An 'Algorithmic Links with Probabilities' Crosswalk for USPC and CPC Patent Classifications with an Application Towards Industrial Technology Composition

Nathan Goldschlag
Center for Economic Studies,
U.S. Census Bureau

Travis J. Lybbert
University of California, Davis,
Department of Agricultural and
Resource Economics

Nikolas J. Zolas
Center for Economic Studies,
U.S. Census Bureau

February 28, 2016[1]

Patents are a useful proxy for innovation, technological change, and diffusion. However, fully exploiting patent data for economic analyses requires patents be tied to measures of economic activity, which has proven to be difficult. Recently, Lybbert and Zolas (2014) have constructed an International Patent Classification (IPC) to industry classification crosswalk using an 'Algorithmic Links with Probabilities' approach. In this paper, we utilize a similar approach and apply it to new patent classification schemes, the U.S. Patent Classification (USPC) system and Cooperative Patent Classification (CPC) system. The resulting USPC-Industry and CPC-Industry concordances link both U.S. and global patents to multiple vintages of the North American Industrial Classification System (NAICS), International Standard Industrial Classification (ISIC), Harmonized System (HS) and Standard International Trade Classification (SITC). We then use the crosswalk to highlight changes to industrial technology composition over time. We find suggestive evidence of strong persistence in the association between technologies and industries over time.

[1] All opinions and views expressed are those of the authors and do not represent those of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. We thank Javier Miranda, Shawn Klimek, Asrat Tesfayesus, Lars Vilhuber, participants at the CES brown bag seminar series, and participants at the 2015 FCSM conference for helpful comments.

INTRODUCTION

Innovation and the diffusion of technological change are key drivers of economic growth (Romer 1990; Aghion and Howitt 1992). Measuring innovation and technology transfer has proven to be difficult. Patent data has been used in a number of studies as a proxy of technological change (Griliches 1998). One advantage of patent data is its richness—patent data contains information on the inventor(s), the associated firm, the ideas themselves, and antecedent ideas in the form of prior art. Fully leveraging these data, however, requires the ability to disaggregate and combine patent statistics with other measures of economic activity. These other measures often use classification systems other than the United States Patent Classification (USPC), which is the native technological classification system in U.S. patent data. In more recent years, the focus of classification efforts has shifted to the Cooperative Patent Classification (CPC) scheme; a cooperative effort between the United States Patent and Trademark Office (USPTO) and the European Patent Office (EPO) to develop a common, internationally compatible technology classification system. It is important to be able to translate between both USPC and CPC to other industry and product classification systems in order to assign economic values and measures to patent data. This type of translation is key to conducting analysis of policies aimed at affecting innovation and economic development.

In this research, we develop such a linkage using the Algorithmic Links with Probabilities (ALP) approach first described by Lybbert and Zolas (2014). We provide concordances that translate USPC and CPC codes into multiple vintages of the International Standard Industrial Classification (ISIC), the North American Industrial Classification System (NAICS), the Standard International Trade Classification (SITC), and the Harmonized System (HS) product codes. The ALP methodology is an automated and generalizable approach that utilizes a variety of text mining techniques and readily facilitates revision as classification systems are updated and new patents are issued. The resulting ALP concordances provide direct probabilistic linkages between classification systems by leveraging the textual content of documents themselves. These probabilities can then be used as weights in joint analyses of patent and economic data, which directly supports policy relevant research questions related to innovation and technological diffusion.

After introducing the methodology and demonstrating the validity of the concordances using external data sources, we use these concordance to investigate how the relationship between technologies and industries has changed over time. We find suggestive evidence that while the relationship changes on a yearly basis, there is strong persistence in the cumulative technologies associated with each industry.

BACKGROUND

Patents are a powerful source of information on innovative activity partly because of the detailed information they contain. Patent documents include information on the inventor(s), such as name and location, the name and location of the assigned firm (if applicable), detailed descriptions of the innovation, related innovations in the form of prior art, and the technological classification of the innovation. Moreover, experienced patent examiners curate these data elements, ensuring their accuracy and quality.

The U.S. Patent Classification System (USPC), first developed in 1900, is used by the USPTO to organize all U.S. patent documents into collections of common subject matter. The USPC is organized in

a hierarchical structure with more than 450 classes and more than 150,000 sub classes. The IPC classification scheme, in contrast, was established in 1971 Strasbourg Agreement and contains over 71,000 subgroup classifications (Harris et al. 2010). The CPC classification system is the result of a partnership between the USPTO and EPO to harmonize existing classification schemes. The agreement, which was announced in 2010, has been utilized to classify patents granted since 2013. The CPC is similar in structure to the IPC classification system with some minor modifications. Going forward, the CPC will be the main identification system for international patents with concordances used to apply them to older patents. This paper provides the first known crosswalk that concords the CPC to a variety of industry classifications and vice versa.

A variety of efforts have been made to translate the USPC into other classification schemes. One of the first attempts to link patent and industry data was Schmookler (1966), which assigned "industries-of-use" to patents organized using the USPC. In addition, the USPTO issued concordances between USPC and IPC, SIC, and NAICS.[2] Though the USPC to IPC concordance is relatively comprehensive, the SIC and NAICS concordances map USPC codes to approximate groupings of industries and focuses exclusively on manufacturing industries. One of the first comprehensive patent to industry classification concordance is the Yale Technology Concordance (YTC) developed in the early 1990s (Evenson & Putnam 1994). The YTC links IPC codes to the Canadian SIC (cSIC) system using a set of Canadian patents granted between 1978 and 1993 that were explicitly assigned a technology field using the IPC as well as an Industry of Manufacture and a Sector of Use according to the Canadian SIC. This set of patents implicitly provides a direct concordance between IPC and cSIC. The YTC has several advantages not least of which being that it relies upon the purposeful consideration of expert patent examiners. On the other hand, one of the primary limitations of the YTC is that it is frozen in time and unable to adapt to a changing technology landscape and continually evolving classification systems. Moreover, the YTC provides a direct linkage only between IPC and cSIC, which necessitates the layering of multiple concordances to integrate data not classified in cSIC. In the case of US patents the YTC also requires layering USPC to IPC concordances.

In addition to the YTC, there are several other concordances between IPC and industry classification schemes including the "DG Concordance" (Schmoch et al. 2003) and the MERIT Concordance (Verspagen et al. 1994), which rely on direct one-to-one manually generated matches. In all three cases, however, to arrive at a consistent USPC to industry classification would require the combination of multiple concordances. This layering introduces additional error and ambiguity into the translation. Instead, this research provides a direct probabilistic many-to-many linkage between USPC and CPC to several vintages of classification systems including ISIC, NAICS, SITC, and HS. In addition, the methodology employed to create these linkages is generalizable, repeatable, can be aggregated/disaggregated and flexible. It can easily accommodate additional patent documents, updated classification systems, and has the added benefit of working in both directions (i.e. from patent-to-industry and industry-to-patent). The resulting suite of concordances provides researchers with a number of different tools to assess important policy questions related to patenting and innovation.

