Measuring Technological Innovation over the Long Run*

Bryan Kelly† Dimitris Papanikolaou‡ Amit Seru§ Matt Taddy ¶

July, 2017

Abstract

We use textual analysis of 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 quality is predictive of future citations and correlates strongly with measures of market value. Our quality measure is unique in that it is available for the entirety of patent documents, spanning approximately two centuries of innovation (1836–2016) and covers innovation by private and public firms, as well as non-profit organizations and the US government.

*We thank Jinpu Yang for excellent research assistance.

†Chicago Booth and NBER

‡Kellogg School of Management and NBER

§Stanford GSB and NBER

¶Chicago Booth and Microsoft Research

Economists broadly agree that advances in technology play a major role in economic growth over the long run. Over the last two centuries, real GDP 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; hence much of economic growth is attributed to improvements in productivity. Similarly, there are large and persistent differences in productivity across firms or establishments. Understanding the link between technological progress and these measures of productivity has been at the centre of the economic agenda that tries to explain these differences. Measuring technological innovation using patents has been an important step in this direction.[1] However, a major obstacle in inferring technological progress from patent statistics is that patents vary greatly in their technical and economic significance. While measures such as citations a patent garners and market value of a patent have been used to address this obstacle, capturing these metrics for patents going back in time has been infeasible (see Kogan, Papanikolaou, Seru, and Stoffman (2016) for a discussion).[2]

In this paper, we propose a new measure of patent quality that is based on textual analysis of the patent documents and is available over almost two centuries. An impactful patent, according to our measure, is one that is similar to future patents but is different from prior patents. Our measure thus aims to capture distinct improvements in the current level of technology that become the new foundation upon which subsequent inventions are built.

We construct similarity between patents using state of the art techniques from textual analysis. This requires no other input besides the text of the patent document. Our measure correlates with existing quality indicators, including patent citations and estimates of economic value. Unlike existing quality indicators, our quality measure is available for the entirety of patent documents, spanning approximately two centuries of innovation (1836–2016).

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 address this issue by leveraging existing advances in 'big data' to create links between each new invention and the set of existing and subsequent patents

[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] Much of the existing literature measures the 'quality' of the underlying patents by citation counts. However, patent citations are consistently recorded in patent documents only relatively recently (after 1946) which makes analyzing long-run trends challenging. More recently, Kogan et al. (2016) propose a new measure of the private, economic value of new innovations that is based on stock market reactions to patent grants and is only available for publicly traded firms after 1927.

that take into account that different terms (words or phrases) occur with different frequency. In particular, we construct measures of similarity that place more weight on 'important' terms. A term that is common in one document but appears rarely in others is assigned more importance. However, this approach—used commonly and termed *inverse document frequency* (IDF)—has an important disadvantage in our case, in that it ignores the temporal ordering of patents.[3] To overcome this issue, we introduce a modification to the standard IDF approach by weighing terms according to the frequency in which they appear in patent documents *up until the patent application is filed*. That is, the appropriate weight that terms receive in our similarity calculation evolves over time as scientific terms become more common or as natural language evolves.

Analyzing the empirical distribution of these similarity scores reveals a sparse matrix of connections. Most patent pairs are dissimilar, but a few are strongly connected. Importantly, these estimated connections are meaningful. Specifically, patent pairs that are linked by a citation are more similar. Further, patents tend be more similar to other patents in the same technology class than patents in other classes.

Armed with a methodology in characterizing similarities between distinct patent documents, we next construct quality indicators at the patent level. Focusing on the subset of patents for which we have information on forward citations (approximately one-half), we see that highly cited patents are those that are both *novel* (they are sufficiently different than previous patents) and *impactful* (they are closely related to subsequent patents). We therefore measure patent quality as the ratio of the patent's similarity to future patents, scaled by its similarity to previous patents. For computational reasons, but also to deal with truncation issues at the start and end of the sample, we restrict the horizons over which these similarities are calculated to be less than 20 years.

Our text-based indicator of patent quality has sensible properties. The relation with patent citations is both statistically and economically significant. When we measure patent quality and citations over the same time period after the patent application is filed, we find that an increase of our patent quality indicator from the median to the 90-th percentile is associated with an increase of 60-80% in citations for the median patent in terms of cites. In addition to be significantly correlated with forward citations, our measure is also correlated with the Kogan et al. (2016) 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. We find that a patent that is in the 90-th distribution in terms of patent quality as per our

[3]Consider for example Nikola Tesla's famous 1888 patent (number 381,968) of an AC motor, one of the first patents to use the bi-gram "alternating current," a phrase used with great frequency throughout the 20th century. The standard IDF approach would sharply de-emphasize this term in the $TFIDF$ vector representing Tesla's patent, and thereby give a misleading portrayal of the patent's technological innovation.

measure is approximately 7.5% to 10% more valuable than the median patent, depending on the horizon over which we measure quality.

An important advantage of our measure is that it allows us to measure the quality of inventions that are discovered in the 19-th and early 20-th century. In general, citation information on these patents is quite limited.[4] To illustrate the usefulness of our measure for this earlier period of American invention, we obtain a list of 110 historically important patents. This list of patents contains most important inventions during this period, such as, anesthesia, the telephone, the internal combustion engine, the phonograph, and the 'calculating machine' (a precursor to the computer). We find that over 40% of these patents are at the top 10% of the distribution in terms of our quality measure.

In addition, our new measure of patent quality has distinct advantages over citation counts that extend beyond data availability. Focusing on the subset of patents for which citation information is available, we find that our quality indicators capture information about the importance of a patent that is complementary to patent citations. In particular, we see that our text-based measure of quality is predictive over the number of times the patent is subsequently cited. That is, we find that, our measure of patent quality based on the similarity with patents filed within the first T years subsequent to the patent application can reliably predict the number of times the patent is cited by patents that are filed *more than T years* after the patent application. This result holds even when controlling for the number of times the patent is cited within the first T years.

We illustrate the usefulness of our measure through several applications. First, we construct an index of major technological innovations that spans the 1840–2010 period. Our index is a simple count of the number of important patents that are filed in each year, where important patents are those that lie in the top 5% of the unconditional distribution of patent quality captured using our measure. Our index correlates strongly with subsequent growth in aggregate measures of productivity, measured either as labor productivity (over the 1889–1957 period) or total factor productivity (over the 1948–2015) period. In addition, we show that our index contains information that is distinct from simple patent citation counts.

Second, we revisit the analysis in Hall, Jaffe, and Trajtenberg (2005) that relates stock of patents and citations garnered by these patents to firms' stock market valuations. We find that constructing stocks of intangibles, that is, accumulated patent counts adjusted for quality using our text-based measure, accounts for a substantial fraction of the cross-sectional dispersion in Tobin's Q across firms. As before, the information contained in our measure is

[4]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 two citations as of March 2017. 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 2017, it has received a total of 10 citations.

complementary to patent citations, and largely comparable in magnitude.

In sum, we propose a new metric of patent quality that is based on textual analysis of patent documents. The information contained in our measure is complementary to patent citations, even when citation information is available. Unlike existing indicators, our measure can be constructed for the entirety of patent documents made available over the 1840–2016 period by USPTO, spanning innovation by private and public firms, as well as non-profit organizations and the US government.

The paper most closely related to our work is Kogan et al. (2016), who propose a new measure of the private, economic value of new innovations that is based on stock market reactions to patent grants. Being an estimate of the economic value of a patent, their measure has the advantage of being readily comparable across time and industries. However, their measure is only available for publicly traded firms after 1927, and hence misses both private firms and research institutions, but also the technological advances associated with the Second Industrial Revolution of the late 19-th century. Further, the two metrics measure different aspects of patent quality. By construction, Kogan et al. (2016) measure the private value of the patent to the firm. By contrast, our indicators measure the scientific novelty and impact of the patent. These two metrics 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. Measures of scientific value are useful in estimating the social return to R&D or the productivity of research personnel (see e.g. Bloom, Jones, Reenen, and Webb, 2017). Nevertheless, our results confirm that the scientific and the (private) economic value of patents are related.

The advantage of using financial data 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. This private value need not coincide with the scientific value of the patent – typically assessed using forward patent citations. For instance, a patent may represent only a minor scientific advance, yet be very effective in restricting competition, and thus generate large private rents. These ex-ante private values are useful in studying firm allocation decisions, estimating the (private) return to R&D spending, and assessing the degree of creative destruction and reallocation that results following waves of technological progress. Further, the fact that our measure of 'quality' is in terms of dollars implies that our estimates are comparable across time and across different industries; in contrast, since patenting propensities could vary, comparing patent counts across industries and time becomes more challenging.

More broadly, 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 level of individual patents and spans the 1840-2016 period.

Second, our work is related to work that links firm patenting activity to stock market valuations (see, e.g. Pakes, 1985; Austin, 1993; Hall et al., 2005; Nicholas, 2008). 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 that is in addition to 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 in addition to R&D, patent and citation counts.

1 Measurement

Here, we describe the construction of our patent quality metrics. First, we briefly describe our data sources in Section 1.1. Appendix A has all the details. Section 1.2 describes our estimation of similarity between patent documents, along with our modification of the leading textual analysis methodology. Section 1.3 contains the properties of the estimated pairwise similarity metric. Last, Section 1.4 contains the bulk of our analysis, which focuses on constructing a patent-level indicator of quality.

1.1 Data

Here, we briefly summarize the data construction process, including the process through which we convert the text of patent documents to a format that is amenable to constructing similarity measures. We relegate all details to Appendix A. Our dataset is built on two

sources. 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. For patents granted prior to 1976, we collect the patent text patents from our second main datasource, Google's patent search engine. 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. Some parts 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. (2016) with adaptations for our more extensive sample.

1.2 Measuring similarity between two patent documents

We begin our analysis by constructing measures of pairwise patent similarity. A key step in obtaining a distance measure between two patents is to devise an appropriate metric that weighs the importance of different words; we want terms like 'electricity' to matter more than common words like 'and' or 'inventor'. The leading approach in text analysis is to weigh term counts by "term-frequency-inverse-document-frequency," or

$$TFIDF_{d,w} \equiv TF_{dw} \times IDF_w. \tag{1}$$

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

$$TF_{dw} \equiv \frac{c_{dw}}{\sum_k c_{dk}}, \tag{2}$$

and describes the relative importance of term w for patent d. It is essentially a count of how many times term w appears in patent d, adjusted for document length. The second component is the inverse document frequency of term w, which is defined as

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

The IDF part of the weight is a measure of the 'informativeness' of term w, and under-weighs common words that appear in many documents.

The product of these two terms, $TFIDF$, describes the importance of a given word or phrase w in a given document d. 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 $TFIDF_{dw}$ indicates that term i appears relatively frequently in document d but does not appear in most other documents, thus

conveying that word w is especially representative of document d's semantic content.

For our purposes, this existing approach is problematic, since the temporal ordering of patents matters. In particular, we are interested in the representative text content of a patent d given the history of innovation leading up to the development of patent d. Consider for example Nikola Tesla's famous 1888 patent (number 381,968) of an AC motor, among the first patents to use the bi-gram "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, and thereby give a misleading portrayal of the patent's technological innovation.

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. This "backward-IDF" of term w as of date t, (denoted by $BIDF_{wt}$), measures log frequency of documents containing w in any patent granted prior to date t. More specifically, backward-IDF is defined as:

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

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 that reflects the history of invention up to but not beyond the new patent's arrival.

We measure the overlap in textual content between a pair of patents i and j as the cosine similarity in their $TFIDF$ vector representations. 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

$$TFIDF_{w,i,t} = TF_{w,i} \times BIDF_{w,t}, \quad t \equiv \min(\text{yr}_i, \text{yr}_j) \tag{5}$$

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

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

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 $TFIDF$'s are not centered before the dot product is applied. Thus, because $TFIDF$ is non-negative, $\rho_{i,j}$ 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 matrix of approximate dimension 9 million $\times$ 9 million, or roughly 30 terabytes of data. To reduce the computational burden when studying similarities, we set the similarities below 5% to zero. This affects 93.4% of patent pairs. Patents with such low text similarity are, for all intents and purposes, completely unrelated, yet would introduce a large computational burden in the types of analyses we pursue. Replacing these approximate zeros with similarity scores of exactly zero achieve large computational gains by allowing us to can work with sparse matrix representations that require substantially less memory.

1.3 Descriptive statistics of the pairwise similarity measure

Next, we describe some of the features of our pairwise similarity measure $\rho_{i,j}$. Panel A of Figure 1 plots the distribution of our similarity score across patent pairs. For computational reasons, we only consider pairs that are at most 20 years apart. We see that the distribution of our similarity scores is substantially skewed to the right: the median similarity score across patent pairs is 7.8%, whereas the average similarity score is 10.2%. The right tail is substantial: the 90-th and 95-th percentile of similarity scores are 17.6% and 22.9%, respectively. These results indicate that the similarity score is able to capture a strong connection among certain patent pairs. For comparison, only 0.007% of patent pairs (with similarity scores above 5%) are linked by citations.

A natural next step is to examine how our similarity score compares to existing measures of similarity. To this end, we examine whether patent pairs that are classified as being similar given our text-based measure $\rho_{i,j}$ are more likely to belong in a citation pair. To do so, we compute the likelihood of patent j citing patent i conditional on their text-based similarity, $E\left[\mathbf{1}_{i,j}|\rho_{i,j}\right]$, 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 before patent j.

Panel B of Figure 1 plots the results. 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. These results constitute a useful external validity check for our procedure, since our computation of similarity scores does not include any information on patent citations.

We next examine how the pairwise similarity score varies depending on whether patents i and j belong in the same technology class, defined as both patents sharing a technology classification code at the 3-digit CPC level. Since technologies may diffuse at different rates within versus between technology classes, we also condition on the distance in years between the year that patent j is filed relative to patent i. For comparison, we perform the same exercise for patent citations.

Figure 2 presents the results. Examining Panel A, we see that the mean similarity score is approximately 15–20% higher if the patent pair i and j share a technology classification, versus if they do not. Further, we see that the mean similarity scores are mildly declining with the difference in the filing years between patents i and j. While part of this decline may simply reflect the evolution of language, it is also conceivable that it captures the fact that the rate of technology diffusion slows down with time. When comparing to fiward patent citations, Panel B reveals that patents that share a technology class are also more likely to cite each other—approximately by a factor of ten—relative to patent pairs that do not share a technology classification. Interestingly, we also see that the likelihood that patent j cites patent i is non-monotonic with respect to the time gap between them, peaking approximately at year 5. Contrasting this pattern with that obtained for our similarity measure, one interpretation is that the text-based measure of similarity is better able to capture links between patents that are filed closely together relative to citations—possibly because patent examiners may not be aware of recently filed patents. Of course, it is also possible that the evolution of language confounds this effect. We return to this issue in Section @@ below.

1.4 Constructing patent-level measures of quality

The next part of our analysis consists of combining these pair-wise similarity scores into a patent level measure of 'scientific' quality. In doing so, it might be important to distinguish

the degree to which a patent is different than its predecessors (its 'novelty') from the degree to which later patents build on this invention (its 'impact').

1.4.1 Methodology

Our goal in this section is to construct a measure of the scientific impact and novelty of a patent. Conceptually, a scientifically impactful patent is one that opens the the way for more innovation. We would therefore expect these impactful patents to be closely related to *future* patents.

