Tracking the Technological Composition of Industries with Algorithmic Patent Concordances[*]

Nathan Goldschlag[†] Travis J. Lybbert[‡] Nikolas J. Zolas[§]

July 9, 2019

Abstract

Patents are a useful proxy for innovation, technological change, and diffusion. However, fully exploiting patent data for economic analyses requires linking patents to measures of economic activity, which has proven to be difficult. We construct and update the probabilistic linkages between the U.S. Patent Classification (USPC) system and Cooperative Patent Classification (CPC) system and industry and product classifications including the North American Industrial Classification System (NAICS), International Standard Industrial Classification (ISIC), Harmonized System (HS) and Standard International Trade Classification (SITC). We use these concordances to evaluate the persistence of technology-industry relationships over time by generating linkages over different years of patent data. We find strong persistence in technology usage within industries and, until recently, relatively little change in the technology composition of industries over time. As the technology composition of industries becomes more stable, we find evidence of increased specialization. Finally, we show that industries that exhibit changing technology composition also show shifting occupational composition.

Keywords: patents, concordance, classifications, USPC, CPC, technological change, occupational composition

[*] Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views 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, Daniele Moschella, participants at the CES brown bag seminar series, participants at the 2015 FCSM conference and two anonymous referees for their helpful comments. The previous version of this paper was titled, "An 'Algorithmic Link with Probabilities' Crosswalk for USPC and CPC Patent Classifications".

[†] U.S. Census Bureau - nathan.goldschlag@census.gov

[‡] University of California, Davis - tlybbert@ucdavis.edu

[§] U.S. Census Bureau (corresponding author) - nikolas.j.zolas@census.gov

1 Introduction

Systematically linking technologies and industries has proven to be difficult (Dosi and Nelson, 2010; Antonelli, 2014), hampering tests of technology diffusion and technological change as a key driver of economic growth (Romer, 1990; Aghion and Howitt, 1992). Due to their wide availablity and rich technological detail, patent data has frequently served as a primary source for describing the technological composition of industries (Griliches, 1990). For the purpose of systematically linking technologies and industries, the technology classification systems used to organize patents are particularly promising due to their high resolution and hierarchical structure. In addition, these classification systems are useful as research tools, as they can offer a unique dynamic perspective on the evolving application of technologies new and old in specific industries (Lafond and Kim, 2017).

There is a long history of using patent data to measure how the technology used by industries has changed over time. Historically, many of these measures relied on citation weighted patent counts.[1] These changes are often incremental and difficult to measure unless a large-scale or sudden technological shock transforms an industry. Identifying these small, incremental changes is challenging for a variety of reasons. If we think of different types of technologies as ingredients (i.e. the "recipe") for the production of a product or service, as in Dosi and Nelson (2010), measuring direct technological change would be akin to quantifying changes to a recipe. Hence, the majority of case studies have relied on the emergence and spread of a radical innovation within a narrow set of industries such as semiconductors (Dosi, 1984), the tire industry (Klepper and Simons, 2000), the typesetting industry (Tripsas, 1997), or some small disparate collection of industries as in Klepper and Simons (2005a).

However, recent improvements in data collection, natural language processing techniques, and network analysis tools have led to a number of indices that leverage the textual content of patents and citation networks to create new indices of innovative activity (Kelly,

[1] See Griliches (1990) for an early description of how patents can be used as measure of innovation. See Kortum and Lerner (1998) and Hall, Jaffe, and Trajtenberg (2001) for discussions of the heterogeneity of patent value and the use of citation measures to characterize the impacts of individual patents.

Papanikolaou, Seru, and Taddy, 2018) and capture the gradual changes of technology-use within an industry (for instance, (Krafft, Quatraro, and Saviotti, 2011) use patent citation analysis to look at technology use in biotech industries). Our approach to measuring technological change within industries falls within this latter group, extrapolating the types of industries that a patent relates to using the text of the patent. We create a comprehensive measure of technology change for that is capable of capturing relatively modest changes not driven by radical innovations. Though these changes may seem small over short periods of time, they can accumulate and manifest as significant over several decades.[2]

To classify our measure of technological change, we rely on commonly used industry classifications developed by statistical agencies and the technology classifications provided by patent examiners. The United States Patent and Trademark Office (USPTO) has historically used two classification schemes to organize patents by the types of technologies embodied within them: the United States Patent Classification (USPC) and the Cooperative Patent Classification (CPC). The USPC system was for many years the technological classification system used by the USPTO to classify patents. In recent years the CPC scheme, the result of a cooperative effort between the USPTO and the European Patent Office (EPO) to develop an internationally compatible technology classification system, has been used to classify patents.

In this paper we use a probabilistic linkage methodology first developed by Lybbert and Zolas (2014) to create concordances between USPC and CPC technology codes and industry and product classifications including 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. We then use those concordances to analyze how technology-industry relationships have changed by allowing the set of patents contributing to the concordance to vary over time. We find

[2] For example, memory storage devices have undergone incremental changes each year in the speed and amount of data one can store. However, there has also been significant long-term change in the underlying technology as the industry moves away from hard disk drives (HDD) to solid state drives (SSD).

that the link between technologies and industries is remarkably persistent and the amount of change, which has been declining for decades, began increasing in recent years. We also provide suggestive evidence that our measure of technological change is economically meaningful. Changes in the industry technology composition correlates with changing occupational composition, which is consistent with the literature on the labor market effects of new technologies.

One of the primary advantages of the probabilistic linkage methodology used is that it can be continually updated as new patent data become available. This also means we are able "rewind the clock" to measure the persistence of industry-technology linkages over time by utilizing different sets of patents. Doing this over a 35 year period we find that the technology-industry relationships are very persistent. For nearly 40% of industries, the top (most strongly associated) technology in 1980 remains at the top in 2010. Moreover, when probability links are calculated within 5-year windows, 92% of the year-to-year variation in linkage weights can be explained by weights in the previous period and nearly 50% of the variation can be explained by the weights from 30 years earlier.[3]

By computing many different concordances based on different sets of patents, we are able to quantify the changing technological composition of industries over time. We do this by calculating the Euclidean distance between industry-technology links between periods. We find that this within-industry measure of technical change has been steadily declining through the mid 2000s. Between 1987 and 2005, the 10-year change in technology composition fell 8% within manufacturing industries, recovering completely from that decline by 2012. This decline in within-industry technology change has coincided with a marked increase in industry-level technology specialization. The concentration of technologies within industries (less disperse industry-technology links) increased more than 60% between 1982 and 2008. However, we also find that immediately following recessions, there is a relatively steep decline in technological concentration, which is perhaps indicative of increased ex-

[3] These findings are consistent with Colombelli and Quatraro (2014), who find persistence within firm-level knowledge production functions.

perimentation taking place following an industry shakeout. These findings are consistent with theories on the life-cycle of industrial innovation first proposed in Gort and Klepper (1982) and more recently in Klepper and Simons (2005b) where periods of rapid industrial change are followed by industrial shakeout and increased concentration before a new cycle of innovation occurs.

Finally, we explore whether our measure of technology change correlates in meaningful ways with shifting labor demand. We provide suggestive evidence that our measure of changing technology composition is positively associated with changing occupational composition. Industries that experience more significant changes to their associated technologies also saw changes in the types of workers employed in those industries. These findings are consistent with the literature on skill-biased technology change and job polarization that has shown how changes in technology affect labor demand (Katz and Autor, 1999; Acemoglu and Autor, 2011a; Goos, Manning, and Salomons, 2014).

The remainder of the paper is structured as follows. Section 2 provides background on existing technology concordances and the dynamics of technological change. Section 3 describes how the concordances are constructed and how they are used to measure technology persistence, change, and composition. Section 4 provides the results of several exercise to validate the concordances and analyses of our new measures of technological persistence, change, and composition and how technological change relates to occupational dynamics. Section 5 concludes.

2 Background

Previous research linking technologies to industries has focused on R&D (Scherer, 1984) and/or patents (Ernst, 1997). Patents have been used in a number of studies as a proxy of technological change (Griliches, 1990). 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 in the abstract and claims, 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 USPC classification system, 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 by 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, Arens, and Srinivasan, 2010) in a hierarchical structure. 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 all patents with concordances used to link them to older patents.

While patents in the U.S. typically list both a USPC and IPC/CPC classification schemes, there are several reasons why a strictly USPC-based concordance may be of interest. First, there is no clear disambiguated link between USPC and IPC/CPC technology classes. Second, along with being well documented, the USPC is the oldest and one of the most frequently used patent classifications. For these reasons, we believe a strictly USPC-based concordance will be of use for researchers studying innovation, particularly for longitudinal studies, despite the fact that the USPC is being discontinued. In addition to the USPC concordance, and in part because it is being discontinued, we also develop a concordance for CPC codes. To our knowledge, there are no previous efforts to directly translate CPC codes to industry and product classifications.

