Research Policy 43 (2014) 530–542

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Getting patents and economic data to speak to each other: An 'Algorithmic Links with Probabilities' approach for joint analyses of patenting and economic activity☆,☆☆

Travis J. Lybbert[a,*], Nikolas J. Zolas[b]

[a] Department of Agricultural & Resource Economics, University of California, Davis, United States
[b] Center for Economic Studies, United States Census Bureau, United States

ARTICLE INFO

Article history:
Received 24 August 2012
Received in revised form 1 May 2013
Accepted 4 September 2013
Available online 1 October 2013

Keywords:
Patents
Trade
Industry
Concordances
Technology
International Patent Classification

ABSTRACT

International technological diffusion is a key determinant of cross-country differences in economic performance. While patents can be a useful proxy for innovation and technological change and diffusion, fully exploiting patent data for such economic analyses requires patents to be tied to measures of economic activity. In this paper, we describe and explore a new algorithmic approach to constructing concordances between the International Patent Classification (IPC) system that organizes patents by technical features and industry classification systems that organize economic data, such as the Standard International Trade Classification (SITC) and the International Standard Industrial Classification (ISIC). This 'Algorithmic Links with Probabilities' (ALP) approach mines patent data using keywords extracted from industry descriptions and processes the resulting matches using a probabilistic framework. We compare the results of this ALP concordance to existing technology concordances. Based on these comparisons, we discuss advantages of this approach relative to conventional approaches. ALP concordances provide a meso-level mapping to industries that complements existing macro- and firm-level mappings – and open new possibilities for empirical patent analysis.

1. Introduction

International technological diffusion is an important driver of technological change, which is in turn a key determinant of cross-country differences in income and economic growth (Romer, 1990; Aghion and Howitt, 1992; Grossman and Helpman, 1991; Keller, 2004). International trade and foreign direct investment are often considered to be key catalysts of technology transfer (Coe and Helpman, 1995; Eaton and Kortum, 2002; Branstetter et al., 2006; Acharya and Keller, 2009), but directly studying this process is often hampered by the fact that measuring transferred technology empirically is challenging. While patent data often serve as useful proxies for technological change (Griliches, 1990; Basberg, 1987) and diffusion (Jaffe et al., 1993), fully exploiting patent data in economic analyses would require that patents be linked to economic activity at a level of disaggregation that allows for different technological, industrial and spatial patterns. Such a detailed link between technological and economic activity would further improve our assessment of policies that aim to promote innovation, as well as assess the relationship between technological change and economic development. In this paper, we propose an algorithmic approach to constructing such a link.

Patent statistics have frequently been used as both technological and economic indicators due to the widespread availability of patent data and the assumption that patents reflect direct inventive activity and innovation. Basberg (1987) describes how patents have been incorporated into innovation models to measure technology diffusion and to evaluate the output of research activity. In a similar survey, Griliches (1990) documents how patents have been used as economic indicators for inter alia R&D output, stock market activity and total factor productivity. Within this literature, however, the empirical validity of patents as technological or economic indicators remains a matter of debate largely because of concerns about how patents are used, enforced and valued differently across

☆ *Note*: The ALP concordances described in this paper can be downloaded from the WIPO website at http://www.wipo.int/econ_stat/en/economics/publications.html.

☆☆ We thank Prantik Bhattarchayya for superb research assistance. We are grateful for the assistance and guidance provided by researchers and programmers at the World Intellectual Property Organization (WIPO), including Carsten Fink, Hao Zhou, Christophe Mazenc, Sacha Wunsch-Vincent, and others. We thank seminar participants at WIPO and Morrison Foerster and attendees at the 2012 "Patent Statistics for Decision-Makers" Conference at the OECD-Paris. We also acknowledge the financial support the project received from the National Science Foundation. All opinions and views expressed are those of the authors and do not represent those of the NSF or the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

* Corresponding author.
E-mail address: tlybbert@ucdavis.edu (T.J. Lybbert).

0048-7333/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.respol.2013.09.001

different industries, jurisdictions and time periods. We believe that more disaggregated analyses of patent statistics – particularly when matched with equally disaggregate economic data – will alleviate some of these concerns and open new research possibilities.

In general, there are three levels at which patents can be linked to economic activity. At the coarsest macro-level, aggregate patent data taken from a specific country in a specific year can be associated with aggregate economic data, respectively. Linking patent and economic data at this aggregate level is based simply on the country-year unit of analysis and has enabled research on questions such as measuring the rate of innovation (Porter and Stern, 2000), a country's innovative capacity (Furman et al., 2002) and the effects of patent harmonization (McCalman, 2001). Analyses of foreign patent flows and economic activity (Eaton and Kortum, 1996; Xu and Chiang, 2005; Falvey and Foster, 2006; Harhoff et al., 2009) is similarly based on an aggregate association of patents to economic data through a shared country-year unit of analysis.

At the finest level, patents and economic activity can be linked at the firm-level. While this micro-linkage between patent and economic data enables rigorous and insightful research on patenting as part of firm-level strategies (Brouwer and Kleinknecht, 1999; Austin, 1993), constructing and maintaining such a firm-level database requires substantial effort, is only feasible for a fraction of the firms represented in patent databases, and may miss broader considerations regarding related products, competitors and industrial dynamics. Although progress will continue to be made at this level, these limitations constrain our ability to link patents to economic activity at the firm-level, especially in emerging economies where firm-level data is relatively sparse.

Between these macro- and micro-level linkages is a meso- or industry-level linkage that associates patents and economic data based on the domain of goods and services they represent. At this level, patents on biomedical and semiconductor inventions, for example, are linked to industry or product classes that use biomedical and semiconductor inventions, respectively. We argue that a robust industry-level linkage – perhaps in conjunction with macro- and micro-level analyses – will enable researchers to better understand the relationship between patenting and economic activity over time and across space and technology classes. Most industry-level linkages are based on concordances. For example, the Yale Technology Concordance (YTC) (Kortum and Putnam, 1997) links the International Patenting Classification code (IPC) to the Canadian Standardized Industrial Classification system. Thus, with the YTC a researcher can link patent data organized by IPC, country and year to the value of production organized by Canada SIC, country and year. Unfortunately, conventional concordance approaches like the YTC suffer from a host of flaws that limit their usefulness in empirical research. After describing these limitations, we propose new methods for constructing concordances and, thereby, industry-level linkages between patent and economic data. These methods use data mining and probabilistic matching to build links that can be applied broadly or narrowly across time and space, can be easily updated, and can create direct linkages between patent data and a variety of economic classification schemes.

We refer to the general approach we develop in this paper as an Algorithmic Links with Probabilities (ALP) approach to constructing concordances. This approach identifies patents in the PATSTAT database that contain keywords extracted from industry classifications in the text of the title and abstract. Tabulated by IPC code, these retrieved patents reveal frequency matches between the industry and IPC classifications. We then process these frequencies to generate a probabilistic mapping that works in two directions: from IPC to the industrial classifications and vice versa. Researchers can use these direct ALP concordances for industry and technology-level analyses of the relationships between patents and economic activity organized by different classification schemes such as SITC, ISIC, North American Industry Classification System (NAICS), and Harmonized System (HS). Given that these methods require minimal manual or subjective intervention, the concordances they generate are also easy to update and refine when new patent data becomes available or when classification systems undergo revisions as they frequently do.

After providing a brief background of related patent concordance research, we discuss the prevailing IPC concordances in some detail and describe their limitations when applied to economic data. We then describe our ALP approach to constructing more useful concordances and generate IPC concordances for both trade (SITC) and industry (ISIC) classification schemes. To test our approach, we compare the ALP concordance with two prevailing concordances, including the YTC.

2. Background

Patents are a potentially powerful data source for technology and innovation analyses because the patents themselves contain a wealth of information, including the names of the inventee, date, prior art, technologies used, as well as a full description of the embedded technology with numerous figures and references. Recently, there has been a large push initiated by the private sector to develop novel ways of analyzing, organizing and making this patent information accessible to firms interested in exploiting or diversifying their patent portfolios and formulating R&D strategies (Trippe, 2003; Moehrle et al., 2010). This form of patent analysis – called "patinformatics" – aims to reveal relationships between individual patents and broader technological fields in order to inform commercial, legal and policy decisions and includes grouping similar concepts and technologies, creating patent landscape maps, tracking the evolution of these maps over time, and analyzing and interpreting citation networks. These approaches typically use recent developments in text analysis and text clustering software, and then use the findings from these programs to create different visualization and mapping schemes (see Yang et al., 2008, for a good overview of the various softwares that exists). Among some of the studies that use patent-based indicators and relate it to economic phenomena are the numerous patent citation studies used to assess everything from knowledge spillovers (Jaffe et al., 1993; Jaffe and Trajtenberg, 1999), firm evaluation (Hall et al., 2005) and institution types (Trajtenberg et al., 1992; Jaffe et al., 1998). Other studies have used patent-based indicators by tabulating IPCs to assess the networks of technologies (Leydesdorff, 2008) and quantifying inventor competence (Moehrle et al., 2005).