Our first measure of quality—impact—is then defined as the total forward similarity, that is,

$$FS_j^{0,\tau} = \sum_{i \in \mathcal{F}} \rho_{j,i}, \tag{8}$$

where $\rho_{i,j}$ is the pairwise measure of similarity between patents i and j defined above in equation 7, while $\mathcal{F}_{j,\tau}$ denote the set of all "forward" patents, that is, patents filed in the τ calendar years following patent j's application year. The forward similarity measure in (8) is an estimate of the strength of association between the patent and future technological innovation over the next τ years.

Similarly, we can define the notion of a novel patent would be an invention that is a discrete advance relative to the state of the art—and would therefore be dissimilar to the existing patent stock. This notion can be captured by a measure of backward similarity,

$$BS_j^{0,\tau} = \sum_{i \in \mathcal{B}} \rho_{j,i}, \tag{9}$$

where now $\mathcal{B}_{j,\tau}$ denote the set of "backward" patents granted in the τ calendar years prior to j's application year. We will consider $\tau = 5$ as our baseline case, though our results are similar if we use shorter windows. Here, backward similarity is a measure of the novelty of the patent relative to the existing patent stock.

Table 1 reports the distribution of the forward similarity measures across different horizons τ. By symmetry, the distribution of the novelty score is similar up to truncation lags. For comparison, we also report the distribution of forward citations across the same horizons, and the KPSS measure of patent value, which is based on the dollar stock market reaction around the days that the patent is issued to the firm. Examining the Table, we note that our text-based impact measure is highly skewed, with the mean typically being 1.5 to 2 times the value of the median. This pattern is consistent with the well-known fact that forward patent citations are also highly skewed, and the previously documented skewness in the KPSS patent value measure. These facts are consistent with the presence of a small number of

highly valuable patents. In the sections that follow, we examine the correlations between these measures of patent quality.

1.4.2 Impact, novelty and forward citations

The existing literature on innovation mostly relies on patent citations as a measure of the 'quality' of the underlying patent. As a first pass, we first examine how our patent impact measure relates to forward patent citations. In particular, we estimate

$$\log\left(1+CITES_j^{0,\tau}\right) = a + b\log\left(FS_j^{0,\tau}\right) + c\log\left(BS_{0,5}^{j}\right) + Z_j + \varepsilon_j, \tag{10}$$

where we measure patent impact and citations over the next τ years after the patent is filed. To reduce space, we only focus on estimating backward similarity over the last 5 years, but our results are robust to alternative horizons. Here, Z_j is a vector of controls that includes dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm and the interaction of firm and year effects. Including firm fixed effects dramatically reduces the number of observations since we have firm identifiers only for firms that are matched to CRSP/Compustat. That is, in our most conservative specification we compare patents in the same technology class that are granted to the same firm in the same year. Since patent citations are only consistently documented after 1945, we restrict the sample to the 1946–2016 period. Last, we cluster the standard errors by the patent grant year.

Table 2 presents the results. The table reveals two broad patterns, that are consistent across horizons τ and choice of controls Z. First, the forward similarity measure is positively and significantly related to the number of times the patent gets cited over the same period. That is, patents that are likely to be related to subsequent patents in terms of text similarity, are also more likely to receive more citations. Second, patents that are more novel, in the sense that are more dissimilar to earlier patents, are also more likely to be cited more in the future. Interestingly, the estimated coefficients b and c are of similar magnitude—but opposite sign. Hence, it appears that the ratio between the forward and the backward similarity may be a useful summary statistic of the scientific value of a patent, at least as measured by patent citations. Next, we explore this idea further to construct a summary measure of patent quality.

1.4.3 A summary measure of quality

Here, we build a summary measure of patent quality that incorporates both the patent's impact (forward similarity) and novelty (backward similarity). Specifically, motivated by the fact that the estimated coefficients b and c are of similar magnitudes, but opposite signs, we

infer that a variable that likely summarizes the information content in both FS and BS is the (log) difference between them. Hence, we construct a measure of the scientific importance of a patent, as the ratio of the patent's future impact FS to its novelty BS,

$$RSIM_j^{0,\tau} = \frac{FS_j^{0,\tau}}{BS_j^{0,5}}. \tag{11}$$

We refer to the measure in (11) as "relative forward similarity" and interpret it as an overall measure of patent quality. We choose a horizon $\tau = 5$ for the denominator, but we obtain similar results using shorter horizons of one year.

In particular, our summary measure (11) attaches higher scientific value to patents that are more novel relative to their predecessors but are related to subsequent research. Forward similarity measures the strength of association between the patent and future technological innovation, and normalizing by backward similarity emphasizes the novelty of the patent. A patent may have high forward similarity because it is a "follower" in a technology area with many other followers, in which case it is likely to also have a high backward similarity as well. On the other hand, its high forward similarity may indicate a new and impactful breakthrough, in which case it is likely to have low backward similarity, and thus an especially high relative forward similarity. Further, another reason why a patent might have high forward similarity is that it uses general language that is not distinct to any particular technology but is stylistically common. In this case, normalizing by backward similarity counteracts the effect of general language on measured impact.

To obtain a sense of the time-series properties of our quality measure, the top panel of Figure 4 plots the cross-sectional distribution over time. Examining the figure, we see that the average patent quality is relatively high in the 1840–1870 period, coinciding with the beginning of the Second Industrial Revolution, and also in the 1980–2000 period, which coincides with the Information Age. Further, even though the mean is not particularly high, the top percentiles of patent quality are also relatively high in the 1920–30s and 1950–70s, periods that have been identified as technologically progressive (Field, 2003). In Section 2 we revisit this issue, constructing long-run indices of technological change.

The bottom panel of the same Figure 4 plots the cross-sectional distribution of patent citations over time. We see that, because citations suffer a truncation issue, not only in the end but also in the beginning of the sample —the pattern looks rather different. Hence, without some type of adjustment, one cannot easily compare the number of citations a patent receives across different cohorts. For instance, we see that the median patent prior to 1910 has zero citations. Nevertheless, the patents at the very top of the distribution receive a considerable number of citations, even if they were issued in the 19-th century. Hence, even

with this truncation bias, citations are still informative about the quality of the patent (Moser and Nicholas, 2004; Nicholas, 2008).

1.4.4 Relation with patent citations

To illustrate how (11) performs as a metric of patent quality, we next re-estimate a similar specification as (10)

$$\log\left(1 + CITES_j^{0,\tau}\right) = a + b \log RSIM_j^{0,\tau} + Z_j + \varepsilon_j. \tag{12}$$

As we see in Table 3, the measure constructed in (11) largely summarizes the information contained in both text-based measures that is related to citations. Further, the magnitude of these correlations is substantial. Focusing on our most conservative specification, that compares two patents filed in the same year, are in the same class, and are issued to the same firm in the same year, we find that increasing the quality measure from the median to the 90-th percentile results in an increase in the dependent variable of 0.11 to 0.45 log points. Focusing on a 5 (10) year horizon, these numbers imply an increase of 0.80 (1.7) additional citations relative to the median of 2 (3) citations. For comparison, the 90-50th range of patent citations over the next 5 (10) years following the patent application date is 5 (11).

The results in Table 3 are based on citations measured over the same window as our quality indicator. Consequently, they are based only on patents granted after 1945, which is when the patent office started recording patent citations. However, patent citations may still be informative about the relative importance of older patents, as long as they are measured over the entire sample (Moser and Nicholas, 2004; Nicholas, 2008). Indeed, we saw in Figure 4 that even among the patents that were granted in the 19-th century, there are some (relatively) highly cited patents.

Table 4 therefore estimates a modified version of equation (12), in which patent citations are measured over the entire sample. Since only within-cohort comparisons are possible as a result, all specifications include year fixed effects. We see that even during this period, our quality indicator is still significantly correlated with patent citations. More importantly, the magnitudes are comparable to the post-1945 sample. Focusing on our most conservative specification, that compares two patents filed in the same year, are in the same class, and are issued to the same firm in the same year, we find that increasing the quality measure from the median to the 90-th percentile results in an increase in the dependent variable of approximately 0.06 to 0.15 log points. Focusing on a 5 (10) year horizon, these numbers imply an increase of 0.32 additional citations relative to the median value of 1 citation during this period.

In sum, we see that our text-based measures are strongly significantly related to the most

commonly-used indicator of patent quality, forward citations. Importantly, our proposed quality measures have several distinct advantages in measuring the scientific value of the patent relative to patent citations. The first advantage is that, unlike citations, our quality measures do not suffer from truncation bias, except at the very ends of the sample. As a result, our quality indicators are useful not only in comparing patents of the same cohort, but also *across* cohorts; in Section 2 we will exploit this advantage to construct indices of technological change that span two centuries.

Second, our quality measure likely has an advantage over patent citations even in the post-1945 sample. In particular, citations are a discrete event, whereas our similarity measures are continuous. The discreteness of patent citations may make it a rather noisy measure of patent quality, especially when citations are measured over short horizons. For example, Table 2 shows that the median patent in the post-1945 sample receives no citations in the first year following its filing date, 1 citation 0-5 years out, and 2 citations 6-10 years out. Further, an advantage of our text-based measure is that it does not rely 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. Figure 3 illustrates this pattern more clearly. In panel A, we plot the mean increase in the total forward similarity $\Delta FS_{0,t}$ across horizons of $t = 1 \ldots 20$ years. We see that the amount by which the total forward similarity $FS_{0,t}$ increases is strongly declining across horizons — that is, $FS_{0,t}$ is concave in t. This pattern is reminiscent of a similar pattern documented in Figure 2 for patent pairs, but this is aggregated at the patent level. By contrast, the increase in forward citations $\Delta C_{0,t}$ is non-monotonic, again peaking at about 5 years.

One interpretation of the patterns in Figure 3 is that our text-based quality measure captures information about the quality of a patent *earlier* than patent citations. To explore this idea further, we estimate predictive regressions of the form,

$$\log\left(1 + CITES_j^{\tau,\tau+T}\right) = a + b \log RSIM_j^{0,\tau} + c \log\left(1 + CITES_j^{0,\tau}\right) + Z_j + \varepsilon_j. \qquad (13)$$

That is, we examine whether our predictive measure computed over a fixed horizon $t \in [0, \tau]$ is predictive of future citations to the same patent after τ, while controlling for the number of citations the patent receives in $t \in [0, \tau]$. As before, we include a variety of fixed effects, including year and technology class dummies.

Our main coefficient of interest is b, which captures the predictive relation between our impact measure and future citations. We present the results in Table 5. We see that our impact measure predicts future citations, even controlling for the number of contemporaneous citations. The relation is both statistically as well as economically significant. Focusing on the middle row of the table, and our most conservative specification that includes application,

grant, and class fixed effects, we see that an increase in the patent quality (measured over 5 years) from the median to the 90-th percentile is associated with an increase of 0.13 log points of the dependant variable, which predicts an increase in forward citations over the next 5 years (6 to 10 years out) of 0.28 relative to the median of 1 citation. Hence, these magnitudes are not only statistically significant but also economically meaningful.

We conclude that our text-based measure of patent quality contains economically meaningful information relative to forward citations, even when both measures are computed over the same horizon. Our conjecture is that this result is driven by the increased granularity of our impact measure relative to citation counts, which are discrete events.

2 Innovation over the long run

The results in the previous sections illustrate that our quality indicator is highly correlated with patent citations. An important advantage our quality indicator has over forward citations however, is that it is not subject to truncation lags — except at the very end of the sample. By contrast, since patent citations are recorded in patent documents only since 1945, they are subject to truncation lags that cover most of the 1840–2010 sample. In this section, we exploit this advantage of our quality indicators to create time-series indicators of the degree of technological progress during this period. Specifically, we focus on the number of 'breakthrough' patents—that is, patents that score highly in terms of the unconditional distribution of quality given our measures. To minimize the truncation at the end of the sample, we focus our analysis on the measure $RSIM_j^{0,10}$, which only uses information over the next 10 years of a patent grant.

We begin by first discussing the composition of these breakthrough patents in Section 2.1 and focus on several prominent examples. Section 2.2 describes a validation exercise using a list of historically important patents. Section 2.3 describes the construction of our index of technological progress and documents the correlation with measured productivity.

2.1 Composition of important patents

Figure 6 presents the composition of patents across technology classes in the 1840 to 2010 period by decade. Examining Panel A of the figure, we see that the composition of patents across technology classes is relatively stable across decades. However, each of these classes were responsible for important inventions at different points in time.

Panel B of Figure 6 illustrates this point more clearly. Here, for each decade, we plot the class composition of the top 1% of patents in terms of our quality measure. We see that the technology classes in which important inventions originated has varied quite a bit over the

last 170 years. In the 1840–70 period, we see that some the most important inventions took place in engineering and construction, consumer goods, and manufacturing. An example of an invention in construction that score high in terms of our quality measure is the 'Bollman Bridge' (patent number 8,624), named after its creator Wendell Bolman, 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). Interestingly, many of the important patents in consumer goods are also related to new items clothing.

Starting around 1870, many more patents that score high in terms of our measure are related to electricity. Many of the most important patents given our measure are related 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 score in the top 1% of the unconditional distribution of our quality measure.[5]

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 1% 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's 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, another area that is responsible for important patents is chemistry; many of those patents are related to the invention of plastic compounds. Our quality indicators identify the patent for bakelite (942,699), the world's first fully synthetic plastic as particularly important—it ranks in the top 10% of all patents in terms of our quality indicators. This innovation opened the floodgates to a torrent of now-familiar synthetic

[5]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

plastics, including the invention in the 1930's of plasticized polyvinyl chloride by Waldo Semon (patents 1,929,453 and 2,188,396) and nylon by Wallace H Carothers (patent 2,071,250), all of which are important patents according to our measure. Other important patents in chemistry continue through the 1950's. Patents that score in the top few percentiles according to our measure, include 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 practical integrated circuit made of silicon 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. Indeed, Xerox has been responsible for a substantial fraction of these innovations; some of the patents identified as important by our measure include patent 4,558,413 for a software version management system; patent 4,899,136 for improvements in user interface; patent 4,437,122 for bitmap (raster) 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 also fall in the top 1% of all patents. [6]

Improvements in genetics also comprise a significant fraction of the patents in the 1980–2000 period. A few early examples of important patents in this area, 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 and 4,683,195 for the polymerase chain reaction (PCR) method, a technique for making copies of DNA segments quickly, with high fidelity, easily, and at relatively low cost; patent 4,736,866 for transgenic (genetically modified) animals; and patent 4,889,818 for heat-stable DNA-replication enzymes.

2.2 A validation exercise

As we discuss above, a key advantage of our measure is that it can be used to compare the quality of patents across different cohorts, even in the earlier part of the sample. This period is particularly relevant for innovation. It includes not only the Second Industrial Revolution (1870–1910), but also the 1920–1940 period, which according to Field (2003) contained major

[6] See, 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.

technological breakthroughs, this advantage provides us an unique opportunity to expand the period over which novelty of innovation can be explored.

A potential challenge in using the text of these earlier patent documents is that the print quality—and hence the accuracy of our similarity measure—may be lower than in the modern period. One way to provide some external validity for our measure of patent quality is to examine how historically important patents are scored according to our quality indicator. In the absence of other independent measures of patent quality, we use the list of 110 patents complied by the patent historian Jim Bieberich, available through USPAT.COM.[7] This is a subjective list, that however includes most breakthrough inventions of the 19-th and 20-th century: the Sewing Machine, Anesthesia, Machine Guns, the Telephone, the Automobile, and the Radio are just a few examples. The full list is shown in Tables A.5–A.6 in the Appendix.

