NBER WORKING PAPER SERIES

MEASURING TECHNOLOGICAL INNOVATION OVER THE LONG RUN

Bryan Kelly
Dimitris Papanikolaou
Amit Seru
Matt Taddy

Working Paper 25266
http://www.nber.org/papers/w25266

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2018, Revised February 2020

We thank Pierre Azoulay, Nicholas Bloom, Diego Comin, Carola Frydman, Kyle Jensen, and seminar participants at AQR, Harvard, and the NBER Summer Institute for valuable comments and discussions. We are grateful to Kinbert Chou, Inyoung Choi, Jinpu Yang and Jiaheng Yu for excellent research assistance and to Enrico Berkes and Cagri Akkoyun for sharing their data. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w25266.ack

Measuring Technological Innovation over the Long Run
Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy
NBER Working Paper No. 25266
November 2018, Revised February 2020
JEL No. E22,E32,N1,O3,O4

ABSTRACT

We use textual analysis of high-dimensional data from patent documents to create new indicators of technological innovation. We identify significant patents based on textual similarity of a given patent to previous and subsequent work: these patents are distinct from previous work but are related to subsequent innovations. Our measure of patent significance is predictive of future citations and correlates strongly with measures of market value. We identify breakthrough innovations as the most significant patents – those in the right tail of our measure – to construct indices of technological change at the aggregate, sectoral, and firm level. Our technology indices span two centuries (1840-2010) and cover innovation by private and public firms, as well as non-profit organizations and the US government. These indices capture the evolution of technological waves over a long time span and are strong predictors of productivity at the aggregate and sectoral level.

Bryan Kelly
Yale School of Management
165 Whitney Ave.
New Haven, CT 06511
and NBER
bryan.kelly@yale.edu

Dimitris Papanikolaou
Kellogg School of Management
Northwestern University
2211 Campus Drive, Office 4319
Evanston, IL 60208
and NBER
d-papanikolaou@kellogg.northwestern.edu

Amit Seru
Stanford Graduate School of Business
Stanford University
655 Knight Way
and NBER
aseru@stanford.edu

Matt Taddy
Amazon
mataddy@gmail.com

Over the last two centuries, real output per capita in the United States has increased substantially more than the growth of inputs to production, such as the number of hours worked or the amount of capital used. Thus, much of economic growth is attributed to improvements in productivity—which however appears to have slowed down in the recent decades (Gordon, 2016). Similarly, there are significant differences in productivity across firms or establishments, which are rather persistent. Understanding the economic factors behind these differences in productivity across time and space has been at the forefront of the economic agenda (Syverson, 2011). Models of endogenous growth ascribe most of these movements to fluctuations in the rate of technological progress. However, both this link and the underlying economic forces are hard to pin down due to difficulty in measuring degree of technological progress over time. Our goal is to fill this gap by constructing indices of technological progress at the aggregate and sectoral level that are consistently available—and comparable—over long periods of time.

Patent statistics are a useful starting point. Though not all innovations are patented, patent statistics are by definition related to inventiveness.[1] A major obstacle in inferring the degree of technological progress from patent data is that patents vary greatly in their technical and economic significance. While measures such as citations a patent receives in the future have been used to address this obstacle, these metrics are not uniformly and consistently available over time, making it difficult to compare citation counts of patents across cohorts.[2] More recently, Kogan et al. (2017) propose a new measure of the private, economic value of new innovations that is based on stock market reactions to patent grants. However, their measure is only available for patents that are assigned to publicly traded firms after 1927. Hence, time-series fluctuations in indices derived from their measure could be affected by shifts in innovative activity between public firms and other entities—which include private firms, research institutions or government agencies.

We apply state-of-the-art techniques in textual analysis on the high-dimensional data from patent documents to construct indices of breakthrough innovations. Breakthrough innovations represent distinct improvements in the technological frontier and which become the new foundation upon which subsequent inventions are built. If citation data were objectively determined and consistently available, a breakthrough innovation would receive a large number

[1] Griliches (1998) writes on statistics that are based on patents: "they are available; they are by definition related to inventiveness, and they are based on what appears to be an objective and only slowly changing standard. No wonder that the idea that something interesting might be learned from such data tends to be rediscovered in each generation."

[2] Patent citations are only consistently recorded by the USPTO in patent documents after 1945. Prior to 1945, citations sometimes appear inside the text of the patent document, but they are much less common than in the post-war era. For instance, consider patent 388,116 issued to William Seward Burroughs on August 1888 for a 'calculating machine', one of the precursors to the modern computer. Burroughs' patent has just three citations as of March 2018. Similarly, patent 174,465 issued to Graham Bell for the telephone in February 1876 has the first recorded citation in 1956 (from patent 2,807,666). Until March 2018, it has received a total of 10 citations. These issues are not confined to the pre-1945 period: one of the first computer patents 2,668,661 issued in 1954 to George Stibitz at Bell Labs has just 15 citations as of March 2018.

of future citations. Given the absence of consistently available citation data, we instead propose a measure that is similar in spirit that can be constructed by analyzing the text of patent documents. We use advances in textual analysis to create links between each new invention and the set of existing and subsequent patents. Specifically, we construct measures of textual similarity to quantify commonality in the topical content of each pair of patents. We then identify significant (high quality) patents as those whose content is distinct from prior patents (is novel), but is similar to future patents (is impactful). Since our indicators of the significance of a patent require no other inputs besides the text of the patent document, they are consistently available for the entire history of US patents spanning nearly two centuries of innovation (1840–2010).

We validate our indicator of a significance of a patent along several dimensions. We first focus on the sample when citation data is available. We find that our indicator is significantly correlated with patent citations. More importantly however, we find that our text-based patent indicators are significant *predictors* of future citations—indicating that they provide a (much) more timely assessment of a patent's quality than citation counts. Within a few years of a patent's arrival, text-based similarity measures are able to reach an assessment of patent quality that predicts citation counts decades henceforth.

To examine how our quality indicator performs in evaluating older patents, we identify a set of major technological breakthroughs of the 19^{th} and 20^{th} century using the help of research assistants. Our indicators of patent significance perform substantially better than citation counts in identifying these major technological breakthroughs—especially when citations are measured over the same horizon as our indicator, but often even when they are measured using the entire sample. These breakthroughs include watershed inventions such as the telegraph, the elevator, the typewriter, the telephone, electric light, the airplane, frozen foods, television, plastics, computers and advances in modern genetics. This superior performance is not only driven by the fact that citations are sparsely recorded prior to 1945. Even in the more recent period, we find that our indicators often perform better than citations (over the same horizon) in identifying major technological breakthroughs—including for instance, recent advances in molecular biology and genetics.

As a further validation of our indicators we explore their relation to measures of private values. We emphasize that we view our indicators as more likely to be measuring the scientific value of a patent, given that it captures the extent to which novel contributions are adopted by subsequent technologies. That said, prior work has documented a strong correlation between patent citations (which form the inspiration for our measure) and measures of market value (e.g. Hall et al., 2005; Kogan et al., 2017).[3] Along these lines, we show that our quality indicator is

[3] The scientific and private value of a patent need not coincide. For instance, a patent may represent only a minor scientific advance, yet be very effective in restricting competition, and thus generate large private rents. That said, models of innovation with endogenous markups (Aghion and Howitt, 1992; Grossman and Helpman,

significantly correlated with the Kogan et al. (2017) measure of each patent's economic value. Our most conservative specification compares two patents that are granted to the same firm in the same year: in this case, a one standard deviation increase in our quality measure is associated with a 0.4 to 1.2 percentage point increase in patent value.

Armed with a consistent measure of the significance of a patent, we next set out to analyze long-run trends in innovation. We begin by identifying breakthrough innovations—patents that lie at the right tail of our measure. We construct time-series indices that describe the arrival intensity of breakthrough innovations, which requires us to compare patents of different cohorts in terms of quality. To ensure that the time-variation in our measure is not driven by changes in language—or measurement error due variances in the quality of the optical recognition algorithm applied to the text document—we remove calendar year-specific average from our measure. Our operating assumption is that such shifts in language (or measurement error) likely affect all patents symmetrically. We then construct indices of breakthrough innovation—at the aggregate, sectoral, and firm level—by counting the number of patents each year whose quality is in the top fifth percentile of our quality measure (net of year fixed effects). For comparison, we also construct corresponding indices using forward citation counts (net of year fixed effects), measured either over specific horizons or over the entire sample.

Our aggregate innovation index uncovers three major technological waves: the second Industrial Revolution (mid- to late 19^{th} century), the 1920s and 1930s, and the post–1980 period. Examining the technology areas where these breakthrough innovations occurred, we find that advances in electricity and transportation play a role in the 1880s; agriculture in the 1900s; chemicals and electricity in the 1920s and 1930s; and computers and communication in the post-1960s. Our innovation index is a strong predictor of aggregate total factor productivity during this period: a one-standard deviation increase in our index is associated with a 0.5 to 2 percentage points higher annual productivity growth over the next five to ten years.

We create sectoral indices of technological breakthroughs that span the entire sample by mapping technology areas to industries. Sectors that have breakthrough innovations experience faster growth in productivity than sectors that do not. In specifications that examine within-industry fluctuations in productivity (that is, net of industry and time effects), we find that a one-standard deviation increase in our innovation index is associated with 1 percentage point higher annual productivity growth over the next five years. In contrast to our text-based breakthrough index, the citations-based index is not statistically significantly related to industry productivity.

In sum, our paper provides a measure of technological innovation that is consistent across time and space. Our text-based indicator of patent quality are complementary to forward

1991) imply that the markup a technology leader can charge is related to the improvement in quality relative to the second-best alternative.

citations and have distinct advantages. First, it is consistently available for the entire 1840–2010 period, which allows us to construct indices of the level of technological change by comparing patents across cohorts. Second, it incorporates information faster than patent citations. Our indicator predicts future citations and, estimated over relatively short horizons post patent filing date (up to 5 years), it often shows a stronger correlation with real outcomes than citations measured over the same period.

Our work is connected to several strands of the literature. First, patent statistics offer a promising avenue in constructing indices of technological progress. Shea (1999) constructs direct measures of technology innovation using patents and R&D spending and finds a weak relationship between TFP and technology shocks. The results in Shea (1999) likely illustrate a shortcoming of simple patent counts, since they ignore the wide heterogeneity in the economic value of patents (Griliches, 1998; Kortum and Lerner, 1998). Furthermore, fluctuations in the number of patents granted are often the result of changes in patent regulation, or the quantity of resources available to the US patent office (see e.g. Griliches, 1990; Hall and Ziedonis, 2001). As a result, a larger number of patents does not necessarily imply greater technological innovation (for more details, see the discussion in Griliches, 1998). Alexopoulos (2011) proposes an alternative measure that is based on books published in the field of technology. Though the measure in Alexopoulos (2011) overcomes many of the shortcomings of patent counts, it is only available at the aggregate level and for only the later part of the 20^{th} century. By contrast, our measure is available at the individual patent level and is available since the 1840s.

Second, our analysis is related to work on patent valuation (see, e.g. Pakes, 1985; Austin, 1993; Hall et al., 2005; Nicholas, 2008; Kogan et al., 2017). The advantage of using financial data in inferring the (private) value of patents is that asset prices are forward-looking and hence provide us with an estimate of the private value to the patent holder that is based on ex-ante information. In particular, Pakes (1985) examines the relation between patents and the stock market rate of return in a sample of 120 firms during the 1968–1975 period. His estimates imply that, on average, an unexpected arrival of one patent is associated with an increase in the firm's market value of $810,000. Hall et al. (2005) finds that the current stock of patent citations carries information for firms' market valuations beyond that in past R&D expenditures and simple patent counts. Our results are similar; measures of intangibles constructed using our quality indicators contain information on firm values that is not captured by R&D, patent counts, or citation counts. Closest to our paper, Kogan et al. (2017) propose a new measure of the private, economic value of new innovations that is based on stock market reactions to patent grants. Kline et al. (2017) extrapolate their measure to a broader sample of patents to private firms. By construction, our indicators measure the scientific novelty and impact of the patent, which need not perfectly coincide with the private value of a patent.

Our paper is part of a recent but growing effort in applying advances in textual analysis to patent documents. Closest to our work is Balsmeier et al. (2018), who as part of a broader effort in disambiguating assignee and inventor names, also construct a patent-level measure of novelty starting in 1975. They define a novel patent as one that contains words that did not previously appear in the entire set of patent documents in their sample period. As a part of our definition of breakthrough patents over last two centuries, we also construct a measure of novelty. While the two measures are related, our construction of novel patent is somewhat different. We define a novel patent as one that is textually dis-similar from recent patents, defined as those within five years of the patents application date, where our similarity calculation overweighs uncommon words. As our analysis shows, breakthrough patents, which builds on our measure of novelty, strongly relate with metrics that might be associated with innovative activity.

Last, our paper makes a methodological contribution to estimating document similarity. Specifically, a key challenge in analyzing the textual similarity between documents is separating differences in writing style (language) from differences in content. Patent documents have the advantage that they largely contain scientific and legal terms, whose use has changed only slowly. However, given that our analysis spans almost two centuries of data, this is an important concern. We follow the literature on text analysis and construct measures of similarity that place more weight on important terms—that is, terms that are relatively uncommon across documents based on the *inverse document frequency* (IDF) (for a survey of existing methods, see e.g., Gentzkow et al., 2017). This static approach is ill-suited to our purposes; the process of innovation is often associated with the introduction of new scientific terminilogy. Hence, we introduce a dynamic modification to the existing approach that is crucial to our purposes. Specifically, we instead weigh terms according to the frequency in which they appear in patent documents *up until the patent document is filed.* As a result, the appropriate weight that terms receive in our similarity calculation evolves over time as scientific terms become more common or as natural language evolves.

I. Measuring the Significance of a Patent

In this section, we describe the construction of our metrics of patent significance. Throughout the paper, we will use the terms significant and high-quality patent interchangeably. We describe our data sources in Section A, then Section B describes our measure of similarity between patent documents. Section C contains the bulk of our analysis, which focuses on constructing a patent-level measure of quality that is based on textual similarity.

A. Data

We briefly overview our conversion of unstructured patent text data into a numerical format suitable for statistical analysis. To begin, we build our collection of patent documents from two sources. The first is the USPTO patent search website, which records all patents beginning from 1976. Our web crawler collected the text content of patents from this site, which includes patent numbers 3,930,271 through 9,113,586. The records in this sample are comparatively easy to process as they are available in HTML format with standardized fields.

For patents granted prior to 1976, we collect patent text from our second main datasource, Google's patent search engine. For the pre-1976 patent records, we recover all of the fields listed above with the exception of inventor/assignee addresses (Google only provides their names), examiner, and attorney. Some parts of our analysis rely on firm-level aggregation of patent assignments. We match patents to firms by firm name and patent assignee name. Our procedure broadly follows that of Kogan et al. (2017) with adaptations for our more extensive sample. In addition to the citation data we scrape from Google, we obtain complementary information on patent citations from Berkes (2016) and the USPTO. The data in Berkes (2016) includes citations that are listed inside the patent document and which are sometimes missed by Google. Nevertheless, the likelihood of a citation being recorded is significantly higher in the post-1945 than in the pre-1945. When this consideration is relevant, we examine results separately for the pre- and post-1945 periods.

To represent patent text as numerical data, we convert it into a *document term matrix* (DTM), denoted C. Columns of C correspond to words and rows correspond patents. Each element of C, denoted c_{pw}, counts the number of times a given one-word phrase (indexed by w) is used in a particular patent (indexed by p), after imposing a number of filters to remove stop words, punctuation, and so forth. We provide a detailed step-by-step account of our DTM construction in Appendix V. Our final dictionary includes 1,685,416 terms in the full sample of over nine million patents.

B. Measuring patent similarity

The basic building block for our patent-level quality measure using patent text is the textual similarity between pairs of patents. Here, we discuss the construction of our textual similarity measure in more detail.

1. Definition of patent similarity

A key consideration in devising a similarity metric for a pair of text documents is to appropriately weigh words by their importance. It is more informative if terms such as 'electricity' and 'petroleum' enter more prominently into the similarity calculation than common words like ‘process’ or ‘inventor.’ In textual analysis, a leading approach to overweighting terms that are most diagnostic of a document’s topical content is the “term-frequency-inverse-document-frequency” transformation of word counts:

$$TFIDF_{pw} \equiv TF_{pw} \times IDF_w. \tag{1}$$

The first component of the weight, term frequency (TF), is defined as

$$TF_{pw} \equiv \frac{c_{pw}}{\sum_k c_{pk}}, \tag{2}$$

and describes the relative importance of term w for patent p. It counts how many times term w appears in patent p adjusted for the patent’s length. The second component is the inverse document frequency (IDF) of term w, which is defined as

$$IDF_w \equiv \log\left(\frac{\#\text{ documents in sample}}{\#\text{ documents that include term } w}\right). \tag{3}$$

IDF measures the informativeness of term w by under-weighing common words that appear in many documents, as these are less diagnostic of the content of any individual document.

The product of these two terms, $TFIDF$, describes the importance of a given word or phrase w in a given document p. Words that appear infrequently in a document tend to have low $TFIDF$ scores (due to low TF), as do common words that appear in many documents (due to low IDF). A high value of $TFIDF_{pw}$ indicates that term w appears relatively frequently in document p but does not appear in most other documents, thus conveying that word w is especially representative of document p’s semantic content.

For our purposes, this traditional weighting scheme is not ideal because it ignores the temporal ordering of patents. In particular, we are interested in the novelty or impact of patent p’s text content given the history of innovation leading up to the development of p. Consider for example Nikola Tesla’s famous 1888 patent (number 381,968) of an AC motor, which was among the first patents to use the phrase “alternating current,” a phrase used with great frequency throughout the 20th century. Standard IDF would sharply de-emphasize this term in the $TFIDF$ vector representing Tesla’s patent because so many patents subsequently used this phrase so intensively. $TFIDF$ would therefore give a misleading, and quite inverted, portrayal of the patent’s innovativeness.

To overcome this issue, we devise and analyze a modified version of the traditional $TFIDF$ measure. In particular, in place of (3), we instead construct a retrospective, or ‘point-in-time’ version of inverse document frequency. Therefore, we define the “backward-IDF” of term w for filing year t, (denoted by $BIDF_{wt}$) as the log frequency of documents containing w in any patent filed *prior* to filing year t. More specifically, backward-IDF is defined as:

$$BIDF_{wt} = \log\left(\frac{\#\text{ patents prior to } t}{1 + \#\text{ documents prior to } t \text{ that include term } w}\right). \tag{4}$$

This retrospective document frequency measure evolves as a term becomes more or less widely used over time, giving a temporally appropriate weighting to a patent's usage of each term. It reflects the history of invention up to, but not beyond, the new patent's arrival.

Continuing with the Tesla example discussed above, consider measuring the similarity between Tesla's AC motor patent, and patent 4,998,526 assigned in 1990 to General Motors Corporation for an "Alternating current ignition system." An important question emerges: What is the most sensible IDF to use when calculating $TFIDF$ similarity of these two patents. One possibility is to use $BIDF$ for the year 1888 in the $TFIDF$ of Tesla's patent, and $BIDF$ as of 1990 for GM's patent. However, over the 102 years between these two patents, "alternating current" appears in tens of thousands of other patents. Thus, the use of "alternating current" by GM would be greatly down-weighted with a 1990 $BIDF$ adjustment, and thus the co-occurrence of "alternating current" in these two patents would have a small contribution to the pair's similarity.

One of the central goals of this paper is to quantify the impact of patents on future technological innovations. To best reflect quantify this impact, we instead calculate pairwise similarity by applying to *both* patent counts the $BIDF$ corresponding to the *earlier* of the two patents. Thus, to calculate the similarity between the patent pair in this Tesla/GM example, the term frequencies of both are normalized by the 1888 backward-IDF.

In sum, we construct the similarity between the patent pair (i, j) as follows. First, for both patents we construct our modified-version of the $TFIDF$ for each term w in patent i as

$$TFBIDF_{w,i,t} = TF_{w,i} \times BIDF_{w,t}, \quad t \equiv \min(\text{filing year for i}, \text{filing year for j}) \tag{5}$$

and likewise for patent j. These are arranged in a W-vector $TFBIDF_{i,t}$ where W is the size of the set union for terms in pair (i, j). Next, each $TFBIDF$ vector is normalized to have unit length,

$$V_{i,t} = \frac{TFBIDF_{i,t}}{||TFBIDF_{i,t}||}. \tag{6}$$

Finally, we calculate the cosine similarity between the two normalized vectors:

$$\rho_{i,j} = V_{i,t} \cdot V_{j,t}. \tag{7}$$

Our similarity measure is closely related to Pearson correlation, with the difference that $TFBIDF$ is not centered before the dot product is applied. Because $TFBIDF$ is nonnegative, $\rho_{i,j}$ lies in the interval [0,1]. Patents that use the exact same set of words in the same proportion will have similarity of one, while patents with no overlapping terms have similarity of zero.

Pairwise similarities constitute a high-dimensional matrix of approximate dimension 9 million × 9 million, which leads to over 800 terabytes of data. To reduce the computational burden when studying similarities, we set similarities below 5% to zero and get roughly 20 terabytes of data. This affects 93.4% of patent pairs. Patents with such low text similarity are, for all intents and purposes, completely unrelated, yet introduce a large computational load in the types of analyses we pursue. Replacing these approximate zeros with similarity scores of exactly zero achieves large computational gains by allowing us to work with sparse matrix representations that require substantially less memory.[4]

2. Descriptive statistics and validation of similarity

Panel A of Figure 1 plots the distribution of our similarity score across patent pairs, and focuses on pairs that are 0–20 years apart. The first observation is that the distribution of pairwise similarities is highly skewed. Patents tend to be highly dissimilar, with only a small fraction of pairs very closely related. The median similarity score across patent pairs is 7.8%, whereas the average similarity score is 10.2%. In the right tail, the 90^{th} and 95^{th} percentiles of similarity scores are 17.6% and 22.9%, respectively. In network terminology, the patent system's connectivity is sparse.

Citations provide a natural external measurement of patent linkages for assessing the text-based similarity measure $\rho_{i,j}$. To this end, we examine whether patent pairs with high $\rho_{i,j}$ are more likely to be linked by a citation. We bin patent pairs i-j in terms of their cosine similarity, and then compute the average propensity of a citation link—that is, we estimate $E\left[\mathbf{1}_{i,j}|\rho_{i,j}\right]$, where $\mathbf{1}_{i,j}$ is a dummy variable that takes the value one if patent j cites patent i (where patent i is filed prior to patent j). Panel B of Figure 1 plots the results. Indeed, patent pairs that are linked by a citation are more similar. The likelihood that patent j cites the earlier patent i is monotonically increasing in the similarity $\rho_{i,j}$ between the two patents. Our similarity score does not rely on any patent citation information, thus the results in Panel B are a powerful external validity check for our measure.

3. Patent similarity: examples

Figure 2 provides a few examples of patents' similarity network. To simplify the presentation, and also illustrate the advantages of our method in the early parts of the sample, we focus on

[4] Our empirical findings are insensitive to this threshold as they are driven primarily by the highest similarity pairs. In experiments with similarity cutoffs ranging from 1% to 10%, we find results that are quantitatively indistinguishable.

four patents from the 19th century. For each of these patents, the figure plots the set of prior and subsequent patents (filed within a period of five years) that have a cosine similarity of 50% or greater with the focal patent.

