AER: Insights 2021, 3(3): 303–320
https://doi.org/10.1257/aeri.20190499

Measuring Technological Innovation over the Long Run[†]

By Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy*

We use textual analysis of high-dimensional data from patent documents to create new indicators of technological innovation. We identify important patents based on textual similarity of a given patent to previous and subsequent work: these patents are distinct from previous work but related to subsequent innovations. Our importance indicators correlate with existing measures of patent quality but also provide complementary information. We identify breakthrough innovations as the most important patents—those in the right tail of our measure—and construct time series indices of technological change at the aggregate and sectoral levels. Our technology indices capture the evolution of technological waves over a long time span (1840 to the present) and cover innovation by private and public firms as well as nonprofit organizations and the US government. Advances in electricity and transportation drive the index in the 1880s, chemicals and electricity in the 1920s and 1930s, and computers and communication in the post-1980s. (*JEL* C43, N71, N72, O31, O33, O34)

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, varies significantly over time and across sectors. 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 the degree of technological progress over time. Our goal is to fill this gap by constructing indices of technological progress at the aggregate and sectoral levels that are consistently available over long periods of time.

Patent statistics are a useful starting point (Griliches 1998). 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

*Kelly: Yale School of Management and NBER (email: bryan.kelly@yale.edu); Papanikolaou: Kellogg School of Management and NBER (email: d-papanikolaou@kellogg.northwestern.edu); Seru: Stanford GSB, Hoover Institution, and NBER (email: aseru@stanford.edu); Taddy: Amazon (email: mataddy@gmail.com). Amy Finkelstein was coeditor for this article. We thank Pierre Azoulay, Nicholas Bloom, Diego Comin, Carola Frydman, Erik Hurst, Pete Klenow, Kyle Jensen, Chad Jones, Chad Syverson, 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.

[†] Go to https://doi.org/10.1257/aeri.20190499 to visit the article page for additional materials and author disclosure statement(s).

citations a patent receives in the future have been used to address this obstacle, these metrics have significant disadvantages. First, patent citations are consistently recorded by the US Patent and Trademark Office (USPTO) in patent documents only after 1947. Prior to 1947, citations sometimes appear inside the text of the patent document, but they are much less common than in the postwar era.[1] Second, citations tend to take discrete values (the median post-1947 patent has three citations in a ten-year forward window). Third, citations rely on the discretion of the inventor or the patent examiner in choosing which prior patents to cite, or on whether they are aware of the existence of closely related patents.[2]

Given these shortcomings, we instead propose a new indicator of patent importance that is similar in spirit and can be constructed by analyzing the text of patent documents. Our indicators require no other inputs besides the text of the patent document; hence they are consistently available for the entire history of US patents, spanning nearly two centuries of innovation (1840–2010). Further, unlike citations, there is limited discretion in how much of the patent document is written (besides the claims), since it pertains to the technical description of the innovation.

We start by leveraging natural language processing techniques 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 an important patent as one whose content is distinct from prior patents (is novel) but similar to future patents (is impactful). These innovations represent distinct improvements in the technological frontier and 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 of future citations.

Several tests confirm the validity of our measure of patent importance. First, we identify a set of major technological breakthroughs of the nineteenth and twentieth centuries using the help of research assistants. Our indicators of patent significance perform quite well in identifying these major technological breakthroughs. Next, focusing on the post-1947 sample when citations data are available, we find 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 more timely assessment of a patent's quality than citation counts. Last, we relate our indicator to measures of private values. Though we view our indicators as more likely to be measuring the scientific value of a patent, 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, Jaffe, and Trajtenberg 2005; Kogan et al. 2017). We find that our quality

[1] 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's 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 received a total of ten citations. These issues are not confined to the pre-1947 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.

[2] As an example, patent 6,368,227 for "method of swinging on a swing," issued to Steven Olson (age five) 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.

indicator is significantly correlated with the Kogan et al. (2017) measure of a patent's economic value.

Next, we construct time series indices that describe the arrival intensity of breakthrough innovations—at the aggregate and sectoral levels—by counting the number of patents each year whose importance is in the top decile of our importance measure (breakthrough patents). Our aggregate innovation index uncovers three major technological waves: the second Industrial Revolution (mid- to late nineteenth century), the 1920s and 1930s, and the post-1980 period. Inventions related to electricity were important in the late nineteenth and early twentieth centuries. Innovations in agriculture played an important role in the beginning of the twentieth 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.

I. Measuring Patent Similarity

Here, we discuss how to measure the similarity between pairs of patent documents, aggregate these similarities into a patent-level measure of importance, and construct a time series index of breakthrough innovations.

Data.—We build our dataset from two sources: the USPTO and Google's patent search engine. Our final dataset includes the full text of over nine million patents over the period 1840–2010. The online Appendix provides additional details on our data collection process as well as the conversion of unstructured patent text data into a numerical format suitable for statistical analysis.