For each one these breakthrough inventions we compute our patent quality measure 11 over horizons of 1, 5, 10 and 20 years. We then examine how our quality indicator for each patent compares to the distribution of quality indicators for all patents filed in the same year. Table 6 summarizes these results. We find that these important patents are substantially more likely to have higher text-based quality scores. Specifically, when we measure our patent quality over at least 10 years following the patent application day, approximately 40% of these patents are in the top 10% of the distribution of quality, while approximately 30% are in the top 5% of all patents in terms of the unconditional distribution of patent quality.

Next, we repeat the same exercise using forward citations, measured over the entire sample — since citations are only recorded since 1945. This comparison is naturally skewed in favor of forward citations, not only because they use much more information than the first 10 years of the patent filing date, but also because the number of citations was likely to be a criterion for compiling this list. Despite these drawbacks, we see that our quality indicators do about as well — and often better — than citations in identifying these patents as being important.

In sum, these results confirm that our text-based measure of patent quality captures meaningful information about the importance of a patent, even during the earlier parts of the sample in which the quality of the digitized patent documents was worse than the later parts. Specifically, we see that historically important patents consistently score in the top percentiles of our quality distribution. In the next section, we exploit this result to build time-series of technological progress that span the 1840–2016 period.

[7]The list is available here: http://www.uspat.com/historical/index.shtml

2.3 A time series index of technological progress

Constructing a time-series index of technological progress presents a challenge for several reasons. A first approach, taken by Shea (1999) is to construct and index based on patent counts. Such an index implicitly assumes that all patents are equally valuable. However, 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. Kogan et al. (2016) take a step towards this end by constructing a time-series index that is based on the estimated market values of patents that are granted. However, a short-coming of their index is that 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.

Here, we build upon the results of the previous section to construct a long time series of technological improvements that spans the entire length of the USPTO data. We do so by counting the number of 'breakthrough' inventions, that is, patents that have quality scores in the top 1% of the (unconditional) distribution. In particular, our index is constructed as

$$\xi_t^\tau = \log\left(1 + \sum_{j \in P_t} \mathbf{1}_{RSIM_j^\tau \geq RSIM_{0.99}^\tau}\right), \tag{14}$$

where P_t is the set of patents that are filed in year t and $q_{0.99}^\tau$ is the 99-th percentile of the patent quality index constructed in (11), in which the numerator (forward similarity, or impact) is measured over years 0 to τ following the patent application date.

We plot the resulting time series for $\tau = 10$ in Panel A of Figure 7. Both indices display considerable fluctuations at relatively low frequencies, indicating four major periods of technological innovation: the 1850–1880 period, the 1920's, the 1950–1970's, and the 1980–2000's. These periods line up with the major waves of technological innovation in the U.S. The first peak in our indices corresponds to the beginning of the Second Industrial Revolution, which saw numerous technological advances, such as the telephone and electric lighting. Second, both indices suggest high values of technological innovation in the 1920s, consistent with the evidence compiled in Field (2003) regarding the advances in manufacturing during this period. Third, our measure suggests higher innovative activity from the mid–1950's to the early 1970s – a period commonly recognized as a period of high innovation in the U.S (see, e.g. Laitner and Stolyarov, 2003). Finally, developments in computing and telecommunication

have brought about the latest wave of technological progress in the 1980s to the beginning of the 2000s, which coincides with the high values of both of our measures. Contrasting Panel A with Panel B, which plots the total number of successful patent applications per capita, reveals that our indices display different behavior than the total number of patents.

As a further validity check of our indices, we examine their correlations with measures of productivity during the period. The most commonly source of productivity measures for the later 19-th century to the mid 20-th century is Kendrick (1961). We use his series on labor productivity in manufacturing, measured as output per effective labor input (Kendrick, 1961, see pages 465–466 in). For the later period, we use the series on total factor productivity constructed by Basu, Fernald, and Kimball (2006) which starts in 1948. Panel C of Figure 7 plots the times series of these two productivity estimates.

We relate our technology indices to measures of (log) productivity x_t using the following specification,

$$x_{t+k} - x_t = a_k + b_k \, \xi_t^{\tau} + \rho_k \, x_t + c_k \, log N_t \varepsilon_{t+k}. \tag{15}$$

Here, x denotes log productivity, ξ^{τ} our technology index defined in (14), and N_t denotes the number of new patent applications filed in year t normalized by population. As Jorda (2005) demonstrates, these local projections confer significant advantages relative to traditional VARs, including being relatively more robust to misspecifications. We adjust the standard errors in (15) for overlapping observations using the Hodrick (1992) procedure. To conserve space, we focus on results for $\tau = 10$, but results using other horizons are qualitatively similar.

The top row of Figure A.3 presents the baseline results that do not control for the number of patents applications (i.e. imposing $c_k = 0$). We see that our innovation indices are positively related to future TFP improvements, and the relation is statistically significant at horizons than 10 years. In terms of magnitudes, a one-standard deviation increase in our index is associated with approximately a 5% increase in productivity over 15 years. We see that the point estimates are similar across the two samples, though they are more precisely estimated in the 1948-2016 period. Part of this difference could be attributed to differences in the quality of the data (both productivity as well as our text-based measure). However, this difference could also be driven by the fact that the two series are not directly comparable. Measures of labor productivity confound changes in the productivity of capital and labor with capital deepening (changes in capital-labor ratios). According to Field (2003), a substantial component of economic growth during the second half of the 19-th century could be attributed to capital deepening, which might for the somewhat weaker correlation between our technology index and labor productivity during this period. By contrast, Field (2003) argues that changes in capital-labor ratios played only a minor role in productivity growth during the second half of the 20-th century. Indeed, we obtain similar results if we

replace TFP with measures of labor productivity (output per hours) in the post-1948 period.

We next examine whether these results are driven mostly by fluctuations in the total number of patent applications. That is, we re-estimate (15), but now include controls for the (log) number of patent applications per capita N_t filed in year t. As we see in the bottom row of Figure A.3, the point estimates are largely similar, indicating that the associated movements in TFP are related to the important inventions, rather than just an increase in overall patenting activity.

We performed several additional robustness checks to our analysis. Specifically, we (a) experimented with an alternative threshold of 90% in (14); (b) we considered alternative measures of productivity in both periods, specifically, output per worker hours. None of these made a material difference in these results, and are therefore related to the Online Appendix.

3 Market value and quality

In this section, we discuss the relation between our quality indicators and market valuations. In examining these relations, one has to keep in mind that market values measure, by construction, the present value of pecuniary benefits to the holder of the patent. By contrast, our quality measure is most likely correlated with the scientific importance of the patent. In general, the relation between the two can be ambiguous. For instance, a patent may represent only a minor scientific advance, yet be very effective in restricting competition, and thus generate 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.[8]

In what follows, we revisit some of the evidence using our new indicator of patent quality. We do so at two levels of granularity. In section 3.1 we do so at the patent level. The advantage is that we can do so at a higher level of granularity than Hall et al. (2005). The disadvantage is that the estimated market value of each patent is based on stock market reactions around a very narrow window around the issuance date, and hence may omit the part of the market value that was incorporated into the stock price prior to the patent grant. In section 3.2 we perform a similar exercise at the firm level, following Hall et al. (2005). The advantage of this approach is that it is identified using differences in the firms' patent portfolio and market valuations, hence it does not suffer from the 'missing value' problem of KPSS. The disadvantage is that the analysis relies on comparing otherwise similar firms, and

[8]For instance, Hall et al. (2005) and Nicholas (2008) document that firms owning highly cited patents have higher stock market valuations. Harhoff, Narin, Scherer, and Vopel (1999) and Moser, Ohmstedt, and Rhode (2011) provide estimates of a positive relation using smaller samples that contain estimates of economic value. By contrast, Abrams, Akcigit, and Popadak (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.

hence may be contaminated by unobservables.

3.1 Patent-level evidence

We next discuss the relation between our text-based measure of the quality of a patent and the market value of a patent using the measure of KPSS. The latter is based on the stock market reaction to a patent grant, and can therefore be interpreted as a measure of the private value of the patent—that is, the present value of cashflows to the patent assignee that can be attributed to the patent.

We relate patent impact to market values using the following specification,

$$\log V_j = a + b \log RSIM_j^{0,\tau} + Z_j + \varepsilon_j. \tag{16}$$

As before, we saturate our specifications with controls, including year fixed effects, technology-class dummies, firm, and firm-year fixed effects. When estimating (16), if multiple patents are issued to the same firm in the same day, we collapse these observations into one by averaging across patents. We do so because the KPSS measure cannot 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.

We present the results in Table 7. The estimated coefficient b reveals a strong, statistically significant relation between the KPSS measure of market value and our text-based measure of impact. Focusing on the most conservative specification—column (5)—which compares patents in the same class, that are issued to the same firm in the same year, we see that increasing the quality measure from the median to the 90-th percentile results in a 7.5% to 10% increase in the estimated log patent value, where quality is measured across horizons of 1 to 10 years subsequent to the patent application date.

In Table 8, we repeat the same exercise with the number of forward citations included as controls. We again see that our text-based measure of impact is significantly related to patent market values, even when citations are included. This pattern reinforces our earlier conclusion that our text-based measure of impact captures information on the importance of a patent that is complementary to citation counts. In most specifications, both impact as well as citations are significantly related to estimates of patent value. The exception is column (5) which includes firm–year dummies. In this case, citations enter with a negative sign, suggesting that they perhaps help 'clean up' some of the measurement error in our quality measure. Consistent with this idea, the estimated magnitudes are somewhat larger: increasing the quality measure from the median to the 90-th percentile now results in a 8.5% to 12% increase in the estimated log patent value.

In sum, these results confirm our earlier findings that our patent quality measure captures information that is complementary to forward citations.

3.2 Firm-level evidence

Next, we examine the extent to which our innovation measure can account for differences in firm valuations. Prior work has validated measures of R&D productivity, captured by patent based metrics, by assessing if they are consistently related to firm value. In particular, Hall et al. (2005) assess the relationship between a firm's Tobin's Q and its "knowledge stock" that is constructed based on investment in R&D, number of patents and number of citations. We modify the specification of Hall et al – a log linearized firm-level market value function – by including another explanatory variable that captures the "knowledge stock" based on our measure of similarity.

More specifically, following Hall et al. (2005), the knowledge stock for investment in R&D, number of patents granted and number of citations received by patents granted in a given year is constructed based on a declining balance formula as:

$$SX_{f,t} = (1-\delta)\, SX_{f,t-1} + X_{f,t} \tag{17}$$

where, $X_{f,t}$ is the flow of new R&D, patent applications of successful patents by firm f in year t and citations received by patents of firm f in year t and $SX_{f,t}$ refers to the accumulated (stock) measure. We follow Hall et al. (2005) and use a depreciation rate of $\delta = 15\%$.

The new measure in the market value specification relates to our similarity measure. We first construct a measure of patent quality for firm f in year t as:

$$RSIM_{f,t} = \sum_{k=1}^{P_{f,t}} RSIM_k^{0,\tau} \tag{18}$$

where, $P_{f,t}$ is the total number of patents applied for firm f in year t; . $RSIM_k^{0,\tau}$ is the quality of patent k that is granted to the firm f for a patent application in year t—constructed in equation (11) for different horizons τ ranging from 1 to 10 years.

Next, we construct similarity knowledge stocks by accumulating the firm measure of patent quality (18) similarly to (17). We apply the same depreciation rate $\delta = 15\%$ as in equation (17). As a robustness check, we also experiment with rates of 5%, 10%, 20% and 25% and we obtain similar results to those reported in this section.

We estimate the firm's market value as a function of various explanatory variables as:

$$\log Q_{f,t} = \log\left(1 + \gamma_1 \frac{SRD_{f,t}}{A_{f,t}} + \gamma_2 \frac{SPAT_{f,t}}{SRD_{f,t}} + \gamma_3 \frac{SCITES_{f,t}}{SPAT_{f,t}} + \gamma_4 \frac{SRSIM_{f,t}}{SPAT_{f,t}}\right) \tag{19}$$

$$+q_t + D\left(SRD_{f,t} = 0\right) + \varepsilon_{f,t} \tag{20}$$

where $SRD_{f,t}$, $SPAT_{f,t}$, $SCITES_{f,t}$, and $SRSIM_{f,t}$ are the stocks of R&D expenditure, number of patents, patent citations, and the patent quality measures constructed as in (17). As in Hall et al. (2005), q_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. Other variables in the specification are defined in the Appendix. We estimate the market value regressions, computing the similarity and citations stocks over horizons τ of 1, 5, 10, and 20 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. We cluster standard errors by firm.

Our main coefficient of interest is γ_4 which estimates the relation between our accumulated patent quality measure and Tobin's Q. Table 9 presents the results. We see a a strong, statistically significant relation between Tobin's Q and similarity stock after conditioning on other measures of knowledge stock. In particular, this pattern is obtained after we account for knowledge stock of citations. The economic magnitudes are large as well. When estimated over the 0-1 year horizon, the variation in a firm's Tobin's Q that is explained by variation in its similarity stock is almost twice as large as the variation explained by its citation stock.[9] At longer horizons, reported in other columns, the effect of similarity stock remains important, though it declines in terms of its importance relative to citation stock.

We repeat the same exercise basing our sample only on the Manufacturing firms (SIC 2000-3999) and report the results in Table A.8. As before, within manufacturing firms, the relation between Tobin's Q and similarity stock are also consistently strong and significant. Similar to our baseline results, the variation in a firm's Tobin Q that is explained by variation in its similarity stock is comparable to that explained by its citation stock, with larger effects at shorter horizons.

Overall, the results of this section suggest that the variation in knowledge stock at the firm level that is constructed based on our measure of similarity, significantly relates to its Tobin Q. Importantly, we find this relation after accounting for other measures of knowledge stock, including its R&D stock, patent stock and citation stock. These patterns demonstrate that similarity stock carries important information about a firm's market value and should be a part of standard controls that researchers use when assessing various factors that are related to firm value.

[9] = 0.210/0.112, see Table 9 column (1) "normalized coefficients" based on 1 standard deviation changes in variables.

4 Conclusion

We use textual analysis of 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 quality is predictive of future citations and correlates strongly with measures of market value. Our quality measure is unique in that it is available for the entirety of patent documents, spanning approximately two centuries of innovation (1836–2016) and covers innovation by private and public firms, as well as non-profit organizations and the US government.

A Data Construction Appendix

Here, we describing 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.1 Text Data Collection

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.

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. The records in this sample are easy to process because they are provided in HTML format with standardized fields. 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

[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 only available information that we do not collect 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.

A.2 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.[11] 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.

[11]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, 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.

A.3 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.

A.4 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.[12]

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.[13]

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.[14]

[12] 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)

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

[14] 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.

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.

A.5 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. (2016) 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.

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.'s (2016) patent valuation data for patents granted between 1926 and 2012.

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.

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.

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

Bloom, N., C. I. Jones, J. V. Reenen, and M. Webb (2017). Are ideas getting harder to find? working paper, Stanford University.

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

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.

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.

Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *The Review of Financial Studies 5*(3), 357.

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.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2016). Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics forthcoming*.

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.

Laitner, J. and D. Stolyarov (2003). Technological change and the stock market. *The American Economic Review 93*(4), pp. 1240–1267.