The patent at the top left part of the figure (US 4,750) is one of the first patents associated with the sewing machine, issued to 1846 to Elias Howe Jr. The patent is for the lockstitch, an efficient and sturdy stitch mechanism, which continues to be used today. The figure shows that this patent is not significantly connected to any prior patents. By contrast, it is relatively closely related to sixteen patents, all for improvements in the sewing machine, that were filed over the next five years. Many of these subsequent patents were owned by either Elias Howe, or three companies, Wheeler & Wilson, Grover and Baker, and I. M. Singer, who together formed the first patent pool in American industry in 1856 (Lampe and Moser, 2010).

The patent on the top right (US 493,426) is one of the earliest patents associated with cinematography. The patent is issued to Thomas Edison, for exhibiting 'photographs of moving objects', by Thomas Edison, and is essentially one of the first film projectors. The patent is highly similar to two prior patents and twelve subsequent patents, filed within five years apart. Most of the subsequent patents are related to cinematography–among them Among the subsequent patents, three are fo a 'kinetographic' camera, one of the early precursors of the film capera.

The patent at the bottom, left part of the figure (US 161,739) is one of the early patents issued to Graham Bell, for multiplexing intermittent signals on a single wire, that eventually led to the invention of the telephone. We can see that it is quite similar to four prior patents filed over the previous five years, all of which are related to the telegraph. It is also related to eleven patents filed over the next five years, one of which is Graham Bell's famous 'telephone' patent (174,465). Last, the patent on the bottom right is a random patent (US 222,189) for improvements in the cover of petroleum lamps. Within a five-year span, it is related to seven prior patents and five subsequent patents, all of which refer to improvements in lamps.

In brief, our examples show that our similarity measure identifies meaningful connections between patents. We next examine additional validation checks using an external measure of connection—patent citations.

C. Measuring Significant Patents

We aggregate a patent's pairwise similarity with other patents into a single indicator of significance of a patent—also referred to as the quality of a patent. Our main idea is that a significant patent is one that is both novel and impactful. Novel patents are those that are conceptually distinct from their predecessors, and therefore rely less on prior art. Impactful patents are those influence future scientific advances, manifested as high similarity with subsequent innovations.

1. Significant patents: definition

Our definition of patent significance combines both novelty and impact. As a novel patent is one that is distinct from prior art, we measure a patent's novelty as the (inverse of) its similarity with the existing patent stock at the time it was filed. We refer to this as "backward similarity," and define it as

$$BS_j^{\tau} = \sum_{i \in \mathcal{B}_{j,\tau}} \rho_{j,i}, \tag{8}$$

where $\rho_{i,j}$ is the pairwise similarity of patents i and j defined in equation (7) and $\mathcal{B}_{j,\tau}$ denotes the set of "prior" patents filed in the τ calendar years prior to j's filing. Patents with low backward similarity are dissimilar to the existing patent stock. They deviate from the state of the art and are therefore novel. We will consider a backward-looking window of $\tau = 5$ years in our baseline quality measure—-henceforth denoted by BS_j. That said, our results are insensitive to other window choices.

Next, we measure a patent's impact by its "forward similarity," defined as

$$FS_j^{\tau} = \sum_{i \in \mathcal{F}_{j,\tau}} \rho_{j,i}, \tag{9}$$

where $\mathcal{F}_{j,\tau}$ denotes the set of patents filed over the next τ calendar years following patent j's filing. The forward similarity measure in (9) estimates of the strength of association between the patent and future technological innovation over the next τ years.

A patent might have high forward similarity because it changes the course of future innovation. Or, it might be part of scientific regime shift that was catalyzed by a predecessor patent. The "alternating current" example highlights this difference. Nikola Tesla's patent has a high forward similarity because it dictated the course of future electronics, but was very different from any prior patents. The General Motors patent's similarity with future AC-related patents merely reflects that it is part of a mainstream technology—it has a high similarity both backward and forward. The distinction between these two patents emerges when we compare forward versus backward similarity for a given patent.

Thus, our indicator of patent significance combines forward and backward similarity to identify patents that are both novel and impactful in the following way:

$$q_j^{\tau} = \frac{FS_j^{\tau}}{BS_j}. \tag{10}$$

Our indicator (10) attaches higher scientific value to patents that are both novel relative to their predecessors and are influential for subsequent research. A patent may have high forward similarity because it is a “follower” in a technology area with many other followers, in which case it will have a high backward similarity as well. In normalizing by backward similarity, our quality measure adjusts for this. Highly significant patents—those with a large influence on future technologies and that deviate from the status quo—are more likely to represent scientific breakthroughs.

Our indicator of the significance of a patent largely follows the logic behind indicators based on future citations. Specifically, the numerator in (10) is the sum over similarity with future patents—which is directly analogous to the sum of future citations. The numerator in (10) scales the forward similarity score by the novelty of the patent—since, presumably, patents should be citing the earliest relevant prior patents that are related to the invention, that is, novel patents. However, given our interest in constructing time-series indices of innovation, one worry is that time-series fluctuations in (10) are also affected by mechanical factors, such as shifts in language; the fact that the retrospective document frequency measure (4) is changing over time so terms become less novel over time; and the fact that the number of patents is rapidly expanding over time. Given that these issues likely affect most patents symmetrically, when constructing time-series indices in Section III, we will adjust (10) by removing time fixed effects.

2. Significant patents: descriptive statistics

Figure 3 compares the cross-sectional distribution of quality, and citations, and its evolution over time. We can immediately see that the vast majority of patents receive very few citations in the pre-1947 period. For instance, even patents in the 90-th or 95-th percentile receive almost no citations over the next 10 years. Even when we examine their total citations in the entire sample, patents in the 95-th percentile typically receive between 2 to 10 citations in the pre-1947 period—compared to 20 citations in the 1960s or 50 citations in the 1980s. Part of this shift in the distribution of citations is mechanical, since the USPTO only started officially recording citations after 1947. However, we see that shifts in the propensity for patents to cite earlier patents could have played a role.

II. Validation

Next, we conduct three validation checks for our quality measure. First, we identify a list of important patents and examine how they score in terms of our quality indicators. Second, we relate our quality measure to forward patent citations, a common measure of patent quality in the innovation literature. Last, we examine the correlation between our quality indicators and market values.

A. Historically important patents

Our first validation exercise examines how historically important patents score in terms of our quality indicator. We compile a list of approximately 250 historically important patents based on online lists of 'important patents', for instance, the USPTO's "Significant Historical Patents of the United States" list. Our list targets indisputable important and radical inventions of the last 200 years, beginning with the telegraph and internal combustion engine, and ending with stem cells, Google's Pagerank algorithm and gene transfer. The full list of patents and sources is provided in Appendix Table A.7.

For each of these radical inventions we report their rank in terms of our patent quality measure (10) and forward citations. We focus on horizons of 10 years after the filing date for measuring quality and citations. For each patent, we compute its percentile rank based on quality or citations; for instance, a value of 0.90 indicates that the patent is in the top 10%. In addition to computing percentile ranks using the unconditional distribution, we perform two adjustments with the aim of removing time-series variation in these indicators that is unrelated to technical change. First, we rank patents based on cohort (issue year) demeaned values of these indicators. Removing cohort fixed effects helps eliminate factors that affects patents symmetrically, such as shifts in language; variation in the quality of the digitized patent documents; or changes in citation patterns. Second, we compute ranks within cohort. Though this comparison is not very useful in constructing a time-series index of technological change, it clarifies the extent to which these indicators are useful for purely cross-sectional comparisons.

Figure 4 summarize our findings. Panel A shows that, in terms of unconditional comparisons, our similarity-based quality indicator significantly outperforms citations: the average rank assigned to these important patents is 0.74, compared to 0.37 for citations. In Panel B, we see that the difference shrinks when these indicators are demeaned using year-fixed effects, but is not fully eliminated—0.78 for quality versus 0.70 for citations. Comparing Panel B to Panel C shows that removing time fixed effects leads to similar results as comparing patents within cohorts.

Appendix Table A.2 performs additional comparisons between our quality indicators and citations for different measurement horizons. In sum, these historically important patents rank at least as high using our patent quality measure than citations, even when the latter are measured over the entire sample. A key driver of behind the out-performance of our text-based quality indicators is that the texts of the underlying patent document have been uniformly available throughout the entire sample. By contrast, patent citations have been consistently recorded in patent documents only after 1945. Given our goal of constructing indices of technological change entails comparisons across patent cohorts, our text-based indicators have a significant advantage—which we exploit in Section III.

B. Patent Significance and Citations

The existing literature on innovation mostly relies primarily on patents' citations to measure their impact. We next investigate the power of our text-based quality measure for explaining patent citations. In particular, we estimate the following specification at the patent level (indexed by j):

$$\log\left(1 + CITES_j^{0,\tau}\right) = \alpha + \beta \log q_j^\tau + \gamma\, \mathbf{Z}_j + \varepsilon_j. \tag{11}$$

For this regression, we restrict attention to the sample of patents issued after 1945, as this is the period for which citations are recorded consistently by the USPTO. We measure patent quality and citations over the τ years since patent filing. The vector $\mathbf{Z}_j$ includes dummies controlling for technology class (defined at the 3-digit CPC level), grant year, assignee and the interaction of assignee and year effects. Including assignee fixed effects reduces the number of observations since many patents have no assignees. Nevertheless, in our most conservative specification we compare patents in the same technology class that are granted to the same assignee in the same year. Lastly, we cluster the standard errors by patent grant year.

Panel A of Figure 5 shows binned scatter plots of citations versus our text-based quality measure and reveal a strong positive correlation between the two. We collect observations into 50 bins (cutoff at every other percentile of the quality distribution). Within each bin, we average citation and text-based quality measures after controlling for technology class and assignee-by-grant year fixed effects, and consider contemporaneous forward windows of τ =1, 5, and 10 years for both citations and text similarity. Table 1 reports corresponding regression estimates. The contemporaneous explanatory power of our patent quality for citations is consistent across horizons τ and choice of controls Z. Importantly, the magnitude of these correlations is substantial. Focusing on our most conservative specification, which compares two patents filed in the same year, are in the same class, and are issued to the same entity in the same year, we find that increasing the quality measure from the median to the 90^{th} percentile results in 0.7 (1.5) additional citations, relative to the median of 2 (3) citations, when quality and citations are measured over the next 5 (10) years after the patent application is filed.

In short, our text-based measure of patent quality is highly correlated with patent citations over the same measurement horizon. Perhaps more interestingly, text-based quality measure is predictive of future citations. The left-most figure in Figure 5, Panel B plots the predictive relation between our text-based quality measured in the 0-1 year window after filing, versus all citations in years 2 and beyond. Likewise, we plot quality over years 0-5 versus citations in years 6+, and quality over 0-10 versus citations in years 11+. In all cases, we find an unambiguously strong positive association between our near-term quality measure and long-term future citations.

Similarly, we estimate the same predictive relation via regression while controlling for the information in lagged citations:

$$\log\left(1 + CITES_j^{\tau+}\right) = \alpha + \beta \, \log q_j^{0,\tau} + c \, \log\left(1 + CITES_j^{0,\tau}\right) + \gamma \, \mathbf{Z}_j + \varepsilon_j. \qquad (12)$$

This specification uses patent quality from years 0 through τ to forecast citations in year $\tau + 1$ and beyond, controlling for citations in the 0 to τ window. As before, the control vector $\mathbf{Z}$ includes fixed effects for year, technology class, and assignee. Our main coefficient of interest is b, which captures the predictive relation between our impact measure and future citations. The results in Table 3 show that our impact measure predicts future citations after controlling for the number of citations over the same period for which text-based quality is measured. The relation is statistically as well as economically significant. Focusing on the most conservative specification that includes the full set of fixed effects, we see that an increase in the patent quality from the median to the 90^{th} percentile is associated with 20-25% more citations relative to the median. Similar results obtain when we expand the sample to include patents issued prior to 1945 (see Appendix Table A.3).

To explore their individual roles, we estimate a variant of equation (11) that decomposes our quality measure into the numerator (impact) and the denominator (novelty). Table 2 shows that patent impact—as measured by the patent's forward similarity—is positively and significantly related to the number of times the patent gets cited over the same period. Second, patents that are more novel, that is, they are more dissimilar to earlier patents, are also more likely to be cited more in the future. Interestingly, the estimated coefficients on the log backward and forward similarity are of similar magnitude—and opposite sign. These estimate support the one-to-one ratio between the forward and the backward similarity that we use in our baseline indicator of quality.

Our text-based measures are strongly related to the most commonly-used indicator of patent quality, forward citations. Yet our quality measure has important advantages over patent citations. First, unlike citations, text-based quality does not suffer from truncation bias. Citations, on the other hand, are limited to the latter portion of the patent sample.

Second, citations tend to take small, discrete values (the median patent has one citation in a 10-year forward window), while our quality measure is continuous. This property of citations makes it a noisy measure for inferring patent quality, and the issue is exacerbated over short horizons (the median citation count drops to zero with a five year post-filing window).

Third, our text-based measure has the advantage of not relying on the discretion of the inventor or the patent examiner in choosing which prior patents to cite, or whether they are aware of the existence of closely related patents. This could introduce biases and idiosyncratic variation in the nature of which patents are cited and by whom. As an example, patent 6,368,227 for "Method of swinging on a swing", issued to Steven Olson (aged 5) in April 2002, has 11 citations as of June 2018. It is cited, for example, by patent 8,420,782 for "Modular DNA-binding domains and methods of use"; patent 8,586,526 for "DNA-binding proteins and uses thereof"; and patent 8,697,853 for "TAL effector-mediated DNA modification". Many of these citations were added by the patent examiner.

Fourth, the results of Table 3 indicate that our quality measure incorporates information much more quickly than forward citations. To further illustrate this point, Figure 6 reports the rate at which text-based quality (and also patent citations) behave over the measurement horizon τ. Specifically, the figure plots the average patent quality $q^{0,t}$ over different measurement horizons ($t = 1, \ldots, 20$ years) as a fraction of quality measured over the next 20 years $q^{0,20}$. We perform the same exercise for forward citations. We see that the amount by which the total forward similarity $FS_{0,t}$ increases is strongly declining across horizons — that is, $q_{0,t}$ as a fraction of $q^{0,20}$ is concave in t. By contrast, over short horizons, forward citations $C_{0,t}$ are convex in t. We also see that, over short horizons (0–5 years), measured quality accounts for a higher fraction of the total than citations, which is consistent with the view that our quality measure incorporates information faster than forward citations.

C. Patent Significance and Market Values

In this section, we discuss the relation between patent quality and market valuations. Market values are by definition private values; they measure the present value of pecuniary benefits to the holder of the patent. By contrast, our quality measure is designed to ascertain the scientific importance of the patent. The relationship market value and scientific importance can be ambiguous. For instance, a patent may represent only a minor scientific advance while being very effective in restricting competition, thus generating large private rents. The relation between the private and the scientific value of innovation—as measured by patent citations—has been the subject of considerable debate in the literature.[5]

As an estimate of the market value of a patent we use the measure of Kogan et al. (2017)—henceforth KPSS. The KPSS measure, $\hat{V}_j$, infers the value of patent j (in dollars) from stock market reaction to the patent grant. KPSS interpret this measure as an ex-ante measure of the private value of the patent. To investigate how text-based patent quality associates with

[5] For instance, Hall et al. (2005) and Nicholas (2008) document that firms owning highly cited patents have higher stock market valuations. Harhoff et al. (1999) and Moser et al. (2011) provide estimates of a positive relation using smaller samples that contain estimates of economic value. By contrast, Abrams et al. (2013) use a proprietary dataset that includes estimates of patent values based on licensing fees and show that the relation between private values and patent citations is non-monotonic.

estimated private value, we estimate the regression

$$\log \hat{V}_j = \alpha + \beta \, \log q_j^\tau + \gamma \, \mathbf{Z}_j + \varepsilon_j. \tag{13}$$

As before, we saturate our specifications with controls $\mathbf{Z}_j$, including fixed effects for grant year, technology class, and, in this case, firm. The vector of control variables also includes characteristics of the public firm that generates the patent, including the firm's log market capitalization prior to the patent grant (as larger firms may produce more influential patents) and the firm's log idiosyncratic volatility (fast-growing firms have more volatile returns and may produce higher quality patents). Our most stringent specification also the interaction of firm and year effects to account for the possibility that unobservable firm effects may influence our results. We cluster standard errors by grant year to account for correlation in citations among patents granted in the same given year. If multiple patents are issued to the same firm in the same day, we collapse them to a single observation by averaging the dependent and independent variables across patents.[6]

We present the results in Table 4. Columns (1) to (3) show a strong, statistically significant relation between our text-based measure of impact and the KPSS measure of market value. Their association strengthens as we increase the horizon over which we measure quality from 1 to 10 years after the filing date. In column (4), we include as an additional control the number of forward citations the patent receives over the same horizon that quality is measured. Doing so has little effect our point estimates, supporting the conclusion that our quality measure incorporates information that patent citations fail to capture. In terms of magnitudes, our estimates imply that an increase in $\log q$ from the median to the 90-th percentile is associated with approximately 0.4–1.2% increase in market values. Though these estimates may appear relatively modest, they are comparable in magnitude to the relation between patent values and forward citations.

In sum, these results confirm our earlier findings that our patent quality indicators are systematically related to market values, even controlling for patent citations. In Appendix Table A.6, we provide additional evidence that the quality of firms' patent portfolios correlates with their market valuation ratios (Tobin's Q), following the analysis of Hall et al. (2005). Given that these estimates are based on data from the later part of the sample, when citation data are broadly available, these results reinforce the view that our text-based measure captures information about patent quality that is not fully incorporated in patent citations.

[6] The KPSS measure does not differentiate between two patents that are issued to the same firm on the same day—it effectively assigns an equal fraction of the total dollar reaction to multiple patents in a given day to each patent. Estimating (13) at the patent level thus effectively overweighs firms that file a large number of patents. That said, this choice does not materially affect our findings. Appendix Table A.5 shows that results are very similar when estimating (13) at the patent level.

III. Measuring Innovation Over the Long Run

So far, our analysis has focused on developing and validating our patent quality measure. In this section, we use our measure to create time-series indices of the intensity of technological progress at the firm, sector, and aggregate economy levels, and investigate how these indices associate with measured productivity growth.

A. Breakthrough Patents

Here, we construct indices of technological progress at firm, sector and aggregate level by identifying and tracking breakthrough patents defined by our quality measure. Our findings so far—particularly those in Section A—suggest that our quality measure is more useful than forward citations in comparing patents across cohorts and is available over a longer time period. In aggregating patent quality into time series indices, it is important to confront shifts in language (or in the quality of the scanned patent documents) that may introduce systematic errors and unduly influence the comparison of patents across cohorts. To address this concern, we adjust our quality measure removing patent cohort year fixed effects. The implicit assumption in doing so is that shifts in language are likely to symmetrically affect all patents and will thus be absorbed by the fixed effect.

After this adjustment, we define a 'breakthrough' patent as one that falls in the top 10% of the quality distribution (among all patents in all years). Our baseline results use quality with a 10-year forward window. We also compare against an alternative definition of breakthrough patents based on the 10% of patents with the most forward citations over the same horizon (and likewise adjusted for year fixed effects).

B. Aggregate Index of Technological Progress

From our definition of breakthrough patents, we construct a time series of technological improvements that spans the USPTO sample (1840–2010). Our index is defined as the number of breakthrough inventions granted in each year, divided by the the US population. Panel A of Figure 7 plots the resulting time-series of breakthroughs per capita. Our index displays considerable fluctuations at relatively low frequencies. It identifies three main innovation waves, lasting from 1870 to 1880; 1920 to 1935; and from 1985 to the present. These periods line up with the major waves of technological innovation in the U.S. The first peak corresponds to the beginning of the second industrial revolution, which saw technological advances such as the telephone and electric lighting. The second peak corresponds to advances in manufacturing, particularly in plastics and chemicals, consistent with the evidence of Field (2003). The latest wave of technological progress includes revolutions in computing, genetics, and telecommunication.

For comparison, Panel B plots the resulting time-series when our index methodology is instead constructed from forward citations (over the next five years after the patent is filed, line in black). We see that this series essentially identifies no innovation prior to 1940s. Only when citations are measured over the entire sample (blue line) does the index take non-zero values in the pre-WW2 period, but even then the levels dwarf the values of the index post-1980. Given that the importance of inventions in the 1850–1940 era are at least comparable to the those in the last two decades (see, e.g. Gordon, 2016), this pattern mostly reflects the limitations of forward citations as a measure of quality.

Constructing an innovation index has proven challenging in the past. In one approach, Shea (1999) constructs an index of total patent counts, scaled by population growth. This series is plotted in Panel C. Total patents per capita is essentially flat from 1870–1930, dips from 1930–1980, and displays significant spike post-1980. There are reasons to be skeptical that such an index indeed measures the degree of underlying progress, since it implicitly assumes that all patents are equally valuable. Kortum and Lerner (1998) show that there is wide heterogeneity in the economic value of patents. Furthermore, fluctuations in the number of patents granted are often the result of changes in patent regulation, or the quantity of resources available to the US patent office (see e.g. Griliches, 1990; Hall and Ziedonis, 2001). As a result, a larger number of patents does not necessarily imply greater technological innovation. One common adjustment to simple patent counts is to weigh patents by their forward citations. As we see in Panel B however, such an index is contaminated by the fact that citation propensities vary over time.

Kogan et al. (2017) construct a time-series index that is based on the estimated market values of patents that are granted. Their index is plotted in Panel D. Their index has the advantage that it provides a dollar estimate of the value of innovation output in a given year. However, it has several shortcomings. First, it is based on a measure that is confined to the universe of publicly traded firms. Consequently, it omits not only innovations by private firms, non-profit institutions and the government, but also innovation prior to 1927 since reliable information on stock prices is available only after this year. Further, a direct corollary is that its time-series behavior may be influenced by shifts in the fraction of firms in the economy that are public, or variations in the degree of market efficiency.

1. Breakdown across technology classes and specific examples

Panel A of Figure 8 plots the breakdown across technology class of these breakthrough patents. We see that the technology classes in which breakthrough inventions originated has varied quite a bit over the last 170 years. By contrast, we see that the composition of technology classes among all patents has remained relatively stable over time.

In the 1840–70 period, we see that the most important inventions took place in engineering and construction, consumer goods, and manufacturing. An example of an invention in construction that scores high in terms of our quality measure is the 'Bollman Bridge' (patent number 8,624), named after its creator Wendell Bollman, which was the first successful all-metal bridge design to be adopted and consistently used on a railroad. In terms of manufacturing processes, many of the important advances occur in textiles. Specifically, examples of the important patents include various versions of sewing and knitting machines (patent numbers 7,931; 7,296; 7,509; and 60,310). Many of the important patents in consumer goods are also related to new clothing items.