Definition: 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" (*TFIDF*) 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

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

Note, *IDF* measures the informativeness of term w by underweighting 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. Younge and Kuhn (2016) use this method to measure the pairwise patent-to-patent similarity across a large subset of the USPTO patents.

For our purposes, this weighting scheme is not ideal, since 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 (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 twentieth century. Standard *IDF* would sharply deemphasize this term in the *TFIDF* vector representing Tesla's patent because so many patents subsequently used this phrase so intensively. Therefore, *TFIDF* would give a misleading, and quite inverted, portrayal of the patent's importance.

We therefore develop a modified version of the traditional *TFIDF* measure. In place of (3), we instead construct a retrospective version of inverse document frequency. We define the "backward-*IDF*" of term w for patent p, (denoted by $BIDF_{wp}$) as the log frequency of documents containing w in any patent granted *prior* to patent p:

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

This frequency measure evolves as a term becomes more or less widely used over time, reflecting 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 General Motor'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 General Motors would be greatly downweighted 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. Given our goal of quantifying the impact of patents on future technological innovations, we calculate pairwise similarity by applying the *BIDF* corresponding to the *earlier* of the two patents to *both* patent counts. Thus, to calculate the similarity between the patent pair in this

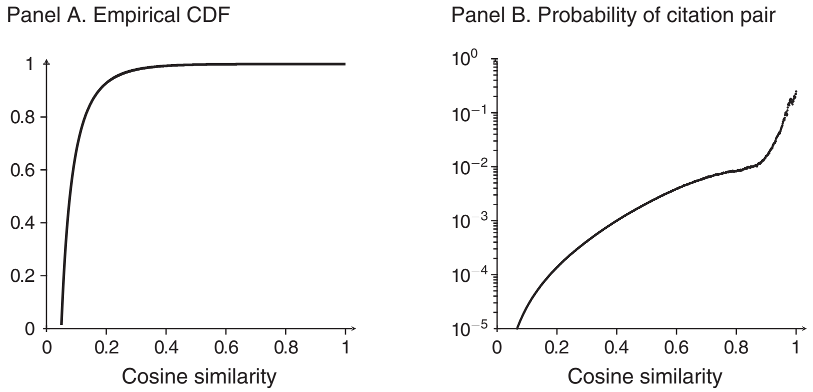


Figure 1. Pairwise Similarity and Citation Linkages

Notes: 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 i 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. Specifically, 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 $E[\mathbf{1}_{i,j} | \rho_{i,j}]$, where $\mathbf{1}_{i,j}$ is a dummy variable that takes the value one if patent j cites patent i (where patent i is led prior to patent j). For computational reasons, we exclude similarity pairs with $\rho_{i,j} \leq 5$ percent. Figure uses data only post 1945, since citations were not consistently recorded prior to that year.

Tesla/General Motors 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

$$(5) \qquad TFBIDF_{w,i,t} = TF_{w,i} \times BIDF_{w,t}, \qquad t \equiv \min(i,j)$$

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:

$$(6) \qquad V_{i,t} = \frac{TFBIDF_{i,t}}{\|TFBIDF_{i,t}\|}.$$

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

$$(7) \qquad \rho_{i,j} = V_{i,t} \cdot V_{j,t}.$$

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.

Descriptive Statistics: Panel A of Figure 1 plots the distribution of our similarity score across patent pairs that are 0–20 years apart. We see that 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 percent, whereas the average similarity score is 10.2 percent. In the right tail, the ninetieth and ninety-fifth percentiles of similarity scores are 17.6 percent and 22.9 percent, respectively. In network terminology, the patent system's connectivity is sparse.

Citation linkages provide external validation for assessing the text-based similarity measure $\rho_{i,j}$. Panel B examines whether patent pairs with high $\rho_{i,j}$ are more likely to be linked by a citation. Indeed, we see that the likelihood that patent *j* cites the earlier patent *i* is monotonically increasing in the similarity $\rho_{i,j}$ between the two patents.

Examples: Figure 2 provides a few examples of patents' similarity networks. To simplify the presentation, and also illustrate the advantages of our method in the early parts of the sample, we focus on four patents from the nineteenth 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 percent or greater with the focal patent.

The patent in panel A (4,750) is one of the first patents associated with the sewing machine, issued in 1846 to Elias Howe Jr. The patent is for the lockstitch, a manufacturing process still in use today. This patent is not significantly connected to any prior patents. By contrast, it is relatively closely related to 16 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 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 in panel B (493,426) is one of the earliest patents associated with cinematography. The patent is issued to Thomas Edison for exhibiting "photographs of moving objects" and is one of the earliest film projectors. The patent is highly similar to 2 prior patents and 12 subsequent patents, filed within five years. Most of the subsequent patents are related to cinematography. Among them, three are for a "kinetographic" camera, one of the early precursors of the film camera.