Moser, P. and T. Nicholas (2004). Was electricity a general purpose technology? Evidence from historical patent citations. *The American Economic Review, Papers and Proceedings 94*(2), 388–394.

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.

Tables and Figures

Table 1: Distribution of patent similarity scores

Variable	mean	sd	p1	p5	p10	p25	p50	p75	p90	p95	p99
Impact (FS), 0–1 years	0.7	0.8	0.0	0.1	0.1	0.2	0.4	0.9	1.8	2.6	4.0
Impact (FS), 0–5 years	3.5	4.1	0.1	0.3	0.4	0.9	2.0	4.3	8.6	12.5	19.9
Impact (FS), 0–10 years	6.4	7.7	0.1	0.4	0.8	1.7	3.7	7.9	15.5	22.8	38.1
Impact (FS), 0–20 years	10.4	12.2	0.1	0.7	1.3	2.9	6.3	13.1	24.2	35.0	61.3
Quality (FS/BS), 0–1 years	0.2	0.1	0.0	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.4
Quality (FS/BS), 0–5 years	1.1	0.4	0.0	0.4	0.7	0.9	1.1	1.2	1.4	1.6	2.2
Quality (FS/BS), 0–10 years	2.1	1.0	0.0	0.4	0.8	1.7	2.1	2.6	3.2	3.7	5.3
Quality (FS/BS), 0–20 years	4.0	2.8	0.0	0.4	0.8	2.3	3.9	5.2	6.8	8.4	13.5
Citations, 0–1 years	0.3	1.1	0	0	0	0	0	0	1	2	4
Citations, 0–5 years	2.9	6.8	0	0	0	0	1	3	7	11	29
Citations, 0–10 years	5.8	14.4	0	0	0	0	2	6	13	23	62
Citations, 0–20 years	8.8	22.7	0	0	0	1	3	9	20	34	95
KPSS patent value ($1982m)	11.5	37.4	0.0	0.0	0.1	0.8	3.6	10.2	24.4	42.5	130.9

Table shows the distribution of our patent level similarity scores: impact (forward similarity) and novelty (backward similarity); forward citations; and KPSS patent values (in 1982 million USD). Time period is 1840–2016 for the similarity scores; 1946–2016 for citations; and 1927–2016 for the KPSS patent value measure. Forward similarity scores are scaled by 1,000.

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

Forward citations, 0-1 yr	(1)	(2)	(3)	(4)	(5)
Forward similarity, 0-1 yr	0.493***	0.629***	0.479***	0.557***	0.546***
	(8.15)	(13.66)	(15.77)	(18.03)	(17.42)
Backward similarity, 0-5 yr	-0.440***	-0.597***	-0.458***	-0.542***	-0.532***
	(-7.40)	(-13.71)	(-15.90)	(-18.47)	(-17.75)
Observation	5,156,699	5,156,699	5,121,721	1,775,138	1,758,319
R^2	0.057	0.114	0.147	0.187	0.215
Forward citations, 0-5 yr	(1)	(2)	(3)	(4)	(5)
Forward similarity, 0-5 yr	1.447***	1.438***	1.184***	1.198***	1.185***
	(18.64)	(38.97)	(72.41)	(65.69)	(69.75)
Backward similarity, 0-5 yr	-1.276***	-1.344***	-1.100***	-1.133***	-1.119***
	(-16.12)	(-36.25)	(-66.59)	(-58.99)	(-63.11)
Observation	4,355,594	4,355,594	4,323,134	1,463,683	1,448,640
R^2	0.168	0.235	0.282	0.341	0.367
Forward citations, 0-10 yr	(1)	(2)	(3)	(4)	(5)
Forward similarity, 0-10 yr	1.530***	1.258***	1.111***	1.097***	1.107***
	(26.66)	(51.40)	(97.20)	(90.01)	(81.92)
Backward similarity, 0-10 yr	-1.381***	-1.171***	-1.017***	-1.020***	-1.029***
	(-24.40)	(-45.12)	(-83.34)	(-76.93)	(-71.79)
Observation	3,528,612	3,528,612	3,499,566	1,151,385	1,138,754
R^2	0.208	0.266	0.311	0.373	0.399
Application Year FE		Y	Y	Y	Y
Grant Year FE		Y	Y	Y	
Class			Y	Y	Y
Firm FE				Y	
Grant Year × Firm FE					Y

Table reports the results of estimating equation (10) 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 (8) and (9), respectively. As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm and the interaction of firm and year effects. Since patent citations are only consistently recorded after 1945, we restrict the sample to the 1946–2016 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 3: Patent citations and patent quality

Forward citations, 0-1 yr	(1)	(2)	(3)	(4)	(5)
Patent quality ($\log FS/BS$), 0-1 yr	0.515***	0.641***	0.471***	0.547***	0.536***
	(7.33)	(13.53)	(15.93)	(18.58)	(17.87)
Observation	5,156,699	5,156,699	5,121,721	1,775,138	1,758,319
R^2	0.040	0.110	0.145	0.187	0.215
Forward citations, 0-5 yr	(1)	(2)	(3)	(4)	(5)
Patent quality ($\log FS/BS$), 0-5 yr	1.560***	1.470***	1.156***	1.160***	1.147***
	(16.02)	(34.05)	(62.43)	(60.70)	(63.06)
Observation	4,355,594	4,355,594	4,323,134	1,463,683	1,448,640
R^2	0.134	0.227	0.277	0.339	0.365
Forward citations, 0-10 yr	(1)	(2)	(3)	(4)	(5)
Patent quality ($\log FS/BS$), 0-10 yr	1.596***	1.274***	1.082***	1.062***	1.069***
	(22.42)	(47.76)	(82.40)	(87.47)	(81.88)
Observation	3,528,612	3,528,612	3,499,566	1,151,385	1,138,754
R^2	0.190	0.260	0.307	0.370	0.397
Application Year FE		Y	Y	Y	Y
Grant Year FE		Y	Y	Y	
Class			Y	Y	Y
Firm FE				Y	
Grant Year × Firm 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 to our measures of patent quality, which combines the patent's impact and novelty, constructed in equation (11). As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm and the interaction of firm and year effects. Since patent citations are only consistently documented after 1945, we restrict the sample to the 1946–2016 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 citations and patent quality: older patents

Forward citations, full sample	(1)	(2)	(3)	(4)
Patent quality ($\log FS/BS$), 0-1 yr	0.0571**	0.126***	0.356***	0.358***
	(3.22)	(7.71)	(5.65)	(5.54)
Observation	2,406,093	2,402,036	124,286	123,318
R^2	0.228	0.276	0.206	0.243
Forward citations, full sample	(2)	(3)	(4)	(5)
Patent quality ($\log FS/BS$), 0-5 yr	0.132***	0.196***	0.410***	0.433***
	(6.79)	(11.81)	(6.67)	(5.99)
Observation	2,406,101	2,402,044	124,286	123,318
R^2	0.228	0.277	0.207	0.244
Forward citations, full sample	(2)	(3)	(4)	(5)
Patent quality ($\log FS/BS$), 0-10 yr	0.165***	0.218***	0.391***	0.421***
	(8.65)	(14.98)	(6.39)	(5.88)
Observation	2,406,101	2,402,044	124,286	123,318
R^2	0.228	0.277	0.208	0.245
Application Year FE	Y	Y	Y	Y
Grant Year FE	Y	Y	Y	
Class		Y	Y	Y
Firm FE			Y	
Grant Year × Firm FE				Y

Table reports the results of estimating equation (12) in the main text for patents that were issued prior to 1946. The regression relates the log of (one plus) the total number of patent citations over the 1946–2017 period, to our measures of patent quality, which combines the patent's impact and novelty, constructed in equation (11). As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm and the interaction of firm and year effects. We restrict the sample to the 1840–1945 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 5: Patent impact predicts citations

Log cites, 2-5 yr	(1)	(2)	(3)
Log patent quality, 0-1yr	1.324***	1.021***	0.813***
	(13.78)	(19.40)	(22.16)
Log cites, 0-1 yr	0.804***	0.721***	0.658***
	(40.21)	(34.83)	(36.48)
Observation	4355590	4355590	4323130
R^2	0.242	0.294	0.331
Log cites, 6-10 yr	(1)	(2)	(3)
Log patent quality, 0-5yr	0.763***	0.457***	0.440***
	(19.47)	(12.74)	(17.48)
Log cites, 0-5 yr	0.554***	0.532***	0.510***
	(27.29)	(20.49)	(20.50)
Observation	3528612	3528612	3499566
R^2	0.381	0.413	0.436
Log cites, 11-20 yr	(1)	(2)	(3)
Log patent quality, 0-10yr	0.259**	0.210***	0.327***
	(2.33)	(4.71)	(12.51)
Log cites, 0-10 yr	0.491***	0.432***	0.410***
	(27.50)	(22.59)	(22.88)
Observation	2414970	2414970	2392257
R^2	0.247	0.309	0.347
Application Year FE		Y	Y
Grant Year FE		Y	Y
Class			Y

Table reports the results of estimating equation (13) 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 (11) 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. Since patent citations are only consistently documented after 1945, we restrict the sample to the 1946–2016 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 6: Quality of 'historically important' patents

	Fraction of patents in the top				
	50 %	75 %	90 %	95 %	99 %
Quality, 0–1 years forward	0.67	0.48	0.29	0.16	0.04
Quality, 0–5 years forward	0.82	0.62	0.36	0.18	0.05
Quality, 0–10 years forward	0.88	0.66	0.31	0.23	0.10
Quality, 0–20 years forward	0.94	0.78	0.43	0.26	0.14
Citations, 0–1 years forward	0.12	0.12	0.07	0.07	0.04
Citations, 0–5 years forward	0.21	0.17	0.13	0.12	0.05
Citations, 0–10 years forward	0.21	0.21	0.17	0.13	0.07
Citations, 0–20 years forward	0.25	0.24	0.16	0.13	0.07
Citations, full sample	0.45	0.33	0.21	0.16	0.10

Table reports the results of the external validation exercise described in Section 2.2 in the paper. Specifically, we obtain a list of 110 'historically important' patents issued in the 1840–1962 period, compiled by the patent historian Jim Bieberich, available through USPAT.COM (http://www.uspat.com/historical/index.shtml). We then report the fraction of these patents that fall in the top x% in terms of quality using our measure(s) as well as patent citations. The breakpoints are computed using the unconditional distribution over the entire sample.

Table 7: Patent impact and value

KPSS value	(1)	(2)	(3)	(4)	(5)
Log patent quality, 0-1 years	0.893***	0.451***	0.397***	0.190***	0.0692***
	(2.98)	(2.79)	(5.20)	(4.59)	(4.11)
Observations	585819	585819	480250	478935	461791
R^2	0.006	0.054	0.199	0.825	0.958
KPSS value	(1)	(2)	(3)	(4)	(5)
Log patent quality, 0-5 years	0.926***	0.452***	0.403***	0.171***	0.0901***
	(4.31)	(3.04)	(6.32)	(5.29)	(4.99)
Observations	511331	511331	416581	415329	399983
R^2	0.011	0.051	0.200	0.831	0.959
KPSS value	(1)	(2)	(3)	(4)	(5)
Log patent quality, 0-10 years	0.409***	0.148	0.233***	0.118***	0.0709***
	(2.82)	(1.06)	(3.81)	(5.46)	(9.01)
Observations	431085	431085	349488	348328	335435
R^2	0.004	0.040	0.194	0.831	0.959
Grant Year FE		Y	Y	Y	
Class			Y	Y	Y
Firm FE				Y	
Grant Year × Firm FE					Y

Table reports the results of estimating equation (16) in the main text. The regression relates the log of the Kogan et al. (2016) estimate of the market value of the patent to our measures of patent quality, which combines the patent's impact and novelty, constructed in equation (11). As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm and the interaction of firm and year effects. Since patent citations are only consistently documented after 1945, we restrict the sample to the 1946–2016 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 8: Patent impact and value (cont)

KPSS value	(1)	(2)	(3)	(4)	(5)
Patent quality, 0-1 years	0.721**	0.385**	0.345***	0.189***	0.0784***
	(2.53)	(2.42)	(4.51)	(4.64)	(4.48)
Log Cites, 0-1 years	0.250***	0.106***	0.132***	0.00408	-0.0391***
	(7.24)	(4.70)	(12.06)	(0.46)	(-20.44)
Observations	585819	585819	480250	478935	461791
R^2	0.008	0.054	0.200	0.825	0.958
KPSS value	(1)	(2)	(3)	(4)	(5)
Patent quality, 0-5 years	0.652***	0.358**	0.326***	0.171***	0.108***
	(3.66)	(2.45)	(5.15)	(5.31)	(6.04)
Log Cites, 0-5 years	0.153***	0.0660***	0.0778***	0.000830	-0.0263***
	(4.85)	(3.42)	(7.38)	(0.28)	(-23.30)
Observations	511331	511331	416581	415329	399983
R^2	0.015	0.052	0.201	0.831	0.959
KPSS value	(1)	(2)	(3)	(4)	(5)
Patent quality, 0-10 years	0.224*	0.103	0.187***	0.116***	0.0888***
	(1.83)	(0.71)	(3.05)	(5.28)	(11.03)
Log Cites, 0-10 years	0.105***	0.0348***	0.0447***	0.00168	-0.0214***
	(3.22)	(2.73)	(7.43)	(0.67)	(-17.03)
Observations	431085	431085	349488	348328	335435
R^2	0.006	0.040	0.194	0.831	0.959
Grant Year FE		Y	Y	Y	
Class			Y	Y	Y
Firm FE				Y	
Grant Year × Firm FE					Y

Table reports the results of estimating a modified version of equation (16) in the main text. The regression relates the log of the Kogan et al. (2016) estimate of the market value of the patent to our measures of patent quality, which combines the patent's impact and novelty, constructed in equation (11) while controlling for the log of (one plus) the number of forward citations. As additional controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm and the interaction of firm and year effects. Since patent citations are only consistently documented after 1945, we restrict the sample to the 1946–2016 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 9: Market Value as a Function of R&D, Patents and Similarity Stocks (All patenting firms)