While the methods we develop are conceptually similar to these tools and could ultimately provide a valuable economic layer to patent landscapes, networks and other patinformatic analyses, the ALP concordances we construct are designed to go beyond descriptive analysis and to enable more rigorous econometric analysis at the industry-level. By doing this, we continue to build on other efforts to link patent and economic data through technology-industry associations. While these industry-level linkages are facilitated by the fact that the IPC and economic classification systems share a detailed hierarchical structure, they are complicated by the fact that these classification systems are motivated by different objectives. Whereas economic classification systems are intended to disaggregate goods and services into meaningful and related sub-groups, the IPC system is intended to facilitate the patent examination process by enabling patent examiners to precisely identify the novel technical features of the disclosed invention and to define the prior art against which they can assess novelty. Since goods or services in very different economic classifications can use the same technical feature (e.g., an electronic motion control device may be used in washing machines and satellites), this difference in intended usage implies that linking patents

to economic data through a concordance of their respective classification systems is never straightforward. Whereas one could manually construct a one-to-one concordance between two industrial classification schemes that share the same unit of analysis (i.e., industry), constructing a concordance between the IPC and an economic classification at any useful level of resolution is effectively a many-to-many mapping that may not be amenable to a manual approach.

The first attempt to link patent data with industry data was conducted by Schmookler in 1966 (Comanor and Scherer, 1969) who assigned "industries-of-use" to patents organized by the US patent class (USPC). The classification scheme used in this early concordance assigned patent classes to industries where at least 2/3 of patents in that class were used for that particular industry. A later concordance developed by a branch of the US Patent and Trademark Office (USPTO) used a similar methodology and assigned equal weighting to patent classes which related to multiple industries. The first comprehensive concordance, the YTC, emerged in the early 1990s (Evenson and Putnam, 1994; Kortum and Putnam, 1997). The YTC was constructed by leveraging a useful feature of the roughly 250,000 patents issued in Canada between 1978 and 1993. For each of these patents, the Canadian Patent office examiners were required to assign a technology field from the IPC system (standard practice worldwide) and to indicate the Industry of Manufacture (IOM) and Sector of Use (SOU) of the invention according to the Canadian Standard Industrial Classification (1980 cSIC-E Version). The patents examined in this window implicitly concord IPC to cSIC since examiners were assigning patents to both systems concurrently. The YTC tabulated these assignments to make this an explicit IPC-cSIC concordance.

Because it is based on assignments made by patent examiners – presumably, experts in the field – the YTC is ostensibly based on hundreds of thousands of hours of thoughtful, expert consideration. Furthermore, this structure implies that the YTC comprehensively covers all technologies and industries included in the 250,000 patents that were cross-classified. An additional benefit is that the YTC uses probabilistic rather than subjective weights, which allows for the same technical feature to be used in multiple sectors. On the other hand, the YTC suffers from some serious limitations.[1] First, it is only possible to directly link to one classification system, the cSIC, which is not commonly used in industry-level studies. Bridging to any other economic classification system introduces noise and can hopelessly atrophy the resulting composite concordance (as discussed below). Second, it is frozen in time and space, as it were, because it will always be based on Canadian patents examined between 1978 and 1993. This introduces potential technological, temporal and spatial biases (Schmoch et al., 2003).

3. The IPC and Prevailing IPC-Industry Concordances

In this section, we describe in more detail the structure of the prevailing concordances that attempt to link the IPC to industry classification systems. First, we describe briefly the structure of the IPC system and contrast it with existing economic classification systems. We then differentiate between the prevailing concordances that build on the YTC and those that chart a different path entirely.

The IPC was established in 1971 by the Strasbourg Agreement to provide a harmonized, language independent, hierarchical system for classifying technology embedded in patents and utility models.[2] Given its role in defining the scope of prior art considered in patent examination, the IPC is a central feature to the global network of national patent systems. The current version of the IPC divides technology into eight sections, which are further divided into a total of nearly 70,000 "subgroups". To illustrate the structure of the IPC, consider the example of IPC "subgroup" B64C 11/18, which covers "Aerodynamic features of propellers used in aircraft." This group number is composed of section B ("Performing operations; Transporting"), class B64 ("Aircraft; Aviation; Cosmonoautics"), subclass B64C ("Aeroplanes; Helicopters"), main group B64C 11/00 ("Propellers"), and subgroup B64C 11/18. We construct our concordance at the four-digit subclass level (e.g., B64C, A21B, etc.), of which a total of 639 exist (in the most recent version). In terms of how the IPC is used in practice, patent examiners around the world classify the inventions claimed by the patents they examine. Where multiple inventive features are evident in an invention, examiners often cross-list the patent in multiple IPCs.[3]

With this brief description of the IPC in mind, consider the structure of existing IPC-industry concordances. Two of these concordances, the "DG Concordance" (Schmoch et al., 2003) and the MERIT Concordance (Verspagen et al., 1994), chart a different path than the YTC. Both of these concordances attempt to match IPC subclasses to ISIC industry classifications using the official descriptions of these respective categories. In order to do this manually, both efforts are based on one-to-one matches, which is only feasible at a relatively coarse resolution. Specifically, the DG concordance assigns 625 IPC subclasses to one of 44 different manufacturing sectors, of which one or more ISICs are associated. The MERIT Concordance matches IPC subclasses to 22 industrial classes based on a mix of 2- and 3-digit ISIC codes. Both approaches are notable for their attempt to manually and directly (i.e., one-to-one) translate the IPC to the ISIC industry classification system. While the mapping to the ISIC that emerges from these efforts is undeniably coarse, it can nevertheless enable some useful empirical and policy analysis.

For more rigorous analysis, higher resolution economic data can be particularly useful – but leveraging these higher resolution data requires a higher resolution concordance. To construct a higher resolution concordance, researchers have had little choice but to trod the YTC path and rely on the same narrow base of Canadian patents. Two other prevailing concordances take this approach and seek to build on the YTC. Specifically, the OECD Concordance (Johnson, 2002) and PATDAT Concordance used by Silverman[4] simply layer an additional concordance to translate the IPC to more commonly used industry classification systems such as ISIC (used in OECD) and the US Standard Industrial Classification (SIC) (used in PATDAT). This conventional composite concordance approach introduces additional complications, such as causing the strength of the technology-industry linkage to atrophy. To illustrate this problem, Table 1 takes a random IPC subclass, B64D "Aircraft; Aviation; Cosmonautics Equipment for Fitting In or To Aircraft", and shows what happens during the layering process. Whereas the initial concordance is sensible, the composite concordance has clearly atrophied – even when the additional concordance layer (cSIC-ISIC in this case) is itself quite robust. Obviously, the severity of this problem intensifies with additional concordance layers.

In summary, any effort to analyze the relationship between patents and economic activity at the industry-level faces a serious concordance dilemma. While there is rich, high resolution data for

[1] In addition to the two limitations described here, it seems that examiners did not cross-classify all patents. While it is unclear how they selected which patents to cross-classify, this selection potentially biases the associated YTC.

[2] For a complete guide to the IPC, including useful training resources, see http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc_2009.pdf. To explore the IPC interactively with complete notes see http://www.wipo.int/ipcpub.

[3] In some jurisdictions, examiners must designate a primary IPC and list the remaining IPCs as secondary. The PATSTAT database compiles patent data from many jurisdictions, only some of which follow this convention, so a primary IPC designation is not always available when multiple IPCs are listed on a patent.

[4] See http://www.rotman.utoronto.ca/~silverman/ipcsic/documentation_ipc-sic_concordance.html for documentation and procedure.

Table 1
Concordance for IPC subclass B64D which is "aircraft; aviation; cosmonautics/equipment for fitting in or to aircraft".

Initial concordance: IPC-cSIC		Composite concordance: IPC-cSIC-ISIC	
Description	Weight	Description	Weight
Aircraft and Aircraft Parts Industry	43.2%	Manufacture of other fabricated metal products; metal working service activities	10.8%
Other Communication and Electronic Equipment Industries	9.4%	Manufacture of motor vehicles	10.8%
Other Machinery and Equipment Industries	6.3%	Manufacture of bodies (coachwork) for motor vehicles; manufacture of trailers and semi-trailers	10.8%
Indicating, Recording and Controlling Instruments Industry	5.8%	Steam and air conditioning supply	10.8%
Other Textile Products Industries	5.0%	Freshwater fishing	1.4%
Electrical Switchgear and Protective Equipment Industry	2.9%	Marine aquaculture	1.4%

both patents and economic activity, and these data would seem to enable a host of insightful empirical analyses, jointly harnessing the high resolution on both sides requires a robust, accurate and high resolution concordance. Manual, one-to-one concordances are too crude for many research questions, but up-to-date more sophisticated concordances have little choice but to build on a relatively narrow set of Canadian patents that are effectively frozen in time, space and technology. Furthermore, since very few (if any) datasets are described with the cSIC classification system, additional concordance layers are required to construct more broadly useful concordances from this narrow patent base, which quickly atrophies the integrity of the concordance.