There have been several efforts to translate USPC to other classification schemes including (Schmookler, 1966), Putnam and Evenson (1994), (Schmoch, Laville, Patel, and Frietsch, 2003), and (Hirabayashi, 2003). The resulting concordances have a number of limitations including manual curation and one-to-many (or many-to-one) mappings (see Lybbert and Zolas (2014) for a discussion of these limitations). None of the existing methodologies implemented in these earlier studies are flexible enough to incorporate newer data or classification schemes. In some cases, multiple concordances must be layered to achieve the desired USPC to industry links. Layering concordances yields a noisier linkage as each layer has its own measure of uncertainty.[4] The noise gets compounded when the links are one-to-many or many-to-one.

The USPC patent classification has undergone significant changes since its inception, with new technologies continuously added and outdated technologies removed. One of the major benefits of the patent classification systems used by the USPTO is that changes to the classification system are backcasted onto patents classified prior to the changes.[5] Thus, if a new technology class is created in a given year but is applicable to a patent application filed prior to that date, the patent is “corrected” with the new technology class. Several studies that have analyzed the long-run evolution of patent classification systems, primarily for the USPC (Strumsky, Lobo, and Van der Leeuw, 2012; Strumsky and Lobo, 2015; Youn, Strumsky, Bettencourt, and Lobo, 2015; Lafond and Kim, 2017). For our purposes this consistency in the patent-level classification scheme allows us to develop a consistent mapping between different vintages of technology, industry, and product codes.[6] It is important to note that the backcasting of newer classification vintages does not imply the level of persistence of industry-technology linkages we describe in the following sections. There is no reason to suspect that, simply because patents are classified using a single

[4] Layering concordances is akin to translating Japanese into English by first translating it into French.

[5] This same benefit may also be seen as a drawback as information about the original technological intent of the patent is lost.

[6] The fact that technology classification systems change over time, and that those changes are applied to the entire patent corpus, also underscores the importance of a generalizable concordance that can be continuously updated as new technology classes emerge and are incorporated into the existing patent database. Our methodology allows us to do just that, re-calculating the concordance weights as new data become available.

vintage of the classification system, the text associated with certain technology codes ought to remain unchanged.

The flexibility of the probabilistic linkage methodology allows us to generate concordances from any arbitrary subset of patents, providing a natural way to measure the persistence of industry-technology links. Varying degrees of persistence in industry-technology links have the potential to affect the structure of firms and industries by altering labor demand and employment dynamics. The literature on skill-biased technology change and job polarization has shown how changes in technology shift labor demand away from low skilled workers towards higher skill, higher education workers, which can "hollow out" the middle of the earnings distribution (Katz and Autor, 1999; Acemoglu and Autor, 2011a; Goos, Manning, and Salomons, 2014). Technology can both augment and substitute labor depending on the type of tasks being performed. Information technologies, for example, induce a number of indirect effects on labor, beyond the rising demand of IT skilled workers, through changes in organizational practices, changes in products and services, and complementarities with workers performing complex tasks (Bresnahan, Brynjolfsson, and Hitt, 2002a; Autor, Levy, and Murnane, 2003). In the case of automation technologies, the decline in labor demand due to machines replacing labor can be counteracted by productivity enhancing effects, capital deepening, and the creation of new labor-intensive tasks (Acemoglu and Restrepo, 2018). These studies, taken together, suggest that industries experiencing greater technology change are likely to see more significant changes in occupational composition.

3 Methodology

3.1 Building USPC and CPC Concordances

The methodology we use to construct the linkages between USPC and CPC codes and industry and trade classifications follows the Algorithmic Links with Probabilities (ALP) approach first developed by Lybbert and Zolas (2014). Since the methodology is not novel, a detailed description of our implementation of the ALP methodology and matching results can be found in Appendix A and B respectively. The methods we implement here are virtually identical to those found in Lybbert and Zolas (2014) with a similar set of search terms and reweighting scheme. However, the underlying patent set and technology classifications are different. We also include updated keywords and search terms for detailed industry, product, and trade classifications.[7]

We use keyword and search term databases to match industry, product, and trade classifications to patent titles and abstracts derived from two patent databases: the USPTO PatentView database[8] and PATSTAT database[9]. The PatentView database provides us the frame against which we match for the USPC concordance, while the PATSTAT provides the frame used for the CPC concordance. Once we combine the two databases, we filter the matches, calculate, and reweigh linkages.

The full suite of concordances for each technology classification are available for researchers. One of the most significant advantages of the ALP methodology is its flexibility and repeatability. This allows us to regenerate the concordance to accommodate changes in the technological landscape or updated classification schemes. We leverage this flexibility to explore the dynamics of the linkages by comparing how the concordance changes as the underlying set of patents evolves over time. Specifically, we can use the ALP methodology to create numerous crosswalks that mine the text of different sets of patents over time. As the text of the patent documents changes over time the ALP weights will shift to reflect different technology-industry configurations. For the CPC concordance, we develop a second suite of crosswalks (described below) utilizing a rolling window of patents. This allows us to capture the evolving relationship between patents and industries characterized by the changing concordance weights. We use these changing weights over time to measure technological

[7] We include updated search terms and key words for: 6-digit North American Industrial Classification System (NAICS), versions 1997, 2002 and 2007, 4-digit International Standard Industrial Classification (ISIC), versions 2.0, 3.0, 3.1 and 4.0, 6-digit Harmonized System (HS), versions 1997, 2002 and 2007, 4 and 5-digit Standard International Trade Classification (SITC), versions 2, 3 and 4.

[8] See http://www.patentsview.org (accessed 2/2/2016))

[9] PATSTAT Global bulk download was purchased and accessed in the second-half of 2016)

persistence and technological change.

3.2 Technological Persistence

We measure the persistence in the technological weights of industries by quantifying how much the crosswalk changes from year-to-year and over long time periods. Using both yearly and cumulative crosswalks we can calculate the rank of each technology within each industry using the ALP weights for 2002 6-digit NAICS industries and 4-digit CPC technologies.[10] The "yearly" concordance weights are calculated by generating industry-technology links using only patents granted within a 5-year window surrounding a given year (e.g. a moving average). The "cumulative" concordance, on the other hand, utilizes all patents granted between 1976 and a given year. Using granular industry and technology codes allows us to capture relatively small changes in technology-industry associations that may only be observable within relatively narrow categories.

Using cumulatively calculated concordance weights we rank technologies by concordance weight. The top-ranked technology for each industry can be thought of as the technology most strongly linked, which contains patents with descriptions that are closely associated with the activities occurring in that industry. Using these ranks we calculate rank-rank transition matrices over long time horizons (1980 to 2010) and annual measures of the probability that a top rank (1 to 5) technology in 1980 is still top ranked each year between 1981 and 2010. This type of analyses, assessing the rank correlation over different time periods, provides insight on the persistence of technology-industry relationships over time.

Focusing on changes in the discrete rankings of technologies within industries may mask significant heterogeneity in the distribution of the continuously measured probability weights. To further analyze the persistence of industry-technology links, we estimate the predictive power of prior technology weights for current period links. This can be accomplished by

[10] We elect to construct the yearly weights using the CPC classification because we have a larger patent corpus to match against (PATSTAT). We have performed an identical analysis using USPC codes and find similar levels of persistence and change. Also, the choice of industry classification is not of consequence as similar findings were made using alternative classification schemes and years.

estimating an autoregressive models of varying length using the yearly concordances as shown in Equation 1.[11]

$$Weight_{i,a,t} = \alpha + \beta_1 Weight_{i,a,t-\delta} + \epsilon_{i,a,t} \tag{1}$$

where $Weight_{i,a,t}$ and $Weight_{i,a,t-\delta}$ are the ALP probability weights between industry i, technology a, in year t and $(t-\delta)$ respectively. The estimates from this autoregressive specification allow us to quantify the amount of variation in the probability weights for industry-technology pairs that can be explained by the weights assigned 1, 5, 10, 15, and even 30 years lagged.

3.3 Composition and Technological Change

Using the yearly and cumulative crosswalks, we quantify how much the technological composition of each industry changes over short and long time periods. Specifically, we calculate the Euclidean distance between the vector of yearly concordance weights for each industry at time $(t-\delta)$ and t using 4-digit CPC technologies and 6-digit NAICS codes. Given the ALP weight $w_{i,t}$ for an industry i at time t, and technology classes $a = 1, ...A$, we take the square root of the changes in technology weights across all technologies between time t and $t-\delta$, shown in Equation 2.

$$d(w_{i,t}, w_{i,t-\delta}) = \sqrt[2]{(w_{i,1,t} - w_{i,1,t-\delta})^2 + (w_{i,2,t} - w_{i,2,t-\delta})^2 + ... + (w_{i,A,t} - w_{i,A,t-\delta})^2} \tag{2}$$

This measure captures the total change across all technologies linked to an industry.[12]

[11] This is similar to what is known as measuring the "persistence parameter" in the macro literature. A study looking at the persistence of innovative activities using a similar methodology was done by Malerba, Orsenigo, and Peretto (1997). The main idea is that the larger the coefficient value of β_1 and the more variation that can be explained by the prior weights, the higher the degree of persistence.