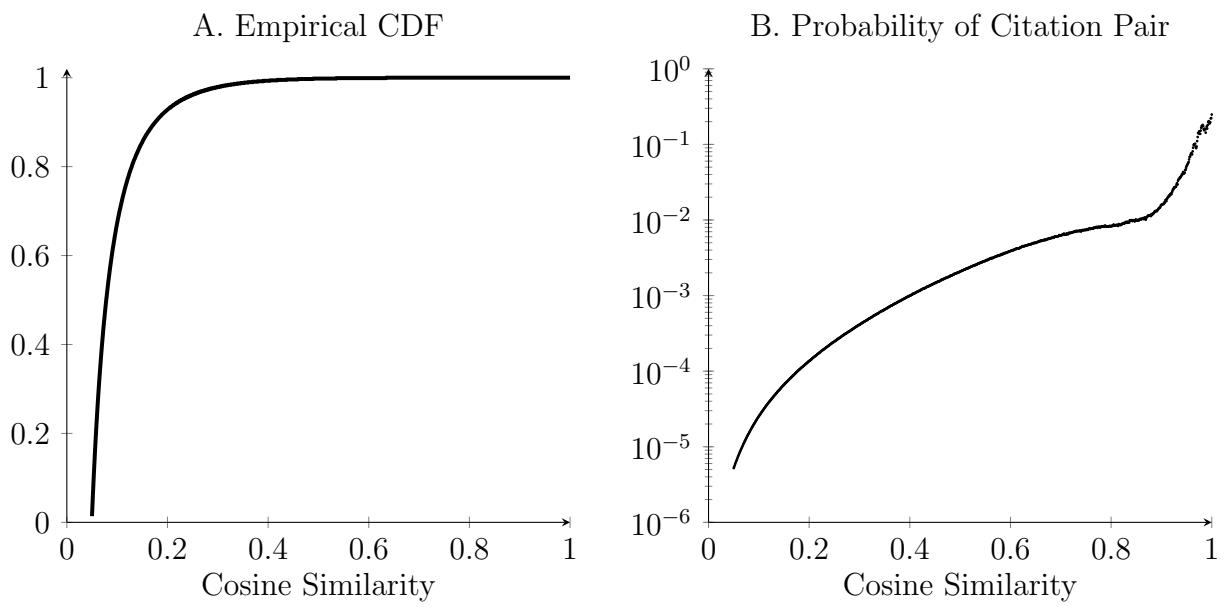
$\log Q$	(1)	(2)	(3)	(4)
Horizon τ	(0,1)	(0,5)	(0,10)	(0,20)
$SRD_{f,t}/A_{f,t}$	0.830***	0.924***	1.007***	1.234***
	(14.59)	(14.01)	(12.28)	(8.48)
$SPAT_{f,t}/SRD_{f,t}$	0.001***	0.013***	0.109	0.333**
	(17.28)	(9.89)	(1.20)	(2.04)
$SCIT_{f,t}/SPAT_{f,t}$	0.103***	0.033***	0.021***	0.017***
	(6.33)	(10.63)	(12.60)	(9.48)
$SRSIM_{f,t}/SPAT_{f,t}$	4.866***	0.672***	0.258***	0.076***
	(12.02)	(10.94)	(11.12)	(8.67)
$D(SRD = 0)$	-0.076***	-0.052***	-0.030***	0.018
	(-7.67)	(-5.20)	(-2.85)	(1.45)
Normalized coefficients: $SCIT_{f,t}/SPAT_{f,t}$	0.112	0.321	0.465	0.665
Normalized coefficients: $SRSIM_{f,t}/SPAT_{f,t}$	0.210	0.231	0.263	0.231
N	93,739	82,036	68,295	41,241
R^2	0.479	0.483	0.472	0.350

t statistics in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

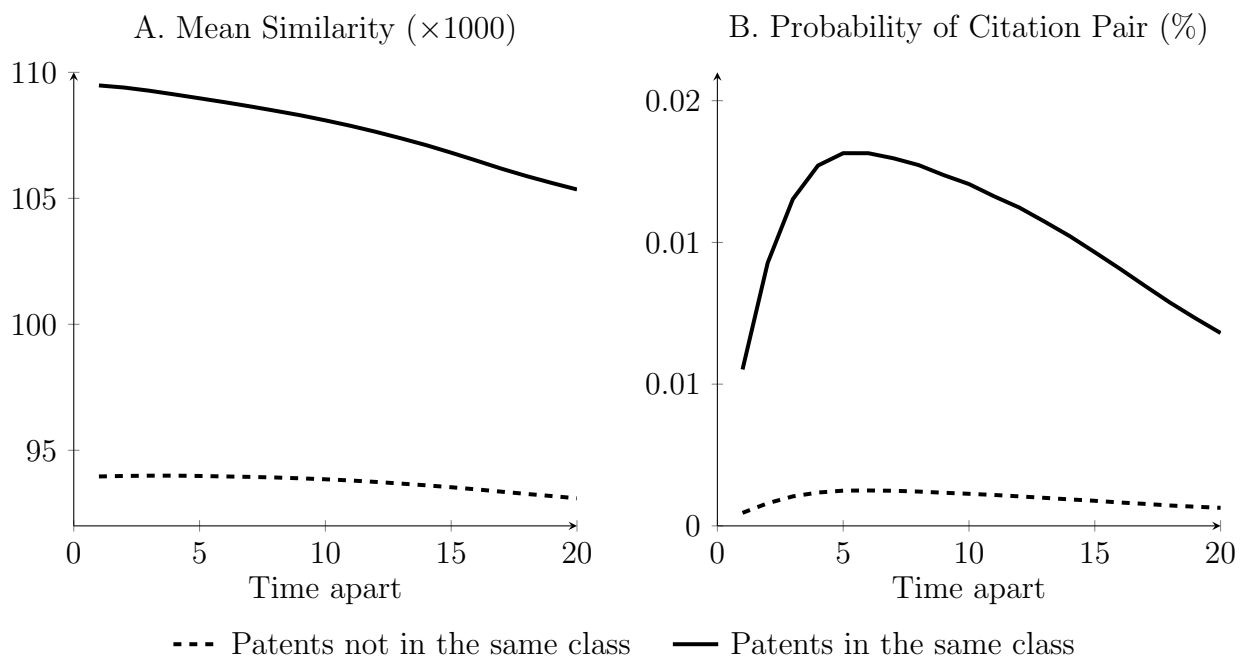
Table reports estimates of equation (19) in the text. The equation relates the logarithm of a firm's Tobin's Q to the stocks of R&D expenditure ($SRD_{f,t}$), number of patents ($SPAT_{f,t}$), patent citations ($SCITES_{f,t}$), and the patent quality measures ($SRSIM_{f,t}$) — constructed as in (17) 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.

Figure 1: Pairwise similarity and citation linkages



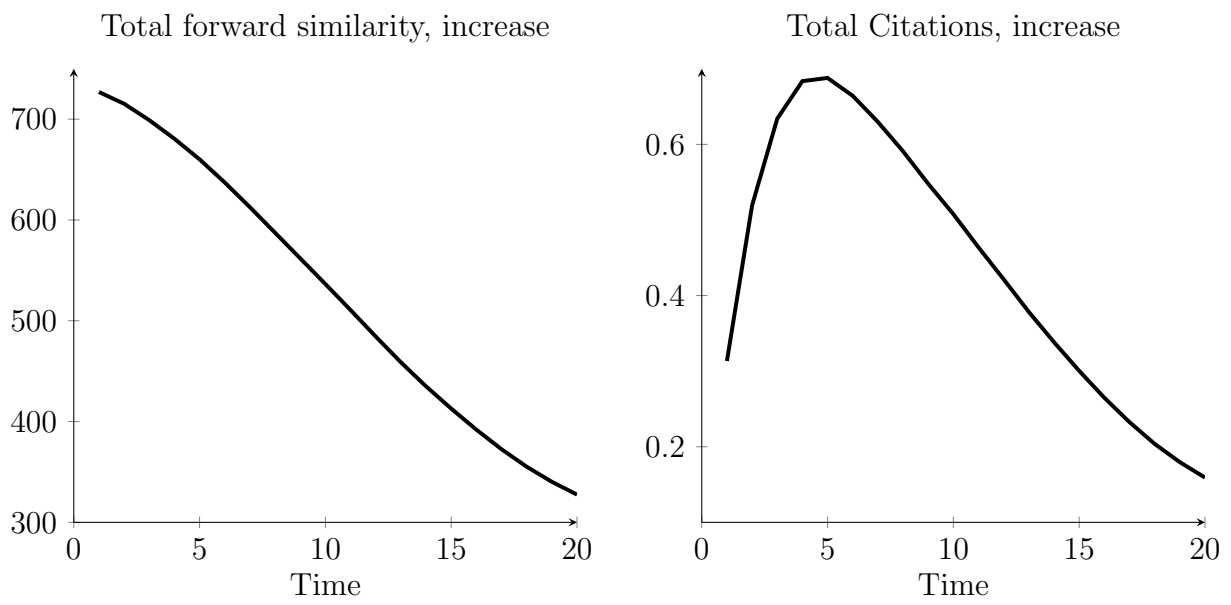
The figure on the left panel (A) plots the empirical CDF of our similarity measure $\rho_{i,j}$ across patent citation pairs. The figure on the right panel plots the conditional probability that patent j cites an earlier patent j as a function of the text-based similarity score between the two patents,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

Figure 2: Pairwise similarity and citation linkages over time and across tech class



The left panel plots the mean similarity across patent pairs i and j as a function of the distance in filing years between the two patents, and whether the two patents belong in the same tech class or not. The right performs the same exercise for the mean number of citations across pairs. Similarity refers to 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 $\rho_{i,j} \leq 0.5\%$. Panel B uses data only post 1945, since citations were not consistently recorded prior to that year.

Figure 3: Patent Citations and Forward Similarity Lags



The left panel plots the mean increase in the total forward similarity as a function of years following the patent filing date. The right panel performs the same exercise for mean citation counts; for the right panel, we restrict attention to the 1946-2016 sample, since prior to 1945, citations are not consistently recorded in patent documents.

Figure 4: Distribution of patent quality and citations over time

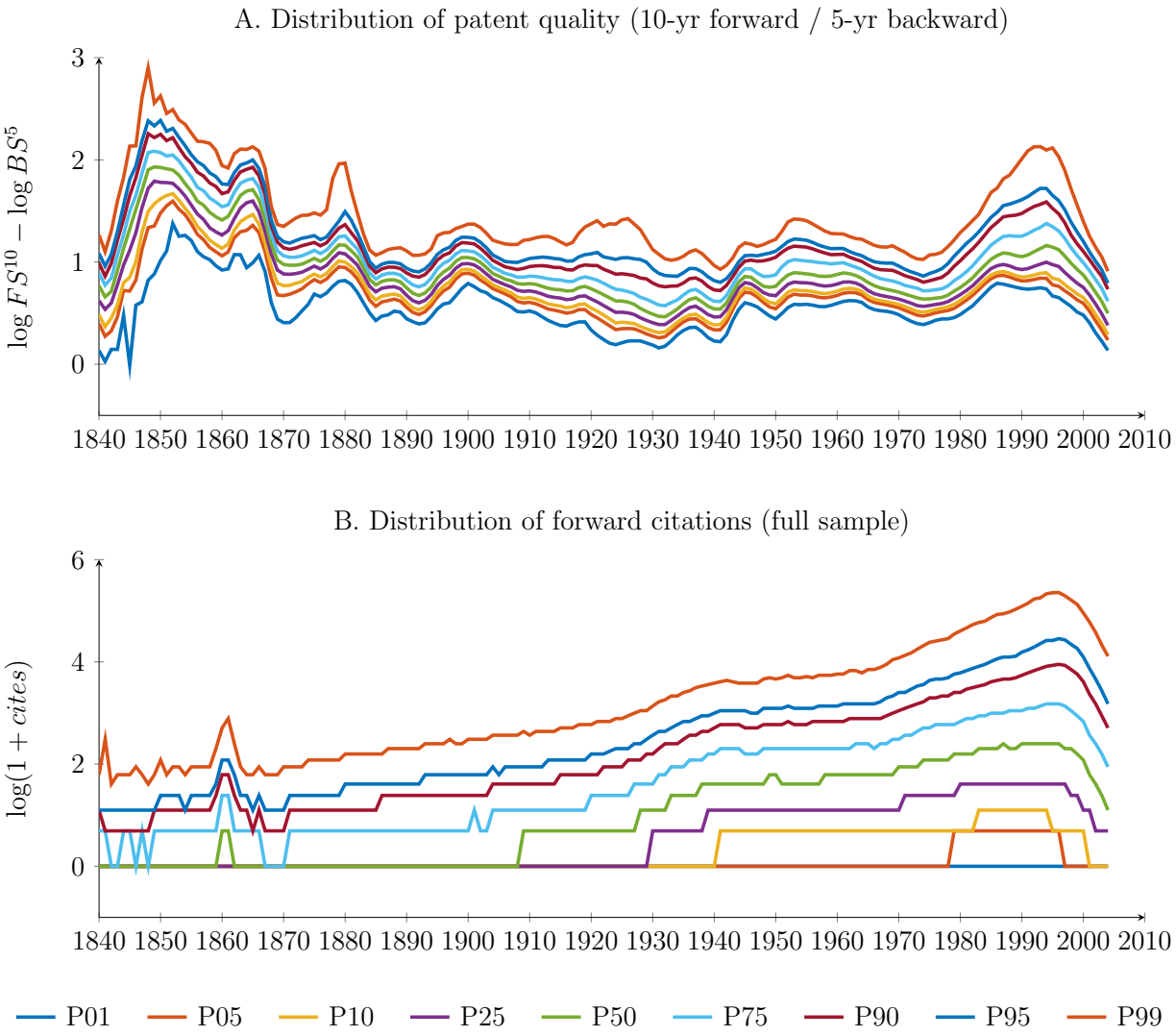


Figure plots the cross-sectional distribution of our patent quality measure (estimated using 10-year forward similarity, in panel A) and the number of forward citations (using the entire sample, in panel B).

Figure 5: Patent quality and citations

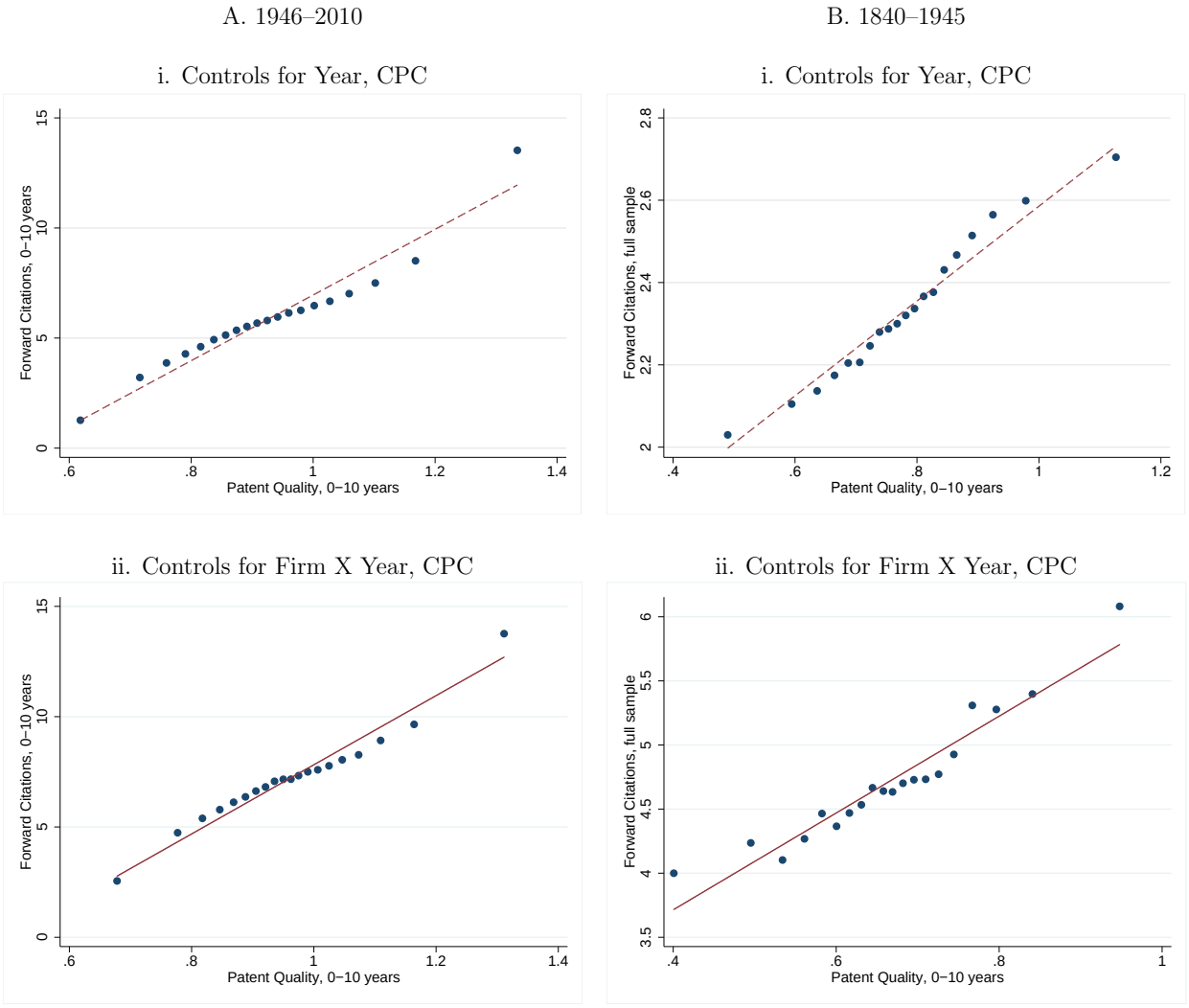


Figure presents the regression results of Tables 3 (panel A) and 4 (panel B) as a binned scatterplot.