[2] See Hirabayashi (2003) for details.

METHODOLOGY

The methodology we use to construct the linkages between USPC and CPC codes and industry and trade classifications follows the ALP approached first described by Lybbert and Zolas (2014). The ALP methodology, which relies on keyword extraction and text mining, has several important advantages over existing approaches. First, the ALP method builds up from the textual content of individual documents to develop aggregate concordances. Second, the ALP method yields direct probabilistic linkages, eliminating the need to layer concordances and accommodating the many-to-many linkages that often appear between industry and product concordances. Finally, the ALP approach relies on a generalizable automated process, allowing for the rapid processing of millions of documents and minimizing the need for manual intervention. This process is both flexible and repeatable, allowing each concordance to be re-executed to accommodate changes in the technological landscape or updated classification schemes.

The programs that perform these tasks yield linkages that approximate manual assignment of industry and trade classifications by searching through each patent's abstract for key words associated with industry and trade codes. As with any algorithmic search technique, these methods cannot perfectly replicate careful manual classification. By processing millions of documents, however, this approach relies on the Law of Large Numbers, improving with the size of the patent corpus. It is important to note that the nature of technologies changes over time and the set of patents used includes US patents granted between 1976 and 2014 and international patents from the PATSTAT database granted in the same time period. By pooling patents across years, our resulting matches reflect the relationship between technologies and industries on average over the entire period. For example, if the technologies in a given USPC/CPC code are associated with one industry in the 1980s and a different industry in the 1990s, our method would capture both relationships and treat them equally..

The ALP approach relies on the text mining of patent abstracts and keywords extracted from industry and product classification descriptions. Whereas for patents we have access to the text of millions of abstracts, unfortunately there is no comparably rich set of qualitative information for industry classifications. Therefore, for industry and product classification schemes we exploit the only available source of qualitative information: the brief descriptions used to characterize each industry or product category. The interpretation of the final technology-industry crosswalk is critically dependent on the way in which the industry and product classification schemes are constructed. For example, NAICS is used to classify the primary activity performed by business establishments, where activity is understood to be the processes involved in transforming resources such as equipment, labor, manufacturing techniques or intermediate products into goods and services[3]. Therefore, our translation of USPC to NAICS will capture industries that use or implement technologies rather than industries that perform research and development activities.

We extract search terms associated with 4- and 5-digit SITC, 4-digit ISIC and 6-digit HS industry descriptions provided by the United Nations, along with 6-digit NAICS descriptions provided by the BEA, BLS and Census Bureau. These descriptions often include a single or multiple sentences that lists the products and/or services that are included in the category. A combination of algorithmic and manual approaches are used to curate a set of keywords that retrieve patents relevant for the corresponding

[3] See Census Economic Classification Policy Committee – Issues Paper No. 1

category. The algorithmic methods include the keyword extraction algorithm, Topia Term Extract[4], which determines the keywords using a simple Parts-Of-Speech (POS) tagging algorithm. These keywords are also modified to be robust to typical syntactic concerns including plurals and word phrases. We expand the keyword set to include synonyms found in the WIPO's PATENTSCOPE, which generates synonyms based on the full text of patents in different languages. Finally, we manually inspect the final set of keywords and incorporate "not" terms that exclude erroneous matches.

The final curated set of keywords are used to query the patent abstracts of over 5 million patents granted in US between 1976 and 2014 found in the PatentsView database[5] for the USPC crosswalk and over 40 million patents applied for worldwide PATSTAT database for the CPC crosswalk[6]. These data provide both USPC and CPC codes associated with each patent granted between 1976 and 2014[7]. We select for each classification all patents that contain at least one of the keywords and zero of the "not" terms. Patents that contain multiple keywords across multiple industries are counted multiple times. This process yields many-to-many matches from classification to patents. We then tabulate the number of patents for each USPC/CPC to industry/product classification combination. We filter out obviously incorrect matches, e.g. Pharmaceuticals to Concrete Manufacturing, and exclude matches to service industries. We then reweight the results using a Bayesian weighting scheme described in Lybbert and Zolas (2014). The purpose of the reweighting of the frequencies is to minimize both Type I and Type II errors. This weighting scheme takes into account the number of possible technologies and how frequently each technology class is matched to a given industry/product category. Specifically, we rely on the hybrid weighting scheme that combines the raw and specificity weights to balance Type I and Type II errors (Lybbert and Zolas 2014). The formula for this weighting scheme is:

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

Where A_j is the outcome of being matched to technology j and B_i is the outcome of being matched to industry i. W_{ij}^{R} is the raw Bayesian weights given by

$$W_{ij}^{R} = \Pr(A_j|B_i) = \frac{\Pr(B_i|A_j)\Pr(A_j)}{\Pr(B_i|A_1)\Pr(A_1) + \cdots + \Pr(B_i|A_J)\Pr(A_J)}$$

In the hybrid approach, we substitute the $\Pr(A_j)$ found in the raw Bayesian approach with $\Pr(A_j)=W_{ij}^{R}/J$, which has the effect of discounting widely matched technologies (i.e. patents/technologies that are matched across a wide variety of industries) and increasing the weights of more specific technology-industry/product matches (i.e. frequent matches within relatively few technologies/patents). This more

[4] A full description of the program can be found here: https://pypi.python.org/pypi/topia.termextract/ (accessed 2/2/2016).
[5] See http://www.pat7entsview.org (accessed 2/2/2016) for more details.
[6] In order to maintain consistency, we limit the patents used in the CPC crosswalk to those applied for between 1976 and 2014, along with the patent actually being granted.
[7] For PATSTAT patents, we utilize the first application date as the date for the patent and only included granted patents. As a result, patents in the later years of the PATSTAT (2011 and later) will be limited due to the average time between application and granting (typically 3-5 years).

"balanced approach" does not completely discount widely applicable technologies, but instead tries to focus on the unique identifying technology of each industry.[8]

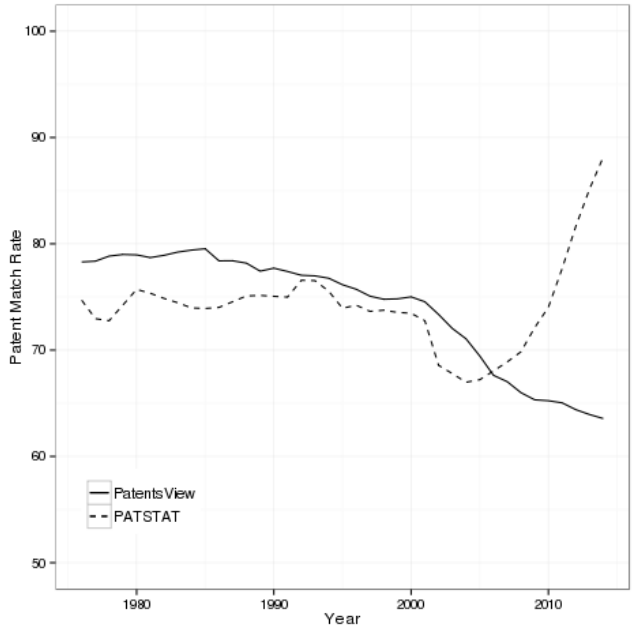
After reweighting the frequencies, we introduce a cutoff condition (2%) for the weights in order to further reduce Type I errors. The cutoff condition sets all weights below a certain threshold to zero before renormalizing. Imposing a threshold helps reduce noise associated with rare or idiosyncratic matches and focuses the concordance on common patterns. After the cutoff condition is implemented, several industries and technologies may drop due to their maximum weight being below the cutoff threshold. In this instance, we keep the maximum weight for the dropped industry/technology in order to ensure full coverage across all industry and technology types. Finally, we can then renormalize the remaining positive matches such that they sum to one.