The patent in panel C (161,739) is one of the early patents issued to Graham Bell and eventually led to the invention of the telephone. We can see that it is quite similar to four prior "telegraph" patents filed over the previous five years. It is also related to 11 patents filed over the next five years, one of which is Bell's famous "telephone" patent (174,465). Last, the patent in panel D is a random patent (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.

II. Important Patents

Novel patents are those that are conceptually distinct from their predecessors and therefore rely less on prior art. Impactful patents are those that influence future scientific advances, manifested as high similarity with subsequent innovations. The main idea in this paper is that an important patent is one that is *both novel and impactful*.

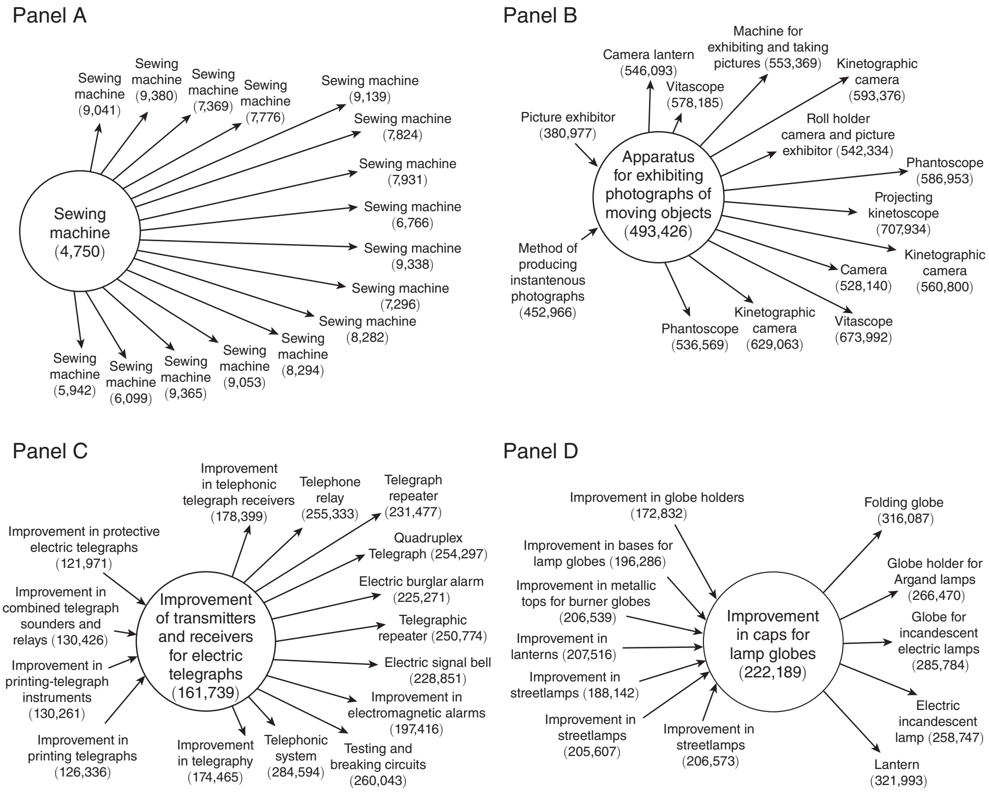

Panel A
Sewing machine (4,750)
Sewing machine (9,041)
Sewing machine (9,380)
Sewing machine (7,369)
Sewing machine (7,776)
Sewing machine (9,139)
Sewing machine (7,824)
Sewing machine (7,931)
Sewing machine (6,766)
Sewing machine (9,338)
Sewing machine (7,296)
Sewing machine (8,282)
Sewing machine (8,294)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

FIGURE 2. EXAMPLES OF SIMILARITY NETWORKS

Notes: Figure displays the similarity networks for four patents: the patent for the first sewing machine (panel A), one of the earlier patents for moving pictures (panel B), one of the early patents that led to the telephone (panel C), and a randomly chosen patent from the 1800s (panel D). In plotting the similarity links, we restrict attention to patent pairs led at most five years apart and with a cosine similarity greater than 50 percent.

A. *Definition*

We measure a patent's novelty as its dissimilarity with the existing patent stock at the time it was filed. We start from a measure of "backward similarity," defined 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 consider a backward-looking window of $\tau = 5$ years in our baseline importance measure—henceforth denoted by BS_j.

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 a scientific regime shift that was catalyzed by a predecessor patent. The "alternating current" example highlights this difference. Nikola Tesla's patent has 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 high similarity both backward and forward. Majorly important patents—those with a large influence on future technologies and that deviate from the status quo—are more likely to represent scientific breakthroughs.

The distinction between these two patents emerges when we compare forward versus backward similarity for a given patent. That is, our indicator of patent importance combines forward and backward similarity to identify patents that are both novel and impactful:

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

This indicator attaches higher scientific value to patents that are both novel relative to their predecessors and influential for subsequent research.