4. Guiding principles and methodology

To escape the dilemma described in the previous section, an ideal concordance would replicate the human process of reviewing each patent and assigning industry codes based on the information contained within the patent, while also including a much broader set of patents from around the world, allowing for direct translation into multiple economic classification schemes, and facilitating updates to reflect technological and classification system changes. In this section, we formalize a set of guiding principles based on this ideal and then describe the methods we develop to approximate an ideal concordance according to these principles.

Three principles have guided our effort to approximate an ideal concordance to link patents to economic activity:

1. *Use the descriptive content of individual patents as the basis for the concordance*. Since technical features classified in the IPC can pertain to several different classes of economic activity, it is important to consider each patent individually. An ideal concordance would be based on an effective evaluation of the content of each patent, including how and where the underlying invention may be used. The patent applicant is best suited to assess the potential uses of the invention and, in most jurisdictions, has an incentive to discuss this industrial usefulness in the application.
2. *Eliminate the need for concordance layering by constructing direct concordances*. To avoid the composite concordance problem, we aim to devise methods that can be directly applied to the most common economic classification schemes, including SITC (Rev. 2 and 3), ISIC (Rev. 2, 3, 3.1 and 4), NAICS, HS and SIC. As new versions of these concordances or the IPC are released, new direct concordances are preferable to indirect ones that update the older to the newer version via a concordance.
3. *Automate the construction process as much as possible*. Technology changes rapidly, and the concordance should reflect these changes. A proper concordance will therefore need continuous updating to reflect new technologies as they emerge. Automating the process implies that it should:
   a. *Involve minimal manual work in order to rapidly process millions of patents at a time*. The process should not require, for example, manually sifting through patents or classification schemes.
   b. *Be relatively easy to implement and flexible enough to capture changing technologies and industries*. Through the process, generating a new version of the concordance should be relatively cheap and easy to do. The process should also be flexible enough to allow for adjustments in the technological focus or years considered to tailor the concordance as needed.
   c. *Rely more on objective algorithms than subjective judgments*. This helps to reduce the manual workload of constructing the concordance, but can also provide a critical objective basis on which to construct the weights in a many-to-many concordance.

The ALP methodology we describe below is guided by these principles. Programs that perform tasks such as keyword extraction and text mining allow for specific bits of information to be extracted from individual patents, making it possible to approximate a manual assignment of industry classifications. As with any algorithmic search technique, our methods cannot perfectly replicate careful manual inspection and assignment, but because they can sift through millions of patents they may be able to converge on accurate implied linkages. Because our ALP approach statistically relies on the Law of Large Numbers, we expect the resulting concordances to improve as the number of patents processed increases.

Patents are a natural candidate for mining and clustering techniques because of the wealth of information they contain. We use the PATSTAT database available from the European Patent Office (EPO) as the source of our patent data. The PATSTAT database contains patent data for 86 countries since 1990 and contains details for more than 100 million patent applications, some of which relate to the same invention in different jurisdictions. Included in this database are almost 20 million unique patent abstracts and titles. In contrast, there is no comparable information-rich source of qualitative data on economic activity by industry or product classification. We exploit the only source of qualitative information available by economic classification: the brief descriptions used to characterize a particular industry, product or service category.

To showcase these mining and matching methods, we focus on directly mapping four-digit IPC subclasses to four-digit SITC and ISIC classifications and vice versa. This same process can be replicated for other classification schemes such as HS, NAICS, SIC, and the Central Product Classification (CPC). The next two sections describe our preferred ALP approach. Appendix B briefly describes a second 'probability matching' approach that stems from the same general ALP methodology.[5]

[5] This alternative approach extracts keywords from the patents and then matched these keywords to industry descriptions. Based on the tests we run on the resulting concordances in later sections of this paper, this probability matching approach contains more noise and is therefore less useful than our preferred 'data mining' approach. As described in the appendix, however, with improvements this probability matching approach could be quite promising.

Table 2
Example search terms used for SITC Industry Descriptions.

SITC code	SITC full description	Search terms	"Not" search terms
8484	Headgear and fitting thereof	"Headgear", "Head Gear", "Helmet"	
8510	Footwear	"Footwear"	
8710	Optical instruments and apparatus	"Optical Instruments", "Eyeglasses"	
8720	Medical instruments and appliances	"Medical Instrument", "Medical Appliance"	
8731	Gas, liquid and electricity supply or production meters; etc.	"Gas Meter", "Liquid Meter", "Electric Meter"	"Part"
8732	Counting devices non-electrical; stroboscopes	"Counting Device", "Stroboscope"	"Part", "Electric"
8741	Surveying, navigational, compasses, etc., instruments, nonelectrical	"Surveying Equipment", "Surveying Instrument"	

4.1. Methodology

Our preferred ALP approach relies on data mining the patent abstracts and titles included in the PATSTAT database using keywords from the industry classification descriptions. Specifically, we generate search terms for each industry description and then comb through the all of the patent text (which in this case includes titles and abstracts) for these search items. We generate a list of all the patents where these search terms were found, and from these patents, we are able to compile a frequency of IPC subclasses that are matched to every industry. The results are then reweighted to reduce noise and possible bias.

We extracted search terms corresponding to each 4- and 5-digit SITC and ISIC industry descriptions provided by the United Nations.[6] In most cases, these industry descriptions consist of a single sentence that lists numerous products/services that fall into the category. We use a combination of algorithmic and manual techniques to extract keywords that retrieve patents that are specific, relevant to the corresponding economic category and robust to standard keyword construction concerns (e.g., plurals, word phrases, etc.). In order to expand this initial set of keywords to include relevant synonyms, we used the Cross-Lingual Expansion tool embedded in WIPO's PATENTSCOPE.[7] This tool is ideal for our purposes because it generates synonyms based on the full text of patents in different languages and therefore expands our keywords based on words and phrases that are commonly used in patent documents. We then manually inspect and refine the resulting set of keywords. The final result is a list of one to dozens of keywords corresponding to each 4-digit classification with additional "not" terms.[8] Table 2 provides an example of the search terms generated for a grouping of SITC industry codes.

Once the search terms are generated, we then query the PATSTAT database using these terms and retrieve patents that contain the exact phrases of each search term in either their title or abstract. Although it would be straightforward to limit this query to specific countries, regions or years if the specific use of the resulting ALP concordance necessitated such specificity, we impose no such restrictions in order to retrieve a pool of relevant patents that is as large and varied as possible.[9] Finally, we tabulate the IPC subclasses listed on the retrieved patents to generate raw frequencies on which to base probabilistic linkages from SITC to IPC and vice versa.

4.2. Weighting schemes

To illustrate how we process these raw frequencies to generate ALP concordances, it is useful to introduce some notation and a simple stylized example. We use m_{ij} to denote the total number of patents from technology (i.e., IPC) class j retrieved using the keywords for industry i. Assuming $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, our tabulated frequencies will generate an $I \times J$ matrix of match combinations. The total number of matches, $\sum_i \sum_j m_{ij}$, in this matrix is not tied to the number of patents in the queried database because a single patent can be retrieved several times[10] and some patents may never be retrieved.

Let M_i be the total number of matches for industry i and N_j be the total number of matches for technology j. Given the number of matches by industry and technology, we assign probabilities to each ij combination using Bayes rule, which takes into account the number of possible technologies that exist and how frequently each technology class is matched to a given industry. Table 3 provides a concrete, stylized example with two industries (1 and 2) and two technologies (X and Y). The raw match counts tell us that both industries rely heavily on technology Y. Furthermore, a small proportion of technology Y (X) patents are matched to industry 1 (2). Our goal is to reweight the matches in such a way that minimizes Type I errors and takes into account both the raw frequencies and the specificity of each technology class. As we reviewed our results, we found that the number of times a patent was matched coincided with how broad/specific the technology class was. Since each industry has a defining (or specific) characteristic that separates that industry from others, we found the need to rebalance our results to consider the technologies that are specific to the industry and place less emphasis on the broad technologies used across numerous industries.

[6] To download of each SITC and ISIC revision industry description, see http://unstats.un.org/unsd/cr/registry/regdnld.asp (accessed March 2013).

[7] This tool is available here: http://www.wipo.int/PATENTSCOPE/search/clir/clir.jsp?interfaceLanguage=en (accessed 23.04.13).