The final result is patent-to-industry and industry-to-patent crosswalk that can be aggregated and disaggregated from the 1-digit to 6-digit level, and can be continuously updated as technology continues to evolve.

RESULTS

We run the full ALP methodology across both USPC and CPC codes using more than 5 million US Patents and 40 million global patents found in the PATSTAT database and make the full version of the crosswalk available for researchers to use for free. In order to assess the validity of the crosswalk, we compare the patent counts generated by the USPC concordance with existing USPTO publications and with results from an IPC to industry concordance. We also discuss the resulting CPC concordances and compare it to an existing EUROSTAT concordance that links IPC to NACE Rev. 2.

Figure 1: Utility Patent Match Rates by Year



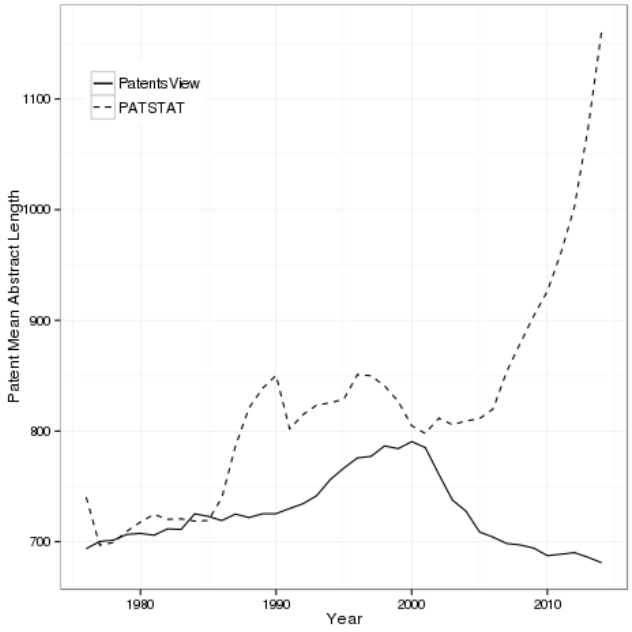
Source: Matched PatentsView and PATSTAT patent abstracts to industry and product search terms.

[8] For further discussion, see Lybbert and Zolas (2014)

Notes: Includes patent counts by grant year for PatentsView patents and patent counts for granted patents by application year for PATSTAT patents. The recent uptick in the PATSTAT match rate is due to the low number of patents granted for patents applied for in the most recent years (2011 and later, see footnote 7). Includes utility patents from PatentsView database and granted patents with English language abstracts from PATSTAT database. Vertical axis is not zero adjusted, ranging from 50 to 100.

With over 155 thousand search terms and more than 45 million patent abstracts to search, we perform over 7 trillion comparisons. The overall PatentsView (USPC) match rate for US utility patents to search terms is 71%. For PATSTAT (CPC) we match 69% of all patents and 75% of granted patents with English language abstracts[9]. As shown in Figure 1, both match rates vary by year, with the PatentsView exhibiting a notable decline between 1976 and 2014 and PATSTAT showing a significant increase after 2008. One possible explanation for match rates changing by year is that the structure or amount of underlying abstract text might be changing over time. Fundamentally, our methodology is sensitive to the amount of text available within the patent abstracts. To investigate this possibility, Figure 2 shows average annual abstract length of patents both from the PatentsView and PATSTAT databases. The trends we observe in average abstract length are suggestive that the increasing textual content found in the PATSTAT data may partially account for the dramatic rise in match rates. Average abstract length in the PatentsView data may only partially account for the slow and steady decline in match rates. The PatentsView data exhibits a mild rise in average abstract length through 2000, before reversing in 2001, which coincides with the timing of the fall in match rates.

Figure 2: Mean Annual Patent Abstract Length by Year



Source: PatentsView and PATSTAT patent abstracts.
Notes: Abstract length in characters. Includes utility patents from PatentsView database and granted patents with English language abstracts from PATSTAT database. The PATSTAT data after 2011 contains fewer patents since our sample consists of granted patents and the year is based on the first application date (see footnote 7).

Investigating the unmatched patents we find that the five most frequent USPC codes among unmatched PatentsView patents (which account for almost 13% of the unmatched cases) are "514-Drug, Bio-

[9] There are over 19.4 million granted patents in the PATSTAT database with English language abstracts.

Affecting and Body Treating Compositions", "435-Chemistry Molecular Biology and Microbiology", "370-Multiplex communications", "257-Active Solid-State Devices (e.g. Transistors, Solid-State Diodes)", and "455-Telecommunications". These technologies cover innovations for which the lexicon is both very specific and rapidly changing. The set of search terms utilized in the concordance construction, on the other hand, is relatively static, which likely accounts for the declining match rates.

It could be the case that the remaining unmatched patents are incorrectly categorized because their abstracts contain terms that are associated with multiple industries but also contain not terms for those industries. For example, suppose industry A's in-terms include silicon and not-terms includes aluminum. Similarly, industry B's in-terms include aluminum with not-term silicon. Then a patent that references both silicon and aluminum would effectively exist between the two industry categories, but is matched to neither. To asses the extent of this phenomenon we rematch the residual unmatched patents to the search terms without considering not terms. Of the 1.4 million US patents that did not match to any search terms only slightly more than 22 thousand matched to the search terms in the absence of exclusionary terms. This suggests that it is not the case that the unmatched patents exist between the industry definitions. Next we address robustness of the resulting crosswalks by comparing to external data sources.

USPTO TO ALP COMPARISON

As part of a report describing industry patenting trends, the USPTO developed a concordance between the USPC and a set of 30 product fields. These reports provide a potential external source of validation for our USPC to industry concordances. The product fields, or NAICS-based categories, were loosely based on 2002 NAICS industries[10]. The categories include three and four digit manufacturing industries. These NAICS-based categories were manually generated based upon the patent's primary or 'original' classification[11]. Each USPC is assigned between one and seven NAICS-based categories, each given equal probability weighting. The concordance is based on the type of establishment producing a product or implementing a process. Finally, it is important to note that these NAICS-based tabulations rely heavily on the earlier USPC to SIC concordances originally developed in 1974[12].