Our indicator of patent importance largely follows the logic behind indicators based on future citations. Specifically, the numerator in (10) is the total similarity with future patents—which is directly analogous to the sum of future citations. The denominator scales the forward similarity score by the novelty of the patent—since, presumably, patents should be citing the earliest relevant prior patents.[3]

The key advantage of our indicator is that it relies only on the text of the patent document and is therefore broadly available. However, it also has limitations. Existing computational constraints limit the window over which we can compare patents; hence, our measure may underweigh innovations whose impact took time to materialize. Further, our algorithm is reliant on the digitization quality of the patent document; patents with inaccurate text will be less similar to other (prior or subsequent) patents. Last, our algorithm in identifying impactful patents may be affected by shifting within-field propensity to patent (for instance, the recent rise of

[3] In contemporaneous work, Ashtor (2019) also constructs a measure of quality using textual similarity (estimated using latent semantic analysis). In contrast to this paper, Ashtor (2019) identifies high-quality patents as those that have high similarity to contemporaneous and prior patents, use only the claims portion of the document, and do not use any information on future similarity.

software patents). Such patents may appear to be impactful—as they are related to subsequent similar patents—and are more likely to be cited. Our measure shares this potential shortcoming with patent citations. One possibility is to extend our definition of impact (9) to include nonpatent literature, which we leave to future work.

B. *Validation*

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

Historically Significant Patents.—We compile a list of approximately 250 "historically significant patents" based on online lists. For instance, the USPTO has a "Significant Historical Patents of the United States" list. Our list targets major 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 online sources is provided in online Appendix Table A.1.

For each of these radical inventions, we report their percentile rank in terms of our importance measure (10); for instance, a value of 0.90 indicates that the patent is in the top 10 percent of most important patents. We compute a patent's rank using three approaches. First, we use the unconditional distribution. Second, we rank patents after subtracting the mean importance measure within each cohort (issue year). Removing cohort fixed effects helps eliminate factors that affect patents symmetrically, such as shifts in language or variation in the quality of the digitized patent documents. 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.

Overall, these 250 patents score highly in terms of our importance indicator. The mean rank of these patents is 0.74 when using the unconditional distribution, or 0.78 when performing either adjustment. That said, our importance measure does miss some important inventions, such as pasteurization and Morse code, which our importance measure ranks at the bottom 20 percent.

One way to assess the performance of our measure is to compare how these patents rank in terms of their citations. Since many of these patents are filed during the period when citation data are not broadly available, we extend the horizon for citations and measure them using the entire sample. Naturally this skews the comparison in favor of citations since they are measured over a significantly longer horizon than our indicators.

We find that our importance measure moderately outperforms citations: the average rank assigned to these important patents is 0.74, compared to 0.54 for citations when citations are measured using the full sample. The difference shrinks when these indicators are demeaned using year fixed effects but is not fully eliminated—0.78 for importance versus 0.75 for citations. Removing time fixed effects leads to similar results as comparing patents within cohorts (mean rank 0.78 for importance versus 0.69 for citations). Online Appendix Figure A.1 summarizes these findings. We conclude that our text-based importance indicators are considerably more informative than patent citations in comparing patents across different cohorts, especially once we consider that our importance indicator can be computed using only ten years as opposed to several decades—or centuries—of data in many cases.