Starting around 1870, many more patents that score high in terms of our measure are related to electricity, with some of the most important patents (based on our measure) relating to the production of electric light (203,844; 210,380; 215,733; 210,213; 200,545; 218,167). Most importantly, the same period saw the invention of a revolutionary method of communication: the telephone. It is comforting that most of the patents associated with the telephone are among the breakthrough patents we identify.[7]

Another industry that accounted for a significant share of the most important patents during the 1860-1910 period is transportation. Many of the patents that fall in the top 5% in terms of our measure include improvements in railroads (e.g., patents 207,538; 218,693; 422,976; and 619,320), and in particular, their electrification (patents 178,216; 344,962; 403,969; 465,407). Most importantly, the turn of the century saw the invention of the airplane. In addition to the Wright brothers' original patent (821,393), several other airplane patents also score highly in terms of our quality indicator (1,107,231; 1,279,127; 1,307,133; 1,307,134). Our measure also identifies other patents related to air transportation based on air balloons that are similar to the Zeppelin (i.e., 678,114 and 864,672). Last, innovations in construction methods continue to play a role in the 1870-1910 period. Among the patents that score in the top 1% in terms of our quality indicator are those that are related to the use of concrete (618,956; 647,904; 764,302; 654,683; 747,652; and 672,176) as a material in the construction of buildings, roads and pavements.

In the first half of the 20th century, chemistry emerges as a new area responsible for important patents, many describing inventions of plastic compounds. Among our breakthrough inventions is the patent for bakelite (942,699), the world's first fully synthetic plastic. This innovation opened the floodgates to a torrent of now-familiar synthetic plastics, including the invention in the 1930's of plasticized polyvinyl chloride (PVC) by Waldo Semon (patents 1,929,453 and 2,188,396) and nylon by Wallace H. Carothers (patent 2,071,250), all of which

[7] Specifically, the following patents associated with the telephone rank in the top 5% in terms of our baseline quality measure among the patents granted in the same decade: 161,739; 174,465; 178,399; 186,787; 201,488; 213,090; 220,791; 228,507; 230,168; 238,833; 474,230; 203,016; 222,390. Source: https://en.wikipedia.org/wiki/Invention_of_the_telephone#Patents

are score highly according to our measure. Other important patents in chemistry continue through the 1950's in the form of drug patents, including Nystatin (2,797,183); improvements in the production of penicillin (2,442,141 and 2,443,989); Enovid, the first oral contraceptive (2,691,028); and Tetracyline, one of the most prescribed broad spectrum antibiotics (2,699,054).

Subsequent to the 1950's, a large fraction of the important patents identified by our measure are in the area of Instruments and Electronics, and are related to the arrival of the Information Age. One of the most important patents according to our measure is the invention of the first microchip by Robert Noyce in 1961 (patent 2,981,877). During the 1970s, firms such as IBM, Xerox, Honeywell, AT&T, and Sperry Rand are responsible for some of the major innovations in computing. Xerox, for example, is responsible for several high-scoring inventions such as patent 4,558,413 for a management system software; patent 4,899,136 for improvements in computer user interface; patent 4,437,122 for bitmap graphics; and patents 3,838,260 and 3,938,097 for improvements in the interface between computer memory and the processor. In the 1980s and 1990s, several important patents that pertain to computer networks emerge among the set of breakthrough patents—for instance, patents 4,800,488; 4,823,338; 4,827,411; 4,887,204; 5,249,290; 5,341,477; 5,544,322; and 5,586,260.

Improvements in genetics comprise a significant fraction of high quality patents in the 1980–2000 period. A few early examples that fall in the top 1% of the unconditional distribution according to our quality indicator are: patent 4,237,224 for recombinant DNA methods (that is, the process of forming DNA molecules by laboratory methods of genetic recombination, such as molecular cloning, to bring together genetic material from multiple sources); patents 4,683,202; 4,683,195, and 4,965,188 for the polymerase chain reaction (PCR) method for rapidly copying DNA segments with high fidelity and at low cost; patent 4,736,866 for genetically modified animals; and patent 4,889,818 for heat-stable DNA-replication enzymes.

IV. Innovation and Measured Productivity

We next relate our innovation indices to measured productivity.

A. Aggregate Productivity

We begin by focusing on aggregate productivity growth. For the post-war sample, we use the TFP measure constructed by Basu et al. (2006), which is available over the 1948-2018 period. For the earlier sample, we measure productivity using output per hour using the data collected by Kendrick (1961), which is available for the 1889 to 1957 period. Following Jorda (2005), we estimate the following specification,

$$\frac{1}{\tau}\left(x_{t+\tau}-x_{t}\right)=a_{0}+a_{\tau}\,\log\text{BreakthroughIndex}_{t}+\rho_{\tau}\,x_{t}+c_{\tau}\,\mathbf{Z}_{t}+u_{t+\tau}, \tag{14}$$

where x_t is log productivity, $\text{BreakthroughIndex}_t$ refers to our innovation index, and Z_t is a vector of controls that includes the log number of patents per capita and the level of productivity. We consider horizons of $\tau = 1 \ldots 10$ years and adjust the standard errors using the Newey-West procedure. All independent variables are normalized to unit standard deviation. To ensure that we are not capturing pre-existing trends, we also examine negative values of τ.

We plot the estimated coefficients in Figure 10. Panel A presents the results of estimating (14) for the post-war sample. Focusing on horizons of five to ten years, we see that a one-standard deviation increase in our technology index is associated with an increase in TFP of 0.5 percent per year—which is substantial given that the standard deviation in measured TFP growth over this period is 1.8%. Importantly, there is no statistically significant correlation between past changes in productivity and our innovation index. Panel B shows the results for the earlier sample. Again focusing on horizons of five to ten years, we see that a one-standard deviation in our innovation index is associated with an increase in labor productivity growth of approximately 1.5–2% per year—compared to an annual standard deviation of 5.2% for labor productivity growth.

B. Sector-level Analysis

We next construct indices of innovation at the sector level. One issue that arises is how to map patents to industries in a way that is independent of the presence of an explicit assignee. We do so by exploiting the mapping between patent technology classifications (CPC) and various industry classifications constructed by Goldschlag et al. (2016). Because this is a probabilistic mapping (there is no one-to-one correspondence between CPC and industry codes), we assign a fraction of each patent to industry codes based on the given probability weights associated with its (4-digit) CPC technology classification. Goldschlag et al. (2016) provide mappings to NAICS industry definitions, at different levels of granularity.[8]

We begin by constructing long time-series indices of innovation using the 3-digit NAICS classification. Figure 9 plots our industry indices. Our industry indices reveal that the origin of breakthrough patents has varied considerably over time, consistent with our prior results. Inventions related to electricity were important in the late 19^{th} and early 20^{th} century.

[8] Here, two caveats are in order. First, this mapping is based on post-1970 data, whereas our analysis spans the entire period since the 1840s. Hence, there might be measurement error in our index since we assign a fraction of patents to each of the industries that map to a CPC classification based the weights estimated from only part of the sample. Second, this mapping is primarily available for manufacturing industries–which are however the industries that patent most heavily.

Innovations in agriculture played an important role in the beginning of the 20^{th} century, while advances in genetically modified food have peaked in the last two decades. Chemical and petroleum-related innovations were particularly important in the 1920s and 1930s. Computers and electronic products have peaked since the early 1990s. We next examine whether our industry indices are related to measured productivity.

Panel A Figure 11 presents our results for the period from 1987 to the present. We use the estimates of productivity at the NAICS 4-digit level from the Bureau of Labor Statistics (BLS), which covers 86 manufacturing industries. For this period, we then estimate a panel analogue of equation (14),

$$\frac{1}{\tau}\left(x_{t+\tau}-x_t\right)=a_0+a_\tau\,\log \mathrm{BreakthroughIndex}_{i,t}+\rho_\tau\,x_{i,t}+c_\tau\,\mathbf{Z}_{i,t}+u_{i,t+\tau}, \tag{15}$$

where, as above, $x_{i,t}$ denotes (log) multi-factor productivity; $\mathrm{BreakthroughIndex}_{i,t}$ is our industry innovation index (count of breakthrough patents, scaled by population); and $Z_{i,t}$ is a vector of controls that includes time and industry fixed effects, the log total number of patents, scaled by population and the level of productivity. Standard errors are clustered by industry. Given the shorter time-dimension of the data, we consider horizons of $\tau = 1 \dots 5$ years. To ensure that we are not capturing pre-existing trends at the industry level, we also examine the relation between innovation and past productivity growth, that is, negative values of τ.

We find a strongly statistically positive relation between our innovation index and future productivity growth—while the relation with past productivity growth is insignificant. In terms of magnitudes, a one-standard deviation increase in our innovation index is associated with approximately 1–1.2% higher productivity growth per year, over the next 5 years.

Panel B performs a similar exercise for the earlier sample. We use the labor productivity data collected by Kendrick (1961), which covers 62 manufacturing industries for the years 1899, 1909, 1919, 1937, 1947, and 1954. Since the data is only available at discrete periods, we modify our approach as follows: for each period $(t, t+\tau)$, we regress the annualized difference in log labor productivity on the log of the accumulated level of innovation (number of breakthrough patents) in $t \pm 2$ years—controlling for period, industry dummies, the log number of patents during the same period, and the log level of productivity at t.[9]

Examining Panel B, we again see a strong and statistically significant relation between our industry innovation indices and measured productivity: a one standard deviation increase in our innovation index is associated with a 1.4% higher growth rate in measured productivity over the next period.

For comparison, Figure 12 performs the same exercise using a corresponding index based on

[9] To construct innovation indices for the Kendrick industries, which are defined at the SIC code level, we use the concordance between 1997 NAICS and 1987 SIC codes from the Census Bureau. If NAICS industries map into multiple SIC codes, we assign an equal fraction to each.

citations (measured over a 10 year horizon). Examining Panels A and B, we see that there is no statistically significant relation between the citations-based index and industry productivity in either sample period.

V. Conclusion

We use textual analysis of high-dimensional data from patent documents to create new indicators of patent quality. Our metric assigns higher quality to patents that are distinct from the existing stock of knowledge (are novel) and are related to subsequent patents (have impact). These estimates of novelty and similarity are constructed using a new methodology that builds on recent advances in textual analysis. Our measure of patent significance is predictive of future citations and correlates strongly with measures of market value.

We identify breakthrough innovations as the most significant patents—that is, patents in the right tail of our measure—to construct indices of technological change at the aggregate and sectoral. Our technology indices span two centuries (1840-2010) and cover innovation by private and public firms, as well as non-profit organizations and the US government. These indices capture the evolution of technological waves over a long time span and are strong predictors of productivity.

References

Abrams, D. S., U. Akcigit, and J. Popadak (2013). Patent value and citations: Creative destruction or strategic disruption? Working Paper 19647, National Bureau of Economic Research.

Aghion, P. and P. Howitt (1992, March). A Model of Growth through Creative Destruction. *Econometrica 60*(2), 323–51.

Alexopoulos, M. (2011). Read all about it!! What happens following a technology shock? *American Economic Review 101*(4), 1144–79.

Austin, D. H. (1993). An event-study approach to measuring innovative output: The case of biotechnology. *American Economic Review 83*(2), 253–58.

Balsmeier, B., M. Assaf, T. Chesebro, G. Fierro, K. Johnson, S. Johnson, G.-C. Li, S. Luck, D. O'Reagan, B. Yeh, G. Zang, and L. Fleming (2018). Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *Journal of Economics & Management Strategy 27*(3), 535–553.

Basu, S., J. G. Fernald, and M. S. Kimball (2006). Are technology improvements contractionary? *American Economic Review 96*(5), 1418–1448.

Berkes, E. (2016). Comprehensive universe of u.s. patents (cusp): Data and facts. Working paper, Northwestern University.

Fama, E. F. and K. R. French (1997). Industry costs of equity. *Journal of Financial Economics 43*(2), 153–193.

Field, A. J. (2003). The most technologically progressive decade of the century. *American Economic Review 93*(4), 1399–1413.

Gentzkow, M., B. T. Kelly, and M. Taddy (2017, March). Text as data. Working Paper 23276, National Bureau of Economic Research.

Goldschlag, N., T. J. Lybbert, and N. J. Zolas (2016). An 'algorithmic links with probabilities' crosswalk for uspc and cpc patent classifications with an application towards industrial technology composition. CES Discussion Paper 16-15, U.S. Census Bureau.

Gordon, R. (2016). *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War*. The Princeton Economic History of the Western World. Princeton University Press.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature 28*(4), 1661–1707.

Griliches, Z. (1998, January). *Patent Statistics as Economic Indicators: A Survey*, pp. 287–343. University of Chicago Press.

Grossman, G. M. and E. Helpman (1991). Quality ladders in the theory of growth. *Review of Economic Studies 58*(1), 43–61.

Hall, B. and R. Ziedonis (2001). The patent paradox revisited: An empirical study of patenting in the U.S. semiconductor industry, 1979-1995. *The RAND Journal of Economics 32*(1), 101–128.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2005). Market value and patent citations. *The RAND Journal of Economics 36*(1), pp. 16–38.

Harhoff, D., F. Narin, F. M. Scherer, and K. Vopel (1999). Citation frequency and the value of patented inventions. *The Review of Economics and Statistics 81*(3), 511–515.

Jorda, O. (2005, March). Estimation and inference of impulse responses by local projections. *American Economic Review 95*(1), 161–182.

Kendrick, J. W. (1961). *Productivity Trends in the United States.* National Bureau of Economic Research, Inc.

Kline, P., N. Petkova, H. Williams, and O. Zidar (2017). Who profits from patents? Rent sharing at innovative firms. Working paper.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017). Technological innovation, resource allocation, and growth*. *The Quarterly Journal of Economics 132*(2), 665–712.

Kortum, S. and J. Lerner (1998). Stronger protection or technological revolution: what is behind the recent surge in patenting? *Carnegie-Rochester Conference Series on Public Policy 48*(1), 247–304.

Lampe, R. and P. Moser (2010). Do patent pools encourage innovation? evidence from the nineteenth-century sewing machine industry. *The Journal of Economic History 70*(4), 898–920.

Moser, P., J. Ohmstedt, and P. Rhode (2011). Patents, citations, and inventive output - evidence from hybrid corn.

Nicholas, T. (2008). Does innovation cause stock market runups? Evidence from the great crash. *American Economic Review 98*(4), 1370–96.

Pakes, A. (1985). On patents, R&D, and the stock market rate of return. *Journal of Political Economy 93*(2), 390–409.

Shea, J. (1999). What do technology shocks do? In *NBER Macroeconomics Annual 1998, volume 13*, NBER Chapters, pp. 275–322. National Bureau of Economic Research, Inc.

Syverson, C. (2011). What determines productivity? *Journal of Economic Literature 49*(2), 326–65.

Tables and Figures

Table 1: Patent citations, impact and novelty, contemporaneous correlations

$\log(1 + \text{Forward citations, 0-1 yr})$	(1)	(2)	(3)	(4)
$\log(\text{Patent quality, 0-1yr})$	0.432***	0.255***	0.174*	0.127
	(6.09)	(4.17)	(2.64)	(1.73)
R^2	0.072	0.106	0.177	0.235
Observations	6,017,673	5,981,174	4,492,964	4,054,639
$\log(1 + \text{Forward citations, 0-5 yr})$	(1)	(2)	(3)	(4)
$\log(\text{Patent quality, 0-5yr})$	1.295***	0.962***	0.780***	0.752***
	(34.15)	(22.58)	(14.29)	(13.57)
R^2	0.195	0.243	0.330	0.375
Observations	4,964,003	4,930,423	3,535,656	3,169,209
$\log(1 + \text{Forward citations, 0-10 yr})$	(1)	(2)	(3)	(4)
$\log(1 + \text{Forward citations, 0-10 yr})$	1.273***	1.040***	0.885***	0.879***
	(46.97)	(61.98)	(33.32)	(30.91)
R^2	0.263	0.311	0.397	0.437
Observations	4,135,358	4,104,591	2,811,353	2,509,928
Grant Year FE	Y	Y	Y	
Tech Class FE		Y	Y	Y
Assignee FE			Y	
Grant Year × Assignee FE				Y

Table reports the results of estimating equation (11) in the main text. The regression relates the log of (one plus) the number of patent citations to our measures of patent impact (forward similarity) and lack of novelty (inverse of backward similarity) constructed in equations (9) and (8), respectively. As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm (assignee) and the interaction of firm and year effects. Since patent citations are only consistently recorded after 1947, we restrict the sample to the 1947–2016 period. As patents can be assigned to multiple assignees, observations are at the patent–assignee level. Last, we cluster the standard errors by the patent grant year. See main text for additional details on the specification and the construction of these variables.

Table 2: Patent citations, impact and novelty, contemporaneous correlations

log(1 + Forward citations, 0-1 yr)	(1)	(2)	(3)	(4)
log(Patent impact (FS), 0-1yr)	0.389***	0.247***	0.170**	0.127
	(6.02)	(4.26)	(2.78)	(1.88)
log(Patent novelty (1/BS), 0-5yr)	0.351***	0.221***	0.149*	-0.105
	(5.68)	(3.96)	(2.52)	(1.61)
R^2	0.076	0.107	0.178	0.235
Observations	6,017,673	5,981,174	4,492,964	4,054,639
log(1 + Forward citations, 0-5 yr)	**(1)**	**(2)**	**(3)**	**(4)**
log(Patent impact (FS), 0-5yr)	1.169***	0.917***	0.740***	0.708***
	(31.04)	(19.66)	(13.11)	(12.50)
log(Patent novelty (1/BS), 0-5yr)	1.075***	0.833***	0.671***	0.638***
	(30.47)	(18.37)	(12.03)	(11.41)
R^2	0.200	0.247	0.331	0.376
Observations	4,964,003	4,930,423	3,535,656	3,169,209
log(1 + Forward citations, 0-10 yr)	**(1)**	**(2)**	**(3)**	**(4)**
log(Patent impact (FS), 0-10yr)	1.183***	1.009***	0.853***	0.841***
	(52.15)	(50.70)	(29.05)	(27.06)
log(Patent novelty (1/BS), 0-5yr)	-1.092***	-0.910***	-0.769***	-0.757***
	(-46.07)	(-43.81)	(-25.77)	(-23.88)
R^2	0.267	0.315	0.399	0.438
Observations	4,135,358	4,104,591	2,811,353	2,509,928
Grant Year FE	Y	Y	Y	
Tech Class FE		Y	Y	Y
Assignee FE			Y	
Grant Year × Assignee FE				Y

This Table is the counterpart to Table 1, in which we disaggregate our measure of patent quality into patent impact (forward similarity) and of novelty (inverse of backward similarity) constructed in equations (9) and (8), respectively. Table reports the results of estimating equation (11) in the main text. The regression relates the log of (one plus) the number of patent citations to our measures of patent impact (forward similarity) and lack of novelty (inverse of backward similarity) constructed in equations (9) and (8), respectively. As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm (assignee) and the interaction of firm and year effects. Since patent citations are only consistently recorded after 1947, we restrict the sample to the 1947–2016 period. As patents can be assigned to multiple assignees, observations are at the patent–assignee level. Last, we cluster the standard errors by the patent grant year. See main text for additional details on the specification and the construction of these variables.

Table 3: Patent quality and citations: predictive relation

log(1 + Forward citations, 2+ yr)	(1)	(2)	(3)	(4)
log(Patent quality, 0-1yr)	1.201***	1.009***	0.940***	0.981***
	(15.82)	(17.18)	(15.97)	(15.55)
log(1 + Forward citations, 0-1 yr)	0.656***	0.604***	0.514***	0.506***
	(33.23)	(36.25)	(38.82)	(35.30)
R^2	0.314	0.368	0.479	0.517
Observations	6,017,673	5,981,174	4,492,964	4,054,639
log(1 + Forward citations, 6+ yr)	(1)	(2)	(3)	(4)
log(Patent quality, 0-5yr)	0.635***	0.724***	0.722***	0.782***
	(11.10)	(13.72)	(10.19)	(10.56)
log(1 + Forward citations, 0-5 yr)	0.614***	0.581***	0.540***	0.547***
	(36.70)	(36.66)	(39.30)	(37.56)
R^2	0.319	0.377	0.472	0.506
Observations	4,964,003	4,930,423	3,535,656	3,169,209
log(1 + Forward citations, 11+ yr)	(1)	(2)	(3)	(4)
log(Patent quality, 0-10yr)	0.186***	0.391***	0.401***	0.418***
	(4.41)	(14.42)	(10.44)	(9.67)
log(1 + Forward citations, 0-10 yr)	0.573***	0.539***	0.510***	0.512***
	(37.42)	(37.37)	(38.83)	(37.88)
R^2	0.300	0.363	0.448	0.482
Observations	4,135,358	4,104,591	2,811,353	2,509,928
Grant Year FE	Y	Y		
Class		Y		
Assignee FE			Y	
Grant Year × Assignee FE				Y

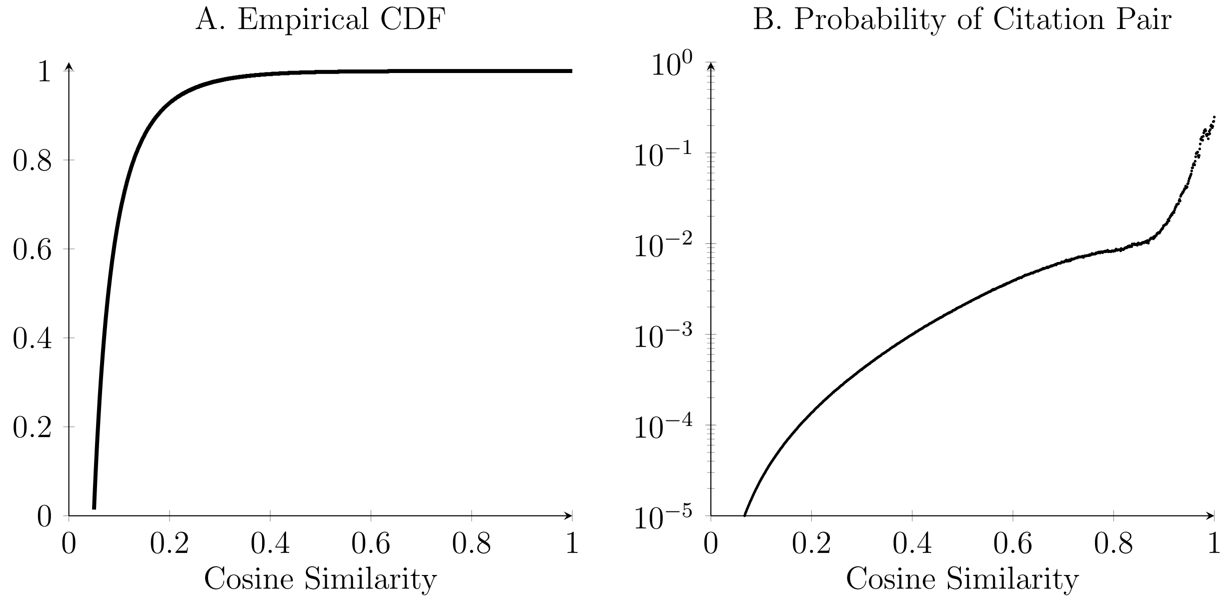
Table reports the results of estimating equation (12) in the main text. The regression relates the log of (one plus) the number of patent citations after time t to our measures of patent quality (10) measured over a horizon $[0, t]$ and citations measured over the same interval $[0, t]$. As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), assignee and issue year effects. Since patent citations are only consistently documented after 1947, we restrict the sample to the 1947–2016 period. Last, we cluster the standard errors by the patent grant year. See main text for additional details on the specification and the construction of these variables.