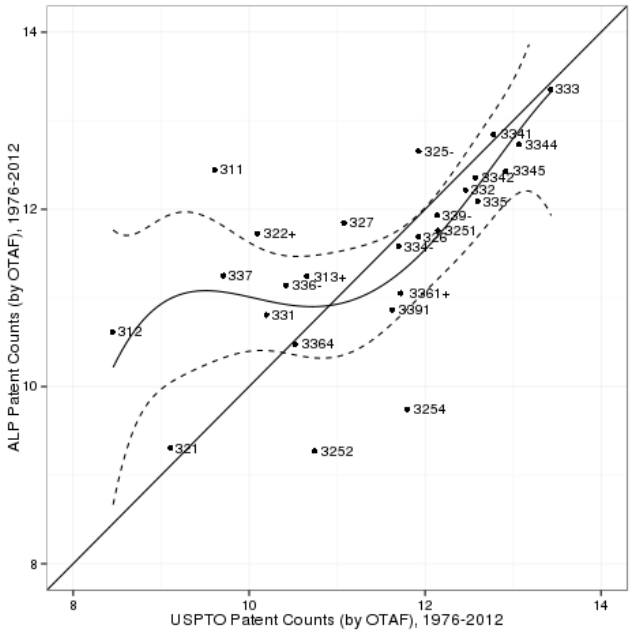
To compare the ALP and USPTO concordances we must first translate the ALP concordance into the NAICS-based categories, or what the USPTO labels OTAF codes, that aggregate 3 and 4 digit 2002 NAICS industries. We then apply the ALP concordance to the primary classifications of utility patents between 1976 and 2012, the range of years covered by the USPTO report. In agreement with the USPTO methodology, we equally weight all associated industry codes. This process yields weighted patent counts by 4-digit 2002 NAICS industries, which can be translated to the OTAF aggregate industry groupings found in the USPTO reports.

[10] http://www.uspto.gov/web/offices/ac/ido/oeip/taf/naics/doc/naics_info.htm website on the existing USPC-NAICS crosswalk

[11] http://www.uspto.gov/web/offices/ac/ido/oeip/taf/naics/stc_naics_faall/all_stc_naics_fa.htm

[12] http://www.uspto.gov/web/offices/ac/ido/oeip/taf/data/misc/patenting_trends/otaf_concordance_review_Jan1985.pdf

Figure 3: Comparing Patent Counts using the ALP with USPTO Counts by OTAF, 1976-2012



Source: USPTO 2012 US Patenting Trends by NAICS report, PatentsView patent database, ALP USPC to 4-digit 2002 NAICS concordances, author's calculations.

Notes: USPTO utility patent counts by NAICS based categories (OTAF codes) 1976-2012 and counts of utility patents by primary USPC code via PatentsView database concorded to 4-digit 2002 NAICS industries. Both axes show logged patent counts, 45-degree line shown with 4th degree fitted polynomial and 95% confidence bands.

As shown in Figure 3, our results compare quite favorably with the existing USPC concordance given the significantly different methodology employed. The largest discrepancies occur in OTAF industries 311 ("Food") and 312 ("Beverages") where the ALP estimates are greater than the OTAF estimates and OTAF industries 3252 ("Resin, Synthetic Rubber, and Artificial and Synthetic Fibers and Filaments") and 3254 ("Pharmaceutical and Medicines"), which the ALP concordance generates fewer patent counts. The explanation for these differences is likely due to the words used in the search terms. Industries with high product varieties such as food and beverages have on average a much more diverse and larger set of search terms, resulting in more hits and therefore higher weights towards these industries. On the other hand, industries composed of raw or synthetic materials (such as resin and rubber), or industries with extreme specialization in terms (such as pharmaceuticals), will likely have fewer search terms associated with them, thereby resulting in slightly lower weights. However, these differences appear to be somewhat minor and it is not entirely conclusive that the USPTO patent counts represent the "true" value of industry patents.

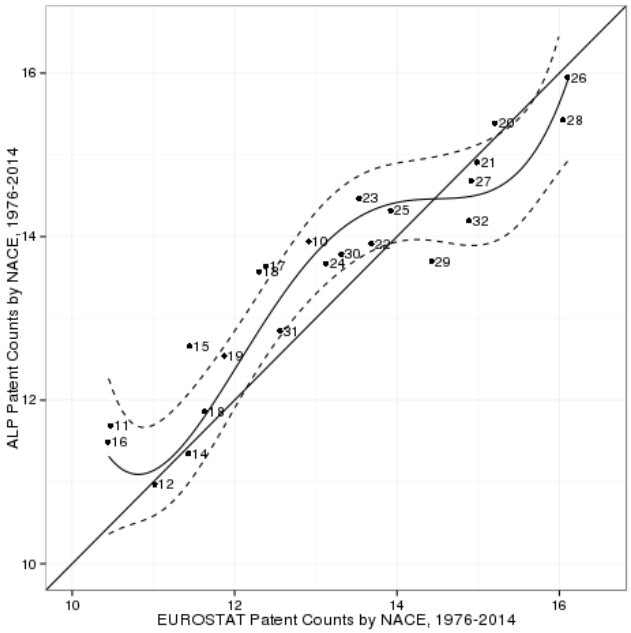
The benefit of our approach is that the ALP concordances are not limited to the aggregated 3 and 4 digit NAICS categories used in the USPTO's concordance, but can be drilled down deeper to 6-digit NAICS. Moreover, our ALP crosswalks also cover industries outside of manufacturing, including agriculture and forestry, mining, utilities, and construction. The ALP concordances are also not "stuck in time", allowing analysis between the 1997, 2002 and 2007 NAICS versions. Finally, concordances created using the ALP methodology rely on a generalized approach to directly translate USPC to NAICS, leveraging the textual content of both classification schemes and patents themselves, rather than manual interpretation based on previous SIC-based vintage crosswalks.

EUROSTAT TO ALP COMPARISON

Beginning in 2015, the PATSTAT database included industry-specific NACE codes assigned to each application that contained an IPC. The industry concordance is based on a revised EUROSTAT IPC to NACE crosswalk developed by Van Looy, Vereyen and Schmoch (2014). The crosswalk is an update to the industry-technology "DG" concordance originally developed in Schmoch et al. (2003) which uses the main industry of firm-owned patents to generate a mostly 1-to-1 match been an IPC and NACE. The new version of the crosswalk incorporates additional technology categories (IPC) that have been introduced since 2003, and converts the IPCs to NACE Rev. 2 (the original converted IPCs to NACE Rev. 1.1). There are also a handful of 4-digit IPCs that receive a proportional distribution into different NACE categories (for instance, B65D "Containers" is allocated into NACE 22.22, 23.13, 17.21, 25.91, 13.9 and 16.24). A detailed comparison and discussion of the methodologies and outcomes can be found in Lybbert and Zolas (2014).