[8] This process involved a basic tradeoff. On the one hand, we would like to include as many patent matches as possible to ensure proper coverage of the industry. However, increasing the scope of possible matches tends to introduce more noise and reduced accuracy (increase potential for Type I errors). Therefore, the process requires careful treatment and we remove all terms that have multiple meanings or are considered too general (for instance, the word "machine" or "tool"). We also incorporate the use of "not" terms, since many industry descriptions include "not elsewhere specified" or refer to a particular sub-group within an industry. On the other hand, we do not want to be too restrictive and miss so many possible matches so that the actual matches are biased. The keyword searches are the only part of the process that requires "manual" updating. During the development of this process, we experimented with keyword extraction programs and synonym programs to automate this process, but found the Type I error rates to be too high due to the reasons explained above. In some cases, the synonyms were not very reliable, or the keyword extraction algorithm extracted the incorrect keyword (for instance if the word followed "Not included") and if the word proved to be too general. The advantage of this method is that we can continue to refine this process and gather feedback from users to increase the validity and accuracy of the process. On the other hand, this process is rather time consuming and requires substantial subjectivity.

[9] All of the search terms are in English, whereas some of the patents that are being queried are in different languages. In this sense, there is a bias toward North American and European patents. It is possible to query the database using other languages, but it is unclear what the value added of this extension would be.

[10] Many patents list multiple IPCs, which increases multiple counting. A patent that lists three IPCs and is retrieved for a given industry keyword query, for example, creates three match observations, one for each IPC.

Table 3
Stylized example of weighting schemes.

Industry i	Technology j	m_{ij}	M_i	N_j	Industry-to-technology			Technology-to-industry		
					W_{ij}^R	W_{ij}^S	W_{ij}^H	W_{ji}^R	W_{ji}^S	W_{ji}^H
1	X	98	998	100	0.0982	0.9151	0.5400	0.9800	0.9977	1.0000
1	Y	900	998	9900	0.9018	0.0849	0.4600	0.0909	0.4742	0.0827
2	X	2	9002	100	0.0002	0.0215	0.0000	0.0200	0.0023	0.0000
2	Y	9000	9002	9900	0.9998	0.9785	1.0000	0.9091	0.5258	0.9173

4.2.1. Raw weights

Let A_j be the outcome of being matched with technology j and B_i be the outcome of being matched with industry i. Bayes rule gives the probability of A_j conditional on observing B_i as

$$Pr(A_j|B_i) = \frac{Pr(B_i|A_j)Pr(A_j)}{Pr(B_i|A_1)Pr(A_1) + \ldots + Pr(B_i|A_J)Pr(A_J)} \quad (1)$$

For our example, the conditional probability that Technology X is relevant to Industry 1 is

$$Pr(A_X) = \frac{N_X}{N_X + N_Y} = \frac{100}{100 + 9900} = 0.01$$

$$Pr(A_Y) = \frac{N_Y}{N_X + N_Y} = \frac{9900}{100 + 9900} = 0.99$$

$$Pr(B_1|A_X) = \frac{m_{1X}}{N_X} = \frac{98}{100} = 0.98$$

$$Pr(B_1|A_Y) = \frac{m_{1Y}}{N_Y} = \frac{900}{9900} = 0.\overline{09}$$

$$W_{1X}^R = Pr(A_X|B_1) = \frac{(0.98)(0.01)}{(0.98)(0.01) + (0.\overline{09})(0.99)} \cong 0.098$$

This "Raw" weight W^R implies that roughly 10% of the technologies that match to industry 1 are from technology X.

4.2.2. Specificity weights

An alternate weighing scheme corrects for how specifically or broadly a given technology class maps to industries. This correction can be important because there are more and more diverse patents filed under some technology classes than others. Thus, some technology classes are widely matched across a broad set of industries while others are narrowly matched to a few specific industries. In such a setting, some matches may contain more information about relevant industry-technology linkages than others. Matches with high specificity within a given technology classes may indicate particularly important linkages relative to broad classes that have many more total matches across many industries. To correct for technological specificity, we impose the condition that each industry has the same *ex ante* probability of matching with each of the J technology classes – i.e., $Pr(A_j) = 1/J$. This has the effect of overweighting classes with relatively low $Pr(A_j)$ and underweighting those with relatively high $Pr(A_j)$.

$$W_{ij}^S = \frac{Pr(B_i|A_j)(1/J)}{Pr(B_i|A_1)(1/J) + \ldots + Pr(B_i|A_J)(1/J)} \quad (2)$$

This "Specificity" weight W^S reflects how specifically or how broadly a given technology class matches to industries. For technologies that are very specific to a given industry (i.e., most of the total matches for the technology is found in that particular industry), we expect their weight to increase relative to the raw weights. On the other hand, for technologies that are not specific to the industry, we expect their weight to be reduced. In our running example, the specificity weight of industry 1 with technology X is nearly an order of magnitude bigger than the raw weight above:

$$W_{1X}^S = \frac{(0.98)(0.5)}{(0.98)(0.5) + (0.\overline{09})(0.5)} \cong 0.915$$

In this case, this specificity correction dramatically increases the weight because a relatively high proportion of technology X's patents match *specifically* to industry 1. In other words, this weighting scheme overweighs narrowly relevant IPC subclasses relative to broadly relevant IPC subclasses. More obscure or specific technologies tend to have smaller values of N_j, which increases their specificity weights relative to widely matched technologies.

4.2.3. Hybrid weights

Raw weights take direct matches at face value and may lead to high Type I errors for technology classes that match to a broad range of industries. In other words, the raw weights tend to reward the high frequency technologies without penalizing them for being too broad. To reduce these errors, specificity weights discount technologies that are dispersed broadly among industries and increase the influence of technologies that map to very specific industries. However, the cost of reducing Type I errors is to increase Type II errors by completely discounting widely matched technologies. As a balanced alternative, we formulate an ad hoc "Hybrid" weighting scheme that blends the raw and specificity weights using $Pr(A_j) = W_{ij}^R/J$, which yields

$$W_{ij}^H = \frac{Pr(B_i|A_j)(W_{ij}^R/J)}{Pr(B_i|A_1)(W_{i1}^R/J) + \ldots + Pr(B_i|A_J)(W_{ij}^R/J)} \quad (3)$$

Note that this is not a weighted average of the raw and the specificity weights. Instead, this hybrid weight accounts simultaneously for both widely matched and narrowly specific technologies. For our example, the hybrid weight of technology X in industry 1 is given by

$$W_{1X}^H = \frac{(0.98)(0.0491)}{(0.98)(0.5)(0.0491) + (0.\overline{09})(0.4509)} \cong 0.54$$

This hybrid weight rewards specificity by increasing the weights of the specific technology, but not at the cost of completely discounting the frequent technology. In cases where the technology is both frequent and specific, the weights will always be higher relative to the raw weights. In other cases where the technology is specific and infrequent, or frequent and broad, the hybrid weight will mediate between the raw and specificity weights.

We can use this same methodology to generate raw, specificity and hybrid weights to map in the reverse direction – from technology-to-industry – by simply switching N_j and M_i in the formulas above. For comparison, Table 3 includes all three weights for our running example and for each possible industry-to-technology and technology-to-industry mapping. Based on the hybrid weights, industry 1 relies 54% on technology X and 46% on technology Y, while industry 2 relies 100% on technology Y. From the technology-to-industry hybrid weights, 100% of technology X is utilized by

Table 4
Comparison of different cutoff levels for SITC Rev. 2 to IPC ALP concordance.

	Cutoff level				
	0%	1%	2%	5%	10%
Total number of different industries	871	871	871	871	867
Total number of different IPCs	629	611	582	528	468
Average number of IPCs to industry	206.51	8.45	5.70	3.25	2.05
Max number of IPCs to industry	625	29	18	8	5
Min number of IPCs to industry	2	1	1	1	1
Hybrid weight of IPCs					
Average	0.0048	0.1184	0.1753	0.3073	0.4857
Herfindahl index	0.3982	0.4430	0.4704	0.5427	0.6583
Max	1.0000	1.0000	1.0000	1.0000	1.0000
Min	4.6E-12	0.0101	0.0204	0.0540	0.1046

Table 5
IPC frequency for industry group, "Headgear and Fitting Thereof" (SITC 8484).

IPC	Untrimmed	Trimmed (2% cutoff)			IPC description
		W^R	W^S	W^H	
A42B	0.4310	0.6977	0.5345	0.9783	Hats; head coverings
A42C	0.0148	–	0.2338	–	Manufacturing or trimming hats
A62B	0.0516	0.0835	0.0993	0.0217	Devices for life-saving
A61F	0.0529	0.0856	–	–	Stents, orthopedic devices
G02B	0.0393	0.0636	–	–	Optical elements
B68B	0.0008	–	0.0512	–	Harness; whips or the like
F41H	0.0175	–	0.0508	–	Armor; camouflage
A61M	0.0219	0.0355	–	–	Devices for administering food or medicine
A41D	0.0211	0.0342	–	–	Outerwear
B63C	0.0158	–	0.0304	–	Life-saving in water

industry 1 and 92% of technology Y is utilized by industry 2.[11] Although each weight has conceptual advantages and disadvantages, whether one is superior than the others is an empirical question we address below.