Table 4: Patent quality and value

log KPSS value	(1)	(2)	(3)	(4)
Log patent quality, 0-1 years	-0.0028	0.0020	0.0041***	0.0041***
	(-1.10)	(0.96)	(3.37)	(3.37)
Log forward citations, 0-1 years				-0.0002
				(-0.37)
R^2	0.947	0.956	0.965	0.965
Observations	559,669	558,329	539,309	539,309
log KPSS value	(1)	(2)	(3)	(4)
Log patent quality, 0-5 years	0.0035	0.0052***	0.0084***	0.0077***
	(1.24)	(2.91)	(5.03)	(4.59)
Log forward citations, 0-5 years				0.0044***
				(5.93)
R^2	0.951	0.959	0.967	0.967
Observations	496,844	495,541	478,049	478,049
log KPSS value	(1)	(2)	(3)	(4)
Log patent quality, 0-10 years	0.0112***	0.0091***	0.0120***	0.0100***
	(5.33)	(6.01)	(7.49)	(5.99)
Log forward citations, 0-10 years				0.0091***
				(9.29)
R^2	0.953	0.960	0.966	0.966
Observations	430,211	428,948	413,458	413,458
Controls:				
Grant Year FE	Y	Y		
Class FE	Y	Y	Y	Y
Firm Size (market cap)	Y	Y	Y	Y
Firm Volatility	Y	Y	Y	Y
Firm FE		Y	Y	Y
Grant Year × Firm FE			Y	Y

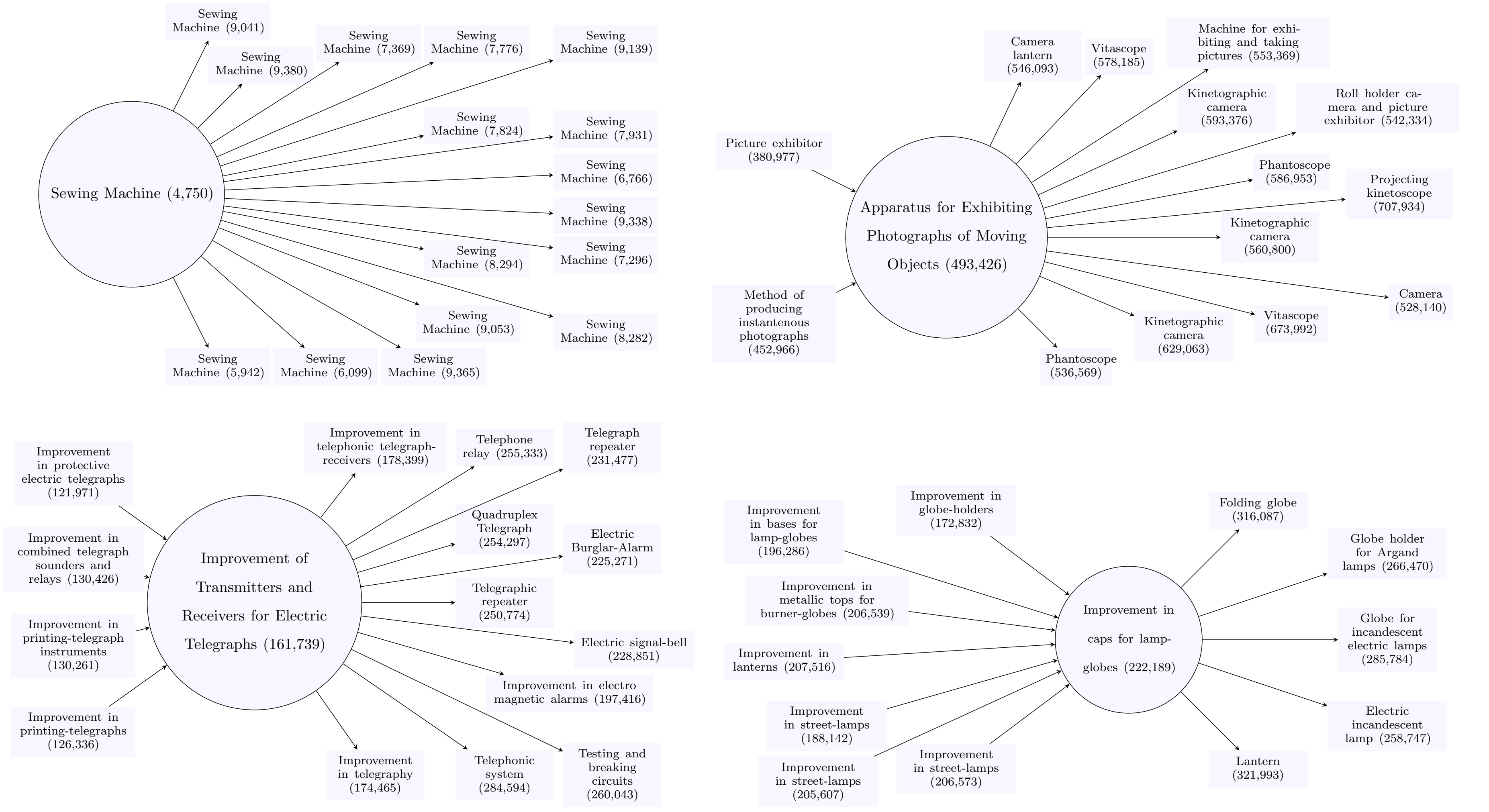
Table reports the results of estimating equation (13) in the main text. The regression relates the log of the Kogan et al. (2017) estimate of the market value of the patent to our (log) measures of patent quality, which combines the patent's impact and novelty, constructed in equation (10). As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm (CRSP: permco) and the interaction of firm and year effects. Since multiple patents can be issued to a given firm in a given day (which implies the same kpss value for these patents) we collapse the observations at the firm-date level. See Appendix Table A.5 for the corresponding regressions at the patent level. We cluster the standard errors by the patent grant year. All independent variables are normalized to unit standard deviation. See main text for additional details on the specification and the construction of these variables.

Figure 1: Pairwise similarity and citation linkages



Panel A plots the empirical CDF of our similarity measure $\rho_{i,j}$ across patent citation pairs. Panel B plots the conditional probability that patent j cites an earlier patent j as a function of the text-based similarity score between the two patents, $\rho_{i,j}$, computed in equation (7) in the main text. For computational reasons, we exclude similarity pairs with $\rho_{i,j} \leq 0.5\%$. Figure uses data only post 1945, since citations were not consistently recorded prior to that year. We use data only post 1945, since citations were not consistently recorded prior to that year.

Figure 2: Similarity Networks, Examples



Sewing Machine (4,750)
Sewing Machine (9,041)
Sewing Machine (9,380)
Sewing Machine (7,369)
Sewing Machine (7,776)
Sewing Machine (9,139)
Sewing Machine (7,824)
Sewing Machine (7,931)
Sewing Machine (6,766)
Sewing Machine (9,338)
Sewing Machine (8,294)
Sewing Machine (7,296)
Sewing Machine (9,053)
Sewing Machine (8,282)
Sewing Machine (5,942)
Sewing Machine (6,099)
Sewing Machine (9,365)
Apparatus for Exhibiting Photographs of Moving Objects (493,426)
Picture exhibitor (380,977)
Method of producing instantenous photographs (452,966)
Camera lantern (546,093)
Vitascope (578,185)
Machine for exhibiting and taking pictures (553,369)
Kinetographic camera (593,376)
Roll holder camera and picture exhibitor (542,334)
Phantoscope (586,953)
Projecting kinetoscope (707,934)
Kinetographic camera (560,800)
Camera (528,140)
Vitascope (673,992)
Kinetographic camera (629,063)
Phantoscope (536,569)
Improvement of Transmitters and Receivers for Electric Telegraphs (161,739)
Improvement in protective electric telegraphs (121,971)
Improvement in combined telegraph sounders and relays (130,426)
Improvement in printing-telegraph instruments (130,261)
Improvement in printing-telegraphs (126,336)
Improvement in telephonic telegraph-receivers (178,399)
Telephone relay (255,333)
Telegraph repeater (231,477)
Quadruplex Telegraph (254,297)
Electric Burglar-Alarm (225,271)
Telegraphic repeater (250,774)
Electric signal-bell (228,851)
Improvement in electro magnetic alarms (197,416)
Improvement in telegraphy (174,465)
Telephonic system (284,594)
Testing and breaking circuits (260,043)
Improvement in caps for lamp-globes (222,189)
Improvement in bases for lamp-globes (196,286)
Improvement in globe-holders (172,832)
Improvement in metallic tops for burner-globes (206,539)
Improvement in lanterns (207,516)
Improvement in street-lamps (188,142)
Improvement in street-lamps (205,607)
Improvement in street-lamps (206,573)
Folding globe (316,087)
Globe holder for Argand lamps (266,470)
Globe for incandescent electric lamps (285,784)
Electric incandescent lamp (258,747)
Lantern (321,993)

Figure displays the similarity network for four patents: the patent for the first sewing machine (top left); one of the earlier patents for moving pictures (top right); one of the early patents that led to the telephone (bottom left) and a randomly chosen patent from the 1800s (bottom right). In plotting the similarity links, we restrict attention to patents pairs filed at most five years apart and with a cosine similarity greater than 50%.

Figure 3: Distribution of Quality and Citations over time

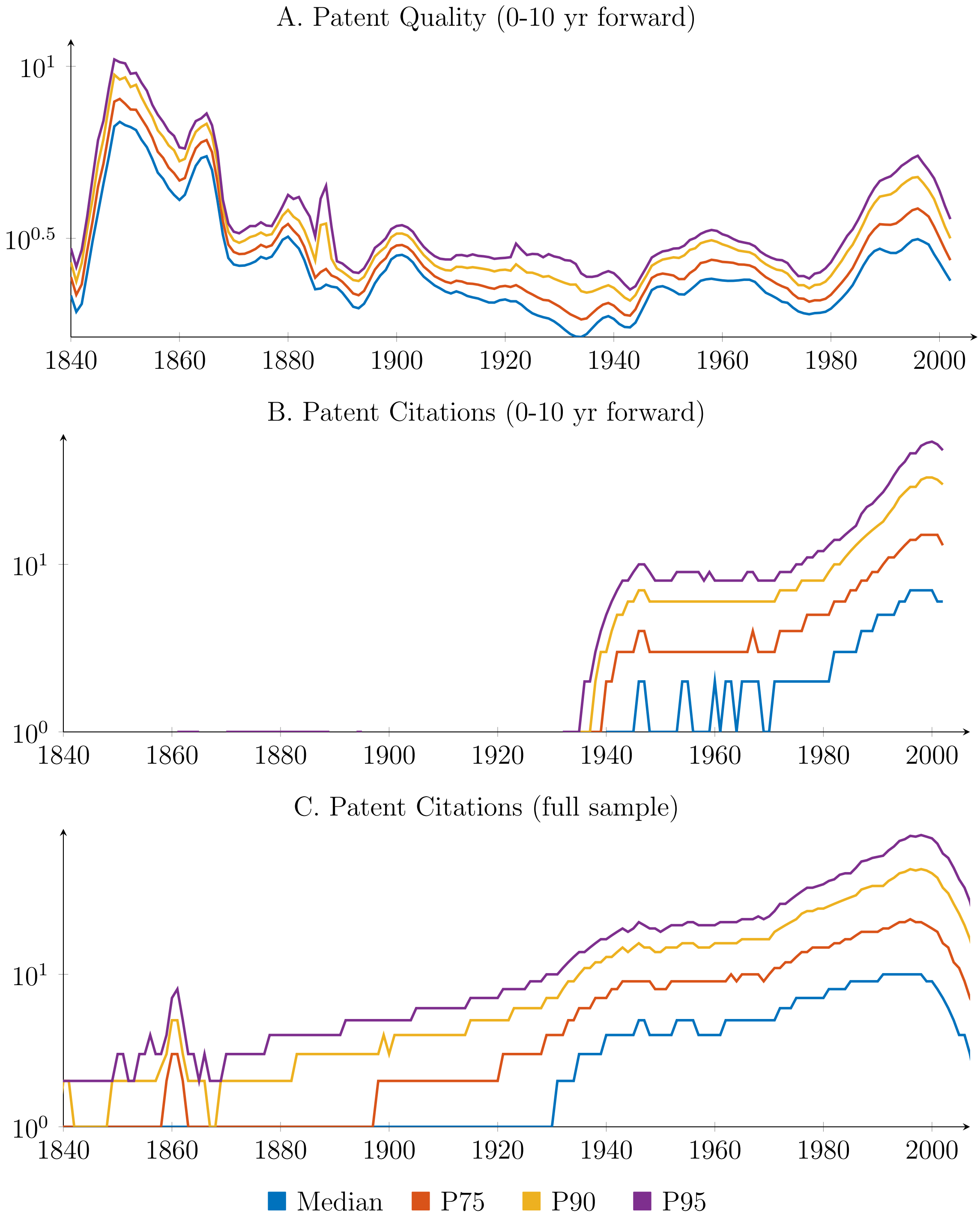
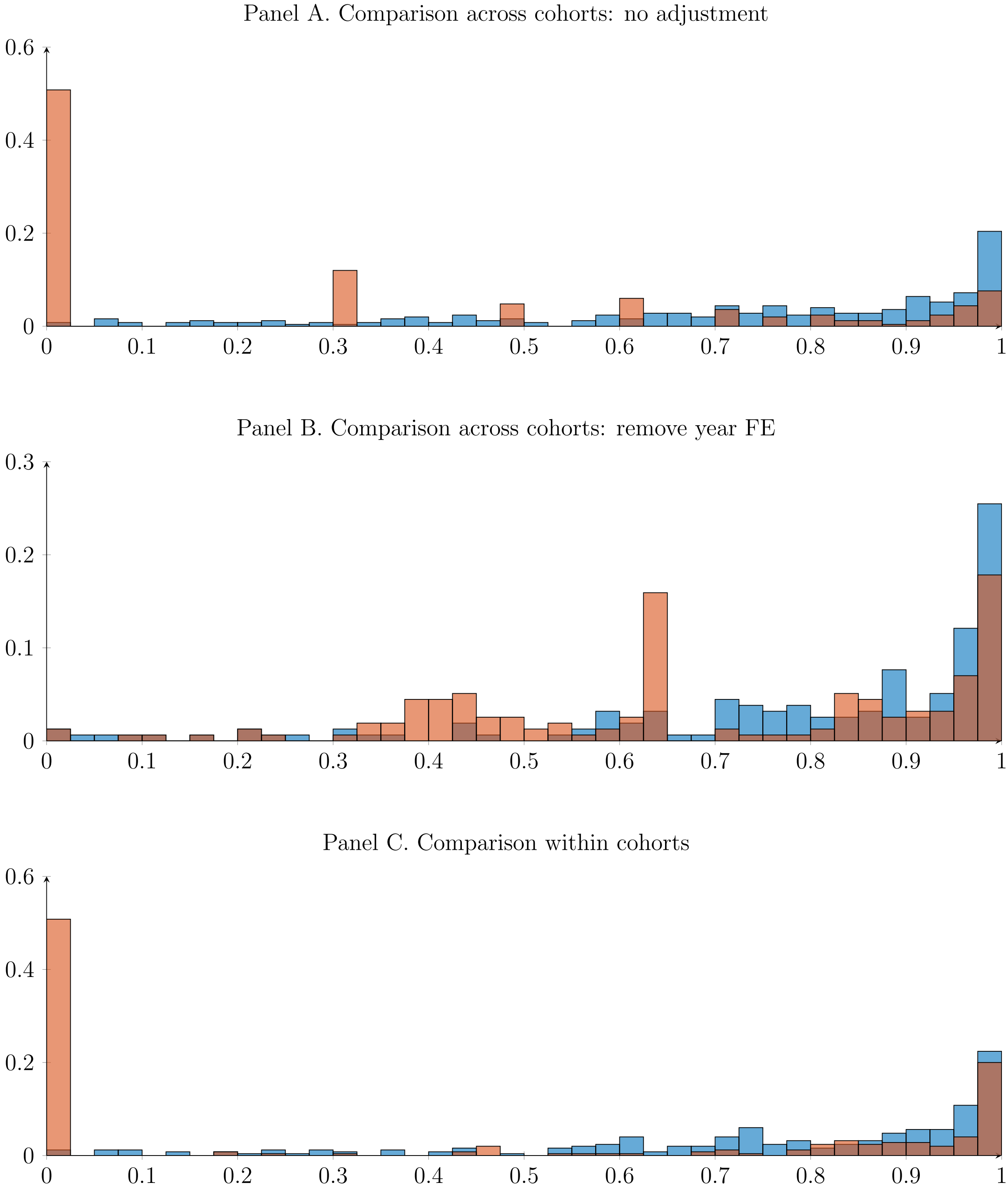


Figure plots the cross-sectional distribution of our quality measure (Panel A) and forward citations (Panels B and C) over time.

Figure 4: Important Patents: Quality vs Citations



Panel A. Comparison across cohorts: no adjustment

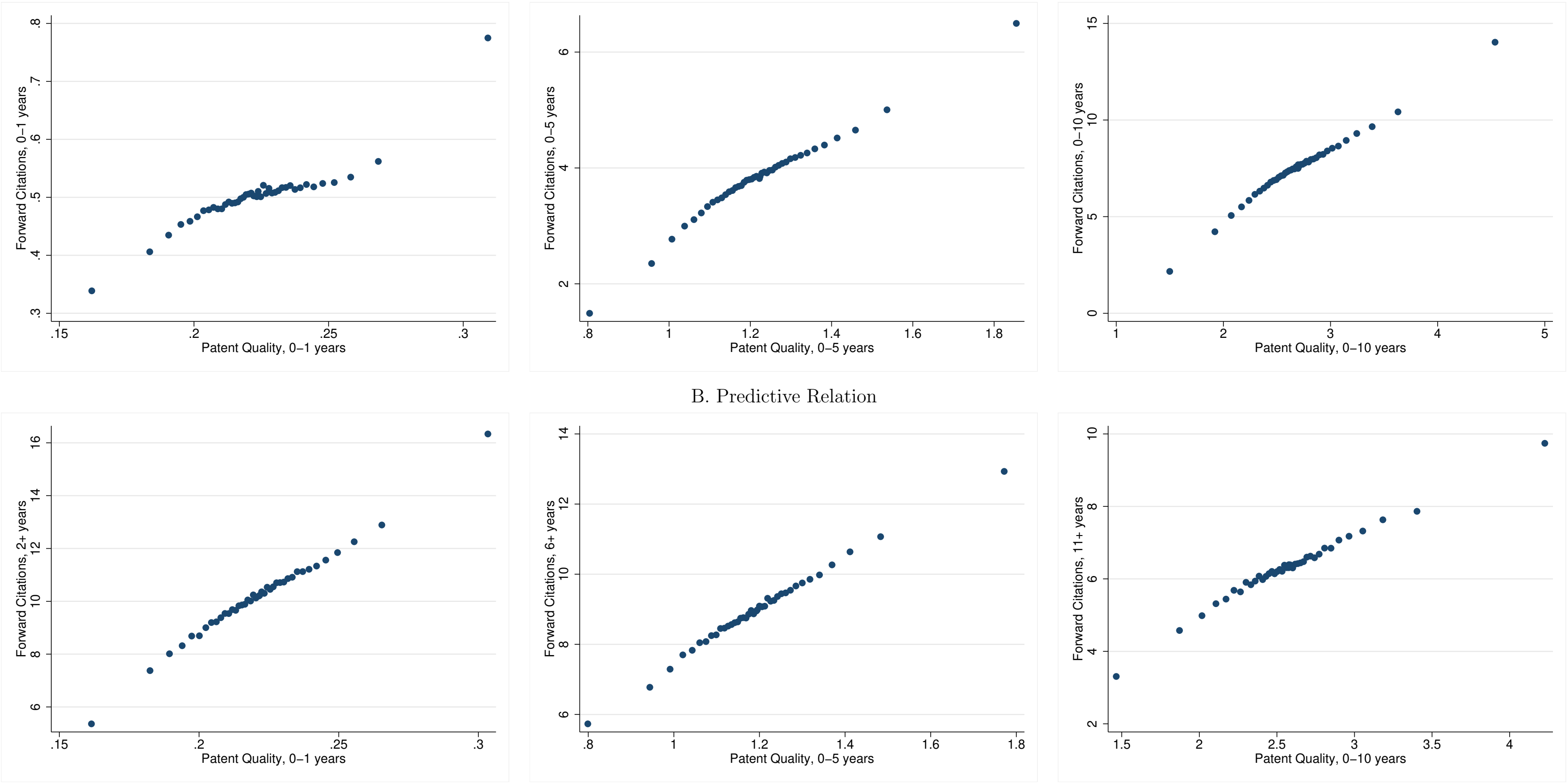
Panel B. Comparison across cohorts: remove year FE

Panel C. Comparison within cohorts

Figure compares the extent to which our quality indicator successfully identifies historically important patents, and compares with patent citations. The figure plots the distribution of patent percentile ranks based on our quality indicator (blue) and forward citations (light red) measured over a horizon of 10 years. A value of x% indicates that a given patent scores higher than x% of all other patents unconditionally (panel A); unconditionally, but adjust quality and citations by removing year-fixed effects (Panel B); or relative to patents that are issued in the same year (panel C). The list of patents, along with their source, appears in Appendix Table A.7

Figure 5: Patent quality and citations

A. Contemporaneous Relation



B. Predictive Relation

Figure plots the relation between the number of forward citations to our quality measure (both in levels). Panel A relates our quality measure to patent citations, when both are measured over the same horizon. The binned scatter plots control for fixed effects for technology class, and the interaction between assignee and patent grant year. Panel B plots the predictive relation between our quality measure and future citations; in addition to technology and assignee-issue year fixed effects, we also control for the number of citation the patent has received over the same horizon that our quality measure is computed.