[12] We restrict our analysis to industries who have patent coverage in both the $(t-\delta)$ and t. This allows us to disregard technology-industry entry and exit, which have the potential to bias our distance weighting.

In the following section we build intuition for what these measures capture by presenting examples of industries that experienced significant amounts of change over the past several decades.

In addition to instructive examples, we also examine how the average technological change and concentration of links has evolved over time. We calculate the average difference in technological composition over 5, 10, and 20 year intervals and explore whether industries are becoming more or less specialized by measuring the relative concentration of industry-technological linkages using a Herfindahl-Hirschman index (HHI) of the concordance probability weights. If industries are becoming more specialized in their technological composition we would expect the weights associated with top ranked technologies to increase relative to lower ranked technologies, increasing the concentration of our probability weights. In contrast, if industries are becoming more generalized in terms of technologies, we would expect the ALP weights to become less concentrated. This has implications on the life-cycle of industrial innovation. Industries that undergo a period of rapid innovation experience an industrial shakeout as the firms and technology become more concentrated (Gort and Klepper, 1982; Klepper and Simons, 2005b).

Canonical models of skill and labor demand suggest that the factor-augmenting properties of technology lead to shifts in the demand for skill (Katz and Murphy, 1992). Task-based models tell us that technological change can alter the mapping of skills and tasks and change the task content of occupations (Acemoglu and Autor, 2011b). For example, Bresnahan, Brynjolfsson, and Hitt (2002a) document how rapid changes in information technology generated capital-labor complementarities that led to changes in job responsibilities and more decentralized organizational practices. With these models and empirical evidence in mind, we test whether our measure of technological change is associated with changing labor demand via shifts in the occupational composition of industries.

We utilize occupational data from the Bureau of Labor Statistics Occupational Employment Statistics (OES), which provides employment by occupation within four-digit NAICS industries. We construct occupational shares for each industry and calculate the Euclidean distance of these changing shares over time similar to the industry-technology link distance in Equation 2. We estimate the regression specification in Equation 3 to assess the correlation between our technology change measure and changes in the occupation mix of industries.[13]

$$\Delta OccDist_{i,t} = \beta_1 \Delta TechDist_{i,t} + \gamma_t + \delta_i + \upsilon_{i,t} \tag{3}$$

where $\Delta OccDist_{i,t}$ is the 5-year log Euclidean distance of the occupational composition of industry i at time t and time $t-5$. $\Delta TechDist_{i,t}$ is the 5-year log Euclidean distance of technology composition for industry i at time t using the yearly weights. γ_t and δ_i are year and industry effects respectively. The year and industry effects are used to capture unobserved heterogeneity in how the occupations are classified and tabulated within each industry over time, as well as unobserved heterogeneity in how the technology weights are captured and reported.

In Equation 3, our measure of $TechDist_{i,t}$ may be sensitive to the amount of innovative activity in an industry. Some industries are more reliant on patents as an appropriability mechanism than others. Large changes in technology weights for an industry that has a substantial number of associated patents may have different implications than for an industry with relatively few patents. To address these issues we estimate activity-weighted regressions based on the patent counts for each industry i. Another concern with the above estimation specification is the difference in magnitudes of changes in $OccDist_{i,t}$ and changes in $TechDist_{i,t}$. Changes in the occupation distance are significantly smaller than the magnitude of changes in technological distance. As a robustness exercise, we estimate a rank-rank specification where we rank the industries by changes in occupational distance and by changes in technology distance and run a similar estimation. If the association between technological change and occupational change is positive, we expect to see the industries with the largest

[13] We limit the analysis to this simple, reduced form equation to keep the association as general as possible and focus primarily on within-industry changes to account for unobserved year heterogeneity.

technological changes to be among the industries with the largest occupational changes.

4 Results

4.1 External Validity of New Concordances

Before analyzing our new measures of technology persistence and compositional change, we validate the accuracy and robustness of the concordances. The USPTO generated a hand-curated concordance between the USPC and a set of 30 product fields as part of a report describing industry patenting trends. These reports provide an external validation of our USPC-industry concordances. The product fields, or NAICS-based categories, were loosely based on 2002 NAICS industries. 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. Each USPC is assigned between one and seven NAICS-based categories, each equally weighting.[14]

As shown in Figure 1, our results compare quite favorably with the USPTO's concordance despite significant differences in how they were constructed.[15] The largest discrepancies occur in industries 311 ("Food") and 312 ("Beverages") where the ALP estimates are greater than the USPTO's estimates and industries 3252 ("Resin, Synthetic Rubber, and Artificial and Synthetic Fibers and Filaments") and 3254 ("Pharmaceutical and Medicines"), for which the ALP concordance generates lower patent counts compared to the USPTO estimates. These outliers likely reflect limitations that arise from the construction of search terms that are the basis for the ALP concordances. Industries with significant product heterogeneity such

[14] These NAICS-based tabulations rely heavily on the earlier USPC to SIC concordances originally developed in 1974.

[15] A valid comparison with the concordance developed by the USPTO requires us to first translate our ALP concordances into the NAICS-based categories (OTAF codes) used by USPTO. We next ensure that our sample of patents are as close to the sample as that used by USPTO, meaning that we primarily rely on utility patents granted 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, which yields us our comparison.

as food and beverages will tend to have a much more diverse set of search terms, resulting in more matches 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 extremely specialized terms (such as pharmaceuticals), will likely have fewer search terms associated with them, thereby resulting in slightly lower weights.

The PATSTAT database provides an additional source of external 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 by Schmoch, Laville, Patel, and Frietsch (2003), which uses the main industry of firm-owned patents to generate a mostly 1-to-1 match been an IPC and NACE.[16] Figure 2 shows the comparison of patent counts by NACE using the EUROSTAT and ALP concordances.[17] 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.

4.2 Persistence of Industry-Technology Relationships

Next we use CPC concordances to analyze the persistence of industry-technology relationships.[18] We begin by examining changes in the ranks of the technology weights within industries for the 5-year window of patents centered around 1980 and the 5-year window of

[16] 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).

[17] For this comparison, we extract the patents that can be concorded using both methodologies and assign equal weights to all 4-digit CPCs. We then concord them to the 2-digit NACE using the EUROSTAT concordance. For the ALP counts, we must first assigne the CPCs to ISIC Rev. 4, which the NACE Rev. 2 is based upon. We can then easily convert them to 2-digit NACE Rev. 2 using the correspondence table found in UN Statistics Division website.

[18] The set of patents used for the CPC concordance is significantly larger and based on worldwide patents (as opposed to only U.S. patents). The results also hold for USPC-based concordances.

patents centered around 2010. Table 1 shows the probability that a 4-digit CPC technology with a given rank in 1980 will be in each corresponding rank in 2010. The top ranked technologies in 1980 have a 39% probability of remaining the top technology and more than 50% probability of being in the top-2 technologies in 2010. Lower ranked technologies exhibit more churning. Technologies ranked 5th in 1980 have a similar chance of making it to rank 1 as they do of staying rank 5, with probabilities ranging from 4.5% to 6.3% respectively. Comparing the transition matrices between 1980 and each year up to 2010, the Spearman's rank-order correlation coefficients of the probability weight rank between subsequent years are all significant with a mean if 0.75, suggesting a strong positive relationship between a technology's rank from one year to the next.

An alternative way of measuring persistence is to allow the patent corpus to grow, capturing the cumulative effect of changes in the relationship between technologies and industries. Figure 3 shows the probability that the technologies ranked 1 through 5 in 1980 remain in the top 5 ranked technologies using cumulative crosswalks. 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, flattening out after the first few years. In most cases, the probabilities become more stable after about 5-10 years, continuing to decline but at a slower rate. This type of stasis may signal persistence in the structural relationship between technologies and industries.

Table 2 shows the estimation results of the autoregressive specification in Equation 1 for different lags for both yearly and cumulative crosswalks. In the top panel, using yearly crosswalks, estimates suggest that 92% of the variation in the weights is explained by the weights in the previous period. This explanatory power remains high even after significant time has elapsed, with the thirty year lagged weight explaining 49% of the variation in the current weights and hinting at a nearly one-to-one relationship between the weights in 1980 and 2010 (coefficient value of 0.865). The results suggest that despite the tremendous amount of technological change between 1980 and 2010, the technologies associated with industries moves quite slowly and is quite persistent over time. The bottom panel shows estimates using cumulative concordances. The estimates are even stronger than the rolling window version and show yet greater persistence when we include prior patents as we add new years and patents. While the cumulative concordances by construction place more weight on past technology-industry links, the relative importance of these patents should naturally fade if there were significant changes in technology-industry relationships.