We compare the patent counts by NACE generated by the EUROSTAT concordance and the ALP concordance as an additional validation exercise. To do this, we first extract the patents that can be concorded using both methodologies from PATSTAT. We assign equal weights to all 4 digit CPCs within a patent application and concord them to the 2 digit NACE using the EUROSTAT concordance supplied by PATSTAT. For the ALP counts, we similarly assign equal weight to all 4-digit CPCs and convert them first to 4 digit ISIC Rev. 4, which the NACE Rev. 2 is based off. We then convert them to 2 digit NACE Rev. 2 using the correspondence table found in UN Statistics Division website. Once the patents have been concorded, we collapse them and sum the overall weights by NACE Rev. 2. Figure 4 presents a comparison of patent counts by NACE using the EUROSTAT and ALP concordances. The patent counts generated using the ALP concordances compare quite favorably to those created using the EUROSTAT data with fewer outliers than the comparison to USPTO reports.

Figure 4: Comparing Patent Counts by NACE using ALP and EUROSTAT, 1976-2012



Source: PatentsView patent database, ALP USPC to 4-digit 2002 NAICS concordances, author's calculations.
Notes: ALP patent counts by NACE translated from patent counts by CPC concorded to ISIC Rev. 4, then to NACE. EUROSTAT patent counts tabulated directly from the PATSTAT database. Both axes show logged patent counts, 45-degree line shown with 4th degree fitted polynomial and 95% confidence bands.

In addition to the favorable comparison, the added benefit of the ALP approach includes more detailed industry breakdown (the EUROSTAT crosswalk converts 4-digit IPCs into 2-4 digit NACE), backward and forward linkages (i.e. technology-to-industry and industry-to-technology), probabilistic distribution, a larger number of industry classifications and more frequent updates.

COMPARING USPC AND IPC CONCORDANCES

As a final verification exercise we ensure that a similar group of patents are consistently concorded to the same industry groups when using either the USPC, IPC, or CPC concordances. In this exercise, we compare counts of utility patents granted between 1976 and 2014 created using the USPC concordances to counts created using the IPC concordance developed in Lybbert and Zolas (2014). The utility patents found in PatentsView contain both a USPC and IPC codes. Hypothetically, we should find perfect overlap when we sum the patent counts by industry using a USPC-based concordance versus using an IPC-based concordance since the exact same patents are being considered. However, there may be differences due to the weighting of the USPC or IPC codes prior to the conversion (for instance, it may be the case that the same patent has only 1 USPC code and multiple IPC codes, since the USPC codes are more disaggregate).

To do this exercise, we start by assigning equal weights to each of the technology classifications by application. We then sum the weights by USPC class and by 4-digit IPC before converting them to the respective industry group using both the USPC ALP Concordance and IPC ALP Concordance. The end result contains a summation of the total weight by industry. Table 1 compares the correlation and gives the R-squared in the industry-weights by USPC Concordance and IPC Concordance.

Table 1: Comparison of Patent counts by Industry for USPC Concordance and IPC Concordance

Industry	Aggregation	Correlation	R^2
ISIC Rev. 2	4-digit	0.9851	0.9704
ISIC Rev. 3	4-digit	0.9843	0.9686
ISIC Rev. 3.1	4-digit	0.9843	0.9688
ISIC Rev. 4	4-digit	0.9743	0.9493
SITC Rev. 2	4-digit	0.9454	0.8939
SITC Rev. 3	4-digit	0.8137	0.6621
SITC Rev. 4	4-digit	0.8394	0.7045
NAICS 1997	6-digit	0.9728	0.9464
NAICS 2002	6-digit	0.9699	0.9407
NAICS 2007	6-digit	0.9700	0.9409
HS2002	6-digit	0.9239	0.8537
HS2007	6-digit	0.9092	0.8266

Source: Suite of ALP USPC and IPC concordances, author's calculations

Table 1 shows the similarity of patent counts generated using the USPC and IPC concordances. On average, the correlation across all industries and levels of aggregation are 0.9405, with an average R^2 of 0.8558, meaning that less than 15% of variation in industry weights between the USPC concordance and IPC concordance remains unexplained. The similarity between the results is unsurprising since both rely on the ALP methodology and use a similar corpus of industry and product keywords. Some deviation, however, is to be expected if only due to the subtle differences inherent between the IPC and USPC classification schemes.

TEMPORAL COMPOSITION OF TECHNOLOGIES

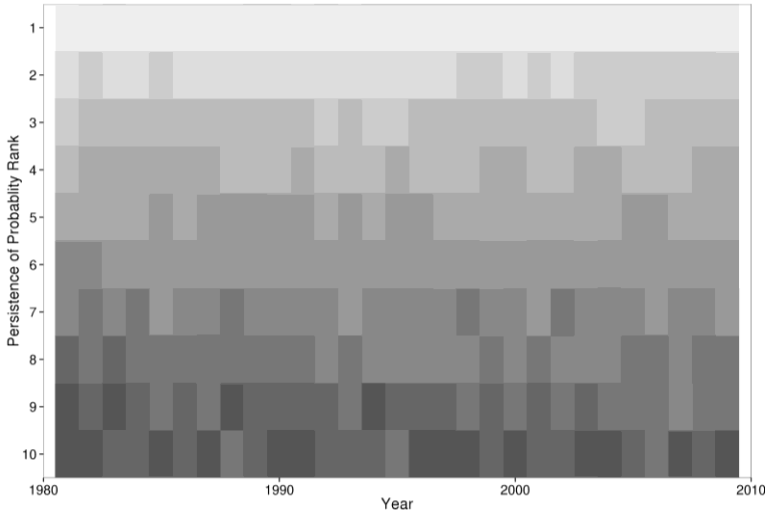
One major benefit of the ALP methodology is the flexibility to update the concordance with each new data release. As the underlying set of patents evolves over time, the concordance is able to "correct" itself by incorporating new information to calculate updated weights. This allows us to provide an up-to-date crosswalk that reflects how newer technologies are associated with industries and products. This flexibility also allows us to go backwards in time and analyze how the relationship between industries and technologies has changed over time. In this section, we analyze the composition of technologies associated with industries in 1980 and assess how the composition and rankings of these technologies evolves over time. We find preliminary evidence of changes in the technological composition of industries on a yearly basis but that there exists strong persistence in the cumulative technology-industry associations between 1980 and 2010.

Cross-Sectional Technology Variation

To examine how the technological composition of industries change over time we first generate cross sectional snap shots of the concordances each year between 1980 and 2010. For each year, we calculate the ALP weights linking industries to 4-digit CPC codes, pooling the 6-digit 2002 NAICS and 4-digit

ISIC Rev 3 industries.[13] We then compare how the ranking of technologies for each industry changes year to year. Figure 5 shows a heat map where the shade of each year-rank cell is scaled by the average 1980 rank within each year-rank cell. The scale of shades is relative to the ranking in 1980. The figure suggests significant reordering of technologies from year to year, but mainly across the lower ranked technologies. Lighter shaded rank 1 and 2 cells become slightly darker over time as lower ranked technologies make their way to rank 1. Similarly, rank 6 to rank 10 cells become lighter and darker as technologies move up and down in the rankings.