Figure 6: Pairwise similarity and citation linkages

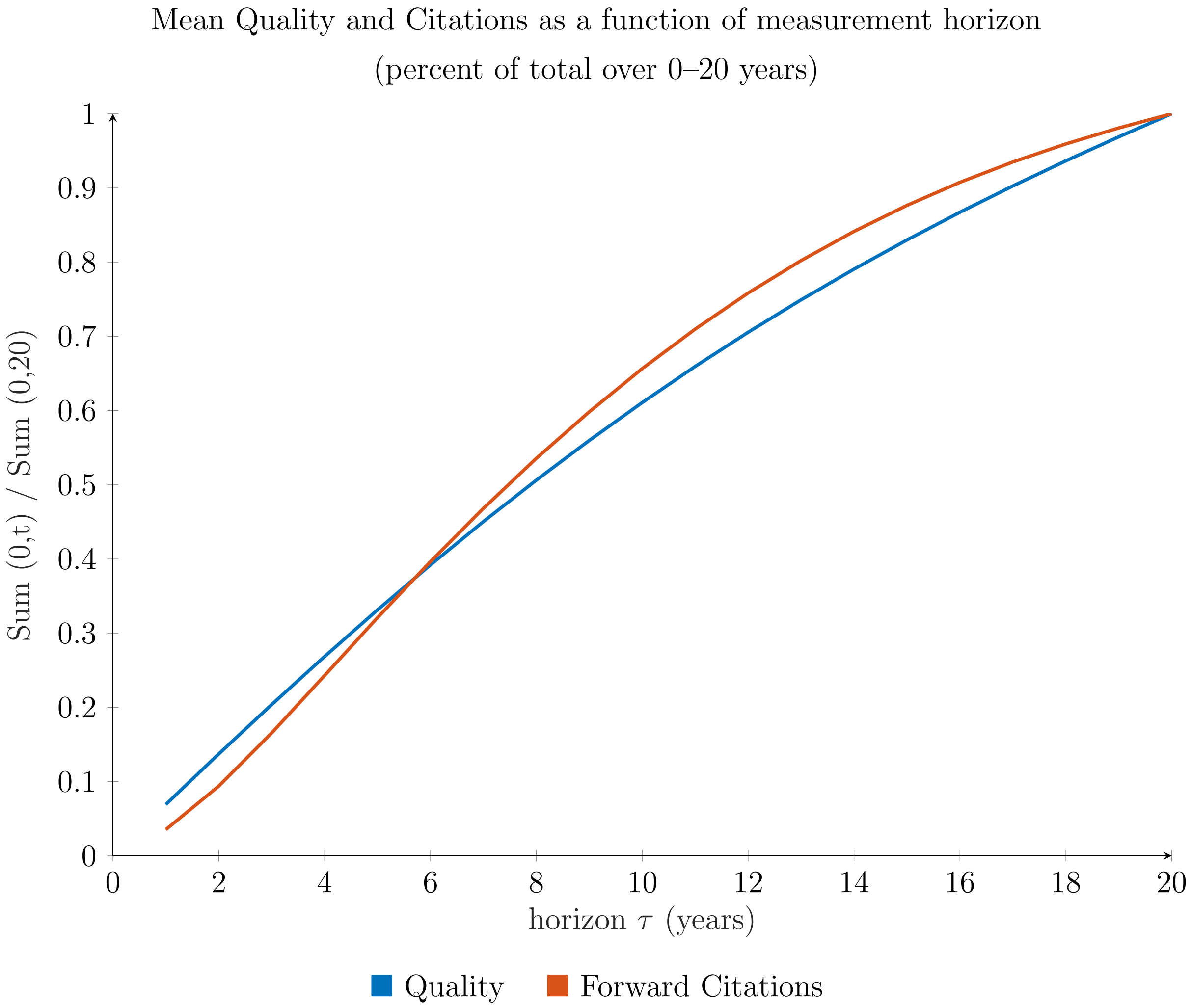
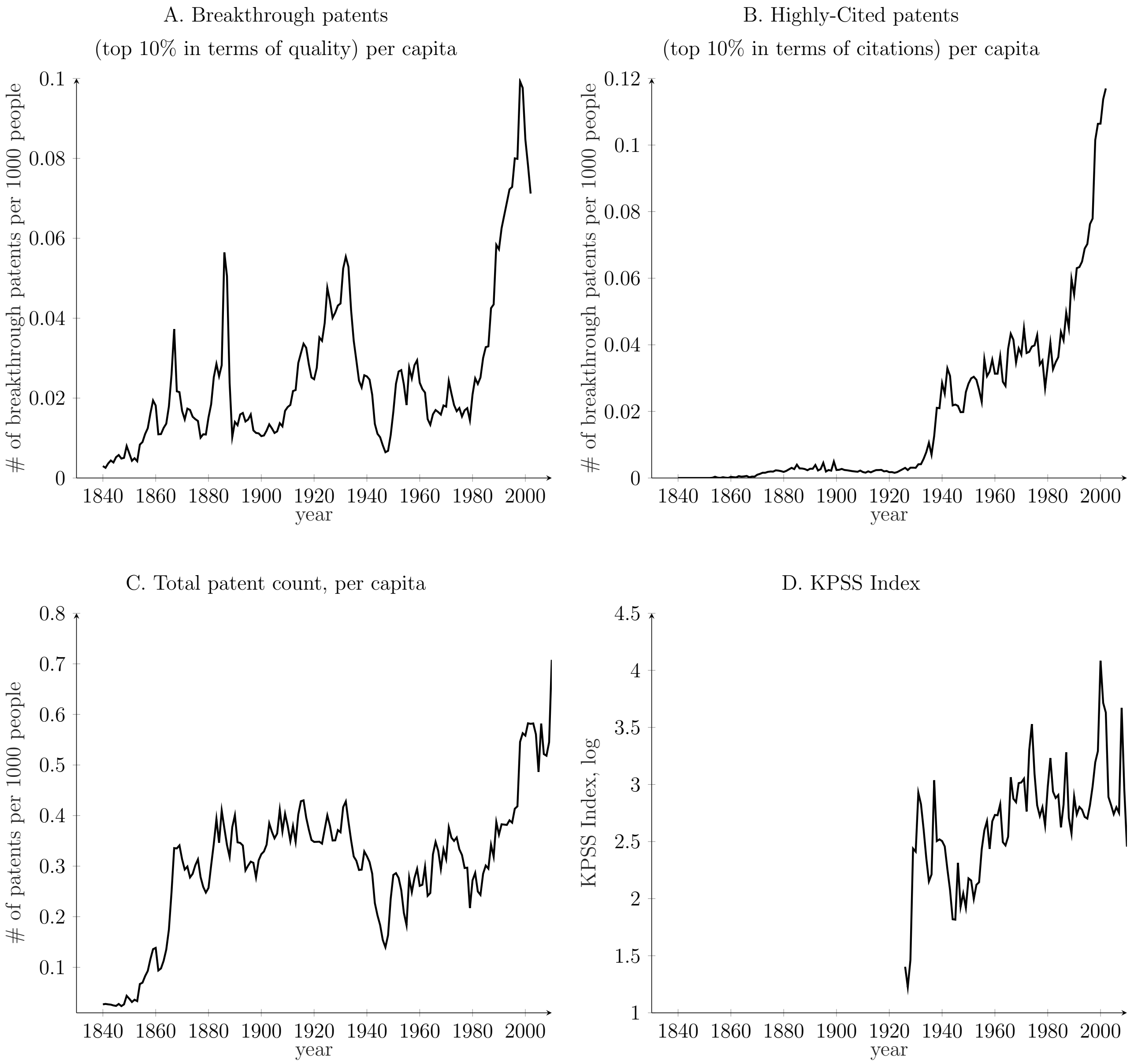


Figure examines the speed at which information about the quality of the patent is reflected in our quality measure and in forward citations. Specifically, we plot the mean across patents of the ratio of $x_{0,\tau}$ where x refers to either our quality indicator or forward citations measured over τ years subsequent to the patent, scaled by $x_{0,20}$.

Figure 7: Technological Innovation over the Long Run



Panel A plots the number of breakthrough patents, defined as the number of patents per year that fall in the top 10% of the unconditional distribution of our quality measures—defined as the ratio of the 10-yr forward to the 5-yr backward similarity, net of year fixed effects. We normalize by US population. In Panel B we plot the number of patents that fall in the top 10% of the unconditional distribution of forward citations—measured over the next 10 years, net of year fixed effects—again scaled by US population. Panel C plots the total number of patents, scaled by population, while Panel D plots the KPSS Index (the sum of the estimated market value of patents scaled by the total capitalization of the stock market.

Figure 8: Breakdown by Technology Classes

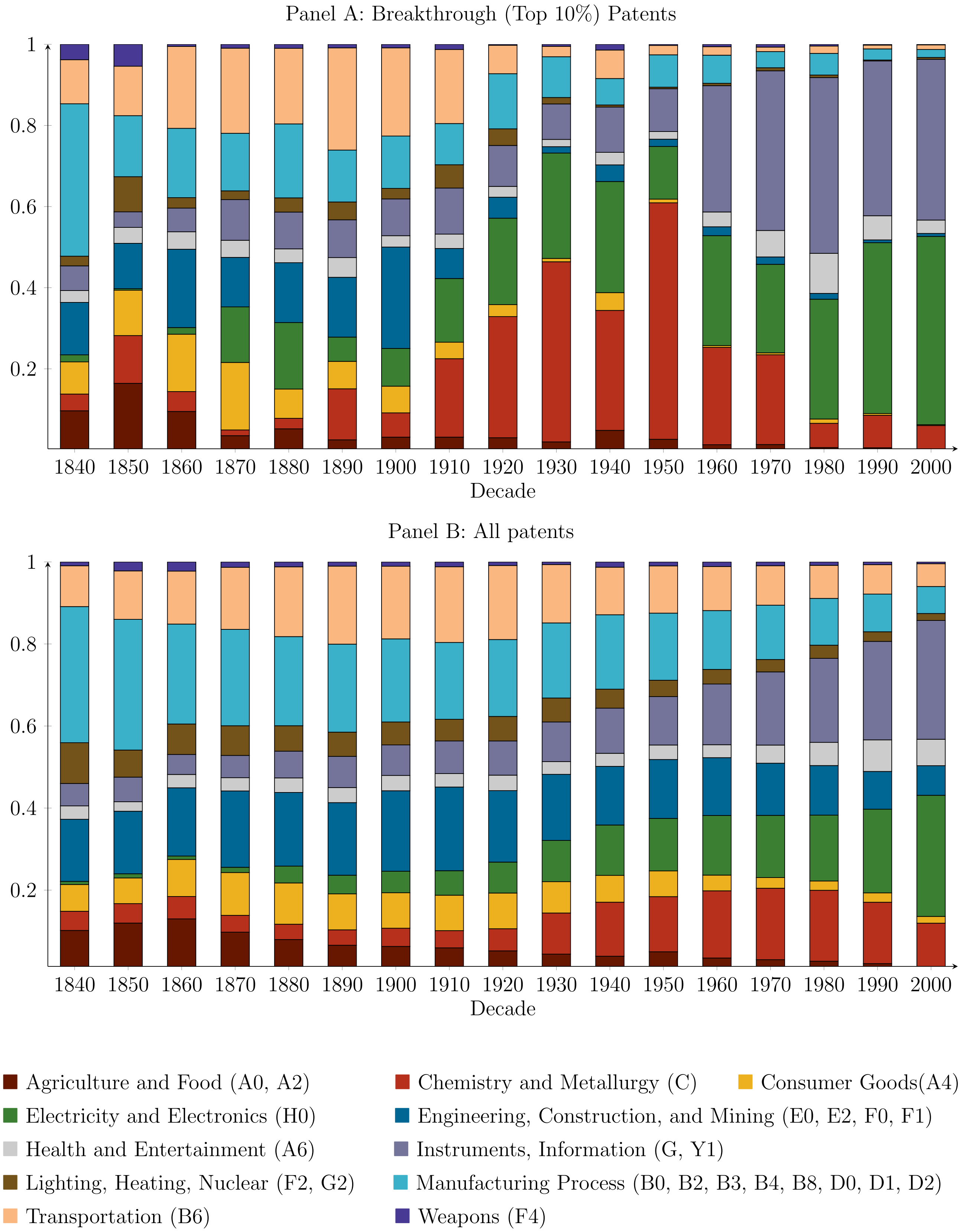
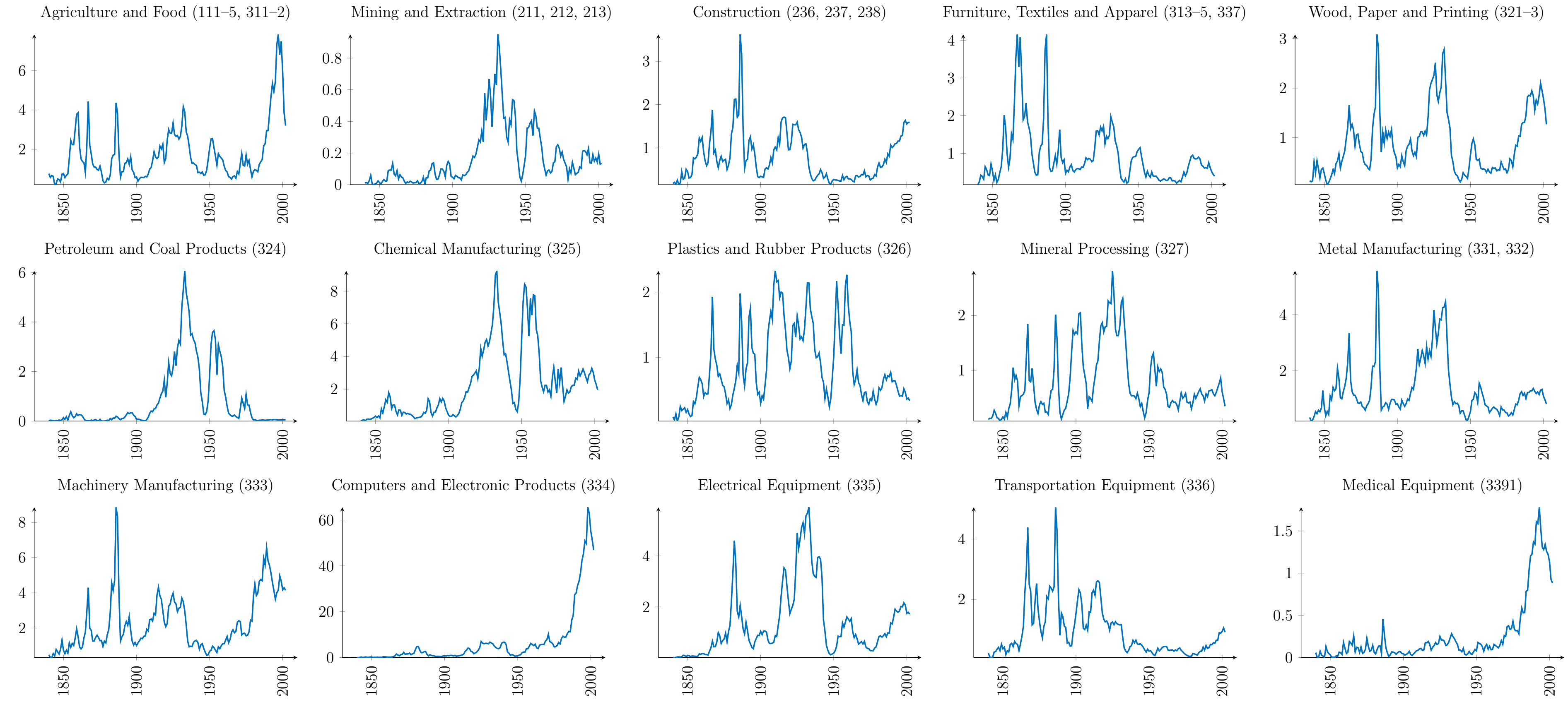


Figure 9: Innovation across industries: Breakthrough patents


Agriculture and Food (111–5, 311–2)

Mining and Extraction (211, 212, 213)

Construction (236, 237, 238)

Furniture, Textiles and Apparel (313–5, 337)

Wood, Paper and Printing (321–3)

Petroleum and Coal Products (324)

Chemical Manufacturing (325)

Plastics and Rubber Products (326)

Mineral Processing (327)

Metal Manufacturing (331, 332)

Machinery Manufacturing (333)

Computers and Electronic Products (334)

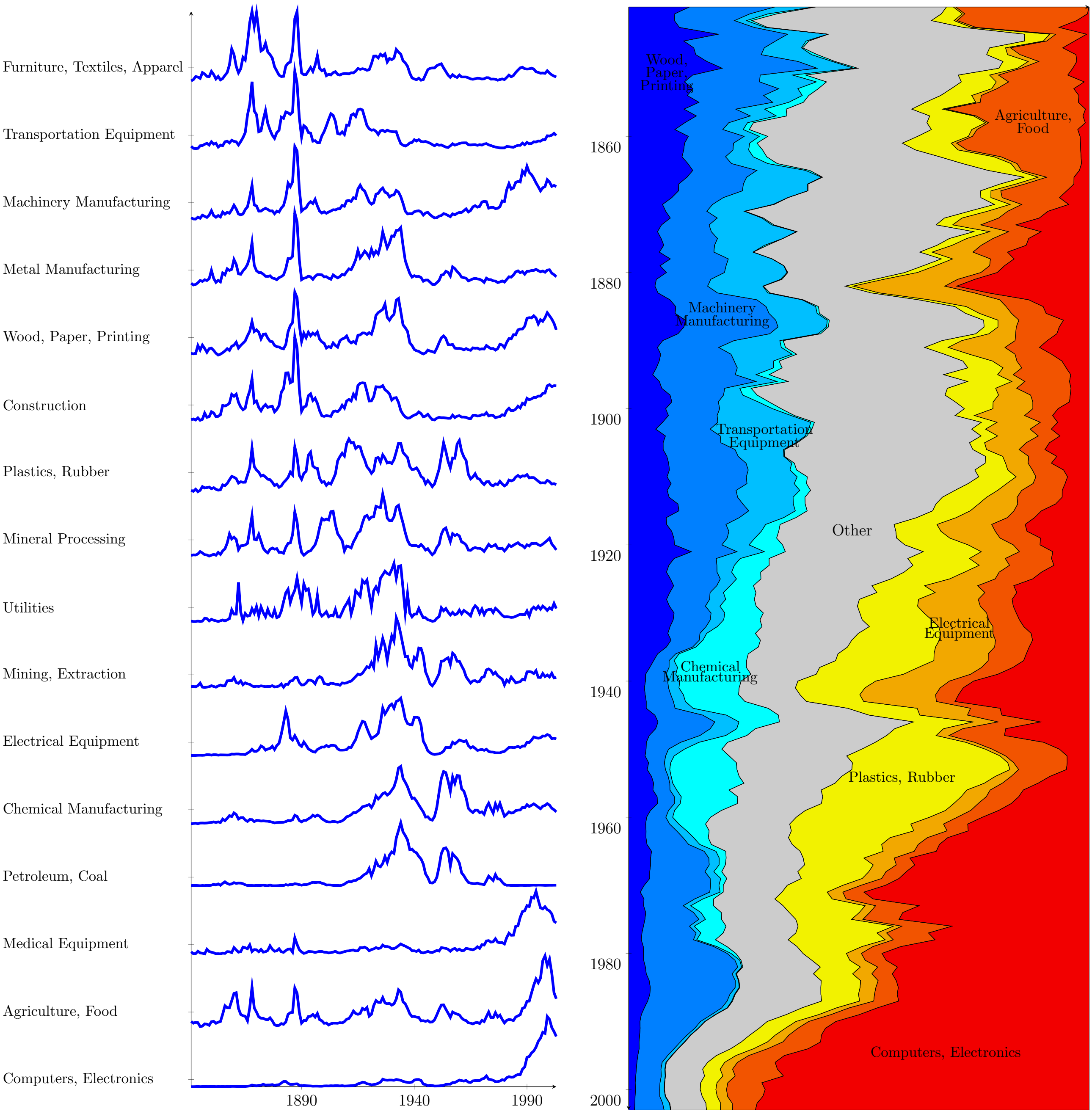
Electrical Equipment (335)

Transportation Equipment (336)

Medical Equipment (3391)

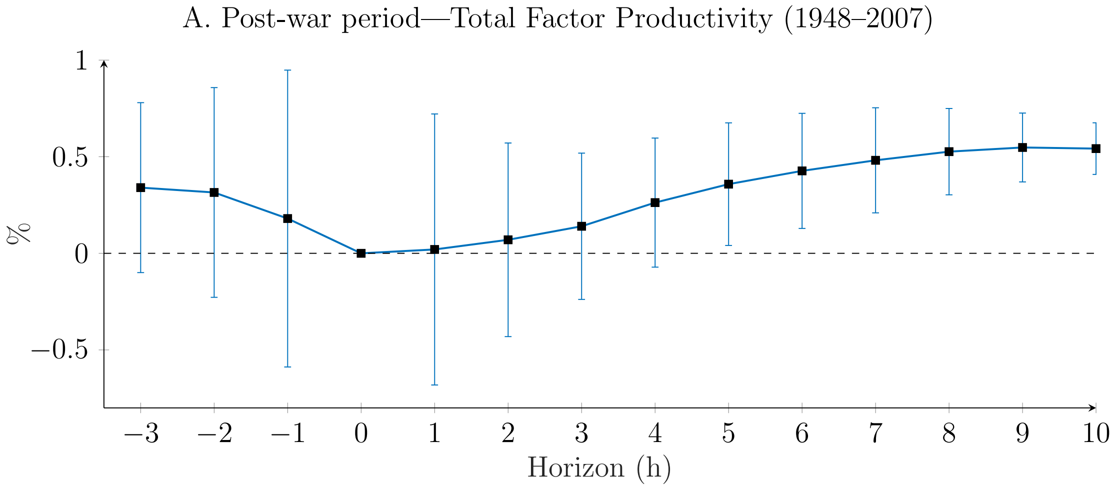
Panel plots the number of breakthrough patents across industries. Industries are defined based on NAICS codes. Breakthrough patents are those that fall in the top 10% of our baseline quality measure (defined as the ratio of the 10-yr forward to the 5-yr backward similarity) net of year fixed effects. We construct industry indices using the CPC4 to NAICS crosswalk constructed by Goldschlag et al. (2016).

Figure 10: Breakthrough Innovation Across Industries



We plot the per capita number of breakthrough patents across industries. Industries are defined based on NAICS codes. Breakthrough patents are those that fall in the top 10% of our baseline importance measure (defined as the ratio of the 10-yr forward to the 5-yr backward similarity) net of issue year fixed effects. We construct industry indices using the CPC4 to NAICS crosswalk constructed by Goldschlag et al. (2016).

Figure 11: Breakthrough patents and Aggregate TFP



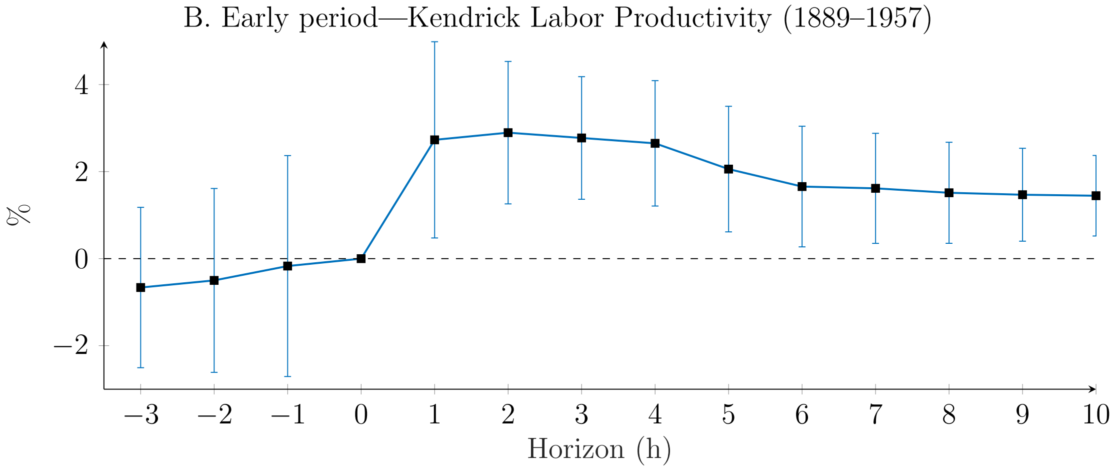
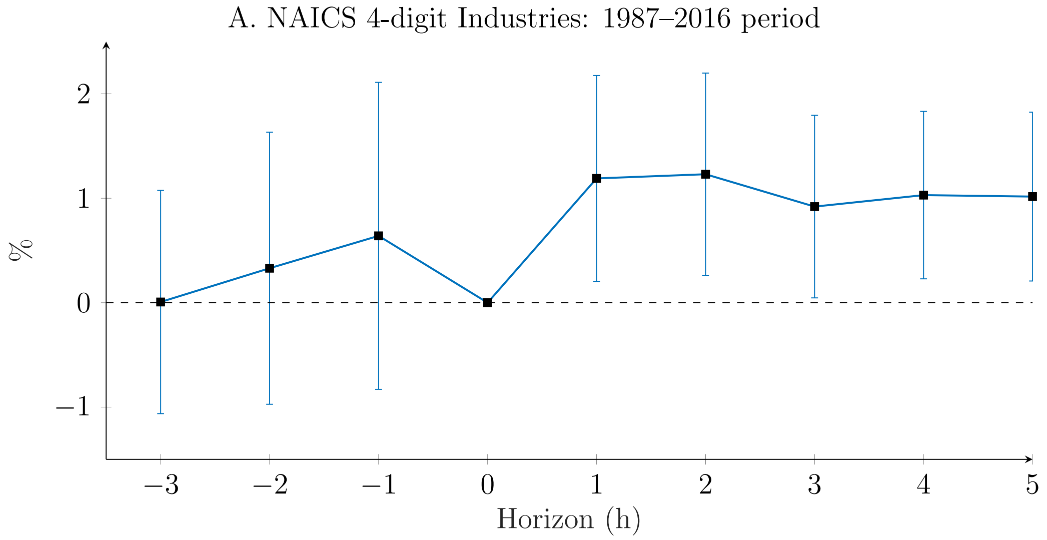


Figure plots the response of measured productivity to a unit standard deviation shock to our technological innovation index (in logs). In Panel A, productivity is measured using total factor productivity from Basu et al. (2006). In Panel B, productivity is measured by output per manhour in manufacturing (Kendrick, 1961, Table D-II). We include 90% confidence intervals, computed using Newey-West standard errors. All specifications control for the lag level of productivity.