4.2.4. Cutoff condition

As the final step in processing the matches that emerge from our keyword queries, we use a cutoff condition to reduce Type I errors. To do this, we impose a cutoff weight (e.g., 2%) and set all weights below this cutoff to zero, then renormalize the remaining positive weights so they sum to one. Imposing a cutoff condition decreases the noise introduced by rare, idiosyncratic matches (i.e., infrequent, nonspecific matches) and isolates more common mapping patterns. Although trimming the raw results in this way is statistically compelling, the actual cutoff used is ultimately arbitrary. To characterize the effect of different cutoff levels, we present different results for different cutoffs.

After generating hybrid weights for mapping IPC subclasses to 4-digit SITC classes, we impose four different cutoff levels, renormalize these weights and summarize the comparison of these cutoff levels in Table 4. Introducing cutoffs significantly reduces the average number of technologies associated with each industry. With no cutoff (0%), each industry utilizes more than 200 different IPCs on average, which is likely to be unrealistic. Once we introduce even a 1% cutoff, the average number of IPCs per industry falls to single digits. Deciding which cutoff level to use involves a basic tradeoff between including enough different technologies within each industry to ensure proper coverage and including too many rare or noisy matches. As a default, we use the 2% cutoff level.[12] Once the cutoff is implemented and the little-used technologies are discarded, we renormalize the weights so that they sum to 1.

To illustrate the results of the full process, we provide the results for SITC code 8484, which is described as "Headgear and fitting thereof", in Table 5. We first queried the PATSTAT database using the search terms found in Table 2. This initial query yielded 11,660 unique patents that listed 379 unique IPCs. We then reweighed the matches using the methodology described, expunged of the low-frequency IPCs and renormalized. To construct a full ALP concordance for IPC to SITC, we repeat these steps for each SITC description. The same ALP methodology can be directly applied to other classification systems without the need for layering concordances.

5. Comparison with existing concordances

In this section, we construct ALP concordances that are structurally comparable to two existing concordances and use these comparisons to test the relative performance of the ALP concordances. Given that the two comparison concordances – the YTC and the DG concordance – are structurally very different, we view these tests as complementary. Specifically, we consider the YTC test to be the best high resolution test of how well the ALP approach can match careful human classification since it is based on patent examiners' classification of patent applications into high resolution industrial categories and provides probabilistic weights that are directly comparable to ALP concordances. The DG concordance

[11] Note that in the industry-to-technology bridge, the hybrid weight appeared to mediate between raw and specificity weights. This is due to the equal importance of frequency and specificity built into the formula. However, in the technology-to-industry bridge, we see that the hybrid weight skewed the raw weights even more for technology Y. This is because industry two was more specific to technology Y than industry one (9000/9002 matches versus 900/998 matches). Therefore, the hybrid would favor Industry 2 for being both more frequent and more specific.

[12] A few reasons underlie this default. Intuitively, we believe that an average of five technologies per industry seems plausible, particularly given the relatively high level of disaggregation used to describe each industry. Second, as we experimented with higher cutoffs, we noticed that the number of different IPCs that met the cutoff criteria diminished by roughly 15–30 per percentage point increase in the cutoff. Not wanting to reduce the total variety of IPCs in the concordance by too much, we elected to stick with the 2%.

Table 6
Cross-tabs of zero and positive values of the YTC and the comparable ALP concordance with hybrid weights.

Sector of use (N = 232,498)				Industry of manufacture (N = 232,361)			
	ALP = 0	ALP > 0			ALP = 0	ALP > 0	
YTC = 0	73.9%	19.7%	93.5%	YTC = 0	75.0%	21.3%	96.3%
YTC > 0	2.5%	4.1%	6.5%	YTC > 0	1.3%	2.4%	3.7%
	76.3%	23.7%	100%		76.4%	23.7%	100%

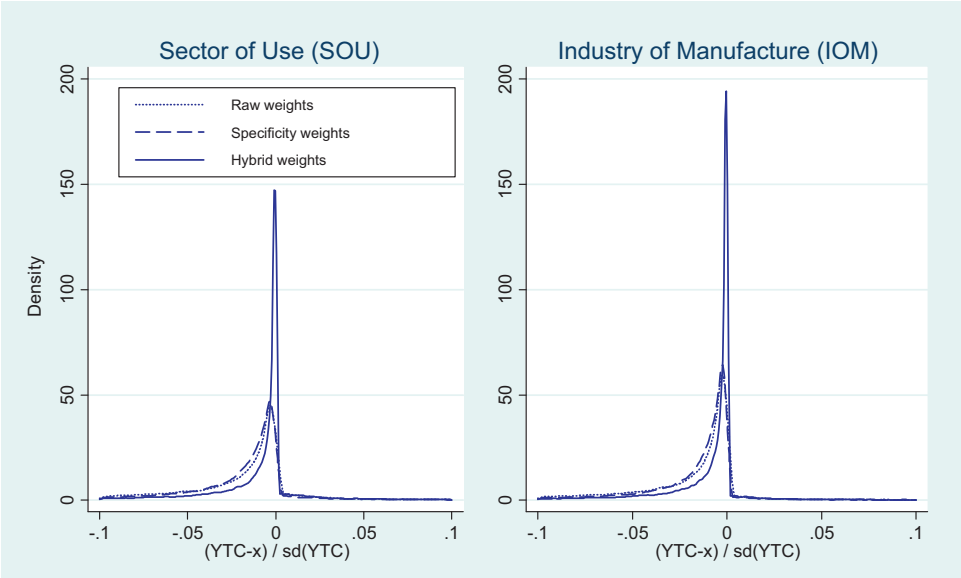

Fig. 1. Kernel densities of differences between the YTC and ALP concordances excluding matching zero values. Differences are measured in YTC standard deviation units.

provides a test of how well the ALP concordance can match more aggregate, one-to-one matches. Since the ALP approach to generating concordances is preferable to alternative approaches for the reasons described earlier, evidence that ALP concordances can statistically replicate these familiar alternatives is strong evidence in favor of the ALP approach.

5.1. YTC comparison

To construct an ALP concordance that is comparable to the YTC, we run the algorithm on the Canadian patents that served as the basis for the YTC. While we cannot identify the exact patents used in the YTC, we can limit our ALP methodology to all the Canadian patents issued in the same time period between 1978 and 1993. This provides us coverage of more than 350,000 Canadian patents and abstracts (30% more than was used in the YTC). We then convert the IPC's from those patents into the Canadian SICs using the ALP methodology. Note that our algorithm is more heavily weighted toward tradable goods, since the specific purpose behind our approach is to convert technology data into specific product-types, which were the most common industry descriptions used to generate search terms. The Canadian SICs are comprised of both tradable and non-tradable goods (e.g. services), so we expect there to be some inherent differences between the two approaches.

We compare the YTC to the ALP concordances based on each of the three weighting options (raw, specificity and hybrid weights). We generate these ALP weights at the 4-digit level of cSIC-E. Since the YTC mixes 3- and 4-digit cSIC concordances with 4-digit IPC, we aggregate both the YTC and ALP concordances to the 3-digit cSIC. Therefore, in all of the comparisons that follow, our ALP results and the YTC results all concord 3-digit cSIC to 4-digit IPC. Lastly, the YTC provides a concordance for both Sector of Use (SOU) and Industry of Manufacture (IOM). Since the current version of the ALP approach cannot differentiate between these two versions of linkages, we test the comparable ALP concordance against both YTC versions.[13]

The first ALP-YTC comparison we conduct is provided in Table 6, a simple cross-tabulation of zero and positive values of the respective results where the off-diagonal elements provide a crude measure of errors. The ALP approach generates matching zero values roughly 75% of the time and matching positive values 2.4–4.1% of the time. Conditional on YTC = 0, the probability that ALP correctly generates a zero weight is 78–79%. When YTC > 0, the probability that this approach correctly generates a positive weight is 62–65%.

Next, we compute the difference between the YTC and our ALP results. Given that the majority of these differences are zero due to matching zero values (see Table 6), we compute these differences across all possible combinations of 3-digit cSIC and 4-digit IPC excluding matching zero values. This provides a strong test of our results against the YTC (since the differences can only be smaller when matched zeros are included). Fig. 1 shows the distribution of these differences in standard deviation (of the YTC) units. Several things are noteworthy in this figure. First, these differences are extremely small relative to the standard deviation of the YTC. Even after excluding matching zero values, the vast majority of these differences are less than 10% of the standard deviation of the YTC. Second, the ALP approach and weighting noticeably affects the fit of the ALP results to the YTC results. Hybrid weights provide the best fit to the YTC. The bulk of these hybrid weight differences are within 3% of the standard deviation of the YTC. Finally, although it is

[13] We are exploring alternative ALP methods that could provide this level of differentiation. Since the benefits of discerning between SOU and IOM loom large, this remains an important (and equally challenging) research objective.