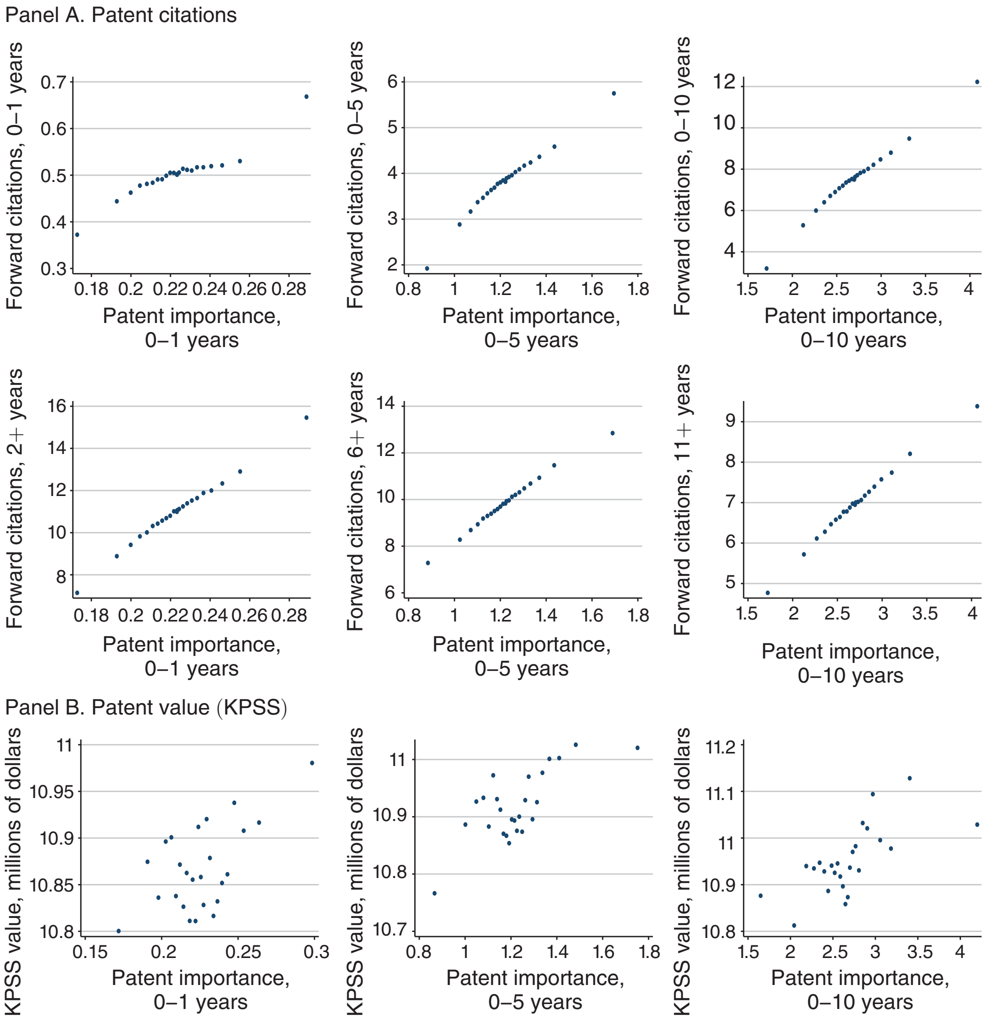
Patent Citations.—We next investigate the relation between our importance measure and patent citations, a commonly used metric of impact. We focus on patents issued after 1947, as this is the period when citations are consistently recorded by the USPTO. Panel A of Figure 3 illustrates the correlation using binned scatter plots; online Appendix Table A.2 reports the corresponding regression estimates.

The first row of Figure 3, panel A, reveals a strong positive contemporaneous correlation between patent importance and forward citations. Specifically, we first consider forward windows of $\tau = 1$, 5, and 10 years for both citations and importance. The correlation is consistently economically significant across horizons τ. Comparing two patents in the same technology class that are issued to the same entity in the same year, we find that increasing the importance measure from the median to the ninetieth percentile results in 1.5 additional citations, relative to the median of 3 citations, when importance and citations are measured over the next ten years after the patent application is filed.

One way to examine the additional information content of our measure relative to citation counts measured over the same horizon is to examine whether it predicts *future* patent citations. The second row of panel A shows that this is indeed the case. We plot the predictive relation between our text-based quality measured in the $0 - \tau$ year window after filing, versus all citations in years $\tau + 1$ and beyond. We control for the number of citations over the same period for which importance is measured. In all cases, we find a strong positive association between our near-term quality measure and long-term future citations. Comparing two patents in the same class, issued to the same entity in the same year, we see that an increase in the patent importance from the median to the ninetieth percentile predicts 20–25 percent more future citations relative to the median. The results strongly suggest that our importance measure incorporates information faster than forward citations.

Estimates of Market Value.—We next discuss the relation between patent importance 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 importance measure is designed to ascertain the scientific importance of the patent. The relation between 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 (see, for example, Abrams, Akcigit, and Grennan 2013). With that caveat in mind, we next examine the relation between our importance measure and the estimate of patent value 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.

Panel B of Figure 3 graphically illustrates this correlation; online Appendix Table A.3 reports the corresponding regression estimates. Patent importance is



Panel A. Patent citations

Panel B. Patent value (KPSS)

FIGURE 3. PATENT IMPORTANCE: VALIDATION

Notes: Figure plots the relation between our importance measure (the ratio of forward to backward patent similarity) to the number of forward citations (panel A) or the Kogan et al. (2017) estimate of patent value (panel B). In the first row of panel A, both the patent importance measure and forward citations are measured over the same horizon. In the second row of panel A, we plot the predictive relation between our importance measure and future citations; for these specifications, we also control for the number of citations the patent has received over the same horizon that our importance measure is computed. To construct the figure, we group observations into 25 bins (cut off at every other percentile of the quality distribution). Within each bin, we average citation and text-based importance measures after controlling for technology class and assignee-by-grant year fixed effects. See online Appendix Tables A.2 and A.3 for the corresponding regression tables.

positively and statistically significantly correlated to the KPSS estimate of market value. Focusing on two patents in the same class that are issued to the same firm in the same year, increasing the importance measure from the median to the ninetieth percentile results in a 0.23–0.47 percent increase in patent values. Though these estimates may appear relatively modest, they are comparable in magnitude to the relation between patent values and forward citations (see, for example, KPSS).

