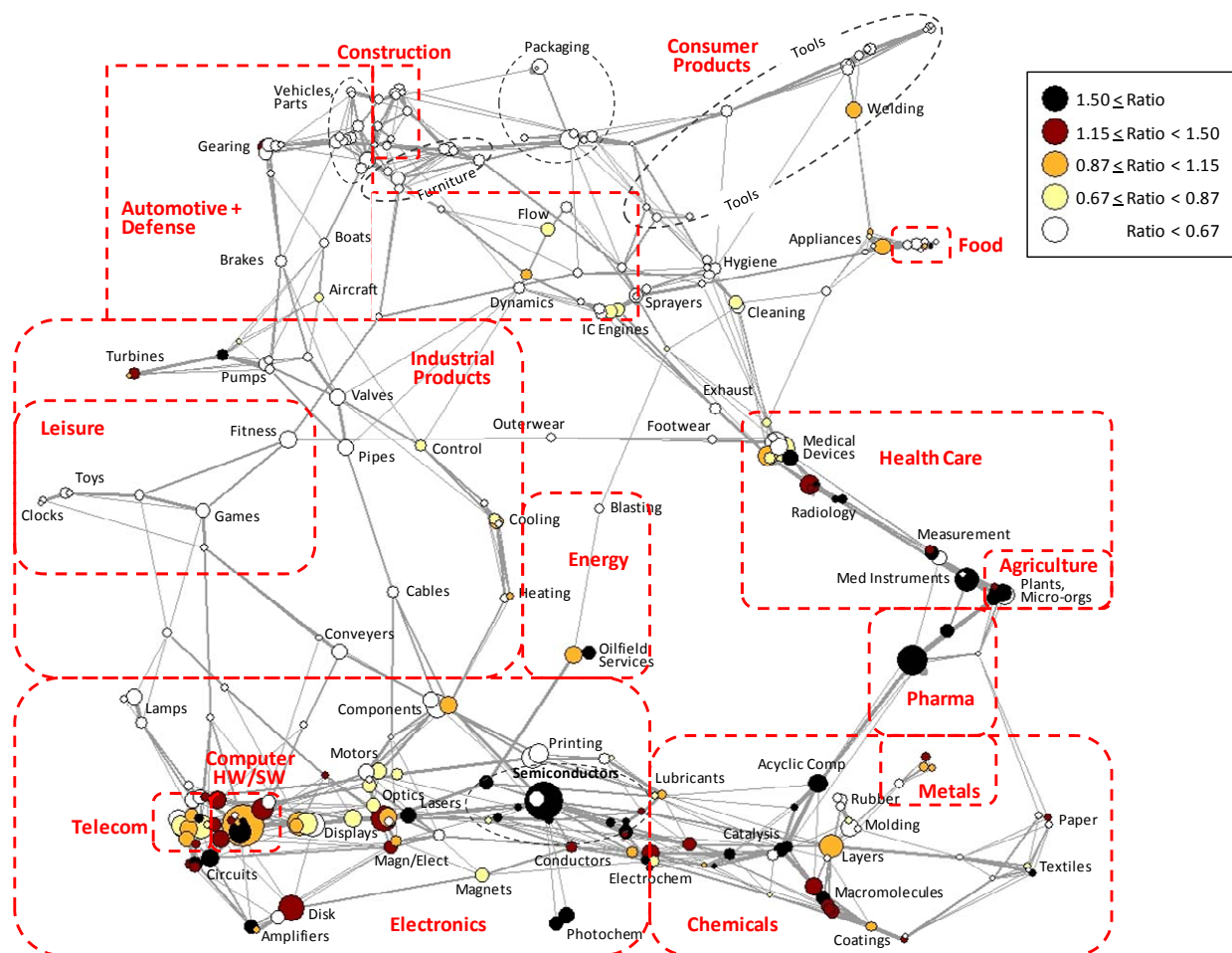


Figure 2. Map of IPC subclasses (3-character) based on co-classification. Smaller labels list the dominant technology in the clusters they are closest to, while larger labels and dashed partitions group the smaller clusters by industry.

The IPC classes with high science orientation are not uniformly distributed throughout the map, but are concentrated mostly in the lower half of the map. The industries with high science orientation nodes thus include *Electronics*, *Computer*, *Telecommunications*, *Chemicals*, *Metals*, *Pharmaceuticals*, *Agriculture*, *Energy*, and *Health Care*. This does not mean that all areas of these industries, or all patent subclasses that contribute heavily to these industries, have high science orientation. There are many light colored nodes (low science orientation) in these industries. By contrast, nearly all nodes in the other industry regions of the map (*Automotive & Defense*, *Construction*, *Industrial Products*, *Leisure*, *Consumer Products*, and *Food*) have a low science orientation.

5. Summary and Future Work

The purpose of this work has been to investigate science-technology interaction over a broad range of science and technology. We have done this by identifying and validating a set of over 13,000 inventor-authors and over 42,000 patents by matching rare names obtained from paper and patent data. Matching of rare names alone only

resulted in an approximate 25% match rate. Institutional matching was required to validate the matches. Analysis of the distribution of our 5% sample of patents has shown that it corresponds reasonably well to the distribution of patents containing non-patent references, and thus is a suitable sample for further studies of science-technology interaction.