The persistence of technology weights and rankings suggests that annual updates of the ALP crosswalks with newly granted patents will have only modest effects on linkage weights. As the corpus of patents changes over time, the technology weights have changed relatively little over time, suggesting useful resilience in concordances based on the ALP methodology.

4.3 Technological Change, Composition, and Occupation Dynamics

Technology weights and rankings from the past have significant explanatory power of present-day technology weights and rankings, even when that gap spans 30 years. In this section, we describe the results from our new measures of technological change and how these new measures relate to industry occupation dynamics. As described previously, we calculate the Euclidean distance between the vector of probability weights for an industry at time $(t - \delta)$ and t. This measure captures the total change across all technologies linked to an industry. For intuition on this measure, Figure 4 and Figure 5 show the technological evolution for two manufacturing industries–one that exhibited significant change and another that remained relatively unchanged in its technology composition.

Figure 4 shows the evolution of Magnetic and Optical Media Manufacturing (NAICS 3346), which exhibited a relatively high-degree of technology change. Patents affiliated with Magnetic and Optical Media Manufacturing underwent a relatively dramatic transformation, where the primary three-digit CPC code throughout the 1980s was Information Storage (G11). By 2010, the primary CPC code for Magnetic and Optical Media Manufacturing changed to Computing, Calculating, Counting (G06), followed by Electric Communication Technique (H04). This change is mostly gradual with consistent movement throughout the time period. On the other hand, Figure 5 shows how the Apparel Accessories and Other Apparel Manufacturing industries (NAICS 3159) only saw small year-to-year fluctuations in weights across its two primary CPC codes, Wearing Apparel (A41) and Headwear (A42). The technology composition of this apparel manufacturing industry changed very little over this 35 year period.

The amount of change in industry-technology relationships varies over time. Figure 6 plots the mean 5-year, 10-year and 20-year Euclidean distances in technology composition over time. The figure shows that the average change in technology weights over the time period has been slowly declining until recent years. Between 1987 and 2005 the 10-year change technology composition fell 8% before recovering almost completely by 2012. The decline in mean technology change suggests increased stability in the types of technologies associated with industries, which may be linked to industry maturity. These patterns are consistent with an industry dynamic in which relatively newer industries, based on relatively newer technologies, exhibit more churn in their industry-technology linkages while the underlying technologies used in the industry are more fluid. Over time, industry standards and technological best practices become more clearly defined, leading to increased concentration in the weights as certain types of technology become dominant within maturing industries.

One way to measure such specialization is a Herfindahl-Hirschman index (HHI) of the concordance probability weights, which will capture the relative concentration of the mapping of technologies to industries. Figure 7 plots the mean HHI with the 5-year changes in technology composition over the same time period. The rise in the HHI of technology composition coincides with the decline in the changes to technology composition, suggesting a strong link between changes in technology composition and increased specialization.[19] As the technology composition of industries becomes more stable it also becomes less diverse. This fits nicely with the life-cycle of innovations proposed in Klepper and Simons (2005b). Also interesting, is the steep drop in concentration following the 2008 recession. Similar, but smaller drops also occur immediately following the 2001 recession and 1991 recession. Looking more closely at the data, we find that the drop is mainly due to highly concentrated industries becoming significantly less concentrated, with much of the dispersion being spread through relatively low weights across many technologies. This seems to be indicative of increased experimentation following the industrial shakeout caused by the 2008 recession.

Turning to the relationship between our new measure of technological evolution and industry occupational composition, Table 3 shows estimation results from Equation 3. The top panel shows estimates for our continuous Euclidean distance measure and the bottom panel shows estimates for the discrete rank-rank specification. There are a couple patterns worth noting. First, changes in technology composition across industries has a positive and significant association with changes to occupational composition both unweighted (column 1) and weighted (column 2) without year and industry fixed effects. The weighted specification has a much larger point estimate, which implies that the relationship between technological change and occupational change are not as strong in industries with few innovations. A ten percentage point change in the 5-year Euclidean distance in the technology composition of an industry corresponds to a 1.3 - 3.6 percentage point change in the 5-year Euclidean distance in employment composition. In other words, industries that experienced larger changes in the associated patents also experienced changes in the occupational mix of workers in the industry.

Estimates in column 3 and 4 of Table 3 incorporate industry and year fixed effects respectively. When we focus on the within-industry changes (column 3) and within industry, within-year changes (column 4), we find a similar pattern but with smaller magnitudes

[19] The correlation between these two measures is -0.88.

and less precision. This suggests that at least some of the relationship between changing technological and occupation composition is driven by other industry characteristics. It is important to note that our findings are not causal, as there exists a number of omitted variables that are not controlled for that may potentially cause both the occupational distribution to change, as well as technology usage. However, these correlations are consistent with models that link technology and labor demand.[20].

The bottom panel of Table 3 shows a rank-rank specification, which assesses how the ranks in changes to occupational composition are correlated with the ranks in changes to technology composition. The positive association continues to hold in the rank-rank measures, both across industries (columns 1 and 2) and within industry (columns 3 and 4), indicating that industries that experienced the largest changes in technology composition also saw the largest changes in occupational composition. These analyses confirm that our measure of changing technology composition varies in ways that are reasonable and consistent with the literature on the labor market impacts of changing technology. We are also able to provide new estimates, in addition to how the evolution of industry technology composition is changing over time, of the correlation between changing technology and occupational mix.

5 Conclusion

This paper uses data mining and natural language processing techniques to develop concordances between technology classifications (USPC and CPC) and industry, product, and trade classifications. Our analyses indicate that the concordances are robust, in broad agreement with similar concordances. The crosswalks we develop, in contrast to those against which we compare, rely on probabilistic, automated, and easily updateable methodologies as opposed to hand-curated linkages that rely on stagnant information both in terms of technologies and classification schemes. Moreover, the concordances we generate cover a variety of different

[20] For example, see Acemoglu and Autor (2011b), Autor, Levy, and Murnane (2003), Bresnahan, Brynjolfsson, and Hitt (2002b) Berman, Bound, and Griliches (1994)

industry classifications and aggregation levels, allowing researchers to more fully leverage U.S. and international patent data.

We utilize the flexibility of our methodology to explore the relationship of technologies and industries over time, highlighting how changes to the patent corpus reflect changes to the technology composition of industries. We find that technology-industry relationships are remarkably persistent. The top technology for an industry in 1980 has a 40% chance of remaining the top technology 30 years later. Furthermore, the amount of change in technology-industry links fell between the 1980s and early 2000s but has increased in recent years. Using the composition of concordance weights, we also find a marked increase in industry-level technology specialization, followed by a marked decrease following the Great Recession. These concentration levels are consistent with theories on the life-cycle of industrial innovation. Finally, we find that change in technology composition correlates positively with changes in occupational composition. Industries that experience more churn in their associated technologies also saw churn in their occupational mix, consistent with the notion that the introduction of new technologies induces shifting demand of skilled labor. While our analysis was limited to estimating partial correlations, we believe that this topic can be investigated further with a more fully developed model out-of-scope for our work here.

Nevertheless, our findings shed light on several measurement and policy issues and on promising lines for further research. The stability of industry-technology relationships suggests that industries can largely be characterized as technology sets that show remarkable persistence over time on average. In practice, the distinction between technology-using and technology-producing sectors is clearly important, yet our approach is notably agnostic on this distinction. We focus on technology-industry associations revealed by simultaneously considering descriptions of industry activities and the codification of technologies found in patented innovations. Determining whether these associations emerge from using technology or generating technology (or both) would be a fruitful direction for future work.

Our analyses also speak to the relationship between technological change and labor demand. These changes are often incremental with most industries undergoing only modest changes over time that are barely perceptible. Disruptive technological changes tend to receive more attention in the literature since they are more readily observed, but incremental changes may have a similarly important effects that accumulate after many years. We demonstrate that as the technology-industry links change, the occupational distribution of workers in those industries change as well and are strongly associated. Changing occupational composition, to the extent that it is driven by disruptive technologies, is an important dimension of creative destruction and reallocation related to technological innovation. With rapid advances in robotics and artificial intelligence, these critical relationships between technology and workforce composition are set to change dramatically in the near future. Developing and refining empirical tools to understand these relationships and their evolution over time is an important and policy-relevant research priority.

These differences in the propagation of new technologies across industries also opens new lines of inquiry as to the productivity and economic effects of the speed of technological change. Is it better for industries to undergo short periods of radical change with dramatic shifts in the labor demand, or longer transition periods, where the workforce has more time to adjust? To answer these types of questions one might investigate the industry productivity, entry, and exit dynamics during periods of technological change.