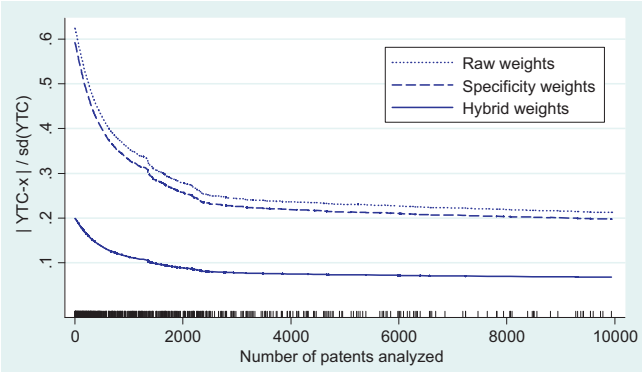

Fig. 2. Non-parametric LOWESS regression of the normalized absolute deviation of ALP results from the YTC (IOM) as a function of the number of patents analyzed (i.e., the number of Canadian patents in the 1978–1993 window by IPC subclass (4-digit)). Tick marks along *x*-axis depict the distribution of the number of patents analyzed.

not clear the differences are significant, the weighted ALP approach appears to better fit IOM than SOU results.

As a final comparison of our ALP results and the YTC, we assess how the fit between the two changes with the number of patents available to process, which is determined by the number of Canadian patents in each IPC subclass (4-digit) from 1978 to 1993. Since the ALP approach is a statistical approach that relies on the Law of Large Numbers, we hypothesize that it will more closely approximate the human classification-based YTC as the number of patents processed increases. For future use of ALP approaches, it is important to demonstrate this pattern and to characterize how the number of patents processed affects the quality of the results. The YTC comparison offers a convenient test of this hypothesis since the number of patents in different IPC subclasses varies widely in these Canadian patents (see *x*-axis in Fig. 2). To exploit this variation, we non-parametrically regress the absolute deviation of the YTC with our ALP results–normalized again by the standard deviation of the YTC–on the number of patents processed. This regression (Fig. 2) confirms that the fit improves as the number of patents processed increases. When the number of patents processed is less 2000, the rate of improvement is very apparent. Beyond this threshold, doubling or tripling the number of patents analyzed does not improve the fit. This result provides a useful benchmark for future applications of the ALP approach, which, incidentally, will almost always include many more patents than are contained in this subset of Canadian patents.

Overall, the comparison of the ALP concordances with the YTC shows some systematic differences that are mainly attributable to the methodological construction of the concordance. Two such differences are noteworthy. First, the differences are asymmetric because the ALP approach is more likely to estimate positive weights when YTC weights are zero than vice versa. This feature is apparent in the asymmetric distributions in Fig. 1 and in convergence to a positive difference instead of zero in Fig. 2 and stems from the fact that algorithmic data mining casting a broader net than precise manual classification. This could imply that the ALP approach is either more prone to Type 1 errors or able to detect relationships that the YTC does not. Second, our concordance matches to tradable classes better than non-tradable classes.[14] While these differences can be seen occasionally at high resolution (e.g., 4-digit), the differences quickly fade with aggregation. There may be more that could be done to refine the matches on the margin, but we expect these improvements to be modest at best and will instead focus our attention on applying the ALP methodology to other trade and industry classifications.

5.2. DG concordance comparison

As a second check, we compare the results of our concordance with the DG Concordance constructed by Schmoch et al. (2003). The DG Concordance linked IPCs to both the NACE and ISIC (Rev. 3) classification system using a one-to-one mapping of 4-digit IPC groups into 44 different manufacturing fields, which are then assigned to one or more ISICs. The assignment of IPCs to manufacturing fields was based on the industry of operation of firms filing patent applications. In some ways, the methodology used is similar to our own. The DG Concordance used more than 3000 applicant firms that accounted for more than 150,000 patents from 1997 to 1999. Once they identified the industry of the firm, they summed up the IPC counts of the patents filed by the firm and assigned the largest IPC weight a one-to-one match (100% weight) with the manufacturing field. The concordance only works in one direction, from IPC to ISIC, with multiple IPCs assigned to one of the 44 manufacturing fields. The manufacturing fields consist of either 2, 3 or 4-digit ISIC classes.

To compare an ALP concordance to the DG concordance, we first generate an IPC-ISIC (Rev. 3) concordance using our ALP methodology and, consistent with the DG concordance, excluded the non-manufacturing ISIC groups.[15] When necessary, we aggregate the resulting concordance to match the 44 industry fields. In addition to our preferred 2% cutoff level with hybrid weights, we experiment with 5% and 10% cutoff levels to cull the weak (i.e., low weight) matches that are explicitly excluded in the DG concordance and give higher weights to the top matches. For the same reason, we also report comparison results for a one-to-one match version of the ALP concordance that assigns the top match for each IPC a weight of 100%. Finally, we report the weighted average (weighted by number of patents in a given IPC subclass) of all the matches within each of the 44 fields in Table 7.

Across all of the DG fields, the average weight of the DG assignments ranges from 0.38 to 0.51, indicating that the DG assignments make up a substantial match in our concordance. The overall correlation between our ALP weights and the DG weights across all IPCs ranges from 0.46 to 0.71. Given the structural differences between these approaches (i.e., one-to-one matching versus probabilistic), these correlations are quite encouraging. This structural difference also seems to be directly responsible for the lowest percentage matches (e.g., in fields 12, 15, 24, 25, 33, 34, 38 and 39). In nearly all of these instances (with the exceptions of field 12 and field 39),[16] the ALP approach matched a large number of IPCs (at least a dozen or more, and in some cases hundreds of IPCs) to only one or two 4-digit ISIC classes. The ALP methodology performs best when the IPC and economic classification are matched at comparable levels of resolution, which is why we match 4-digit IPCs to 4-digit industries

[14] As we pushed further into the comparison with the YTC, we ran some basic fixed-effect regressions on the 4-digit weights to identify any specific differences between certain class levels. We found that our algorithmic approach tends to under-weigh most of the non-tradable cSIC-E (these are cSIC1 greater than 5). This is unsurprising since our algorithm relies most frequently on identifying specific products and goods, and it is much more difficult to match specific services.

[15] We kept ISIC fields 1500–3700, which are classified as manufacturing.

[16] For field 12, the DG concordance only assigned one 4-digit IPC (B27K) to ISIC 2422, which is described as "Manufacturing of paints, varnishes and similar coatings". The results from our concordance assign most of the weight of this ISIC to the IPC "C09D", which is described as "Coating Compositions, e.g. Paints, Varnishes, Lacquers, etc.. . ." Unfortunately, this IPC falls under the DG field 10, which is described as "Basic Chemicals". For field 39, the ALP concordance matched ISIC 3313 with very similar IPCs to the DG assignments. For instance, the DG assignment includes the following IPCs: G01K, G01L, G05B and G08C. The ALP concordance assigns most of the weight to:G01G, G01N, G01R, G01S, G03B and G03G, which are all very similar to the DG assignments.

Table 7
Comparison of the DG and ALP Concordances across industrial field.

Field	Description		ALP concordances (weighted average)			
		DG (1)	2% cutoff (2)	5% cutoff (3)	10% cutoff (4)	1-to-1 match (5)
1	Food	100%	91%	94%	96%	97%
2	Tobacco	100%	100%	100%	100%	100%
3	Textiles	100%	46%	51%	61%	100%
4	Wearing	100%	25%	28%	30%	0%
5	Leather	100%	22%	26%	31%	6%
6	Wood products	100%	45%	51%	57%	63%
7	Paper	100%	72%	75%	80%	96%
9	Petroleum	100%	18%	20%	34%	53%
10	Basic chemicals	100%	45%	56%	65%	84%
11	Pesticides	100%	93%	93%	100%	100%
12	Paint	100%	0%	0%	0%	0%
13	Pharmaceuticals	100%	47%	56%	58%	86%
14	Soaps	100%	89%	89%	90%	88%
15	Other chemicals	100%	10%	10%	9%	0%
16	Man-made fibers	100%	23%	25%	25%	0%
17	Plastic products	100%	12%	15%	18%	10%
18	Mineral products	100%	55%	66%	72%	76%
19	Basic metals	100%	50%	54%	59%	73%
20	Metal products	100%	23%	27%	34%	47%
21	Energy machinery	100%	57%	65%	72%	73%
22	Non-specific machinery	100%	21%	24%	25%	32%
23	Agricultural machinery	100%	33%	36%	39%	36%
24	Machine tools	100%	3%	3%	2%	4%
25	Special machinery	100%	5%	5%	7%	7%
26	Weapons	100%	80%	87%	98%	96%
27	Domestic appliances	100%	44%	54%	57%	78%
28	Computers	100%	48%	57%	67%	70%
29	Electric motors	100%	18%	26%	42%	26%
30	Electrical distribution	100%	24%	35%	27%	6%
31	Accumulators	100%	97%	100%	100%	100%
32	Lightening	100%	41%	62%	92%	94%
33	Other electrical	100%	4%	4%	1%	0%
34	Electronic components	100%	2%	0%	0%	0%
35	Telecommunications	100%	14%	15%	15%	23%
36	Television	100%	38%	44%	62%	98%
37	Medical equipment	100%	13%	15%	20%	39%
38	Measuring instruments	100%	6%	2%	0%	0%
39	Industrial control	100%	2%	3%	0%	0%
40	Optics	100%	24%	27%	37%	57%
41	Watches	100%	97%	100%	100%	100%
42	Motor vehicles	100%	27%	31%	40%	53%
43	Other transport	100%	49%	56%	66%	83%
44	Consumer goods	100%	26%	32%	40%	28%