IPC patent subclasses (4-character) were mapped using patent co-classification, and science-technology interaction values based on our 5% sample of patents were shown on the map. Fifteen different industry-based partitions of IPC subclasses were identified. Of these, nine industries have a high science orientation, while the other six have low science orientation. IPC subclasses with high science orientation can be distinguished from those with low science orientation.

There is much work that can be done with a set of validated inventor-authors data such as that identified in this study. Although we identified individual patents associated with the inventors, we did not identify the specific papers associated with the authors. This could easily be done, and would enable studies of overlap from the science perspective. If coupled with funding data, such studies could show the multiple overlaps of funding, science production, and patent production. Such studies would have distinct policy implications. Mining of citation count data for papers and patents could enable studies to show productivity and impact effects of inventor-authors with respect to their non-publishing or non-patenting peers, much like the study of Meyer (2006), but on a broader scale. Such results should have great impact in the policy community in the years to come.

References

- Balconi, M., Breschi, S., & Lissoni, F. (2004). Networks of inventors and the role of academia: An exploration of Italian patent data. *Research Policy*, 33, 127-145.
- Bassecouard, E., & Zitt, M. (2004). Patents and publications: The lexical connection. In H. F. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems* (pp. 665-694). Dordrecht: Kluwer Academic Publishers.
- Batagelj, V., & Mrvar, A. (1998). Pajek - A program for large network analysis. *Connections*, 21(2), 47-57.
- Bonaccorsi, A., & Thoma, G. (2007). Institutional complementarity and inventive performance in nano-science and technology. *Research Policy*, 36(6), 813-831.
- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.
- Carpenter, M. P., & Narin, F. (1983). Validation study: Patent citations as indicators of science and foreign dependence. *World Patent Information*, 5(3), 180-185.
- Cassiman, B., Glennisson, P., Verbeek, A., & van Looy, B. (2007). Measuring industry-science links through inventor-author relations: A profiling methodology. *Scientometrics*, 70(2), 379-391.
- Coward, H. R., & Franklin, J. J. (1989). Identifying the science-technology interface: Matching patent data to a bibliometric model. *Science, Technology & Human Values*, 14(1), 50-77.
- Glänzel, W., & Meyer, M. (2003). Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations. *Scientometrics*, 58(2), 415-428.

Hicks, D., Tomizawa, H., Saitoh, Y., & Kobayashi, S. (2004). Bibliometric techniques in the evaluation of federally funded research in the United States. *Research Evaluation*, 13(2), 78-86.

King, D. A. (2004). The scientific impact of nations. *Nature*, 430, 311.

Klavans, R., & Boyack, K. W. (2006a). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2), 251-263.

Klavans, R., & Boyack, K. W. (2006b). Quantitative evaluation of large maps of science. *Scientometrics*, 68(3), 475-499.

Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy*, 29, 409-434.

Meyer, M. (2002). Tracing knowledge flows in innovation systems. *Scientometrics*, 54(2), 193-212.

Meyer, M. (2003). Academic patents as an indicator of useful research? A new approach to measure academic inventiveness. *Research Evaluation*, 12(1), 17-27.

Meyer, M. (2006). Are patenting scientists the better scholars? An exploratory comparison of inventor-authors with their non-inventing peers in nano-science and technology. *Research Policy*, 35, 1646-1662.

Meyer, M. (2001). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. *Scientometrics*, 51(1), 163-183.

Murray, F. (2002). Innovation as co-evolution of scientific and technological networks: Exploring tissue engineering. *Research Policy*, 31, 1389-1403.

Narin, F., & Olivastro, D. (1992). Status report: Linkage between technology and science. *Research Policy*, 21, 237-249.

Narin, F., & Olivastro, D. (1998). Linkage between patents and papers: An interim EPO/US comparison. *Scientometrics*, 41(1-2), 51-59.

Noyons, E. C. M., Buter, R. K., van Raan, A. F. J., Schmoch, U., Heinze, T., Hinze, S., et al. (2003). *Mapping excellence in science and technology across Europe: Nanoscience and nanotechnology*: Centre for Science and Technology Studies (CWTS), Leiden University, Netherlands.

Noyons, E. C. M., van Raan, A. F. J., Grupp, H., & Schmoch, U. (1994). Exploring the science and technology interface: Inventor-author relations in laser medicine research. *Research Policy*, 23(4), 443-457.

Tijssen, R. J. W., & Korevaar, J. C. (1997). Unravelling the cognitive and interorganisational structure of public/private R&D networks: A case study of catalysis research in the Netherlands. *Research Policy*, 25(8), 1277-1293.

Verbeek, A., DeBackere, K., Luwel, M., Andries, P., Zimmermann, E., & Deleus, F. (2002). Linking science to technology: Using bibliographic references in patents to build linkage schemes. *Scientometrics*, 54(3), 399-420.

World Intellectual Property Organization (WIPO) (2006). *International Patent Classification, Eighth edition*, from <http://www.wipo.int/classifications/ipc/ipc8/>