Figure 6: Breakdown by Technology Classes

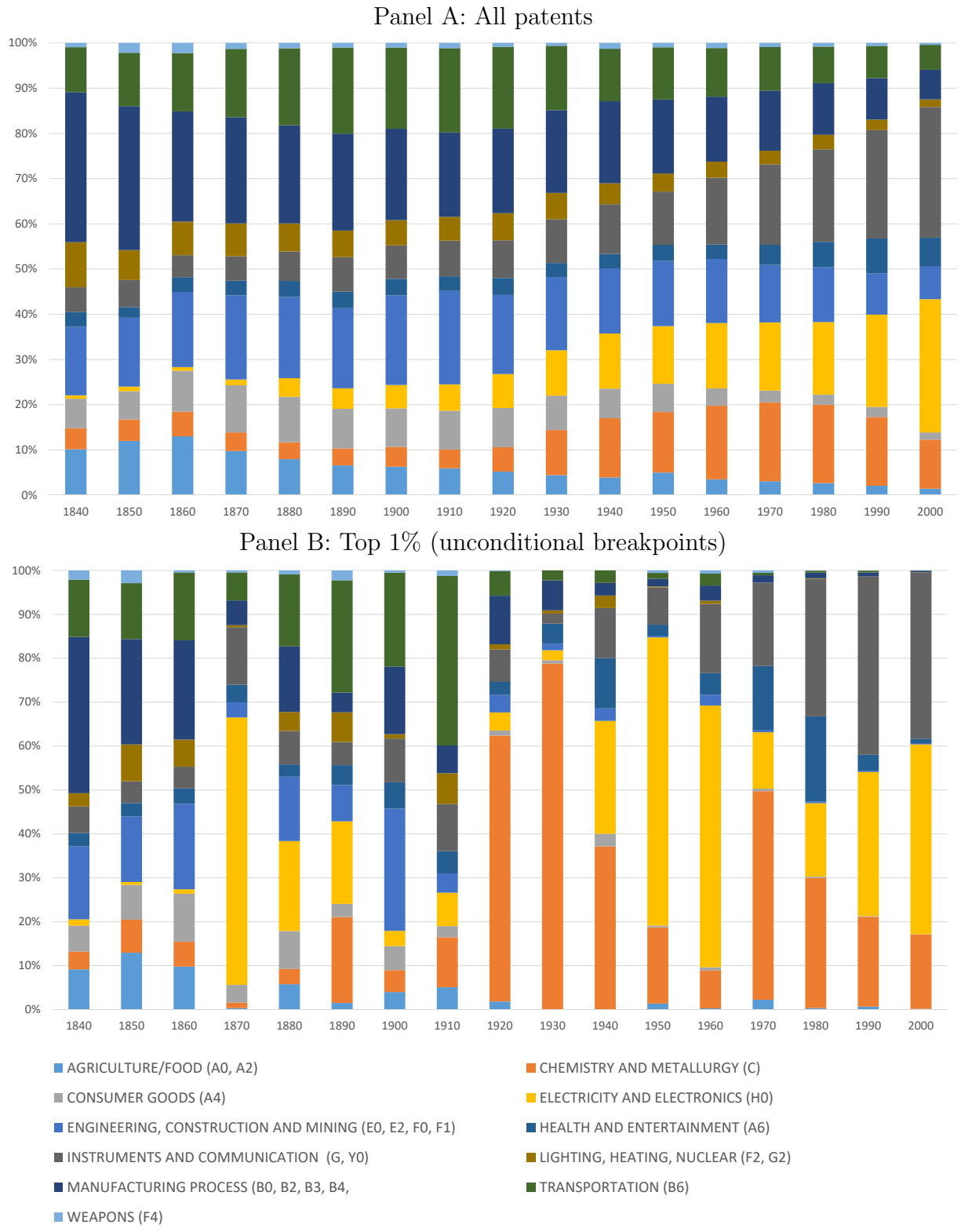
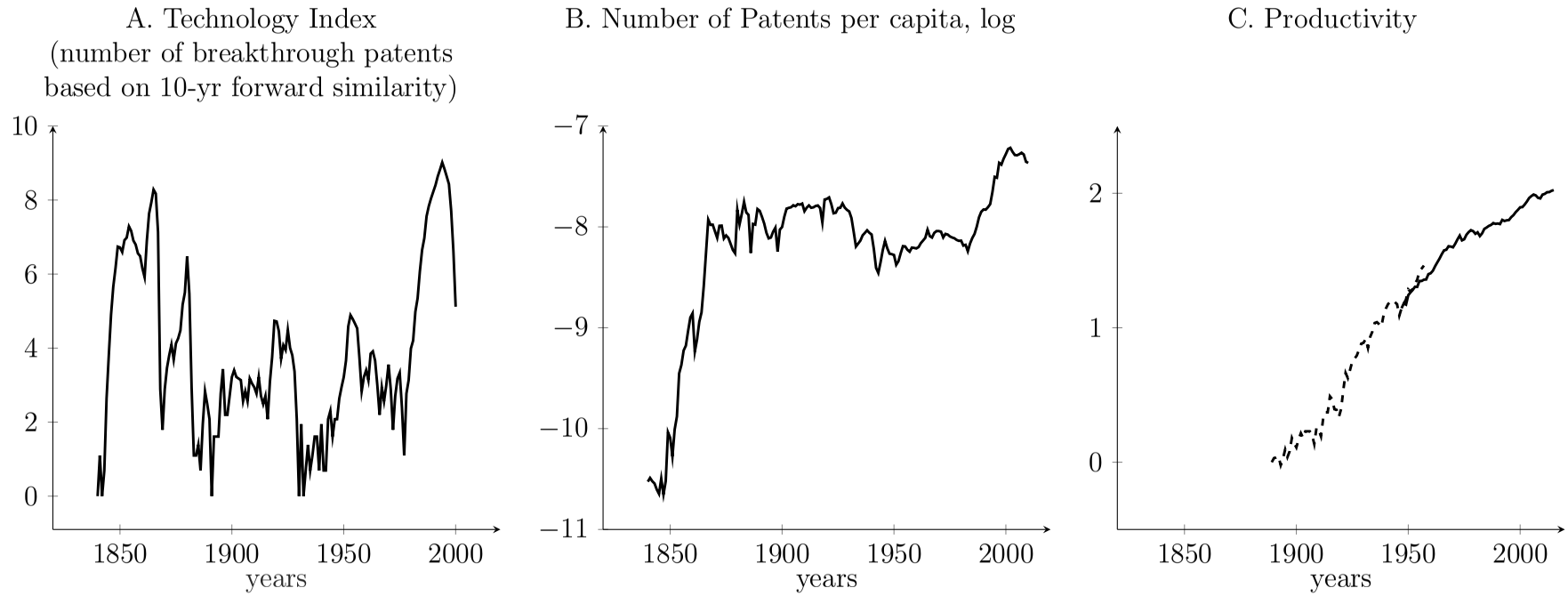
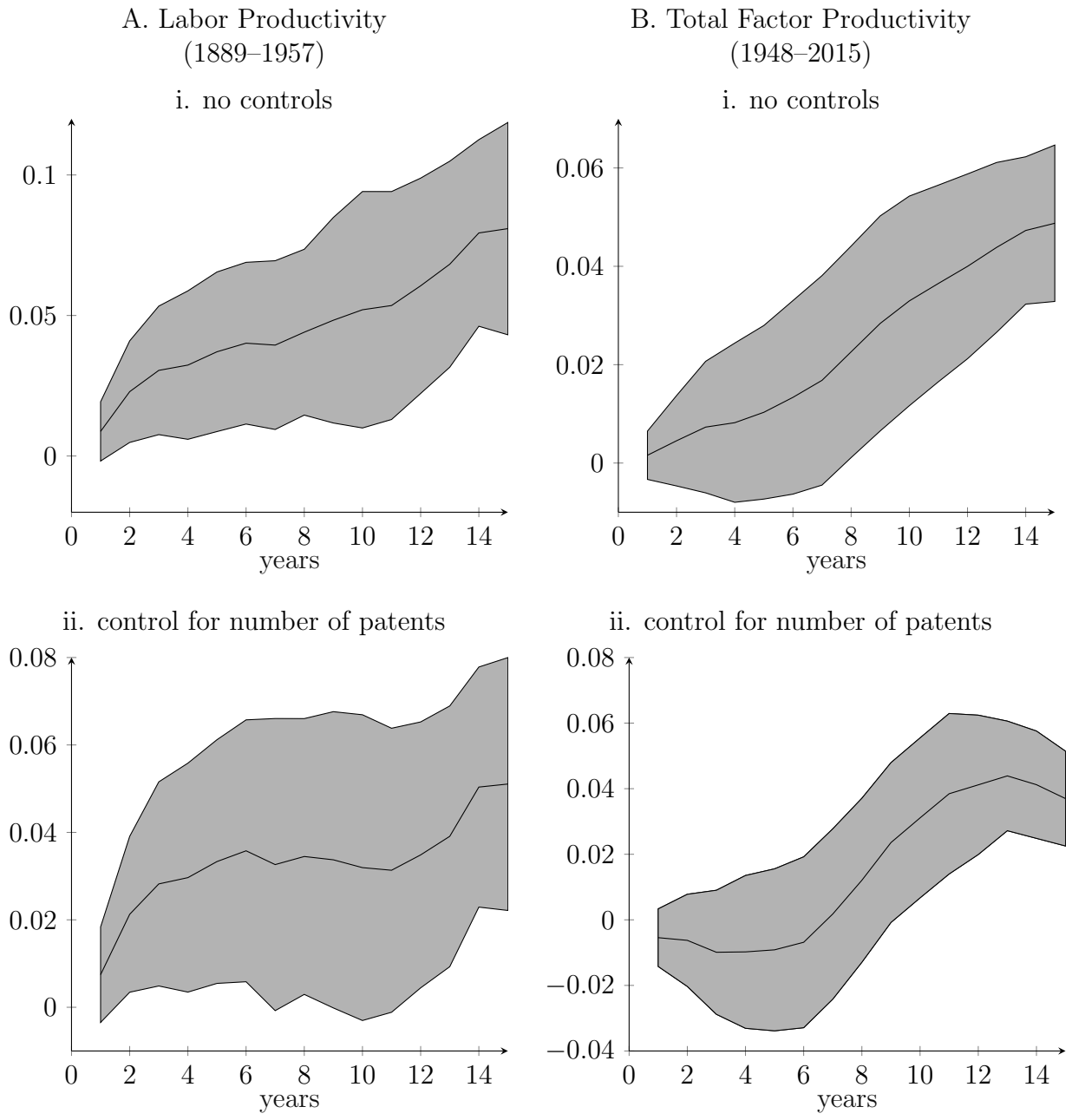


Figure 7: Index of technological progress and productivity growth



Figures plot the technology indices computed using 10-yr forward similarity (Panel A), along with the total number of (successful) patent applications divided by population in each year (Panel B), and measures of productivity (Panel C). The technology indices represent the logarithm of one plus the number of patents that have quality scores in the top 1% of the unconditional distribution (see equation (14) in the main text). The series on productivity in the 1889 to 1957 period (dotted line) is output per effective labor input in the Manufacturing sector, and is from Kendrick (1961) (pages 465–466). The series on total factor productivity during the 1948 to 2015 period (solid line) is from Basu et al. (2006), and available through the website of the San Francisco FRB.

Figure 8: Index of technological progress and future productivity growth



This set of figures plots the estimated coefficients $B(H)$ from equation (15) in the main text, which correspond to the response of measured productivity on our technology index. The series on labor productivity in the 1889 to 1957 period is from Kendrick (1961), pages 465–466. The series on total factor productivity during the 1948 to 2015 period is from Basu et al. (2006), available through the San Francisco FRB. The top row presents responses to a one-standard deviation shock to our technology index plotted in Panel A of Figure 7, that is, the log of one plus the number of patents that fall in the top 1% of the unconditional distribution of patent quality, as measured by $FS^{0,10}/BS^{0,5}$. The bottom row includes controls for the total number of patents per capita that are filed in each year. Standard errors are adjusted for overlapping observations following Hodrick (1992).