Further, online Appendix Table A.3 shows that the correlation remains significant once we include as additional controls the number of forward citations the patent receives over the same horizon that importance is measured—which supports the conclusion that our measure incorporates additional information to patent citations.

III. Indices of Technological Progress

Our goal here is to construct indices of breakthrough innovations—that is, innovations that are in the right tail of our importance measure—that span the USPTO sample (1840–2010).

A. *Construction*

One challenge is that time series fluctuations in (10) are mechanically affected by 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, we adjust (10) by removing patent cohort issue year fixed effects. After this adjustment, we define a "breakthrough" patent as one that falls in the top 10 percent of the unconditional distribution of importance estimated as the ratio of 10-year forward to 5-year backward similarity. We then construct a time series index as the number of breakthrough inventions granted in each year, divided by US population. The implicit assumption in our methodology is that shifts in language are likely to symmetrically affect all patents and will thus be absorbed by the fixed effect—which is mainly identified by the nonbreakthrough patents. We also construct indices of innovation at the sector level using the probabilistic mapping between patent technology classifications (CPC) and industry classifications constructed by Goldschlag, Lybbert, and Zolas (2016).

To validate our methodology, we show that our technology indices are significantly related to measured productivity, both at the aggregate as well as sectoral levels. As we discuss in online Appendix Section F, a one standard deviation increase in our index is associated with 0.5 percent to 2 percent higher annual productivity growth over the next 10 years. Similarly, sectors that have breakthrough innovations experience faster growth in productivity than sectors that do not: a one standard deviation increase in our innovation index is associated with 1 percent higher annual productivity growth over the next 5 years.

B. *Aggregate Index*

Panel A of Figure 4 plots the resulting time series of breakthroughs per capita. Our index identifies three main innovation waves, lasting from 1870 to 1880,1920 to 1935, and 1985 to the present. The first peak corresponds to the beginning of the second industrial revolution, which saw technological advances such as the telephone and electric lighting and improvements in railroads. 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.

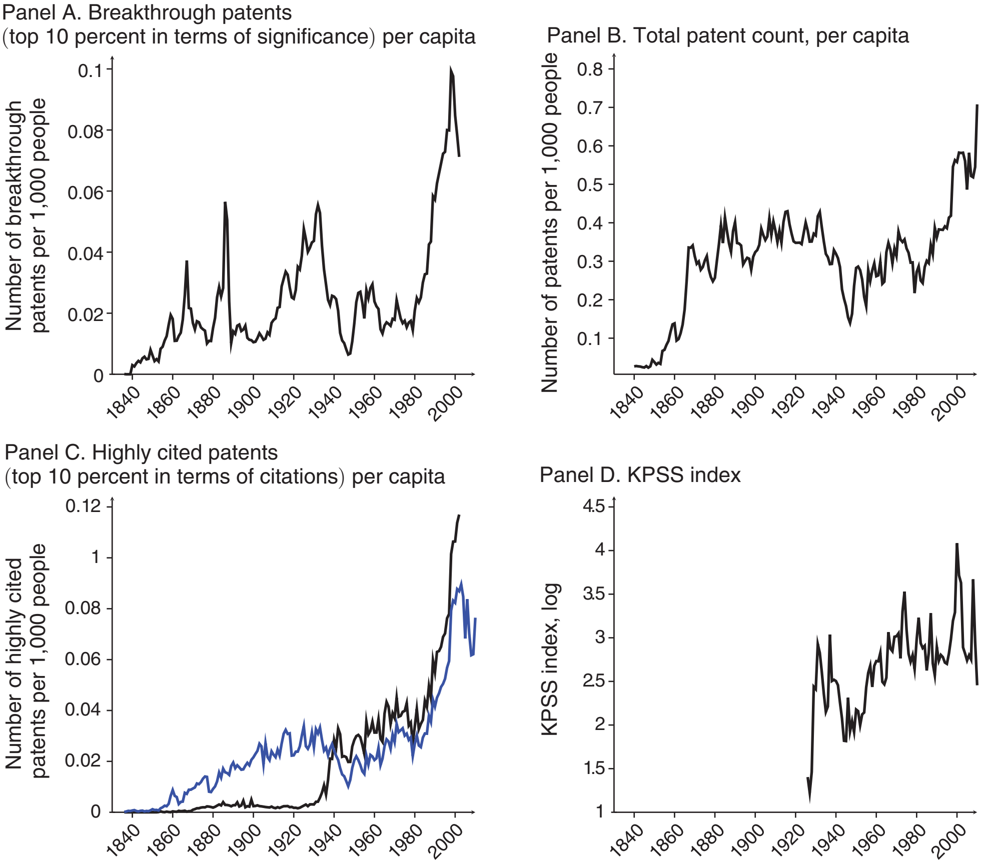


Figure 4. Technological Innovation over the Long Run: New versus Existing Indicators

Notes: Panel A plots the number of breakthrough patents per capita. Breakthrough patents are those that fall in the top 10 percent of the unconditional distribution of our importance measure, where importance is defined as the ratio of the 10-year forward to the 5-year backward similarity, net of year fixed effects. Panel B plots the total number of patents, scaled by population. In panel C we plot, in black (blue), the number of patents that fall in the top 10 percent of the unconditional distribution of forward citations—measured over the next 10 years (entire sample), net of year fixed effects—again scaled by US 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).