References

Acemoglu, D., and D. Autor (2011a) “Skills, tasks and technologies: Implications for employment and earnings,” in *Handbook of labor economics*, vol. 4, pp. 1043–1171. Elsevier.

——— (2011b) “Skills, tasks and technologies: Implications for employment and earnings,” in *Handbook of labor economics*, vol. 4, pp. 1043–1171. Elsevier.

Acemoglu, D., and P. Restrepo (2018) “Artificial Intelligence, Automation and Work,” Discussion paper, National Bureau of Economic Research.

Aghion, P., and P. Howitt (1992) “A model of growth through creative destruction,” *Econometrica*, 60(2), 323–351.

Antonelli, C. (2014) *The Economics of Innovation, New Technologies and Structural Change.* Routledge.

Autor, D. H., F. Levy, and R. J. Murnane (2003) “The skill content of recent technological change: An empirical exploration,” *The Quarterly journal of economics*, 118(4), 1279–1333.

Berman, E., J. Bound, and Z. Griliches (1994) “Changes in the demand for skilled labor within US manufacturing: evidence from the annual survey of manufactures,” *The Quarterly Journal of Economics*, 109(2), 367–397.

Bresnahan, T. F., E. Brynjolfsson, and L. M. Hitt (2002a) “Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence,” *The Quarterly Journal of Economics*, 117(1), 339–376.

——— (2002b) “Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence,” *The Quarterly Journal of Economics*, 117(1), 339–376.

Colombelli, A., and F. Quatraro (2014) “The persistence of firms knowledge base: a quantile approach to Italian data,” *Economics of Innovation and New Technology*, 23(7), 585–610.

Dosi, G. (1984) *Technical Change and Industrial Transformation: The Theory and an Application to the Semiconductor Industry.* Springer.

Dosi, G., and R. Nelson (2010) “Technical Change and Industrial Dynamics as Evolutionary Processes,” in *Handbook of the Economics of Innovation.* North-Holland.

Ernst, H. (1997) “The Use of Patent Data for Technological Forecasting: The Diffusion of CNC-Technology in the Machine Tool Industry,” *Small Business Economics*, 9(4), 361–381.

Goos, M., A. Manning, and A. Salomons (2014) “Explaining job polarization: Routine-biased technological change and offshoring,” *American Economic Review*, 104(8), 2509–26.

Gort, M., and S. Klepper (1982) “Time paths in the diffusion of product innovations,” *The economic journal*, 92(367), 630–653.

Griliches, Z. (1990) “Patent Statistics as Economic Indicators: A Survey,” *Journal of Economic Literature*, 28(4), 1661–1707.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001) “The NBER patent citation data file: Lessons, insights and methodological tools,” Discussion paper, National Bureau of Economic Research.

Harris, C. G., R. Arens, and P. Srinivasan (2010) “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.*

Katz, L. F., and D. Autor (1999) "Changes in the wage structure and earnings inequality," in *Handbook of labor economics*, vol. 3, pp. 1463–1555. Elsevier.

Katz, L. F., and K. M. Murphy (1992) "Changes in relative wages, 1963–1987: supply and demand factors," *The quarterly journal of economics*, 107(1), 35–78.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2018) "Measuring Technological Innovation over the Long Run," Discussion paper, National Bureau of Economic Research.

Klepper, S., and K. Simons (2000) "The Making of an Oligopoly: Firm Survival and Technological Change in the Evolution of the U.S. Tire Industry," *Journal of Political Economy*, 108(4), 728–760.

——— (2005a) "Industry shakeouts and Technological Change," *International Journal of Industrial Organization*, 23(1–2), 23–43.

Klepper, S., and K. L. Simons (2005b) "Industry shakeouts and technological change," *International Journal of Industrial Organization*, 23(1-2), 23–43.

Kortum, S., and J. Lerner (1998) "Stronger protection or technological revolution: what is behind the recent surge in patenting?," in *Carnegie-Rochester Conference Series on Public Policy*, vol. 48, pp. 247–304. Elsevier.

Krafft, J., F. Quatraro, and P. P. Saviotti (2011) "The knowledge-base evolution in biotechnology: a social network analysis," *Economics of Innovation and New Technology*, 20(5), 445–475.

Lafond, F., and D. Kim (2017) "Long-run dynamics of the US patent classification system," *Journal of Evolutionary Economics*, pp. 1–34.

Lybbert, T. J., and N. J. Zolas (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.

Malerba, F., L. Orsenigo, and P. Peretto (1997) "Persistence of innovative activities, sectoral patterns of innovation and international technological specialization," *International Journal of Industrial Organization*, 15(6), 801–826.

Putnam, J., and R. E. Evenson (1994) "Inter-sectoral technology flows: Estimates from a patent concordance with an application to Italy," *Mimeograph, Yale University, New Haven, CT.*

Romer, P. M. (1990) "Endogenous technological change," *Journal of political Economy*, 98(5, Part 2), S71–S102.

Scherer, F. M. (1984) "Inter-Industry Technology Flows and Productivity Growth," *The Review of Economics and Statistics*, 64(4), 627–634.

Schmoch, U., F. Laville, P. Patel, and R. Frietsch (2003) "Linking technology areas to industrial sectors," *Final Report to the European Commission, DG Research*, 1(0), 100.

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

Strumsky, D., and J. Lobo (2015) "Identifying the sources of technological novelty in the process of invention," *Research Policy*, 44(8), 1445–1461.

Strumsky, D., J. Lobo, and S. Van der Leeuw (2012) "Using patent technology codes to study technolocial change," *Economics of Innovation and New Technology*, 21(3), 267–286.

Tripsas, M. (1997) "Surviving Radical Technological Change through Dynamic Capability: Evidence from the Typesetter Industry," *Industrial & Corporate Change*, 6(2).

Van Looy, B., C. Vereyen, and U. Schmoch (2014) “Patent Statistics: Concordance ipc V8-nace, 2. Eurostat,” .

Youn, H., D. Strumsky, L. Bettencourt, and J. Lobo (2015) “Invention as a combinatorial process: Evidence from U.S. patents,” *Journal of the Royal Society Interface*, 12(106).

Tables

Table 1: Probability Matrix of Rank Transitions between 1980 and 2010, Yearly Weights

		2010 Ranking									
											Not
		1	2	3	4	5	6	7	8	9	10+	Ranked
	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
1980	5	4.53	4.18	5.23	5.57	6.27	3.48	3.83	1.39	2.44	1.05	62.02
Ranking	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: ALP USPC and CPC concordances, author's calculations.

Notes: Observations are 6-digit NAICS industry and 4-digit CPC technology codes where we observe a non-missing ALP probability weight in either 1980 and 2010. Includes granted patents with an application date between 1980 and 2010.

Table 2: Autoregressive Model of Yearly Calculated Weights, Varying Lag Length

	(1) Weight at t, $\delta = 5$	(2) Weight at t, $\delta = 10$	(3) Weight at t, $\delta = 15$	(4) Weight at t, $\delta = 30$
5-Year Window				
Weight at $t - \delta$	0.918***	0.905***	0.900***	0.865***
	(0.00835)	(0.00992)	(0.0119)	(0.0223)
Observations	2.81M	2.27M	1.73M	108K
R-squared	0.784	0.703	0.634	0.487
Cumulative				
Weight at $t - \delta$	0.952***	0.915***	0.880***	0.795***
	(0.00352)	(0.00629)	(0.00876)	(0.0140)
Observations	4.94M	4.21M	3.48M	1.31M
R-squared	0.843	0.735	0.643	0.424

Source: ALP USPC and CPC concordances, author's calculations.
Notes: Clustered robust standard errors clustered by industry in parentheses. δ signifies the number of years prior to time t of the independent variable. The 5-year window tabulates the probability weights of all patents applied for and granted between time periods $t - 2$ and $t + 2$, centered around time t. *, **, and *** denote significance at 5, 1 and 0.1% confidence level.