Figure 5: Heat Map of Average 1980 Rank by Year to 2010



Source: Annually calculated ALP concordances.
Notes: ALP concordances include 6-digit 2002 NAICS, 4-digit ISIC Rev 3, 6-digit 2002 HS, and 4-digit SITC Rev 3. Year-rank cells shaded by the mean 1980 rank of technologies. Shades normalized to the 1980 rank.

The transition matrix in Table 2 shows the probability that each 4-digit CPC technology in a given rank in 1980 will be in each corresponding rank in 2010 on a cross-sectional basis (i.e. comparing 1980 patents and 2010 patents). Note that the top ranked technologies in 1980 have only a 39% probability of remaining the top technology in 2010. These probabilities tend towards the single digits beyond the second and third ranked technologies. Overall, these transition probabilities suggest a significant amount of churn in the relationship between technologies and industries in our concordances, with 34% of the top technologies dropping completely out of the rankings by 2010. However, despite the churn, we can see that more than 60% of the top technology in 1980 is still within the Top 5 technologies by 2010. The Spearman's rank-order correlation coefficients of the probability weight rank between subsequent years from 1980 to 2010 are all significant with a mean if 0.75, again suggesting a positive association between a technology's rank from one year to the next.

[13] We find similar results in the relationship between products and technologies using 6-digit 2002 HS and 4-digit SITC Rev 3 product codes, see Appendix A for details.

Table 2: Probability Matrix of Rank Transitions between 1980 and 2010, Yearly Weights

		2010 Ranking										
		1	2	3	4	5	6	7	8	9	10+	Not Ranked
1980 Ranking	1	39.12	12.43	5.67	3.84	1.83	0.73	0.73	0.91	0.18	0.55	34.00
	2	13.56	21.26	8.50	4.45	2.63	1.62	1.01	0.81	0.40	0.81	44.94
	3	6.55	13.35	11.65	6.07	2.91	2.43	2.18	0.97	0.73	1.21	51.94
	4	4.31	7.18	8.33	7.76	5.75	2.30	0.86	1.44	1.15	1.15	59.77
	5	4.53	4.18	5.23	5.57	6.27	3.48	3.83	1.39	2.44	1.05	62.02
	6	2.10	5.46	5.04	3.36	5.04	4.20	2.10	0.84	1.26	4.20	66.39
	7	3.26	2.72	1.63	7.07	1.63	1.63	2.72	2.17	1.63	1.63	73.91
	8	2.13	2.84	4.96	3.55	2.13	7.09	2.84	1.42	2.84	0.71	69.50
	9	3.42	1.71	2.56	0.85	1.71	1.71	2.56	6.84	3.42	5.13	70.09
	10+	1.61	0.65	1.94	0.65	1.94	2.26	0.65	0.65	0.97	7.10	81.61
	Not Ranked	11.86	12.50	13.84	12.79	10.75	8.53	6.95	5.55	4.03	13.20	-

Source: Annually calculated ALP concordances.
Notes: Includes granted patents with an application date between 1980 and 2010. ALP concordances include 6-digit 2002 NAICS and 4-digit ISIC Rev 3. Results are similar for 6-digit HS and 4-digit SITC product codes.

Focusing on changes in the discrete ranking of technologies within industries may mask significant variation in the continuously measured probability weights. One way to see how the underlying weights between industries and technologies change over time is to measure the extent to which lagged weights predict current weights. Table 3 shows the estimation results of autoregressive models using different lags. The first lagged value, the weight in t-1, explains 40% of the variance in the weight at time t. The predictive power of lagged weights falls as we increase the length of the lag, dropping to 26% after 30 years. One interesting facet about the table is how dramatically the predictive power of the previous period's technology drops in the first period, before quickly leveling off. In the first period, we lose more than 60% of the predictive power, but by year 30, the additional loss is less than 15%.

Table 3: Autoregressive Model of Annually Calculated Weights, Varying Lag Length

	Yearly Weight at t, δ=1	Yearly Weight at t, δ=5	Yearly Weight at t, δ=10	Yearly Weight at t, δ=15	Yearly Weight at t, δ=30
Yearly Probability Weight, t-δ	0.630***	0.597***	0.584***	0.573***	0.529***
	(0.0146)	(0.0156)	(0.0159)	(0.0164)	(0.0223)
Observations	4,354,860	3,774,212	3,048,402	2,322,592	145,162
R-squared	0.396	0.352	0.330	0.313	0.261

Source: Annually calculated ALP concordances.
Notes: Standard errors are clustered at the four or six digit industry code in parentheses * p<0.05, ** p<0.01, *** p<0.001. Includes granted patents with an application date between 1976 and 2014.

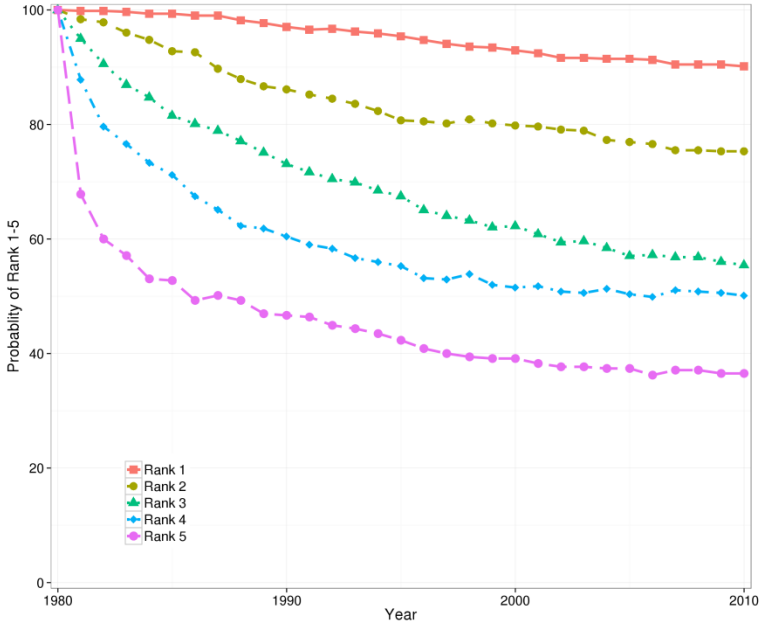
The cross sectional relationship between industries and technologies highlights significant churn in the yearly composition of which technologies are most strongly associated with each industry. However,

there is evidence that while the year-to-year changes may be dramatic, it tends to level off almost immediately afterwards, with limited changes to the weights and rankings after the first period, which may be indicative of persistence in the technology composition of industries. To further test this idea of persistence, we next look at the cumulative industry-technology weights each year. This measure will inherently be more stable because as the corpus of patents grows each additional patent has less of an effect on the concordance weights. This type of cumulative measure, however, allows us to take into account the impacts of past technologies while updating each year with additional information from newer innovations.