Figure 12: Breakthrough patents and Industry TFP



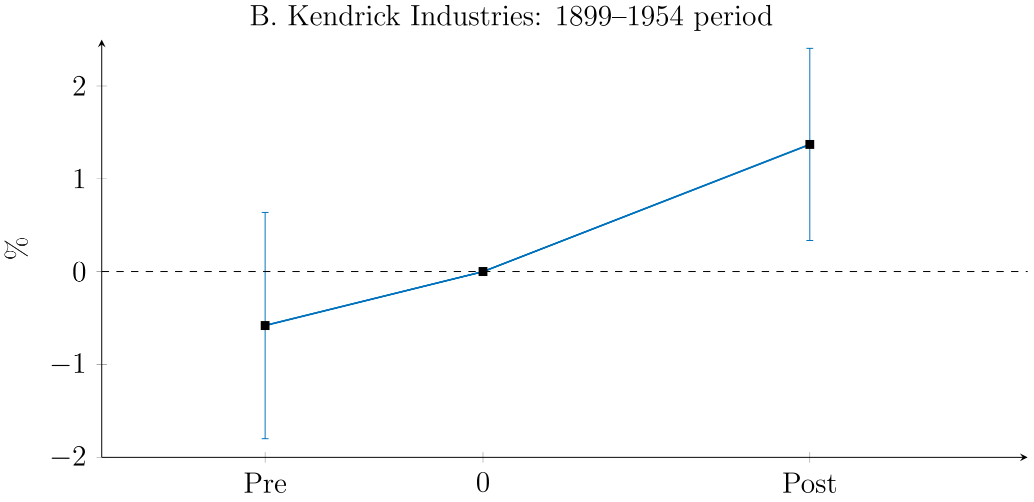
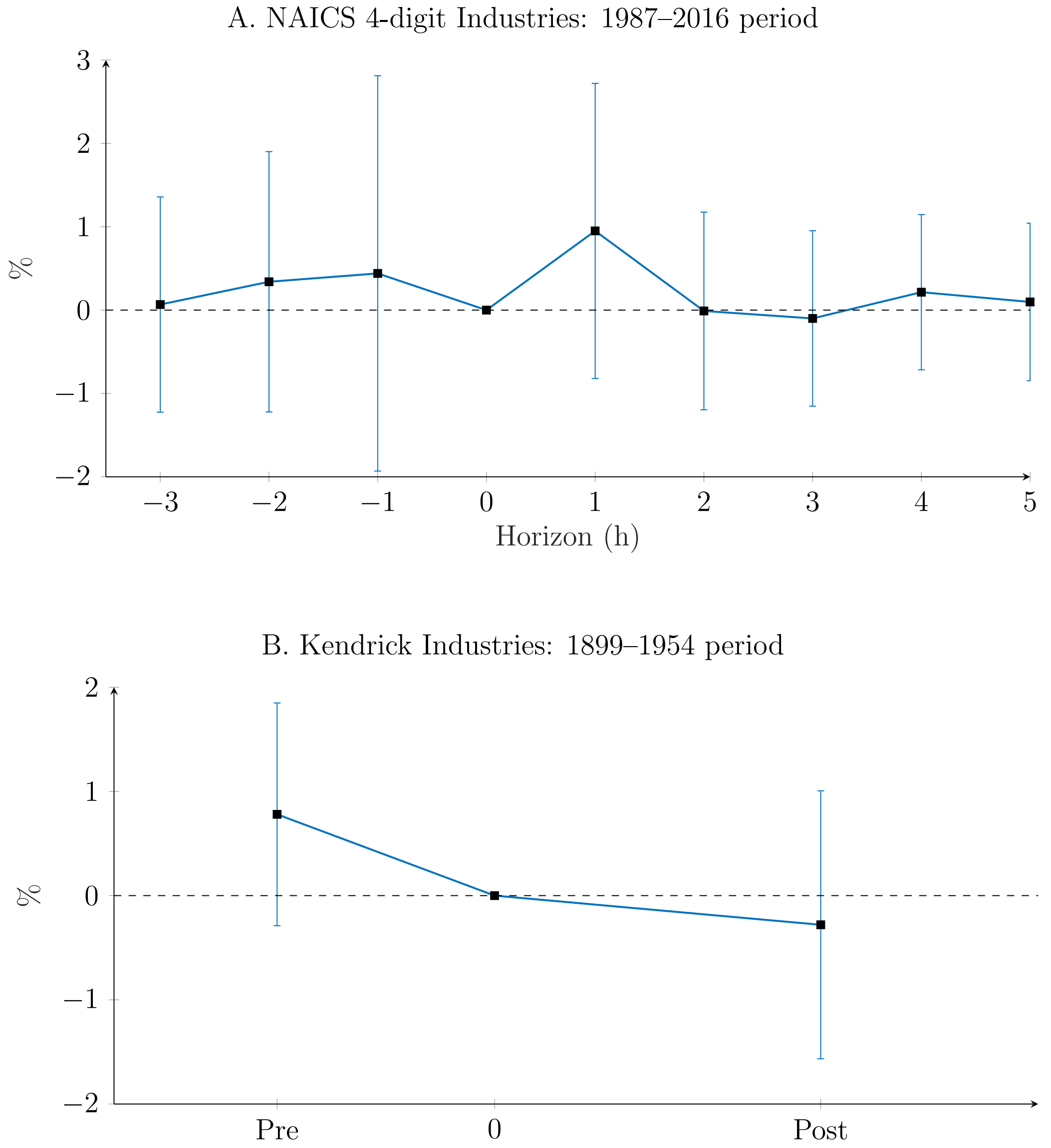


Figure plots the response of industry total factor productivity to a unit standard deviation shock to our technological innovation index. Panel A presents results for 86 manufacturing industries at the NAICS 4-digit level using data from the Bureau of Labor Statistics. The data for Panel B (Kendrick data) is from Table D-V in Kendrick (1961), and includes information for the level of labor productivity (output per manhour) for 62 manufacturing industries for the years 1899, 1909, 1919, 1937, 1947, and 1954. For each period (t, s), we regress the annualized difference in log labor productivity on the log of the accumulated level of innovation (number of breakthrough patents) in $t \pm 2$ years—controlling for period, industry dummies, the log number of patents during the same period, and the log level of productivity at t. Standard errors are clustered by industry. To construct industry innovation indices for NAICS industries, we use the probabilistic mapping from CPC codes to NAICS codes from Goldschlag et al. (2016). To construct innovation indices for the Kendrick industries, which are defined at the SIC code level, we use the concordance between 1997 NAICS and 1987 SIC codes from the Census Bureau. If NAICS industries map into multiple SIC codes, we assign an equal fraction to each.

Figure 13: Breakthrough patents and Industry TFP—comparison to Citations


A. NAICS 4-digit Industries: 1987–2016 period

B. Kendrick Industries: 1899–1954 period

Figure performs the same exercise as Figure 11, except that we now construct the industry innovation indices based on citation counts.

Appendix

We briefly overview our conversion of unstructured patent text data into a numerical format suitable for statistical analysis. To begin, we build our collection of patent documents from two sources. The first is the USPTO patent search website, which records all patents beginning from 1976. Our web crawler collected the text content of patents from this site, which includes patent numbers 3,930,271 through 9,113,586. The records in this sample are comparatively easy to process as they are available in HTML format with standardized fields.

For patents granted prior to 1976, we collect patent text from our second main datasource, Google's patent search engine. For the pre-1976 patent records, we recover all of the fields listed above with the exception of inventor/assignee addresses (Google only provides their names), examiner, and attorney. Some parts of our analysis rely on firm-level aggregation of patent assignments. We match patents to firms by firm name and patent assignee name. Our procedure broadly follows that of Kogan et al. (2017) with adaptations for our more extensive sample. In addition to the citation data we scrape from Google, we obtain complementary information on patent citations from Berkes (2016) and the USPTO. The data in Berkes (2016) includes citations that are listed inside the patent document and which are sometimes missed by Google. Nevertheless, the likelihood of a citation being recorded is significantly higher in the post-1945 than in the pre-1945. When this consideration is relevant, we examine results separately for the pre- and post-1945 periods.

To represent patent text as numerical data, we convert it into a *document term matrix* (DTM), denoted C. Columns of C correspond to words and rows correspond patents. Each element of C, denoted c_{pw}, counts the number of times a given one-word phrase (indexed by w) is used in a particular patent (indexed by p), after imposing a number of filters to remove stop words, punctuation, and so forth. We provide a detailed step-by-step account of our DTM construction in Appendix V. Our final dictionary includes 1,685,416 terms in the full sample of over nine million patents.

The next section provides additional details on the data construction, including the process through which we convert the text of patent documents to a format that is amenable to constructing similarity measures.

A. Text Data Collection, Additional Details

The Patent Act of 1836 established the official US Patent Office and is the grant year of patent number one.[10] We construct a dataset of textual content of US patent granted during the 180 year period from 1836-2015. Our dataset is built on two sources.

[10] The first patent was granted in the US in 1790, but of the patents granted prior to the 1836 Act, all but 2,845 were destroyed by fire.

The first is the USPTO patent search website. This site provides records for all patents beginning in 1976. We designed a web crawler collect the text content of patents over this period, which includes patent numbers 3,930,271 through 9,113,586. We capture the following fields from each record:

1. Patent number (WKU)
2. Application date
3. Granted date
4. Inventors
5. Inventor addresses
6. Assignees
7. Assignee addresses
8. Family ID
9. Application number
10. US patent class
11. CPC patent class
12. Intl. patent class
13. Backward citations
14. Examiner
15. Attorney
16. Abstract
17. Claims
18. Description

The only information available from USPTO that we do not store are image files for a patent's "figure drawing" exhibits.

For patents granted prior to 1976, the USPTO also provides bulk downloads of .txt files for each patent. The quality of this data is inferior to that provided by the web search interface in three ways. First, the text data is recovered from image files of the original patent documents using OCR scans. OCR scans often contain errors. These generally arise from imperfections in the original images that lead to errors in the OCR's translation from image to text. Going backward in time from 1976, the quality of OCR scans deteriorates rapidly due to lower quality typesetting. Second, the bulk download files do not use a standardized format which makes it difficult to parse out the fields listed above.

Rather than using the USPTO bulk files, we collect text of pre-1976 patents from our second main datasource, Google's patent search engine. Like post-1976 patents from USPTO, Google provides patent records in an easy-to-parse HTML format that we collect with our web crawler. Furthermore, inspection of Google records versus 1) OCR files from the USPTO and 2) pdf images of patents that are the source of the OCR scans, reveals that in this earlier period Google's patent text is more accurate than the OCR text in USPTO bulk data. From Google's pre-1976 patent records, we recover all of the fields listed above with the exception of inventor/assignee addresses (Google only provides their names), examiner, and attorney.

B. Cleaning Post-1976 USPTO Data

Next, we conduct a battery of checks to correct data errors. For the most part, we are able to capture and parse of patent text from the USPTO web interface without error. When there are errors, it is almost always the case that the patent record was incompletely captured, and this occurs for one of two reasons. The first reason is that the network connection was interrupted during the capture and the second is that the patent record on the UPSTO website is itself incomplete (in comparison with PDF image files of the original document, which are also available from USPTO via bulk download).

Our primary data cleaning task was to find and complete any partially captured patent records. First, we find the list of patent numbers (WKUs) that are entirely missing from our database, and re-run our capture program until all have been recovered. Many of the missing records that we find are explicitly labeled as "WITHDRAWN" at the USPTO. Withdrawn information can be found at https://www.uspto.gov/patents-application-process/patent-search/withdrawn-patent-numbers. Next, we identify WKUs with an entirely missing value for the abstract, claims, or description field. Fortunately, we find this to be very infrequent, occurring in less than one patent in 100,000, making it easy for us to correct this manually.

Next, a team of research assistants (RA's) manually checked 3,000 utility patent records, 1,000 design patent records, and 1,000 plant patents records against their PDF image files. The RA task is to identify any records with missing or erroneous information in the reference, abstract, claims, or description fields. To do this, they manually read the original pdf image for the patent and our digitally captured record. We identify patterns in partial text omission and update our scraping algorithm to reflect these. We then re-ran the capture program on all patents and confirmed that omissions from the previous iteration were corrected.

C. Cleaning Pre-1976 Google Data

Fortunately, we find no instances of missing WKU's or incomplete text from Google web records. Next, we assess the accuracy of Google's OCR scans by manually re-scanning a random sample of 1,000 pre-1976 patents using more recent (and thus more accurate) ABBYY OCR software than was used for most of Google's image scans. We compare the ABBYY scan to the pdf image to confirm the scan content is complete, the compare the frequency of garbled terms in our scan versus that OCR text from Google. The distribution of pairwise cosine similarities in our ABBYY text and Google's OCR is reported below.

Cosine Similarity	
mean	0.957
std	0.073
P1	0.701
P5	0.863
P10	0.900
P25	0.951
P50	0.977
P75	0.991
P90	0.996
P95	0.998
P99	0.999
N	1000

Only 10% of sampled Google OCR records have a correlation with ABBYY below 90%.

Next, we manually compare both our OCR scans and those from Google against the pdf image. We find that garble rate for ABBYY OCRed is 0.025 on average, with standard deviation of 0.029. We find that Google has only slightly more frequent garbling than our ABBYY scans. Of the term discrepancies in the two sets of scans, around 52% of these correspond to a garbled ABBYY records and 83% to a garbled Google record. We ultimately conclude that Google's OCR error frequency is acceptable for use in our analysis.

D. Conversion from Textual to Numeric Data

We convert the text content of patents into numerical data for statistical analysis. To do this, we use the NLTK Python Toolkit to parse the "abstract," "claims," and "description" sections of each patent into individual terms. We strip out all non-word text elements, such as punctuation, numbers, and HTML tags, and convert all capitalized characters to lowercase. Next, we remove all occurrences of 947 "stop words," which include prepositions, pronouns, and other words that carry little semantic content.[11]

[11] We construct our stop word list as the union of terms in the following commonly used lists:

http://www.ranks.nl/stopwords
https://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html
https://code.google.com/p/stop-words/
http://www.lextek.com/manuals/onix/stopwords1.html
http://www.lextek.com/manuals/onix/stopwords2.html
http://www.webconfs.com/stop-words.php
http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html
http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_170.html
https://pypi.python.org/pypi/stop-words
https://msdn.microsof,t.com/zh-cn/library/bb164590
http://www.nltk.org/book/ch02.html (NLTK list)

The remaining list of "unstemmed" (that is, without removing suffixes) unigrams amounts to a dictionary of 35,640,250 unique terms. As discussed in Gentzkow, Kelly, and Taddy (2017), an important preliminary step to improve signal-to-noise ratios in textual analysis is to reduce the dictionary by filtering out terms that occur extremely frequently or extremely infrequently. The most frequently used words show up in so many patents that they are uninformative for discriminating between patent technologies. On the other hand, words that show up in only a few patents can only negligibly contribute to understanding broad technology patterns, while their inclusion increases the computational cost of analysis.[12]

We apply filters to retain influential terms while keeping the computational burden of our analysis at a manageable level, and focus on the number of distinct patents and calendar years in which terms occur. Table A.1 reports the distribution across terms for number of patents and the number of distinct calendar years in which a term appears. A well known attribute of text count data is its sparsity—most terms show up very infrequently—and the table shows that this pattern is evident in patent text as well. We exclude terms that appear in fewer than twenty out of the more than nine million patents in our sample. These eliminate 33,954,834 terms, resulting in a final dictionary of 1,685,416 terms.[13]

After this dictionary reduction, the entire corpus of patent text is reduced in a $D \times W$ numerical matrix of term counts denoted C. Matrix row d corresponds to patent (WKU) d. Matrix column w corresponds the w^{th} term in the dictionary. Each matrix element c_{dw} the count of term w in patent d.

E. Matching Patents to Firms

Much of our analysis relies on firm-level aggregation of patent assignments. We match patents to firms by merging firm names and patent assignee names. Our procedure broadly follows that of Kogan et al. (2017) with adaptations for our more extensive sample.

The first step is extracting assignee names from patent records. For post-1976 data we use information from the USPTO web search to identify assignee names. Due to the high data quality in this sample, assignee extraction is straightforward and highly accurate. For pre-1976, we use assignee information from Google patent search. While it is easy to locate the assignee name field thanks to the HTML format, Google's assignee names are occasionally garbled by the OCR.

[12] Filtering out infrequent words also removes garbled terms, misspellings, and other errors, as their irregularity leads them to occur only sporadically.

[13] The table also shows that there are some terms that appear in almost all patents. Examples of the most frequently occurring words (that are not in the stop word lists) are "located," "process," and "material." Because these show up in most patents they are unlikely to be informative for statistical analysis. These terms are de-emphasized in our analysis through the $TFIDF$ transformation.

Next, we clean the set of extracted assignee names. There are 766,673 distinct assignees in patents granted since 1836. Most of the assignees are firm names and those that are not firms are typically the names of inventors. We clean assignee name garbling using fuzzy matching algorithms. For example, the assignee "international business machines" also appears as an assignee under the names "innternational business machines," "international businesss machines," and "international business machiness." Garbled names are not uncommon, appearing for firms as large as GE, Microsoft, Ford Motor, and 3M.

We primarily rely on Levenshtein edit distance between assignees to identify and correct erroneous names. There are two major challenges to overcome in name cleaning. The first choosing a distance threshold for determining whether names are the same. As an example, the assignees "international business machines" (recorded in 103,544) and "ibm" (recorded in 547 patents) have a large Levenshtein distance. To address cases like this, we manually check the roughly 3,000 assignee names that have been assigned at least 200 patents, correcting those that are variations on the same firm name (including the IBM, GE, Microsoft, Ford, and 3M examples). Next, for each firm on the list of most frequent assignees, we calculate the Levenshtein distance between this assignee name and the remaining 730,000+ assignee names, and manually correct erroneous names identified by the list of assignees with short Levenshtein distances.

The second challenge is handling cases in which a firm subsidiary appears as assignee. For example, the General Motors subsidiary "gm global technology operations" is assigned 8,394 patents. To address this, we manually match subsidiary names from the list of top 3,000+ assignees to their parent company by manually searching Bloomberg, Wikipedia, and firms' websites.

After these two cleaning steps, and after removing patents with the inventor as assignee, we arrive at 3,036,859 patents whose assignee is associated with a public firm in CRSP/Compustat, for a total of 7,467 distinct cleaned assignee firm names. We standardized these names by removing suffixes such as "com," "corp," and "inc," and merge these with CRSP company names. Again we manually check the merge for the top 3,000+ assignees, and check that name changes are appropriately addressed in our CRSP merging step. Finally, we also merge our patent data with Kogan et al. (2017) patent valuation data for patents granted between 1926 and 2012.

F. Patent Quality and Firm Valuation Ratios

Here, we examine the extent to which our text-based patent quality measure accounts for differences in firm value. Our analysis closely follows that of Hall et al. (2005), who estimate the relation between a firm's Tobin's Q and its "knowledge stock." Hall et al. (2005) define knowledge stock as a depreciating balance of the firm's investment in R&D, its number of patents, or its patent citation count, according to the formula

$$SX_{f,t} = (1 - \delta) \, SX_{f,t-1} + X_{f,t} \tag{16}$$

where $X_{f,t}$ represents either the flow of new R&D, successful patent applications, or citations received by patents, for firm f in year t. $SX_{f,t}$ is thus the firm's accumulated stock of X. We use the same depreciation rate of $\delta = 15\%$ as Hall et al. (2005).

We introduce a fourth knowledge stock variable based on our patent quality measure. First, we define firm-level patent quality for firm f in year t as:

$$q^{\tau}_{f,t} = \sum_{j \in J_{f,t}} q^{\tau}_{j} \tag{17}$$

where, $J_{f,t}$ is the set of patents filed for firm f in year t. We then create a "quality-weighted" patent stock that accumulates (17) according to (16) (again using $\delta = 15\%$).[14]

Our firm-level regression specification, following Hall et al. (2005), is

$$\begin{aligned}\log Q_{f,t} = \log &\left(1 + \gamma_1 \frac{SRD_{f,t}}{A_{f,t}} + \gamma_2 \frac{SPAT_{f,t}}{SRD_{f,t}} + \gamma_3 \frac{SCITES^{\tau}_{f,t}}{SPAT_{f,t}} + \gamma_4 \frac{Sq^{\tau}_{f,t}}{SPAT_{f,t}}\right) \\ &+ a_t + D\left(SRD_{f,t} = 0\right) + \varepsilon_{f,t}\end{aligned} \tag{18}$$

where $SRD_{f,t}$, $SPAT_{f,t}$, $SCITES_{f,t}$, and $q_{f,t}$ are the stocks of R&D expenditure, number of patents, patent citations, and the patent quality measures constructed as in (16). We follow the Hall et al. (2005) choices for scaling knowledge stock variables, scaling R&D stock by total assets ($A_{t,t}$), patent stock by R&D stock, and citation stock by patent stock. We scale our patent quality stock by the stock of patents by count, giving it an interpretation as the average quality of patents held by the firms. We estimate the market value regressions using quality and citation stocks over horizons τ of 1, 5, or 10 years after the application date. For our baseline results, we restrict the sample to patenting firms (that is, firms that have filed at least one patent). As in Hall et al. (2005), a_t is the fixed effect for year t and accounts for any time specific effect that moves around the value of all the firms in a given year. We also include a dummy variable for missing R&D observations. Depending on the specification, we also include industry-fixed effects, based on the 49 industry classification of Fama and French (1997). We cluster standard errors by firm.