Field 8 is not included since the DG concordance did not assign any IPCs to Field 8. All ALP concordances use hybrid weights.

rather than to 1-, 2- or 3-digit industries. When there is a mismatch between these resolutions, the resulting ALP weights will either be very small or very large. For example, when there are a large number of 4-digit IPCs assigned to only one or two 4-digit ISICs, they will tend to have much lower weights than the DG assignments because we are taking the average weight of this 4-digit industry across many different technologies. In order to generate a high average weight, the industry would need to be a significant user of dozens or even hundreds of technologies, which is very unlikely.

As a final test, we compare the ALP weights of the DG assigned matches (DG = 1) versus the non-DG matches (DG = 0) in Table 8. Across the 623 IPCs, the average ALP weight of the DG assigned matches ranges from 0.32 to 0.50. The average ALP weights when DG = 0 are all 0.01 or less. Also shown in Table 8, we find that nearly 86% of the DG assigned matches are matched in the ALP concordance at the 2% cutoff level, with an average ranking of 2.87. At this cutoff level, the DG match is listed in the top 3 in almost three-quarters of the IPCs.

Taken together, these comparisons seem to indicate that the ALP concordance provides a strong match to the DG concordance. With this comparison in mind, it is worth noting that there are added benefits to the ALP approach relative to the DG approach. If one

Table 8
Summary of comparison of ALP concordance with the DG concordances where DG = 1 indicates that the DG assigns a particular IPC subclass to an industrial field.

	Mean ALP Weight across all IPCs			For DG = 1		
	DG = 0	DG = 1	*t*-Statistic	% matched in ALP	If matched, mean rank	If matched, % in Top 3
2% cutoff	0.0101	0.32	63.8	85.7%	2.87	74.0%
5% cutoff	0.0094	0.37	>100	70.8%	1.93	86.8%
10% cutoff	0.0087	0.42	>100	57.9%	1.45	98.6%
1-to-1 match	0.0074	0.50	>100	38.4%	1.00	100%

Mean ALP weight is weighted by the number of patents in each IPC subclass.

is interested primarily in the 44 fields contained in the DG concordance, the ALP approach generates a probability structure that in many contexts is preferable to the one-to-one binary matches of the DG concordance. Potentially even more important, the ALP approach provides much more disaggregated linkages that enable economic data to speak to patent data at a much higher industrial resolution if necessary. Finally, the DG concordance only works in one direction allowing technologies to be bridged into industries, where the ALP concordance works in both directions.

6. Conclusion

There is a long and important literature that uses patents to understand the innovation and diffusion of technology. While economists have made important contributions to this field of inquiry, economic analyses of patents have often been constrained by the mismatch between patent and economic data. Using the algorithmic approach we propose to bridging patents and economic data, researchers can analyze important relationships between patents and a wide range of economic activities at an unprecedented level of disaggregation.

There are many policy-relevant questions that could be addressed by joint, high resolution analyses of patent and economic data, including both descriptive exercises (e.g., enhanced patent landscapes) and more rigorous model estimation (e.g., dynamics models of the economic impacts associated with innovation, international technology transfer and patenting strategies, industry evolution, etc.). By making the ALP concordances we have constructed available to researchers and continuing to refine these methods as yet more powerful algorithmic tools are developed, we hope to enable these kinds of industry-level analyses in order to complement the insightful but relatively limited firm-level analyses that exist. With ongoing efforts to make such firm-level data more readily available, the complementarities between micro-analyses of firms and meso-analyses of industries will enable a richer understanding of relationships between patenting and economic activity in different sectors, countries, or regions – including particularly important insights into the dynamics of these relationships over time and in response to policy and institutional changes.

In this paper, we have used tests against familiar concordances to demonstrate the robustness of the linkages captured in ALP concordances and have refined these algorithmic methods. Our preferred, hybrid weight-based ALP concordances for both IPC-SITC and IPC-ISIC are available to researchers to serve as a platform for empirical patent analysis. With continued advances in text processing and semantic analysis tools and ever richer databases, new possibilities will emerge for building these linkages at yet greater levels of disaggregation. For example, an enhanced ALP approach may soon be able to differentiate between Sector of Use and Industry of Manufacture or to match individual patents to economic classifications. Although effectively leveraging high resolution linkages like this will demand real research creativity, we believe the potential gains associated with a flurry of creative work on this frontier are extraordinary.

Appendix A.

The full results for the concordance can be found at: http://www.wipo.int/export/sites/www/econ_stat/en/economics/zip/wp5_concordance.zip. Since the ALP concordance is constructed as a probability distribution, the weights represent the probability that the origin classification system (Column 1) was matched into the destination classification system (Column 2).

We provide three columns for each type of concordance. Column 1 indicates the origin classification system used, which will be either the International Patent Classification system (IPC, Version 2006), International Standard Industrial Classification (ISIC, Versions 2, 3, 3.1 or 4) and Standard International Trade Classification (SITC, Versions 2, 3 or 4). Column 2 indicates the destination classification system. Column 3 is the ALP-DM (Data Mining) Hybrid weight.

For other weights described in the paper, please contact the authors.

Use of the concordance is straightforward. One simply multiplies the values of the origin classification system by the weights to get the new values as measured by the destination classification scheme.

Appendix B.

Indexing and probabilistic matching approach (ALP-PM)

As an alternative approach, we experimented with what we call a "Probabilistic Matching Approach", which instead of taking key words from the industry descriptions as the ALP-DM does, extracts key words from the patents and then "matches" it to the industry descriptions using a probability matching algorithm. Based on our comparison tests with the YTC and DG concordances, the ALP-DM approach performed better across all measures so that there was little reason to include the method in the paper. However, we still believe that there is potential use of the methodology. Hence, we describe the approach in detail here to illustrate our procedure and compare the results with the ALP-DM approach. With ongoing software improvements, we may revisit this approach in future revisions of the concordance.

Methodology

This approach uses a similar methodology as the ALP-DM approach, but incorporates a separate matching process. In this case, we first extract keywords from the patents and then match them to the industry descriptions using probability weights. While the data mining approach would typically be used to translate industries into technologies, this approach might better be used in the opposite direction and match technologies to industries. This approach may also ultimately enable patent-specific matching to economic classifications, although this would require further refining.

In the initial step of this approach, we order the patents by IPC cluster. We then run each of the patents through a keyword extraction program. For our initial approach, we utilize an open-source Python-based keyword extraction program called "Topia Term Extract 1.10."[17] This extraction program is a generalized text extraction program that identifies the important terms within written content. The benefit of this program is that it also uses language patterns and statistical analysis to determine the strength of each keyword, so that it is possible to rank the keywords by order of importance. There are many other keyword extraction programs in existence, each with their own niche and specialty. While the results from each program will differ slightly, the programs generate very similar results on the whole.

Because of the large quantity of words contained in both the patent abstracts and titles (especially when compared to the quantity of words found in the industry descriptions), it makes sense to weigh the keywords extracted from each patent according to relative importance. In this case, we weigh the keywords from the title

[17] The program package and description can be found at The program package and description can be found at The program package and description can be found at http://pypi.python.org/pypi/topia.termextract/.

Table B.1
ALP-PM and ALP-DM approach example for IPC class A42B (2% cutoff).