Additional Tables for Appendix

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: Patent quality predicts citations

Log cites, 2-5 yr	(1)	(2)	(3)
$\log FS$, 0-1yr	1.244***	1.019***	0.848***
	(14.69)	(19.02)	(22.46)
$\log BS$, 0-5yr	-1.124***	-0.949***	-0.785***
	(-12.98)	(-17.74)	(-20.80)
Log cites, 0-1 yr	0.765***	0.707***	0.652***
	(45.08)	(33.36)	(35.66)
Observations	4355590	4355590	4323130
R^2	0.259	0.295	0.330
Log cites, 6-10 yr	(1)	(2)	(3)
$\log FS$, 0-5yr	0.757***	0.457***	0.450***
	(18.99)	(12.78)	(17.40)
$\log BS$, 0-5yr	-0.736***	-0.462***	-0.431***
	(-18.61)	(-13.44)	(-16.95)
Log cites, 0-5 yr	0.550***	0.533***	0.508***
	(26.01)	(20.16)	(20.33)
Observations	3528612	3528612	3499566
R^2	0.381	0.403	0.427
Log cites, 11-20 yr	(1)	(2)	(3)
$\log FS$, 0-10yr	0.258**	0.203***	0.325***
	(2.33)	(4.95)	(12.70)
$\log BS$, 0-5yr	-0.263**	-0.249***	-0.328***
	(-2.31)	(-5.50)	(-12.42)
Log cites, 0-10 yr	0.491***	0.437***	0.411***
	(26.65)	(22.24)	(22.80)
Observations	2414970	2414970	2392257
R^2	0.247	0.307	0.344
Grant Year FE		Y	Y
Class			Y

This table is the counterpart to Table 5 in the main text, in which we disaggregate our patent quality measure into impact (forward similarity) and novelty (the inverse of backward similarity), constructed in equations (8) and (9), respectively. The regression relates the log of (one plus) the number of patent citations over a horizon $[t, s]$ to our measures of patent impact and novelty 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. Since patent citations are only consistently documented after 1945, we restrict the sample to the 1946–2016 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 A.3: Patent impact and value

KPSS value	(1)	(2)	(3)	(4)	(5)
FS, 0-1 years	0.872***	0.451***	0.354***	0.176***	0.0553***
	(3.03)	(2.77)	(4.42)	(4.19)	(3.36)
BS, 0-5 years	-0.800***	-0.598***	-0.417***	-0.198***	-0.0772***
	(-2.84)	(-4.00)	(-5.52)	(-4.84)	(-4.66)
Observations	585819	585819	480250	478935	461791
R^2	0.007	0.057	0.199	0.825	0.958
KPSS value	(1)	(2)	(3)	(4)	(5)
FS, 0-5 years	0.905***	0.455***	0.381***	0.158***	0.0764***
	(4.49)	(3.01)	(5.68)	(4.91)	(4.32)
BS, 0-5 years	-0.856***	-0.554***	-0.415***	-0.178***	-0.0980***
	(-4.75)	(-3.91)	(-6.54)	(-5.53)	(-5.51)
Observations	511331	511331	416581	415329	399983
R^2	0.011	0.053	0.200	0.831	0.959
KPSS value	(1)	(2)	(3)	(4)	(5)
FS, 0-10 years	0.403***	0.149	0.218***	0.104***	0.0599***
	(2.91)	(1.06)	(3.52)	(4.84)	(7.77)
BS, 0-5 years	-0.377***	-0.228*	-0.247***	-0.130***	-0.0810***
	(-3.16)	(-1.76)	(-3.98)	(-5.99)	(-9.97)
Observations	431085	431085	349488	348328	335435
R^2	0.004	0.041	0.194	0.832	0.959
Grant Year FE		Y	Y	Y	
Class			Y	Y	Y
Firm FE				Y	
Grant Year × Firm FE					Y

This Table is the counterpart to Table 7 in the main text, in which we disaggregate our measure of patent quality into patent impact (forward similarity) and lack of novelty (inverse of backward similarity) constructed in equations (8) and (9), respectively. Table reports the results of estimating equation (16) in the main text. The regression relates the log of the Kogan et al. (2016) estimate of the market value of the patent to our measures of patent quality, which combines the patent's impact and novelty, constructed in equation (11). As controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm and the interaction of firm and year effects. Since patent citations are only consistently documented after 1945, we restrict the sample to the 1946–2016 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 A.4: Patent impact and value (cont)

KPSS value	(1)	(2)	(3)	(4)	(5)
FS, 0-1 years	0.721**	0.368**	0.297***	0.174***	0.0648***
	(2.60)	(2.28)	(3.70)	(4.23)	(3.78)
BS, 0-5 years	-0.669**	-0.522***	-0.365***	-0.197***	-0.0858***
	(-2.45)	(-3.54)	(-4.79)	(-4.89)	(-4.99)
Log Cites, 0-1 years	0.227***	0.136***	0.137***	0.00547	-0.0379***
	(8.11)	(5.52)	(12.42)	(0.61)	(-19.52)
Observations	585819	585819	480250	478935	461791
R^2	0.009	0.058	0.200	0.825	0.958
KPSS value	(1)	(2)	(3)	(4)	(5)
FS, 0-5 years	0.652***	0.342**	0.295***	0.156***	0.0952***
	(3.66)	(2.27)	(4.39)	(4.91)	(5.41)
BS, 0-5 years	-0.642***	-0.451***	-0.337***	-0.177***	-0.115***
	(-3.98)	(-3.24)	(-5.33)	(-5.52)	(-6.50)
Log Cites, 0-5 years	0.150***	0.0801***	0.0812***	0.00208	-0.0251***
	(5.15)	(3.90)	(7.51)	(0.71)	(-21.81)
Observations	511331	511331	416581	415329	399983
R^2	0.015	0.054	0.201	0.831	0.959
KPSS value	(1)	(2)	(3)	(4)	(5)
FS, 0-10 years	0.224*	0.0933	0.166**	0.101***	0.0777***
	(1.82)	(0.64)	(2.67)	(4.61)	(9.86)
BS, 0-5 years	-0.221**	-0.177	-0.200***	-0.128***	-0.0971***
	(-2.03)	(-1.32)	(-3.22)	(-5.76)	(-11.81)
Log Cites, 0-10 years	0.104***	0.0430***	0.0472***	0.00328	-0.0203***
	(3.34)	(3.16)	(7.42)	(1.33)	(-15.78)
Observations	431085	431085	349488	348328	335435
R^2	0.006	0.040	0.194	0.831	0.959
Grant Year FE		Y	Y	Y	
Class			Y	Y	Y
Firm FE				Y	
Grant Year × Firm FE					Y

This Table is the counterpart to Table 8 in the main text, in which we disaggregate our measure of patent quality into patent impact (forward similarity) and lack of novelty (inverse of backward similarity) constructed in equations (8) and (9), respectively. Table reports the results of estimating a modified version of equation (16) in the main text. The regression relates the log of the Kogan et al. (2016) estimate of the market value of the patent to our measures of patent quality, which combines the patent's impact and novelty, constructed in equation (11). We control for the number of citations the patent receives over that period. As additional controls, we include dummies controlling for technology class (defined at the 3-digit CPC level), grant year, firm and the interaction of firm and year effects. Since patent citations are only consistently documented after 1945, we restrict the sample to the 1946–2016 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 A.5: Patent quality: Historically important patents

Patent	Year	Inventor	Patent Name	Cites	Percentile Rank				
					Patent quality				Citations
					(0-1)	(0-5)	(0-10)	(0-20)	(full)
1647	1840	Samuel F. B. Morse	Morse Code	2	1.8	0.3	5.0	93.1	55.7
3633	1844	Charles Goodyear	Vulcanized Rubber	3	94.2	98.0	99.1	99.9	65.1
4750	1846	Elias Howe, Jr.	Sewing Machine	1	1.0	98.2	99.9	100.0	41.7
4834	1846	Benjamin F. Palmer	Artificial Limb	0	17.1	96.6	99.0	99.9	
4848	1846	Charles T. Jackson	Anesthesia	0	90.6	94.7	98.6	99.9	
4874	1846	Christian F. Schonbein	Guncotton	0	96.4	96.7	98.4	99.7	
5199	1847	Richard M. Hoe	Rotary Printing Press	0	63.7	97.7	99.6	99.6	
6281	1849	Walter Hunt	Safety Pin	0	99.2	99.7	100.0	100.0	
9300	1852	Lorenzo L. Langstroth	Beehive	1	82.4	95.9	99.9	100.0	41.7
13661	1855	Isaac M. Singer	Shuttle Sewing Machine	1	80.1	98.6	97.6	99.0	41.7
15553	1856	Gail Borden, Jr.	Condensed Milk	0	73.9	99.1	99.7	99.8	
17628	1857	William Kelly	Iron and Steel Manuf.	0	98.6	98.0	99.3	99.4	
26196	1859	James J. Mapes	Artificial Fertilizer	1	96.2	94.7	99.4	99.2	41.7
31128	1861	Elisha Graves Otis	Elevator	1	18.7	95.8	98.5	98.2	41.7
31278	1861	Linus Yale, Jr.	Lock	10	2.3	84.4	98.1	98.2	90.7
36836	1862	Richard J. Gatling	Machine Gun	2	95.3	97.8	97.8	97.9	55.7
59915	1866	Pierre Lallement	Bicycle	0	99.9	99.9	99.5	98.9	
78317	1868	Alfred Nobel	Dynamite	3	98.6	93.5	75.5	80.8	65.1
79265	1868	C. Latham Sholes	Typewriter	1	98.3	97.7	96.8	97.4	41.7
79965	1868	Alvin J. Fellows	Spring Tape Measure	2	86.0	83.6	90.0	94.5	55.7
91145	1869	Ives W. McGaffey	Vacuum Cleaner	3	78.8	88.5	83.6	90.2	65.1
110971	1871	Andrew Smith Hallidie	Cable Car	0	80.3	84.5	88.9	95.9	
135245	1873	Louis Pasteur	Pasteurization	0	33.9	28.5	64.4	64.3	
157124	1874	Joseph F. Glidden	Barbed Wire	1	49.0	91.9	97.3	97.4	41.7
174465	1876	Alexander Graham Bell	Telephone	4	8.3	99.3	99.8	99.6	71.9
194047	1877	Nicolaus August Otto	Internal Combustion Engine	1	13.0	69.0	82.9	87.8	41.7
200521	1878	Thomas Alva Edison	Phonograph	10	87.8	96.6	96.1	95.3	90.7
223898	1880	Thomas Alva Edison	First Incandescent Light	20	99.8	99.9	99.5	98.7	97.7
224573	1880	Emile Berliner	Microphone	0	97.4	95.7	97.8	99.2	
237664	1881	Frederic E. Ives	Halftone Printing Plate	0	93.5	95.4	95.4	91.9	
304272	1884	Ottmar Mergenthaler	Linotype	0	96.6	93.6	95.8	94.7	
347140	1886	Elihu Thomson	Electric Welder	2	5.9	73.6	68.6	69.8	55.7
371496	1887	Dorr E. Felt	Adding Machine	2	70.7	90.5	81.0	77.7	55.7
372786	1887	Emile Berliner	Phonograph Record	1	90.9	93.2	82.3	96.5	41.7
373064	1887	Carl Gassner, Jr.	Dry Cell Battery	3	86.2	81.8	34.1	31.8	65.1
382280	1888	Nikola Tesla	A. C. Induction Motor	1	88.7	96.2	89.1	93.7	41.7
388116	1888	William S. Burroughs	Calculator	1	10.8	87.5	85.1	80.0	41.7
388850	1888	George Eastman	Roll Film Camera	1	98.5	96.1	92.9	95.0	41.7
395782	1889	Herman Hollerith	Computer	0	29.7	53.4	63.8	66.2	16.7
400664	1889	Charles M. Hall	Aluminum Manufacture	12	5.7	56.6	63.2	79.8	93.2
430212	1890	Hiram Stevens Maxim	Smokeless Gunpowder	0	90.8	74.4	64.0	79.6	
468226	1892	William Painter	Bottle Cap	5	66.3	85.1	82.7	90.4	77.2
492767	1893	Edward G. Acheson	Carborundum	11	14.9	10.9	30.0	52.2	92.1
493426	1893	Thomas A. Edison	Motion Picture	1	44.7	65.0	86.3	94.4	41.7
504038	1893	Whitcomb L. Judson	Zipper	5	24.0	23.2	30.0	62.9	77.2
549160	1895	George B. Selden	Automobile	0	33.2	58.6	79.1	86.2	
558936	1896	Joseph S. Duncan	Addressograph	2	11.9	13.7	39.7	71.6	55.7
586193	1897	Guglielmo Marconi	Radio	0	10.2	84.4	90.9	90.4	
589168	1897	Thomas A. Edison	Motion Picture Camera	0	46.3	42.9	73.9	89.0	
608845	1898	Rudolf Diesel	Diesel Engine	8	63.5	76.1	86.4	90.4	86.9
621195	1899	Ferdinand Graf Zepplin	Dirigible	1	56.3	87.7	81.9	81.2	41.7

The table reports the full list of patents used in the validation exercise in Section 2.2 in the paper. In addition to the number of forward citations, we report for each patent its rank percentile in terms of our quality measure(s) and the number of forward patent citations. Source: http://www.uspat.com/historical/index.shtml.

Table A.6: Patent quality: Historically important patents (continued)

Patent	Year	Inventor	Patent Name	Cites	Percentile Rank				
					Patent quality				Citations
					(0-1)	(0-5)	(0-10)	(0-20)	(full)
644077	1900	Felix Hoffmann	Aspirin	1	92.8	91.9	81.0	84.2	41.7
661619	1900	Valdemar Poulsen	Magnetic Tape Recorder	7	63.0	93.9	92.1	87.8	84.4
708553	1902	John P. Holland	Submarine	1	91.4	89.9	84.7	90.5	41.7
745157	1903	Clyde J. Coleman	Electric Starter	1	96.4	96.4	95.9	96.1	41.7
766768	1904	Michael J. Owens	Glass Bottle Manuf.	0	91.5	89.6	85.5	79.6	
808897	1906	Willis H. Carrier	Air Conditioning	20	52.6	70.1	76.1	82.2	97.7
821393	1906	Orville Wright	Airplane	18	97.9	99.9	100.0	100.0	97.0
841387	1907	Lee De Forest	Triode Vacuum Tube	4	41.2	19.9	35.9	61.6	71.9
942809	1909	Leo H. Baekeland	Bakelite	1	92.8	95.0	95.1	97.6	41.7
971501	1910	Fritz Haber	Ammonia Production	0	98.6	98.9	98.4	98.8	
1005186	1911	Henry Ford	Automotive Transmission	2	48.9	64.4	69.8	71.3	55.7
1008577	1911	Ernst Alexanderson	High Frequency Generator	0	17.9	37.0	60.3	84.1	
1030178	1912	Peter Cooper Hewitt	Mercury Vapor Lamp	1	88.8	93.7	92.5	94.6	41.7
1082933	1913	William D. Coolidge	Tungsten Filament Light Bulb	17	82.8	84.0	82.7	87.3	96.6
1102653	1914	Robert H. Goddard	Rocket	55	15.1	57.1	50.9	47.2	99.8
1113149	1914	Edwin H. Armstrong	Wireless Receiver	7	85.6	91.8	94.2	96.6	84.4
1115674	1914	Mary P. Jacob	Brassiere	1	91.2	74.2	63.7	45.9	41.7
1180159	1916	Irving Langmuir	Gas Filled Electric Lamp	3	93.7	87.6	87.1	86.8	65.1
1203495	1916	William D. Coolidge	X-Ray Tube	4	83.8	78.3	85.8	94.2	71.9
1279471	1918	Elmer A. Sperry	Gyroscopic Compass	8	92.8	96.3	97.4	96.9	86.9
1413121	1922	John Arthur Johnson	Adjustable Wrench	0	48.1	13.0	9.0	3.2	
1420609	1922	Glenn H. Curtiss	Hydroplane	2	87.6	80.7	77.9	68.7	55.7
1573846	1926	Thomas Midgley, Jr.	Ethyl Gasoline	2	32.9	38.6	43.6	35.4	55.7
1773080	1930	Clarence Birdseye	Frozen Food	18	71.3	83.0	70.3	47.3	97.0
1773980	1930	Philo T. Farnsworth	Television	8	90.9	94.9	91.0	88.1	86.9
1848389	1932	Igor Sikorsky	Helicopter	5	93.6	55.1	49.7	46.2	77.2
1941066	1933	Edwin H. Armstrong	FM Radio	0	75.5	44.7	65.6	73.2	
1948384	1934	Ernest O. Lawrence	Cyclotron	96	23.1	31.8	46.9	51.1	100.0
2021907	1935	Vladimir K. Zworykin	Television	16	33.7	44.9	66.0	65.4	96.1
2059884	1936	Leopold D. Mannes	Color Film	11	9.4	23.9	31.9	29.0	92.1
2071250	1937	Wallace H. Carothers	Nylon	186	61.7	72.4	87.8	85.0	100.0
2153729	1939	Ernest H. Volwiler	Pentothal	2	58.4	88.3	76.5	73.6	55.7
2206634	1940	Enrico Fermi	Radioactive Isotopes	97	42.4	88.8	86.9	89.9	100.0
2297691	1942	Chester F. Carlson	Xerography	736	66.0	10.1	16.2	44.7	100.0
2329074	1943	Paul Muller	DDT - Insecticide	48	9.9	8.6	20.9	53.3	99.8
2404334	1946	Frank Whittle	Jet Engine	35	3.9	17.1	22.7	32.8	99.4
2451804	1948	Donald L. Campbell	Fluid Catalytic Cracking	9	89.0	74.5	73.9	80.6	89.0
2524035	1950	John Bardeen	Transistor	132	51.4	69.5	87.5	95.0	100.0
2543181	1951	Edwin H. Land	Instant Photography	116	32.4	52.3	71.5	93.5	100.0
2569347	1951	William Shockley	Junction Transistor	140	32.2	52.6	74.0	84.5	100.0
2668661	1954	George R. Stibitz	Modern Digital Computer	14	98.8	97.0	98.0	99.4	94.9
2682050	1954	Andrew Alford	Radio Navigation System	3	32.3	72.8	78.2	86.0	65.1
2682235	1954	Richard B. Fuller	Geodesic Dome	86	30.0	56.5	67.5	77.6	99.9
2691028	1954	Frank B. Colton	First Oral Contraceptive	4	87.7	93.4	95.8	96.3	71.9
2699054	1955	Lloyd H. Conover	Tetracycline	38	93.4	95.9	96.1	96.0	99.6
2708656	1955	Enrico Fermi	Atomic Reactor	196	99.9	98.7	97.9	98.4	100.0
2708722	1955	An Wang	Magnetic Core Memory	76	36.1	79.2	90.1	93.1	99.9
2717437	1955	George De Mestral	Velcro	258	48.6	51.9	46.9	60.5	100.0
2816721	1957	R. J. Taylor	Rocket Engine	25	82.1	80.1	80.9	80.6	98.7
2835548	1958	Robert C. Baumann	Satellite	16	82.9	88.3	87.4	79.6	96.1
2866012	1958	Charles P. Ginsburg	Video Tape Recorder	30	78.5	85.2	88.9	89.5	99.2
2879439	1959	Charles H. Townes	Maser	22	72.4	81.0	82.9	79.5	98.1
2929922	1960	Arthur L. Shawlow	Laser	122	79.5	89.4	90.0	89.1	100.0
2956114	1960	Charles P. Ginsburg	Wideband Magnetic Tape	11	67.9	77.2	82.6	85.7	92.1
2981877	1961	Robert N. Noyce	Semiconductor Device	152	96.1	97.6	97.7	96.8	100.0
3093346	1963	Maxime A. Faget	First Manned Space Capsule	19	87.7	94.0	93.5	85.0	97.4
3118022	1964	Gerhard M. Sessler	Electret Microphone	39	66.7	79.2	78.8	72.2	99.6
3156523	1964	Glenn T. Seaborg	Americium (Element 95)	1	93.0	88.7	91.5	94.7	41.7