Table 3: Impact of Technology Composition Changes on Employment Composition, 2007-2011

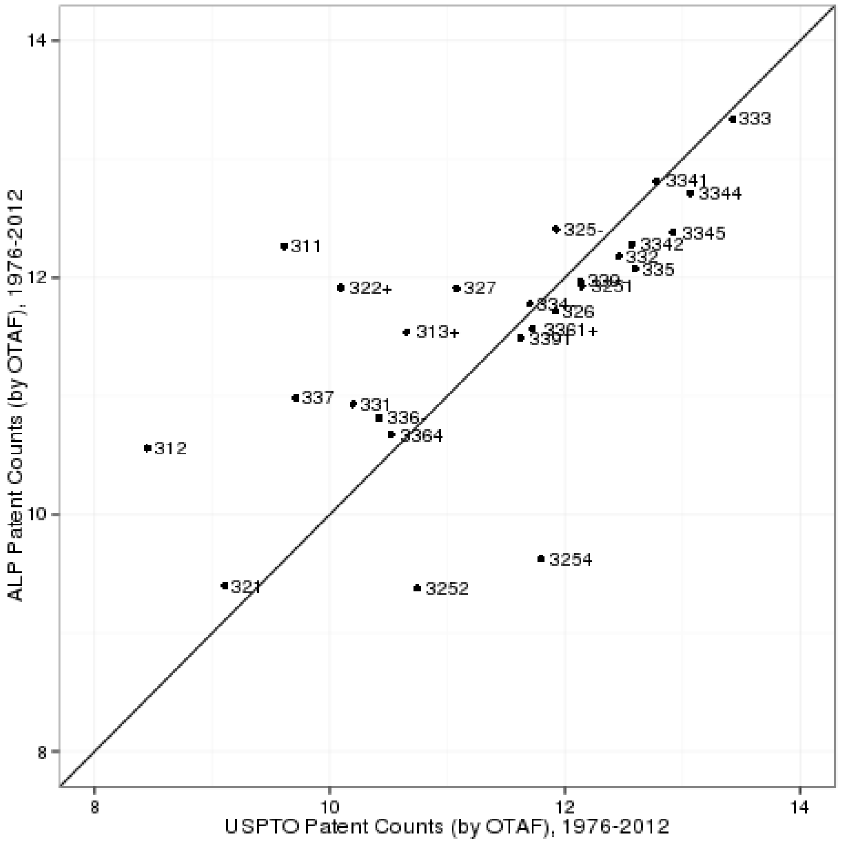
	(1) $\Delta Occ_{i,t}$	(2) $\Delta Occ_{i,t}$	(3) $\Delta Occ_{i,t}$	(4) $\Delta Occ_{i,t}$
$\Delta Tech_{I,t}$	0.127***	0.362***	0.0507*	0.0425*
	(0.0229)	(0.0545)	(0.0244)	(0.0200)
Industry FE	No	No	Yes	Yes
Yearly FE	No	No	No	Yes
Observations	425	425	425	425
Weighted	No	Yes	Yes	Yes
R-squared	0.315	0.360	0.945	0.958
	$R(\Delta Occ_{i,t})$	$R(\Delta Occ_{i,t})$	$R(\Delta Occ_{i,t})$	$R(\Delta Occ_{i,t})$
$R(\Delta Tech_{I,t})$	0.734***	0.745***	0.255	0.281*
	(0.0583)	(0.128)	(0.156)	(0.126)
Industry FE	No	No	Yes	Yes
Yearly FE	No	No	No	Yes
Observations	425	425	425	425
Weighted	No	Yes	Yes	Yes
R-squared	0.549	0.669	0.927	0.935

Source: ALP USPC and CPC concordances, BLS OES, author's calculations.
Notes: Clustered robust standard errors in parentheses clustered by four-digit industry. Regressions measure the 5-year change in occupational composition ($\Delta Occ_{i,t}$) on the change in technological composition ($\Delta Tech_{i,t}$). $R(.)$ signifies the ranking of the change across industries. Each of the variables (both dependent and independent) are five-year averages centered around a given year. There are 85 four-digit manufacturing industries included over 5-years for 425 observations. Industry fixed effects are at the four-digit level. *, **, and *** denote significance at 5, 1 and 0.1% confidence level.

Source: Suite of ALP CPC concordances, author's calculations.

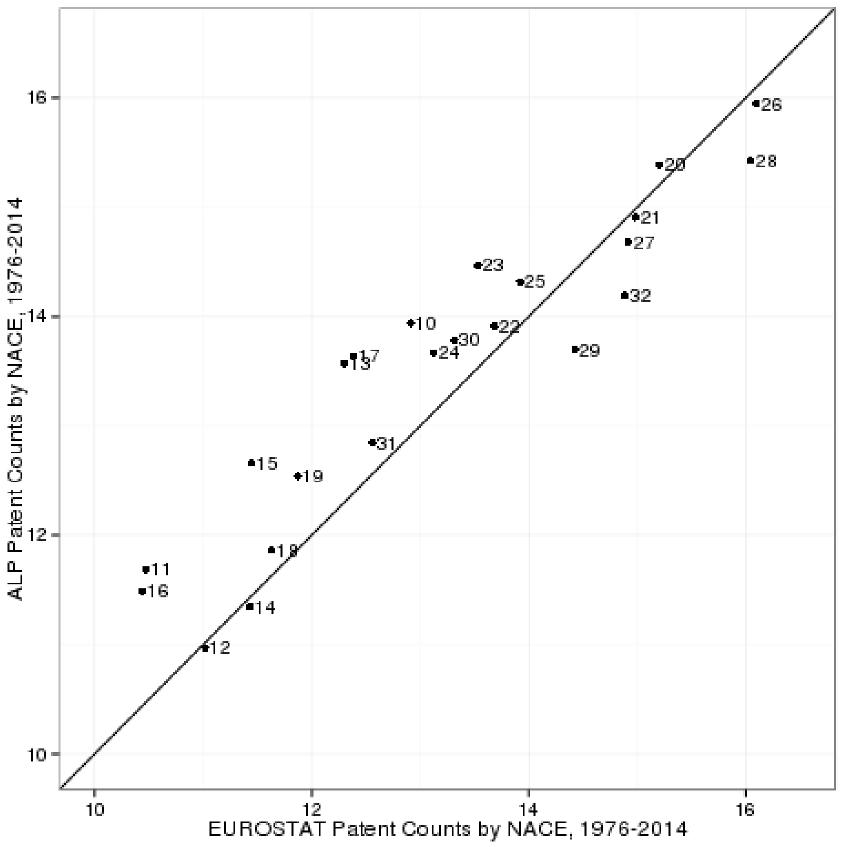
Figures

Figure 1: 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, authors 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.

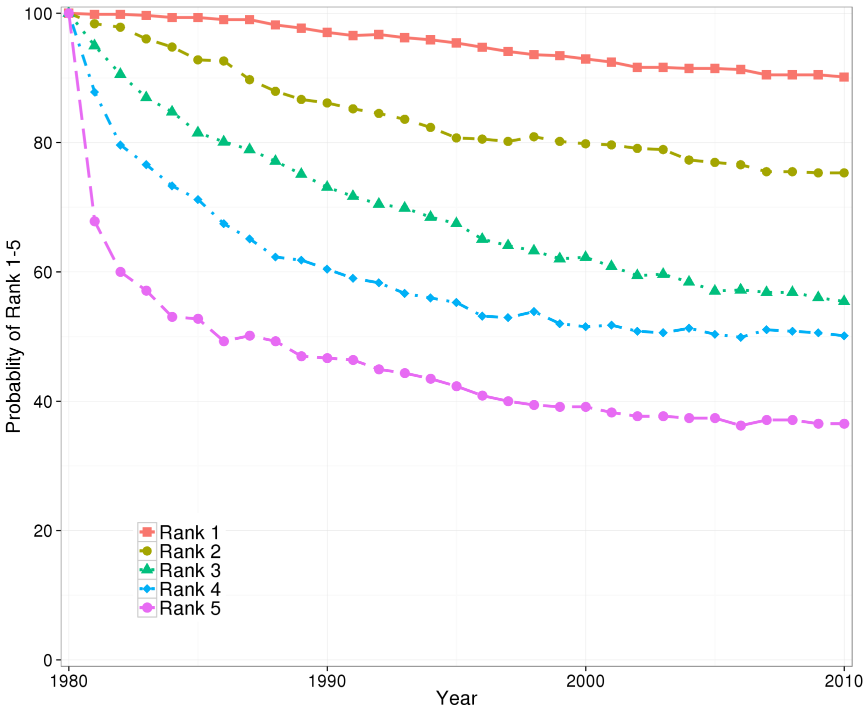
Figure 2: Comparing Patent Counts by NACE using ALP and EUROSTAT, 1976-2012



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.