Cumulative Technology Variation

To start, we first compare the persistence of the top ranked technologies by industry over time. Figure 6 shows the probability that the technologies ranked 1 through 5 in 1980 remain in the top 5 ranked technologies. The top 4-digit CPC technology in 1980 has more than an 89% probability of remaining in the top 5 in 2010. Even the rank 5 technologies in 1980 have nearly 40% chance of being in the top 5 thirty years later. The probability of remaining a top ranked technology decays over time, but as shown in the chart, the majority of the movement occurs very early on, before levelling off. In most cases, the probabilities seem to level out after 1990, continuing to decline but at a slower rate. The figure suggests strong persistence in the relationship between technologies and industries over time.

Figure 6: Cumulative Probability Rank of 1980 Rank



Source: Cumulatively calculated ALP concordances.
Notes: Captures probability that a technology in a given rank in 1980 will appear in ranks 1 through 5 in subsequent years.

As with the cross sectional yearly weights we can calculate rank-to-rank transition probabilities between 1980 and 2010, as shown in Table 4. Overall, the results suggest greater persistence in the rankings for the cumulatively calculated weights compared to the annually calculated results in Table 2. There is a 66% chance that the top ranked 4-digit CPC in 1980 is still the top ranked technology in 2010 and nearly 90% chance that the top ranked 4-digit CPC in 1980 is in the top 5 ranked technologies in 2010. This

persistence continues to hold for some of the lower ranked technologies as well. The average year-to-year Spearman's rank-order correlation coefficient is 0.98 with all estimates statistically significant. These findings again point to the stability of in the relationship between technologies and industries.

Table 4: Probability Matrix of Cumulative Rank Transitions between 1980 and 2010, Cumulative Weights

		2010 Ranking										
		1	2	3	4	5	6	7	8	9	10+	Not Ranked
1980 Ranking	1	66.34	12.81	6.73	3.12	1.15	0.82	1.48	0.82	-	0.82	5.91
	2	12.43	37.48	14.41	5.59	5.41	1.44	1.26	0.54	0.36	0.18	20.90
	3	5.62	11.45	22.09	10.64	5.62	3.82	1.20	1.00	0.80	1.20	36.55
	4	3.51	7.03	10.30	16.39	12.88	3.75	3.51	1.64	1.64	1.41	37.94
	5	2.03	4.64	7.54	10.43	11.88	6.38	5.51	1.74	2.03	2.32	45.51
	6	0.74	2.59	5.19	7.78	7.41	7.78	7.78	5.19	0.37	4.07	51.11
	7	1.34	2.23	3.13	6.25	4.46	8.48	5.36	5.80	2.23	4.02	56.70
	8	1.84	3.07	3.07	3.68	3.68	2.45	5.52	4.91	3.07	8.59	60.12
	9	1.64	0.82	0.82	3.28	1.64	7.38	7.38	6.56	3.28	7.38	59.84
	10+	1.08	1.08	1.43	0.36	2.87	2.51	1.79	1.79	4.30	19.71	63.08
	Not Ranked	9.56	14.30	12.97	12.17	10.23	10.29	7.15	6.28	4.48	12.57	-

Source: Cumulatively calculated ALP concordances.
Notes: Includes granted patents with an application date between 1980 and 2010. ALP concordances include 6-digit 2002 NAICS and 4-digit ISIC Rev 3. Results are similar for 6-digit HS and 4-digit SITC product codes.

To quantify the extent of this persistence in the probability weights we return to an autoregressive model to measure the extent to which lagged weights predict current weights. Table 5 shows the results of estimating autoregressive models of the weights between technologies and industries with lags of various lengths. The estimates show that 97% of the variation in the weights is explained by the weights in the previous period. Interestingly, this explanatory power remains high even after significant time has passed, with the thirty year lagged weight explaining 54% of the variation in the current weights.

Table 5: Autoregressive Model of Cumulatively Calculated Weights, Varying Lag

	Cumulative Weight at t, δ=1	Cumulative Weight at t, δ=5	Cumulative Weight at t, δ=10	Cumulative Weight at t, δ=15	Cumulative Weight at t, δ=30
Cumulative Probability Weight, t-δ	0.987***	0.952***	0.915***	0.880***	0.795***
	(0.000868)	(0.00352)	(0.00629)	(0.00876)	(0.0140)
Observations	5,516,156	4,935,508	4,209,698	3,483,888	1,306,458
R-squared	0.969	0.882	0.796	0.722	0.542

Source: Cumulatively calculated ALP concordances.
Notes: Standard errors are clustered at the four or six digit industry code in parentheses * p<0.05, ** p<0.01, *** p<0.001. Includes granted patents with an application date between 1980 and 2010.

One of the natural consequences of the ALP methodology is the flexibility to calculate weights over different subsets of patents across different time frames. This allows us to explore the relationship between technologies and industries by calculating concordance weights yearly and cumulatively. This

investigation hints at remarkable persistence in the associated technologies within industries, providing suggestive evidence that industries can largely be defined by the technologies embedded within them since the use of these technologies remains relatively constant over time.

CONCLUSION

To conclude, this paper extends the ALP methodology to a new set of data and new patent classification schemes to develop a technology-to-industry crosswalk useful to researchers working with U.S. and PATSTAT patents. The results indicate that the crosswalk is robust and has the added benefit of relying on probabilistic, automated, and easily updateable methodologies. Many existing concordances, on the other hand, use manual classification or are "stuck in time"—relying on stagnant information both in terms of technologies and classification schemes. We utilize this flexibility to explore the relationship of technologies and industries over time, finding suggestive evidence that the association of technologies and industries does not change dramatically over time. Furthermore, by leveraging text analysis and data mining techniques, the ALP methodology yields concordances to a variety of different industry classifications and aggregation levels–concordances that will enable researchers to leverage more fully U.S. and international patent data. More specifically, such joint analysis of patents and economic data may shed light on differences between USPC and CPC classification schemes.

REFERENCES

Aghion, P., & Howitt, P. (1992). A Model of Growth Through Creative Destruction. *Econometrica*, *60*, 323-351.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not) (No. w7552). National Bureau of Economic Research.

Evenson, R. and J. Putnam. (1994). Inter-Sectoral Technology Flows: Estimates from a Patent Concordance with an Application to Italy, Yale University Mimeo.

Griliches, Z. (1998). Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence* (pp. 287-343). University of Chicago Press.

Harris, C. G., Arens, R., & Srinivasan, P. (2010, October). Comparison of IPC and USPC classification systems in patent prior art searches. In *Proceedings of the 3rd international workshop on Patent Information Retrieval* (pp. 27-32). ACM.

Hirabayashi, J. (2003). Revisiting the USPTO Concordance Between the US Patent Classification and the Standard Industrial Classification Systems. In *WIPO-OECD Workshop on Statistics in the Patent Field, Geneva, Switzerland*.