Our main coefficient of interest is γ_4 which estimates the relationship between quality-weighted patent stock and firm value. Table A.6 presents the results. Examining column (2), we see a strong and statistically significant relation between Tobin's Q and the patent quality stock.

[14] We have experimented with depreciation rates of 5, 10. 20 and 25% and found similar results.

A one-standard deviation increase in the (per-patent) quality stock is associated with a 0.15 log point increase in Tobin's Q—evaluated at the median—which is economically significant given that the unconditional standard deviation in log Tobin's Q is equal to 0.63. For comparison, a one-standard deviation increase in the citation-weighted stock in column (3) is associated with a 0.13 log point increase. Column (4) shows that the our quality indicator contains information that is complementary to citations, both variables are statistically significant and account for a comparable share of the overall variation in Q—approximately 0.1 and 0.11 log points, respectively. Column (5) shows that both variables also account for within-industry variation in Tobin's Q. Last, columns (6) through (8) show that both indicators of quality are jointly statistically and economically significant when we restrict attention to manufacturing, pharmaceutical, and the high-tech industry.

Appendix Material

Table A.1: Distribution of document terms across patents

	# Patents	# Years
mean	124.03	3.33
std	12465.99	9.29
min	1	1
50%	1	1
75%	2	2
90%	7	6
95%	24	14
98%	69	24
max	8399814	182

Table reports the distribution across terms for number of patents and the number of distinct calendar years in which a term appears.

Table A.2: Historically Important Patents: Quality vs Citations

Mean Percentile Rank	Quality		Citations		
	(0–5years)	(0–10years)	(0–5years)	(0–10years)	(full sample)
A. Comparison across cohorts, no adjustment	0.739	0.742	0.334	0.369	0.548
	(0.017)	(0.016)	(0.024)	(0.025)	(0.024)
B. Comparison across cohorts, remove year FE	0.774	0.780	0.680	0.699	0.758
	(0.016)	(0.016)	(0.016)	(0.017)	(0.016)
C. Comparison within cohorts	0.772	0.779	0.425	0.490	0.688
	(0.016)	(0.016)	(0.029)	(0.029)	(0.023)

Table compares the extent to which our quality indicator successfully identifies historically important patents, and compares with patent citations. The presents mean patent percentile ranks based on our quality indicator (Column 1) and forward citations (Columns 2 and 3). A value of x% indicates that a given patent scores higher than x% of all other patents unconditionally (row A); unconditionally, but adjust quality and citations by removing year-fixed effects (row B); or relative to patents that are issued in the same year (row C). Standard errors are in parentheses. The list of patents, along with their source, appears in Appendix Table A.7

Table A.3: Patent impact and novelty predicts citations (includes old patents)

$\log(1 + \text{Forward citations, 2+ yr})$	(1)	(2)	(3)	(4)
log(Patent quality, 0-1yr)	0.788^{***}	0.683^{***}	0.709^{***}	0.810^{***}
	(9.42)	(11.20)	(11.03)	(12.90)
$\log(1 + \text{Forward citations, 0-1 yr})$	0.660^{***}	0.610^{***}	0.516^{***}	0.511^{***}
	(32.77)	(37.00)	(40.97)	(37.35)
Observations	8460384	8422712	3619813	3173149
R^2	0.397	0.439	0.520	0.556
$\log(1 + \text{Forward citations, 6+ yr})$	(1)	(2)	(3)	(4)
log(Patent quality, 0-5yr)	0.451^{***}	0.529^{***}	0.563^{***}	0.668^{***}
	(8.17)	(11.38)	(9.15)	(10.38)
$\log(1 + \text{Forward citations, 0-5yr})$	0.611^{***}	0.581^{***}	0.532^{***}	0.541^{***}
	(35.22)	(36.02)	(34.24)	(33.01)
Observations	7432397	7397785	3165185	2756793
R^2	0.398	0.442	0.522	0.557
$\log(1 + \text{Forward citations, 11+ yr})$	(1)	(2)	(3)	(4)
log(Patent quality, 0-10yr)	0.148^{***}	0.309^{***}	0.326^{***}	0.375^{***}
	(4.44)	(13.25)	(9.07)	(9.43)
$\log(1 + \text{Forward citations, 0-10yr})$	0.561^{***}	0.531^{***}	0.503^{***}	0.508^{***}
	(35.20)	(36.22)	(32.21)	(31.99)
Observation	6619620	6587879	2802615	2429734
R^2	0.338	0.388	0.476	0.515
Grant Year FE	Y	Y		
Class		Y		
Assignee FE			Y	
Grant Year × Assignee FE				Y

Table reports the results of estimating equation (12) in the main text. The regression relates the log of (one plus) the number of patent citations over a horizon $[t, s]$ to our measures of patent quality (10) measured over a horizon $[0, t]$ and citations measured over the same interval $[0, t]$. As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), application and grant year effects. Sample covers the entire 1840–2015 period. Last, we cluster the standard errors by the patent grant year. See main text for additional details on the specification and the construction of these variables.

Table A.4: Patent quality predicts citations (all patents)

log(1 + Forward citations, 2+ yr)	(1)	(2)	(3)	(4)
log(Patent impact (FS), 0-1yr)	1.111***	0.952***	0.858***	0.908***
	(15.53)	(17.38)	(14.85)	(16.03)
log(Patent novelty (BS), 0-5yr)	-1.107***	-0.908***	-0.825***	-0.873***
	(-15.33)	(-16.66)	(-14.14)	(-15.30)
log(1 + Forward citations, 0-1 yr)	0.658***	0.602***	0.513***	0.509***
	(33.39)	(36.27)	(39.55)	(36.59)
Observations	5959978	5922791	2788578	2495354
R^2	0.310	0.366	0.462	0.501
log(1 + Forward citations, 6+ yr)	(1)	(2)	(3)	(4)
log(Patent impact (FS), 0-5yr)	0.621***	0.681***	0.651***	0.718***
	(10.26)	(14.14)	(10.97)	(12.05)
log(Patent novelty (BS), 0-5yr)	-0.696***	-0.679***	-0.647***	-0.716***
	(-11.41)	(-13.84)	(-10.79)	(-11.89)
log(1 + Forward citations, 0-5yr)	0.623***	0.581***	0.536***	0.545***
	(36.96)	(36.49)	(34.57)	(32.88)
Observations	4931983	4897863	2333989	2079027
R^2	0.321	0.375	0.468	0.504
log(1 + Forward citations, 11+ yr)	(1)	(2)	(3)	(4)
log(Patent impact (FS), 0-10yr)	0.204***	0.371***	0.366***	0.398***
	(4.39)	(13.98)	(10.13)	(10.38)
log(Patent novelty (BS), 0-5yr)	-0.313***	-0.393***	-0.383***	-0.418***
	(-6.98)	(-14.90)	(-10.54)	(-10.87)
log(1 + Forward citations, 0-10yr)	0.582***	0.540***	0.514***	0.517***
	(38.63)	(37.62)	(33.14)	(32.32)
Observation	4119206	4087993	1971429	1751975
R^2	0.306	0.362	0.448	0.483
Grant Year FE	Y	Y		
Class		Y		
Assignee FE			Y	
Grant Year × Assignee FE				Y

Table reports the results of estimating equation (12) in the main text. The regression relates the log of (one plus) the number of patent citations over a horizon $[t, s]$ to our measures of patent quality (10) measured over a horizon $[0, t]$ and citations measured over the same interval $[0, t]$. As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), application and grant year effects. Sample covers the entire 1840–2015 period. Last, we cluster the standard errors by the patent grant year. See main text for additional details on the specification and the construction of these variables.

Table A.5: Patent impact and value — patent-level regressions

log KPSS value	(0-1)	(0-5)	(0-10)
Log patent quality	0.0015	0.0029**	0.0042***
	(1.58)	(2.48)	(2.83)
R^2	0.948	0.947	0.940
Breakthrough Patent (quality, top 5%)	0.0025	0.0051***	0.0046*
	(1.13)	(2.71)	(1.94)
R^2	0.948	0.947	0.940
log KPSS value	(0-1)	(0-5)	(0-10)
Log patent quality	0.0016	0.0026**	0.0032**
	(1.59)	(2.15)	(2.05)
Log forward citations	-0.0003	0.0017***	0.0039***
	(-0.68)	(2.85)	(4.15)
R^2	0.948	0.947	0.940
log KPSS value	(0-1)	(0-5)	(0-10)
Breakthrough Patent (quality, top 5%)	0.0026	0.0048**	0.0038
	(1.16)	(2.55)	(1.59)
Breakthrough Patent (citations, top 5%)	-0.0009	0.0033**	0.0065***
	(-0.64)	(2.42)	(3.33)
R^2	0.948	0.947	0.940
N	1923629	1723891	1407564
Controls:			
Class FE	Y	Y	Y
Firm Size (market cap)	Y	Y	Y
Firm Volatility	Y	Y	Y
Grant Year × Firm FE	Y	Y	Y

Table reports the results of estimating equation (13) in the main text. The regression relates the log of the Kogan et al. (2017) estimate of the market value of the patent to our (log) measures of patent quality, which combines the patent's impact and novelty, constructed in equation (10). As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level) and the interaction of firm (CRSP: permco) and grant year effects. The unit of observation is a patent. See Table 4 for a specification in which the unit of observation is at the firm-patent grant date level. We cluster the standard errors by the patent grant year. Independent variables are normalized to unit standard deviation. See main text for additional details on the specification and the construction of these variables.

Table A.6: Market Value and Patent Quality

$\log Q$	All Patenting Industries					Manuf	Health	HiTech
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Capitalized R&D / Assets	0.491***	1.314***	0.542***	0.951***	0.262***	1.236***	0.347***	0.192***
	(17.20)	(9.04)	(16.53)	(10.93)	(8.54)	(7.65)	(7.13)	(3.52)
Capitalized # Patents / Capitalized R&D	0.061***	0.208***	0.087***	0.166***	0.124***	0.182	10.266	7.250
	(5.48)	(6.82)	(9.98)	(8.94)	(9.13)	(0.64)	(1.16)	(0.67)
Patent portfolio, quality-weighted (novelty/impact)		0.602***		0.297***	0.103***	0.446***	0.075***	0.211***
		(7.02)		(6.89)	(6.17)	(5.08)	(2.80)	(3.65)
Patent portfolio, citation-weighted			0.287***	0.356***	0.184***	0.855***	0.126***	0.140***
			(14.81)	(9.52)	(8.99)	(7.89)	(2.97)	(4.63)
R&D=0 Dummy variable	-0.067***	-0.062***	-0.052***	-0.054***	-0.010	0.012	0.106**	0.166***
	(-5.84)	(-5.60)	(-4.72)	(-4.97)	(-0.92)	(0.88)	(2.24)	(5.71)
N	70,769	70,769	70,769	70,769	70,769	51,753	9,529	15,425
R^2	0.189	0.227	0.223	0.237	0.317	0.250	0.133	0.203
Year FE	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE					Y			

Table reports estimates of equation (18) in the text. The equation relates the logarithm of a firm's Tobin's Q to the stocks of R&D expenditure ($SRD_{f,t}$), number of patents ($SPAT_{f,t}$), patent citations ($SCITES_{f,t}$), and the patent quality measures ($Sq_{f,t}$) — constructed using a depreciation rate of $\delta = 15\%$. We restrict the sample to patenting firms, that is, firms that have filed at least one patent. We cluster standard errors by firm. All independent variables are normalized to unit standard deviation. Manufacturing includes SIC codes 2000-3999. Health is healthcare services, medical equipment, and pharmaceuticals (industries 11-13 in the Fama and French (1997) 49 industry classification). HiTech is telecommunications, computer hardware and software, and electronic equipment (industries 32, 35–37 in the Fama and French (1997) 49 industry classification).

Table A.7: Important Patents

Patent	Year	Inventor	Invention	Citations	Percentile Ranks						Source
					No Adjustment			Remove year FE			
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
1647	1840	Samuel F. B. Morse	Morse Code	2	0.00	0.00	0.29	0.01	0.65	0.81	Reference
3237	1843	Nobert Rillieux	Sugar Refining	0	0.24	0.00	0.00	0.77	0.65	0.44	Reference
3316	1843	Samuel F. B. Morse	Telegraphy Wire	0	0.78	0.00	0.00	0.98	0.65	0.44	Reference
3633	1844	Charles Goodyear	Vulcanized Rubber	3	0.97	0.00	0.38	0.99	0.65	0.88	Reference
4453	1846	Samuel F. B. Morse	Telegraph Battery	0	0.98	0.00	0.00	0.98	0.65	0.44	Reference
4750	1846	Elias Howe, Jr.	Sewing Machine	1	0.97	0.00	0.17	0.96	0.65	0.70	Reference
4834	1846	Benjamin Franklin Palmer	Artificial Limb	0	0.94	0.00	0.00	0.87	0.65	0.44	Reference
4848	1846	Charles T. Jackson	Anesthesia	0	0.90	0.00	0.00	0.75	0.65	0.44	Reference
4874	1846	Christian Frederick Schonbein	Guncotton	0	0.94	0.00	0.00	0.88	0.65	0.44	Reference
5199	1847	Richard M. Hoe	Rotary Printing Press	0	0.96	0.00	0.00	0.76	0.65	0.42	Reference
5711	1848	M. Waldo Hanchett	Dental Chair	1	1.00	0.00	0.17	0.98	0.65	0.70	Reference
5942	1848	John Bradshaw	Sewing Machine	0	1.00	0.00	0.00	0.97	0.65	0.44	Reference
6099	1849	Morey/Johnson	Sewing Machine	1	1.00	0.00	0.17	0.99	0.65	0.69	Reference
6281	1849	Walter Hunt	Safety Pin	0	1.00	0.00	0.00	0.97	0.65	0.42	Reference
6439	1849	John Bachelder	Sewing Machine	0	1.00	0.00	0.00	0.98	0.65	0.42	Reference
7296	1850	D.M. Smith	Sewing Machine	0	1.00	0.00	0.00	1.00	0.65	0.40	Reference
7509	1850	J. Hollen	Sewing Machine	0	1.00	0.00	0.00	1.00	0.65	0.40	Reference
7931	1851	Grover and Baker	Sewing Machine	0	1.00	0.00	0.00	0.99	0.65	0.40	Reference
8080	1851	John Gorrie	Ice Machine	0	0.99	0.00	0.00	0.27	0.65	0.40	Reference
8294	1851	Isaac Singer	Sewing Machine	0	1.00	0.00	0.00	0.98	0.65	0.40	Reference
9300	1852	Lorenzo L. Langstroth	Beehive	1	0.93	0.00	0.17	0.00	0.65	0.69	Reference
13661	1855	Isaac M. Singer	Shuttle Sewing Machine	1	0.98	0.00	0.17	0.03	0.63	0.63	Reference
15553	1856	Gail Borden, Jr.	Condensed Milk	0	0.99	0.00	0.00	0.78	0.64	0.34	Reference
17628	1857	William Kelly	Pneumatic Process of Making Steel	0	0.97	0.00	0.00	0.65	0.63	0.35	Reference
18653	1857	H.N. Wadsworth	Toothbrush	6	0.94	0.00	0.58	0.30	0.63	0.94	Reference
23536	1859	Martha Coston	System of Pyrotechnic Night Signals	1	0.89	0.00	0.17	0.82	0.64	0.58	Reference
26196	1859	James J. Mapes	Artificial Fertilizer	1	0.90	0.00	0.17	0.85	0.64	0.58	Reference
31128	1861	Elisha Graves Otis	Elevator	1	0.92	0.00	0.17	0.74	0.42	0.46	Reference
31278	1861	Linus Yale, Jr.	Lock	10	0.76	0.00	0.72	0.20	0.42	0.94	Reference
31310	1861	Samuel Goodale	Moving Picture Machine	0	0.98	0.00	0.00	0.96	0.42	0.18	Reference

Table A.7: Important Patents (cont)

Patent	Year	Inventor	Invention	Citations	Percentile Ranks						Source
					No Adjustment			Remove year FE			
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
36836	1862	Richard J. Gatling	Machine Gun	3	0.97	0.31	0.38	0.43	0.85	0.82	Reference
43465	1864	Sarah Mather	Submarine Telescope	0	0.96	0.00	0.00	0.02	0.41	0.40	Reference
46454	1865	John Deere	Plow	0	0.99	0.00	0.00	0.36	0.44	0.41	Reference
53561	1866	Milton Bradley	Board Game	2	1.00	0.00	0.29	1.00	0.49	0.81	Reference
59915	1866	Pierre Lallement	Bicycle	0	1.00	0.00	0.00	0.96	0.49	0.41	Reference
78317	1868	Alfred Nobel	Dynamite	4	0.88	0.00	0.46	0.27	0.64	0.92	Reference
79265	1868	C. Latham Sholes	Typewriter	1	0.96	0.00	0.17	0.81	0.64	0.69	Reference
79965	1868	Alvin J. Fellows	Spring Tape Measure	2	0.75	0.00	0.29	0.06	0.64	0.82	Reference
88929	1869	George Westinghouse	Air Brake	1	0.91	0.00	0.17	0.81	0.64	0.69	Reference
91145	1869	Ives W. McGaffey	Vacuum Cleaner	4	0.81	0.00	0.46	0.53	0.64	0.92	Reference
110971	1871	Andrew Smith Hallidie	Cable Car	1	0.76	0.00	0.17	0.71	0.42	0.67	Reference
113448	1871	Mary Potts	Sad Iron	3	0.72	0.00	0.38	0.63	0.42	0.87	Reference
127360	1872	J.P. Cooley, S. Noble	Toothpick-making machine	0	0.67	0.00	0.00	0.69	0.41	0.39	Reference
129843	1872	Elijah McCoy	Improvements in Lubricators for Steam-Engines	1	0.63	0.00	0.17	0.63	0.41	0.66	Reference
135245	1873	Louis Pasteur	Pasteurization	0	0.24	0.00	0.00	0.20	0.37	0.38	Reference
141072	1873	Louis Pasteur	Manufacture of Beer and Treatment of Yeast	1	0.15	0.00	0.17	0.11	0.37	0.66	Reference
157124	1874	Joseph F. Glidden	Barbed Wire	1	0.86	0.00	0.17	0.95	0.39	0.65	Reference
161739	1875	Alexander Graham Bell	Telephone	7	0.95	0.00	0.62	0.98	0.40	0.96	Reference
171121	1875	George Green	Dental Drill	2	0.52	0.31	0.29	0.54	0.84	0.79	Reference
174465	1876	Alexander Graham Bell	Telephone	6	0.99	0.50	0.58	1.00	0.92	0.95	Reference
178216	1876	Alexander Graham Bell	Telephone	0	0.97	0.00	0.00	0.99	0.42	0.38	Reference
178399	1876	Alexander Graham Bell	Telephone	2	0.98	0.31	0.29	0.99	0.85	0.79	Reference
186787	1877	Alexander Graham Bell	Electric Telegraphy	0	1.00	0.00	0.00	1.00	0.38	0.37	Reference
188292	1877	Chester Greenwood	Earmuffs	17	0.92	0.00	0.84	0.93	0.38	0.99	Reference
194047	1877	Nicolaus August Otto	Internal Combustion Engine	1	0.60	0.00	0.17	0.37	0.38	0.65	Reference
200521	1878	Thomas Alva Edison	Phonograph	12	0.94	0.50	0.77	0.87	0.92	0.98	Reference
201488	1878	Alexander Graham Bell	Telephone	2	1.00	0.00	0.29	1.00	0.36	0.78	Reference
203016	1878	Thomas Alva Edison	Speaking Telephone	15	1.00	0.50	0.82	1.00	0.92	0.99	Reference
206112	1878	Thaddeus Hyatt	Reinforced Concrete	0	0.83	0.00	0.00	0.48	0.36	0.36	Reference
220925	1879	Margaret Knight	Paper-Bag Machine	4	0.92	0.62	0.46	0.56	0.95	0.90	Reference

Table A.7: Important Patents (cont)

Patent	Year	Inventor	Invention	Citations	Percentile Ranks						Source
					No Adjustment			Remove year FE			
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
222390	1879	Thomas Alva Edison	Improvement in carbon telephones	16	1.00	0.00	0.83	1.00	0.37	0.99	Reference
223898	1880	Thomas Alva Edison	First Incandescent Light	20	1.00	0.00	0.87	1.00	0.43	0.99	Reference
224573	1880	Emile Berliner	Microphone	0	0.92	0.00	0.00	0.44	0.43	0.36	Reference
228507	1880	Alexander Graham Bell	Electric Telephone	3	1.00	0.50	0.38	1.00	0.93	0.85	Reference
237664	1881	Frederic E. Ives	Halftone Printing Plate	1	0.92	0.31	0.17	0.64	0.85	0.64	Reference
304272	1884	Ottmar Mergenthaler	Linotype	0	0.89	0.00	0.00	0.92	0.40	0.35	Reference
312085	1885	Edward J. Claghorn	Seat Belt	13	0.28	0.00	0.79	0.25	0.38	0.98	Reference
322177	1885	Sarah Goode	Folding Cabinet Bed	3	0.44	0.00	0.38	0.49	0.38	0.84	Reference
347140	1886	Elihu Thomson	Electric Welder	16	0.64	0.94	0.83	0.58	1.00	0.99	Reference
349983	1886	Gottlieb Daimler	Four Stroke Combustion Engine	4	0.99	0.00	0.46	0.99	0.39	0.89	Reference
371496	1887	Dorr E. Felt	Adding Machine	6	0.84	0.71	0.58	0.79	0.97	0.94	Reference
372786	1887	Emile Berliner	Phonograph Record	4	0.88	0.62	0.46	0.86	0.95	0.89	Reference
373064	1887	Carl Gassner, Jr.	Dry Cell Battery	3	0.73	0.00	0.38	0.59	0.38	0.84	Reference
382280	1888	Nikola Tesla	A. C. Induction Motor	2	0.93	0.31	0.29	0.95	0.84	0.76	Reference
386289	1888	Miriam Benjamin	Gong and Signal Chair for Hotels	0	0.66	0.00	0.00	0.55	0.40	0.34	Reference
388116	1888	William S. Burroughs	Calculator	3	0.80	0.00	0.38	0.78	0.40	0.84	Reference
388850	1888	George Eastman	Roll Film Camera	1	0.93	0.00	0.17	0.95	0.40	0.62	Reference
395782	1889	Herman Hollerith	Computer	1	0.45	0.31	0.17	0.31	0.85	0.61	Reference
400665	1889	Charles M. Hall	Aluminum Manufacture	2	0.86	0.31	0.29	0.89	0.85	0.76	Reference
415072	1889	William Starley, Herbert Owen	Tandem Bicycle	1	0.74	0.00	0.17	0.73	0.43	0.61	Reference
430212	1890	Hiram Stevens Maxim	Smokeless Gunpowder	0	0.65	0.00	0.00	0.75	0.45	0.34	Reference
430804	1890	Herman Hollerith	Electric Adding Machine	2	0.91	0.31	0.29	0.96	0.85	0.76	Reference
447918	1891	Almon B. Strowger	Automatic Telephone Exchange	81	0.74	0.00	0.98	0.91	0.46	1.00	Reference
453550	1891	John Boyd Dunlop	Pneumatic Tyres	1	0.75	0.31	0.17	0.92	0.85	0.61	Reference
468226	1892	William Painter	Bottle Cap	7	0.77	0.00	0.62	0.96	0.35	0.94	Reference
472692	1892	G.C. Blickensderfer	Typewriting Machine	4	0.23	0.31	0.46	0.58	0.84	0.88	Reference
492767	1893	Edward G. Acheson	Carborundum	12	0.07	0.00	0.77	0.33	0.44	0.98	Reference
493426	1893	Thomas Alva Edison	Motion Picture	1	0.56	0.00	0.17	0.92	0.44	0.60	Reference
504038	1893	Whitcomb L. Judson	Zipper	6	0.19	0.00	0.58	0.65	0.44	0.93	Reference
536569	1895	Charles Jenkins	Phantoscope	0	0.79	0.00	0.00	0.97	0.34	0.31	Reference