Constructing an innovation index has proven challenging in the past. In one approach, Shea (1999) constructs an index of per capita patent counts, which is plotted in panel B. Patents per capita is essentially flat from 1870–1930, dips from 1930–1980, and displays a 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. One common adjustment to simple patent counts is to weigh patents by their forward citations. Panel C (black line) plots the resulting time series when our index methodology is instead constructed from ten-year forward citations. Due to the limitations of citation data, this series essentially identifies no innovation prior to 1940s. Only when citations are measured over the entire sample (blue line) does the index take nonzero values in the pre-WWII 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, for example, Gordon 2017), this pattern mostly reflects the limitations of forward citations as a measure of patent importance.

KPSS 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 is confined to the universe of publicly traded firms, thereby omitting innovations by private firms, nonprofit institutions, and the government. Moreover, it is not available prior to 1927, since information on stock prices is readily available only after this year.

C. *Sectoral Indices*

Figure 5 plots time series indices of industry innovation at the three-digit NAICS level. We see that the origin of breakthrough patents has varied considerably over time. In the 1840–1870 period, we see that the most important inventions took place in engineering and construction, consumer goods, and manufacturing. An example of an important invention in construction according to our importance measure is the "Bollman Bridge" (patent 8,624), the first successful all-metal bridge design widely used for railroads. Other important advances in this period occur in textiles; examples include various versions of sewing and knitting machines (patents 7,931; 7,296; 7,509; and 60,310).

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 relating to the production of electric light (203,844; 210,380; 215,733; 210,213; 200,545; and 218,167). 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 top 1 percent.[4]

Another industry that accounted for a significant share of important patents during the 1860–1910 period is transportation. Many of the patents that fall in the top 1 percent include improvements in railroads (207,538; 218,693; 422,976; and 619,320), and their electrification (178,216; 344,962; 403,969; and 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 importance indicator (1,107,231; 1,279,127; 1,307,133; and 1,307,134) as well as patents related to air balloons and the Zeppelin (678,114 and 864,672). Innovations in construction methods continue to play a role in this period, such as those that are related to the use of concrete (618,956; 647,904; 764,302; 654,683; 747,652; and 672,176) in the construction of buildings, roads, and pavements.

In the first half of the twentieth century, chemistry emerges as a major generator of important patents, many describing inventions of plastic compounds. Among our breakthrough inventions is the patent for bakelite (942,699), the world's first fully

[4] Patents 161,739; 174,465; 178,399; 186,787; 201,488; 213,090; 220,791; 228,507; 230,168; 238,833; 474,230; 203,016; and 222,390.