Levin, R. C., Klevorick, A. K., Nelson, R. R., Winter, S. G., Gilbert, R., & Griliches, Z. (1987). Appropriating the returns from industrial research and development. Brookings papers on economic activity, 783-831.

Lybbert, T. J., & Zolas, N. J. (2014). Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity. Research Policy, 43(3), 530-542.

Mansfield, E., Schwartz, M., & Wagner, S. (1981). Imitation costs and patents: an empirical study. The Economic Journal, 907-918.

National Science Foundation. 2014. "Business Research and Development and Innovation: 2011." NSF 15-307.

National Science Foundation. 2013. "Business Research and Development and Innovation: 2008-10." NSF 13-332.

National Science Foundation. 2011. "Research and Development in Industry: 2006-07." NSF 11-301.

Romer, P. M. (1990). Endogenous Technological Change. *Journal of Political Economy*, *98*(5 pt 2).

Schmoch, U., F. LaVille, P. Patel, and R. Frietsch. (2003). Linking Technology Areas to Industrial Sectors: Final Reports to the European Commission, DG Research.

Schmookler, J. (1966). Invention and economic growth.

Van Looy, B., C. Vereyen and U. Schmoch. (2014). Patent Statistics: Concordance IPC V8 – NACE Rev. 2., EUROSTAT.

Verspagen, B., T. van Moergastel, and M. Slabbers. (1994). MERIT Concordance Tables: IPC-ISIC (Rev. 2), MERIT Research Memorandum.

APPENDIX A

In this section, we focus on the technology composition of products, as opposed to industries. The product categories are much more varied and contain many more observations. For instance, there exists between 200-400 4 and 6 digit production categories of ISIC and NAICS, there are more than 1,000 product categories in the SITC and HS classifications. Hence, we separated the results from the main paper. However, the results largely conform to the industry results shown in the paper. While there is significantly more churn in the product categories on a yearly basis, the cumulative technology variation still remains surprisingly stable.

Table A1: Industry-Product Concordances, Probability Matrix of Rank Transitions between 1980 and 2010, Yearly Weights

		2010 Ranking										
		1	2	3	4	5	6	7	8	9	10+	Not Ranked
1980 Ranking	1	29.16	8.89	4.07	2.36	1.47	0.79	0.73	0.20	0.12	0.22	52.01
	2	11.18	10.81	6.51	3.27	2.04	1.25	0.71	0.52	0.27	0.32	63.11
	3	7.66	7.41	6.95	3.73	2.83	1.45	1.03	0.13	0.23	0.39	68.19
	4	5.75	5.18	6.12	4.15	3.45	2.26	1.11	1.56	0.74	0.70	68.98
	5	3.46	5.25	5.93	3.83	2.99	1.99	0.79	1.21	0.31	0.94	73.28
	6	2.76	3.25	3.11	3.32	3.61	2.40	2.05	1.49	1.20	1.77	75.04
	7	1.50	3.93	2.80	3.46	3.18	1.96	1.87	1.59	0.84	1.12	77.76
	8	3.46	3.46	4.12	2.66	2.53	2.13	2.26	1.73	1.06	1.46	75.13
	9	2.98	1.68	2.42	3.91	3.17	2.23	0.56	2.61	0.56	1.49	78.40
	10+	1.79	1.20	2.09	1.89	3.19	1.99	3.49	1.89	1.20	2.59	78.66
	Not Ranked	18.47	19.30	15.57	13.20	9.57	7.55	5.61	3.74	2.70	4.30	-

Source: Annually calculated ALP concordances.
Notes: Includes granted patents with an application date between 1980 and 2010. ALP concordances include 6-digit HS and 4-digit SITC product codes.

Table A1 highlights that more than 45% of the top technology in 1980 products is still within the top 5 technologies in 2010. On the other hand, more than 50% of the top technologies drop out completely on a yearly basis. Table A2 provides the cumulative ranking probability matrix.

Table A2: Industry-Product Concordances, Probability Matrix of Rank Transitions between 1980 and 2010, Cumulative Weights

		2010 Ranking										
		1	2	3	4	5	6	7	8	9	10+	Not Ranked
1980 Ranking	1	58.72	15.69	6.55	3.96	2.64	1.72	0.65	0.69	0.49	0.35	8.54
	2	15.63	31.02	12.31	6.72	3.28	1.92	0.90	0.90	0.81	0.79	25.72
	3	6.97	12.18	16.34	9.63	4.80	5.07	2.02	1.69	0.55	0.86	39.89
	4	3.82	7.58	9.68	11.04	7.65	4.25	3.31	1.55	0.70	1.00	49.42
	5	3.59	5.14	7.50	10.12	8.31	5.26	3.32	1.82	0.93	1.43	52.59
	6	3.24	2.88	5.49	7.53	7.06	5.81	3.87	2.82	1.57	2.30	57.43
	7	1.96	1.26	5.52	3.98	5.59	4.19	6.35	3.00	2.51	3.00	62.64
	8	1.95	2.34	4.88	3.80	3.90	5.56	4.88	4.20	2.63	2.44	63.41
	9	1.75	2.56	2.97	4.32	4.32	5.67	6.07	3.37	3.64	6.48	58.84
	10+	1.52	2.28	2.28	2.28	2.87	2.96	2.11	3.80	4.48	8.20	67.20
	Not Ranked	14.16	16.97	16.57	13.60	11.27	8.23	6.56	4.71	3.20	4.75	-

Source: Annually calculated ALP concordances.
Notes: Includes granted patents with an application date between 1980 and 2010. ALP concordances include 6-digit HS and 4-digit SITC product codes.

In Table A2, we find that more than 85% of the number 1 technology in 1980 is still in the top 5 technologies in 2010, with nearly 60% remaining in the top position. This again, highlights the dramatic persistence of associated technologies into product categories over time. Finally, in Table A3, we provide the Yearly and Cumulative autoregressive models over time.

Table A3: Industry-Product Concordances, Autoregressive Model of Cumulatively Calculated Weights, Varying Lag

	Yearly Weight at t, δ=1	Yearly Weight at t, δ=15	Yearly Weight at t, δ=30	Cumulative Weight at t, δ=1	Cumulative Weight at t, δ=15	Cumulative Weight at t, δ=30
Probability Weight, t-δ	0.477***	0.407***	0.360***	0.983***	0.833***	0.705***
	(0.00448)	(0.00490)	(0.00678)	(0.000273)	(0.00280)	(0.00461)
Observations	23,333,370	12,444,464	777,779	29,555,602	18,666,696	7,000,011
R-squared	0.228	0.161	0.132	0.958	0.637	0.415

Source: Annual and cumulatively calculated ALP concordances.
Notes: Standard errors are clustered at the four or six digit industry code in parentheses * p<0.05, ** p<0.01, *** p<0.001. Yearly calculated concordances include granted patents with an application date between 1980 and 2010. Cumulatively calculated concordances include granted patents with an application date between 1976 and 2014.

The auto-regressive models show more variation than the industry counterparts, but are still strongly predictive of future weights.