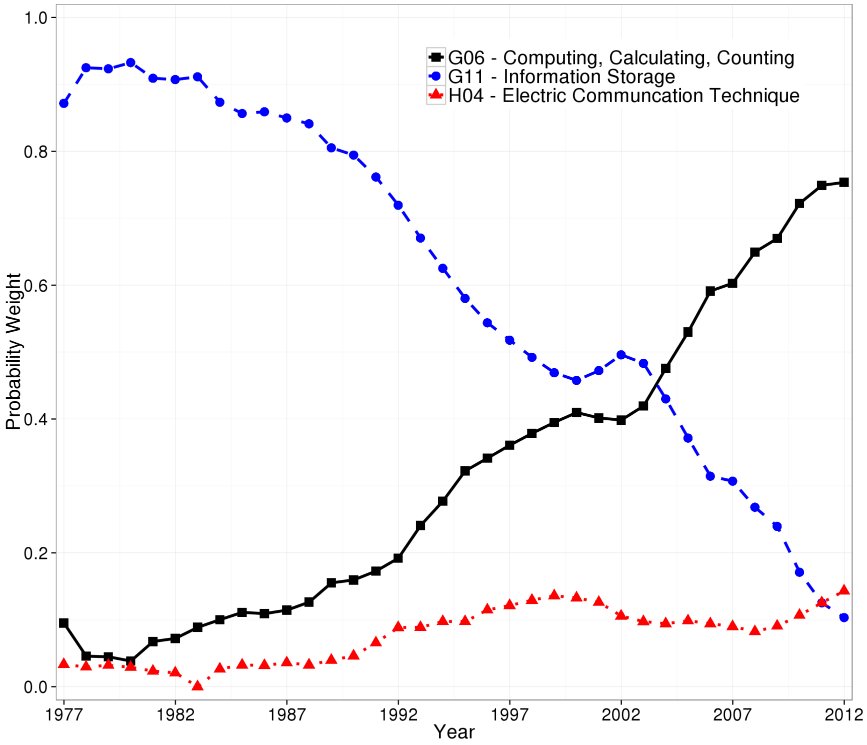
Figure 3: Cumulative Probability Rank of 1980 Rank



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.

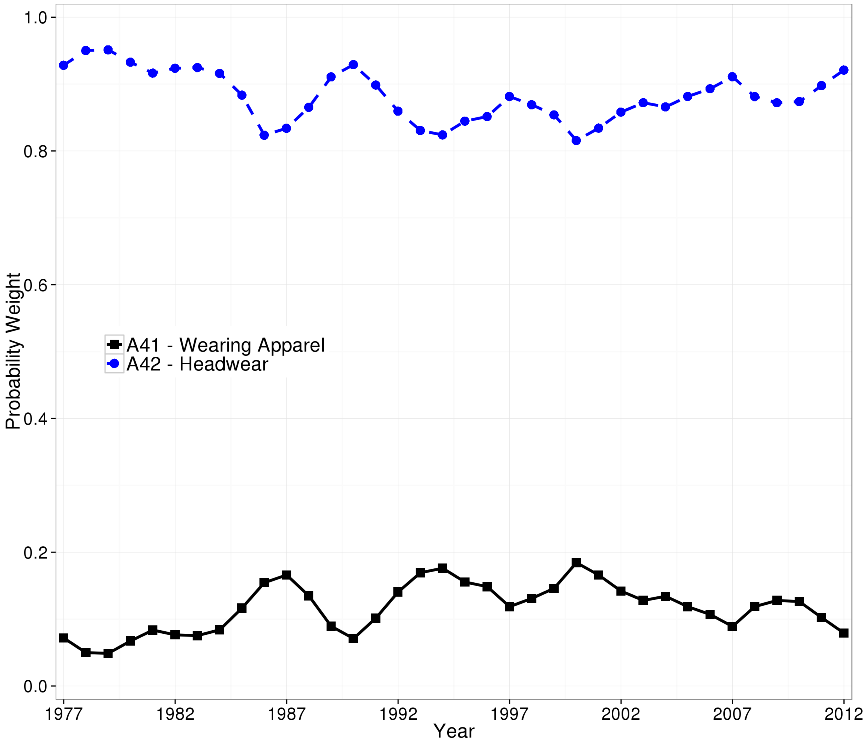
Figure 4: Technology Evolution for Magnetic and Optical Media Manufacturing (NAICS 3346)



Source: Calculate 5-year ALP concordances.

Notes: Probability weights calculated pooling for year t patents applied for in $t-2$ to $t+2$.

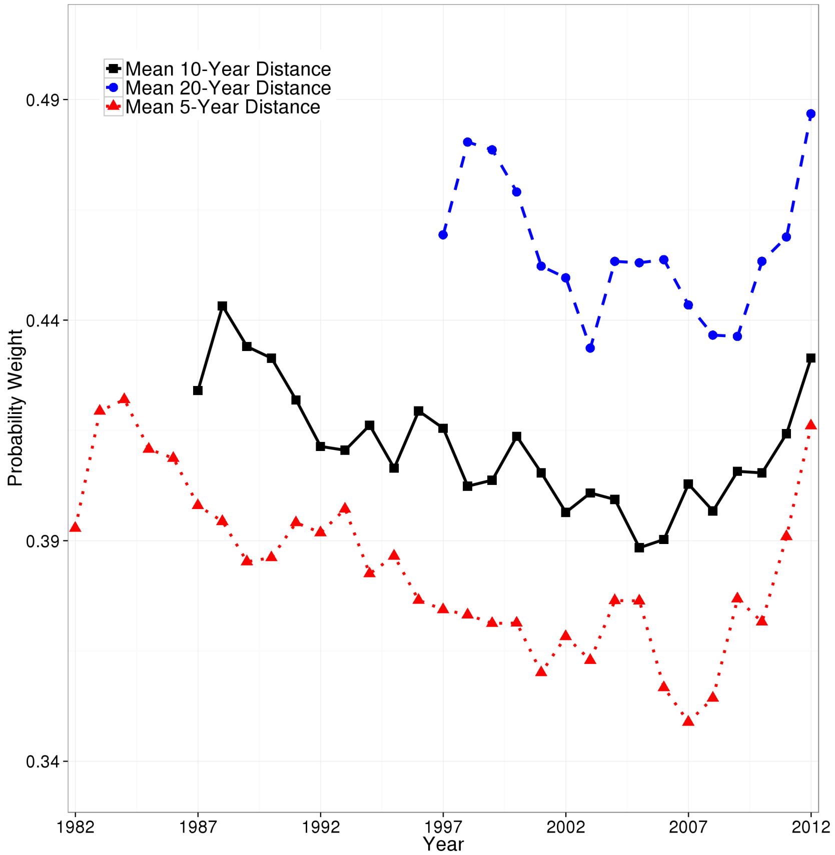
Figure 5: Technology Evolution for Apparel Accessories and Other Apparel Manufacturing (NAICS 3159)



Source: Calculate 5-year ALP concordances.

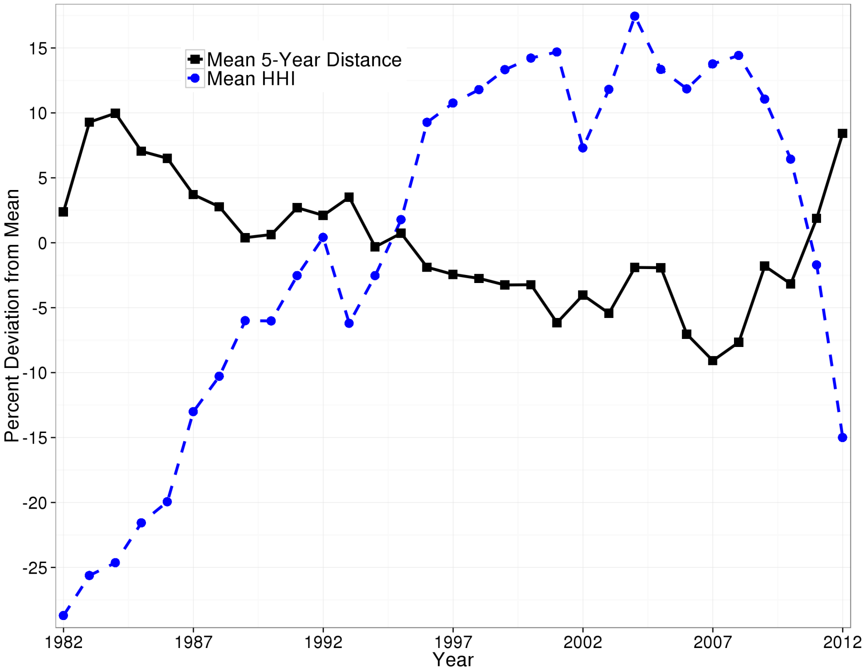
Notes: Probability weights calculated pooling for year t patents applied for in $t-2$ to $t+2$.

Figure 6: Mean Technology Composition Differences



Source: Calculate 5-year ALP concordances for manufacturing industries each year beginning in 1976 (current year, ±2 years) and take 5, 10 and 20-year differences. Concordances are at the 6-digit NAICS level.

Figure 7: Technology HHI and Composition Differences



Source: Calculate 5-year ALP concordances for manufacturing industries each year beginning in 1976 (current year, ±2 years) and take 5-year differences for the technology composition difference.
For the HHI, we first calculate the Herfindahl-Hirschman Index (HHI) based on the technology weights for each industry-year. We then take the mean HHI for every year across all manufacturing industries.