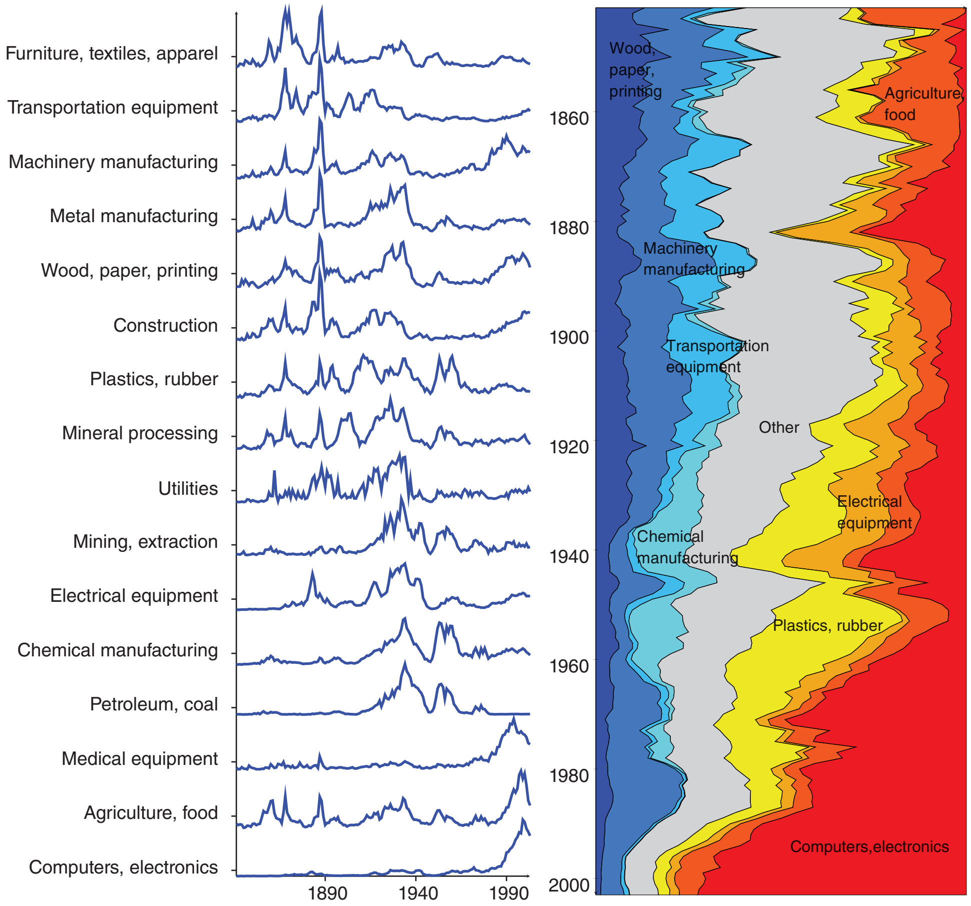


Figure 5. Breakthrough Innovation across Industries

Notes: 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 percent of our baseline importance measure (defined as the ratio of the 10-year forward to the 5-year backward similarity) net of issue year fixed effects. We construct industry indices using the CPC4 to NAICS crosswalk constructed by Goldschlag, Lybbert, and Zolas (2016).

synthetic plastic. This innovation opened the floodgates to a torrent of now-familiar synthetic plastics, including the invention in the 1930s of 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 score in the top 5 percent. Other important chemistry patents in the 1950s include drugs such as 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).

The 1950s are marked by the harnessing of nuclear energy for civilian purposes. Enrico Fermi's patents on the development of the nuclear reactor all score highly.[5] Subsequent to the 1950s, a large fraction of the important patents identified by

[5] Patents 2,206,634; 2,836,554; 2,524,379; 2,852,461; 2,708,656; 2,768,134; 2,780,595; 2,798,847; 2,807,581; 2,807,727; 2,813,070; 2,837,477; and 2,931,762.

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 developed some of the major innovations in computing. Xerox, for example, produced 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 emerged among the set of breakthrough patents.[6] Improvements in genetics comprise a significant fraction of the most important patents in the 1980–2000 period. A few early examples that fall in the top 1 percent of the unconditional distribution according to our importance indicator are patent 4,237,224 for recombinant DNA methods; patents 4,683,202, 4,683,195, and 4,965,188 for the 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. Conclusion

We use textual analysis of high-dimensional data from patent documents to create a new measure of technological innovation that allows us to characterize the evolution of technological waves over the entire 1840–2010 period across a broad set of sectors.

REFERENCES

Abrams, David S., Ufuk Akcigit, and Jillian Grennan. 2013. "Patent Value and Citations: Creative Destruction or Strategic Disruption?" National Bureau of Economic Research Working Paper 19647.

Ashtor, Jonathan H. 2019. "Investigating Cohort Similarity as an Ex Ante Alternative to Patent Forward Citations." *Journal of Empirical Legal Studies* 16 (4): 848–80.

▶**Field, Alexander J.** 2003. "The Most Technologically Progressive Decade of the Century." *American Economic Review* 93 (4): 1399–1413.

Goldschlag, Nathan, Travis J. Lybbert, and Nikolas J. Zolas. 2016. " An 'Algorithmic Links with Probabilities' Crosswalk for USPC and CPC Patent Classifications with an Application towards Industrial Technology Composition." Center for Economic Studies Discussion Paper 16-15.

Google Patents. 2016. Google. https://patents.google.com/ (accessed July 2016).

Gordon, Robert J. 2017. *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: Princeton University Press.

Griliches, Zvi. 1998. "Patent Statistics as Economic Indicators: A Survey." In *R&D and Productivity: The Econometric Evidence*, 287–343. Chicago: University of Chicago Press.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg. 2005. " Market Value and Patent Citations." *RAND Journal of Economics* 36 (1): 16–38.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. 2021. "Replication Data for: Measuring Technological Innovation over the Long Run." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E119043V1.

[6] 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.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman. 2017. “Technological Innovation, Resource Allocation, and Growth.” *Quarterly Journal of Economics* 132 (2): 665–712.

Lampe, Ryan, and Petra Moser. 2010. “Do Patent Pools Encourage Innovation? Evidence from the Nineteenth-Century Sewing Machine Industry.” *Journal of Economic History* 70 (4): 898–920.

Shea, John. 1999. “What Do Technology Shocks Do?” In *NBER Macroeconomics Annual 1998*, Vol. 13, edited by Ben S. Bernanke and Julio Rotemberg, 275–322. Cambridge, MA: MIT Press.

Younge, Kenneth A., and Jeffrey M. Kuhn. 2016. “Patent-to-Patent Similarity: A Vector Space Model.” https://dx.doi.org/10.2139/ssrn.2709238.