Table A.7: Important Patents (cont)

					Percentile Ranks						
Patent	Year	Inventor	Invention	Citations	No Adjustment			Remove year FE			Source
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
549160	1895	George B. Selden	Automobile	0	0.50	0.00	0.00	0.88	0.34	0.31	Reference
558393	1896	John Harvey Kellogg	Cereal	3	0.41	0.00	0.38	0.69	0.49	0.83	Reference
558719	1896	C.B. Brooks	Street Sweeper	2	0.37	0.50	0.29	0.65	0.93	0.75	Reference
558936	1896	Joseph S. Duncan	Addressograph	3	0.09	0.00	0.38	0.16	0.49	0.83	Reference
586193	1897	Guglielmo Marconi	Radio	4	0.76	0.71	0.46	0.89	0.97	0.88	Reference
589168	1897	Thomas A. Edison	Motion Picture Camera	0	0.36	0.00	0.00	0.46	0.48	0.31	Reference
608845	1898	Rudolf Diesel	Diesel Engine	8	0.67	0.00	0.66	0.73	0.47	0.95	Reference
621195	1899	Ferdinand Graf Zepplin	Dirigible	1	0.80	0.00	0.17	0.70	0.35	0.57	Reference
644077	1900	Felix Hoffmann	Aspirin	1	0.86	0.00	0.17	0.72	0.46	0.58	Reference
661619	1900	Valdemar Poulsen	Magnetic Tape Recorder	15	0.89	0.71	0.82	0.80	0.97	0.98	Reference
708553	1902	John P. Holland	Submarine	1	0.83	0.00	0.17	0.61	0.45	0.57	Reference
743801	1903	Mary Anderson	Windscreen Wiper	2	0.29	0.00	0.29	0.04	0.50	0.73	Reference
745157	1903	Clyde J. Coleman	Electric Starter	1	0.94	0.00	0.17	0.92	0.50	0.57	Reference
764166	1904	Albert Gonzales	Railroad Switch	0	0.77	0.00	0.00	0.68	0.50	0.30	Reference
766768	1904	Michael J. Owens	Automatic Glass Bottle Manufacturing	7	0.83	0.50	0.62	0.78	0.93	0.94	Reference
775134	1904	KC Gillette	Razor (with removable blades)	4	0.91	0.31	0.46	0.92	0.85	0.87	Reference
808897	1906	Willis H. Carrier	Air Conditioning	21	0.61	0.00	0.88	0.58	0.54	0.99	Reference
815350	1906	John Holland	Submarine	0	0.64	0.00	0.00	0.63	0.54	0.28	Reference
821393	1906	Orville Wright	Airplane	19	1.00	0.31	0.86	1.00	0.85	0.99	Reference
841387	1907	Lee De Forest	Triode Vacuum Tube	5	0.16	0.00	0.52	0.08	0.56	0.90	Reference
921963	1909	Leonard H. Dyer	Automobile Vehicle	0	0.58	0.00	0.00	0.71	0.54	0.26	Reference
942809	1909	Leo H. Baekeland	Bakelite	3	0.91	0.00	0.38	0.97	0.54	0.80	Reference
970616	1910	Thomas A Edison	helicopter (never flown)	2	0.98	0.00	0.29	0.99	0.58	0.71	Reference
971501	1910	Fritz Haber	Ammonia Production	1	0.99	0.31	0.17	1.00	0.85	0.54	Reference
1000000	1911	Francis Holton	Non-Puncturable Vehicle Tire	2	0.79	0.00	0.29	0.89	0.58	0.71	Reference
1005186	1911	Henry Ford	Automotive Transmission	3	0.55	0.00	0.38	0.65	0.58	0.80	Reference
1008577	1911	Ernst F. W. Alexanderson	High Frequency Generator	6	0.31	0.62	0.58	0.31	0.96	0.92	Reference
1030178	1912	Peter Cooper Hewitt	Mercury Vapor Lamp	1	0.89	0.00	0.17	0.96	0.55	0.54	Reference
1082933	1913	William D. Coolidge	Tungsten Filament Light Bulb	28	0.76	0.00	0.92	0.90	0.61	0.99	Reference
1102653	1914	Robert H. Goddard	Rocket	58	0.48	0.62	0.97	0.71	0.96	1.00	Reference

Table A.7: Important Patents (cont)

Patent	Year	Inventor	Invention	Citations	Percentile Ranks						Source
					No Adjustment			Remove year FE			
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
1103503	1914	Robert Goddard	Rocket Apparatus	29	0.39	0.62	0.92	0.59	0.96	0.99	Reference
1113149	1914	Edwin H. Armstrong	Wireless Receiver	11	0.86	0.31	0.75	0.96	0.85	0.97	Reference
1115674	1914	Mary P. Jacob	Brassiere	1	0.65	0.00	0.17	0.85	0.61	0.53	Reference
1180159	1916	Irving Langmuir	Gas Filled Electric Lamp	13	0.80	0.62	0.79	0.94	0.96	0.97	Reference
1203495	1916	William D. Coolidge	X-Ray Tube	11	0.69	0.62	0.75	0.88	0.96	0.96	Reference
1211092	1917	William D. Coolidge	X-Ray Tube	7	0.91	0.00	0.62	0.98	0.55	0.92	Reference
1228388	1917	Frederick C Bargar	Fire Extinguisher	2	0.51	0.00	0.29	0.74	0.55	0.68	Reference
1254811	1918	Charles F. Kettering	Engine Ignition	1	0.65	0.00	0.17	0.85	0.60	0.51	Reference
1279471	1918	Elmer A. Sperry	Gyroscopic Compass	9	0.93	0.00	0.69	0.98	0.60	0.95	Reference
1360168	1920	Ernst Alexanderson	Antenna	4	0.91	0.00	0.46	0.97	0.62	0.83	Reference
1394450	1921	Charles P Strite	Bread Toaster	2	0.60	0.00	0.29	0.82	0.62	0.66	Reference
1413121	1922	John Arthur Johnson	Adjustable Wrench	0	0.09	0.00	0.00	0.10	0.63	0.20	Reference
1420609	1922	Glenn H. Curtiss	Hydroplane	2	0.72	0.00	0.29	0.89	0.63	0.65	Reference
1573846	1926	Thomas Midgley, Jr.	Ethyl Gasoline	3	0.33	0.31	0.38	0.57	0.85	0.72	Reference
1682366	1928	Charles F. Brannock	Foot Measuring Device	4	0.22	0.00	0.46	0.37	0.51	0.78	Reference
1699270	1929	John Logie Baird	Television / TV	11	0.62	0.00	0.75	0.88	0.52	0.94	Reference
1773079	1930	Clarence Birdseye	Frozen Food	10	0.73	0.31	0.72	0.95	0.85	0.93	Reference
1773080	1930	Clarence Birdseye	Frozen Food	18	0.75	0.00	0.86	0.95	0.49	0.97	Reference
1773980	1930	Philo T. Farnsworth	Television	29	0.91	0.62	0.92	0.98	0.96	0.99	Reference
1800156	1931	Erik Rotheim	Aerosol Spray Can	30	0.76	0.31	0.93	0.97	0.85	0.99	Reference
1821525	1931	Nielsen Emanuel	Hair Dryer	11	0.13	0.00	0.75	0.55	0.50	0.93	Reference
1835031	1931	Herman Affel	Coaxial cable	15	0.46	0.77	0.82	0.90	0.98	0.96	Reference
1848389	1932	Igor Sikorsky	Helicopter	5	0.47	0.00	0.52	0.94	0.47	0.78	Reference
1867377	1932	Otto F Rohwedder	Bread-Slicing Machine	2	0.16	0.00	0.29	0.75	0.47	0.52	Reference
1925554	1933	John Logie Baird	Color Television	1	0.37	0.00	0.17	0.92	0.44	0.33	Reference
1929453	1933	Waldo Semon	Rubber	56	0.79	0.93	0.97	0.98	1.00	1.00	Reference
1941066	1933	Edwin H. Armstrong	FM Radio	0	0.38	0.00	0.00	0.93	0.44	0.10	Reference
1948384	1934	Ernest O. Lawrence	Cyclotron	96	0.27	0.00	0.99	0.87	0.42	1.00	Reference
1949446	1934	William Burroughs	Adding and Listing Machine	1	0.06	0.31	0.17	0.55	0.85	0.31	Reference
1980972	1934	Lyndon Frederick	Krokodil	1	0.76	0.00	0.17	0.98	0.42	0.31	Reference

Table A.7: Important Patents (cont)

Patent	Year	Inventor	Invention	Citations	Percentile Ranks						Source
					No Adjustment			Remove year FE			
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
2021907	1935	Vladimir K. Zworykin	Television	18	0.38	0.00	0.86	0.89	0.39	0.95	Reference
2059884	1936	Leopold D. Mannes	Color Film	15	0.20	0.50	0.82	0.59	0.92	0.93	Reference
2071250	1937	Wallace H. Carothers	Nylon	231	0.63	0.50	1.00	0.89	0.92	1.00	Reference
2087683	1937	PT Farnsworth	Image Dissector	1	0.68	0.00	0.17	0.92	0.36	0.23	Reference
2153729	1939	Ernest H. Volwiler	Pentothal (General Anesthetic)	2	0.81	0.00	0.29	0.96	0.33	0.38	Reference
2188396	1940	Waldo Semon	Rubber	59	0.97	0.00	0.97	1.00	0.32	0.99	Reference
2206634	1940	Enrico Fermi	Radioactive Isotopes	99	0.82	0.31	0.99	0.98	0.82	1.00	Reference
2230654	1941	Roy J. Plunkett	Teflon	49	0.43	0.89	0.96	0.93	0.99	0.99	Reference
2258841	1941	Jozsef Bir— Laszlo	Fountain Pen	20	0.02	0.77	0.87	0.23	0.97	0.94	Reference
2292387	1942	Hedwig Kiesler Markey	Secret Communication System	71	0.45	0.31	0.98	0.95	0.76	0.99	Reference
2297691	1942	Chester F. Carlson	Xerography	738	0.06	0.71	1.00	0.62	0.95	1.00	Reference
2329074	1943	Paul Muller	DDT - Insecticide	48	0.05	0.99	0.96	0.56	1.00	0.98	Reference
2390636	1945	Ladislo Biro	Ball Point Pen	27	0.34	0.97	0.92	0.79	1.00	0.95	Reference
2404334	1946	Frank Whittle	Jet Engine	35	0.13	0.94	0.94	0.23	0.99	0.97	Reference
2436265	1948	Allen Du Mont	Cathode Ray Tube	18	0.65	0.81	0.86	0.74	0.96	0.91	Reference
2451804	1948	Donald L. Campbell	Fluid Catalytic Cracking	9	0.65	0.50	0.69	0.74	0.81	0.77	Reference
2495429	1950	Percy Spencer	Microwave	15	0.22	0.87	0.82	0.21	0.98	0.89	Reference
2524035	1950	John Bardeen	Transistor	132	0.60	1.00	0.99	0.75	1.00	1.00	Reference
2543181	1951	Edwin H. Land	Instant Photography	116	0.44	0.99	0.99	0.63	1.00	1.00	Reference
2569347	1951	William Shockley	Junction Transistor	140	0.45	1.00	0.99	0.63	1.00	1.00	Reference
2642679	1953	Frank Zamboni	Resurfacing Machine	16	0.36	0.50	0.83	0.55	0.82	0.89	Reference
2668661	1954	George R. Stibitz	Modern Digital Computer	14	0.95	0.31	0.80	0.98	0.71	0.86	Reference
2682050	1954	Andrew Alford	Radio Navigation System	3	0.63	0.00	0.38	0.77	0.22	0.39	Reference
2682235	1954	Richard Buckminster Fuller	Geodesic Dome	86	0.48	0.77	0.99	0.60	0.94	0.99	Reference
2691028	1954	Frank B. Colton	First Oral Contraceptive	4	0.88	0.00	0.46	0.96	0.22	0.48	Reference
2699054	1955	Lloyd H. Conover	Tetracycline	38	0.92	0.98	0.95	0.97	1.00	0.97	Reference
2708656	1955	Enrico Fermi	Atomic Reactor	196	0.98	1.00	1.00	0.99	1.00	1.00	Reference
2708722	1955	An Wang	Magnetic Core Memory	76	0.70	0.97	0.98	0.78	1.00	0.99	Reference
2717437	1955	George De Mestral	Velcro	258	0.44	0.62	1.00	0.43	0.88	1.00	Reference
2724711	1955	Gertrude Elion	Leukemia-fighting drug 6-mercaptopurine	1	0.74	0.31	0.17	0.82	0.71	0.13	Reference

Table A.7: Important Patents (cont)

Patent	Year	Inventor	Invention	Citations	Percentile Ranks						Source
					No Adjustment			Remove year FE			
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
2752339	1956	Percy L. Julian	Preparation of Cortisone	11	0.84	0.62	0.75	0.88	0.88	0.81	Reference
2756226	1956	Brandl/Margreiter	Oral Penicillin	7	0.70	0.71	0.62	0.71	0.92	0.67	Reference
2797183	1957	Hazen/Brown	Nystatin	13	0.86	0.31	0.79	0.90	0.69	0.85	Reference
2816721	1957	R. J. Taylor	Rocket Engine	25	0.71	0.77	0.91	0.72	0.94	0.95	Reference
2817025	1957	Robert Adler	TV remote control	27	0.70	0.96	0.92	0.71	1.00	0.95	Reference
2835548	1958	Robert C. Baumann	Satellite	16	0.81	0.92	0.83	0.85	0.99	0.89	Reference
2866012	1958	Charles P. Ginsburg	Video Tape Recorder	30	0.77	0.97	0.93	0.81	1.00	0.96	Reference
2879439	1959	Charles H. Townes	Maser	24	0.72	0.96	0.90	0.77	0.99	0.94	Reference
2929922	1960	Arthur L. Shawlow	Laser	122	0.82	1.00	0.99	0.89	1.00	1.00	Reference
2937186	1960	Burckhalter/Seiwald	Antibody Labelling Agent	8	0.83	0.31	0.66	0.89	0.69	0.72	Reference
2947611	1960	Francis P. Bundy	Diamond Synthesis	62	0.71	0.00	0.98	0.77	0.19	0.99	Reference
2956114	1960	Charles P. Ginsburg	Wideband Magnetic Tape System	11	0.68	0.62	0.75	0.74	0.88	0.81	Reference
2981877	1961	Robert N. Noyce	Semiconductor Device-And-Lead Structure	152	0.96	1.00	1.00	0.98	1.00	1.00	Reference
3057356	1962	Greatbatch Wilson	Pacemaker	127	0.88	0.93	0.99	0.93	0.99	1.00	Reference
3093346	1963	Maxime A. Faget	First Manned Space Capsule-Mercury	19	0.89	0.87	0.86	0.93	0.97	0.91	Reference
3097366	1963	Paul Winchell	Artificial Heart	23	0.48	0.62	0.89	0.41	0.87	0.93	Reference
3118022	1964	Gerhard M. Sessler	Electret Microphone	39	0.70	0.50	0.95	0.70	0.80	0.97	Reference
3156523	1964	Glenn T. Seaborg	Americium (Element 95)	1	0.82	0.00	0.17	0.85	0.17	0.13	Reference
3174267	1965	Edward C Bopf, Deere & Co	Cotton Harvester	4	0.62	0.71	0.46	0.55	0.91	0.47	Reference
3220816	1965	Alastair Pilkington	Manufacture of Flat Glass	25	0.83	0.31	0.91	0.85	0.67	0.94	Reference
3287323	1966	Stephanie Kwolek, Paul Morgan	Kevlar	1	0.70	0.00	0.17	0.70	0.15	0.12	Reference
3478216	1969	George Carruthers	Far-Ultraviolet Camera	3	0.71	0.31	0.38	0.84	0.70	0.39	Reference
3574791	1971	Patsy Sherman	Scotchguard	81	0.66	0.84	0.98	0.82	0.97	0.99	Reference
3663762	1972	Edward Joel Amos Jr	Cellular Telephone	112	0.59	0.87	0.99	0.78	0.97	1.00	Reference
3789832	1974	Raymond V. Damadian	MRI	59	0.42	0.71	0.97	0.74	0.90	0.98	Reference
3858232	1974	William Boyle	Digital Eye	51	0.39	0.95	0.97	0.71	0.99	0.98	Reference
3906166	1975	Martin Cooper	Cellular Telephone	219	0.38	0.81	1.00	0.71	0.95	1.00	Reference
4136359	1979	Stephen Wozniak, Apple	Microcomputer	37	0.79	0.62	0.95	0.97	0.84	0.94	Reference
4229761	1980	Valerie Thomas	Illusion Transmitter	3	0.59	0.00	0.38	0.92	0.11	0.21	Reference
4237224	1980	Boyer/Cohen	Molecular chimeras	301	1.00	1.00	1.00	1.00	1.00	1.00	Reference

Table A.7: Important Patents (cont)

Patent	Year	Inventor	Invention	Citations	Percentile Ranks						Source
					No Adjustment			Remove year FE			
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
4363877	1982	Howard M. Goodman	Human Growth Hormone	51	0.99	0.71	0.97	1.00	0.88	0.96	Reference
4371752	1983	Gordon Matthews	Digital Voice Mail System	223	0.75	0.93	1.00	0.94	0.98	1.00	Reference
4399216	1983	Richard Axel	Co-transformation	482	0.99	0.97	1.00	1.00	0.99	1.00	Reference
4437122	1984	Walsh/Halpert	Bitmap graphics	178	0.99	0.90	1.00	1.00	0.97	1.00	Reference
4464652	1984	Apple	Lisa Mouse	112	0.70	0.98	0.99	0.89	1.00	0.99	Reference
4468464	1984	Boyer/Cohen	Molecular chimeras	109	1.00	0.50	0.99	1.00	0.74	0.99	Reference
4590598	1986	Gordon Gould	Laser	20	0.70	0.31	0.87	0.58	0.29	0.80	Reference
4634665	1987	Richard Axel	Co-transformation	183	0.99	0.62	1.00	0.99	0.77	1.00	Reference
4683195	1987	Kary B. Mullis	polymerase chain reaction	2884	0.97	1.00	1.00	0.97	1.00	1.00	Reference
4683202	1987	(several)	polymerase chain reaction	3328	0.95	1.00	1.00	0.94	1.00	1.00	Reference
4736866	1988	Leder/Stewart	transgenic (genetically modified) animals	370	1.00	0.81	1.00	1.00	0.90	1.00	Reference
4744360	1988	Patricia Bath	Cataract Laserphaco Probe	81	0.94	0.81	0.98	0.91	0.90	0.98	Reference
4799258	1989	Donald Watts Davies	Packet-switching technology	153	0.96	0.95	0.99	0.95	0.98	0.99	Reference
4816397	1989	Michael A. Boss	recombinant antibodies	567	0.97	0.81	1.00	0.98	0.90	1.00	Reference
4816567	1989	Shmuel Cabilly	immunoglobulins	1785	0.99	0.77	1.00	0.99	0.87	1.00	Reference
4838644	1989	Ellen Ochoa	Recognizing Method	22	0.94	0.81	0.89	0.92	0.90	0.81	Reference
4889818	1989	(several)	polymerase chain reaction	366	0.98	0.99	1.00	0.98	1.00	1.00	Reference
4965188	1990	(several)	polymerase chain reaction	1176	0.97	0.99	1.00	0.98	1.00	1.00	Reference
5061620	1991	Ann Tsukamoto	Method for isolating the human stem cell	252	0.99	1.00	1.00	1.00	1.00	1.00	Reference
5071161	1991	Geoffrey L Mahoon	Airbag	23	0.81	0.96	0.89	0.67	0.98	0.81	Reference
5108388	1992	Stephen L. Troke	Laser Surgery Method	125	0.97	0.00	0.99	0.97	0.04	0.99	Reference
5149636	1992	Richard Axel	Co-transformation	6	0.99	0.31	0.58	0.99	0.24	0.36	Reference
5179017	1993	Richard Axel	Co-transformation	131	1.00	0.96	0.99	1.00	0.98	0.99	Reference
5184830	1993	Saturo Okada, Shin Kojo	Compact Hand-Held Video Game System	201	0.98	0.98	1.00	0.98	0.99	1.00	Reference
5194299	1993	Arthur Fry	Post-It Note	76	0.87	0.00	0.98	0.73	0.04	0.97	Reference
5225539	1993	Gregory P. Winter	Chimeric, humanized antibodies	671	1.00	0.99	1.00	1.00	1.00	1.00	Reference
5272628	1993	Michael Koss	Core Excel Function	94	0.99	0.92	0.99	0.99	0.95	0.98	Reference
5747282	1998	Mark H. Skolnick	BRCA1 gene	15	0.98	0.71	0.82	0.97	0.72	0.67	Reference
5770429	1998	Nils Lonberg	human antibodies from transgenic mice	248	0.91	0.84	1.00	0.61	0.84	1.00	Reference
5837492	1998	(several)	BRCA2 gene	5	0.95	0.00	0.52	0.83	0.01	0.26	Reference

Table A.7: Important Patents (cont)

Patent	Year	Inventor	Invention	Citations	Percentile Ranks						Source
					No Adjustment			Remove year FE			
					Quality	Citations		Quality	Citations		
				(total)	(0-5)	(0-5)	(total)	(0-5)	(0-5)	(total)	
5939598	1999	(several)	Transgenic mice	262	1.00	0.31	1.00	1.00	0.09	1.00	Reference
5960411	1999	Hartman/Bezos/Kaphan/Spiegel	1-click buying	1387	1.00	1.00	1.00	1.00	1.00	1.00	Reference
6230409	2001	Patricia Billings	Geobond	7	0.86	0.62	0.62	0.75	0.33	0.46	Reference
6285999	2001	Larry Page	Google Pagerank	689	0.98	1.00	1.00	0.99	1.00	1.00	Reference
6331415	2001	Shmuel Cabilly	Antibody molecules	243	0.98	0.00	1.00	0.99	0.01	1.00	Reference
6455275	2002	Richard Axel	Co-transformation	7	0.97	0.31	0.62	0.98	0.12	0.52	Reference
6574628	2003	Robert Kahn, Vinton Cerf	Packet-Switching Knowbot	61	0.99	0.95	0.97	1.00	0.96	0.98	Reference
6955484	2005	Nicholas D. Woodman	Harness system for attaching camera to user	15	0.59	0.84	0.82	0.78	0.89	0.87	Reference
6985922	2006	Janet Emerson Bashen	LinkLine	47	0.81	0.95	0.96	0.93	0.98	0.98	Reference