Appendices

A Construction of ALP Weights

The methodology for developing the ALP weights comes from Lybbert and Zolas (2014). This updated version includes a new set of industry classifications (the Harmonized System (HS)) and a slightly modified set of search terms for all industries. Outside of these updates, the methodology is almost identical where we query a patent database and reweight the matches to minimize errors. One alteration to the methodology now includes formerly missing industries and technologies that were previously dropped for not meeting the cutoff condition (see below).

A.1 ALP Steps

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 sentence or paragraph that describes the set of products, services and/or activities 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 category. The algorithmic methods include the keyword extraction algorithm, Topia Term Extract, which determines the keywords using a simple Parts-Of-Speech (POS) tagging algorithm.[21] 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 the US between 1976 and 2014 found in the PatentsView database for the USPC crosswalk and over 40 million patents applied for worldwide PATSTAT database for the CPC crosswalk.[22] These data provide both USPC and CPC codes associated with each patent granted between 1976 and 2014.[23] We select for each classification all patents

[21] A full description of the program can be found here: https://pypi.python.org/pypi/topia.termextract/ (accessed 2/2/2016).

[22] See http://www.patentsview.org (accessed 2/2/2016) for more details about the PatentsView database. 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.

[23] For PATSTAT patents, we utilize the first application date as the date for the patent and only included

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.[24] We then reweight the results using a modified Bayesian weighting scheme (Lybbert and Zolas, 2014), dubbed the "hybrid" weighting approach. 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 a combination of the raw and specificity weights to counter balance keywords that are generalizable across a number of patents (indicating that the industry/technology shows up very frequently when it should not) versus keywords that are so precise that when they do show up, we can be relatively certain the match is of high quality (Lybbert and Zolas, 2014). The hybrid approach considers both by increasing the weights of the specific matches and reducing the weights of the generalizable matches. The formula for this weighting scheme between industry i (an industry code) and technology j (either a USPC or CPC code) is:

$$W_{ij}^{H} = Pr(A_j|B_i) = \frac{Pr(B_i|A_j)(\frac{W_{ij}^R}{J})}{(\frac{W_{i1}^R}{J})Pr(B_i|A_1) + ... + (\frac{W_{iJ}^R}{J})Pr(B_i|A_J)} \tag{4}$$

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:

$$W_{ij}^{R} = Pr(A_j|B_i) = \frac{Pr(B_i|A_j)Pr(A_j)}{Pr(B_i|A_1)Pr(A_1) + ... + Pr(B_i|A_J)Pr(A_J)} \tag{5}$$

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 "balanced approach" does not completely discount widely applicable technologies, but instead tries to focus on the unique identifying technology of each industry.[25]. This equation provides a probabilistic match from industry-

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).

[24] The filter consistently removes between 20 and 25 percent of matches by industry group, reducing noise in our final weights.

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

to-technology and we perform this set of calculations for every combination of industry code i and technology code j (many will be zeros). We can likewise perform a probabilistic match from technology-to-industry by switching the placement of i and j.

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.

B ALP Match Results

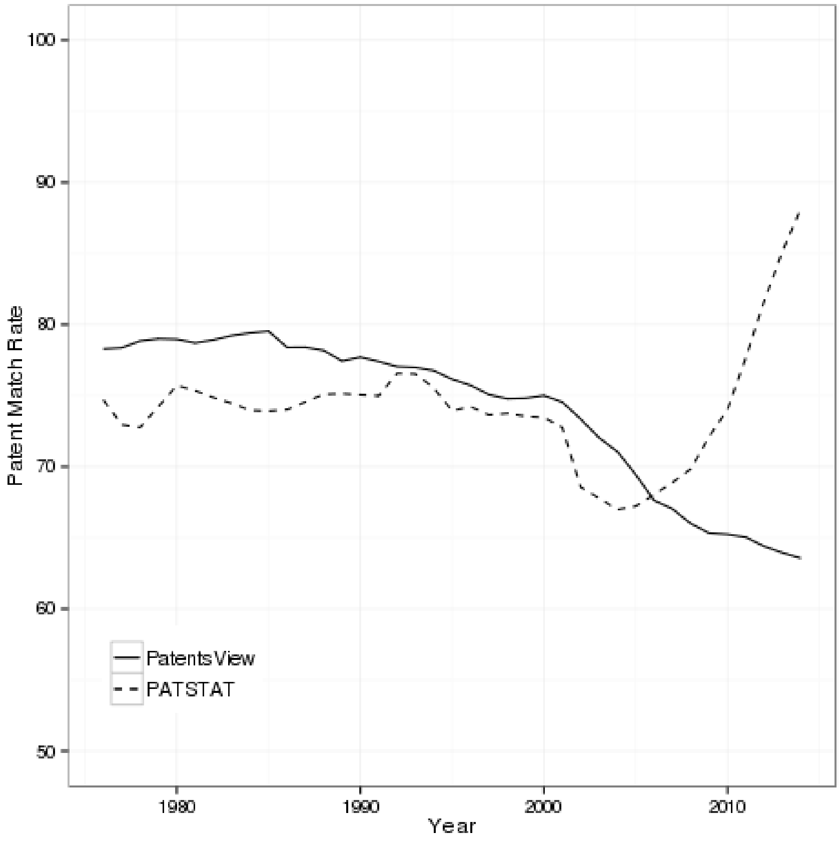
We run the full ALP methodology across both USPC and CPC codes using more than 5 million US Patents and 40 million global patents found in the PATSTAT database. The resulting suite of ALP crosswalks are available for download.[26]. In order to assess the validity of the crosswalk, we compare the patent counts generated by the USPC concordance with existing USPTO publications and with results from an IPC to industry concordance (IPC is the pre-cursor to the CPC and are very comparable). We also discuss the resulting CPC concordances and compare it to an existing EUROSTAT concordance that links IPC to NACE Rev. 2.

With over 155 thousand search terms and more than 45 million patent abstracts to search, we perform over 7 trillion comparisons. The overall number of patents from PatentsView (USPC) linked to at least one search term is 71%. For PATSTAT (CPC) we match 69% of all patents and 7% of granted patents with English language abstracts.[27]

[26] The latest versions can be found here: `https://sites.google.com/site/nikolaszolas/PatentCrosswalk`

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

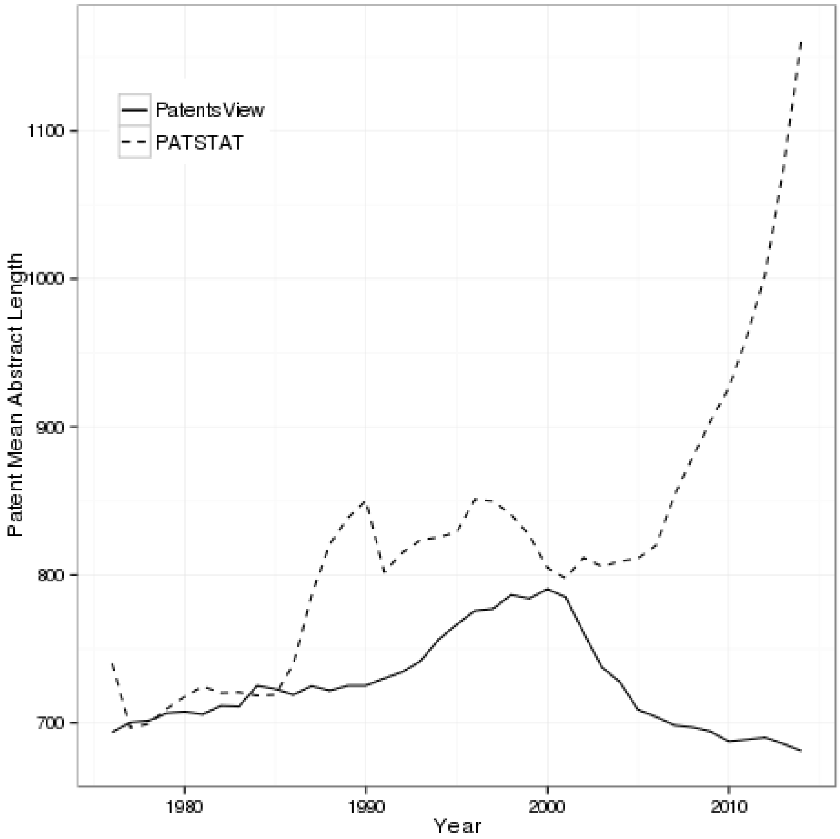
Figure 8: Utility Patent Match Rates by Year



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

As shown in Figure 8, 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 9 shows average annual abstract length of patents both from the PatentsView and PATSTAT databases.

Figure 9: 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.

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 steepening fall in match rates.

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-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 as the industry definitions typically undergo only minor modifications every 5-10 years, which may account for the declining match rates. Nevertheless, as these are all technologies who patent very frequently, we are confident that the set of patents that match with a keyword is mostly representative of the underlying technology composition, as denoted by the patent class.

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 assess 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. This concludes the matching exercise used in the construction of the USPC and CPC concordances.