IPC number	A42B
IPC description	Headwear – hats; head coverings
Top keywords	"Helmet", "Utility Model", "Cap", "Hat", "Head"
# of patents analyzed	51,864

SITC description	ALP-PM method			ALP-DM method		
	Raw weight	Specificity weight	Hybrid weight	Raw weight	Specificity weight	Hybrid weight
8484-Headgear and fitting thereof	67.53%	25.92%	84.42%	87.67%	58.30%	100.00%
6576-Hat shapes, forms and bodies	17.59%	12.53%	10.63%		13.49%	
8483-Fur clothing (not headgear)		28.51%	4.95%			
6571-Articles of felt	7.47%					
8421-Overcoats and other coats	7.42%					
6582-Tarpaulins, sails, tents		9.78%				
8481-Clothing accessories of leather		6.16%				
2690-Old clothing and textile articles		5.97%				
2692-Rags of twine, wool		5.68%				
8462-Cotton undergarments		5.45%				
8471-Clothing accessories, not knitted					13.42%	
8459-Non-elastic outerwear, knitted					10.86%	
8515-Other footwear with textile				5.00%	3.93%	
7851-Motorcycles				4.06%		
7853-Invalid carriages				3.28%		

to be twice the weight of the extracted keywords from the abstract. This is due to our belief that a single word from the title will provide a better clue as to the real nature of the invention rather than a single word from the abstract. We also limit the number of keywords extracted from each patent to be 10 total words from both the title and abstract. Patent titles and abstracts vary greatly in length, so in order for all patents to receive equal weighting, it is important to limit the matching process to the ten strongest keywords so that certain patents are not more influential.

Another more nuanced step in the keyword extraction process is the use of a "blacklist." Early in our analysis, we found that certain words kept appearing on the keyword extractions that were too general to be used in the matching process, such as "system", "device", "model", "invention" and more. To construct this blacklist of keywords, we ran the keyword extraction program over 500,000 random patents and tabulated the keywords. We looked at the top 100 keywords and ran the PATENTSCOPE cross-lingual expander on certain keywords, which left us with a blacklist of roughly 250–300 keywords. We remove all of the blacklisted words from the extraction results.

Once all of the keywords have been extracted and tabulated for the IPC cluster, we are left with a list of keywords and weights, which were obtained by summing the number of times each keyword appeared in all of the analyzed patents. Each of the keywords and weights are then matched against the industry classification descriptions generated in the ALP-DM approach with additional augmentations. For our initial runs, we used "exact string" matching, although it is possible to do "like" matching and set the tolerance level. For the "exact string" matching portion, we used an expanded word list based on the ALP-DM search terms, full industry descriptions, PATENTSCOPE synonyms and additional plurals, root words and alternative spellings. The reason for this augmentation of the industry terms is that the pool of possible industry matches is much smaller than the pool of patent matches (a couple hundred versus almost 20 million), so we wanted to maximize the quantity of matches and utilize a filtering system and reweighting process to reduce the false positives and thereby improve quality.

For each match, we weighed the importance of the match by the weight of each keyword. The industries that matched with the keywords that have the highest weight after the extraction process were weighed the most. Once the industry and weights have been tabulated, we are left with our raw results.

Next, to reduce the number of spurious matches, we employed a filtering process to the raw results. The first filtering process involved assigning allowable IPC-SITC correspondences. To implement this filter, we assigned lower level IPC's (3-digit) with lower-level SITCs (2-digit). If the correspondence did not make sense, i.e., agricultural production with steel technology, then we disregarded the weights for that specific match. We did this for all 3-digit IPC's and 2-digit SITCs. The next filter involved the 2% cutoff condition, which was similarly employed in the ALP-DM approach. All weights that represented less than 2% of the total weights between IPC and SITC were disregarded and the remaining weights were retabulated and normalized. We then implemented the same "Specificity" and "Hybrid" weighting schemes to these results.

To better illustrate the results, we run the full approach for IPC subclass A42B which is described as being "Headwear/Hats; Head Coverings". Ideally, we would expect SITC product code 8484, which is described as "Headgear and fitting thereof" to be the sole utilizers of this technology. The results can be found in Table B.1 below. Overall, there are 51,864 patents that contain this particular IPC subclass. After running the keyword extraction program through these patents, we find that the five most common keywords are "utility model" (which is excluded from subsequent matching as a blacklisted term), "cap", "hat" and "helmet" and "head". We then used exact string matching to match all the (non-blacklisted) extracted keywords to corresponding SITCs. While the end results of the ALP-PM method matches relatively close with our own preconceptions of the industries that use headwear technology, we see that the ALP-DM method is superior based on the higher weightings in SITC industry 8484.

References

Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. Econometrica 60 (March (2)).

Acharya, R., Keller, W., 2009. Technology transfer through imports. Canadian Journal of Economics 42 (November (4)).

Austin, D., 1993. An event-study approach to measuring innovative output: the case of biotechnology. American Economic Review 83 (2).

Basberg, B., 1987. Patents and the measurement of technological change: a survey of the literature. Research Policy 16 (August (2–4)).

Branstetter, L., Fisman, R., Fritz Foley, C., 2006. Do stronger intellectual property rights increase international technology transfer? Evidence from U.S. firm-level panel data. Quarterly Journal of Economics 121 (February (1)).

Brouwer, E., Kleinknecht, A., 1999. Innovative output, and a firm's propensity to patent: an exploration of CIS micro data. Research Policy 28 (6).
Coe, D., Helpman, E., 1995. International R&D Spillovers. European Economic Review 35 (5).
Comanor, W., Scherer, F.M., 1969. Patent statistics as a measure of technical change. Journal of Political Economy 77 (May–June (3)).
Eaton, J., Kortum, S., 1996. Trade in ideas: patenting and productivity in the OECD. Journal of International Economics 40 (May (3–4)).
Eaton, J., Kortum, S., 2002. Technology, geography and trade. Econometrica 70 (September (5)).
Evenson, R., Putnam, J., 1994. Inter-Sectoral Technology Flows: Estimates from a Patent Concordance with an Application to Italy. Yale University Mimeo.
Falvey, R., Foster, N., 2006. The Role of Intellectual Property Rights in Technology Transfer and Economic Growth: Theory and Evidence. UNIDO Working Paper.
Furman, J., Porter, M., Stern, S., 2002. The determinants of national innovative capacity. Research Policy 31 (6).
Griliches, Z., 1990. Patent statistics as economic indicators: a survey. Journal of Economic Literature 28 (December).
Grossman, G., Helpman, E., 1991. Innovation and Growth in the Global Economy. MIT Press, Cambridge, MA.
Hall, B.H., Jaffe, A., Trajtenberg, M., 2005. Market value and patent citations. RAND Journal of Economics 36.
Harhoff, D., Hoisl, K., Reichl, B., Van Pottelsberghe De La Potterie, B., 2009. Patent validation at the country level: the role of fees and translation costs. Research Policy 38 (9).
Jaffe, A., Trajtenberg, M., 1999. International knowledge flows: evidence from patent citations. Economics of Innovation and New Technology 8.
Jaffe, A., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Quarterly Journal of Economics 108 (August (3)).
Jaffe, A.B., Fogarty, M.S., Banks, B.A., 1998. Evidence from patents and patent citations on the impact of NASA and other federal labs on commercial innovation. Journal of Industrial Economics 46.
Johnson, D., March 2002. The OECD Technology Concordance (OTC): Patents by Industry of Manufacture and Sector of Use, OECD Science, Technology and Industry Working Papers.
Keller, W., 2004. International technology diffusion. Journal of Economic Literature 42.
Kortum, S., Putnam, J., 1997. Assigning patents to industries: tests of the Yale technology concordance. Economic Systems Research 9 (2).
Leydesdorff, L., 2008. Patent classifications as indicators of intellectual organization. Journal of the American Society for Information Science and Technology 59 (10).
McCalman, P., 2001. Reaping what you sow: an empirical analysis of international patent harmonization. Journal of International Economics 55 (1).
Moehrle, M., Walter, L., Bergmann, I., Bobe, S., Skrzipale, S., 2010. Patinformatics as a business process: a guideline through patent research tasks and tools. World Patent Information 32.
Moehrle, M., Walter, L., Geritz, A., Muller, S., 2005. Patent-based inventor profiles as a basis for human resource decisions in research and development. R&D Management 35 (5).
Porter, M.E., Stern, S., 2000. Measuring the Ideas Production Function: Evidence From International Patent Output. NBER Working Paper 7891.
Romer, P., 1990. Endogenous technological change. Journal of Political Economy 98 ((October) 5).
Schmoch, U., LaVille, F., Patel, P., Frietsch, R., 2003. Linking technology areas to industrial sectors: final reports to the European commission. DG Research November.
Trajtenberg, M., Henderson, R., Jaffe, A., 1992. Ivory tower versus corporate lab: an empirical study of basic research and appropriability. National Bureau of Economic Research, No. 4146.
Trippe, A., 2003. Patinformatics: tasks to tools. World Patent Information 25 (3).
Verspagen, B., van Moergastel, T., Slabbers, M., 1994. MERIT concordance tables: IPC-ISIC (Rev. 2). MERIT Research Memorandum February.
Xu, B., Chiang, E., 2005. Trade, patents and international technology diffusion. Journal of International Trade and Economic Development 14 (1).
Yang, Y.Y., Akers, L., Klose, T., Yang, C., 2008. Text mining and visualization tools: impressions of emerging capabilities. World Patent Information 30 (4).