Table A.7: Firm-level data — descriptive statistics

	A. All patenting firms							
	Observations	Mean	sd	p10	p25	p50	p75	p90
Tobin's Q	94433	2.08	3.04	0.82	1.00	1.33	2.11	3.73
Book Assets (USDb)	94433	4.58	44.80	0.01	0.04	0.15	0.88	4.62
SRD, R&D Stock (USDb)	94433	0.25	1.69	0.00	0.00	0.01	0.05	0.23
SPAT, Patent Stock (1000's)	94433	0.10	0.59	0.00	0.00	0.01	0.02	0.14
SRSIM / SPAT, 0-1 yr	94433	0.23	0.04	0.19	0.20	0.22	0.24	0.27
SRSIM / SPAT, 0-5 yr	94433	1.23	0.34	0.95	1.02	1.15	1.32	1.57
SRSIM / SPAT, 0-10 yr	94433	2.61	1.02	1.80	2.02	2.38	2.90	3.68
SRSIM / SPAT, 0-20 yr	94433	5.25	3.03	2.37	3.76	4.63	6.09	8.48
SCIT / SPAT, 0-1 yr	94433	0.44	1.09	0.00	0.00	0.20	0.46	1.00
SCIT / SPAT, 0-5 yr	94433	4.72	9.74	0.71	1.34	2.43	4.94	9.65
SCIT / SPAT, 0-10 yr	94433	10.83	22.14	1.59	2.80	5.19	11.00	23.24
SCIT / SPAT, 0-20 yr	94433	18.66	39.12	3.00	4.88	9.00	18.16	39.00
	B. Manufacturing (2000-3999)							
	Observations	Mean	sd	p10	p25	p50	p75	p90
Tobin's Q	65795	2.04	2.88	0.81	1.00	1.34	2.11	3.68
Book Assets (USDb)	65795	2.18	11.82	0.01	0.03	0.13	0.66	3.09
SRD, R&D Stock (USDb)	65795	0.30	1.84	0.00	0.00	0.01	0.07	0.31
SPAT, Patent Stock (1000's)	65795	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SRSIM / SPAT, 0-1 yr	65795	0.23	0.04	0.19	0.20	0.22	0.24	0.26
SRSIM / SPAT, 0-5 yr	65795	1.20	0.29	0.94	1.01	1.14	1.30	1.50
SRSIM / SPAT, 0-10 yr	65795	2.51	0.87	1.77	2.00	2.34	2.84	3.48
SRSIM / SPAT, 0-20 yr	65795	5.05	2.66	2.28	3.71	4.55	5.95	8.05
SCIT / SPAT, 0-1 yr	65795	0.37	0.67	0.00	0.03	0.21	0.43	0.87
SCIT / SPAT, 0-5 yr	65795	3.92	5.33	0.78	1.36	2.39	4.51	8.26
SCIT / SPAT, 0-10 yr	65795	9.02	13.54	1.69	2.79	5.04	10.00	19.55
SCIT / SPAT, 0-20 yr	65795	15.77	25.73	3.00	4.83	8.74	16.50	33.02

Table reports descriptive statistics for the firm-level stock variables constructed in equation (17).

Table A.8: Market Value as a Function of R&D, Patents and Similarity Stocks (Manufacturing only)

$\log Q$	(1)	(2)	(3)	(4)
Horizon	(0,1)	(0,5)	(0,10)	(0,20)
$RD_{i,t}/A_{i,t}$	0.777***	0.878***	0.855***	1.126***
	(12.77)	(11.99)	(10.31)	(6.64)
$PAT_{i,t}/RD_{i,t}$	-0.014	-0.013	-0.063	0.066
	(-0.51)	(-0.60)	(-0.94)	(0.42)
$CIT_{i,t}/PAT_{i,t}$	0.163***	0.056***	0.027***	0.020***
	(5.76)	(11.02)	(11.29)	(9.46)
$RSIM_{i,t}/PAT_{i,t}$	5.014***	0.686***	0.247***	0.094***
	(11.32)	(9.79)	(9.35)	(8.83)
$D(R\&D = 0)$	-0.179***	-0.139***	-0.113***	-0.0490***
	(-13.39)	(-10.40)	(-8.24)	(-3.27)
Normalized coefficients: $CIT_{i,t}/PAT_{i,t}$	0.110	0.308	0.370	0.275
Normalized coefficients: $RSIM_{i,t}/PAT_{i,t}$	0.182	0.199	0.219	0.232
N	65,229	57,800	48,943	30,232
R^2	0.491	0.496	0.478	0.362

t statistics in parentheses

* $p < .1$, ** $p < .05$, *** $p < .01$

Table reports estimates of equation (19) in the text, restricted to the sample of manufacturing firms (SIC 2000-3999) . 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 ($SRSIM_{f,t}$) — constructed as in (17) 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.

Figure A.1: Index of technological progress and future productivity growth

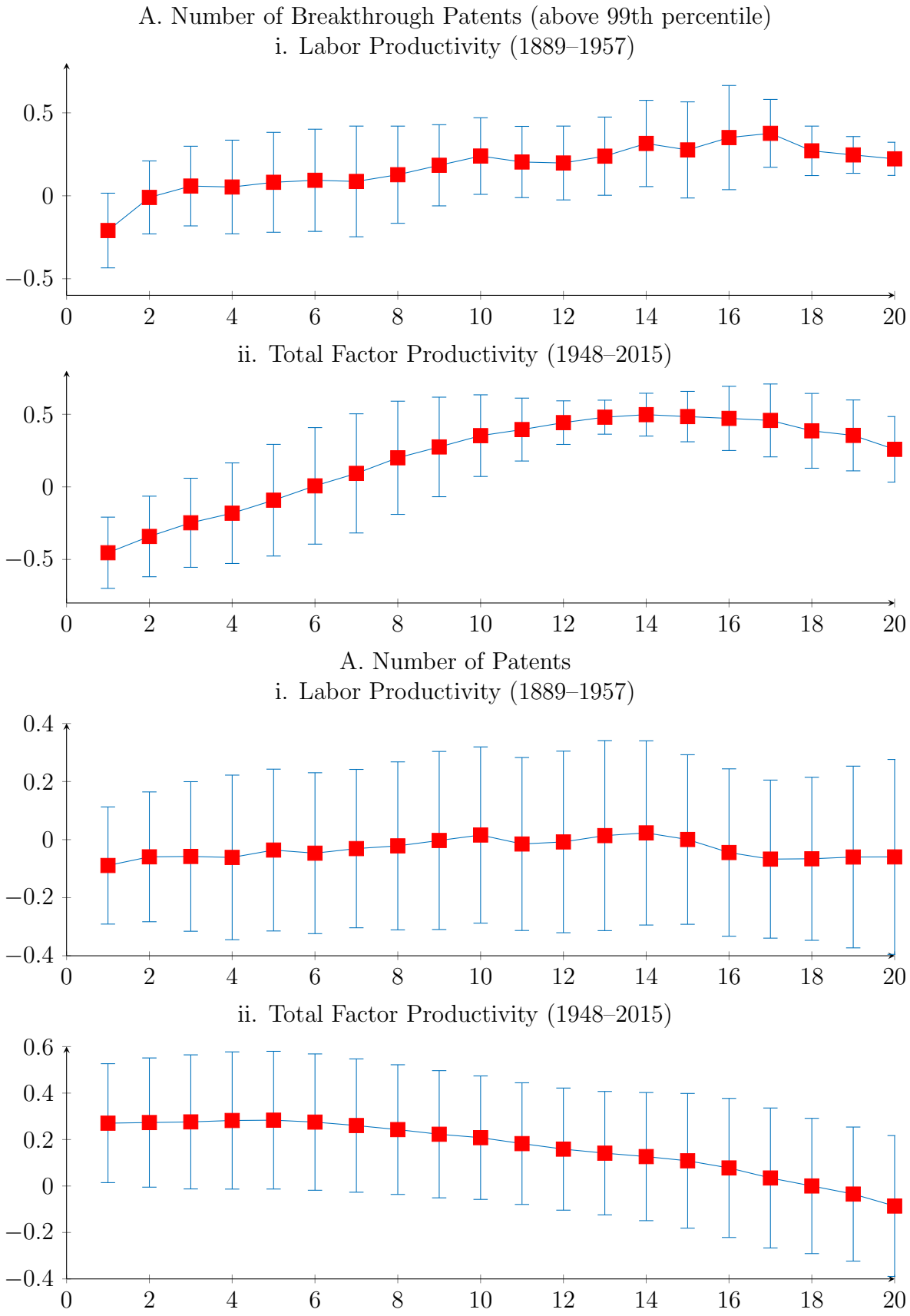


Figure A.2: Index of technological progress and future productivity growth

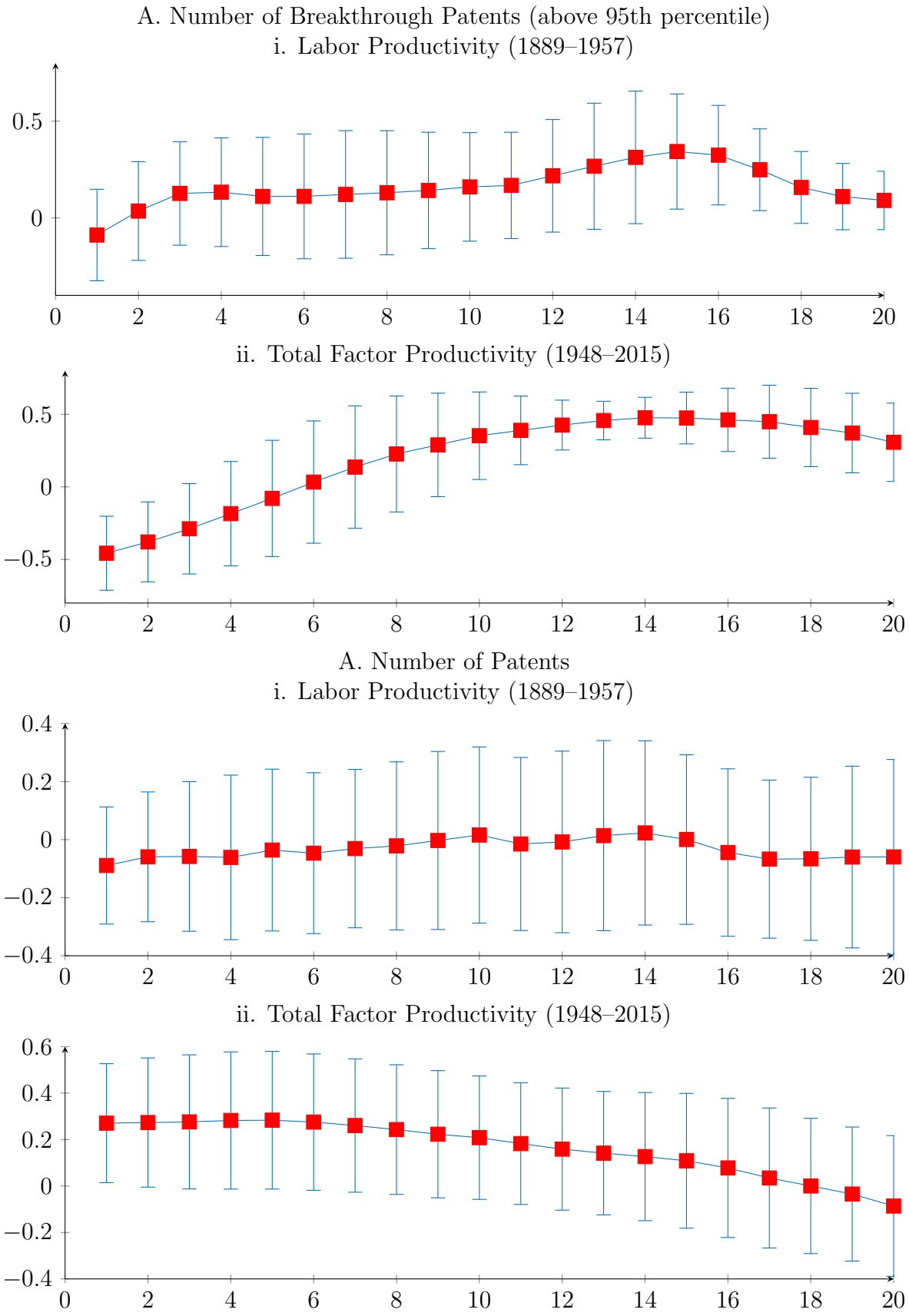


Figure A.3: Index of technological progress and future productivity growth